THE CRITICS HAVE RAVED, AND I’M SURE THEY MUST HAVE BEEN TALKING ABOUT THIS BOOK:

THE PHENOMENAL #1 BEST SELLER!

THE BOOK EVERYBODY’S TALKING ABOUT!

NOW A MAJOR MOTION PICTURE

“Taut, fast, and relentless . . . A ride worth taking. Keeps the reader hooked . . . from the gripping first chapters . . . to the vise-tightening midsection and on to the sensational finale.”

Name withheld by request
.1. Preface

This manuscript represents my effort to produce a comprehensive text on real analysis, in a new organizational format possible only in an online book. The project is far from complete; indeed, it could be described as still being in its infancy. It will continue to evolve and grow in the future.

I have tried to take advantage of the increased organizational flexibility available for an online manuscript. A printed book must be organized linearly, but an online book need not be. Hyperlinks can make it possible to navigate through the various sections in a complicated and organic way. Supplemental material can be included without interfering with the flow of the core material (which can itself be varied according to the interests of the reader). While I am certainly not the first author to think of trying to take advantage of this flexibility in organizing an online book, I have been surprised to find that the possibilities have been largely unexplored so far, and the potentially revolutionary benefits of flexible organization online have been little exploited. Even \( \LaTeX \) (or at least \( \LaTeX \)), which has become the universally-used text editing system for mathematics and related fields (and which I have used for this manuscript), is too inflexibly oriented toward producing linearly-organized printed documents. I hope my approach to organization of this manuscript will help encourage other attempts to develop the possibilities. See () for more detailed comments about organization.

This book is intended as a supplement, not a replacement, for the many fine books on this subject already available. I have consulted existing texts and references extensively, and they have collectively been very influential in both the general approach and specific details of my treatment. I have tried to be careful in giving credit where credit is due; however, in a subject as generally mature as real analysis, there is little originality in the actual mathematics of most books (including this one) – the primary contributions of each author are the exposition and organization, and decisions on what material to emphasize and what to leave out (the online organization of this book has made the decision criteria considerably different for me than for authors of printed books).

No part of this manuscript should be regarded as the last word, or even my last word, on the subject. My approach on many topics is pretty standard, and in a lot of cases will probably remain so; much of this subject has been quite refined for a long time. But my approach to some parts of the subject has changed already during the project (frankly, I have been forced to think critically about the treatment of some topics for the first time in my career), and this process will surely continue.

I regard myself as a competent mathematician who is reasonably qualified for a project of this sort, and I think I have learned something about the right way to present this subject in my 35 years of experience in teaching analysis at the university level. But I am in no way uniquely or even especially well qualified; I think there are many other mathematicians who could do at least as good a job, in some cases much better, on all or parts of the project. The only difference is that I actually have done it.

I expect and welcome comment and criticism concerning my efforts, either through personal communication or public statements. I will take all (informed) criticism to heart, and it will perhaps result in improvements in the manuscript. Certainly simple corrections to errors and misstatements (of which there are likely to be a substantial number despite my best efforts to be careful) will be greatly appreciated and quickly acted on. Broader criticism of the text is also welcome. However, I issue the following challenge to anyone who is dissatisfied with what is here: don’t just criticize me, try to do better yourself. I do not regard myself as being in competition with anyone; my goal is simply to have the best treatment of this subject available to the community at large. Criticism of my manuscript makes a small contribution to this goal, but anything you write which is better than what I write benefits us all much more.
# Contents

.1. Preface ................................................................. ii

### I Introduction

1. Calculus and Analysis ........................................... 1
2. Differential Calculus ........................................... 2
3. Measure and Integration ........................................ 2
   3.1. Measure ......................................................... 2
   3.2. Integration ................................................... 6
4. Other Topics ....................................................... 7
5. Rigor in Mathematics ........................................... 8
   5.1. Semantics: What is Obvious? ................................. 14
   5.2. Set Theory and Logic .......................................... 38
   5.3. Proofs vs. Formal Proofs (or, How Mathematics Really Works) .................. 40
   5.4. How to Understand a Proof .................................. 43
   5.5. Proof by Contradiction ...................................... 43
   5.6. Proofs and Counterexamples: Mathematics as Creative Activity .................. 45
   5.7. Scholarship vs. Practice ..................................... 55
   5.8. Is Mathematics Created or Discovered? ...................... 55
   5.9. Analogies for the Structure of Mathematics ................ 56
   5.10. Pure vs. Applied Mathematics ............................... 60
   5.11. Is Mathematics an Art or a Science? ....................... 63
6. The Foundations of Mathematics .................................. 67
   6.1. Philosophies of Mathematics .................................. 67
   6.2. Mathematical Theories ......................................... 91
   6.3. Languages ..................................................... 91
   6.4. Logic .......................................................... 92
   6.5. Mathematical Structures and Models ........................ 92
7. Printed or Online? .................................................. 93
   7.1. How Much Generality? ......................................... 95
   7.2. History vs. Logic ............................................... 96
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>II.9. Cardinals</td>
<td>184</td>
</tr>
<tr>
<td>II.9.1. Cardinals as Ordinals</td>
<td>184</td>
</tr>
<tr>
<td>II.9.2. General Cardinals and Alephs</td>
<td>186</td>
</tr>
<tr>
<td>II.9.3. Cardinal Arithmetic</td>
<td>187</td>
</tr>
<tr>
<td>II.9.4. The Role of the Axiom of Choice</td>
<td>188</td>
</tr>
<tr>
<td>II.9.5. The Continuum Hypothesis</td>
<td>191</td>
</tr>
<tr>
<td>II.9.6. Large Cardinals</td>
<td>192</td>
</tr>
<tr>
<td>II.9.7. Exercises</td>
<td>193</td>
</tr>
<tr>
<td>II.10. Measurable Cardinals</td>
<td>198</td>
</tr>
<tr>
<td>II.10.1. Introduction</td>
<td>198</td>
</tr>
<tr>
<td>II.10.2. Real-Valued Measurable Cardinals</td>
<td>199</td>
</tr>
<tr>
<td>II.10.3. Atomless Full Real-Valued Measures</td>
<td>201</td>
</tr>
<tr>
<td>II.10.4. Measurable Cardinals</td>
<td>203</td>
</tr>
<tr>
<td>II.10.5. Consistency and Equiconsistency Results</td>
<td>207</td>
</tr>
<tr>
<td>II.11. Categories and Functors</td>
<td>209</td>
</tr>
<tr>
<td>II.11.1. Exercises</td>
<td>211</td>
</tr>
<tr>
<td>III. The Numbers of Analysis</td>
<td>214</td>
</tr>
<tr>
<td>III.1. The Real Numbers – Axiomatic Approach</td>
<td>214</td>
</tr>
<tr>
<td>III.1.1. The Algebraic Axioms of $\mathbb{R}$</td>
<td>215</td>
</tr>
<tr>
<td>III.1.2. The Order Axioms for $\mathbb{R}$</td>
<td>218</td>
</tr>
<tr>
<td>III.1.3. Absolute Value</td>
<td>220</td>
</tr>
<tr>
<td>III.1.4. The Natural Numbers</td>
<td>220</td>
</tr>
<tr>
<td>III.1.5. Induction and Recursion</td>
<td>222</td>
</tr>
<tr>
<td>III.1.6. The Integers</td>
<td>227</td>
</tr>
<tr>
<td>III.1.7. The Rational Numbers</td>
<td>231</td>
</tr>
<tr>
<td>III.1.8. The Number Line</td>
<td>232</td>
</tr>
<tr>
<td>III.1.9. Suprema, Infima, and the Completeness Axiom</td>
<td>233</td>
</tr>
<tr>
<td>III.1.10. The Archimedean Property and Density of $\mathbb{Q}$</td>
<td>238</td>
</tr>
<tr>
<td>III.1.11. Uniqueness of $\mathbb{R}$</td>
<td>239</td>
</tr>
<tr>
<td>III.1.12. Existence of Roots</td>
<td>243</td>
</tr>
<tr>
<td>III.1.13. Exercises</td>
<td>245</td>
</tr>
<tr>
<td>III.2. The Natural Numbers</td>
<td>247</td>
</tr>
<tr>
<td>III.2.1. Introduction</td>
<td>247</td>
</tr>
<tr>
<td>III.2.2. Axiomatic Characterization of $\mathbb{N}$</td>
<td>247</td>
</tr>
<tr>
<td>III.2.3. Induction and Recursion</td>
<td>251</td>
</tr>
<tr>
<td>III.2.4. Algebraic Structure of $\mathbb{N}$</td>
<td>252</td>
</tr>
<tr>
<td>III.2.5. Algebraic Axiomatization of $\mathbb{N}$</td>
<td>255</td>
</tr>
<tr>
<td>III.2.6. Order Structure of $\mathbb{N}$</td>
<td>257</td>
</tr>
<tr>
<td>III.2.7. Subtraction and Division in $\mathbb{N}$</td>
<td>259</td>
</tr>
<tr>
<td>III.2.8. Axiomatization of $\mathbb{N}_0$</td>
<td>260</td>
</tr>
<tr>
<td>III.2.9. Further Properties of $\mathbb{N}$</td>
<td>261</td>
</tr>
<tr>
<td>III.2.10. Summary</td>
<td>264</td>
</tr>
<tr>
<td>III.2.11. Exercises</td>
<td>265</td>
</tr>
<tr>
<td>III.3. The Integers</td>
<td>265</td>
</tr>
<tr>
<td>III.3.1. Introduction and Axiomatization of $\mathbb{Z}$</td>
<td>265</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>III.3.2</td>
<td>Models for $\mathbb{Z}$</td>
</tr>
<tr>
<td>III.3.3</td>
<td>Uniqueness of $\mathbb{Z}$</td>
</tr>
<tr>
<td>III.3.4</td>
<td>Ordering and Subtraction in $\mathbb{Z}$</td>
</tr>
<tr>
<td>III.3.5</td>
<td>Multiplication and Division in $\mathbb{Z}$</td>
</tr>
<tr>
<td>III.3.6</td>
<td>Ring Axiomatization of $\mathbb{Z}$</td>
</tr>
<tr>
<td>III.3.7</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.4.1</td>
<td>Introduction and Axiomatization of $\mathbb{Q}$</td>
</tr>
<tr>
<td>III.4.2</td>
<td>The Standard Model of $\mathbb{Q}$</td>
</tr>
<tr>
<td>III.4.3</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.5.1</td>
<td>Construction of $\mathbb{R}$: the Dedekind Cut Approach</td>
</tr>
<tr>
<td>III.5.2</td>
<td>Construction of $\mathbb{R}$: the Cauchy Sequence Approach</td>
</tr>
<tr>
<td>III.5.3</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.6.1</td>
<td>Axiomatization of $\mathbb{C}$</td>
</tr>
<tr>
<td>III.6.2</td>
<td>The Standard Model for $\mathbb{C}$: the Complex Plane</td>
</tr>
<tr>
<td>III.6.3</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.7.2</td>
<td>Compactness and Continuity</td>
</tr>
<tr>
<td>III.7.3</td>
<td>The Creeping Lemma</td>
</tr>
<tr>
<td>III.7.4</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.8.1</td>
<td>Decimal Expansions</td>
</tr>
<tr>
<td>III.8.2</td>
<td>Continued Fractions</td>
</tr>
<tr>
<td>III.8.3</td>
<td>Measure Theory of Continued Fractions</td>
</tr>
<tr>
<td>III.8.4</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.9.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>III.9.2</td>
<td>Irrationality of $\pi$</td>
</tr>
<tr>
<td>III.9.3</td>
<td>Transcendence of $e$ and Related Numbers</td>
</tr>
<tr>
<td>III.9.4</td>
<td>Twentieth Century Results</td>
</tr>
<tr>
<td>III.9.5</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.10.1</td>
<td>Ordered Fields Containing $\mathbb{R}$</td>
</tr>
<tr>
<td>III.10.2</td>
<td>The Hyperreal Numbers via Ultrafilters</td>
</tr>
<tr>
<td>III.10.3</td>
<td>Internal and External Sets and Functions</td>
</tr>
<tr>
<td>III.10.4</td>
<td>The Transfer Principle</td>
</tr>
<tr>
<td>III.10.5</td>
<td>Other Nonstandard Structures</td>
</tr>
<tr>
<td>III.10.6</td>
<td>Proofs and Applications of Nonstandard Analysis</td>
</tr>
<tr>
<td>III.10.7</td>
<td>Loeb Measures and a Nonstandard Approach to Measure Theory</td>
</tr>
<tr>
<td>III.10.8</td>
<td>Exercises</td>
</tr>
<tr>
<td>III.11.1</td>
<td>Matrix Algebra</td>
</tr>
<tr>
<td>III.11.2</td>
<td>Vector Spaces</td>
</tr>
</tbody>
</table>

vi
V Calculus

V.1. Limits of Functions .................................................. 512
  V.1.1. Ordinary Limits ................................................ 512
  V.1.2. The Sequential Criterion ...................................... 516
  V.1.3. Limit Theorems and Indeterminate Forms ..................... 516
  V.1.4. One-Sided Limits .............................................. 516
  V.1.5. Extended Limits ............................................... 516
  V.1.6. Limit Superior and Limit Inferior ............................ 516
  V.1.7. Exercises ..................................................... 516

V.2. Continuity on \( \mathbb{R} \) .............................................. 517
  V.2.1. Continuity at a Point ......................................... 517
  V.2.2. Continuity on a Set .......................................... 517
  V.2.3. The Intermediate Value Theorem .............................. 519
  V.2.4. The Max-Min Theorem ........................................ 521
  V.2.5. Continuity and Oscillation .................................... 522

V.3. Differentiation and Derivatives ..................................... 524
  V.3.1. The Derivative ................................................ 524
  V.3.2. Elementary Rules of Differentiation ......................... 530
  V.3.3. The Product and Quotient Rules ............................... 532
  V.3.4. The Chain Rule ............................................... 535
  V.3.5. Derivatives of Inverse Functions ............................. 538
  V.3.6. Critical Points and Local Extrema ............................ 538
  V.3.7. Higher Derivatives .......................................... 541
  V.3.8. Continuous but Nondifferentiable Functions ................. 542
  V.3.9. The Bush-Wunderlich-Swift Example ........................... 544
  V.3.10. Exercises .......................................................... 549
  V.3.11. Calculus of Finite Differences ............................... 552

V.4. Extensions of the Derivative ....................................... 557
  V.4.1. One-Sided Derivatives ...................................... 557
  V.4.2. Dini Derivatives .............................................. 558
  V.4.3. Symmetric First and Second Derivatives ..................... 562
  V.4.4. Approximate and Weak Derivatives ............................ 562
  V.4.5. Exercises ..................................................... 562

V.5. Exponential and Trigonometric Functions ........................ 563
  V.5.1. Rational Exponents ........................................... 563
  V.5.2. Irrational Exponents and Exponential Functions ............. 564
  V.5.3. The Number \( e \) .............................................. 566
  V.5.4. Derivatives of Exponential and Logarithm Functions ....... 568
  V.5.5. Exercises ..................................................... 569

V.6. Trigonometric Functions ............................................. 571
  V.6.1. Arc Length of a Circle and the Arcsine Function .......... 571
  V.6.2. The Sine and Cosine Functions ................................ 572
  V.6.3. Other Trigonometric Functions ................................ 574
  V.6.4. The Inverse Trigonometric Functions ......................... 575
  V.6.5. The Area of a Circle ......................................... 576
  V.6.6. The Hyperbolic Functions .................................... 578
<table>
<thead>
<tr>
<th>VIII Curves</th>
<th>781</th>
</tr>
</thead>
<tbody>
<tr>
<td>VII.1 Curves and Parametrized Curves in ( \mathbb{R}^n )</td>
<td>782</td>
</tr>
<tr>
<td>VII.1.1 Parametrized Curves</td>
<td>782</td>
</tr>
<tr>
<td>VII.1.2 Reparametrization</td>
<td>783</td>
</tr>
<tr>
<td>VII.1.3 Paths</td>
<td>784</td>
</tr>
<tr>
<td>VII.2 Smooth and Piecewise Smooth Curves in ( \mathbb{R}^n )</td>
<td>784</td>
</tr>
<tr>
<td>VII.3 Rectifiable Curves and Arc Length</td>
<td>784</td>
</tr>
<tr>
<td>VII.3.1 Arc Length</td>
<td>784</td>
</tr>
<tr>
<td>VII.3.2 Rectifiable Curves</td>
<td>788</td>
</tr>
<tr>
<td>VII.3.3 Length of Arcs</td>
<td>794</td>
</tr>
<tr>
<td>VII.3.4 Exercises</td>
<td>795</td>
</tr>
<tr>
<td>VII.4 Curves in ( \mathbb{R}^2 )</td>
<td>795</td>
</tr>
<tr>
<td>VII.5 Curves in ( \mathbb{R}^3 )</td>
<td>795</td>
</tr>
<tr>
<td>VII.5.1 Exercises</td>
<td>795</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VIII Multivariable and Vector Calculus</th>
<th>796</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIII. Limits and Continuity for Vector Functions</td>
<td>796</td>
</tr>
<tr>
<td>VIII.1 Graphs of Functions of Several Variables</td>
<td>797</td>
</tr>
<tr>
<td>VIII.1.2 Limits of Sequences in ( \mathbb{R}^n )</td>
<td>797</td>
</tr>
<tr>
<td>VIII.1.3 Multivariable Limits</td>
<td>798</td>
</tr>
<tr>
<td>VIII.2 Partial and Directional Derivatives</td>
<td>802</td>
</tr>
<tr>
<td>VIII.2.1 Partial Derivatives</td>
<td>802</td>
</tr>
<tr>
<td>VIII.2.2 Equality of Mixed Partials</td>
<td>804</td>
</tr>
<tr>
<td>VIII.2.3 Directional Derivatives</td>
<td>809</td>
</tr>
<tr>
<td>VIII.2.4 Exercises</td>
<td>810</td>
</tr>
<tr>
<td>VIII.3 Differentiability of Multivariable Functions</td>
<td>811</td>
</tr>
<tr>
<td>VIII.3.1 Differentiability and the Derivative</td>
<td>811</td>
</tr>
<tr>
<td>VIII.3.2 A Criterion for Differentiability</td>
<td>814</td>
</tr>
<tr>
<td>VIII.3.3 The Chain Rule</td>
<td>816</td>
</tr>
<tr>
<td>VIII.3.4 The Gradient and Tangent Planes</td>
<td>818</td>
</tr>
<tr>
<td>VIII.3.5 Functional Independence</td>
<td>821</td>
</tr>
<tr>
<td>VIII.3.6 Exercises</td>
<td>823</td>
</tr>
<tr>
<td>VIII.4 Higher-Order Derivatives and Critical Point Analysis</td>
<td>825</td>
</tr>
<tr>
<td>VIII.4.1 Higher-Order Differentiability</td>
<td>825</td>
</tr>
<tr>
<td>VIII.4.2 Taylor's Theorem</td>
<td>826</td>
</tr>
<tr>
<td>VIII.4.3 Critical Point Analysis</td>
<td>826</td>
</tr>
<tr>
<td>VIII.4.4 Exercises</td>
<td>826</td>
</tr>
<tr>
<td>VIII.5 Vector Analysis</td>
<td>827</td>
</tr>
<tr>
<td>VIII.5.1 Surfaces in ( \mathbb{R}^3 )</td>
<td>827</td>
</tr>
<tr>
<td>VIII.5.2 Divergence and Curl</td>
<td>827</td>
</tr>
<tr>
<td>VIII.6 The Implicit and Inverse Function Theorems</td>
<td>828</td>
</tr>
<tr>
<td>VIII.6.1 Statements of the Theorems</td>
<td>828</td>
</tr>
<tr>
<td>VIII.6.2 Twin Theorems and a Generalization</td>
<td>834</td>
</tr>
<tr>
<td>VIII.6.3 Proof of the Implicit Function Theorem, Case ( n = 1 )</td>
<td>837</td>
</tr>
<tr>
<td>VIII.6.4 Proof of the General Implicit Function Theorem</td>
<td>847</td>
</tr>
<tr>
<td>VIII.6.5 Proof of the Inverse Function Theorem</td>
<td>851</td>
</tr>
</tbody>
</table>
### VIII.6.6. How Much Can the Theorems be Generalized?

853

### VIII.6.7. The Domain and Range Straightening Theorems

854

### VIII.6.8. Analytic Versions

859

### VIII.6.9. Exercises

863

### VIII.7. Power Series and Analytic Functions of Several Variables

868

### IX. Differential Equations

876

#### IX.1. Existence and Uniqueness of Solutions to Ordinary Differential Equations

878

- IX.1.1. Elementary Considerations and Terminology
- IX.1.2. Existence and Uniqueness Theorem for First-Order ODEs
- IX.1.3. Systems of First-Order Differential Equations
- IX.1.4. Higher-Order Differential Equations and Systems
- IX.1.5. Continuous Dependence on Initial Conditions
- IX.1.6. Peano’s Existence Theorem for First-Order ODEs
- IX.1.7. Power Series and Analytic Solutions
- IX.1.8. Linear ODEs
- IX.1.9. Exercises

#### IX.2. Partial Differential Equations

897

- IX.2.1. Elementary Theory
- IX.2.2. Quasilinear Partial Differential Equations
- IX.2.3. The Cauchy-Kovalevskaya Theorem

### X. Fundamentals of Complex Analysis

899

#### X.1. Elementary Theory

899

- X.1.1. Complex Differentiation and Holomorphic Functions
- X.1.2. Rules of Differentiation
- X.1.3. Branch Cuts and Branches
- X.1.4. The Fundamental Theorem of Algebra

#### X.2. Contour Integration

907

- X.2.1. Integration along a Contour

#### X.3. The Big Theorems

910

- X.3.1. Goursat’s Theorem
- X.3.2. The Antiderivative Theorem and Cauchy’s Theorem for Convex Sets
- X.3.3. The Cauchy Integral Formula
- X.3.4. The Homotopy Theorem for Contour Integrals
- X.3.5. Homology of Contours
- X.3.6. Cauchy’s Theorem, General Versions
- X.3.7. The Cauchy Integral Formula, General Versions

#### X.4. Sequences and Series of Holomorphic Functions; Analytic Functions

928

- X.4.1. Sequences and Series of Holomorphic Functions
- X.4.2. Power Series
- X.4.3. Inverse Power Series
- X.4.4. Analytic Functions
- X.4.5. Normal Families and Montel’s Theorem
X.5. Zeroes and Singularities .......................................................... 947
  X.5.1. Zeroes of Holomorphic Functions ...................................... 947
  X.5.2. Singularities ................................................................. 949
  X.5.3. Behavior at Infinity ....................................................... 955
  X.5.4. Meromorphic Functions ................................................... 957
  X.5.5. The Residue Theorem ..................................................... 957
  X.5.6. Principle of the Argument .............................................. 957
  X.6.1. Conformal Equivalence .................................................. 959
  X.6.2. Simply Connected Sets in $\mathbb{C}$ ................................ 959
  X.6.3. The Riemann Mapping Theorem ....................................... 960
X.7. Analytic Continuation and Riemann Surfaces ......................... 963
  X.7.1. Analytic Function Elements ........................................... 963
  X.7.2. The Riemann Surface of a Total Analytic Function ............... 964
  X.7.3. Analytic Continuation .................................................. 968
  X.7.4. The Regular Riemann Surface of a Function .......................... 971
  X.7.5. Branch Points ............................................................. 972
  X.7.6. Points at Infinity and Meromorphic Functions ...................... 973
  X.7.7. Germs of Analytic Functions ......................................... 975
  X.7.8. General Riemann Surfaces ............................................. 976
X.8. Analytic Functions of Several Variables ................................ 977
  X.8.1. Elementary Theory ..................................................... 977
  X.8.2. Exercises ................................................................. 981
X.9. Differential Algebra .......................................................... 986
  X.9.1. Derivations ............................................................... 986
  X.9.2. Differential Fields ..................................................... 989
  X.9.3. Integration in Finite Terms .......................................... 994
  X.9.4. Exercises ................................................................. 998

XI Topology ................................................................................. 1000
  XI.1. Topological Spaces .......................................................... 1003
    XI.1.1. Definitions .............................................................. 1003
    XI.1.2. Bases and Subbases .................................................... 1004
    XI.1.3. Covers and Subcovers ................................................ 1008
    XI.1.4. Relative Topology on a Subspace .................................. 1009
    XI.1.5. Separated Union ....................................................... 1011
    XI.1.6. Exercises ............................................................... 1012
  XI.2. Closed Sets and Limit Points ............................................ 1013
    XI.2.1. Closed Sets ............................................................. 1013
    XI.2.2. Limit Points of a Set ................................................ 1013
    XI.2.3. Closure of a Set ........................................................ 1014
    XI.2.4. Specification of a Topology via Closure .......................... 1018
    XI.2.5. Isolated Points and Derived Sets .................................. 1020
    XI.2.6. Exercises ............................................................... 1021
  XI.3. Convergence ................................................................. 1022
    XI.3.1. Limits of Sequences ................................................ 1022
XI.3.2. Nets and Limits .......................................................... 1024
XI.3.3. Specification of a Topology via Convergence .................. 1030
XI.3.4. Convergence via Filters .............................................. 1034
XI.3.5. Exercises ................................................................. 1038
XI.4. Continuous Functions .................................................... 1040
   XI.4.1. Definition and Properties ...................................... 1040
   XI.4.2. Topological Continuity vs. Metric Continuity .............. 1040
XI.5. Comparison of Topologies ............................................. 1041
   XI.5.1. Stronger and Weaker Topologies .............................. 1041
XI.6. Products and the Product Topology ................................ 1045
   XI.6.1. Exercises ........................................................... 1045
XI.7. Separation Axioms ...................................................... 1046
   XI.7.1. The $T$-Axioms .................................................. 1046
   XI.7.2. $T_{0}$-Spaces .................................................... 1047
   XI.7.3. $T_{1}$-Spaces .................................................... 1048
   XI.7.4. Hausdorff Spaces ............................................... 1049
   XI.7.5. Regular Spaces ................................................ 1049
   XI.7.6. Normal Spaces .................................................. 1050
   XI.7.7. Completely Regular Spaces .................................. 1053
   XI.7.8. Exercises .......................................................... 1054
XI.8. Quotient Spaces and the Quotient Topology ..................... 1060
   XI.8.1. The Quotient Topology ....................................... 1060
   XI.8.2. Semicontinuous Decompositions ................................ 1062
   XI.8.3. Separation Axioms for Quotient Spaces ...................... 1064
   XI.8.4. Quotient Metrics ............................................... 1067
   XI.8.5. Exercises ........................................................ 1068
XI.9. Identification Spaces .................................................. 1073
   XI.9.1. Collapsing a Subset to a Point ................................ 1073
   XI.9.2. Gluing and Attaching ........................................... 1075
   XI.9.3. Group Actions on Spaces ...................................... 1084
   XI.9.4. Local Homeomorphisms and Covering Maps ................. 1087
   XI.9.5. Exercises ........................................................ 1087
XI.10. Metrizable Spaces ..................................................... 1093
   XI.10.1. Metrizability .................................................. 1093
   XI.10.2A Universal Space ............................................... 1097
   XI.10.3. Polish Spaces .................................................. 1101
   XI.10.4. Exercises ...................................................... 1103
XI.11. Compactness ............................................................ 1107
   XI.11.1. Basic Definitions ............................................. 1107
   XI.11.2. Compactness and Convergence ................................ 1109
   XI.11.3. Compact Metrizable Spaces ................................... 1111
   XI.11.4. Separation Properties ....................................... 1114
   XI.11.5. Automatic Continuity and Comparison of Topologies .... 1115
   XI.11.6. Products of Compact Spaces and Tikhonov’s Theorem .... 1115
   XI.11.7. Variations on Compactness ................................... 1118
   XI.11.8. Locally Compact Spaces ...................................... 1120
XI.11.9 Compactifications .................................................. 1126
XI.11.10 Paracompactness and Partitions of Unity ..................... 1130
XI.11.11 Compactly Generated Spaces .................................... 1135
XI.11.12 Exercises ......................................................... 1136
XI.11.13 Dini's Theorem .................................................... 1141
XI.12 Baire Spaces and the Baire Category Theorem ..................... 1142
  XI.12.1 Meager Sets and Baire Spaces .................................. 1142
  XI.12.2 The Baire Category Theorem ................................... 1143
  XI.12.3 Meager Sets vs. Null Sets ..................................... 1145
  XI.12.4 The Baire Property ............................................ 1146
  XI.12.5 Exercises ....................................................... 1148
XI.13 Connectedness ........................................................ 1150
  XI.13.1 Connected Spaces ............................................... 1150
  XI.13.2 Continuous Images of Connected Sets ......................... 1152
  XI.13.3 Unions and Intersections ....................................... 1153
  XI.13.4 Components of a Space ........................................ 1154
  XI.13.5 Products of Connected Spaces ................................ 1155
  XI.13.6 Totally Disconnected and Zero-Dimensional Spaces ........ 1155
  XI.13.7 Locally Connected Spaces ..................................... 1158
  XI.13.8 Path-Connected Spaces ........................................ 1159
  XI.13.9 Exercises ....................................................... 1162
XI.14 Direct and Inverse Limits .......................................... 1166
  XI.14.1 Direct Systems and Direct Limits of Sets .................... 1166
  XI.14.2 Inverse Systems and Inverse Limits of Sets ................. 1170
  XI.14.3 Direct Limits in Topology .................................... 1174
  XI.14.4 Inverse Limits in Topology ................................... 1175
  XI.14.5 Exercises ....................................................... 1181
XI.15 Uniform Structures .................................................. 1185
  XI.15.1 Topologies and Uniform Structures Defined by Pseudometrics 1185
  XI.15.2 General Uniform Structures ................................... 1190
XI.16 Order Topology ........................................................ 1191
  XI.16.1 Definition and Properties ..................................... 1191
  XI.16.2 Completion of an Order Topology .............................. 1191
  XI.16.3 Function Space Notation and the Compact-Open Topology ... 1192
  XI.16.4 Exercises ....................................................... 1196
XI.17 Cantor Sets ............................................................. 1197
  XI.17.1 The Cantor Set ................................................ 1197
  XI.17.2 Other Cantor Sets ............................................. 1197
  XI.17.3 Exercises ....................................................... 1200
XI.18 The Topology of $\mathbb{R}^n$ ....................................... 1201
  XI.18.1 Invariance of Domain and the Jordan Curve Theorem ........ 1201
  XI.18.2 $\mathbb{R}^n$ is $n$-Dimensional ................................. 1205
  XI.18.3 When is $\mathbb{R}^n$ “Similar” to $\mathbb{R}^m$? .................. 1218
  XI.18.4 Compact Subsets of $\mathbb{R}^n$ ................................. 1221
  XI.18.5 Exercises ....................................................... 1250
XI.19 Manifolds .............................................................. 1260
XIV Integration

XIV.1 Introduction ................................. 1481
  XIV.1.1 Origin of Integration Theories ........ 1481
  XIV.1.2 The Cauchy Integral for Continuous Functions 1483
  XIV.1.3 The Riemann Integral .................. 1486
  XIV.1.4 The Lebesgue Approach ................ 1487
  XIV.1.5 The Daniell Approach ................. 1492
  XIV.1.6 Direct Generalizations of Riemann Integration 1493
  XIV.1.7 Oriented and Unoriented Integration .... 1494
XIV.2 The Riemann Integral ................... 1496
  XIV.2.1 Partitions, Tags, Riemann and Darboux Sums 1496
  XIV.2.2 Step Functions and their Integrals .... 1504
  XIV.2.3 Definitions of the Riemann Integral ... 1504
  XIV.2.4 Comparing the Definitions ............ 1508
  XIV.2.5 Integrals of Continuous and Monotone Functions 1514

XIII.3.6 The Suslin Operation ................ 1432
XIII.3.7 Measures on Metric Spaces ............. 1436
  XIII.3.8 Radon Measures on $\mathbb{R}$ ........ 1436
  XIII.3.9 Lebesgue Measure and Radon Measures on $\mathbb{R}^n$ 1436
  XIII.3.10 Nonmeasurable Subsets of $\mathbb{R}$ 1436
  XIII.3.11 Exercises .......................... 1439
XIII.4 Metric Outer Measures and Borel Measures 1442
XIII.4.1 Metric Outer Measures and Borel Measures 1442
  XIII.4.2 Method II Construction of Outer Measures 1443
  XIII.4.3 Hausdorff Measure .................... 1446
  XIII.4.4 Hausdorff Measure on $\mathbb{R}^n$ ..... 1449
  XIII.4.5 Hausdorff Measure and Hausdorff Dimension for Self-Similar Sets 1450
  XIII.4.6 Packing Measure and other Geometric Measures 1450
  XIII.4.7 Radon Measures on Metric Spaces .... 1453
  XIII.4.8 Exercises ........................... 1453
XIII.5 Product Spaces ......................... 1454
  XIII.5.1 Product $\sigma$-Algebras ............. 1454
  XIII.5.2 Product Measures .................... 1455
  XIII.5.3 Lebesgue Measure on $\mathbb{R}^n$ .... 1458
  XIII.5.4 Infinite Product Spaces ............. 1458
  XIII.5.5 Exercises .......................... 1460
XIII.6 Signed and Complex Measures .......... 1462
  XIII.6.1 Signed Measures ..................... 1462
  XIII.6.2 Complex Measures .................... 1467
  XIII.6.3 Integration With Respect to Signed and Complex Measures 1469
  XIII.6.4 Exercises .......................... 1471
XIII.7 Limits of Measures ...................... 1473
  XIII.7.1 General Considerations ............. 1473
  XIII.7.2 Monotone and Dominated Convergence 1474
  XIII.7.3 The Nikodym and Vitali-Hahn-Saks Theorems 1476
  XIII.7.4 Exercises .......................... 1479

XIV Integration

XIV.1 Introduction ................................. 1481
  XIV.1.1 Origin of Integration Theories ........ 1481
  XIV.1.2 The Cauchy Integral for Continuous Functions 1483
  XIV.1.3 The Riemann Integral .................. 1486
  XIV.1.4 The Lebesgue Approach ................ 1487
  XIV.1.5 The Daniell Approach ................. 1492
  XIV.1.6 Direct Generalizations of Riemann Integration 1493
  XIV.1.7 Oriented and Unoriented Integration .... 1494
XIV.2 The Riemann Integral ................... 1496
  XIV.2.1 Partitions, Tags, Riemann and Darboux Sums 1496
  XIV.2.2 Step Functions and their Integrals .... 1504
  XIV.2.3 Definitions of the Riemann Integral ... 1504
  XIV.2.4 Comparing the Definitions ............ 1508
  XIV.2.5 Integrals of Continuous and Monotone Functions 1514

xviii
Chapter I

Introduction

“I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes.”

*Joseph Louis Lagrange*

This chapter is intended as an informal overview of the subject developed in this book. We will describe in general terms what the theory of real analysis is all about, and how it relates to the rest of mathematics.

I.1. Calculus and Analysis

“By 1800 mathematicians and physicists had developed calculus into an indispensable tool for the study of the natural world, and the problems that arose from this connection led to a wealth of new concepts and methods – for example, ways to solve differential equations – that made calculus one of the richest and hottest research areas in the whole of mathematics. The beauty and power of calculus had become undeniable. However, Bishop Berkeley’s criticisms of its logical basis remained unanswered, and as people began to tackle more sophisticated topics, the whole edifice started to look decidedly wobbly. The early cavalier use of infinite series, without regard to their meaning, produced nonsense as well as insights. The foundations of Fourier analysis were non-existent, and different mathematicians were claiming proofs of contradictory theorems. Words like ‘infinitesimal’ were being bandied about without being defined; logical paradoxes abounded; even the meaning of the word ‘function’ was in dispute. Clearly these unsatisfactory circumstances could not go on indefinitely.

Sorting it all out took a clear head, and a willingness to replace intuition by precision, even if there was a cost in comprehensibility. The main players were Bernard Bolzano, Cauchy, Niels Abel, Peter Dirichlet, and, above all, Weierstrass. Thanks to their efforts, by 1900 even the most complicated manipulations of series, limits, derivatives and integrals could be carried out safely, accurately and without paradoxes. A new subject was created: analysis. Calculus became one core aspect of analysis, but more subtle and more basic concepts, such as continuity and

\[^{1}\text{http://www.math.okstate.edu/~wli/teach/fmq.html}\]
limits, took logical precedence, underpinning the ideas of calculus. Infinitesimals were banned, completely.”

Ian Stewart

Calculus, as learned by undergraduates, roughly develops into two distinct, but interconnected, branches of mathematics. One branch, differential calculus, leads to the subject of differential geometry; the other, integral calculus, is the starting point for the subject of measure and integration. Subjects such as Complex Analysis and Ordinary and Partial Differential Equations are also based on calculus, but cannot be cleanly categorized in either branch. The catchall term analysis is used in mathematics to loosely include all subjects which are outgrowths of calculus, covering such a broad and diverse spectrum of mathematics as to defy a good definition, although it roughly means those parts of mathematics which involve infinite processes and limits; it is also often described as the mathematics of change or of continuous processes. (The word “analysis” in nontechnical English has a different and even broader meaning, which would include most of mathematics.)

I.2. Differential Calculus

I.3. Measure and Integration

Measure and integration, besides being a rich and beautiful theory in its own right, is an indispensable tool in most parts of pure and applied mathematics. To the working mathematician, measure and integration are as fundamental a tool as calculus and linear and abstract algebra. Whole parts of mathematics, including Probability Theory and Functional Analysis, are based primarily on the theory of measure and integration; other subjects, such as the modern theory of Ordinary and Partial Differential Equations, also owe their foundations to this theory.

Either measure or integration can be taken as the fundamental concept, and the other developed from it. It is somewhat a matter of taste which way to do this, but it is valuable to be able to think both ways since each approach is natural in some important contexts.

I.3.1. Measure

I.3.1.1. Measure theory, developed in the early twentieth century, gave the first mathematically satisfactory resolution to two foundational problems of analysis and geometry, one which caused great controversy especially in the sixteenth and seventeenth centuries in Europe (see [Ale13] for a study), and one which was only realized much later. (Measure theory later turned out to be good for a lot of other things too, such as formulating a rigorous theory of probability.)

The Paradox of Indivisibles

I.3.1.2. Should a line be regarded as a set of points? Any discussion of two lines intersecting at a point, drawing line through two given points, or especially considering the points between two given points on a line, all basic notions of Euclidean geometry, would seem to implicitly entail that certain points are at least “located” on a given line, almost (but not necessarily quite the same thing as) saying that the line is

\[?, \text{p. 149}\]
composed of a collection of points. Thus it is hard to imagine that it could be controversial that a line is (or can be thought of as) a collection of points, but it was.

I.3.1.3. Actually, we must first consider a more basic question: is a mathematical line (or plane, or three-dimensional space) something which exists in a well-defined and uniquely determined form, reflecting properties of physical space, as apparently believed by everyone until the nineteenth century, or is it a human mental construct which can be made in various ways, with properties depending on the way it is defined? If one believes the former, then observed properties of physical space place limits on what can be believed or contemplated about mathematical space.

I.3.1.4. The first problem to be considered is whether the number of points (“indivisibles”) on a line segment is finite or infinite. If there are only finitely many points on a line segment, either the points must have a positive diameter and touch each other, or there must be gaps between them. Either hypothesis seems untenable, and few mathematicians or even nonmathematicians have taken this possibility seriously; a mathematical line, or mathematical space, is very different in character from physical objects made up of a large but finite number of atoms. And nowadays (and gradually since DESCARTES) we identify points on a line with (real) numbers, and it is easily seen that between any two numbers there are infinitely many more, and in particular, there is no such thing as two consecutive numbers. In fact, in this picture of a line, every subinterval of a line segment, no matter how tiny, looks just like the whole segment, so a line is “infinitely divisible”; but contrary to what most if not all people from the sixteenth century believed, there is no logical contradiction in saying a line is infinitely divisible but that it is made up of points which are indivisible.

I.3.1.5. The other possibility is that a line segment is made up of infinitely many points (indivisibles). The next question then is whether for each point there is a “next” point, and if so whether the points “touch” or whether there is a gap between them. GALILEO thought there were infinitely many points, each of infinitesimal size, with gaps of infinitesimal length between them (he did not attempt to define what he meant by “infinitesimal”). In our picture of a line as the set of (real) numbers, there are clearly not consecutive points, and the questions of whether points on a line “touch” or whether there are “gaps” become very subtle if not meaningless (see (1)).

I.3.1.6. Whether or not there are consecutive points, if it is assumed that a line segment contains infinitely many points there is a basic paradox, which was one of the roots of the controversy over the theory of indivisibles. The length of the interval should be the sum of the sizes (diameters) of the points. Presumably all the points have the same character, and in particular have the same size. If the diameter of each is positive, the length of the interval is the sum of infinitely many equal positive numbers, which is infinite. And if the diameter of each is zero, then the length of the interval is the sum of infinitely many 0’s, so should be zero. How can the length be a finite positive number?

I.3.1.7. Mathematicians (and nonmathematicians) worried about this paradox for a long time, but through the development of Calculus it was glossed over, or if considered was resolved using the vague (at the time) notion of “infinitesimal numbers,” numbers which were positive but equal to 0 for practical purposes. It was not until measure theory was developed in the twentieth century that a satisfactory and rigorous solution to the paradox was found.
The key idea was the sensational discovery by Cantor in the late nineteenth century that there were different sizes of infinity, i.e. infinite sets could have different sizes. The smallest infinity is “countable infinity”: a set is countably infinite if its elements can be listed in a sequence. There are also “uncountable infinities” (actually many different ones!): a set is uncountable if it has “too many” elements to be listed in a sequence, i.e. any sequence of elements from the set necessarily leaves out some (“most”) elements of the set. Cantor showed not only that uncountable sets exist, but even that every interval of numbers is uncountable, i.e. every line segment contains uncountably many points. The resolution of the paradox is that while a sensible definition of the sum of countably many numbers can be made (this is the theory of infinite series), no sensible definition of the sum of uncountably many numbers is possible. Thus it cannot be sensibly said that the length of an interval is the (uncountable) sum of the (zero) diameters of the points.

The theory of Lebesgue measure is built on this resolution. A notion of “size” of subsets of $\mathbb{R}$ is defined, which has the property that the size of an interval is its length, and in particular the size of a point (singleton) is zero, and the size of a countable disjoint union of sets is the sum of the sizes of the sets. There is no strict numerical relationship between the size of an uncountable union of disjoint sets and the sizes of the sets themselves, so in particular there is no inconsistency between the size of an interval, which is an uncountable disjoint union of single points, and the sizes of the points. (Actually, it turns out that not all subsets of $\mathbb{R}$ have a well-defined “size” (Lebesgue measure), but the ones which do not are so bizarre they cannot even be explicitly described, and never come up in practice.)

A variation of the theory of indivisibles is to regard a plane figure as being “made up” of a collection of parallel line segments, or a solid region in $\mathbb{R}^3$ being made up of a collection of parallel slices, each a plane figure, although these line segments or slices are technically not “indivisible.” One has the same paradox as in the one-dimensional case: the area of the plane figure should be the sum of the areas of the line segments, and similarly in the solid case. The paradox is resolved in the same way via measure theory. But even before this can be done, another foundational problem must be resolved.

What Do Area and Volume Mean?

The area of a plane region, or the volume of a solid, seem to have a self-evident meaning. It apparently did not occur to anyone before the twentieth (or perhaps late nineteenth) century — not Euclid, not Archimedes, not Newton, not Euler, not Gauss, not Cauchy — that area or volume had to be defined: they were just there, and merely needed to be calculated. (This issue is on top of the more basic question discussed earlier of whether lines, planes, or space are themselves well defined.) Not only was it implicitly assumed that area (volume) was a meaningful concept, it was also assumed that it had some “obvious” properties: congruent regions have the same area, the area of a disjoint union is the sum of the areas of the components, the area of a rectangle is $\text{length} \times \text{width}$, the area of a curve is 0, and hence the area of the boundary of a region is 0, so the area of the union of “nonoverlapping” regions is the sum of the areas, etc. See I.5.1.1. for a more extensive discussion of the issue.

So how do we define the area of a region, and how do we show that area has the expected properties? Measure theory gives a satisfactory answer. We first define the area of a rectangle to be $\text{length} \times \text{width}$, and then by a somewhat involved process extend the definition to sets which can be built up from rectangles by countably many set operations. As in the case of Lebesgue measure on the line, not every subset of the plane has a well-defined area by this process, but all “reasonable” subsets do, and in
particular all the sets considered during the long development of geometry and calculus (going back at least to Euclid and Archimedes) have an area by this definition, which agrees with the intuitive areas already calculated by algebraic and calculus methods. The same procedure gives a rigorous definition of volume for subsets of \( \mathbb{R}^3 \), agreeing numerically with the informal or assumed “volume” calculated by earlier methods. The expected properties of area must be proved from the definition, but can be shown to hold (although there are some subtleties, e.g. the area of the boundary of a region is not necessarily zero if the boundary is sufficiently complicated).

The Nature of Measure Theory

I.3.1.13. The goal of the subject of measure theory is to associate a numerical “size” to subsets of a set \( X \) in a precise, consistent, and well-behaved manner. The principal property that such a notion of “size” should have is that the “size” of a union of disjoint subsets should be the sum of the sizes of the sets.

I.3.1.14. It turns out that in most applications it is both natural and necessary to restrict attention only to certain well-behaved subsets of \( X \); frequently some subsets are too bizarre to allow a reasonable notion of “size.” In addition, to get a reasonable theory one must only allow countably many operations. The technical term is that the collection of “measurable sets” (subsets with a “size” or measure) form a \( \sigma \)-algebra (see () for a precise definition.)

I.3.1.15. In many instances there is a natural notion of size for a collection of subsets which is not large enough to form a \( \sigma \)-algebra, and the task is then to extend this size notion to a larger \( \sigma \)-algebra of sets.

There are two fundamental examples of this process, which have motivated most of the development of the theory:

I.3.1.16. Example. As discussed above, let \( X \) be the real numbers \( \mathbb{R} \). There is a natural size for certain nice subsets, the intervals: if \( I \) is the interval from \( a \) to \( b \) (with or without endpoints), the measure of \( I \) should be the length \( b - a \). It turns out that this measure can be extended to a \( \sigma \)-algebra of subsets of \( \mathbb{R} \) in such a way that congruent sets have the same measure. The resulting measure is called Lebesgue measure on \( \mathbb{R} \), and gives by far the most natural and important example of a measure space. It turns out that not every subset of \( \mathbb{R} \) can be Lebesgue measurable, but every subset which can be explicitly described, and in particular any set which arises in any reasonable way in applications, is measurable.

Similarly, the plane \( \mathbb{R}^2 \) has a size notion (“area”), as does three-dimensional Euclidean space \( \mathbb{R}^3 \) (“volume”), and more generally \( \mathbb{R}^n \) for any \( n \). These size notions are only defined at first for nice subsets such as rectangular solids, but can be extended in the same manner as in \( \mathbb{R} \). These cases are a little harder intuitively than the one-dimensional case: for example, even finding the area of a set like a disk requires a nontrivial limiting process. However, the general theory applies in just the same manner as in the one-dimensional case to give \( n \)-dimensional Lebesgue measure.

There are other natural measures on Euclidean space. For example, there is a measure on \( \mathbb{R}^2 \) which assigns to a smooth curve its arc length. This measure, called one-dimensional Hausdorff measure, would give infinite size to any two-dimensional set. Hausdorff measures giving fractional dimensions to fractal sets can also be defined. In fact, measure theory underlies the whole study of fractals.
I.3.1.17. Example. The basic setup of probability theory is to analyze a process or “experiment” to numerically describe the possible results. The set of all possible outcomes is called the sample space, which we will denote \( \Omega \). A subset \( E \) of \( \Omega \) is called an event. In the simplest situation, \( \Omega \) is just a finite set, and each point has a certain probability of occurring; the sum of all the individual probabilities is 1. The probability that an event \( E \) occurs is the sum of the probabilities of the points in \( E \).

However, the set \( \Omega \) is often infinite, and individual points have probability 0, but infinite events may have positive probability. For example, suppose the process is to choose a number \( x \) randomly from \([0, 1]\). The sample space \( \Omega \) is thus \([0, 1]\). If \( x_0 \) is any fixed number, the probability that the chosen \( x \) is exactly equal to \( x_0 \) is 0. However, the probability that \( 0 \leq x \leq 1/2 \) (\( E = [0, 1/2] \)) is 1/2. More generally, the probability that \( a \leq x \leq b \) is \( b - a \).

The proper setup for probability theory is to take the set of allowable events to be a \( \sigma \)-algebra of subsets of \( \Omega \), and the probability of events is given by a measure on this \( \sigma \)-algebra with the measure of the whole space \( \Omega \) equal to 1. In the above example, the probability is given by Lebesgue measure on \([0, 1]\).

In fact, most of the elementary concepts of probability theory are measure theory concepts with different terminology. A simple dictionary is all that is needed to make the translation.

I.3.2. Integration

I.3.2.1. Integration is the other basic component of this part of real analysis. The goal of integration is to systematically assign to a real-valued function \( f \) on a set \( X \), and a subset \( A \) of \( X \), a number \( \mathcal{I}_A(f) \) called the integral of \( f \) over \( A \) (the notation \( \int_A f \), or some variation, is usually used; the symbol \( \int \) is called the integral sign), which is usually thought of as a “normalized sum” of the values of \( f \) on the points of the subset \( A \). This number has various interpretations in applied situations as total distance (regarding \( f \) as giving instantaneous velocity), total mass (regarding \( f \) as giving density), etc. Integration in analysis is thus almost always (and always in this book) a generalization or analog of what in elementary calculus is called the “definite integral.” (We will also have occasion to consider analogs of the “indefinite integral” of calculus, but these will have different names and will not be called integrals.)

I.3.2.2. All modern theories of integration are in a sense outgrowths of the so-called Riemann integral (actually developed first by Cauchy for continuous functions and generalized by Riemann to also handle moderately discontinuous functions), the type of integration treated in elementary calculus. The Riemann integral was the first reasonably satisfactory theory of integration, and is still by far the theory most widely used in applications; Riemann integrals can in many cases be exactly and efficiently computed using antiderivatives and the Fundamental Theorem of Calculus, and there are good numerical techniques for approximating Riemann integrals that cannot be computed exactly. In fact, more general integrals are rarely actually computed directly; in practice they are usually approximated by Riemann integrals which are then computed or estimated using standard techniques.

For theoretical purposes, Riemann integration has two serious inadequacies:

1. The class of Riemann integrable functions is too small. For a function to be Riemann integrable on an interval, it must be bounded and “almost continuous” (see () for a precise statement). It is particularly useful to be able to integrate certain unbounded functions. This is reasonably, but not entirely, satisfactorily achieved by using various notions of improper Riemann integrals and principal values. It is also important to be able to integrate over unbounded intervals, which can be done through other types of improper integrals, although not in an entirely satisfactory way. At least for theoretical purposes, it is also important to be able
to integrate functions which are too discontinuous to be Riemann integrable, and to be able to integrate functions over subsets of the real numbers more general than intervals or finite unions of intervals (although the practical value of this type of integration is sometimes somewhat overemphasized as an inadequacy of Riemann integration).

2. It is crucial in many settings to have limit and approximation theorems, i.e. to know that under reasonable conditions, if \( f_n \to f \) in some sense, we have \( I_A(f_n) \to I_A(f) \). It is quite difficult to establish good theorems of this sort for Riemann integration which are applicable even if we know that the limit function is Riemann integrable (which is often not the case due to the limitations described above). It turns out that to get reasonable convergence theorems, more general integration theories are needed.

Another limitation of Riemann integration is that it only (directly) works on \( \mathbb{R} \), or on \( \mathbb{R}^n \). It is very important to have integration theories which work on more general spaces, and it turns out that modern approaches give this flexibility with essentially no additional work.

“Lebesgue’s definition enables us to integrate functions for which Riemann’s method fails; but this is only one of its advantages. The new theory gives us a command over the whole subject which was previously lacking. It deals, so to speak, automatically with many of the limiting processes which present difficulties in the Riemann theory.”

\[ E. \ C. \ Titchmarsh^{3}\]

There are three principal modern approaches to integration, each with many variations. We first discuss the basics of Riemann integration in a form allowing the various generalizations to be made naturally.

I.4. Other Topics

“The object to be attained by the theory of functions of a real variable consists then largely in the precise formulation of necessary and sufficient conditions for the validity of the limiting processes of Analysis. . . . In the literature of the subject, errors are not infrequent, largely owing to the fact that spatial intuition affords an inadequate corrective of the theories involved, and is indeed in some cases almost misleading.”

\[ E. \ W. \ Hobson^{4}\]

“If you have an apple and I have an apple and we exchange apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas.”

\[ George \ Bernard \ Shaw\]
I.5. Rigor in Mathematics

“A simple man believes every word he hears; a clever man understands the need for proof.”  
Proverbs 14:15\(^5\)

“After deserting for a time the old Euclidean standards of rigour, mathematics is now returning to them, and even making efforts to go beyond them. . . . Later developments . . . have shown more and more clearly that in mathematics a mere moral conviction, supported by a mass of successful applications, is not good enough. Proof is now demanded of many things that formerly passed as self-evident. Again and again the limits to the validity of a proposition have been in this way established for the first time. . . . In all directions these same ideals can be seen at work – rigour of proof, precise delimitation of extent of validity, and as a means to this, sharp definition of concepts.”

Gottlob Frege, 1884\(^6\)

A student, particularly a nonmathematics student, may well wonder about two questions:

1. Why is it necessary to spend the time and effort to develop a precise, rigorous, and complete theory of measure and integration (or any other mathematical topic)? Isn’t this just an idle intellectual exercise for mathematicians with nothing better to do, without practical benefit to those using the theory? Does there have to be an elegant general theory behind every useful piece of mathematics?

2. Even given that someone ought to carefully work out the details, isn’t it enough that someone already has? Why can’t we just use the theory without having to immerse ourselves in the details?

Both questions are legitimate ones, and deserve serious answers.

Let us discuss the second question first. Here there is a good analogy. It is quite possible for a person to use a machine without understanding all the technical details of its construction and operation. However, it is quite important that the user, in addition to simply knowing how to operate the machine, understand enough about it to know what it can and cannot do, and what dangers it may pose; failure and even disaster can result from careless operation or trying to use the machine beyond its capabilities. Also, the user might eventually want to have the machine do something it was not designed for, and the machine may have to be modified to accomplish the new task. It would likely be impossible to figure out how to adapt the machine to a new task without a good understanding of its construction.

In addition, the existing theory is so vast that it is often easier to work out what is needed in a particular situation instead of doing a possibly long and involved search for exactly the right “off-the-shelf” result that applies. One needs a good understanding of the theory to do this (and indeed even to recognize when an off-the-shelf result applies and suffices!)

“The problem may be viewed as one of information retrieval; we wish to find a particular piece of information, not knowing for sure that it has ever been recorded, and instead of trying to provide a catalog of known results (which would necessarily be finite and to a great extent represent past

---

\(^5\)New English Bible

interests), we prefer a method of creating it as needed. Instead of information retrieval we prefer information regeneration. Of course this attitude will not work in all areas of knowledge, but in a mature science the principles should show how to recreate the information on demand.”

R. W. Hamming

The first question is almost asking: What is mathematics? To begin with the last part of Question 1, it is certainly not true that every bit of mathematics has to be backed by a rigorous general theory to be useful. For example, a lot of analysis beginning with calculus was developed and extensively used for about 200 years before its logical underpinnings were carefully worked out in the nineteenth century. Even today, where standards of rigor are high and mathematicians rarely if ever use techniques and results whose validity has not been carefully established, there are many heuristic and ad hoc methods not backed up by general theory, especially in applied mathematics. This is in no way a criticism of this work, or of applied mathematics in general; if applied mathematicians had to wait for general theories underlying their techniques before using them, progress would grind nearly to a halt (and the general theories might never come.)

A good example is fractal geometry. Some mathematicians question whether this subject should even be regarded as part of mathematics, since there are almost no theorems and the work is almost entirely descriptive rather than analytical. I personally do not take such a narrow view of mathematics, and I think this subject is unquestionably part of mathematics; after all, did calculus only become part of mathematics after Weierstrass?

Perhaps an even more striking example is the “theory” of quantum groups. Mathematicians seem to know a quantum group when they see one, and there has been a great deal of work on quantum groups over the last several decades, some of which is widely regarded as important, first-rate mathematics. But there is still as yet no generally accepted definition of a quantum group!

Nonetheless, the usefulness and even necessity of rigorous general theories is not diminished by such work, and indeed the desirability and urgency of a general theory explaining and connecting specific results and techniques is greatly enhanced by the applications.

**Authority in Mathematics**

“Mathematics is fundamentally an egalitarian subject. It should in principle be possible for anyone to check the correctness of any argument; nothing needs to be taken on trust.”

J. K. Truss

(See III.1. for a continuation of this quote.)

“Believe nothing on hearsay. Do not believe in traditions because they are old, or in anything on the mere authority of myself or any other teacher.”

Buddha

---

7[T, p. 244]
8[Tru97, p. v]
9[Bel92, p. xi].
“No man can worthily praise Ptolemye... yet muste ye and all men take heed, that both in him and in all mennes workes, you be not abused by their autoritye, but evermore attend to their reasons, and examine them well, ever regarding more what is saiide, and how it is proved, than who saieth it, for autorite often times deceaveth many menne.”

*Robert Recorde, 1556*

“In questions of science the authority of a thousand is not worth the humble reasoning of a single individual.”

*Galileo*

We should never rely entirely on an “authority” to tell us what is true in mathematics, although we should give credence to the guidance of knowledgable people. For one thing, any “authority” can be wrong. Nobody has a corner on the truth in mathematics; some just have a somewhat better handle on it than others.

A simple illustration is the story told [?, p. 73] by JOHN PAULOS, a mathematician and author of the popular book *Innumeracy*, who at age ten calculated the earned run average of a certain baseball pitcher to be 135 (a pitcher’s earned run average, or ERA, is the average number of earned runs allowed per nine innings pitched, calculated by the formula

\[
\text{ERA} = \frac{9 \cdot \text{(number of earned runs allowed)}}{\text{innings pitched}}
\]

This pitcher had allowed 5 earned runs in 1/3 inning pitched.) He showed his calculation to his teacher, who authoritatively replied that he was wrong, that an ERA could never be higher than 27 (the number of outs in 9 innings). When the published official statistics confirmed his calculation, he felt vindicated, and recalls:

“I remember thinking of mathematics as an omnipotent protector. You could prove things to people and they would have to believe you whether they liked you or not.”

Unfortunately, his teacher still did not believe him even when presented with the published proof!

Students, particularly at lower levels, are often far too willing to defer to the “authority” of instructors in mathematics. I once had a calculus student who came to my office because he could not get his answer to an online homework problem to match what the computer demanded (a serious ongoing problem with online homework, although things are slowly improving). Instead of the “right” answer of 2, he kept getting \(\frac{3+\sqrt{1}}{2}\). I told him to simplify the square root to 1. He thought a minute, then said, “Wait a minute. Is \(\sqrt{1}\) 1?” When I assured him it was, he thought a minute more, then said, “Wait a minute. Is \(\sqrt{1}\) always 1, or just in this course?” I refrained from suggesting he ask his Calculus II instructor the next semester what \(\sqrt{1}\) is in that course.

---


\(^{11}\) Quoted in Arago’s Eulogy on Laplace; cf. [Mor93, 1528].
Abstraction, Interpretation, and Point of View

“The more comprehensive the subject is they treat, and the more general and abstract the manner is in which they are considered, the more are [mathematical] principles free of darkness and the easier it is to grasp them.”

Jean d’Alembert, 1743

“The purpose of axiomatization in mathematics is to reduce to a minimum the presuppositions on which a given theory depends, and at the same time to make every presupposition fully explicit.”

B. Rotman and G. T. Kneebone

“The most striking characteristic of the written language of algebra and of the higher form of the calculus is the sharpness of definition, by which we are enabled to reason upon the symbols by the mere laws of verbal logic, discharging our minds entirely of the meaning of the symbols, until we have reached a stage of the process where we desire to interpret our results. The ability to attend to the symbols, and to perform the verbal, visible changes in the position of them permitted by the logical rules of the science, without allowing the mind to be perplexed with the meaning of the symbols until the result is reached which you wish to interpret, is a fundamental part of what is called analytical power. Many students find themselves perplexed by a perpetual attempt to interpret not only the result, but each step of the process. They thus lose much of the benefit of the labor-saving machinery of the calculus and are, indeed, frequently incapacitated for using it.”

Thomas Hill

“Mathematics is a way of thinking that can help make muddy relationships clear. It is a language that allows us to translate the complexity of the world into manageable patterns. In a sense, it works like turning off the houselights in a theater the better to see a movie. Certainly, something is lost when the lights go down; you can no longer see the faces of those around you or the inlaid patterns on the ceiling. But you gain a far better view of the subject at hand.”

K. C. Cole

One of the characteristic and essential features of mathematics is the fact that most parts of the subject can be approached or interpreted from various points of view. Not only does this variety of viewpoints increase the usefulness of mathematical theories, but the very progress of mathematics often depends on this flexibility: problems which appear obscure or difficult from one point of view often have obvious solutions when viewed in a different way.

In fact, modern mathematicians deliberately try to create theories which have flexibility and variety of viewpoints built in. This is one of the principal reasons for the abstraction of today’s mathematics. The usual procedure is to base the entire theory on a few basic and generally abstract assumptions called axioms, and logically deduce all features of the theory from the axioms without making additional assumptions about the “meaning” of the axioms or the terminology used in them. Abstraction serves several purposes:

---

12J. d’Alembert, Traité de dynamique, Paris, 1743.
1. It simplifies the range of information which must be simultaneously considered and used in developing
the theory, including the notation and terminology, to a manageable level.

2. It allows mathematicians to focus their thinking on the essential structural features of the theory
without being tied to one particular interpretation of the theory which may constrain creative thought
by introducing unimportant or even extraneous elements.

3. It allows the precise limits of the validity of the theory to be described.

4. It allows, even invites, new interpretations of the theory to be identified, leading not only to broader
applicability of the theory and results already obtained, but also to new ways of looking at and solving
problems arising in the development of the theory.

5. It makes it easier to notice similarities and connections with other parts of mathematics, leading both
to a better understanding of the big picture of mathematics, and to cross-fertilization: importation of
ideas and techniques from one part of mathematics into another.

“Abstraction gives us the ‘less is more’ approach – fewer assumptions leave us in touch with
more mathematical objects (and thus the theorems we prove apply in more instances). More
importantly, however, it gives us the ‘less is clearer’ approach. Because we haven’t made a myriad
of assumptions, our mathematical choices are more clearly defined, our mathematical arguments
are cleaner, and the conclusions we draw are less likely to be cluttered up by irrelevant issues.
We can, therefore, sort out the essential mathematical ingredients that support a particular fact.”

*Carol Schumacher*\(^{16}\)

“[A] single simple principle can masquerade as several difficult results; the proofs of many theo-
rems involve merely stripping away the disguise.”

*M. Spivak*\(^{17}\)

The Greek word “axiom”, as used in mathematics, originally meant “self-evident truth.” In formalist
mathematics, axioms are regarded as “arbitrary assumptions.” To most modern mathematicians, they are
neither. The axioms used in a theory are a matter of choice and are not necessarily obvious. But they
are not arbitrary; they are chosen to be a reasonable general starting point for a theory with at least one
interpretation or application of interest. The terminology used in the axioms and throughout the theory is
often, but not always, suggestive of a particular interpretation; such suggestive terminology can be beneficial
in guiding a mathematician’s thoughts, but also confining by restricting the mathematician’s ability to
“think outside the box” and to find and use other interpretations of the concepts. Sometimes an entirely
equivalent theory can be developed from a completely different set of equally reasonable axioms, often
using different terminology; when this occurs, the theory tends to be richer because different results can
be more easily or naturally obtained in the two setups (and the different axiom schemes are also evidence
of broader applicability of the theory). Measure theory is an excellent example of this; various approaches
beginning with measure, probability, and integration lead to essentially the same theory, with many varied
tools available from the different approaches.

\(^{16}\)[?, p. 5].

\(^{17}\)[?, p. ix].
“Mathematics is the art of giving the same name to different things. It is enough that these things, though differing in matter, should be similar in form, to permit of their being run in the same mould. When language has been well chosen, one is astonished to find that all demonstrations made for a known object apply immediately to many new objects: nothing requires to be changed, not even the terms, since the names have become the same.”

H. Poincaré

Abstraction for its own sake is rarely worthwhile in mathematics. Even though it is usually impossible to foresee the full implications of a new theory, significant mathematical developments almost always occur when mathematicians have some interpretation or use in mind for the new theory; these can range from direct applications to important problems outside mathematics to simple curiosity about a specific mathematical problem that just seems interesting.

“The axiomatic method has provided deep insights into mathematics, disclosing identities where none had been suspected. In the hands of mathematicians of genius this method has been used to strip away exterior details that seem to distinguish two subjects and to disclose an identical structure whose properties can be studied once for all and applied to the separate subjects.”

Mina Rees

“[T]he users of [mathematics] naturally enough take a somewhat skeptical view of this development [abstraction and generalization] and complain that the mathematicians now are doing mathematics for its own sake. As a mathematician my reply must be that the abstraction process that goes into functional analysis [and other parts of mathematics] is necessary to survey and to master the enormous material we have to handle. . . . Our critics, especially those well-meaning pedagogues, should come to realize that mathematics becomes simpler only through abstraction. The mathematics that represented the conceptual limit for Newton and Leibniz is taught regularly in our high schools, because we now have a clear (i.e. abstract) notion of a function and of the real numbers.”

G. Pedersen

---

18[?], p. 34.
20[Ped89, p. vii-viii].
I.5.1. Semantics: What is Obvious?

“‘Obvious’ is the most dangerous word in mathematics.”

E. T. Bell

“Since people have tried to prove obvious propositions, they have found that many of them are false.”

Bertrand Russell

“Being a mathematician means that you don’t take ‘obvious’ things for granted but try to reason. Very often you’ll be surprised that the most obvious answer is actually wrong.”

Evgeny Evgenievich

“It ain’t what you don’t know that gets you into trouble. It’s what you know for sure that just ain’t so.”

Mark Twain

“Science is always in need of those who are willing, and able, to say ‘Hold on a minute, isn’t there a problem with . . . ?’ ”

Peter R. Law

It is a surprisingly delicate matter, which can confuse not only students but even professional mathematicians, to decide what in mathematics is “obvious” and can simply be asserted, and what must be carefully spelled out. Often the difficulty is a semantic one, stemming from misuse, or at least careless use, of language: language can be subtly suggestive in a misleading way. We will discuss a few such examples related to the foundations of analysis.

Areas in the Plane

I.5.1.1. The first example concerns areas of sets in the plane. Consider the following two statements:

Statement 1. If \( R \) is a rectangle with length \( x \) and width \( y \), then the area \( A(R) \) is \( xy \).

Statement 2. If a rectangle \( R \) is subdivided into a finite number \( R_1, \ldots, R_n \) of nonoverlapping subrectangles (nonoverlapping means two of the subrectangles can have only boundary points in common), then

\[
A(R) = \sum_{k=1}^{n} A(R_k)
\]

(See Figure I.1; the area of the black rectangle is the sum of the areas of the green rectangles.)

---

21[Bel87, p. 16].
22[?].
23[Fre13, p. 16].
24This is one of many quotes often attributed to Twain for which there is no documentation. The first version published of this statement appears to be due to Josh Billings in 1874.
25MathSciNet review of [Pen05].
Let us set Statement 1 aside for the moment, and consider Statement 2. The question we want to discuss is: Is Statement 2 obvious?

Many people would say that Statement 2 is “geometrically obvious.” And not just people untrained in mathematics: for example, on p. 7 (the first page of actual text) of the otherwise carefully written book [SG77], it is asserted without comment that Statement 2, and even its higher-dimensional version, is true, and this assertion is then used as the basis for the whole development of Riemann integration in Euclidean space. The same assertion is made in [KF75, p. 255] (where it is called “obvious”) and in [Sak64]; essentially the same statement (including the “obvious” assertion) occurs in [Bog07, p. 180]. (Is this an Eastern European thing? Note, however, that [Nat61] does not have this problem – see p. 64-65 of Volume 2.)

The real issue is whether Statement 1 is a definition or a proposition. If it is a proposition, then what is the definition of the area of a rectangle, or other plane figure? Are we using some intuitive “definition” of area, along with its accompanying list of “natural” properties, among which would be some general version of Statement 2? What justification do we have that there is a notion at all of area having the geometric properties we would expect? (There is, but it takes some work to show it, even for polygons; cf. [Har00, §23]. For more complicated regions, even a circle, it is still more difficult, requiring limits.) We can bypass these difficulties by taking Statement 1 as the definition of the area of a rectangle; but then we are not justified in assuming that “area” so defined has the expected geometric properties. In particular, Statement 2 is then not obvious, and requires proof. (The proof, while not difficult when the right “trick” is seen, is nontrivial: see XIV.3.1.6. Note that it cannot simply be proved by induction on \( n \): the cases \( n = 3 \) and \( n = 4 \) can be easily reduced to the case \( n = 2 \), which is obvious, but even the case \( n = 5 \) is not entirely obvious and cannot be simply reduced to applications of the smaller cases, since a rectangle can be divided into five subrectangles in such a way that the union of no two, three, or four subrectangles is a rectangle.)
Some references such as [Wyl64] try to define area axiomatically; Statements 1 and 2 are then among the axioms, or at least theorems, of the theory. But this approach does not eliminate the problem: the difficulty then simply becomes showing that the axioms are consistent, a question not really discussed in [Wyl64].

The semantic problem is use of the word “area” without specification of its meaning; it is suggestive to the reader of an intuitive interpretation which was perhaps reinforced by informal discussions of the subject in school, even though the informal or intuitive interpretation may not be appropriate in careful mathematics.

I.5.1.2. Anyone who thinks it is obvious that when a figure is cut up and rearranged, the area stays the same, should try the following example (photo from the website [http://www.moillusions.com/2006/03/impossible-triangle-illusion-no2.html](http://www.moillusions.com/2006/03/impossible-triangle-illusion-no2.html)):

![Image of a triangle cut into pieces and rearranged](http://www.moillusions.com/2006/03/impossible-triangle-illusion-no2.html)

Figure I.2: Where did the extra square come from?

A triangle of area $\frac{65}{2}$ is cut up and rearranged into a figure of area $\frac{63}{2}$ (or is it?) For a great visual version, see [http://www.moillusions.com/2013/04/video-missing-cubes-optical-illusion.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+OpticalIllusions+%28OpticalIllusions%29&utm_content=Google+International](http://www.moillusions.com/2013/04/video-missing-cubes-optical-illusion.html?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+OpticalIllusions+%28OpticalIllusions%29&utm_content=Google+International).

A much more sophisticated case (which is a real counterexample, not just an illusion) is the Banach-Tarski Paradox (): a ball in $\mathbb{R}^3$ can be dissected into five pieces which can be rearranged to form two balls of the same size as the original. The paradox is resolved by the fact that the pieces are nonmeasurable: they are so bizarre that no “volume” can be assigned to them. (In fact, area in $\mathbb{R}^2$, when defined, is preserved under arbitrary finite dissection and rearrangement, but this is quite nontrivial and nonobvious and, as the Banach-Tarski Paradox shows, does not hold for either volume or surface area in $\mathbb{R}^3$. See e.g. [Wag93].)
Circumference of a Circle and the Number $\pi$

I.5.1.3. Area, at least for a rectangle, is easy and elementary to define (or calculate). But here is an example where making a careful definition of an intuitive concept is quite complicated.

How do we define the number $\pi$? It is usually defined as the ratio of the circumference of a circle to its diameter (for this to make sense, it must be shown that the ratio is always the same for all circles, which is not really so obvious; in fact, it is false in noneuclidean geometry!) Slightly more precisely:

**DEFINITION.** The number $\pi$ is the circumference of a circle of diameter 1.

Several things must be made precise before this can be accepted as a real definition: the terms *circumference*, *circle*, and *diameter* (and I suppose even the number 1) must be defined. To get to the heart of the matter, let us tighten up the definition to

**DEFINITION.** The number $\pi$ is half the circumference of the circle

$$\{(x, y) : x^2 + y^2 = 1\} \subseteq \mathbb{R}^2.$$  

The real issue is the definition of the term *circumference* (or *perimeter*). Although people have a (fairly) good intuition as to what this term means, and one could be led to say the meaning is “obvious”, it is not at all easy to give a precise definition, and the meaning of the word is finessed throughout precalculus mathematics (e.g. to paraphrase MANDELBROT, what is the circumference of Britain?) Using calculus, it can be defined as a special case of arc length, which can itself be defined as the supremum of a set of sums (and in good cases exactly calculated as a definite integral). It is then not entirely obvious that a circle even can be defined as a special case of arc length, which can itself be defined as the supremum of a set of sums.

To elaborate on the difficulties, let us consider the traditional (at least since ARCHIMEDES) way circumference is defined and estimated. Fix a circle. We consider inscribed and circumscribed regular polygons. For a fixed $n$, the perimeter $p_n$ of all inscribed regular $n$-gons is the same by symmetry, and similarly for the perimeter $P_n$ of circumscribed regular $n$-gons, and obviously $p_n < P_n$. It is not obvious that $p_n$ increases monotonically as $n$ increases (although this can be shown), but it is obvious, and easily proved, that $p_n < p_m$ if $n$ divides $m$. It is less obvious, but not difficult to prove, that $P_m < P_n$ if $n$ divides $m$. Thus, if $n$ divides $m$, $p_n < p_m < P_m < P_n$. The circumference $p$ of the circle is generally defined to be the supremum (or limit) of the $p_n$ as $n \to \infty$. It is not hard to show that $P_n - p_n \to 0$ as $n \to \infty$ (cf. I.5.1.16.(v)), so $p$ is also the infimum (or limit) of the $P_n$. By explicitly calculating $p_n$ and $P_n$ for large $n$, decent upper and lower estimates for $p$ can be obtained.

Incidentally, some references assert that it is “obvious” that $p < P_n$ for any $n$. I’m sorry, but it is not at all obvious to me. Intuitive arguments for this fact seem to be based on some type of belief in the continuity of arc length (I.5.1.11.). It does follow (using that $P_m < P_n$ if $n$ divides $m$) from the lower semicontinuity of arc length (VI.4.1.30.), but I don’t think that is intuitively obvious. It is intuitively plausible that $p_n < p$ for all $n$, based on the principle that “a straight line is the shortest distance between two points,” and it is not unreasonable to define $p$ as sup $p_n$. The above arguments which show that $\inf P_n$ is equal to $\sup p_n$ imply that $\inf P_n = p$ and hence $p < P_n$ for all $n$ (this follows more easily directly from the fact that $p_m < P_n$ if $n$ divides $m$, which implies that $P_n$ is an upper bound for the $p_m$ for any $n$).

So what is the difficulty with this argument? For one thing, many seemingly small details must be made precise; all the statements above need to be proved, and along the way some terms need to be carefully defined, which is sometimes not as easy as one might think. In addition, one needs to show:
There exist inscribed and circumscribed regular \( n \)-gons. Note that this cannot be done simply by using angle measure since angle measure cannot be carefully defined until concepts from trigonometry are developed, which already requires knowledge about the perimeter of a circle and its subarcs. Instead, either geometric or analytic arguments must be used. (Certain regular polygons inscribed in a circle like a square or hexagon can be constructed by straightedge and compass, and their sides repeatedly bisected to obtain regular polygons with arbitrarily large numbers of sides, but not all regular polygons can be constructed this way, although enough can to make the argument work with additional justification.)

The perimeter of an inscribed polygon must be defined. In order to do this, it must be said what is meant by a finite sequence of points on a circle being listed in cyclic order. It is “obvious” geometrically what this means, but not so simple to explicitly state in words (if angle measure cannot be used). Two (equivalent) ways are: (1) interiors of edges between two pairs of consecutive points never intersect or (2) the sum of the distances between consecutive points (the perimeter) is minimum among all possible orderings of the given points. (The second characterization may be preferable, since it is not obvious that a given finite subset of the circle has any ordering satisfying (1).)

The Completeness Property of \( \mathbb{R} \) is needed to insure that the supremum or limit of the \( p_n \) exists.

In addition, there is another technical difficulty. The arc length of a circle, as defined in calculus, would be the supremum of the perimeters of all inscribed polygons, regular or not. It is not obvious that the same supremum is obtained for all inscribed polygons as for inscribed regular polygons, i.e. that there could not be a nonregular inscribed polygon whose perimeter is greater than the perimeters of all inscribed regular polygons. A way of showing that the two suprema are the same is to fix a point \( a \) on the circle, and show that the vertices of regular inscribed polygons with one vertex at \( a \) are dense in the circle; by continuity it suffices to consider only inscribed polygons with vertices in this set, and any finite number of points in this set are among the vertices of a single inscribed regular polygon. The density of this set of vertices, while “geometrically obvious”, takes some work to prove. As an alternate approach, it can be shown using a compactness argument (based on the Completeness Property) that for any fixed \( n \), there is an inscribed \( n \)-gon of maximum perimeter, and it is fairly easy to show that it must be a regular \( n \)-gon. Or the arguments of [Pó90a, Chapter X] can be adapted to give an “elementary” proof of this fact, although a careful and complete argument involves much more than is discussed in [Pó90a] (cf. I.5.1.5.).

All in all, it is a quite involved matter to make this approach completely rigorous.

I.5.1.5. There are a number of closely related arguments in [Pó90a]. Let us confine the discussion to the topics in Chapters VIII and X, where max-min problems and the Isoperimetric Problem are considered. The arguments in [Pó90a] are clever and elementary, and at first reading seem convincing, but there are serious logical gaps in the arguments which the author seemingly tries to obscure by his choice of language. Among these are:

(i) It is often implicitly assumed, or even explicitly asserted without explanation, that the desired maximum or minimum exists. To justify the existence requires at least an application of the Completeness Property of \( \mathbb{R} \), and in some cases additional continuity or compactness arguments are needed.

This point of justifying existence of a maximum/minimum is not just a pedantic technicality: there are seemingly similar problems, e.g. the Dirichlet principle (), where it turned out that no max/min exists in general, much to the surprise of mathematicians who assumed the contrary as “obvious.”

(ii) In some arguments, it is assumed without discussion that there is a notion of area of regions of the plane bounded by various types of curves, which is additive on “nonoverlapping” regions and invariant under rigid motions, leading to more complicated versions of the difficulties discussed in I.5.1.1..
(iii) In several arguments, use is made of “tangent lines” to various curves without defining what this means, much less discussing whether such lines exist and what properties they must have. When the curve is a circle, tangent lines can be defined and found without any explicit use of calculus (but there is still some nontrivial work to be done for a careful treatment), but for more general curves calculus cannot be avoided.

(iv) Several other similar difficulties are dismissed using a comment like “remembering some elementary geometry . . .” and some of these are theorems that take some work to prove or even to carefully state.

See [Oss78], [BZ88], or [Ber03, p. 26-29] for much more careful and complete discussions of the Isoperimetric Inequality.
The Area of a Circle

I.5.1.6. The circumference of a circle of radius \( r \) is \( 2\pi r \), and the area is \( \pi r^2 \). Even a casual observer would notice a similarity between these two formulas, and a calculus student would observe that the circumference is the derivative of the area (this is of course no accident, and is the basis for a quick calculus proof of the area formula). Explaining to a layman why the formulas are similar, and why one follows from the other, is a challenge, though.\(^{26}\) Since the circumference formula is essentially just true by the definition of \( \pi \), the task amounts to proving the area formula from the circumference formula.

I.5.1.7. The area formula requires calculus in some form for its proof, since some type of limit is necessarily involved. (As explained earlier, limits are involved in even making a careful definition of either area or circumference.) ARCHIMEDES described a beautiful argument which seems to make the area formula “obvious” once the informal limiting operations are made rigorous. This argument has recently been again popularized by S. STROGATZ in his New York Times series ([http://opinionator.blogs.nytimes.com/2010/04/04/take-it-to-the-limit/](http://opinionator.blogs.nytimes.com/2010/04/04/take-it-to-the-limit/)), from which the following pictures are taken.

I.5.1.8. The argument goes like this. Divide the circle (actually disk) into a large number of congruent wedges, say \( 2^n \), and rearrange them into a strip:

![Figure I.3: Rearrangement with 4 wedges](http://opinionator.blogs.nytimes.com/2010/04/04/take-it-to-the-limit/)

For any \( n \), the wavy strip has the same area as the circle, the top and bottom have arc length \( \pi r \), and the sides are line segments of length \( r \).

I.5.1.9. As \( n \to \infty \), the top and bottom wavy curves flatten out into straight line segments, and the end segments become more and more vertical; thus in the limit the figure becomes a rectangle.

This rectangle has area \( \pi r^2 \), so the area of the circle is also \( \pi r^2 \).

\(^{26}\) Teachers don’t always try. PETR BECKMANN [Bec71, p. 17] says: “Most of us first learned this formula [for the area of a circle] in school with the justification that the teacher said so, take it or leave it, but you better take it and learn it by heart; the formula is, in fact, an example of the brutality with which mathematics is often taught to the innocent.”
I.5.1.10. This argument is sometimes represented as a “proof” of the area formula. But it is not (yet) since it has some gaps, one of which is quite serious. The obvious gaps are that

(i) The limiting operations must be made precise.

(ii) It must be shown that area is well defined for the types of regions considered, and that it has the subdivision property discussed in I.5.1.1.

(Actually, (ii) may not be obvious to most observers!) However, there are two more gaps which may not be as obvious:

(iii) It must be shown that such a decomposition into wedges is possible, discussed in I.5.1.4.(i).

(iv) It must be shown that both area and arc length are preserved under the limiting operations of (i).

I.5.1.11. The preservation of area under the limiting operation of passing from the wavy strips to the limiting rectangle takes some work to prove rigorously, but is at least true in great generality, and it is possibly justifiable to regard the proof as simply a technical justification for an “obvious” geometric assertion.

However, the arc length assertion is definitely not “obvious.” It is not even true in general. Actually, to even formulate it carefully it must be said what the limiting operation is. One reasonable way to state it formally is: regard the bottom curve as the graph of a function $f_n$, and the limiting curve the graph of a
(constant) function \( f \). Then \( f_n \to f \) uniformly. (One could alternately think of them as parametrized curves and consider uniform convergence of the parametrizing functions.) However, arc length is not preserved under such a limiting operation (it is only lower semicontinuous (VI.4.1.30.), not continuous):

**I.5.1.12. Example.** On the interval \([0, \pi]\), define a sequence of functions by

\[
f_n(t) = 2^{-n} \sin(2^n t)
\]

and a limit function \( f = 0 \). Then \( f_n \to f \) uniformly on \([0, \pi]\), i.e. the curves \( y = f_n(x) \) approach the straight line from 0 to \( \pi \) on the \( x \)-axis (in the uniform sense). However, the arc lengths of the curves are all the same, as can be seen easily from a geometric argument (the graph of \( f_{n+1} \) consists of two pieces, each similar to the graph of \( f_n \) scaled by a factor of \( \frac{1}{2} \)), and this arc length can be computed by calculus () to be

\[
\int_0^\pi \sqrt{1 + [f_n'(t)]^2} \, dt = \int_0^\pi \sqrt{1 + \cos^2(2t)} \, dt \approx 3.82
\]

which is not the same as the arc length \( \pi \approx 3.14 \) of the limiting curve.

One can do much worse: set

\[
g_n(t) = 2^{-n} \sin(4^n t)
\]

on \([0, \pi]\). Then \( g_n \to 0 \) uniformly, so the graphs again approach the straight line segment from 0 to \( \pi \), but the arc lengths of the graphs of the \( g_n \) go to \( +\infty \).

**I.5.1.13.** A similar example is given in [LN00] (“le trou normand”), where it is regarded as a “paradox.” Someone with a reasonable level of mathematical sophistication (unlike the authors of that book) would realize that this is not a paradox at all: the problem is simply that if \( f_n \to f \) uniformly, then it is generally not true that \( f_n' \to f' \) in any sense (even if the derivatives all exist, which need not be the case).

In fact, with a little more reflection this disconnect should be expected. The absolute value of the derivative of a function says how rapidly the function is changing, not how much. A function can change rapidly, i.e. briefly have a large derivative, even if it is almost constant, if the changes are small enough. The derivative can even be large (in absolute value) most of the time, if it oscillates rapidly between positive and negative.

**I.5.1.14.** Here is perhaps the simplest example of this phenomenon (I am indebted to S. Strogatz for pointing it out). The diagonal of the unit square in \( \mathbb{R}^2 \) can be arbitrarily approximated uniformly by a “staircase” curve consisting of short alternating horizontal and vertical segments. Such a staircase curve clearly has arc length 2. So why doesn’t the diagonal also have length 2?

**I.5.1.15.** The nontrivial extra feature in the case of Archimedes’ argument (which incidentally was not his standard proof, which was a rigorous argument by double reductio ad absurdum, cf. ()) is that if \( f_n \) is the function whose graph is the lower wavy curve at the \( n \)'th stage, then not only does \( f_n \to 0 \) uniformly, but the derivatives of the \( f_n \) (actually the one-sided derivatives since \( f_n \) is not differentiable everywhere) also converge uniformly to 0, which is sufficient to conclude continuity of the arc lengths. To make the argument into a rigorous proof, some version of this additional argument is necessary.
I.5.1.16. There is an alternate way to phrase the argument which reduces (but does not eliminate) the technicalities, once the previous arguments about circumference have been done, necessary in any event for a full logical treatment. Inscribe a regular polygon with \( n \) sides in a circle of radius \( r \), and connect each vertex to the center of the circle by a line segment (see Figure I.6, from http://www.amsi.org.au/teacher_modules/the_circle.html).

![Figure I.6: Inscribed regular polygon with \( n = 12 \), divided into triangles.](image)

This divides the polygon into \( n \) congruent isosceles triangles of short side \( b_n \) and height \( h_n \). Each of these triangles thus has area \( \frac{1}{2}b_nh_n \). We have that \( nb_n = p_n \), the perimeter of the polygon, so the area of the polygon is

\[
A_n = n \cdot \frac{1}{2}b_nh_n = \frac{1}{2}p_nh_n.
\]

As \( n \to \infty \), \( p_n \to 2\pi r \) and \( h_n \to r \), so the area \( A \) of the circle is

\[
A = \lim_{n \to \infty} A_n = \lim_{n \to \infty} \frac{1}{2}p_nh_n = \frac{1}{2} \cdot 2\pi r \cdot r = \pi r^2.
\]

For this argument, we need to verify:

(i) Such a decomposition into triangles is possible, discussed in I.5.1.4 (i).

(ii) The area of the polygon is indeed the sum of the areas of the triangles, i.e. area is additive.

(iii) The area of a triangle with base \( b \) and height \( h \) is \( \frac{1}{2}bh \), which has an elementary proof by dissection once the area of a rectangle is known (in some approaches such as in [Har00], areas of triangles are the primitive data and no proof of this is necessary).

(iv) \( p_n \to 2\pi r \), the circumference argument already done.

(v) \( h_n \to r \), which takes some work to do carefully but is relatively easy. Indeed, by the Pythagorean Theorem,

\[
0 < r - h_n = \sqrt{\left(\frac{p_{2n}}{2n}\right)^2 - \left(\frac{p_n}{2n}\right)^2} = \sqrt{\frac{(p_{2n} + p_n)(p_{2n} - p_n)}{4n^2}} < \sqrt{\frac{4\pi r (2\pi r - p_n)}{4n^2}} < \frac{\sqrt{4\pi r (2\pi r)}}{2n} = \frac{\pi r \sqrt{2}}{n}
\]

which can be made arbitrarily small by choosing \( n \) large enough.
(vi) $A_n \to A$, one of the same difficulties as in Archimedes’ argument, but easier since it is essentially a monotone argument, traditionally called exhaustion (cf. XIII.1.2.7.).

(vii) The elementary limit fact (IV.1.3.21.(ii)) that a limit of a product is the product of the limits.

I.5.1.17. If $n$ is even, the triangles can be rearranged into a true parallelogram of length $p_n/2$ and slant height $r$ sitting inside the wavy parallelogram of Figure I.5. As $n \to \infty$, by (iv) the length of the base approaches $\pi r$, so the arc length difficulty in Archimedes’ argument is avoided (or, really, converted into the previous circumference argument). But the other difficulties remain.

Equal Division

I.5.1.18. On the subject of area, here is a problem which is described in some popular books as having an “obvious” solution. Let $E$ be a (bounded) region in the plane, and $L$ a line in the plane. Then there is a line parallel to $L$ which cuts $E$ into two regions of equal area.

The “obvious” solution is to begin with a line parallel to $L$ with $E$ entirely on one side. Then move this line “continuously” toward and past $E$ until it lies entirely on the other side of $E$. Since the area on the back side of the line begins at 0 and increases continuously to the area $A(E)$ of $E$, there must have been some point where this area is exactly $\frac{1}{2}A(E)$.

I.5.1.19. What is the difficulty? Some of these references gloss over this question entirely, suggesting there is none at all. But some acknowledge that the result depends on the Intermediate Value Theorem, which is a nontrivial (but perhaps geometrically “obvious”) fact. And there is another difficulty which never seems to be mentioned: if $A(t)$ is the area of the part $E_t$ of $E$ behind the line moved a distance $t$, why is $A(t)$ a continuous function of $t$? It is, but this fact is essentially the upward and downward continuity of Lebesgue measure in the plane (.), again a nontrivial fact, although if $E$ is bounded, continuity of $A(t)$ can be proved more easily by bounding $E_t \setminus E_s$ in a rectangle of fixed length and small width if $s$ is slightly less than $t$. These difficulties are, of course, in addition to the question of whether $A(E)$ and the $A(t)$ are even well defined and whether $A(E) = A(E_t) + A(E \setminus E_t)$ for all $t$, discussed earlier.

I.5.1.20. Some popular books go on to discuss the next version (the two-dimensional version of the ham sandwich theorem ()): if $D$ and $E$ are (bounded) regions of the plane, then there is a line which simultaneously cuts both $D$ and $E$ into regions of equal area. The “obvious” argument goes as follows: by the first part of the argument, for each $\theta$ there is a line $L(\theta)$ making an angle $\theta$ with the horizontal which cuts $E$ into two equal parts. If $\theta$ is varied from 0 to $\pi$, the line $L(\theta)$ passes across $D$ as before, and the division is reversed for $\theta = \pi$ from the division at $\theta = 0$, so there must be a $\theta$ for which the division of $D$ is equal.

There is an additional difficulty in this version, rarely if ever identified in popular references: the line $L(\theta)$ is not necessarily unique, e.g. if $E$ is not connected, so an $L(\theta)$ must be chosen for each $\theta$. It is not obvious, and is rather tricky to prove even in the case where the $L(\theta)$ are unique, that the lines $L(\theta)$ can be chosen to vary “continuously” with $\theta$. However, it is easier to show (although still a little tricky to prove rigorously) that if $D$ and $E$ are bounded, the area of the part of $D$ on the one side of $L(\theta)$ can be chosen to be a continuous function of $\theta$.
Sequences and Pattern Recognition

I.5.1.21. What is the next term in the sequence \((0, 1, 2, \ldots)\)? Most people would immediately answer "3". The "simplest" formula for the \(n\)th term \(x_n\) giving these values for \(n \leq 3\) would be \(x_n = n - 1\), so this is the “obvious” answer. However, there are many other possible formulas, and the “correct” answer depends on guessing what the poser of the problem had in mind. For example, if \(r\) is any real number \(\neq 3\), it is a simple exercise (solving a system of four linear equations in four unknowns) to find a cubic polynomial \(f\) such that \(f(1) = 0, f(2) = 1, f(3) = 2,\) and \(f(4) = r\).

I.5.1.22. There are other possible rules: for example, \(x_n\) could consist of \(n - 1\) followed by \(n - 1\) factorial signs (I am indebted to Vaughan Jones for this example.) Then

\[
\begin{align*}
  x_1 &= 0 \\
  x_2 &= 1! = 1 \\
  x_3 &= 2!! = (2!)! = 2 \\
  x_4 &= 3!!! = 720! \approx 2.6 \times 10^{1746}.
\end{align*}
\]

I.5.1.23. Here is another example. It appears different since it is not stated in purely mathematical terms, but in fact it is essentially identical. Suppose at 6 AM the temperature is 0°, at 9 AM it is 10°, and at noon it is 20°. What will the temperature be at 3 PM? 6 PM? 9 PM? Does anyone believe the temperature will be 50° at 9 PM? I suppose it’s not impossible, but it’s certainly not likely!

Actually, the only difference between this problem and the previous one is that there is more information: how do you know in the first problem that the terms of the sequence aren’t the temperatures (measured in 10° units) at three-hour intervals starting at 6 AM a certain day and location? Or maybe the \(n\)th term is the number of hits a certain baseball player gets in his first \(n\) at bats of the season, or the number of Heads in the first \(n\) tosses of a coin, or \(\ldots\). How do we know there is any rule or pattern at all - why couldn’t it just be a random sequence?

In the second problem meteorologists could predict with considerable accuracy what these temperatures will be, but they would need more information than is stated in the problem. However, in none of the problems can anyone figure out definitively what the next term will be without more information than is given in the problem (and in the coin tossing case no one could determine the next term even with more information).

I.5.1.24. The situation seems to be similar to a television comedy sketch which once appeared on Saturday Night Live, about a game show. The final question was: “I’m thinking of a color. What is it?”

I.5.1.25. But is it really similar? Being able to recognize a pattern in a sequence is an important mathematical skill. The above example is too easy to properly illustrate the point. A better example is the following sequence:

\[(2, 5, 10, 17, \ldots)\]

It is a bit more challenging, although not difficult, to come up with a pattern here. In fact, there are several natural ways to generate a sequence beginning with these terms, and it is not obvious what rule is “simplest” in this case.
I.5.1.26. This sequence appeared on a TerraNova standardized test for high school students. Unfortunately, it was phrased as a multiple choice question, asking for the next term; three of the choices for the answer were 24, 26, and 28.

The solution, as described in the Reno Gazette-Journal, is that the difference between successive terms increases by 2 at each step. The “correct” answer is thus 26.

I.5.1.27. Equivalently, this rule can be phrased as an algebraic formula giving the terms rather than a way of obtaining each new term from the last one. Use the function $f(x) = x^2 + 1$; then the $n$’th term of the sequence is the value of $f(x)$ for $x = n$. This formula gives the same fifth term, 26, as the published rule (in fact, all the terms of the sequences are the same.)

I.5.1.28. One can give a completely different rule. Let the $n$’th term be the sum of the first $n$ prime numbers. We then obtain the sequence $(2, 5, 10, 17, 28, 41, 58, \ldots )$, so the fifth term is one of the “incorrect” answers.

I.5.1.29. One can, of course, play the polynomial game here too. For example, let

$$f(x) = x^4 - 10x^3 + 36x^2 - 50x + 25.$$  

We then have $f(1) = 2$, $f(2) = 5$, $f(3) = 10$, $f(4) = 17$, and $f(5) = 50$. For another example,

$$g(x) = -\frac{1}{12} x^4 + \frac{5}{6} x^3 - \frac{23}{12} x^2 + \frac{25}{6} x - 1$$

has $g(1) = 2$, $g(2) = 5$, $g(3) = 10$, $g(4) = 17$, and $g(5) = 24$, another of the “incorrect” answers from the problem.

I.5.1.30. Most mathematicians are very critical of problems like this one. The objection is not to the problem itself, but the way it is phrased: as written, it is a psychology problem, not a mathematics problem; it is almost like asking, “In what country are the Alps located?” A much preferable way to phrase the problem is: “Describe a rule or pattern giving a sequence of numbers whose first four terms are 2, 5, 10, 17, and find the fifth term in the sequence according to your rule.” Unfortunately, there does not seem to be any reasonable way to phrase this type of problem as a multiple-choice question; the only acceptable multiple-choice version would be one with a choice (the correct answer!) “Cannot be determined from the information given.”

Questions like this are really very much like questions where letters must be unscrambled to form a word, if only one answer is “right”; see figure I.7.27

I.5.1.31. The widespread acceptance of the legitimacy of this type of question is symptomatic of a general human problem: there is often more than one explanation for a phenomenon, but an individual or community often settles on one explanation as the “right” one and rejects any alternatives or even the

---

27The company which produced the instructional software we use in calculus classes was bought out by another company. To use the software, I needed an account with the new company. When I tried to create one, I found that one had already been set up automatically, without anyone telling me. To log in, I needed the password which no one had told me either. When I clicked on “forgot password,” I was required to give the answer to my “Secret Question.” It would not tell me what the question was, presumably because it was secret.
possibility of an alternative, even sometimes in light of new evidence. Scientific theories which contradict, or appear to contradict, established religious beliefs are a notorious example. In the sequence problem case, people untrained in mathematics may have the excuse that they think such ambiguity cannot happen in mathematics and that any mathematically phrased problem has a definitive answer; mathematicians know better!

I.5.1.32. The question remains, however, whether it makes sense in general to try to make precise the notion of the “simplest” rule for a given initial sequence segment. This notion would appear to be highly subjective and mathematically imprecise; however, we should not reject out of hand the possibility that it could have some meaning and that the ability to recognize simple patterns is indicative of general mathematical ability. Thus the proper formulation and use of this type of question on standardized tests is a matter for legitimate debate and study.

I.5.1.33. One possible “simplest” formula for a sequence in which \( n \) terms are given is the unique polynomial of degree \( \leq n - 1 \) giving these values at 1, \ldots, \( n \) (cf. V.11.1.1.), especially if (as in the above cases) the degree of this polynomial is smaller than \( n - 1 \). However, this type of definition excludes simple and natural recursive rules which can be found for certain sequences. A good example is the famous Fibonacci sequence, which arises naturally in many applied problems such as in biology. The first few terms are (1, 1, 2, 3, 5, 8, \ldots), and each term, beginning with the third, is the sum of the two previous terms.

I.5.1.34. Here is another example, taken from [CG96, p. 76-79], where the “simplest” formula is not a polynomial. What is the next term in the sequence (1, 2, 4, 8, 16, \ldots)? The “obvious” answer is 32, which is correct if each term is twice the preceding term (i.e. the \( n \)'th term is \( 2^{n-1} \)). But here is another natural
(but somewhat more complicated) sequence beginning with the same terms. To get the $n$’th term $x_n$, place $n$ points on a circle and connect each pair of points with a line segment (chord), and count the number of regions the disk bounded by the circle is divided into. It is not too hard to see that the answer does not depend on how the points are arranged on the circle, as long as no three of the chords pass through the same interior point (so, for example, the points could not be evenly spaced around the circle if $n$ is even and more than 4). It is easy to check that $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 8$, and $x_5 = 16$. But $x_6 = 31$, not 32, and the difference between $x_n$ and $2^{n-1}$ becomes more dramatic as $n$ increases (it is easily seen that $2^{n-1}$ is far too large if $n$ is, say, 20 or 30). In fact, it can be shown that

$$x_n = \binom{n-1}{0} + \binom{n-1}{1} + \binom{n-1}{2} + \binom{n-1}{3} + \binom{n-1}{4}$$

(where $\binom{n}{k}$ is the binomial coefficient (V.17.7.1.), the number of ways of choosing $k$ elements from a set of $m$ elements), which is a fourth-degree polynomial

$$x_n = \frac{1}{24}(n^4 - 6n^3 + 23n^2 - 18n + 24)$$

This is the unique polynomial of degree $\leq 4$ giving the specified terms for $n \leq 5$, and for the circle division rule gives $x_n$ for all $n$.

**I.5.1.35.** Here is yet another example. Consider the sequence $(1, 4, 6, 13, \ldots)$. These are the first terms of the busy beaver function $\Sigma$, where $\Sigma(n)$ is the maximum number of 1’s an $n$-state Turing machine with alphabet \{0, 1\} can write on a blank tape (all 0’s) and then halt (there are only finitely many such Turing machines, so the maximum exists). $\Sigma$ is an uncomputable function since there is no algorithm for determining whether a Turing machine halts, and in fact grows faster than any computable function. No values of $\Sigma(n)$ are known for $n > 4$; $\Sigma(5)$ is known to be at least 4098, $\Sigma(6)$ is at least $3.5 \times 10^{18267}$, and

$$\Sigma(7) \geq 10^{10^{10^{18705352}}}$$

(probably much larger). In general, $\Sigma(2k) > 3^{4k-2}$ (II.8.6.16.), and is asymptotically much larger.

Not only is the whole function $\Sigma$ uncomputable\(^{28}\), but there are even specific $n$’s for which $\Sigma(n)$ is in principle uncomputable (the smallest $n$ known to have this property is 1919, since a Turing machine of this size which just writes 1’s can be constructed which halts if and only if ZF set theory is inconsistent; but it is generally believed there are much smaller such $n$, perhaps as small as 10).

**I.5.1.36.** Readers interested in exploring mathematical sequences should check out the website

http://oeis.org

This site lets the user search for naturally-occurring sequences beginning with any specified terms. For example, the search gives more than 70 mathematical sequences beginning with (or including as a segment) 2,5,10,17. For example, if $x_n$ is the length of the longest uncrossed knight’s path on an $(n + 2) \times (n + 2)$ chessboard, then the first few terms are $(2, 5, 10, 17, 24, 35, \ldots)$.

---

\(^{28}\)There is no universally accepted definition of a computable function, but we can take this to mean Turing computable. A function $f : N \to N$ is Turing computable if there is a Turing machine with alphabet \{0, 1\} such that, for any $n \in N$, if the machine is given an input tape with $n$ 1’s it outputs $f(n)$ 1’s on the tape and then halts. This is equivalent to several other natural notions of computability for functions. Computability of $f(n)$ for a fixed $n$ is more difficult to define precisely, but means there is a definite finite procedure (perhaps dependent on $n$) for calculating the numerical value from the definition of the function, or a finite proof that a specific number $m$ (specified by a particular algorithmic or inductive numerical notation, e.g. II.8.7.4.) is equal to $f(n)$. 

28
Euclid’s Parallel Postulate

I.5.1.37. Next, we consider an issue which vexed mathematicians for centuries: Euclid’s Parallel Postulate. I believe one of the important reasons it took so long to understand this situation was that mathematicians were misled and constrained in their thoughts by subtle implications of the terms Euclid used in the formulation. (Euclid himself was misled: in some of his proofs he used additional assertions which must have seemed “obvious” to him, but which do not follow from his axioms.)

I.5.1.38. One of Euclid’s postulates (or axioms) for geometry, called the Parallel Postulate, states that “if \( L \) is a line and \( p \) a point not on \( L \), there is one and only one line through \( p \) parallel to \( L \).” (Actually Euclid stated this postulate in a rather different and more complicated way; the version we have stated is often called Playfair’s axiom, although it was known and used at least as far back as Proclus. See [Wil, I.1] for a discussion of the origin and demise of the distinction between axioms and postulates.) In Euclid’s time, and for a very long time afterwards, axioms were supposed to be “self-evident truths,” and the Parallel Postulate is indeed true in ordinary (\( \mathbb{R}^2 \)) geometry. Mathematicians long wondered not whether the Parallel Postulate was “true”, but whether it could be proved from Euclid’s other axioms (since the actual statement was sufficiently complicated as not to be “self-evident.”)

I.5.1.39. Although over the centuries various mathematicians assumed the Parallel Postulate was false, and even proved what are now regarded as theorems of noneuclidean geometry, it was always done with the goal of obtaining a contradiction; it apparently never occurred to mathematicians until the nineteenth century (with the possible exception of J. Lambert in the eighteenth) that there really could be “geometries” satisfying the rest of Euclid’s axioms but not the Parallel Postulate, or, in modern language, models for Euclidean geometry with the Parallel Postulate replaced by its negation. Once the possibility was contemplated, it did not take long for such geometries to be found, in fact independently by several mathematicians: Gauss, Bolyai, and Lobachevskii for hyperbolic geometry, and Riemann for elliptic (some of Euclid’s other postulates arguably have to be modified a little in this case – see below); an actual model of hyperbolic geometry was not constructed until later in the century, by Beltrami. It follows immediately that the Parallel Postulate is independent of the other axioms of Euclidean geometry, i.e. cannot be proved (or disproved!) from them. (In fact, we know now, via model theory, that the converse is also true: if a statement such as the Parallel Postulate is independent of the other axioms, then there is a model for the other axioms in which the statement is false, although this was far from clear in the nineteenth century.)

I.5.1.40. While the mathematical atmosphere in the nineteenth century was much more conducive to the thought processes leading to these examples than in earlier times, I cannot help but think that the language used by Euclid had a stifling effect on creative thought about geometry. Euclid used the terms “point” and “line” as essentially undefined objects (he did give “definitions” which would not be considered true definitions today, e.g. “defining” a point to be “that which has no part” and a line to be “that which has length but no breadth”), but the terms used were strongly suggestive of a particular interpretation in which the Parallel Postulate is rather “obviously” true. It finally took some very smart and creative thinkers to realize that the terms “point” and “line” could be interpreted to mean something other than what Euclid evidently had in mind, e.g. what we would now call geodesics on surfaces could play the role of lines. This thinking has evolved further (pioneered by Poincaré, and culminating in Hilbert’s treatment of the Foundations of Geometry), to the degree that we now realize that “point” and “line” can refer to any kind of mathematical objects at all, so long as the axioms (with suitable interpretation) hold for them.
I.5.1.41. I wonder what would have happened if Euclid had used some non-suggestive terms (such as “bxptq” and “zkmr”) instead of “point” and “line.” One of two things would probably have happened: either models of non-Euclidean geometry would have been discovered sooner, or (more probably) no one would have continued to study Euclid at all since his work would not have seemed to be “about” anything! Suggestive terminology has undeniable advantages so long as the pitfalls are understood.

I.5.1.42. Incidentally, I am pretty tired of reading uninformed nonsense about how the development of non-Euclidean geometry somehow divorced geometry from physical reality and precipitated a “crisis” in mathematics, especially in its foundations. Such an argument is more worthy of Alfred E. Neuman than John von Neumann, and is based on a very simplistic view of “reality.” Although for a brief time the new ideas were evidently unsettling to some mathematicians, I do not think it was a prevalent attitude among mathematicians to be upset by the developments and fear that mathematics might be a house of cards which was about to collapse; in fact, it seems that most good mathematicians thought it was an exciting time in mathematics when the frontiers of knowledge and understanding were expanding rapidly. (There was a controversy in the German mathematical community in the 1920s called the “foundational crisis,” but this was really a dispute over which parts of mathematics, if any, could be considered “true” or “meaningful” and not whether mathematics held together logically.) It is certainly true that the geometric discoveries, along with the roughly simultaneous “arithmetization” of analysis, led to a profound revolution in the nature and conception of mathematics. But even before Einstein came along to definitively discredit the idea that the geometry of physical space is Euclidean (Poincaré, Hilbert, and Minkowski played important roles too), it could be seen that Riemann’s spherical (elliptic) geometry was in fact the geometry of the surface of the earth, physical geometry of the most basic sort (recall that the very word geometry comes from Greek meaning “measuring the earth”). Hyperbolic geometry has less immediate direct application to the familiar world, but it cannot be reasonably argued that it has no connection with physical reality. Indeed, general relativity could not have been developed without geometric ideas which were outgrowths of the discovery of non-Euclidean geometry.

“The value of non-Euclidean geometry lies in its ability to liberate us from preconceived ideas in preparation for the time when exploration of physical laws might demand some geometry other than the Euclidean.”

B. Riemann

Anyone who thinks non-Euclidean geometry is divorced from reality should read [Pen05] and [MTW73]. See [Gra08] for a thorough and intelligent study of the “modernist” revolution in mathematics without resort to the “crisis” characterization.

“[T]he development of mathematics may seem to diverge from what it had been set up to achieve, namely to simply reflect physical behaviour. Yet, in many instances, this drive for mathematical consistency and elegance takes us to mathematical structures and concepts which turn out to mirror the physical world in a much deeper and broad-ranging way than those that we started with.”

Roger Penrose

29 cf. [Gre93, p. 371].
30 [Pen05, p. 60].
Euclid’s Postulates

I.5.1.43. Here are EUCLID’s five postulates, according to [Euc02]31 (other sources give slightly different wording):

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.
5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

Postulate 5 is EUCLID’s version of the Parallel Postulate (note that EUCLID neglected to specify that the two lines not both pass through the same point of the line falling on them! Maybe this is implicit in the “interior angle” term which is actually not defined.) It is conventional wisdom, stated in many references, that Playfair’s Axiom is equivalent to EUCLID’s Postulate 5, in the presence of Postulates 1–4; however, as we shall see, this equivalence requires that some things be read into Postulates 1–3 which are (arguably) not really there. These additional assertions evidently seemed “obvious” to everyone from EUCLID’s time until the nineteenth century (M. PASCH seems to have been the first to notice) and apparently escaped the notice of everyone who tried to deduce Postulate 5 or its believed equivalent Playfair’s Axiom from the other postulates (some seem to escape the notice of even some contemporary writers!)

Does Spherical Geometry Satisfy Euclid’s Postulates?

A good case study to illustrate the principles is spherical geometry. Whether spherical geometry satisfies EUCLID’s postulates is a matter of interpretation which also illustrates how unwarranted consequences can be read into statements.

I.5.1.44. In spherical geometry, a sphere $S$ is fixed. “Points” of the geometry are just points of $S$. “Lines” (“straight lines”) of the geometry are great circles. “Line segments” (“finite straight lines”) are (usually minor) arcs of great circles. “Circles” are ordinary circles on $S$, not necessarily great circles; we will also call singleton sets circles. The center of a circle has its usual interpretation (a circle actually has two antipodal centers!); the center of a singleton circle is the point itself (or the antipodal point). “Angles,” particularly “right angles,” are interpreted in the usual way using tangent lines.

Note that it does not actually contradict the postulates that certain circles are also “straight lines” (or single points), or that circles with different centers can coincide – it only contradicts our intuition about what these terms should “mean.” (Note that there is an ambiguity in EUCLID’s definitions between “circle” and “disk”; in spherical geometry, distinct disks can have the same boundary circle.)

---

I.5.1.45. The first observation is that spherical geometry actually satisfies Postulate 5 as Euclid stated it. Since it obviously does not satisfy Playfair’s Axiom (according to Euclid’s definition, “parallel” means “nonintersecting,” so there are no parallel lines in spherical geometry), either spherical geometry fails to satisfy all of the other four postulates or Playfair’s Axiom is not really equivalent to Postulate 5.

I.5.1.46. Next look at Postulate 4. What this postulate evidently means is that there is an assumed notion of congruence. An easy way to accomplish this (although not the most general way) is to assume there is a set (group) of sufficiently many “rigid motions” available to translate and rotate figures (specifically, which is transitive on points and lines). There is such a group (SO(3)) for the sphere.

I.5.1.47. The suspect postulates in spherical geometry are 1–3. First consider 1. If \( p \) and \( q \) are distinct points on a sphere, there is clearly a line through \( p \) and \( q \), so if Postulate 1 is read strictly as Euclid stated it, it is satisfied in spherical geometry. However, most readers, beginning with Euclid himself, include into Postulate 1 an implicit requirement that the line be unique, which fails on the sphere if \( p \) and \( q \) are antipodal points. So whether spherical geometry satisfies Postulate 1 is a matter of interpretation. (In the projective plane, the quotient of the sphere by the antipodal map, the line through two given points is unique, although the line segment between two points is not necessarily unique. Euclid does not explicitly distinguish in his postulates between lines and line segments, but there is a clear distinction between the two in Postulate 2 and perhaps implicitly in Postulate 1.)

I.5.1.48. Now look at Postulate 2, which leads to similar but more interesting interpretation questions. The statement is somewhat vague, and can be interpreted in several reasonable ways. It might mean simply that every line segment is contained in a line (which may even be equal to the line segment). But the word produce might implicitly mean that the line is larger than the segment. We might expect that it means that the line segment between, say, \( p \) and \( q \) can be “extended beyond \( q \),” whatever that means from the Postulates, and similarly beyond \( p \). It might then mean that a line segment from \( p \) to \( q \) could always be continued to a point \( r \) giving a line segment from \( p \) to \( r \) containing \( q \), with the line segment from \( q \) to \( r \) the magnitude of a fixed reference segment, and similarly beyond \( p \). With this interpretation, “lines” in spherical geometry, i.e. great circles, satisfy the postulate. But someone thinking about ordinary Euclidean geometry reading the postulate would probably interpret it as meaning that the extension of the line segment consists entirely of distinct points, i.e. the extension beyond \( q \) does not intersect the extension beyond \( p \), which fails on the sphere. Euclid himself made and used this assumption, sometimes called the “infinitude of lines,” without mentioning it in his postulates. The “infinitude of lines” interpretation is necessary if one wants to have a notion of one point being “between” two others on a line, which is important in Euclidean geometry; there is no such notion in spherical geometry. “Betweenness” is not mentioned in Euclid’s postulates, but is a fundamental concept in Hilbert’s more careful axioms for Euclidean geometry. Actually, “betweenness” is implicit in the notion of line segment with endpoints; in the presence of Postulates 1 and 2, along with a weak uniqueness statement that three distinct points can lie on at most one line (which does not follow from Euclid’s postulates as stated), the “infinitude of lines” is equivalent to the statement that if three distinct points are on a line, exactly one is “between” the other two.

The “infinitude of lines” is distinct from the question of whether the extension in Postulate 2 is unique, i.e. whether two different lines can have a segment in common, another assumption made by Euclid not included, but arguably implicit, in the postulate; extensions are unique in spherical geometry. (There is another interpretation ambiguity: what does the word “continuously” mean in this postulate? Any reasonable interpretation would seem to be satisfied in spherical geometry, although some authors have suggested that
"continuously" refers to the infinitude of lines, which seems doubtful to me.) I think it is somewhat harder to argue that the “infinitude of lines” is implicit in Postulate 2 than that extensions are unique. Thus whether spherical geometry satisfies Postulate 2 is again a matter of interpretation.

I.5.1.49. Postulate 3 appears to be false in spherical geometry; there is a limit to the radius of circles on a sphere. But this conclusion depends on a modern interpretation of the phrase “any distance.” Euclid did not use the notion of numerical distance between points; in fact, in ancient Greece a clean distinction was made between numbers (natural numbers), which denoted quantity, and “magnitudes,” which denoted size. To Euclid, a “distance” would have meant the “magnitude” of a line segment between two points. Thus he arguably would have agreed that Postulate 3 could have been equivalently phrased:

3’. Given points \( p \) and \( q \), a circle can be drawn with center \( p \) passing through \( q \).

In fact, this is the form of Postulate 3 always used in his proofs (and in straightedge-and-compass constructions); cf. [Gre93]. With this interpretation, spherical geometry does indeed satisfy Postulate 3. There is a slight problem if \( p \) and \( q \) are antipodal points on the sphere; but this problem evaporates if we allow a circle to be a single point, or if we pass to the projective plane, although this could cause (slight) difficulties with Postulate 5 since a line does not divide the projective plane.

I.5.1.50. Note also that the notion of “magnitude” of a line segment, as well as Postulate 4, only make sense if there is an assumed notion of congruence. As previously noted, there is such a notion for the sphere (and for the projective plane).

I.5.1.51. Thus it can be argued, although perhaps somewhat weakly, that spherical geometry satisfies all of Euclid’s postulates (it can be argued somewhat less weakly that the projective plane satisfies all the postulates), and also that the equivalence of Playfair’s Axiom with Postulate 5 depends on reading extra assertions into the other postulates.

Other Interpretations of Euclid’s Postulates

I.5.1.52. One could also consider the set \( P \) of points in \( \mathbb{R}^2 \) with rational coordinates (rational points). Lines in \( P \) will consist of all rational points on an ordinary line through two rational points, and circles in \( P \) will consist of rational points on an ordinary circle with rational center passing through a rational point. (There is a dense set of rational points on any such line or circle.) Then Postulates 1, 2, and 5 hold in \( P \), including uniqueness and the “infinitude of lines,” and Postulate 3 holds in the form of 3’, with no bound on the radii of circles. The (slight) difficulty this time is with Postulate 4, since the group of rigid motions is too small to allow comparison of all pairs of right angles; but the group of similarities is large enough to allow such comparison in a well-defined way, so Postulate 4 can be said to be satisfied in \( P \). (Comparison of magnitudes in \( P \) is also not possible in general for the same reason.) But in this geometry Euclid’s first proposition, construction of an equilateral triangle with two given vertices, fails since the circle with center \( p \) passing through \( q \) and the circle with center \( q \) passing through \( p \) do not intersect even though they “overlap.” Euclid assumed in his proof that such circles intersect, an assumption which does not follow from his postulates. (The assumption also fails in spherical geometry if \( p \) and \( q \) are too far apart.) In fact, \( P \) contains no equilateral triangles.
Consider also the following simple system. The space consists of exactly four points. Lines are defined to be sets containing exactly two points. A circle is a set of exactly three points; the fourth point is called the center of the circle. A right angle is a set of three points with one of them designated as the vertex. Two lines are parallel if they are disjoint. The angle between two nonparallel lines is defined to be the right angle consisting of the union of the lines with the intersection as vertex. Euclid’s postulates 1, 2, and 3’ are satisfied (except possibly for the word “continuously” in Postulate 2, whatever it is supposed to mean), including uniqueness in 1 and 2 (and in 3’); the “infinitude of lines” is also trivially satisfied since the extension of a line segment is just the segment itself. If the group of rigid motions is taken to be the group of all permutations of the space, Postulate 4 is also satisfied. And Postulate 5 is vacuously satisfied since the angle between any two nonparallel lines is always a right angle. Playfair’s Axiom also holds in this space. This example and the previous one show that Euclid’s postulates do not uniquely determine a geometry.

Note that it is also consistent with the postulates that there be no points, and any number of lines (including none), or that there be exactly one point and any number of lines, or that there be any number of points and exactly one line containing all of them. In these cases, some or all of the postulates will hold vacuously. These situations could be ruled out by assuming that there exist three points not all on the same line, another assumption Euclid made in many places.

There was a counterexample staring everyone in the face who tried to deduce Euclid’s Parallel Postulate from the other postulates, if they had been open-minded enough to realize it! (To be fair, it would be unreasonable to expect them to have recognized it before the abstraction of the nineteenth or even twentieth century.) Just take \( \mathbb{R}^3 \) with the usual points, lines, and line segments, and interpret a “circle” to be a sphere. Euclid’s first four postulates hold, even with the extra uniqueness provisions added. But if we use Euclid’s definition that “parallel” means not intersecting, Playfair’s Axiom fails. There is a semantic problem with Euclid’s fifth postulate because it refers to a “side” of a line, but this can be bypassed simply by declaring that a line has only one side (i.e. the side of the line consists of all points not on the line); with this interpretation, the fifth postulate also fails.

This example would surely have been regarded as cheating (even if abstract interpretations were accepted allowing a sphere to be called a “circle”), since Euclid’s postulates were clearly designed to define plane geometry. But there is nothing in any of Euclid’s postulates or common notions which would force two-dimensionality or rule out examples of this sort (he could have easily done so with an additional postulate). From modern hindsight, this example would immediately show that neither Euclid’s fifth postulate nor Playfair’s Axiom can be proved from the other assumptions Euclid spelled out.

If Euclid’s definitions, imperfect as they are, are closely examined, it could be argued that they include implicit additional postulates which rule out examples such as I.5.1.52–I.5.1.55, and even force the space considered to be at least what we would call a surface or 2-manifold. He defines a “line” (meaning what we would call a “curve”; he goes on to define a “straight line” to be a line satisfying a condition I find incomprehensible) to be “that which has length but no breadth.” Length and breadth are not defined, but one could argue that this definition means that lines must be topologically one-dimensional, an interpretation reinforced by the word “continuously” in Postulate 2 which is otherwise mysterious, and must have a well-defined arc length. In particular, lines (curves) must be isometric to intervals in \( \mathbb{R} \), anticipating modern
principles of completeness and measure which were far in the future in ancient Greece.

Euclid’s definition of a circle is “a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure equal one another.” This actually defines what we would call a disk, but if we take the term “circle” to mean the boundary of such a figure, it must be a line or curve in the above sense, and must implicitly have an interior and exterior as well as a center. (Note though that this definition, like some of his others, is circular in that it refers to a “plane figure” although the whole theory is intended to establish what a “plane figure” is.)

I.5.1.57. Thus, although Euclid’s postulates and axioms themselves fall far short of adequacy to describe plane geometry, when the additional postulates arguably implicit in his definitions are added, as most interpretations over the centuries have done, they come remarkably close.

“Proclus shared the human problem of distinguishing between the obvious and that which only appears to be obvious.”

George E. Martin^{32}

“One of the pitfalls of working with a deductive system is too great familiarity with the subject matter of the system. It is this pitfall that accounts for most of the blemishes in Euclid’s Elements.”

Howard Eves^{33}

I.5.1.58. My purpose in discussing these matters is not to find fault with Euclid, but simply to illustrate how easy it is to jump to conclusions not supported rigorously. (In [?] it is stated that only “little people” like to point out flaws in Euclid; then the author himself describes some of the flaws over the next several pages!) Although we may find flaws in it, Euclid’s Elements was a remarkable intellectual achievement (not just by Euclid, but also by his predecessors who developed much of the theory) which is a leading candidate for the greatest mathematical work of all time, and remains the most widely read, studied, and influential non-religious book ever written on any subject. No mathematical work, with the arguable exception of Archimedes’ writings, rivaled it in importance at least until Newton’s Principia two thousand years later (and to call Principia a mathematical work might be a bit of a stretch), and the Elements was more “modern” and rigorous than anything else written before about the nineteenth century.

“If mathematics does not teach us how to think correctly, it at least teaches us how easy it is to think incorrectly.”

Harald Bohr^{34}

“The genius of Euclid is insufficiently celebrated. When he sat down to write his Elements of Geometry, he probably did not know that he was going to create a work that would last for centuries and would forever change how the human race thought, not just about geometry but about truth and how to get at it. He probably did not realize that he was one of the chief figures

---

^{32}Ped89, p. 138.
^{33}Ped89, p. ix.

in a revolution that, in a mere three hundred years or so, made our world possible by elevating reason to its proper place. Probably he was thinking only that it would be a good idea to gather up all the pieces of mathematics that were scattered around (some of them, he may have thought, shockingly badly written) and put them in one place, neatly and logically organized. After all, Euclid worked in a library and may have had some of the habits of mind that librarians are either born with or quickly absorb from the atmosphere of libraries. But whatever his motives, Euclid deserves to be remembered.

Euclid’s book has been a model of deductive reasoning ever since it appeared. Deductive reasoning is highly prized as a method of getting at truth, since from truth only truth can follow by deduction. Science is based on it, and even lawyers and politicians attempt to use it at times. It is such a part of our civilization that we cannot imagine a real civilization without it. Deductive reasoning has made us what we are. That is why we insist that high-school students be exposed to a version of Euclid: it is not that any of them will ever have to prove a theorem about triangles, it is just our way of acknowledging our debt to deductive reasoning.”

_Underwood Dudley_\textsuperscript{35}

\textsuperscript{35}[Dud92, p. 78].
“Obvious” is Subjective

I.5.1.59. Finally, we turn to a somewhat different issue: it is a highly subjective matter what details of an argument an author or lecturer feels need to be explained and justified, and which ones can be safely considered “obvious.” Opinions about this can vary widely from person to person, and even from time to time with the same person, and are not always justifiable. Rather than discuss this issue further, let us just recall a standard joke which illustrates it nicely:

A mathematician was giving a lecture, and at one point he said, “It is obvious that . . .” He then paused in mid-sentence, a concerned look came over his face, and he paced back and forth, scratching his head and stroking his chin, deep in thought. After about five minutes, his face suddenly brightened and he exclaimed, “Ah, yes! It is obvious.”

If an instructor ever tells you something is obvious and you aren’t so sure, or it is not obvious to you, don’t be reluctant or embarrassed to ask for an explanation. It may turn out that it is not so obvious to the instructor either!

Does Familiarity Breed Contempt?

“I thought at first you had done something clever. But [after Holmes’s explanation] I see that there was nothing in it after all.”

_Jabez Wilson (to Sherlock Holmes)_

I.5.1.60. It is a pervasive human characteristic to be more in awe of things we do not understand. Once we see a rational explanation of something, the wonder of it tends to fade and it becomes more mundane or commonplace. We, like Jabez Wilson, may even be fooled into thinking it is obvious or easy, and we often underappreciate the effort and/or genius required to find the phenomenon or its explanation in the first place.

Mathematicians are not immune from this tendency. Consider, for example, the wonderful (possibly apocryphal) description by William Lilly of the first encounter between Henry Briggs, one of the principal developers of the theory of logarithms, and John Napier, their inventor, at Napier’s home in 1615:

“At first, when the Lord Napier, or Marchiston, made publick his Logarithms, Mr. Briggs, then reader of the astronomy lecture at Gresham-College in London, was so surprized with admiration of them, that he could have no quietness in himself, until he had seen that noble person the Lord Marchiston, whose only invention they were: he acquaints John Marr herewith, who went into Scotland before Mr. Briggs, purposely to be there when these two so learned persons should meet. Mr. Briggs appoints a certain day when to meet at Edinburgh: but failing thereof, the Lord Napier was doubtful he would not come. It happened one day as John Marr and the Lord Napier were speaking of Mr. Briggs: ‘Ah, John,’ saith Marchiston, ‘Mr. Briggs will not now come:’ at the very instant one knocks at the gate; John Marr hasted down, and it proved Mr. Briggs, to his great contentment. He brings Mr. Briggs up into my Lord’s chamber, where almost one

---

37 _History of his Life and Times._
quarter of an hour was spent, each beholding the other almost with admiration, before one word was spoke: at last Mr. Briggs began.

‘My Lord, I have undertaken this long journey purposely to see your person, and to know by what engine of wit or ingenuity you came first to think of this most excellent help unto astronomy, viz. the Logarithms; but, my Lord, being by you found out, I wonder no body else found it out before, when, now known, it is so easy.’

Mathematicians tend to become particularly blasé about parts of elementary mathematics as they attain higher levels of knowledge and sophistication in the subject (perhaps including some of the mathematics we discuss in this book). The fact that in the not-too-distant past some of this mathematics represented the pinnacle of human thought can be overlooked in its familiarity to us.

“A mathematician’s first steps into unknown territory constitute the first phase of a familiar cycle.

After the darkness comes a faint, faint glimmer of light, just enough to make you think that something is there, almost within reach, waiting to be discovered. . . . Then, after the faint, faint glimmer, if all goes well, you unravel the thread – and suddenly it’s broad daylight! You’re full of confidence, you want to tell anyone who will listen about what you’ve found.

And then, after day has broken, after the sun has climbed high into the sky, a phase of depression inevitably follows. You lose all faith in the importance of what you’ve achieved. Any idiot could have done what you’ve done, go find yourself a more worthwhile problem and make something of your life. Thus the cycle of mathematical research . . .” [Emphasis original]

C. Villani38

I.5.1.61. There is also a human tendency to need some sort of explanation of mysterious phenomena. This is the origin of most religion; many people are satisfied and comforted with religious explanations for things with no apparent rational explanation. However, I have observed that mathematicians often have quite the opposite instinct: they demand a rational explanation, and do not properly appreciate something, especially a mathematical result, until they understand it. (I do not mean to suggest that mathematics and religion are incompatible; indeed, many mathematicians, including great ones, have been deeply religious.) In fact, the drive to understand phenomena and revel in the understanding might tend to be one characteristic of people with a predilection towards mathematics, although I do not want to attempt a universal generalization here; this tendency may more be characteristic in distinguishing people who are intellectually capable and energetic, whatever their academic interests. In any event, a healthy curiosity about the world, both concrete and abstract, seems to be an almost necessary attribute of anyone who is successful in mathematics.

I.5.2. Set Theory and Logic

One of the main goals of many (but not all) mathematicians in the early 20th century was to carefully axiomatize all of mathematics, beginning with set theory. The hope was that a complete list of all theorems of mathematics could be generated, and there would be a mechanical process by which the truth or falsity of any mathematical statement could be determined. This hope turned out to be impossible, but at least set theory has been put on a generally accepted axiomatic footing.

38[Vil16, p. 243].
*****Rewrite*****: Through the early 20th century, and to a small extent even today, a debate raged between two schools of mathematics, the “intuitionists” and the “formalists”. The current consensus is that the truth lies somewhere between the two camps. On the one hand, the intuitionists’ belief that Cantor’s theory of the transfinite, and the subsequent formalizations of set theory and mathematical logic, are unsound and have no place in real mathematics has been thoroughly discredited; it is now almost universally accepted that these theories have a legitimate and even central place in the world of mathematics. On the other hand, the intuitionists’ main point seems valid: many (perhaps most) mathematical concepts appear to have an intrinsic meaning within the human intellect quite beyond just being a collection of symbols.

Gödel’s Incompleteness Theorem (1931) showed that it is not possible to completely carry out the program described in the first paragraph of this section. Perhaps it is just as well: while it is certainly a worthwhile endeavor to try to put the foundations of mathematics on a sound logical footing, attempting to reduce all of mathematics to mechanical symbol manipulation would rob mathematics of much of its richness, beauty, and usefulness. As Ian Stewart puts it in his book, Nature’s Numbers [?]:

“Textbooks of mathematical logic say that a proof is a sequence of statements, each of which either follows from previous statements in the sequence or from agreed axioms – unproved but explicitly stated assumptions that in effect define the area of mathematics being studied. This is about as informative as describing a novel as a sequence of sentences, each of which either sets up an agreed context or follows credibly from previous sentences. Both definitions miss the essential point: that both a proof and a novel must tell an interesting story. They do capture a secondary point, that the story must be convincing, and they also describe the overall format to be used, but a good story line is the most important feature of all.”

Here are other trenchant observations on the role of proof in mathematics:

“If a mathematical statement is false, there will be no proofs, but if it is true, there will be an endless variety of proofs, not just one! Proofs are not impersonal, they express the personality of their creator/discoverer just as much as literary efforts do. If something important is true, there will be many reasons that it is true, many proofs of that fact. ... And each proof will emphasize different aspects of the problem, each proof will lead in different directions. Each one will have different corollaries, different generalizations ... Mathematical facts are not isolated, they are woven into a vast spider’s web of interconnections.”

Gregory Chaitin\textsuperscript{39}

“Appreciating a theorem in mathematics is rather like watching an episode of Columbo: the line of reasoning by which the detective solves the mystery is more important than the identity of the murderer.”

C. Villani\textsuperscript{40}

\textsuperscript{39}[Cha05, p. 23-24].
\textsuperscript{40}[Vil16, p. 32].
I.5.3. Proofs vs. Formal Proofs (or, How Mathematics Really Works)

“A proof is a proof. What kind of a proof? It’s a proof. A proof is a proof. And when you have a good proof, it’s because it’s proven.”

Jean Chrétien, Prime Minister of Canada

I.5.3.1. The standard logical foundation of mathematics is that everything is defined within set theory, beginning with a set of axioms, and that all arguments and proofs are done using the standard rules of logic. Every mathematical statement can be phrased in this way, and every theorem or proposition needs to have a formal proof from the axioms in which every statement follows from the previous ones by an application of one of the rules of formal logic. Free use can be made of previously proved statements (whose proofs technically then become pieces of the proof of the new statement, in the nature of a “subroutine”). In this way, a large edifice is gradually created.

In practice, it is impossibly complicated and tedious to translate all but the simplest mathematical statements into foundational form and write out formal proofs, and most working mathematicians rarely if ever contemplate doing this with their mathematics. In fact, one can say that much of mathematics consists of developing efficient notation and terminology to allow rigorous analysis of complex mathematical ideas without reducing to the foundational level.

“By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and, in effect, increases the mental power of the race.”

A. N. Whitehead

I.5.3.2. Mathematicians understand the term “proof” to mean something different from “formal proof”. Mathematical proofs are in essence outlines of formal proofs, and the usual (although not universal) unwritten rule is that the proof should be written in sufficient detail that a competent (and perhaps masochistic) reader, with sufficient time and effort, could translate it into a formal proof. As in the case of formal proofs, free use can be made of other results whose proofs have already been accepted. (It is a gray area whether the author of a new proof is expected to carefully check the validity of the proofs of other results used.) This standard is, of course, subjective, and also depends in part on the audience; a proof written by a specialist to be read by other specialists can be written more concisely than would be possible if intended for nonspecialists or students. The standard for acceptance of a proof of a new result is generally a consensus of experts that the proof is correct and complete enough to meet the formal translation criterion.

“The essence of a proof is that it compels belief.”

Pierre de Fermat


43 [Wei07, p. 50].
I.5.3.3. It should be noted, however, that mathematicians sometimes use and accept arguments of a type for which there does not seem to be any possibility of a formal translation, so the question of what a mathematical proof really is is more complicated than we have suggested here. J. R. Lucas [Luc00] regards a proof as a dialog (the view of a mathematical argument as a dialog goes back at least to Plato; a notable recent proponent was IMRE LAKATOS [Lak76]. Indeed, the very term “argument” suggests a dialog.) Lucas then considers a formal proof to be the extreme special case when the dialog is held with a “recalcitrant fool,” and adds:

“An argument is not really cogent unless it works even on morons and skeptics.”

Even cogency may not suffice in such a situation! In the more common case where the discussion is with an intelligent and reasonable fellow truth-seeker, an argument of a different type or level of detail may suffice. See [?] and [?] for further discussion of the nature of mathematical proofs.

I.5.3.4. There are two types of flaws a purported proof can have. The first is actual errors. An error in an argument is a step which is demonstrably false. One error in one step invalidates the entire argument. An argument with an error can be repaired in either of two ways:

(1) By finding a correct logical path around the error, a valid justification for the conclusions drawn from the erroneous assertion.

(2) By adding extra hypotheses to the statement of the result, in effect restricting to a special case where the erroneous statement is valid.

The first type of repair is preferable when it is possible. Both types of repair must sometimes be used together.

I.5.3.5. A more common flaw in “proofs” offered by mathematicians is a gap. A gap is a step in the argument which the reader is unable to bridge. Gaps are often in the eye of the beholder: what constitutes a gap for one reader may not be a gap for the author or other readers.

“I never come across one of Laplace’s ‘Thus it plainly appears’ without feeling sure that I have hours of hard work before me to fill up the chasm and find out and show how it plainly appears.”

N. Bowditch

The most serious gaps are ones which, once identified, cannot be quickly bridged by either the author or other readers. The gap then must be repaired before the argument can be accepted as a valid proof. A gap can be repaired in either of the two ways described above, or by simply finding an argument which bridges the gap directly (the word “simply” must be taken with a grain of salt; it is often not simple at all to do this!)

Purported formal proofs can also contain errors or gaps, but these are rather easy to identify. In fact, a computer can be programmed to check the correctness of a formal proof. There is no such definite procedure for mathematical proofs.

---

44Quoted in Cajori, *Teaching and History of Mathematics in the U.S.* (Washington, 1896); cf. [Mor93, 985]. According to W. W. R. Ball [Mor93, 986], “Biot, who assisted Laplace in revising [the Mécanique Céleste] for the press, says that Laplace himself was frequently unable to recover the details in the chain of reasoning, and if satisfied that the conclusions were correct, he was content to insert the constantly recurring formula, ‘Il est aisè à voir.’ ”
For a proof to be accepted as valid, there must be general agreement among specialists that it contains no errors or gaps. The process of accepting a proof is thus a somewhat subjective one which is not entirely rigorous; it can happen (but, fortunately, is very rare) that an error or gap is noticed in an argument long after it has been generally accepted as valid. In defense of mathematicians, it must be said that the process almost always works well, and acceptance of a mathematical result is much less dependent on the reputation or force of personality of the author than is the situation in most other academic subjects. And no interesting or useful result can stand the test of time for very long unless it is fully and correctly established. Standards of rigor have evolved over time: many arguments accepted as proofs as late as the early nineteenth century would not pass muster today, although the statements of the results were rarely wrong (but sometimes exact technical hypotheses were not carefully spelled out, or even understood), and the arguments can usually be filled out to be acceptable as proofs today. Mathematicians as a group are very careful and critical in testing the validity of proofs, especially since a proved statement becomes a “building block” in the “edifice of mathematics” ( ), and a flawed block could cause a whole section of the building to collapse. Besides, once something in mathematics is really proved \footnote{I do not just mean here that a proof has been accepted; mathematicians’ opinions about what constitutes a proof can be fallible! For contrary views about the sentiment expressed here, see J. Crowe, Ten Misconceptions about Mathematics, in [AK88], p. 267-271.} it stays proved forever; it can never be invalidated by new theories or points of view. Proofs found by ancient Greeks, when written carefully according to modern standards, are just as good today as they were 2500 years ago!

“I sometimes think that we worry ourselves overmuch about the enduring consequences of our errors. They may work a little confusion for a time. In the end, they will be modified or corrected or their teachings ignored. The future takes care of such things. In the endless process of testing and retesting, there is a constant rejection of the dross, and a constant retention of whatever is pure and sound and fine.”

Benjamin N. Cardozo\footnote{The Nature of the Judicial Process, p. 179.}

“We must not be surprised, however, that many minds apply themselves to other studies [than arithmetic and geometry], or to philosophy. Indeed everyone allows himself more freely the right to make his guess if the matter be dark than if it be clear, and it is much easier to have on any question some vague ideas than to arrive at the truth itself on the simplest of all.”

Descartes\footnote{Rules for the Direction of the Mind: Torrey’s Philosophy of Descartes (New York, 1892), p. 63; cf. [Mor93, 1628].}

“It has long been a complaint against mathematicians that they are hard to convince: but it is a far greater disqualification both for philosophy, and for the affairs of life, to be too easily convinced; to have too low a standard of proof. The only sound intellects are those which, in the first instance, set their standards of proof high. Practice in concrete affairs soon teaches them to make the necessary abatement: but they retain the consciousness, without which there is no sound practical reasoning, that in accepting inferior evidence because there is no better to be had, they do not by that acceptance raise it to completeness.”

J. S. Mill\footnote{An Examination of Sir William Hamilton’s Philosophy (London, 1878), p. 611; cf. [Mor93, 811].}
I.5.4. How to Understand a Proof

I believe I am fairly typical of mathematicians in my approach to understanding a proof I try to read. After reading through any previous explanatory material to be sure I understand the setup, context, notation, and terminology, I first read the statement of the result and try to think of my own proof. I usually fail, especially if it is a research-level result in my field or a result outside my specialty, and I often do not spend a great deal of time trying. If I do succeed, or if I think I have succeeded, I read through the written argument to see if it is the same as mine. Sometimes it is, and I then understand the proof because I have internalized it. Sometimes it is a different argument, and after seeing the general form of the author’s argument I put the written proof down and again try to come up with my own argument along those lines (and I also rethink my first argument to see if it is really correct, especially if it is simpler than the written proof.) Sometimes the written argument is basically the same as mine, but treats a technical point which I have overlooked. Occasionally I find that the author has overlooked a technical point which I noticed, and if so I have found a gap or error in the written proof.

It frequently happens that I can identify an overall strategy for proving the result, but there is a crucial step in the proof that I cannot figure out, or I can reduce the statement of the result to another apparently simpler or more elementary statement which I cannot see how to prove. Sometimes, especially if it is a known problem within my specialty, there may even be a well-known strategy or reduction of the problem that I am aware of. In these cases, I hone in on the author’s treatment of the step I regard as the essential difficulty, or look to see if the author has taken a new or different approach entirely.

In the more common case where I have not been able to get very far in finding my own proof, I look at the written proof and try to identify its overall outline, and then using that I again look for my own proof. There are usually several iterations of this process (each of which can take anywhere from a few minutes to several hours). After a couple of rounds I frequently have most of the argument down, but I am still puzzled by one part (or more) of the argument; I then concentrate the process on the part(s) of the argument I am having difficulty with. At the end of the process I (hopefully) really understand the proof, and the test I use to make sure is whether I can explain the whole argument to someone else with reference to notes only for technical items like formulas or calculations.

I.5.5. Proof by Contradiction

“When you have excluded the impossible, whatever remains, however improbable, must be the truth.”

*Sherlock Holmes*49

One of the commonly used schemes of proof is proof by contradiction, sometimes called *reductio ad absurdum*. If one is trying to prove a statement $P$, it is assumed that $P$ is false and from this a contradiction or absurdity is obtained; thus $P$ cannot be false and is therefore true. This technique is a variant of proving an implication $P \Rightarrow Q$ by assuming the conclusion $Q$ is false and proving that at least one hypothesis cannot hold, i.e. that $P$ is also false (that is, proving the implication $P \Rightarrow Q$ by proving the contrapositive $(not Q) \Rightarrow (not P)$); however, a proof by contradiction does not necessarily involve a proof of a contrapositive, at least explicitly.

A good example of a proof by contradiction is the famous ancient proof that $\sqrt{2}$ is irrational, which dates from about 500 BCE. The proof is often attributed to Pythagoras, but is likely due to one of his

followers (Hippasus is frequently credited; some attribute this proof to Aristotle). We will write the proof in modern algebraic notation, which did not exist in ancient Greece.

**Theorem.** $\sqrt{2}$ is irrational: there is no rational number $r$ such that $r^2 = 2$.

**Proof:** Suppose there is such an $r$. Then $r = \frac{a}{b}$ with $a, b \in \mathbb{N}$ (the natural numbers, or positive integers). By repeatedly dividing out common factors of 2, we may assume $a$ and $b$ are not both even. We have

$$\frac{a^2}{b^2} = 2$$

so $a^2$ and hence $a$, is even. Then $a = 2c$ for some $c \in \mathbb{N}$, so

$$a^2 = 4c^2 = 2b^2$$

and $b^2$, hence also $b$, is even, contradicting our assumption that $a$ and $b$ are not both even. The actual contradiction is to the assumption that $r$ exists, i.e. that $a$ and $b$ exist, since if they do the reduction to the case where they are not both even follows from basic properties of $\mathbb{N}$ (cf. III.1.7.4.).

See III.1.12.6. and III.1.13.6. for other elementary proofs by contradiction of the irrationality of $\sqrt{2}$.

It should be noted that intuitionist mathematicians of the Brouwer school do not accept the Principle of the Excluded Middle, i.e. that $(P \text{ or } \neg P)$ is a tautology for every statement $P$, and hence generally reject the validity of proof by contradiction (they do not entirely reject *reductio ad absurdum*, however: for example, the above proof of the irrationality of $\sqrt{2}$ is acceptable to most intuitionists). But almost all mathematicians embrace proof by contradiction as a legitimate method of proof, and agree with the sentiment in the famous statement of G. H. Hardy:

“*Reductio ad absurdum*, which Euclid loved so much, is one of a mathematician’s finest weapons.

It is a far finer gambit than any chess play: a chess player may offer the sacrifice of a pawn or even a piece, but a mathematician offers the game.”

A careful distinction must be made between proof by contradiction and an *impossibility proof*, a proof that a certain mathematical object does not exist. (Theorem I.5.5. is an impossibility theorem: “There does not exist a rational number $r$ with $r^2 = 2$.”) Proof by contradiction is a technique of proof which can be used to prove any sort of mathematical statement. Impossibility results (of which there are many) are commonly proved by contradiction, but not necessarily so: for example, the standard abstract algebra proof that a general fifth-degree polynomial is not solvable by radicals is usually expressed as a direct argument, not a proof by contradiction. Fermat’s Last Theorem is another example (although the proof there is very deep). See I.5.6.10. for more on impossibility results.

Most, if not all, proofs by contradiction can in principle be converted into direct proofs, but it may be very clumsy and unnatural to do so in practice. It is also true that almost any direct mathematical statement can be converted into an impossibility statement, but also often in a clumsy and unnatural way.

My experience is that beginning students tend to overuse proof by contradiction. It is usually good advice to look for a direct proof first and only use proof by contradiction as a last resort, unless one happens to notice an obvious and simple contradiction argument. As a rule, the only types of mathematical statements where it is worthwhile to begin by looking for a proof by contradiction are impossibility statements.

---

I.5.6. Proofs and Counterexamples: Mathematics as Creative Activity

“All truths are easy to understand once they are discovered; the point is to discover them.”

Galileo

I.5.6.1. As important as it is to be able to understand proofs written by other mathematicians, the true essence of mathematics is the ability to solve problems. Mathematical problems range from very specific to completely general and open-ended, and from straightforward to intractable (the two scales, by the way, are largely uncorrelated; very specific problems can be, or seem, intractable). Solutions to mathematical problems can range from simple numerical calculations or straightforward applications of standard mathematical results up to remarkably intricate, clever, and ingenious creative new arguments which can spawn entire new areas of mathematics. The best solutions to even simple problems are ones which are stimulating and thought-provoking beyond the problem at hand.

“In mathematics resolving of important specific questions and development of an abstract general theory are processes as inseparable as inhaling and exhaling.”

V. Zorich\textsuperscript{51}

“A naive non-mathematician ... understandably concludes [from formal presentations] that in mathematics, axioms come first. ... But anyone who has done mathematics knows what comes first – a problem. Mathematics is a vast network of interconnected problems and solutions. ... Sometimes a solution is a set of axioms! ... When a piece of mathematics gets big and complicated, we may want to systematize and organize it, for esthetics and for convenience. The way we do that is to axiomatize it. Thus a new type of problem (or ‘meta-problem’) arises:

‘Given some specific mathematical subject, to find an attractive set of axioms from which the facts of the subject can conveniently be derived.’

Any proposed axiom set is a proposed solution to this problem. The solution will not be unique. ... In developing and understanding a subject, axioms come late. Then in the formal presentations, they come early.”

Reuben Hersh\textsuperscript{52}

“Mathematical statements, mathematical theorems, are answers to questions. ... The mystery of how mathematics grows is in part caused by looking at mathematics as answers without questions. That mistake is made only by people who have had no contact with mathematical life. It’s the questions that drive mathematics. Solving problems and making up new ones is the essence of mathematical life.”

Reuben Hersh\textsuperscript{53}

\textsuperscript{51}[Zor09, p. V]
\textsuperscript{52}What is Mathematics, Really?, p. 6.
\textsuperscript{53}What is Mathematics, Really?, p. 18.
I.5.6.2. One of the most common kinds of mathematical problems is: Is statement $P$ true? True, of course, means always true: the conclusion holds in every situation in which the hypotheses are satisfied. A solution to a problem of this sort consists of a proof or a disproof of the statement. Note that a statement cannot be proved by verifying that it is true in one or more examples unless every example in which the hypotheses hold is checked (which is indeed sometimes feasible in specific problems); at best, verifying the statement in some examples provides evidence that the statement may be true in general. However, one example where the hypotheses are satisfied but the conclusion does not hold disproves the statement (proves that the statement is false). Such an example is called a counterexample for the statement. It is possible in theory, and occasionally in practice, to disprove a statement without finding a specific counterexample; but a “smoking gun” counterexample is usually the simplest, and often the most satisfying, way to disprove a statement.

“The mathematician as the naturalist, in testing some consequence of a conjectural general law by a new observation, addresses a question to Nature: ‘I suspect that this law is true. Is it true?’ If the consequence is clearly refuted, the law cannot be true. If the consequence is clearly verified, there is some indication that the law may be true. Nature may answer Yes or No, but it whispers one answer and thunders the other, its Yes is provisional, its No is definitive.”

G. Polya

Finding a Solution

“It will seem not a little paradoxical to ascribe a great importance to observations even in that part of the mathematical sciences which is usually called Pure Mathematics, since the current opinion is that observations are restricted to physical objects that make impression on the senses. As we must refer the numbers to the pure intellect alone, we can hardly understand how observations and quasi-experiments can be of use in investigating the nature of the numbers. Yet, in fact, as I shall show here with very good reasons, the properties of the numbers known today have been mostly discovered by observation, and discovered long before their truth has been confirmed by rigid demonstrations. There are even many properties of the numbers with which we are well acquainted, but which we are not yet able to prove; only observations have led us to their knowledge. Hence we see that in the theory of numbers, which is still very imperfect, we can place our highest hopes in observations; they will lead us continually to new properties which we shall endeavor to prove afterwards. The kind of knowledge which is supported only by observations and is not yet proved must be carefully distinguished from the truth; it is gained by induction, as we usually say. Yet we have seen cases in which mere induction led to error. Therefore, we should take great care not to accept as true such properties of the numbers which we have discovered by observation and which are supported by induction alone. Indeed, we should use such a discovery as an opportunity to investigate more exactly the properties discovered and to prove or disprove them; in both cases we may learn something useful.”

L. Euler

54 [Pol90a, p. 10];
55 Opera Omnia, ser. 1, vol. 2, p. 459; cf. [Pol90a, p. 3].
I.5.6.3. In order to solve the problem of whether statement $P$ is true or not, one must usually first form an opinion about the answer, which when firmly held and/or communicated to others is called a conjecture. The next step is to try to verify the conjecture by either looking for a proof (if the conjecture is that $P$ is true) or a counterexample (if the conjecture is that $P$ is false). The search often involves doing both things to some extent: obstacles encountered in trying to prove $P$ may suggest a possible counterexample, and on the other hand inability to construct a counterexample can sometimes suggest a method of proof of $P$.

I.5.6.4. Although mathematics is (for the most part) not an experimental science, there is a good analogy with experimental sciences: in both mathematics and experimental sciences, people form a conjecture (called a “theory” in sciences) based at least in part on evidence from observations or experiments. Then either the creator of the conjecture (theory) or someone else “tests” the conjecture (theory) with “experiments.” A well-designed set of experiments, or candidates for counterexamples in mathematics, provide the most critical tests for consequences or predictions of the conjecture (theory): if the conjecture (theory) “passes the test”, it can be deemed worthy of being taken seriously. Of course, the biggest difference between mathematics and experimental sciences is that theories can never be proved in science (at least not in the same sense as proofs in mathematics), and acceptance of a theory in science is based only on believability, observational or experimental evidence, accuracy of prediction, and/or lack of experimental refutation; such acceptance is always subject to revision in light of new evidence.

I.5.6.5. One need not, and often should not, stop when the problem at hand is solved. It is a much more significant accomplishment to come up with a general scheme for solving a class of related problems, so that new problems as they arise can also be solved more or less mechanically without a similar investment of time and effort. This is the origin of new subdisciplines in mathematics. (One should not be under the illusion, however, that all of mathematics can eventually be reduced to a set of mechanical processes or algorithms; there will always be room for creative genius in mathematics.)

“...In general the position as regards all such new calculi is this – That one cannot accomplish by them anything that could not be accomplished without them. However, the advantage is, that, provided such a calculus corresponds to the inmost nature of frequent needs, anyone who masters it thoroughly is able – without the unconscious inspiration of genius which no one can command – to solve the respective problems, yea, to solve them mechanically in complicated cases in which, without such aid, even genius becomes powerless. Such is the case with the invention of general algebra, with the differential calculus, and in a more limited region Lagrange’s calculus of variations, with my calculus of congruences, and with Möbius’s calculus. Such conceptions unite, as it were, into an organic whole countless problems which would otherwise remain isolated and require for their separate solution more or less application of inventive genius.”

C. F. Gauss\(^{57}\)

---


\(^{57}\)Werke, Bd. 8, p. 298; cf. R. Moritz, Memorabilia Mathematica, 1215.
“How thoroughly it is ingrained in mathematical science that every real advance goes hand in hand with the invention of sharper tools and simpler methods which, at the same time, assist in understanding earlier theories and in casting aside some more complicated developments.”

D. Hilbert

“... [M]ath isn’t the art of answering mathematical questions, it is the art of asking the right questions, the questions that give you insight, the ones that lead you in interesting directions, the ones that connect with lots of other interesting questions – the ones with beautiful answers!”

Gregory Chaitin

“The outcome of any serious research can only be to make two questions grow where only one grew before.”

Thorstein Veblen

Is There a Solution?

I.5.6.6. One question which must be faced, at least in passing, in any mathematical problem is: does the problem have a solution at all? Indeed, in some mathematical problems it is not entirely clear what one would exactly mean by a “solution,” and part of formulating such a problem well is deciding what this should mean.

I.5.6.7. Even if the problem is well-formed and it is obvious what a “solution” would be, it is often not obvious whether there is one. In fact, in many problems the most difficult part of the problem is establishing that there is a solution: if a solution is known to exist, it can sometimes be found relatively easily from the properties it must have. And sometimes knowing the mere existence of a solution can itself have important consequences. Thus considerable mechanisms have been developed in mathematics to establish the existence of solutions to certain kinds of problems:

The notions of completeness and compactness can be thought of as properties insuring the existence of solutions to problems involving limits, maxima, and minima (although not all such problems!)

The Axiom of Choice is a set-theoretic assumption insuring the existence of certain kinds of sets and functions.

I.5.6.8. On the other hand, it can sometimes be shown that a certain problem has no solution. This can also have important consequences quite beyond simply letting mathematicians avoid wasting time and effort looking for one. Proving that a problem has no solution is a type of impossibility result ( ). Actually, though, establishing nonexistence of a solution is often not the end of the story: such results sometimes lead to important expanded theories in which there is a solution. Here are some examples where the absence of a solution led to important advances:

58[Cha05, p. 24].
59The Place of Science in Modern Civilization and Other Essays.
(i) The equation $x^2 = 2$ has no solution in the set of rational numbers. This is one of the examples which eventually led to the study of irrational numbers and formulation of the theory of real numbers, which underlies all of analysis.

(ii) The equation $x^2 = -1$ has no solution in the set of real numbers. The complex numbers were developed specifically to provide a solution to this equation; they then turned out to be important and useful beyond the wildest dreams of the mathematicians who first considered them.

(iii) The “Dirac delta function” is a function $\delta$ on the real numbers $\mathbb{R}$ such that $\delta(x) = 0$ for all $x \neq 0$ and such that $\int_{-\infty}^{+\infty} \delta(x) \, dx = 1$. Such a “function” would have many uses in both pure and applied mathematics. But it is easily proved that there is no such function. This example led to the theory of generalized functions and distributions, where $\delta$ exists as a generalized function and, with care, can be used in many of the applications.

“Remember that when the usual road does not lead to success, one should not be content with this determination of impossibility, but should bestir oneself to find a new and more promising route. Mathematical thought, as such, has no end. If someone says to you that mathematical reasoning cannot be carried beyond a certain point, you may be sure that the really interesting problem begins precisely there.”

F. Klein

I.5.6.9. One must be careful about assuming that simply stated problems either have relatively simple solutions or none at all. For example, does the equation

$$x^2 - 1141y^2 = 1$$

have any solutions in positive integers? An exhaustive search through reasonably, semi-reasonably, and even not-so-reasonably sized pairs of positive integers would not turn up a solution, so one could be led to believe there is no solution. However, it turns out that there is a solution, in fact infinitely many, the smallest of which is

$$x = 1,036,782,394,157,223,963,237,125,215 \approx 10^{27}$$

$$y = 30,693,385,322,765,657,197,397,208 \approx 3 \times 10^{25}.$$  

This is an example of Pell’s equation; see e.g. [Sta78] for a discussion.

Another example is Archimedes’ cattle problem (see [Var98] for a formulation and nice discussion). This is a problem which can be expressed in words without use of mathematical symbols, yet the smallest solution is approximately $7.76 \times 10^{206544}$.

“At any moment there is only a fine layer between the ‘trivial’ and the impossible. Mathematical discoveries are made in this layer.”

A. Kolmogorov

---

60[Kle04, p. 143]. Note that this was written before Gödel; nonetheless Klein’s point is well taken.

61[?, p. 83].
I.5.6.10. It is worth a digression to discuss an endemic problem in mathematics which continues to plague mathematicians: misunderstanding the difference between unsolved and unsolvable.

Many mathematical problems are unsolved; some mathematical problems are unsolvable. The classes are quite distinct. (There is a nonempty overlap: some unsolved problems will undoubtedly turn out to be unsolvable, although “most” will not.) What many nonmathematicians do not understand is that many mathematical problems, including some famous ones, have been proved to be unsolvable; thus they cannot be described as “unsolved.” In fact, impossibility theorems are precisely solutions to “unsolvable” problems. Probably the most famous such problem is trisecting an angle, or more precisely giving a general procedure for trisecting a given angle with straightedge and compass, one of the “big three” problems from ancient Greece. The other two ancient Greek problems, almost as famous, are “squaring the circle” (constructing with straightedge and compass a square with the same area as a given circle, cf. [Bla97]) and “doubling the cube” (constructing with straightedge and compass a cube whose volume is twice the volume of a given cube). These problems were unsolved until the nineteenth century, when some abstract algebra now regarded as elementary (covered in undergraduate abstract algebra courses) was developed which proved that the problems are unsolvable, i.e. there can be no such procedure (the transcendence of $\pi$ is also needed for the circle squaring problem, cf. III.9.3.2.).

Mathematicians and mathematics departments regularly receive purported “solutions” to these problems, especially angle trisection. These arguments are often complicated and clever, but of course must necessarily contain a mistake. Mathematicians are understandably extremely reluctant to make the effort to find the inevitable mistakes, and thus tend to ignore or summarily dismiss such communications, leading to bad public relations and charges that mathematicians are closed-minded. Some mathematics departments have actually assigned a faculty member to respond to them for public relations purposes. The people who submit such attempted solutions are generally not crackpots (although there are some of those – see [Dud92] and Figure I.8), but intelligent people who know enough mathematics to be able to formulate the problems and discuss them but not enough to really understand them. In mathematics, a little knowledge can be a dangerous thing.

The most widespread source of confusion is that the problems are often slightly misstated so they sound unsolved, such as: “No one has ever found a procedure to trisect an angle with straightedge and compass.” This is a true statement. But it does not mean the problem is unsolved! It is also a true statement that “No one has ever found a rational number $x$ such that $x^2 = 2.$” And on a simpler level: “No one has ever found an integer $x$ such that $2x = 5.$” No intelligent person would waste time trying to find such an integer, as it is elementary to see there is none. The problem is solved, proved to be unsolvable. Understanding why the trisection problem has no solution is not elementary enough that a person without the requisite knowledge of field theory could be expected to come up with it (even Euler could not), but to a modern mathematician it is quite as impossible to trisect angles as it is to solve the $2x = 5$ problem.

Here are some examples, in roughly increased order of sophistication. These are unsolvable problems, not unsolved problems (the last two were solved in the twentieth century):

There is no integer $x$ with $2x = 5.$

There is no rational number $x$ with $x^2 = 2.$

There is no procedure to trisect angles with straightedge and compass.

There is no procedure to double a cube with straightedge and compass.
There is no procedure to square a circle with straightedge and compass.

There is no algebraic formula for the roots of a general fifth-degree polynomial.

There is no elementary formula for \( \int e^{-x^2} \, dx \) (X.9.3.18.). [But this does not mean there is no antiderivative!]

The Axiom of Choice and the Continuum Hypothesis are neither provable nor disprovable in ZF set theory.

There are no natural numbers \( x, y, z, n \) with \( n > 2 \) for which \( x^n + y^n = z^n \).

Figure I.8: A famous crackpot publication. The author claims that \( \pi = \left( \frac{\sqrt{2}}{2} \right)^4 = \frac{3}{2} \approx 3.16 \). (Is it a coincidence that HEISEL and FABER were both named “Carl Theodore”?)

The assertion that \( \pi = \left( \frac{\sqrt{2}}{2} \right)^4 = 4 \left( \frac{\sqrt{2}}{2} \right)^2 \) was obtained in ancient Egypt (Rhind Papyrus, circa 1650 BCE); cf. e.g. [Bec71, p. 22–25] or [Bla97]. Things could be worse: there are passages in the Bible (I Kings 7:23 and II Chronicles 4:2) which can be interpreted to say that \( \pi = 3 \), so mathematicians could be up against religious dogma as much as evolutionary biologists are. (I have an opinion why this has not happened, which I think I will not commit to print.) Note that mathematics is not immune from religious dogma: the theory that a line is composed of a collection of indivisible points came under withering attack by the Jesuits in the sixteenth century as contradicting the Eucharistic doctrine of transubstantiation (cf. I.3.1.2., [She18]).
Determination and Hard Work

“It would be a mistake to think that solving problems is a purely ‘intellectual affair’; determination and emotions play an important role. Lukewarm determination and sleepy consent to do a little something may be enough for a routine problem in the classroom. But, to solve a serious scientific problem, will power is needed that can outlast years of toil and bitter disappointment.”

G. Polya

I.5.6.11. One psychologically difficult characteristic of the process of solving mathematical problems is that one can work hard for a long time without apparent progress, and even disappointing regress when mistakes are discovered in previous work. But eventually (not always, but often enough), everything suddenly comes together in an “aha!” moment. This moment can even occur while not actively thinking about the problem: after a lot of thought and concentration on a problem the mind continues to mull it over subconsciously even during other activities. When the moment occurs, we can be fooled into thinking that all the previous effort without progress was wasted and could have been avoided if we had just been smarter about the problem, but the fact is that we would have been extremely unlikely to come up with the right idea without all the preceding thought and effort. Edison’s famous statement that “genius is 1% inspiration and 99% perspiration” may be an exaggeration, but contains a good deal of truth. And there is nothing more uplifting for a mathematician than that moment of understanding, which makes (or should make) all the previous work worthwhile.

“Every mathematician worthy of the name has experienced … the state of lucid exaltation in which one thought succeeds another as if miraculously … this feeling may last for hours at a time, even for days. Once you have experienced it, you are eager to repeat it but unable to do it at will, unless perhaps by dogged work …”

André Weil

This does not just happen in mathematics: for example, athletes often describe being “in the zone” and experiencing something similar. But long hours of previous training are necessary preparation.

I.5.6.12. One song I can really relate to is John Denver’s “Looking for Space.” Although he was not a mathematician, and probably knew little about mathematics or the process of doing it, he beautifully captured the emotional roller-coaster mathematicians can deal with in their work:

“On the road of experience, I’m trying to find my own way. Sometimes I wish that I could fly away When I think that I’m moving, suddenly things stand still I’m afraid ’cause I think they always will

And I’m looking for space And to find out who I am And I’m looking to know and understand

---

62 How to Solve It, p. 93.
63 The Apprenticeship of a Mathematician.
It’s a sweet, sweet dream
Sometimes I’m almost there
Sometimes I fly like an eagle
And sometimes I’m deep in despair

All alone in the universe, sometimes that’s how it seems
I get lost in the sadness and the screams
Then I look in the center, suddenly everything’s clear
I find myself in the sunshine and my dreams

And I’m looking for space
And to find out who I am
And I’m looking to know and understand
It’s a sweet, sweet dream
Sometimes I’m almost there
Sometimes I fly like an eagle
And sometimes I’m deep in despair

On the road of experience, join in the living day
If there’s an answer, it’s just that it’s just that way
When you’re looking for space
And to find out who you are
When you’re looking to try and reach the stars
It’s a sweet, sweet sweet dream
Sometimes I’m almost there
Sometimes I fly like an eagle
But sometimes I’m deep in despair
Sometimes I fly like an eagle, like an eagle
I go flying flying.”

The trick a mathematician must learn is to avoid despair during the dry periods, and persevere to savor the moments of insight and enlightenment.

“Perhaps I can best describe my experience of doing mathematics in terms of a journey through a dark unexplored mansion. You enter the first room of the mansion and it’s completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch, you turn it on, and suddenly it’s all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark. So each of these breakthroughs, while sometimes they’re momentary, sometimes over a period of a day or two, they are the culmination of, and couldn’t exist without, the many months of stumbling around in the dark that precede them.”

Andrew Wiles

64Quoted in interview for the PBS TV Nova program, ‘The Proof’.
“Doing research in mathematics is frustrating and if being frustrated is something you cannot get used to, then mathematics may not be an ideal occupation for you. Most of the time one is stuck, and if this is not the case for you, then either you are exceptionally talented or you are tackling problems that you knew how to solve before you started. There is room for some work of the latter kind, and it can be of a high quality, but most of the big breakthroughs are earned the hard way, with many false steps and long periods of little progress, or even negative progress.”

*Peter Sarnak*

“Thought is only a flash between two long nights, but this flash is everything.”

*Henri Poincaré*

“Mathematical discoveries, small or great, are never born of spontaneous generation. They always presuppose a soil seeded with preliminary knowledge and well prepared by labour, both conscious and subconscious.”

*Henri Poincaré*

“Well, some mathematics problems look simple, and you try them for a year or so, and then you try them for a hundred years, and it turns out that they’re extremely hard to solve.”

*Andrew Wiles*

“Unfortunately, mathematics is often taught as a mere matter of learning instructions and following them. You may have the habit of not trying anything in mathematics until you already ‘know what to do,’ until you feel that you ‘understand’ the problem. If this is your habit, try to overcome it as quickly as possible, for it will be the main obstacle to your progress.”

*Kenneth O. May*

“How does one discover how to prove a theorem? ... You must proceed as any person does who is confronted with a problem for which there are no set rules. ... half the battle will be won if you accept the fact that you are not supposed to know in advance how to prove theorems. The problem is not to know, but to discover.”

*Kenneth O. May*

“In research, if you know what you are doing, then you shouldn’t be doing it.”

*R. Hamming*
I.5.7. Scholarship vs. Practice

For every mathematician, from student to leading research mathematician, there is a constant tradeoff between studying the work of others and working on one’s own problems. Both are necessary for growth and development as a mathematician, although the relative emphasis can vary widely: some mathematicians are primarily scholars of mathematics, with vast knowledge of the work done almost entirely by others, while some of the most brilliant mathematicians have surprisingly sparse knowledge of others’ work, preferring to solve most problems arising in their work themselves as they come (this is definitely the exception rather than the rule among top mathematicians, however). For most of us, study of the work of others provides us not only with inspiration, but also a vast supply of machinery we can use in our own work.

According to P. Sarnak [Vil16, p. 115-116], Paul Cohen did not believe mathematics progresses incrementally, but rather by sudden great leaps, and therefore he strongly discouraged using, or even reading, related previous work on one’s problems. This is roughly the “Moore method” carried to its logical extreme. Although much good can be said about the Moore method, and there are even some positives about Cohen’s attitude (assuming it is accurately portrayed), and although Cohen was unquestionably an outstanding mathematician, I think this attitude is almost completely wrong: some of the biggest advances in mathematics do indeed come out of the blue, but most do not, and it is generally a very bad idea for a research mathematician to be ignorant (willfully or otherwise) of other work in the area. For one famous example, how can it be explained that calculus was simultaneously developed by two people in the latter 17th century? It is not adequate to simply say that two of the greatest mathematical minds of all time coincidentally did this at the same time; this is arguably a true statement, but it is also true that the mathematical terrain and atmosphere at the time was ripe for the development. Calculus did not suddenly arise out of nowhere, and most of the ideas and techniques of calculus already existed in some form; the main contribution of Newton and Leibniz (which I do not at all minimize) can be accurately described as largely “putting the pieces together.” In fact, if Cohen was right, worthwhile collaborative research in mathematics would be impossible.

Attempting to do mathematical research as a “lone wolf” has at least two related drawbacks: (1) You often end up “reinventing the wheel” due to inadequate familiarity with other work, although sometimes the new solutions are better than the previous ones and make an important contribution to the subject; (2) Especially nowadays when so much mathematical research is being done in such varied areas, unless you are an unusually great mathematician there is a good chance that someone at least as good as you are has already thought extensively about the problem or a closely related one, and can impart at least some words of wisdom, if not useful techniques or partial results.

A (rather imperfect) analogy with medicine is perhaps useful. The primary goal of medicine is to treat patients. To do this properly, a practicing doctor must not only have a rigorous medical school training, but must regularly keep up with medical journals and other research reports. On the other hand, there is a place, even a necessity, for some doctors to spend most of their time doing laboratory research, or instruction at medical schools and hospitals; however, almost all research or teaching physicians “keep their hand in practice” by treating at least a few patients.

I.5.8. Is Mathematics Created or Discovered?

One of the most basic controversies in the philosophy of mathematics is the question of whether working mathematicians actually create new mathematics which did not exist before, or whether they simply discover mathematical truths which already existed abstractly, somewhere (or whether it even makes sense to ask these questions.) Belief in the a priori existence of mathematical “objects” is referred to as realism (the term
Platonism is also often used, but Platonism is technically only one form of realism).

One popular version of the question is: if there is an intelligent race of aliens on another planet, would they develop mathematics, and if so would it bear any resemblance to the mathematics developed so far by humans? If mathematics is indeed universal, then it should be essentially the same everywhere in the universe.

“Beethoven and Shakespeare did unique work, and we would not have Fidelio or King Lear if they had not lived. However, symmetric functions would have been discovered by others if Newton had not lived.”

Richard Askey

My own view leans more towards the realist philosophy. Mathematical results are not arbitrary; the (provable) truth of a mathematical statement does not depend on the opinions or desires of the mathematicians who study it. Two mathematicians studying the same problem, making the same assumptions, cannot arrive at contradictory solutions (provided that the assumptions they make are not self-contradictory).

The importance of a new (or newly discovered) piece of mathematics is a much more subjective matter, and there can very well be differences of opinion among mathematicians; however, as time goes on there almost always seems to develop a remarkable consensus of opinion about what is significant and what is the “right” approach to an area of mathematics. This seems to provide evidence that there is some pre-existing mathematics either outside or deep within the human intellect. However, it seems likely that the exact form of the mathematics considered in the human mind or communicated between humans is dependent on specific human characteristics as well as on human history and culture (this is obviously true at least on the superficial level of notation and terminology, but probably goes much deeper). The “continent of mathematics” analogy (I) thus seems appropriate.

Realism as a philosophy of mathematics has definitely been unfashionable among philosophers. However, among mathematicians it is probably the predominant view. (Philosophers tend to be dismissive of the opinions of mathematicians on this issue as unreliable, unsophisticated, and/or uninformed, not entirely without justification.) Mathematicians generally have a strong feeling that the abstract objects they work with are “real,” although this could of course be an illusion. This feeling is reflected in the language mathematicians use to describe advances in mathematics, almost invariably referring to “discovering” or “finding” new results or proofs. The term “invent” is rarely used, and when it is (e.g. “Newton and/or Leibniz invented calculus”) it usually refers to the development of a new subdiscipline or point of view in mathematics, which can be viewed as a way of facilitating analysis of the structure of part of the “continent of mathematics” rather than creation of new mathematics per se.

I.5.9. Analogies for the Structure of Mathematics

There are at least three analogies which have been used to describe the subject of mathematics. All are useful, if of course vast oversimplifications; the differences between them reflect different philosophies of the nature of mathematics (we will not attempt to discuss such philosophies here; see I.6.1.).

---

72 R. Askey, *How can mathematicians and mathematical historians help each other?* [AK88, p. 202]. This is just one version of a line often used with many variations.
The Edifice of Mathematics

This picture or analogy was popularized by Descartes (who introduced the term “edifice”), but it may well not have originated there. It is probably the most widely-used picture of mathematics, although it primarily reflects a formalist or logicist philosophy:

“The discovery that all mathematics follows inevitably from a small collection of fundamental laws is one which immeasurably enhances the intellectual beauty of the whole; to those who have been oppressed by the fragmentary and incomplete nature of most existing chains of deduction this discovery comes with all the overwhelming force of a revelation; like a palace emerging from the autumn mist as the traveller ascends an Italian hill-side, the stately storeys of the mathematical edifice appear in their due order and proportion, with a new perfection in every part.”

— Bertrand Russell

If this view of mathematics is accepted, even partially, one must eventually consider what the building stands on and whether it has a solid foundation. Until the nineteenth century, construction and examination of the edifice of mathematics took place almost entirely “above the surface,” building upward from such surface features as numbers, points, and lines. Mathematicians for the most part did not concern themselves with, or even recognize or acknowledge, the problem (from the “edifice” point of view) that the building might stand on shifting sand instead of solid rock.

In the latter part of the nineteenth century, continuing into the twentieth and to this day, mathematicians started exploring the part of the edifice lying below the surface. The development of set theory seemed to provide at least a basement, if not a foundation, for the structure. Logic itself seemed to lie either in a subbasement or an entire substructure of the building. But as investigations went further, it became less clear, rather than more, that there is a solid and well-formed foundation underlying it all.

The Tree of Mathematics

The analogy of mathematics as a tree instead of a building has also been popular for a long time in various forms. The tree is a living, growing organism instead of a structure built with blocks of stone. As it grows, more and more branches of the tree appear and develop. Sometimes parts of these branches become so intertwined that it is nearly impossible to trace where the different strands come from.

The tree of mathematics not only grows upwards above the surface of the ground, but a root system also develops and grows, giving the tree increased (but not total) stability. If the stability of the tree is threatened in a certain direction, there is an impetus to extend the root system in that direction to restabilize the tree.

One problem with this view of mathematics as a tree, of course, is that it is unclear what role human mathematicians have in the growth and development of the tree; the idea that the tree grows on its own, with only fertile soil, water, and sunlight needed, is rather contrary to the picture of mathematics as a human activity. On the other hand, mathematics could be regarded as having a life of its own:

“Upon superficial observation mathematics appears to be a fruit of many thousands of scarcely related individuals scattered through the continents, centuries and millennia. But the internal logic of its development looks much more like the work of a single intellect that is developing his thought continuously and systematically, using as a tool only the variety of human personalities.

As in an orchestra performing a symphony by some composer, a theme is passing from one
instrument to another, and when a performer has to finish his part, another one is continuing it as if playing from music.”

I. R. Shafarevich

The view of mathematics as a tree makes some of the foundational questions less compelling: what is usually called “foundations of mathematics” is simply the root structure of the tree. Indeed, it can be argued from this point of view that the whole idea of a “foundation” of mathematics is not necessary or appropriate, so long as we have confidence that the tree is growing in some kind of solid ground.

The Continent of Mathematics

An analogy I particularly like, which I first saw in IAN STEWART’S book, Nature’s Numbers (cf. also Bel87, p. 8), compares mathematics to the landscape of a mostly unexplored continent. To paraphrase and expand on STEWART, when man first arrived on this remote continent, he could see very little through the impenetrable jungle beyond the beach. The continent has many resources, but is covered by thick vegetation and crisscrossed by mountains and gorges which are difficult and dangerous to cross. In time, some areas of the continent are explored, and settlements and roads built, and gradually more and more of the continent becomes known and accessible. The towns and roads impose a certain social order on the explored region, and some places become established as important centers of the society (perhaps with universities having mathematics departments.)

Here and there are high peaks which are mostly obscured by fog. On clear days, however, expert climbers can reach the summit and get panoramic views of large areas of the continent. They may be able to see undiscovered nearby verdant valleys which could be inhabited, or passes or natural bridges across mountains or gorges previously thought impassable. They may also be able to see that some areas thought to be far apart and unconnected are only actually separated by a narrow ridge and can easily be connected by a short road. (We assume this society never developed air or space travel.)

“Creating a new theory is not like destroying an old barn and erecting a skyscraper in its place. It is rather like climbing a mountain, gaining new and wider views, discovering unexpected connections between our starting point and its rich environment. But the point from which we started out still exists and can be seen, although it appears smaller and forms a tiny part of our broad view gained by the mastery of the obstacles on our adventurous way up.”

Albert Einstein

Mathematicians are the architects of the creation of a society on the continent of mathematics (which is actually a “virtual continent”, not physically altered by the society and capable of supporting simultaneous independent settlement by other cultures). Some mathematicians explore new regions of the continent; others build roads and bridges for the public to use; others design and build the cities; others invent ways of using the resources. Some operate farms or mines in recently explored regions where the climate or topography allows new crops to be grown or new minerals found. [Farms are probably a better analogy than mines, since mathematical resources, once discovered, cannot be depleted; but mines are apt in the sense that digging below the surface in mathematics can sometimes yield new and unexpected treasures.] Some logicians write the laws governing the society, and others study the scientific principles underlying the formation of the

---

74On some tendencies of the development of mathematics; cf. [Bog07], v. 2, p. 439.
continent. The top mathematicians are the expert climbers who can go up the peaks to get the big picture and see undiscovered regions and connections, although these can also sometimes be discovered at ground level. Most of the mathematicians may not be able to reach the summits without cable cars and guided tours, but there is much worthwhile to be done at all levels of the society.

In this view, the topography of the continent of mathematics is pre-existing and is discovered, not created, by the mathematicians. However, the society created by the mathematicians is not preordained: choices are made about where to build the cities and roads, what kind of political and social order to establish, and which resources to use and how; these choices depend on the intellectual ability and cultural background of those building the society, as well as seemingly random events such as what parts of the continent happen to be discovered or explored at a given time, or whether a sufficiently skilled climber comes along during the right weather conditions to get to the top of a critically located peak to catch the view. It could well happen that a different group of settlers (such as aliens from another planet) would establish a quite different society on the same continent.

“The true mathematician is always a good deal of an artist, an architect, yes, of a poet. Beyond the real world, though perceptibly connected with it, mathematicians have intellectually created an ideal world, which they attempt to develop into the most perfect of all worlds, and which is being explored in every direction. None has the faintest conception of this world, except he who knows it.”

A. Pringsheim

But as E. T. Bell [Bel87, p. 8] observes, exploration of the continent of mathematics is an imperfect process:

“If we marvel at the patience and the courage of the pioneers, we must also marvel at their persistent blindness in missing the easier ways through the wilderness and over the mountains. What human perversity made them turn east and perish in the desert, when by going west they could have marched straight through to ease and plenty? This is our question today. If there is any continuity in the progress of mathematics, the men [sic] of a hundred years hence will be asking the same question about us.”

---

75 Jahresbericht der Deutschen Mathematiker Vereinigung, Bd. 32, p. 381; cf. R. Moritz, Memorabilia Mathematica, 1108.
I.5.10. Pure vs. Applied Mathematics

Everyone recognizes that mathematics has enormous and vital applications to the “real world.” **Eugene Wigner** describes this as “the unreasonable effectiveness of mathematics.” (In my view, this thinking is a little backwards: it is not self-evident to me that there should be physical laws at all, but if there are, and the apparent orderly behavior of the universe suggests there are, they must necessarily, like everything else, operate in accordance with mathematical principles.) But mathematicians and nonmathematicians tend to view the applicability of mathematics in rather different ways. Nonmathematicians often view mathematics as just a tool to be used, a discipline whose importance is primarily in support of scientific and technical subjects (and increasingly other academic areas too); while they recognize its importance, they also understand its limitations in this respect. The following statement by an engineer is fairly typical of this point of view:

> “When the difficulty of a problem lies only in finding out what follows from certain fixed premises, mathematical methods furnish invaluable wings for flying over intermediate obstructions; but whenever the chief difficulty of a problem lies in the multiplicity and dubiousness of the premises themselves, and in reconciling them with each other, there is no safe course but to remain continuously on the solid ground of concrete fact.”

**A. M. Wellington**

Mathematicians, on the other hand, generally regard mathematics as an intellectual pursuit with a life of its own independent of applications; mathematics is no more simply a support subject for science than physics is just a support subject for engineering or chemistry a support subject for medicine. Mathematics as done by mathematicians bears the same relation to mathematics as used by scientists as literature bears to business communications.

> “Consider mathematics as some kind of giant computer with a large number of terminals on its periphery, representing fields of application. A practising scientist is like the terminal user. He is primarily interested in the output and will know something about what the computer can do for him, but he is not involved in what goes on inside the heart of the computer. In the early days of computers, users and designers were frequently the same people, but with their rapid growth and sophistication this is now the exception rather than the rule. Similarly it is the increasing sophistication of mathematics which has led to the large gap between ‘users’ and ‘designers’.”

**M. Atiyah**

The interest of many mathematicians in applications is limited to being pleased that nonmathematicians find mathematics important enough that it is worth encouraging and supporting (e.g. via university mathematics departments and NSF grants), but being scornful of occasional attempts by nonmathematicians or administrators to direct resources to “useful” parts of mathematics. Indeed, most mathematicians would say that no one can predict what parts of mathematics will turn out to be most useful, and anyway utility outside mathematics should play only a minor role, if any at all, in the day-to-day work of a mathematician.

These attitudes are largely shared by both “pure” and “applied” mathematicians. There is really not much systematic difference in the way pure and applied mathematicians work or in their approach to mathematics,
although there is some difference in the areas of theoretical mathematics they work in most heavily. The main difference, which provides the best definition of “pure mathematics” and “applied mathematics”, is simply in the origin of the problems they work on: in pure mathematics the problems come from within mathematics, while in applied mathematics the problems come from outside mathematics. (The applied mathematician also sometimes has the additional task of relating the solution of the mathematical problem to the original nonmathematical setting.) Thus almost all mathematicians are “pure mathematicians” at times and “applied mathematicians” at other times, and it is usually rather artificial to label a mathematician as one or the other.

Here are some more expressions of the mathematicians’ views on applications:

“Practical application is found by not looking for it, and one can say that the whole progress of civilization rests on that principle.”

J. Hadamard

“Only impractical dreamers spent two thousand years wondering about proving Euclid’s parallel postulate, and if they hadn’t done so, there would be no spaceships exploring the galaxy today.”

M. Greenberg

“Mathematics has the unique character of being a scientific discipline with applications to all other kinds of sciences and to practical life. But mathematics is also an art, the beauty lying in the symmetries, patterns, and intricately deep relationships which enchant the beholder. Discoveries that require the invention of new methods and great ingenuity are indeed to be hailed as important – at least from one point of view. Will these be of any practical use someday? Is it a legitimate question? Indeed, numerous are the examples when theories seemed for centuries to be gratuitous speculations, like the study of prime numbers, but today a mainstay of crucial applications in communications. It is the intrinsic quality of a new result which confers its importance.”

Paulo Ribenboim

“The mathematician does not study pure mathematics because it is useful; he studies it because he delights in it and he delights in it because it is beautiful.”

Henri Poincaré

“Understanding is a lot like sex. It’s got a practical purpose, but that’s not why people do it normally.”

Frank Oppenheimer

Scientists, too, express similar thoughts on an excessive preoccupation with applications:

---

78 cf. [Gre93, p. xxix].
79 [Gre93, p. xxix].
“If in the range of human endeavor after sound knowledge there is one subject that needs to be practical it is Medicine. Yet in the field of Medicine it has been found that branches such as biology and pathology must be studied for themselves and be developed by themselves with the single aim of increasing knowledge; and it is then that they can be best applied to the conduct of living processes. So also in the pursuit of mathematics, the path of practical utility is too narrow and irregular, not always leading far. The witness of history shows that, in the field of natural philosophy, mathematics will furnish the more effective assistance if, in its systematic development, its course can freely pass beyond the ever-shifting domain of use and application.”

A. R. Forsyth

“He who seeks for immediate practical use in the pursuit of science, may be reasonably sure, that he will seek in vain. Complete knowledge and complete understanding of the action of forces of nature and of the mind, is the only thing that science can aim at. The individual investigator must find his reward in the joy of new discoveries, as new victories of thought over resisting matter, in the esthetic beauty which a well-ordered domain of knowledge affords, where all parts are intellectually related, where one thing evolves from another, and all show the marks of the mind’s supremacy; he must find his reward in the consciousness of having contributed to the growing capital of knowledge on which depends the supremacy of man over the forces hostile to the spirit.”

H. Helmholtz

“In Plato’s time mathematics was purely a play of the free intellect; the mathematic-mystical reveries of a Pythagoras foreshadowed a far-reaching significance, but such a significance (except in the case of music) was as yet entirely a matter of fancy; yet even in that time mathematics was the prerequisite to all other studies! But today, when mathematics furnishes the only language by means of which we may formulate the most comprehensive laws of nature, laws which the ancients scarcely dreamed of, when moreover mathematics is the only means by which these laws may be understood,– how few learn today anything of the real essence of our mathematics! . . . In the schools of today mathematics serves only as a disciplinary study, a mental gymnastic; that it includes the highest ideal value for the comprehension of the universe, one dares scarcely to think of in view of our present day instruction.”

F. Lindemann, 1904

“What is the usefulness of a new-born child?”

Benjamin Franklin

---

82 Presidential Address, British Association for the Advancement of Science; *Nature* 56 (1897), p. 377; cf. [Mor93, 1540].
83 *Vorträge und Reden* (Braunschweig, 1884), Bd. 1, p. 142; cf. [Mor93, 1545].
84 *Lehren und Lernen in der Mathematik* (München, 1904); cf. [Mor93, 1523].
85 When the usefulness of a new invention was questioned. J. Parton, *Life and Times of Benjamin Franklin* (1864), Part IV, Chapter 17.
I.5.11. Is Mathematics an Art or a Science?

I.5.11.1. Nonmathematicians, and especially scientists, tend to think of mathematics as a science; although there are obvious differences in character between mathematics and experimental sciences, the normal demonstrative nature of visible mathematics has much in common with techniques in science. Because of this, and the many fruitful interactions between mathematics and the experimental sciences, mathematics is often organizationally grouped among the sciences in universities. Many mathematicians are not so sure. Mathematics is a unique discipline with some flavor of science but also a creative aspect which to many more resembles art. Mathematics is not simply a mechanical or purely logical process; imagination plays an essential role, and what distinguishes great mathematicians from ordinary ones is much more a difference in imagination than technical ability. (To be sure, top-quality science also has a strong creative thought component.) And mathematics possesses an intrinsic beauty and elegance which can overwhelm those who delve deeply enough into the subject to experience it.

I.5.11.2. Indeed, mathematics cannot be considered a science at all unless one believes in some form of realism, since science is by definition the systematic study and understanding of some aspect of nature. Thus a logicist, formalist, or intuitionist/constructivist cannot regard mathematics as a science. But even a realist need not entirely subscribe to the notion that mathematics is a science; the Continent of Mathematics picture, for example, portrays mathematicians as far more than just scientists.

“I like to look at mathematics almost more as an art than as a science; for the activity of a mathematician, constantly creating as he is, guided though not controlled by the external world of the senses, bears a resemblance, not fanciful I believe but real, to the activity of an artist, of a painter let us say. Rigorous deductive reasoning on the part of the mathematician may be likened here to technical skill in drawing on the part of the painter. Just as no one can become a good painter without a certain amount of skill, so no one can become a mathematician without the power to reason accurately up to a certain point. Yet these qualities, fundamental though they are, do not make a painter or mathematician worthy of the name, nor indeed are they the most important factors in the case. Other qualities of a far more subtle sort, chief among which in both cases is imagination, go to the making of a good artist or good mathematician.

Maxime Bôcher

“Mathematics, rightly viewed, possesses not only truth, but supreme beauty – a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show.”

Bertrand Russell

---

87 The study of Mathematics: Philosophical Essays; cf. R. Moritz, Memorabilia Mathematica, 1104.
“Surely the claim of mathematics to take a place among the liberal arts must now be admitted as fully made good. Whether we look at the advances made in modern geometry, in modern integral calculus, or in modern algebra, in each of these three a free handling of the material employed is now possible, and an almost unlimited scope is left to the regulated play of fancy.”

J. J. Sylvester

“The true mathematician is always a good deal of an artist, an architect, yes, of a poet. Beyond the real world, though perceptibly connected with it, mathematicians have intellectually created an ideal world, which they attempt to develop into the most perfect of all worlds, and which is being explored in every direction. None has the faintest conception of this world, except he who knows it.”

A. Pringsheim

“Mathematics has beauties of its own – a symmetry and proportion in its results, a lack of superfluidity, an exact adaptation of means to ends, which is exceedingly remarkable and to be found elsewhere only in the works of the greatest beauty. . . . The beauties of mathematics – of simplicity, of symmetry, of completeness – can and should be exemplified even to young children. When this subject is properly and concretely presented, the mental emotion should be that of enjoyment of beauty, not that of repulsion from the ugly and the unpleasant.”

J. W. A. Young

“The beautiful has its place in mathematics as elsewhere. The prose of ordinary intercourse and of business correspondence might be held to be the most practical use to which language is put, but we should be poor indeed without the literature of imagination. Mathematics too has its triumphs of the creative imagination, its beautiful theorems, its proofs and processes whose perfection of form has made them classic. He must be a ‘practical’ man who can see no poetry in mathematics.”

W. F. White

“The mathematician is entirely free, within the limits of his imagination, to construct what worlds he pleases. What he is to imagine is a matter for his own caprice; he is not thereby discovering the fundamental principles of the universe nor becoming acquainted with the ideas of God. If he can find, in experience, sets of entities which obey the same logical scheme as his mathematical entities, then he has applied his mathematics to the external world; he has created a branch of science.”

J. W. N. Sullivan

---

89 Jahresbericht der Deutschen Mathematiker Vereinigung, Bd. 32, p. 381; cf. R. Moritz, Memorabilia Mathematica, 1108.
92 Aspects of Science, 1925.
“[I]t was not its worldly utility that led Euclid or Archimedes or Georg Cantor to devote so much of their energy and genius to mathematics. These individuals did not feel compelled to justify their work with utilitarian applications any more than Shakespeare had to apologize for writing love sonnets instead of cookbooks or Van Gogh had to apologize for painting canvases instead of billboards.”

William Dunham\textsuperscript{93}

“Mathematics is not the rigid and rigidity-producing schema that the layman thinks it is; rather, in it we find ourselves at that meeting point of constraint and freedom that is the very essence of human nature.”

Hermann Weyl

“Proof is the end product of a long interaction between creative imagination and critical reasoning. Without proof the program remains incomplete, but without the imaginative input it never gets started.”

M. Atiyah\textsuperscript{94}

“Mathematics is the most perfect union between exact knowledge and theoretical thought.”

E. Curtius\textsuperscript{95}

“It is a relatively good approximation to truth \ldots that mathematical ideas originate in empirics, although the genealogy is sometimes long and obscure. But, once they are conceived, the subject begins to live a peculiar life of its own and is better compared to a creative one, governed by almost entirely aesthetical motivations.”

J. von Neumann\textsuperscript{96}

“Historically, mathematics developed out of man’s concern with physical problems. And although mathematics is ultimately a construction of the mind alone and exists only as a magnificent collection of ideas, from the earliest times down to the present day it has been stimulated and inspired by contact with the external world. Its problems are often idealizations of problems first encountered by the physicist or the engineer, or, more recently, by the social scientist. Many of its concepts are abstractions from the common experience of all men. And many areas of mathematics stem originally from the needs of scholars for more powerful analytical tools with which to pursue their investigations of the world around them. Surely, without contact with the external world, mathematics, if it existed at all, would be vastly different from what it is today.

\textsuperscript{93}[Dun90, p. vi].
\textsuperscript{94}The Princeton Companion to Mathematics, p. 1002.
\textsuperscript{95}Berliner Monatsberichte, 1873, p. 517; cf. [Mor93, 1409].
But mathematics repays generously her indebtedness to the other sciences. At the mere suggestion of a new problem, mathematics sets to work developing procedures for its solution, generalizing it, relating it to work already done and to results already known, until finally it gives back to scholars in the original field a well-developed theory for their use.

Oftentimes mathematics outruns completely the demands of the physical problem which may have stimulated it, and careless critics scoff at its ‘pure’ or ‘abstract’ or ‘useless’ character. Such criticism is absurd on two quite different counts. In the first place, all mathematics worthy of the name is pure or abstract or even, in a certain sense, useless, just as poetry and music and painting and sculpture are useless except as they bring satisfaction to those who create and to those who enjoy. It is no more appropriate to criticize mathematics for possessing the attributes of one of the creative arts than it is to criticize the arts themselves, unless it be that perhaps the number of those who find enjoyment in mathematics, though considerable, is less than those who enjoy the more conventional arts. Then in the second place, by a remarkable coincidence which is almost completely responsible for the existence of all present-day science and technology and which has no counterpart in the arts, even the most abstract parts of mathematics have turned out again and again to be highly ‘practical,’ as science, becoming ever more sophisticated, finds that it needs mathematical tools of greater and greater refinement.”

C. R. Wylie, Jr.\textsuperscript{97}

“Without mathematics we cannot penetrate deeply into philosophy. Without philosophy we cannot penetrate deeply into mathematics. Without both we cannot penetrate deeply into anything.”

\textit{Leibniz}\textsuperscript{98}

\textsuperscript{97}[Wyl64, p. 5].
\textsuperscript{98}[Cha05, p. 3].
I.6. The Foundations of Mathematics

Despite more than a hundred years of intense investigation, the foundations of mathematics remain almost as shaky as ever.

Until the nineteenth century, construction and examination of the edifice of mathematics took place almost entirely “above the surface,” building upward from such surface features as numbers, points, and lines. Mathematicians for the most part did not concern themselves with, or even recognize or acknowledge, the problem that the building might stand on shifting sand instead of solid rock.

In the latter part of the nineteenth century, continuing into the twentieth and to this day, mathematicians started exploring the part of the edifice lying below the surface. The development of set theory seemed to provide at least a basement, if not a foundation, for the structure. Logic itself seemed to lie either in a subbasement or an entire substructure of the building. But as investigations went further, it became less clear, rather than more, that there is a solid and well-formed foundation underlying it all.

I.6.1. Philosophies of Mathematics

“There is no primal language, or world view, or real line, but there are many languages, world views, and real lines, each of which can be understood to a large extent in terms of another but can only be fully understood on its own terms.”

T. W. Körner

There are, at least among philosophers, four principal philosophies of mathematics, or, more accurately, of the foundations of mathematics: realism or Platonism, intuitionism or constructivism, logicism, and formalism. Each has several variations. While the arcane details are more of interest to philosophers than working mathematicians, it is important for mathematicians to have some understanding of the differences.

Most philosophers admit that none of the four philosophies is adequate by itself as an overall philosophy of mathematics (see, for example, [Goo86]; my extended review of this article can be found at ()). Some modern-day philosophers are also critical of the historical preoccupation of philosophy with the foundations of mathematics, at the expense of studying its actual practice – see, for example, [Tym86] and [Cor03]). Indeed, each of them in its crudest form can be rather easily demolished as an absurdity. Nonetheless, each has some appealing and apparently accurate aspects, and most mathematicians (insofar as they can articulate any coherent personal philosophy of mathematics at all) seem to believe some combination of the philosophies; philosophers too acknowledge that belief in one of the philosophies does not preclude partial acceptance of the others. Indeed, I would say the primary shortcoming of each of the philosophies, at least in their more nuanced versions, is the assertion of exclusivity.

We will only give a bare introduction to these philosophies, which will necessarily be quite superficial and will give short shrift to the complex history of the ideas. Readers are referred to books on the philosophy of mathematics for details. I will also express some personal opinions; I do this advisedly, despite the warning from [Pot04, p. v]:

“... mathematicians have often been tempted, especially in later life, to commit to print philosophical reflections which are either wholly vacuous or hopelessly incoherent.”

99[Kör04, p. 394].
Realism and Platonism

Realism as a philosophy of mathematics has definitely been unfashionable among philosophers, to the extent that many philosophers talk about “the three schools of the philosophy of mathematics,” meaning the other three! However, among mathematicians it is probably the predominant view, and one which I largely subscribe to myself. (Philosophers tend to be dismissive of the opinions of mathematicians on this issue as unreliable, unsophisticated, and/or uninformed, not entirely without justification. Mathematicians are often similarly dismissive of philosophers; cf. I.6.1.30.) Mathematicians generally have a strong feeling that the abstract objects they work with are “real,” although this could of course be an illusion. This feeling is reflected in the language mathematicians use to describe advances in mathematics, almost invariably referring to “discovering” or “finding” new results or proofs. The term “invent” is rarely used, and when it is (e.g. “Newton and/or Leibniz invented calculus”) it usually refers to the development of a new subdiscipline or point of view in mathematics, which can be viewed as a way of facilitating analysis of the structure of part of the “continent of mathematics” per se.

“Mathematical truth has some sort of necessity about it, which contrasts with the merely contingent beliefs we have about the world of sense experience. There is a harshness about mathematical truth which makes it not only ineluctably true, but profoundly true, because it is immune to the changes and chances of this fleeting world of transient phenomena, and tells us about not what just is, but what must be, the case.”

J. R. Lucas

“[W]hile the other sciences search for the rules that God has chosen for this universe, we mathematicians search for the rules that even God has to obey.”

Jean-Pierre Serre

“Platonism seems obvious when you are thinking about mathematical truth, but impossible when you are thinking about mathematical knowledge.”

W. D. Hart

Existence

I.6.1.3. Some critics of realism or Platonism construe “existence” too narrowly, limiting it to actual observable physical objects. By this restricted use of the term most, if not all, mathematical “objects” do not exist. But then other things we commonly talk about also cannot exist. What about gravity, or more generally “force” as used in physics? Gravity is not an object and cannot itself be observed (at least so far – we can only observe effects we attribute to gravity). What about “good” and “evil”? Are all these things just constructs of the human mind, or do they really “exist”?

“Of course it is happening inside your head, Harry, but why on earth should that mean that it is not real?”

Albus Dumbledore

100 [Luc00, p. 20].
102 cf. [Luc00, p. 13].
Actually, what does “existence” mean for an abstract concept? This is a tough question, with many logical pitfalls. For example, while “the set of all sets” is a logical impossibility due to Russell’s Paradox, if we say that the set of all sets does not exist, then how can we talk about it? What is the meaning of the sentence “The set of all sets does not exist”? How can a sentence refer to a nonexistent concept?

We even use nonexistent objects or concepts routinely in mathematics. For example, the ancient proof that $\sqrt{2}$ is irrational, as it is usually written today (I.5.5.), begins with: “Suppose $a$ and $b$ are natural numbers with $\frac{a}{b} = \sqrt{2}$.” The proof goes on to show by contradiction that such $a$ and $b$ cannot exist. So how can we suppose they exist when they do not? Can we make a distinction between a conceivable concept and a concept that exists? If a concept does not exist, how can it be conceivable?

We often do not know whether mathematical objects exist. For example, no one knows whether “the smallest even natural number larger than two which cannot be written as a sum of two primes” exists or not. If it does not exist, how can I talk about it, as I just did?

The question of “existence” of numbers like 1 and 2 seems to me essentially the same as the question of “existence” of hydrogen and helium. Note that the existence of “hydrogen” is a subtly different question than the existence of hydrogen atoms. It is indisputable that hydrogen atoms exist in great abundance, as do helium atoms. But one needs a concept of “hydrogen” to say that two distinct hydrogen atoms are both hydrogen, and that a helium atom is not hydrogen. If “hydrogen” exists only in the human mind, how can hydrogen and helium atoms act differently chemically, and by what mechanism do all hydrogen atoms act identically? Although the concept “hydrogen” may be a human construct, there is an aggregate of properties or characteristics possessed by some atoms and not others constituting “hydrogenness” independently of any involvement of the human intellect. So it is with “oneness.” See I.6.1.12. for more comments.

If “one” and “two” exist as transcendent concepts, then what about the rest of the natural numbers? Do they all exist, and if not, what is the first one which does not?

Some things, if they “exist” at all, are certainly pure constructs of the human mind, e.g. Mickey Mouse, unicorns. But the distinction is not always so clear-cut: did the Piltdown Man ever exist, and if so, in what sense?

We can, of course, open a whole different can of worms by asking whether the empty set exists. On the one hand, the existence of anything would seem not to make sense unless we could contrast it with the existence of nothing; and if there does not exist anything, then nothing exists. On the other hand, if “nothing” exists, what kind of a thing could it be? And is the empty set the same as “nothing”, and is “zero” the same as either? See [Gar01] for an essay on this subject.

Atiyah [?] argues that even the natural numbers are simply constructs of the human mind, and would not necessarily come into the consciousness of an intelligent creature of a different form:

“But let us imagine that intelligence had resided, not in mankind, but in some vast solitary and isolated jellyfish, deep in the depths of the Pacific. It would have no experience of individual objects, only with the surrounding water. Motion, temperature and pressure would provide its basic sensory data. In such a pure continuum the discrete would not arise and there would be nothing to count.”

69
(Such a creature is certainly conceivable, and may actually exist, perhaps on Europa?) But Atiyah confuses existence with familiarity. Just because natural numbers might not be within its sense experience and may be discovered late or not at all, it does not mean that they are any less “real.” Four-dimensional geometry is not within our sense experience and thus began to be accepted and studied only relatively recently, but from a mathematical standpoint it is just as “real” as three-dimensional geometry. And eventually such a creature would find things to count, the dimensions in which it could move if nothing else; if it is intelligent enough to develop some form of calculus, it would also encounter first derivatives, second derivatives, etc.

Atiyah goes on to raise a more troubling point:

“Even more fundamentally, in a purely static universe without the notion of time, causality would disappear and with it that of logical implication and of mathematical proof.”

We can hardly conceive of the absence of time and causality, and mathematics and logic as they exist for us certainly have causality at their root. But mathematics, broadly enough defined, might well have aspects not even dependent on causality.

Atiyah makes some good points and convincingly argues that mathematics as we know and understand it is a creation of the human mind with origins in human sense experience, but his arguments that mathematical objects and truths do not have some type of independent existence do not hold water.

I.6.1.10. A common but misguided way of dismissing realism which I find particularly obnoxious is to equate it to theology or belief in God. Apparently the rationale is that everything must be created by someone, and if mathematics is discovered and not created in the human mind, it must have been created by God. This is simply baloney. One direction of the argument has some validity: realism would evidently be attractive to a religious mathematician, and it is a historical fact that some mathematicians and philosophers (including Plato himself) have been drawn to realism by their religious beliefs. But belief in mathematical realism does not in any way entail belief in a Creator, any more than belief in the existence of physical laws does; a mathematical realist is no more compelled to be religious than an astronomer or biologist.

“[Platonic] ‘existence’ is really just the objectivity of mathematical truth.”

Roger Penrose

I.6.1.11. An interesting and rather different existence question is discussed in [Gow02]: Does the complex number \( i \) exist? This is not the same as asking whether the complex numbers exist: even if we accept that they do, there are two complex numbers whose square is \(-1\), and unlike the two square roots of \( 2 \) in \( \mathbb{R} \) where there is a reasonable way to pick one (the positive one) to call \( \sqrt{2} \), there is no logical way to pick one of the square roots of \(-1\) to be called \( i \). But this is really more a question of definition rather than existence: it is really the fact that there is no well-defined (e.g. continuous) square root function defined on \( \mathbb{C} \). It also relates to the logical problem of making the common definition of \( \mathbb{C} \) as the complex plane: in addition to the logical difficulties discussed in III.6.2.24., there are two equally reasonable ways to identify \( \mathbb{C} \) with \( \mathbb{R}^2 \).

\[104\][Pen05, p. 13].
Invention vs. Discovery vs. ???

I.6.1.12. Returning to the subject of chemical elements, some recent developments in chemistry/atomic physics raise interesting questions related to the existence question of mathematical objects, and whether mathematicians “discover” or “invent” new mathematics. It was reported that scientists had, with high probability, created a few atoms of the element with atomic number 118 (now called oganesson). Although these atoms lasted for only a minute fraction of a second, the existence of this element was transformed from a theoretical possibility to a physical fact.

I.6.1.13. It does not seem accurate to say that these scientists had either “discovered” or “invented” Element 118. Some news accounts referred to the achievement as a “discovery.” This assertion seems rather similar to the common statement in history books that Peary “discovered” the North Pole. While Peary, or a member of his team, was (possibly!) the first person to reach the North Pole, he did not discover it; in fact, the North Pole was actually discovered by the first person who realized that the earth is a round ball and rotates on an axis (the North Star has been known and used for navigation since ancient and possibly even prehistoric times, although it was a different star in ancient times due to precession; once one realizes that the North Star rises higher in the sky as one moves north, it is not much of a leap to expect that it will eventually be straight overhead if one goes far enough north, assuming the earth has no edge, although uniqueness of this point on the earth is not so obvious). After all, did Neil Armstrong discover the moon?

I.6.1.14. It also does not seem accurate to say that the scientists “invented” Element 118. While they were evidently the first humans to actually assemble such an atom, the potential existence of Element 118 has long been recognized in increasingly great detail, and even the likely methods of creation of a 118 atom were widely known to specialists. This is more of a semantic question than the “discovery” one: for example, is it accurate to say that the Wright brothers “invented” the airplane? Of course, they built the first one that actually flew, and one cannot minimize the importance of that achievement; but the concept of the airplane goes back to antiquity, and specific designs of airplanes which could have flown with sufficient technical refinements go back at least to Leonardo da Vinci (in fact, if the internal combustion engine had been available to da Vinci, he might very well have built the first flying airplane.)

I.6.1.15. Perhaps the best verb for this situation is produce: the scientists were the first to produce an element 118 atom. This might be an appropriate term to also describe some advances in mathematics, avoiding the “invent” vs. “discover” question.

I.6.1.16. Scientists have long speculated about even heavier elements. For example, many scientists believe that if the right isotope of an atom with atomic number 126 could be assembled, it would be fairly stable, and they can predict with a fair degree of confidence what its chemical and physical properties would be. So let us ask: does Element 126 exist?

I.6.1.17. Of the three possible answers, Yes, No, and Maybe, the only one which cannot be rationally justified is No. No one can know that there is not a single atom of this element anywhere in the universe. A more interesting question is whether existence of such an atom somewhere is even relevant to the discussion, since it seems there is no possible way for us to ever know whether Element 126 atoms currently exist or existed at one time, say, in the remnants of a star in a galaxy millions of light-years away (according to the theory of relativity, it does not even make sense to ask whether such an atom currently exists.)
I.6.1.18. To answer the existence question, one must really decide what “existence” means in this case. Does it mean having an actual atom, or does it just mean having a very specific and rather well-understood theoretically possible physical object? Without actual atoms at hand, Element 126 is only a theoretical concept, but it seems quite different in character than, say, the postulated extra dimensions of string theory. Even a high-school student who knows about the periodic table can easily speculate about the existence of Element 126 and understand exactly (at least in general terms) what kind of physical object it would be.

I.6.1.19. It would be easy to say that Element 126 thus clearly exists as an abstract concept. After all, otherwise how could scientists even talk about it, much less study its properties? But a note of caution must be given here. Mathematicians have often talked about “the set of all sets,” and sometimes still do, even though it is now well known that such a set is a logical impossibility. Someday someone may discover a scientific principle which makes assembly of an Element 126 atom impossible; how would such a discovery affect the existence of Element 126 as an abstract concept?

I.6.1.20. Furthermore, if Element 126 is an object whose existence can be reasonably contemplated, what about Element 1000 (an atom with atomic number 1000)? I don’t imagine most chemists or physicists think there is any theoretical possibility that atoms this large could ever be created. What is the largest atomic number which is conceivable as a potential physical object? If 1000 is not big enough, what about Element $10^{100}$, which is clearly physically impossible since there are (according to our present knowledge) far less than $10^{100}$ protons and electrons in the universe.

Logicism

Logicism has the most succinct nutshell description of any of the philosophies of the foundations of mathematics: in Bertrand Russell’s words, it is the belief that “Mathematics is logic.” Or, as he later said in a more detailed and perhaps more accurate statement:

“Mathematics and logic, historically speaking, have been entirely distinct studies. . . . But both have developed in modern times: logic has become more mathematical and mathematics has become more logical. The consequence is that it has now become wholly impossible to draw a line between the two; in fact the two are one. They differ as boy and man: logic is the youth of mathematics and mathematics is the manhood of logic.”

Bertrand Russell

Specifically, logicism holds that all mathematical objects and truths, beginning with the natural numbers, are just simple consequences of the laws of logic. All more advanced parts of mathematics are built up from basic logical principles by repeated applications of the same rules of logic.

Russell [?, p. 3] gives the following definition of (pure) mathematics:

“Pure mathematics is the class of all propositions of the form ‘$p$ implies $q,$’ where $p$ and $q$ are propositions containing one or more variables, the same in the two propositions, and neither $p$ nor $q$ contains any constants except logical constants. And logical constants are all notions definable in terms of the following: Implication, the relation of a term to a class of which it is a member, the notion of such that, the notion of relation, and such further relations as may be involved in

---

105 Introduction to Mathematical Philosophy, 1919.
the general notion of propositions of the above form. In addition to these, mathematics uses a notion which is not a constituent of the propositions which it considers, namely the notion of truth.”

Quite apart from the vagueness of parts of this definition, it would surely not be considered an adequate description of mathematics by most modern mathematicians. (It should be noted that this definition was written more than a hundred years ago, and the conception of mathematics and its breadth and depth have changed markedly over this time. Indeed, Russell himself expressed second thoughts about this definition in the preface to the second edition of [? in 1938.]

It is hard to see, for example, how the Axiom of Choice, or even the Axiom of Infinity or Peano’s (second-order) axiom P5 ( ), could be considered as derivable from basic logic. Indeed, Russell admits [?, p. viii]:

“Such existence-theorems, with certain exceptions, are, I should now say, examples of propositions which can be enunciated in logical terms, but can only be proved or disproved by empirical evidence.”

This assertion seems to seriously undermine the argument that mathematics is simply a part of logic.

**Intuitionism and Constructivism**

The terms *intuitionism* and *constructivism* are often used more or less interchangeably to mean a combination of two rather distinct notions. They are usually thought of together since the primary proponent, L. Brouwer, forcefully argued for the combination. However, although there are obvious reasons why constructivism would be attractive to an intuitionist, there does not appear to be any necessary implication between the two, in either direction.

The thesis of intuitionism is that mathematics consists of the study of “mental mathematical constructions.” To an intuitionist, mathematics, or a “mathematical construction”, need not be based on formal logic or even be formalizable in the accepted sense of the term. Acceptable modes of mathematical reasoning are not completely specified, but simply consist of all reasoning the human mind considers convincing. As with all other aspects of human thought, the conception and understanding of a piece of mathematics may vary slightly from person to person, and communication of mathematics (or writing it down) is at best an imperfect representation of the mental construction involved. Thus it cannot be said that any piece of mathematics is entirely rigorous. The question of whether mathematical objects have a transcendent or mind-independent existence is not relevant in intuitionism, and perhaps does not even make sense.

I have considerable (but not complete) sympathy with the philosophy of intuitionism when it is expressed in this way, so long as the term “mental mathematical construction” is given a sufficiently liberal interpretation. For example, I see no reason to think that humans cannot make a “mental construction” of an infinite set, or even the set of real numbers. Most intuitionists have a rather restricted view of what a “mental mathematical construction” can be, and thus reject existence of infinite sets as “completed” objects, although they accept “potentially infinite” mathematical objects such as the natural numbers or even the real numbers. Most intuitionists reject the Law of the Excluded Middle, and some even reject the whole idea of negation (cf. [Gri46]–[Gri51], [Hey66, 8.2]), while others embrace the notion of contradiction in a limited form. As with all philosophies of mathematics, there is considerable variation in the specifics of the theories of various proponents of intuitionism.

Mathematicians often make use of arguments which do not have any obvious formalization, and sometimes accept such arguments as proofs. A good example is given in [Lak76, ]. Formalizable but nonconstructive
Arguments are even more common in modern mathematics, although they were unusual as recently as the nineteenth century: when Hilbert first made a name for himself by giving a fairly simple but totally nonconstructive proof of a theorem on existence of finite systems of invariants, generalizing a result which had previously been proved laboriously by a constructive argument by P. Gordan. Gordan’s reaction was “Das ist nicht Mathematik, das ist Theologie!” Gordan later warmed up to Hilbert’s methods, and admitted, “I have convinced myself that theology also has its advantages.” [Kli72, p. 930].

Even a good formalization of a theory is generally an imperfect representation of the intuitive theory. For example, ZF (and even ZFC) set theory is inadequate to settle the Continuum Hypothesis. In fact, Gödel’s Incompleteness Theorem guarantees that essentially any formalization of a mathematical theory is inadequate in some respect.

Constructivism is a rather different, although not unrelated, idea: that mathematical objects only exist when they have been given an explicit finite construction. In the most common version, the natural numbers (individually, not as a completed totality) are regarded as “given” and all other mathematical objects must be explicitly constructed from them. This is the form of constructivism espoused by Brouwer, but its origins go back at least to Kronecker, to whom is attributed the famous statement

“God made the integers [natural numbers]: all else is the work of man.”

Intuitionists find this form of constructivism natural and/or attractive since they regard the natural numbers (or, more precisely, specific small natural numbers) as being mathematical objects whose mental construction is self-evident and needs no further explanation or justification. Thus constructivism can be regarded as a natural logical consequence, or a somewhat extreme special case, of intuitionism. However, one could embrace constructivism without accepting the overall philosophy of intuitionism, e.g. it would be easy to instead justify constructivism on the basis of some type of formalism. Indeed, most constructivists greatly restrict acceptable modes of reasoning, in a manner arguably more in line with formalism than intuitionism.

The specification of natural numbers, and not, say, real numbers, points and lines, or sets, as the fundamental objects of mathematics out of which everything else must be built, can seem a bit strange and arbitrary, but constructivists have extensive arguments purporting to justify it. See ().

Among mathematicians, the most forceful modern-day proponent of constructivism was Errett Bishop, who showed that a quite remarkable amount of classical analysis can be developed within constructivist mathematics; see [BB85]. (Much of this was developed before Bishop, notably by Brouwer; see, for example, [Hey66]. This development has continued since Bishop; see, for example, [BV06].) Bishop’s book also contains a lengthy exposition of his philosophy of mathematics, which, it must be noted, differed significantly from Brouwer’s. For a less dogmatic introduction to intuitionism and constructivism, see [Hey66] or [Dum00]. Heyting actually presents intuitionist logic in an intriguing way, as a generalization of ordinary logic where statements do not just have a truth value of 0 or 1, but a truth value in a more general Boolean algebra. (This is similar to, but not the same as, fuzzy logic, which despite the name is a mathematically sound theory where truth values can be numbers between 0 and 1; fuzzy logic has yet to become a part of mainstream mathematics.)

The development of the theory of imaginary and complex numbers is an interesting case study for the philosophies of intuitionism and constructivism, although the discussion is purely academic since historically the entire development of the complex numbers predated the formulation of intuitionism or any of the other

---

106 This was said at a meeting of the Berliner Naturforscher-Versammlung in 1886; cf. Jahresbericht der Deutschen Mathematiker-Vereinigung, Bd. 2, p. 19. Bell [Bel87, p. 33] says, “As he said this in an after-dinner speech, perhaps he should not be held to it too strictly.”
modern philosophies of mathematics. Imaginary numbers were contemplated at least as far back as the 16th
century, when CARDANO, and, subsequently and more clearly, BOMBELLI, discovered that even to represent
the real roots of a cubic polynomial with real coefficients by an algebraic formula, square roots of negative
numbers must sometimes be used. They, and most other mathematicians for centuries afterward, regarded
such square roots to be “nonexistent” as numbers and “meaningless” (many, including CARDANO, even felt
the same way about negative real numbers), but found that they could be used algebraically as though they
made sense in order to obtain real results, without apparent logical contradiction. (Not all mathematicians
of the era had this problem; in the seventeenth century, JOHN WALLIS accepted the legitimacy of complex
numbers and even came close to describing the complex plane; cf. [Kli72, p. 594-595].) It was only much
later that anything resembling an actual construction was done of imaginary and complex numbers, and
that they began to be accepted as actual mathematical objects. Modern mathematicians consider the terms
“real” and “imaginary” as historical artifacts which are still used merely for convenience; today, we would
say there is nothing inherently more (or less) “real” about \( \sqrt{2} \) than about \( i = \sqrt{-1} \), although \( \sqrt{2} \) may be
more familiar to many people (but see I.6.1.11.).

An intuitionist might well say (although I doubt most would) that CARDANO and BOMBELLI had made
a “mental construction” of square roots of negative numbers, despite their reservations about whether these
objects represented anything “real”, and that imaginary numbers were part of mathematics from that time
onward. (One must be careful here: as brilliant a mathematician as FREGE made a “mental construction”
of the set of all sets, and worked for ten years on a theory based on this notion before becoming aware
of Russell’s Paradox.) A constructivist, on the other hand, probably would say that complex numbers did
not become mathematical objects until the nineteenth century, and some would say that CARDANO and
BOMBELLI were not justified in using them at all; in fact, if strict constructivist methods had been insisted
on, several hundred years of mathematical progress probably would not have happened.

Constructivists disagree about whether Goldbach’s Conjecture (that every even number greater than 2 is
a sum of two primes) has a truth value at present. Most mathematicians would argue that this statement is
definitely either true or false, and not logically indeterminate, since if it is false there is a finitely constructible
even number for which it fails. The fact that it is not now known which it is, and may never be known, is
not really relevant for most mathematicians. (On the other hand, the same argument can be made about
the consistency of ZF set theory.)

For perhaps a clearer example, suppose \( x \) is a fixed number of 100,000 digits, say with no prime factors of
less than 10,000 digits (beyond our current ability to test primality in practice). It is hard to imagine even
an ardent constructivist saying that the statement “\( x \) is prime” has no definite truth value, since there is an
obvious process of fixed finite length which will definitively establish the truth or falsity of the statement,
even though the process is not one which we are currently able to carry out. What about the statement

\[
35933408596862231041960188598043661065388726959079837 \text{ is prime.}
\]

(which is now known to be true)? Was this statement already true 100 years ago, or did it only become true
when modern computers made it possible to determine the number’s primality? What will happen in the
future if some catastrophe leaves mankind without computers – will this statement still be true?

My view is that constructivist mathematics is good and interesting mathematics, and the techniques
and results are of both theoretical and practical importance. A constructivist proof of a theorem is usually
“better” (albeit sometimes less “elegant”), and says more, than a nonconstructive one. But constructivist
mathematics is by no means all of mathematics!

This is the fundamental point of disagreement I have with (some) constructivists, who believe that
constructivist mathematics is the only legitimate mathematics. For example, one finds the following grandiose
claim made as an apparently serious assertion:
“A presentation of today’s intuitionist mathematics sets out from meaningful mathematics and, following a clear, bright road of mathematical reasons, arrives at a firm vantage from which conventional mathematics is seen to be generally false and set theory, analysis, arithmetic, algebra – all the fields of modern mathematics – reappear, scoured clean of fallacy and error.”

D. C. McCarty\textsuperscript{107}

(I should not accuse all constructivists of this exclusivity attitude – for example, the philosophy of constructivism described in \cite{BV06} is a healthy one with which I almost entirely agree.)

This quote also illustrates a point made in \cite{Luc00}: intuitionists claim that legitimate mathematical arguments are simply ones that mathematicians find convincing; but most mathematicians find classical mathematics convincing, and (many) intuitionists think they are wrong to do so.

“Intuitionists can no more argue that classical mathematicians are wrong to argue as they do than a jazz pianist can argue that Bach is wrong to compose as he did.”

J. R. Lucas\textsuperscript{108}

Until I see that the Law of the Excluded Middle and other nonconstructive arguments lead to some contradiction or absurdity, I see no reason not to use them, although I think it is very worthwhile to see how much of mathematics can be developed and proved without them. In this respect my opinion about the Law of the Excluded Middle (called the “excluded muddle” in \cite{Bel87}) is not very different from my attitude about the Axiom of Choice.

It is sometimes asserted that intuitionists/constructivists have created a “new” or “different” mathematics. I disagree with this assertion; my conception of what mathematics is is broad enough to comfortably include intuitionist mathematics as well as “classical” mathematics, and also potentially more mathematics based on other logical and language schemes.

Frege, although he is considered a logicist, also seemed to leave open the possibility of multiple logics in a statement which also admirably summarized the axiomatic method:

“It cannot be demanded that everything be proved … but we can require that all propositions used without proof be expressly declared as such, so that we can see distinctly what the whole structure rests upon. After that we must try to diminish the number of these primitive laws as far as possible by proving everything that can be proved. Furthermore, I demand – and in this I go beyond Euclid – that all methods of inference employed be specified in advance . . . .”

\textit{Gottlob Frege}\textsuperscript{109}

\section*{Formalism}

The crudest expression of formalism is the assertion that mathematics is simply a “game” of symbol manipulation according to agreed rules. As such, there is no “meaning” or “truth” in mathematics, only formal derivation of new statements from agreed starting assumptions or previously derived consequences.

\textsuperscript{107}\cite[p. 357]{Sha05}
\textsuperscript{108}\cite[p. 171]{Luc00}
\textsuperscript{109}Grundgesetze der Arithmetik, v. 1, introduction, translated by M. Furth; cf. \cite[p. 231]{Gra00}.
“Mathematics, like dialectics, is an organ of the higher sense, in its execution it is an art like eloquence. To both nothing but the form is of value; neither cares anything for content. Whether mathematics considers pennies or guineas, whether rhetoric defends truth or error, is perfectly immaterial to either.”

Goethe

“The results of systematic symbolical reasoning must always express general truths, by their nature; and do not, for their justification, require each of the steps of the process to represent some definite operation upon quantity. The absolute universality of the interpretation of symbols is the fundamental principle of their use.”

William Whewell

“If the question of ‘truth’ of axioms isn’t the mathematician’s business to discuss whose is it?”

Hilary Putnam

There are considerably more complex and nuanced versions of formalism, however, which cannot be discredited nearly as easily as the crude version of the previous paragraph. Hilbert is regarded as the father of formalism, but Hilbert's formalism is almost unrecognizably different from the description above. Hilbert maintained that there are indeed real mathematical objects such as numbers and geometric objects like points and lines, with some sort of Platonic or intuitive existence, but that the right way to derive the properties of these real objects is to abstract them into theories applying to not only the real objects but also to additional “ideal” objects which really have no meaning or existence on their own, but about which formal statements can be derived, in particular giving valid statements about the real objects one began with. When phrased in this way, formalism more or less just appears to be a codification of the abstraction process commonly accepted and used throughout mathematics.

The set of real numbers is a good and relevant example of expanding a set of “meaningful” objects into an ideal theory which can be formalized and axiomatized. Only certain real numbers can be said to have intuitive meaning, arguably only finitely many; in fact, in principle, at most countably many real numbers can even have an explicit (finite) description or characterization. But uncountably many “ideal” numbers must be added to make a theory that can be axiomatized in the standard way we describe in this book. Hilbert actually regarded any infinite set, considered as a “completed” object, to be an idealization, but did not shy away from considering and working with them formally.

Hilbert went on to state that the only arguments which should be accepted as proofs are “finitary” arguments which can be completely and explicitly carried out according to the most basic logical principles. In this respect, the limitations he argued for might not be terribly different than what constructivists insist on, although Hilbert was a severe critic of Brouwer and the intuitionists/constructivists; indeed, Hilbert regarded any kind of potential mathematical objects not leading to a logical contradiction as being things whose existence could be legitimately assumed, or could at least be worked with as though they really existed. Hilbert did not reject the role of intuition and non-finitary arguments in developing and understanding mathematics, but regarded such arguments as no substitute for finitary proofs.

110 Sprüche in Prosa, Natur IV, 946; cf. R. Moritz, Memorabilia Mathematica, 1307.
111 The Philosophy of the Inductive Sciences (London, 1858), Part I, Bk. 2, chap. 12, sect. 2; cf. R. Moritz, Memorabilia Mathematica, 1212.
Hilbert’s goal was the formalization of all of mathematics, with finite mechanical processes to determine the truth or falsity of every mathematical statement. Although his ultimate hopes were dashed by Gödel’s Incompleteness Theorem, Hilbert’s program led to the subject of metamathematics, which could be regarded as the mathematical theory of mathematical theories (including mathematical logic), and which is an important and flourishing branch of mathematics today.

“[Gödel’s Incompleteness Theorem] means that man can never eliminate the necessity of using his own intelligence, no matter how cleverly he tries.”

*Paul Rosenbloom*¹¹³

For a working mathematician, there is little practical difference between logicism and formalism; both assert that the only valid mathematical arguments are completely formal ones, and that intuition and plausibility arguments have no place in mathematics proper (at least officially). There are, however, large philosophical differences between the two, which have led to vigorous criticism of each by proponents of the other. For example, logicists regard mathematics as part of logic, while formalists consider logic, at least “mathematical logic,” to be just a part of mathematics. Formalists, at least at the basic level described above, deny that there is any notion of mathematical “truth”; logicists maintain that there is, but that mathematical “truth” is really just logical “truth.” Formalists (basic ones, not Hilbert) say that mathematics is not really “about” anything, a point of view vehemently rejected by logicists, who view logic, and hence mathematics, as being “about” everything.

**The Practice of Mathematics**

“There is probably no other science which presents such different appearances to one who cultivates it and one who does not, as mathematics. To [the noncultivator] it is ancient, venerable, and complete; a body of dry, irrefutable, unambiguous reasoning. To the mathematician, on the other hand, his science is yet in the purple bloom of vigorous youth, everywhere stretching out after the “attainable but unattained,” and full of the excitement of nascent thoughts; its logic is beset with ambiguities, and its analytic processes, like Bunyan’s road, have a quagmire on one side and a deep ditch on the other, and branch off into innumerable by-paths that end in a wilderness.”

*C. H. Chapman, 1892*¹¹⁴

Mathematics, as practiced by human mathematicians, consists of two rather distinct and almost independent aspects:

1. Ascertaining what is true and significant in mathematics, including good formulation of mathematical ideas and statements.

2. Proving that the statements thought to be true are in fact true.

Both aspects are essential. They are not entirely independent; a third aspect which bridges the other two, of comparable importance, is:


¹¹⁴Bulletin of the New York Mathematical Society, 2, 1892, p. 61; cf. [Bel92, p. v].
1.5. Understanding why the statements which are true are true.

Some mathematicians would say that 1.5 and 2 are really the same thing, but they are not (at least not exactly). One can sometimes prove a statement without having much real understanding of why it is true, except in a formal sense. And one can often gain a pretty good “understanding” of how a mathematical subject works and of what must be true without being able to actually prove it. In fact, much of calculus had this status for about 200 years.

“The logician cuts up, so to speak, each demonstration into a very great number of elementary operations; when we have examined these operations one after the other and ascertained that each is correct, are we to think we have grasped the real meaning of the demonstration? Shall we have understood it even when, by an effort of memory, we have become able to reproduce all these elementary operations in just the order in which the inventor had arranged them? Evidently not; we shall not yet possess the entire reality; that I know not what, which makes the unity of the demonstration, will completely elude us . . .

If you are present at a game of chess, it will not suffice, for the understanding of the game, to know the rules of moving the chess pieces. That will only enable you to recognize that each move has been made conformably to these rules, and this knowledge will truly have very little value. Yet this is what the reader of a book on mathematics would do if he were a logician only. To understand the game is wholly another matter; it is to know why the player moves this piece rather than that other which he could have moved without breaking the rules of the game. It is to perceive the inward reason which makes of this series of moves a sort of organized whole. This faculty is still more necessary for the player himself, that is, for the inventor.”

Henri Poincaré

“We are not very pleased when we are forced to accept a mathematical truth by virtue of a complicated chain of formal conclusions and computations, which we traverse blindly, link by link, feeling our way by touch. We want first an overview of the aim and of the road; we want to understand the idea of the proof, the deeper context.”

Hermann Weyl

It can be argued that 1, and even 1.5, are not really part of mathematics at all. This is exactly what (basic) formalism does: it gives a good, precise, and rather accurate description of 2, but has nothing to say about 1 or 1.5, and in effect excludes them from mathematics proper, and also implies that the history of mathematics has no place in the study of present-day mathematics; indeed, full acceptance of formalism (or logicism) would imply that humans did virtually no actual mathematics before the late nineteenth century (Euclid perhaps excluded). In fact, in 1901 Russell only half-jokingly wrote:

“Pure mathematics was discovered by Boole, in a work which he called The Laws of Thought (1854). This work abounds in assererations that it is not mathematical, the fact being that Boole was too modest to suppose his book the first ever written on mathematics.”

However, 1 and 1.5 are certainly parts, and important parts, of human mathematics, and a philosophy of mathematics which does not recognize and embrace this is an inadequate and empty one. Russell [?, p. 280] acknowledges this point:

“It is a peculiar fact about the genesis and growth of new disciplines that too much rigour too early imposed stifles the imagination and stultifies invention. A certain freedom from the strictures of sustained formality tends to promote the development of a subject in its early stages, even if this means the risk of a certain amount of error.”

André Weil [Wei62] elaborated on and at the same time qualified this view:

“The mathematician who first explores a promising new field is privileged to take a good deal for granted that a critical investigator would feel bound to justify step by step; at times when vast territories are being opened up, nothing could be more harmful to the progress of mathematics than a literal observance of strict standards of rigor. Nor should one forget . . . that the so-called ‘intuition’ of earlier mathematicians, reckless as their use of it may sometimes appear to us, often rested on a most painstaking study of numerous special examples, from which they gained an insight not always found among modern exponents of the axiomatic creed. At the same time, it should always be remembered that it is the duty, as it is the business, of the mathematician to prove theorems, and that this duty can never be disregarded for long without fatal effects. The experience of many centuries has shown this to be a matter on which, whatever our tastes or tendencies, whether ‘creative’ or ‘critical,’ we mathematicians dare not disagree.”

Felix Klein [Kle04, p. 207-208] similarly expounded on the nature of mathematics and formalism:

“You can hear often from non mathematicians, especially from philosophers, that mathematics consists exclusively in drawing conclusions from clearly stated premises; and that, in this process, it makes no difference what these premises signify, whether they are true or false, provided only that they do not contradict one another. But a person who has done productive mathematical work will talk quite differently. In fact those persons are thinking only of the crystallized form into which finished mathematical theories are finally cast. The investigator, however, in mathematics, as in every other science, does not work in this rigorous deductive fashion. On the contrary, he makes essential use of his phantasy and proceeds inductively, aided by heuristic expedients. One can give numerous examples of mathematicians who have discovered theorems of the greatest importance, which they were unable to prove. Should one, then, refuse to recognize this as a great accomplishment and, in deference to the above definition, insist that this is not mathematics, and only the successors who supply polished proofs are doing real mathematics? After all, it is an arbitrary thing how the word is to be used, but no judgment of value can deny that the inductive work of the person who first announces the theorem is at least as valuable as the deductive work of the one who first proves it. For both are equally necessary, and the discovery is the presupposition of the later conclusion.”

M. Postnikov [ZD07, p. 166] described the stages of development of new mathematics:

“In general, mathematical work consists of three stages. The first stage is the most pleasant: thinking up the theorem and an idea for proving it. The second step – less pleasant, but still with an element of satisfaction – is to organize the proof, fill the gaps, check and simplify the
computations, derive the necessary technical lemmas. The distasteful third stage – which, alas!, takes up 90% of one’s time – consists in writing out the proof on paper in an understandable form. It is not surprising that many mathematicians shortchange this last stage, as a result of which their work becomes a mysterious puzzle, and one often finds it easier to think up a proof oneself than to figure out the author’s text.”

Many mathematicians would take issue with the distaste Postnikov expressed about the last step. I personally find considerable satisfaction and pleasure in producing a readable exposition of a mathematical result; admittedly it is much easier with the advent of word processing to do this today, and Postnikov’s last two steps are generally more or less done simultaneously now. I also find that on average the quality of exposition in current mathematics publications is significantly better than in the past. But the separation of the development of the ideas and the formal exposition is something about which I think mathematicians would widely agree.

“Mathematics is regarded as a demonstrative science. Yet this is only one of its aspects. Finished mathematics presented in a finished form appears as purely demonstrative, consisting of proofs only. Yet mathematics in the making resembles any other human knowledge in the making. You have to guess a mathematical theorem before you prove it; you have to guess the idea of the proof before you carry through the details. You have to combine observations and follow analogies; you have to try and try again. The result of the mathematician’s creative work is demonstrative reasoning, a proof; but the proof is discovered by plausible reasoning, by guessing. If the learning of mathematics reflects to any degree the invention of mathematics, it must have a place for guessing, for plausible inference.”

G. Polya

“[I]ntuition comes to us much earlier and with much less outside influence than formal arguments which we cannot really understand unless we have reached a relatively high level of logical experience and sophistication. . . . In the first place, the beginner must be convinced that proofs deserve to be studied, that they have a purpose, that they are interesting.”

G. Polya

“If we say that a theorem has no meaning except as a conclusion from axioms, then do we say that Gauss did not know the fundamental theorem of algebra, Cauchy did not know Cauchy’s integral formula, and Cantor did not know Cantor’s theorem?”

Reuben Hersh

“The object of mathematical rigor is to sanction and legitimize the conquests of intuition, and there was never any other object for it.”

J. Hadamard

---

117[Pol90a, p. vi].
118[Pol90a]; cf. [BB08, p. vii].
119Some proposals for reviving the philosophy of mathematics
120[Pol90a]; cf. [BB08, p. 10].
“I believe the search for an explanation, for understanding, is what we should really be aiming for. Proof is simply part of that process, and sometimes its consequence.”

M. Atiyah\textsuperscript{121}

“Proofs really aren’t there to convince you that something is true – they’re there to show you why it is true.”

Andrew Gleason\textsuperscript{122}

On the other hand, it is quite remarkable that most, if not all, true mathematical statements can be proved simply through formal deductions from accepted axioms, and elegant formal, or at least quasi-formal, proofs are generally very satisfying if not always very illuminating. I personally find this aspect of mathematics rather mysterious, but enormously appealing, and it gives formalism, especially the Hilbert version, much credibility as a partial explanation of the nature of mathematics.

“Mathematical understanding develops by a mysterious interplay between intuitive insight and symbolic manipulation.”

Robert Goldblatt\textsuperscript{123}

It has been said (\textsuperscript{124}) that a typical mathematician is a Platonist during the week and a formalist on Sunday. He or she works all week formulating the problems and understanding the answer, all the time convinced that the problems and results concern real if abstract objects, and then at the end tries to write down a proof which is at least an outline of a formal argument, often leaving out or obscuring most of the ideas and intuition leading to the statement or proof in the first place. (It is fortunate that most mathematicians willingly, even eagerly, reveal such details in conversations and lectures, and sometimes in written surveys; published manuscripts are only part of the important communication which goes on in the mathematical community.)

This somewhat schizophrenic interplay of two disparate views of the nature of mathematics not only aids the progress of mathematics, but is psychologically useful too, and also convenient, as the next quote observes:

“On foundations we believe in the reality of mathematics, but of course when philosophers attack us with their paradoxes we rush behind formalism and say: ‘Mathematics is just a combination of meaningless symbols.’ \ldots Finally we are left in peace to go back to our mathematics and do it as we have always done, with the feeling that each mathematician has that he is working with something real.”

Jean Dieudonné\textsuperscript{124}

We conclude with an extensive quote which nicely summarizes the process of doing and communicating mathematics, and the healthy role formalism can play in the process:

\textsuperscript{121} The Princeton Companion to Mathematics, p. 1003.
\textsuperscript{123}[Ga98, p. v]
\textsuperscript{124}J. Dieudonné, The work of Nicholas Bourbaki, American Mathematical Monthly, 77 (1970), 134-145.
“... Let us imagine that we wish to convince someone of the correctness of a certain statement. We may try to convince him by citing some authority in which he believes (sometimes called ‘proof by intimidation’), by getting him to ‘see’ (‘proof by intuition’), by appealing to his emotions (‘proof by waving the red flag’), by confusing and tricking him with words and faulty logic (‘elastic inference’), or by other methods familiar to everyone. These methods have two serious disadvantages: (a) They can be used to establish false statements just as easily as to establish true ones, and (b) They do not lead to universal agreement.

If we wish to convince a person of the correctness of a statement in such a way that our success does not depend on his weaknesses or our cleverness and so that our argument would be equally convincing to others, we must adopt a method of proof that meets with general agreement. Whatever this method is, it must certainly involve making statements, since it is hard to conceive of convincing anyone of anything without communicating with him. So we must carry on proof by means of statements. But suppose that the listener objects to some of our statements? Then we must convince him of them before we can continue. Now we can only do this by making other statements. Again if the listener does not agree with these, we should have to try to justify them – of course by making still other statements. Evidently, we cannot convince anyone of anything unless we can get him to agree to something! Apparently, we cannot prove all statements any more than we could define all terms. We must begin by assuming some statements without proof.

If we wish to prove any theorems we must begin by assuming some axioms. But what if our listener does not agree that our axioms are truly laws? Is there no way out if we cannot obtain agreement on any axioms we propose? There is a way out. Namely we may ask our listener to agree merely that if our axioms are indeed laws, then our theorems are also laws. If we take this tack we eliminate argument about the truth of both the axioms and the theorems. The only question we ask our listener to agree on is whether the theorems really do follow from the axioms! Now this is precisely the procedure we follow in mathematics. Any formal mathematical theory begins with certain axioms and derives from them certain theorems. The mathematician claims only that the theorems follow from the axioms, and he has no objection if someone prefers to adopt different axioms.

But suppose that our listener objects to the manner in which we derive the theorems from the axioms? Then we will have to come to some agreement with him as to what procedures are legitimate. And if we cannot agree with him on method of proof? Then we take the same way out as before. We say to him, ‘Let us merely agree that if these axioms are accepted and if these methods of proof are used, then these theorems can be obtained.’ In constructing a formal mathematical theory we state in advance the acceptable methods of deriving theorems from axioms. Then all we claim is that the theory is derived from the axioms according to these rules. If the axioms and the rules are acceptable to anyone, then the theorems should be also. If the axioms are applicable to any particular situation (i.e. are true in a particular case), then the theorems can be applied to that situation.

When we construct a theory on the basis of explicitly stated axioms and rules of proof, we have not really eliminated all possibility of controversy. There may, for one thing, be differences of opinion as to whether we have correctly applied the rules of proof. But this is just a question of whether we have made a mistake. Mistakes may be hard to find, but such differences of opinion can be solved by sufficiently careful examination of the theory. Unsolvable disagreement is still possible, however, on whether the axioms and rules of proof should have been adopted at all, whether the theory that results from them is a good theory, and so on. Such questions are not answered by
the theory itself, but they are placed outside the theory when we agree to argue on the basis of the axioms and the methods of proof. Hence we may expect universal agreement within the theory, but no universal agreement about it. ... Evidently, we cannot eliminate disagreement or controversy, but we can construct a theory in such a way that disagreement is possible only about certain parts of it – namely the axioms and the methods of proof. This is a great advantage because it leads to universal agreement over a considerable area, avoids arguing about matters that can be agreed upon, and identifies the really controversial issues."

Kenneth O. May

Proof Beyond a Reasonable Doubt

I.6.1.21. There is an important trend in the standard of rigor in mathematics which is emerging (or, more accurately, reemerging) largely as a result of the advent of powerful computers which can perform extensive and sophisticated mathematical computations. This point of view is articulated rather persuasively, and sometimes entertainingly, in [BB08], [BBG04], and [BD09]. I am, however, not yet personally convinced that this approach to rigor will, or should, be the future of mathematics. See also [Kra11] for an interesting discussion.

I have adopted the legal term “proof beyond a reasonable doubt” for this approach, although mathematicians undoubtedly have a different idea of what this phrase means than the courts. The idea is that it can sometimes be established with very high probability that a result is true, even in the absence of what would be considered a rigorous proof, and sufficiently high probability of truth should be enough for mathematicians to accept the result, although the search for a rigorous proof should not be abandoned.

Such “proof beyond a reasonable doubt” normally consists of a combination of an intuitive, heuristic, and/or nonrigorous argument for the plausibility of the result, along with extensive experimental or calculation evidence supporting the result (similar to standards for accepting a scientific theory).

I.6.1.22. Such arguments are not new, and were quite common before the increased rigor of the nineteenth century. To illustrate, we consider a well-known example due to Euler. The problem was to compute

\[
\sum_{n=1}^{\infty} \frac{1}{n^2}.
\]

This series was already “known” to converge (although the notion of convergence of an infinite series in the modern sense was not yet well understood) and that the numerical value of the sum was approximately 1.645; but the problem was to find what we would call a closed form expression for the exact sum.

Euler combined together some ingenious arguments and “known” facts to convince himself that

\[
\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}
\]

(an outline of his argument is given below). He recognized that his argument was not a rigorous proof (even by the standards of the time), but simply gave a strong candidate for the exact value of the sum. He then calculated both the sum and \(\frac{\pi^2}{6}\) to a number of decimal places (about 20), and found that the numbers agreed.

\[\text{Fundamental Mathematics, preliminary edition, 1955, p. 75-78.}\]
at least to the limits of his calculation. He took this agreement as definitive verification that the formula
is correct (as indeed it is; cf. I.6.1.29.). Whether he regarded the argument plus numerical calculation as
a “proof” of the formula is somewhat unclear; he subsequently made efforts to make parts of his argument
more rigorous (see [Pó99a, p. 17-21] for a further discussion). Indeed, he did say:

“The kind of knowledge which is supported only by observations and is not yet proved must
be carefully distinguished from the truth; it is gained by induction, as we usually say. Yet we
have seen cases in which mere induction led to error. Therefore, we should take great care not to
accept as true such properties of the numbers which we have discovered by observation and which
are supported by induction alone. Indeed, we should use such a discovery as an opportunity to
investigate more exactly the properties discovered and to prove or disprove them; in both cases
we may learn something useful.”

L. Euler

It should be noted that although the modern definition of convergence of a sequence or series was not
detailed until the nineteenth century, it would be unfair and inaccurate to suggest that Euler did not
understand the difference between convergent and divergent series; in fact, he was acutely aware of the
distinction. He said he had grave doubts as to the use of divergent series, but that he had never been led
into error by using his definition of “sum”. Of course, few other mathematicians of the time, or any other
time, have had Euler’s awesome intuition and insight.

I.6.1.23. I am sure few if any modern mathematicians would regard what Euler did as a rigorous proof
of the formula. After all, there are known examples of potential formulas which are supported by heuristic
arguments and which are accurate to more decimal places than Euler calculated, but turn out not to be
exact (some examples can be found in [?]; see also III.9.4.3. and V.14.4.4.).

However, suppose Euler had had access to a supercomputer and found that the sum agreed with \( \sum_{n=1}^{\infty} \frac{1}{n^2} \)
only to a million, or a billion, decimal places. Would we take this as compelling evidence of the truth of the formula? I
imagine most mathematicians would not. But is the chance that the result fails in light of such evidence
really greater than the chance that a standard proof of some standard theorem in mathematics contains
a subtle flaw so far unnoticed by mathematicians? One cannot say for certain that even the best-known
“rigorous” mathematical proofs cannot have some inadequacy or error.

I.6.1.24. An interesting example concerns the classification of finite simple groups, which was one of
the crowning achievements of mathematics in the late twentieth century. Even the statement of the result
is far too complicated to give here, and the full proof (at least at present) consists of many research papers
totaling about 15,000 printed pages. See [GLS94] for a survey of the project and result.

Group theorists take a curious combination of positions on this statement. They generally claim that the
result is proved, i.e. a theorem, while at the same time acknowledging that it is virtually certain that some
parts of the 15,000-page proof contain errors or at least gaps. *****UPDATE*****

---

126 Opera Omnia, ser. 1, vol. 2, p. 459; cf. [Pó99a, p. 3].
127 Letter to N. Bernoulli; cf. [?], p. 322.
The Role of Computers

“The computer is a precision tool. It should not be used as a bludgeon or a substitute for thought.”

Forman S. Acton

I.6.1.25. It is unquestionably true that computers have had a significant and growing influence on the practice of mathematics. It is a matter of some disagreement whether this influence has been entirely beneficial.

There is universal agreement that computers can be very useful in many parts of mathematics, especially applied mathematics. They have long been crucial for numerical calculations needed to turn theoretical results into practical solutions to many problems. They are also valuable in “experiments” testing the plausibility of conjectures, and in helping mathematicians analyze behavior of complex mathematical systems and develop new conjectures. They can even find and identify reasonable conjectures (and not just numerical ones) which would be difficult or impossible for a human mathematician to come up with, and, perhaps even more importantly, refute some apparently reasonable conjectures.

More controversial is the use of computers to actually carry out parts of arguments constituting actual proofs (or at least purported proofs) of mathematical results. The most famous case of this is the proof of the Four-Color Theorem; see () for extensive discussions of this example. This theorem is significant at least in part because it shows that there are theorems in mathematics whose statement is simple enough to be understood by almost anyone, but whose proof is necessarily so long and complicated as to be beyond the ability of any human mathematician to completely understand it or even completely verify its correctness.

Computers and computation have also had an influence on the mindset of mathematics (or at least of mathematicians). An algorithmic approach to much of mathematics has come more into vogue. There have even been predictions that the “age of infinity” in mathematics is drawing to a close! Such an assertion is preposterous, in my opinion: while the algorithmic and computational parts of mathematics will undoubtedly grow greatly in importance and sophistication, the mathematics of infinity, including most of analysis, is and will remain far too important and central to fade away. What will instead happen is that the interplay between these two complementary parts of mathematics will become ever greater and more essential.

Fraud in Mathematics

The nature of mathematics, at least in its present state, has one nice consequence: almost uniquely among academic subjects, there is almost no possibility of fraud in mathematics. (This may change with the spread of large-scale computation as an integral part of mathematics.) Just about the only type of fraud conceivable in mathematics today is plagiarism, and even this is very hard to get away with since the world of mathematics is generally so small and interconnected that experts in a subdiscipline usually have a pretty good idea of what all the other experts are doing. And nowadays most new mathematical results are quickly written up and posted on the arXiv in preprint form, unlike in sciences where confidentiality of a paper is often jealously guarded until actual publication; as a result priority for new results in mathematics can rarely be disputed. (Correctness and completeness of purported proofs of new results may be another matter!)
Euler’s Argument

We now outline Euler’s argument for the above formula.

I.6.1.26. He began with the already “known” formula

\[ \sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \]

which is just the Maclaurin (Taylor) series, and is valid for all \( x \in \mathbb{R} \) (the validity of this formula had not yet been proved to modern standards). Thus if we set

\[ f(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!} \]

we have that

\[ f(x) = \frac{\sin x}{x} \text{ for } x \neq 0. \]

I.6.1.27. Next, he observed that if \( p(x) \) is a polynomial of degree \( m \) with roots \( a_1, \ldots, a_m \), all nonzero (but not necessarily distinct), then

\[ p(x) = p(0) \prod_{n=1}^{m} \left( 1 - \frac{x}{a_n} \right) . \]

Thus, since the roots of \( f(x) \) are \( \{ \pm \pi, \pm 2\pi, \pm 3\pi, \ldots \} \), he concluded by analogy that

\[ \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!} = f(x) = f(0) \prod_{n=1}^{\infty} \left[ \left( 1 + \frac{x}{n\pi} \right) \left( 1 - \frac{x}{n\pi} \right) \right] = \prod_{n=1}^{\infty} \left( 1 - \frac{x^2}{n^2\pi^2} \right) \]

since \( f(0) = 1 \). (This formula turns out to be correct, but the proof of correctness needs a good bit more work than Euler did.)

I.6.1.28. Finally, he multiplied out the right side as though it was a finite product, and compared the coefficients of powers of \( x \) on both sides of the equation. In particular, the coefficient of \( x^2 \) on the right side is \( -\sum_{n=1}^{\infty} \frac{1}{n^2\pi^2} \), and on the left it is \( -\frac{1}{6} \), giving the desired formula. (Comparison of coefficients of other powers of \( x \) gave additional formulas for other infinite series.)

A Proof of Euler’s Formula

I.6.1.29. For completeness, we give a slick proof of the formula, from [LeV56] (where originality is not claimed). This proof and a couple of others, including one closer in spirit to Euler’s argument, can be found in [AZ10]. The argument uses some theorems from integration theory; the special cases used can easily be established directly.

Set

\[ f(x, y) = \frac{1}{1 - xy} . \]
Let $R$ be the unit square $[0, 1]^2$ in $\mathbb{R}^2$. We will calculate

$$I = \iint_R f \, dA$$

two ways. Note that this is an improper Riemann integral since $f$ has a singularity at $(1, 1)$; but the integrand is nonnegative on $R$, so the meaning of the integral is unambiguous.

The infinite series (geometric series)

$$\sum_{k=0}^{\infty} (xy)^k$$

with nonnegative terms converges pointwise to $f$ on $R \setminus \{(1, 1)\}$; thus by the Monotone Convergence Theorem () we have

$$I = \sum_{k=0}^{\infty} \left[ \iint_R (xy)^k \, dA \right].$$

By the continuous version of Fubini’s Theorem (XIV.3.4.7.), these integrals can be calculated as iterated integrals:

$$\iint_R (xy)^k \, dA = \int_0^1 \int_0^1 x^k y^k \, dy \, dx = \left[ \int_0^1 x^k \, dx \right] \left[ \int_0^1 y^k \, dy \right] = \frac{1}{(k+1)^2}$$

and thus

$$I = \sum_{k=0}^{\infty} \frac{1}{(k+1)^2} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

On the other hand, $I$ can be calculated by a change of variable

$$x = u - v, \quad y = u + v$$

i.e. by the Change of Variables Theorem () with $\phi(u,v) = (u-v, u+v)$ we have

$$I = \iint_R f \, dA = \iint_D (f \circ \phi) |J_{\phi}| \, dA$$

where $D = \phi^{-1}(R)$ is the diamond with vertices $(0, 0)$, $(\frac{1}{2}, \frac{1}{2})$, $(\frac{1}{2}, -\frac{1}{2})$, $(1, 0)$. We have that

$$J_{\phi}(u,v) = det \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = 2$$

for all $(u,v)$ (the change of variables is linear, making things easy), so we have

$$I = 2 \iint_D (f \circ \phi) \, dA.$$

We have that

$$(f \circ \phi)(u,v) = \frac{1}{1 - u^2 + v^2}$$

and by Tonelli’s Theorem () the double integral over $D$ can be computed by an iterated integral:

$$I = 2 \left[ \int_0^{1/2} \left( \int_u^1 \frac{1}{1 - u^2 + v^2} \, dv \right) \, du + \int_1^{1/2} \left( \int_{1-u}^1 \frac{1}{1 - u^2 + v^2} \, dv \right) \, du \right].$$
(Note that the singularity of \( f \) at \((1, 1)\) “disappears” under this transformation.) We have

\[
\int \frac{1}{1 - u^2 + v^2} \, dv = \frac{1}{\sqrt{1 - u^2}} \arctan \left( \frac{v}{\sqrt{1 - u^2}} \right) + C
\]

for \( 0 \leq u < 1 \), so we have

\[
I = 4 \int_0^{1/2} \frac{1}{\sqrt{1 - u^2}} \arctan \left( \frac{u}{\sqrt{1 - u^2}} \right) \, du + 4 \int_{1/2}^{1} \frac{1}{\sqrt{1 - u^2}} \arctan \left( \frac{1 - u}{\sqrt{1 - u^2}} \right) \, du .
\]

The first integral can be evaluated by the substitution

\[
s = \arctan \left( \frac{u}{\sqrt{1 - u^2}} \right), \quad ds = \frac{1}{\sqrt{1 - u^2}} \, du
\]

to obtain

\[
\int_0^{1/2} \frac{1}{\sqrt{1 - u^2}} \arctan \left( \frac{u}{\sqrt{1 - u^2}} \right) \, du = \int_0^{\arctan(1/\sqrt{3})} s \, ds = \int_0^{\pi/6} s \, ds = \frac{1}{2} \left( \frac{\pi}{6} \right)^2 .
\]

Similarly, the second integral can be calculated by the substitution

\[
t = \arctan \left( \frac{1 - u}{\sqrt{1 - u^2}} \right) = \arctan \left( \frac{\sqrt{1 - u^2}}{1 + u} \right), \quad dt = -\frac{1}{2 \sqrt{1 - u^2}} \, du
\]

to obtain

\[
\int_{1/2}^{1} \frac{1}{\sqrt{1 - u^2}} \arctan \left( \frac{1 - u}{\sqrt{1 - u^2}} \right) \, du = 2 \int_0^{\pi/6} t \, dt = \left( \frac{\pi}{6} \right)^2 .
\]

Thus we have

\[
I = 4 \cdot \frac{1}{2} \left( \frac{\pi}{6} \right)^2 + 4 \left( \frac{\pi}{6} \right)^2 = 6 \left( \frac{\pi}{6} \right)^2 = \frac{\pi^2}{6} .
\]

Mathematicians on Philosophy

I.6.1.30. Mathematicians tend to look down on philosophy, sometimes excessively, as a subject unworthy of their time and intellect. The only aspect of philosophy which is of any widespread interest among mathematicians is the foundations of mathematics, and even this interest is pretty limited. Here are some of the opinions expressed about philosophy; I emphasize that I do not personally endorse them, but just present them in the interest of full disclosure:

“[P]hilosophy proper is a subject, on the one hand so hopelessly obscure, on the other so astonishingly elementary, that there knowledge hardly counts.”

G. H. Hardy\textsuperscript{129}

\textsuperscript{129}Mathematical Proof, in [7, p. 173].
“Philosophers may be unable to understand how we can know mathematical truths, but they have a bad track record at being able to understand anything.”

J. R. Lucas\textsuperscript{130}

“... People who study philosophy too long, and don’t treat it simply as part of their early education and then drop it, become, most of them, very odd birds, not to say thoroughly vicious; while even the best of them are reduced by this study ... to complete uselessness as members of society.”

Plato\textsuperscript{131}

**Should Mathematicians Care About Foundations?**

“Mathematics changes slowly. After ten years, some problems have been solved and some new problems have arisen, but the subject is recognisably the same. However, over a century it changes in ways which mathematicians at the beginning of that century could not conceive. ... Given this, it seems presumptuous to seek a permanent foundation for mathematics. It is a happy accident that almost all of mathematics, as we now know it, can be deduced from one set of axioms; but, just as history has taught us that there is not one geometry but many, so it is not unlikely that the future will present us with different mathematics deduced from radically different axiom schemes. The art of the mathematician will still consist in the rigorous proof of unexpected conclusions from simple premises, but those premises will change in ways that we cannot possibly imagine.”

T. W. Körner\textsuperscript{132}

I.6.1.31. The question of which, if any, of the philosophies of the foundations of mathematics is correct has rather little to do with the day-to-day work of most mathematicians. Mathematicians generally just take for granted that mathematics has some sort of logical, consistent foundation, whether it is understood yet or not. I think it is fair to say that the only foundational issue that does, or probably even should, concern every mathematician is the consistency of mathematics.

“God exists since mathematics is consistent, and the Devil exists since we cannot prove it.”

André Weil\textsuperscript{133}

Even if it were shown that, say, ZF set theory is inconsistent, most mathematicians would be at least moderately interested, but not particularly upset; this would just mean that ZF set theory was not the right theory to base mathematics on after all. They would then simply go on doing their mathematics as usual.

The only thing that would really upset mathematicians is if mathematics itself (e.g. real analysis as axiomatized in (), or geometry, algebra, etc.) were discovered to be inconsistent. Mathematicians mostly

\textsuperscript{130}Luc00, p. 19.

\textsuperscript{131}The Republic, VII.3; cf. ?.

\textsuperscript{132}Kör04, p. 356.

\textsuperscript{133}In H. Eves, Mathematical Circles Adieu, Boston: Prindle, Weber and Schmidt, 1977.
just take on faith that this is an impossibility; indeed, belief in some form of realism would preclude this possibility. After all, the very existence of mathematics as a subject is entirely based on the belief that if one starts with correct statements and applies valid, careful reasoning, correct conclusions will always be obtained. It seems inconceivable to us that an inconsistency could show up in mathematics, especially since the massive edifice already created has produced no hint of inconsistency. At the very core of every mathematician is a certain conviction of the consistency of mathematics, reinforced by all the mathematics they see and do throughout their careers.

“If mathematical thinking is defective, where are we to find truth and certitude?”

D. Hilbert

I.6.2. Mathematical Theories

I.6.3. Languages

Any theory must be expressed in some sort of language. Actually, it is generally accepted that there must be at the outset some sort of metalanguage which is simply understood, and in which certain phrases and terms have meaning. The metalanguage is not part of the theory, or really part of mathematics at all; it is just a vehicle for communication and specification of the theory. In [Kle02], KLEENE calls this the observer’s language (with the term “metalanguage” reserved for an observer’s language with certain technical restrictions), which seems apt terminology.

In addition, a mathematical theory must have a specified language in which it is expressed, called the object language in [Kle02]. At a minimum, the language of a theory must define some collection of “statements” which are considered “meaningful” for the theory (although no actual meaning in the usual sense of the term need be attached to them), and there needs to be a way to combine together statements or transform them into new statements according to specific rules.

There does not seem to be any necessity that the language consist of finite one-dimensional strings of discrete symbols (human spoken language is an obvious exception), but languages in mathematics, called formal languages, are normally taken to be of this sort: an alphabet of symbols is specified, and certain finite strings of these symbols are the “statements.” This restriction is necessary (at least at present) to make formulation and analysis of mathematical theories manageable. It can be argued that formal languages can satisfactorily encode anything which could be called a “language”, although it must be noted that even ordinary written language is at best an imperfect representation of human spoken language; in spoken language there is a lot of subtle information conveyed in timing, stress and intonation, facial expressions, gestures, and body language which is not captured in the written language (even transmission of spoken language over a telephone or radio causes significant losses of information).

In a good formal language we need to have the ability to make definitions. In effect, a definition is simply the establishment of an “abbreviation” for a (generally more complicated) phrase. In principle, the defined term can always be replaced by the phrase it abbreviates without loss of utility, and thus from a strictly logical point of view definitions are unnecessary. However, in practice it is very useful, often essential, to be able to make definitions: otherwise, statements involving even moderately complex mathematical ideas would become unmanageably or even incomprehensibly complicated.

134 On the Infinite, cf. [?].
I.6.4. Logic

“Logic is the systematic study of the conditions of valid inference.”

\textit{Abraham Wolf}\textsuperscript{135}

“Logic is the systematic study of the structure of propositions and of the general conditions of valid inference by a method which abstracts from the content or \textit{matter} of the propositions and deals only with their logical \textit{form}.”

\textit{Alonzo Church}\textsuperscript{136}

Logic is a discipline which is not merely descriptive, but normative as well, being concerned more with how we \textit{ought} to reason than with how we habitually do so. … We shall seek to construct an ideal theory of correct reasoning, based on a small number of general principles which we postulate as ‘laws of thought.’ ”

\textit{G. T. Kneebone}\textsuperscript{137}

“Few persons care to study logic, because everyone conceives himself to be proficient enough in the art of reasoning already.”

\textit{C. S. Peirce}\textsuperscript{138}

As with languages, a distinction must be made between what may be called \textit{metalogic} (this term is not commonly used in the literature), which is an underlying feature, and \textit{logic}, which is a part of the mathematics. Unlike with language, this distinction is often not carefully made and has led to much confusion and disagreement over the last century, with arguments about whether logic was part of mathematics or vice versa (or neither). The subject of \textit{mathematical logic} has since become established.

Metalogic can be thought of as the rules of inference which are assumed to underly not only mathematics but all rational discourse. An example of a principle of metalogic is transitivity of implication:

\begin{quote}
If $P$ implies $Q$ and $Q$ implies $R$, then $P$ implies $R$.
\end{quote}

(Here $P$, $Q$, and $R$ are statements.) Another example, called \textit{modus ponens}, is

\begin{quote}
If $P$ is true and $P$ implies $Q$, then $Q$ is true.
\end{quote}

To rephrase \textit{modus ponens} without using the loaded term “true”, we could say instead:

\begin{quote}
If we assume or conclude $P$ and we assume or conclude ($P$ implies $Q$), then we conclude $Q$.
\end{quote}

I.6.5. Mathematical Structures and Models

\textsuperscript{135}Encyclopaedia Britannica, fourteenth edition, 1929; cf. [?], p. 5.
\textsuperscript{137}[?], p. 8.
\textsuperscript{138}Collected Papers; cf. [?].
I.7. Printed or Online?

When this book project began, I envisioned the result as a traditional printed volume. But I fairly quickly became convinced that it should be an online book instead. An online book has at least four enormous advantages over a printed one:

(1) An online book is much more widely, easily, and freely available to potential readers. Production, reproduction, and distribution costs are almost nonexistent. The path from author to reader is direct without having to pass through a publisher.

(2) Unlike with a printed book, there is no practical limit to the length of an online book. Redundancy (within reason) does not have to be avoided.

(3) An online book does not have to be organized linearly.

(4) Revisions, corrections, and additions can be made immediately and painlessly.

Advantage (1) is obvious and needs no further comment. As to (2), an author of a printed book usually has to make difficult and delicate decisions about what to include and what to omit, and how much detail should be included. A printed book with too much material or detail not only becomes physically unwieldy and expensive to produce, but also creates psychological problems for the reader: it can be difficult for a reader to plow through a mass of detail to extract the essentials he/she needs, and an abundance of ancillary material can be intimidating to a reader not yet familiar with the subject and can make the theory appear to be more complex and difficult than necessary.

An online book, if properly organized, avoids most (but not all) of these problems. Extra details, technicalities, and side threads can be included in such a way that they are invisible to readers not specifically interested in looking at them. The book can consist of a large number of separate sections connected only by links organized by a complicated directed graph. Each section can have a set of links to the prerequisite sections and another set of links to potential successors. (However, one serious danger in a non-linear organization of a mathematics book is the possibility of circularity in definitions and arguments; great care must be exercised to avoid circularity.)

A reader of a book organized in this way will probably need some guidance through the material, a “road map.” Actually, several road maps will be required for readers of various backgrounds and interests, and I plan to provide some. If this book is used as a text or supplementary text for a course, the instructor will have to settle on a road map appropriate for the course, or provide his/her own. Every instructor has a slightly different idea about how the material of the subject should be organized and presented. This book by and large represents my idea, but I am trying to make the organization flexible enough that an instructor who wants a different flow through the material can develop a satisfactory road map for the desired journey.

One advantage related to (4) is that the book can be placed online before it is “finished,” so long as it is far enough along to be useful. (The currently posted version is already probably about the 1000th edition; it is likely that most of these editions have not been read by anyone but me!) In fact, my expectation is that I will never regard the book as finished. I have a dream that this book will eventually turn into an all-encompassing treatment of real analysis. Of course, I am far from having the personal expertise, or enough time and energy left in my lifetime, to accomplish this goal by myself. I hope eventually that others will contribute to the project by adding treatments of additional material and/or alternative treatments of material already covered. I hope also that experts interested in other parts of mathematics, especially parts of mathematics that are mature enough that a reasonably definitive exposition can be given, will undertake
similar projects in their specialties, and that eventually a massive and interconnected body of knowledge and understanding of a large part of mathematics will exist. Previous attempts have been made along this line, notably by Bourbaki, but I do not regard my efforts or goals to be at all comparable or duplicative of such work. The current and rapidly expanding treatment of mathematics in Wikipedia and other websites is laudable and very useful, but also not comparable in goals. (Wikipedia articles on some other subjects, especially ones with political or nationalistic overtones, are notoriously unreliable; however, I have found most Wikipedia articles on mathematics to be quite good, although they should not be taken as gospel. Articles on mathematicians and/or the history of mathematics are not always as good.)

It must be said that even in an online book of potentially infinite length, the necessity of making decisions about what to include and what to omit is not completely avoided unless one plans to cover all of mathematics. The decision is particularly critical in regard to material which is ancilliary to the main topic of the book, such as the intricacies of set theory or topology in my case. I have necessarily employed subjective standards based on my personal interests and limitations, and I have included more detail than necessary on some topics I like, and (at least so far) much too little detail on other topics. At least I can rationalize some of these decisions by saying I have not yet finished a complete treatment of the subject. Or I suppose I could just invoke the words of Descartes [?]:

“I hope that posterity will judge me kindly, not only as to the things of which I have explained, but also as to those which I have intentionally omitted so as to leave to others the pleasure of discovery.”

I expect that many, if not most, sections of this book will eventually have “short” (“appetizer”) and “long” (“main course”) versions. The short version will generally be more of a survey of the high points of the specific topic, with most proofs and technical details suppressed, and sometimes with definitions and/or statements of results stated somewhat informally. (Current versions of many sections probably fall somewhere between the ultimate short and long versions.) I will insist, though, that even the short treatments be mathematically “honest.”

There are two common kinds of mathematical “dishonesty” frequently encountered in lower-level (and sometimes higher-level) texts:

(1) Suggesting that a certain mathematical statement is true more generally than it actually is.

(2) Suggesting that a certain mathematical statement is “obvious” and does not require proof, when in fact it does.

A variation of (2), related to (1), is

(2') Suggesting that an informal definition of something (e.g. tangent line, area) is mathematically adequate.

A third kind of mathematical dishonesty, somewhat less common but still frequently seen, and arguably more serious than (1) and (2), is

(3) Giving a purported “proof” of something which is not complete or not rigorous, while suggesting it is.

“Few will deny that even in the first scientific instruction in mathematics the most rigorous method is to be given preference over all the others. Especially will every teacher prefer a consistent proof to one which is based on fallacies or proceeds in a vicious circle, indeed it will be morally impossible for the teacher to present a proof of the latter kind consciously and thus
in a sense deceive his pupils. Notwithstanding these objectionable so-called proofs, so far as the foundation and the development of the system is concerned, predominate in our textbooks to the present time. Perhaps it will be answered, that rigorous proof is found too difficult for the pupil’s power of comprehension. Should this be anywhere the case, – which would only indicate some defect in the plan or treatment of the whole, – the only remedy would be to merely state the theorem in a historic way, and forego a proof with the frank confession that no proof has been found which could be comprehended by the pupil; a remedy which is ever doubtful and should only be applied in the case of extreme necessity. But this remedy is to be preferred to a proof which is no proof, and is therefore wholly unintelligible to the pupil, or deceives him with an appearance of knowledge which opens the door to all superficiality and lack of scientific method."

H. Grassmann, 1904

In my opinion, as a general rule a written argument should not be called a “proof” unless it is complete and rigorous (although I must admit that I myself have not always adhered to this principle!)

I certainly do not mean to suggest that it is necessary, or even desirable, that all the technicalities, ramifications, or potential pathologies connected with a particular topic being discussed in an elementary text be treated or even mentioned. However, I believe that mathematical honesty is possible even in, say, a Business Calculus text, although maintaining mathematical honesty while at the same time keeping the discussion at a reasonable and appropriate level of detail and sophistication is a fine art and not easily accomplished.

I.7.1. How Much Generality?

Another constantly encountered problem in preparing an exposition of real analysis (or of many other parts of mathematics) is the question of whether to do things in the utmost generality, or stick to an important special case. There is no simple answer to this dilemma, and I found it impossible (as well as undesirable) to be entirely consistent.

There are several considerations, which are often in tension:

1. It is desirable to have results recorded in the greatest generality that can be reasonably stated, for future reference, even if the immediate applications one has in mind require only a special case.
2. It is often easier to understand and appreciate a proof if the statement of the result contains the exact hypotheses needed to make the proof work.

On the other hand:

3. The proof of the most general case of a result can often be considerably harder or more delicate than that of a crucial special case.
4. Even the statement of the most general result can be much harder to write or digest than a clean statement of a special case.
5. It may seem farfetched in some cases that increased generality would ever be useful, especially if it requires more work to state or prove than the result of obvious importance.

I have had to make many decisions of this type in the manuscript, and have each time had to settle on the best tradeoff between generality, difficulty, and usefulness. My decisions may not have always been the wisest, and will be subject to review and possible revision in the future.

139 cf. [Mor93, #538].
I.7.2. History vs. Logic

“Once a good building is completed its scaffolding should no longer be seen.”

*C. F. Gauss*140

“Whoever wants to get to know a science shouldn’t just grab the ripe fruit – he must also pay attention to how and where it grew.”

*J. C. Poggendorf*141

There are two competing schools of thought as to how mathematics, especially “finished” mathematics such as Real Analysis, should be presented:

(1) The ideas and results should be presented in the most efficient, elegant, and logical order, without regard to their origins or history.

(2) The historical development and context of all results and ideas should be carefully traced and presented.

The ultimate exponent of approach (1) is *Bourbaki*. There are numerous books on Analysis taking the second approach. Most textbooks basically take the first approach, with some lip service to the second.

New mathematics is usually not very elegant at first. The first proof of a result is rarely the most efficient one or done in the optimum generality, and historically it has often happened that the first “proof” of an important result is not even complete or correct. In fact, BESICOVITCH once said that a mathematician should be judged by how many bad proofs he has produced (he meant to imply a positive correlation). It usually takes time and effort, often quite a bit of it, before the “proper” approach or point of view for a part of mathematics is developed.

“The ways of discovery must necessarily be very different from the shortest way, indirect and circuitous, with many windings and retreats. It’s only at a later stage of knowledge, when a new domain has been sufficiently explored, that it becomes possible to reconstruct the whole theory on a logical basis, and show how it might have been discovered by an omniscient being, that is, if there had been no need for discovering it!”

*George Sarton*142

I have tried to steer somewhat of a middle ground between the two approaches, although the result is of necessity much closer to the first than the second since I am a mathematician, not a historian. But I have made considerable efforts to accurately put many things in historical context, not only because it is beneficial in understanding where the mathematics came from, but also because much of the history is fascinating. I have particularly tried to correctly trace the history of some things which I discovered have been commonly misportrayed in modern references. I am certainly not the first to try to do this: for example, in [Kle04, p. 234], *KLEIN* points out that what is (even today!) commonly called the “Maclaurin series” had been previously obtained by TAYLOR, explicitly stated as a special case of the Taylor series, and was acknowledged as such by MACLAURIN himself. *KLEIN* states:

---

140cf. [Rem98, p. viii].

141cf. [Rem98, p. vii].

“Errors of this sort are common. . . . It is difficult, if not impossible, to overcome such deeprooted absurdities. At best, one can only spread the truth in the small circle of those who have historical interests.”

I hope this is unduly pessimistic!

Despite my efforts along these lines, I am sure my amateur historical research is highly inadequate in many cases and probably inaccurate in some respects. I will endeavor to gradually improve it as I learn more.
Chapter II

The Formalities of Mathematics

In this chapter, we have collected together discussions of the basic language and tools of mathematics (not just analysis). These vary widely in sophistication, ranging from elementary (from the point of view of modern mathematics), e.g. formal statements, basic set theory, functions, and relations, to very advanced, e.g. the intricacies of the Axiom of Choice, ordinals, and large cardinals. The sections in this chapter can be regarded primarily as reference material for our treatment of analysis.
II.1. Formal Statements, Negations, and Mathematical Arguments

“Mathematics is often considered a difficult and mysterious science, because of the numerous symbols which it employs. Of course, nothing is more incomprehensible than a symbolism which we do not understand. Also a symbolism, which we only partially understand and are unaccustomed to use, is difficult to follow. In exactly the same way the technical terms of any profession or trade are incomprehensible to those who have never been trained to use them. But this is not because they are difficult in themselves. On the contrary they have invariably been introduced to make things easy. So in mathematics, granted that we are giving any serious attention to mathematical ideas, the symbolism is invariably an immense simplification.”

A. N. Whitehead

“The domain over which the language of analysis extends its sway, is, indeed, relatively limited, but within this domain it so infinitely excels ordinary language that its attempt to follow the former must be given up after a few steps. The mathematician, who knows how to think in this marvelously condensed language, is as different from the mechanical computer as heaven from earth.”

A. Pringsheim

“The prominent reason a mathematician can be judged by none but mathematicians, is that he uses a peculiar language. The language of mathesis is special and untranslatable. In its simplest form it can be translated, as, for instance, we say a right angle to mean a square corner. But you go a little higher in the science of mathematics, and it is impossible to dispense with a peculiar language. It would defy all the power of Mercury himself to explain to a person ignorant of the science what is meant by the single phrase ‘functional exponent.’ . . . But to one who has learned this language, it is the most precise and clear of all modes of expression. It discloses the thought exactly as conceived by the writer, with more or less beauty of form, but never with obscurity.”

Thomas Hill

Mathematical statements are typically given in a hybrid of symbolic language and standard English. For example:

“For every real number \( x \) there is a unique integer \( n \) such that \( n \leq x < n+1 \).”

(Of course, comparable statements can be made equally well using other languages than English.) It is fairly easy to understand what a statement like this means, but many mathematical statements, particularly in analysis, are considerably more complicated, such as:

“For every real number \( a \) and every \( \epsilon > 0 \) there is a \( \delta > 0 \) such that, for every real number \( x \) with \( |x - a| < \delta \), \( |f(x) - f(a)| < \epsilon \).”

---

1. Introduction to Mathematics, 1911, p. 59-60; cf. [Mor93, 1218].
(Some statements are much more complicated than this!) We need to have a systematic way to analyze the meaning of a statement like this. In particular, it is often important to be able to write (and understand) the negation of such a statement. A very useful tool for these problems is to rewrite mathematical statements in a language-independent formal way. This section describes how to do this.

The presentation in this section is intended for students preparing to be working mathematicians, and we will not strictly adhere to the (sometimes rather arcane) conventions of the subject of mathematical logic. Thus our notation and terminology will not necessarily agree with that found in texts on logic.

II.1.1. Formal Statements

II.1.1.1. A formal statement equivalent to the second one above is

\[(\forall a \in \mathbb{R})(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in \mathbb{R})[|x - a| < \delta \Rightarrow (|f(x) - f(a)| < \epsilon)].\]

We will explain what this formal statement means and how to get it.

II.1.1.2. A formal statement (or proposition) will be a string of symbols; however, not every string of symbols will be regarded as a formal statement. The allowable symbols will be of three types: constants, variables, and logical symbols. A formal statement will then be, roughly speaking, a string of allowable symbols which is in a form to which a truth value (true or false) can potentially be attached, perhaps depending on the values assigned to the variables in the statement. We will in due course be more precise.

Constants

II.1.1.3. Constants are symbols standing for previously defined specific objects. The objects may be elements, sets, ordered pairs, functions, relations, binary operations, or any other type of object specifically defined in the theory being discussed (all the types of objects listed are officially sets, so one could just say that the objects are all sets to be precise). Additional constants can be defined in the course of the discussion. Note that some standard objects in mathematics have names consisting of compound symbols, such as \(\sin\), \(\frac{d}{dx}\), or \(\int \cdot \, dx\), which are regarded as single constants. Note also that a function which is a constant in this sense (i.e. simply a specified function) is generally not a “constant function” in the function sense of taking the same value at any input.

Variables

II.1.1.4. Variables are similar to constants, but are symbols which have not been previously defined to represent specific objects, and thus can potentially represent many different objects. The type of object that a particular variable represents can be specified, and can be any of the types of objects that can be constants.

The distinction between variables and constants can be a subtle one. For example, in the statement in II.1.1.1., the \(f\) might be either a constant or a variable depending whether it had previously been defined to be a specific function. Even more subtle is the specification of variable names which are treated essentially like constants although they remain variables; this is discussed in II.1.7.7.

Logical Symbols

II.1.1.5. There are only a few logical symbols we will use. These fall into the categories of relations, set operations, connectives, and quantifiers, and one additional symbol for negation.
Relations:

\[ = \text{ (“equals”) } \]
\[ \in \text{ (“is a member of” or “belongs to”) } \]
\[ \subseteq \text{ (“is a subset of”) } \]

Set Operations:

\[ \cup \text{ (“union”) } \]
\[ \cap \text{ (“intersection”) } \]
\[ \setminus \text{ (“relative complement”) } \]

Connectives:

\[ \land \text{ (“and”) } \]
\[ \lor \text{ (“or”) } \]
\[ \implies \text{ (“implies”) } \]

Quantifiers:

\[ \exists \text{ (“there exists”) } \]
\[ \forall \text{ (“for all”) } \]

Negation:

\[ \neg \text{ (“not”) } \]

The symbols \( \land, \implies, \) and \( \neg \) are sometimes written \&, \( \rightarrow \), and \( \sim \) respectively. The set operation \( \setminus \) is also sometimes written \( \sim \) or \( - \).

II.1.1.6. The set-theoretic operations are not strictly speaking logical symbols, but constant operations (functions) defined in set theory. However, they are so universal that we can regard them as logical symbols. Other set-theoretic operations such as symmetric difference can also be defined and given symbols. The relations \( \in \) and \( \subseteq \) are also technically constant relations defined in set theory.

II.1.1.7. The symbols we have given here are somewhat redundant; not all are needed, since some can be defined in terms of others, as described below. But it is useful to have all of them since equivalent statements written without them can be considerably more complicated. A few other useful logical symbols such as \( \iff \) and \( ! \) will be defined later in terms of the standard ones.
Parentheses

II.1.1.8. Parentheses are technically not symbols of the formal language, and are used merely to make the formal statements more comprehensible. They can be avoided entirely if certain conventions of interpretation are strictly adhered to. But they are very desirable to have in practice, and we will make use of them freely. For clarity, nested parentheses can be written as (), [], or rarely even {}, which all have the same meaning (or lack thereof).

Terms

II.1.1.9. Definition. A simple term of a certain type is a constant or variable of the specified type. A compound term of level 2 of a certain type is a function whose value is of the specified type, whose arguments are simple terms of the appropriate types. A compound term of level $n \geq 2$ of a certain type is a function whose value is of the specified type, whose arguments are simple terms or compound terms of level $< n$ of the appropriate type.

A simple term or a compound term of any (finite) level is called a term.

The level of a compound term is of little if any significance; even the distinction between simple and compound terms is unimportant. The important thing is that terms are objects of a specified type built up from constants and variables through a finite number of steps.

II.1.1.10. Examples. Examples of simple terms of type “real number” are $0$, $\sqrt{2}$, $\pi$, $x$ where $x$ is a variable of type “real number.” If $f$ is a variable which is of type “real-valued function of three real arguments”, and we recall that $+$ is a constant of type “function from pairs of real numbers to real numbers”, then such expressions as $f(x, y, 5)$ and $x + z$ (where $x, y, z$ are variables of type “real number”) are compound terms of level 2 of type “real number”, and $f(x + y, x + 2, f(x, x, y))$ is a compound term of level 3 of type “real number.”

Atomic Statements

II.1.1.11. Definition. An atomic statement, or simple statement, is an expression of the form $R(x_1, \ldots, x_n)$, where $R$ is an $n$-ary constant or logical relation ($n \geq 2$) and $x_1, \ldots, x_n$ are terms (simple or compound) of the appropriate types.

The words sentence and proposition are often used interchangeably with statement.

II.1.1.12. Examples. Expressions of the form $x = y$, $(x + 5) \in A$, $f(x, y, \sin(z)) > 0$, etc., are atomic statements, where $x, y, z$ are variables of type “real number” and $A$ is a constant or variable of type “set of real numbers.”

General Formal Statements

II.1.1.13. Definition. A formal statement is an expression built up from atomic statements by a finite number of connectives, quantifications, and negations. Specifically:

Every atomic statement is a formal statement.

If $P$ and $Q$ are formal statements, then the following are also formal statements:
\( P \land Q \).
\( P \lor Q \).
\( P \Rightarrow Q \).
\( \neg P \).

\((\forall x)(P)\), where \( x \) is a variable not already quantified in \( P \).

\((\exists x)(P)\), where \( x \) is a variable not already quantified in \( P \).

(Quantification is discussed in detail in the next subsection.)

The level of a formal statement can be defined as in II.1.1.9.: atomic statements have level 1, and every connective, quantification, or negation adds one to the level. Every formal statement has level \( n \) for some finite \( n \).

**First-Order vs. Higher-Order Statements**

II.1.1.14. There is an important point, which, however, need not concern us very much at this level. In most contexts, there is a natural set of objects being discussed, usually called the *domain of discourse*. In most of this book, the domain of discourse would be the real numbers.

II.1.1.15. While constants can be not only objects in the domain of discourse, but also subsets of this domain, functions and relations on this domain, etc., logicians often (in fact, usually) restrict variables to represent only actual elements of the domain of discourse. Statements containing only variables taking values in the domain of discourse are called *first-order statements*, and statements with variables ranging over larger sets are *higher-order statements*. Levels of higher-order statements can be defined similarly to II.1.1.9.; for example, a statement containing one or more variables ranging over subsets of the domain of discourse or functions from the domain of discourse to itself is a *second-order statement*. Thus, for example, the statements of II.1.1.1. and II.1.3.11. are first-order statements, but the statement in II.1.3.12. is second-order.

II.1.1.16. The distinction between first-order and higher-order statements can be eliminated by expanding the domain of discourse. However, in practice the domain of discourse is usually established by a set of axioms assumed in the discussion, so cannot be easily expanded. (Also, even expanding the domain of discourse does not convert second-order theories to first-order theories.)

II.1.1.17. A *first-order theory* is a theory in which all statements are first-order. “First-order logic” can be used in first-order theories. “Higher-order logic” is needed for higher-order theories; higher-order logic is not as well behaved, or as well understood, as first-order logic, and indeed “logic” means “first-order logic” to many logicians.

One common criticism of second-order logic, made particularly in [?], is that second-order logic is essentially first-order set theory, and hence is “mathematical” and not “logic”; in [?] second-order logic is called “set theory in sheep’s clothing.” See [?] for an attempt to refute this criticism.

While restriction to first-order logic may make sense to logicians, it does not work well for mathematics in general; second-order logic is really needed to properly describe real analysis and even ordinary arithmetic.
the way mathematicians do it. See e.g. [Luc00] or [Sha05, Chapter 25] for details. As J. Barwise [?, p. 5] puts it:

As logicians, we do our subject a disservice by convincing others that logic is first-order and then convincing them that almost none of the concepts of modern mathematics can really be captured in first-order logic."

II.1.1.18. We will feel free to use higher-order statements (i.e. quantify as we see fit) in our development of analysis, and just assume enough higher-order logic exists to justify this policy (this assumption is in fact justifiable despite the incompleteness of higher-order logic).

II.1.2. Meaning of Statements

II.1.2.1. The “meaning” of a formal statement is not something that can really be defined. We will limit our discussion to the interpretation of truth values for formal statements.

II.1.2.2. The truth value of an atomic formal statement cannot be determined (in most cases) unless values for the variables are specified. Once they are, the statement is true if and only if the resulting n-tuple is one which satisfies the defining relation (technically, the relation is a set of n-tuples, and the relation is “satisfied” by a particular n-tuple if it is in the relation). A statement is false if it is not true.

II.1.2.3. Truth values for compound statements depend on the truth values of the constituent parts in a precise way which we will now describe. The first three are simple and fairly obvious:

\[(P \land Q)\] is true if and only if both \(P\) and \(Q\) are true.

\[(P \lor Q)\] is true if and only if \(P\) is true or \(Q\) is true or both. (Thus \(\lor\) is sometimes called “inclusive or”, as opposed to the “exclusive or” used primarily in computer science.)

\[\neg P\] is true if and only if \(P\) is false.

Implication

II.1.2.4. The meaning of \(\Rightarrow\) is less obvious. \((P \Rightarrow Q)\) is read “\(P\) implies \(Q\),” “If \(P\) then \(Q\),” or “\(P\) only if \(Q\).” Mathematicians have found through experience that the proper meaning of \((P \Rightarrow Q)\) is

\[\neg P \lor Q\] .

The idea is that if \(P\) is true, then \(Q\) must necessarily be true, but if \(P\) is false, no conclusion about the truth of \(Q\) may be drawn. So the implication is only violated if \(P\) is true but \(Q\) is false; the implication is satisfied (not violated) in all other cases. If \(P\) is false, the implication is not violated irrespective of whether \(Q\) is true or false, and if \(Q\) is true, the implication is not violated irrespective of whether \(P\) is true or false.

This interpretation leads to some jarring conclusions in cases where there appears to be no logical relationship between \(P\) and \(Q\). Thus mathematicians would regard the implication

“\(1 = 2\), then the sky is green.”

as being “true.” But there is no other reasonable interpretation in terms of truth values for \((P \Rightarrow Q)\) than \[\neg P \lor Q\].
II.1.2.5. The connective $\Rightarrow$ can thus be defined in terms of $\neg$ and $\lor$, and so is strictly speaking redundant. But we will use it freely since it is natural and many statements are simpler and more comprehensible if it is used.

Converse, Inverse, and Contrapositive

II.1.2.6. Definition. Let $P$ and $Q$ be statements.

(i) The converse of the statement $P \Rightarrow Q$ is $Q \Rightarrow P$.

(ii) The inverse of the statement $P \Rightarrow Q$ is $(-P) \Rightarrow (-Q)$.

(iii) The contrapositive of the statement $P \Rightarrow Q$ is $(-Q) \Rightarrow (-P)$.

II.1.2.7. It is important and basic that the converse of a statement does not have the same meaning; there is no implication either way between a statement and its converse, as almost any example shows. It is a common mistake to confuse a statement and its converse, which requires practice to learn to avoid. However:

II.1.2.8. Proposition. A statement and its contrapositive are logically equivalent (if ordinary logic is used).

Proof: The statement $P \Rightarrow Q$ means $(-P) \lor Q$, and its contrapositive $(-Q) \Rightarrow (-P)$ means $[\neg(-Q)] \lor (-P)$, and $\neg(-Q)$ is equivalent to $Q$ in ordinary logic.

II.1.2.9. Note that similarly, the converse and inverse of a statement are logically equivalent; indeed, the inverse is the contrapositive of the converse.

II.1.2.10. So to prove a statement, it is just as good to prove the contrapositive, and this is frequently easier or more natural. In fact, we will often just prove the contrapositive of a statement without comment.

If and Only If

II.1.2.11. We can define one other convenient logical connective, which can be regarded as merely an abbreviation if desired. If $P$ and $Q$ are statements, we write $P \iff Q$, read “$P$ if and only if $Q$”, to mean

$$[(P \Rightarrow Q) \land (Q \Rightarrow P)] .$$

Rewriting this without the symbol $\Rightarrow$, we obtain

$$[((-P) \lor Q) \land [P \lor (-Q)]] .$$

The statement $P \iff Q$ is also logically equivalent to

$$[(P \land Q) \lor [(-P) \land (-Q)]] .$$
II.1.2.12. The following table summarizes the behavior of truth values under connectives and negation:

<table>
<thead>
<tr>
<th>$P$</th>
<th>$Q$</th>
<th>$\neg P$</th>
<th>$(P \land Q)$</th>
<th>$(P \lor Q)$</th>
<th>$(P \Rightarrow Q)$</th>
<th>$(P \Leftrightarrow Q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

II.1.3. Quantification and Closed Statements

II.1.3.1. The quantifiers $\exists$ and $\forall$ are called the existential quantifier and the universal quantifier respectively. They are used to modify a statement $P$ by specifying the sense in which a certain variable is used in $P$. A variable appearing in $P$ which has not been quantified is called free; free variables give a degree of incompleteness in a statement. For example, if we just state “$x > 3$”, we cannot pass judgment on the truth of the statement without knowing what the $x$ is supposed to stand for. But if we write $(\forall x)(x > 3)$ (“For all $x$, $x > 3$”) or $(\exists x)(x > 3)$ (“There exists an $x$ such that $x > 3$”), we have a self-contained statement and we know what we are talking about, and it makes sense to say the statements are false and true respectively (in the real numbers).

Note, however, that if the statement $x > 3$ appears in the course of a discussion, and $x$ has been previously defined in the discussion to be a specific number (e.g. suppose there has been a previous line “Let $x = \sqrt{10}$.”), then the $x$ in the statement $x > 3$ is a constant, not a variable, and a truth value can be unambiguously assigned (true in this case).

Quantified variables can be regarded as “dummy variables” whose names are really just placeholders in an expression. The $x$ in $(\forall x)(x > 3)$ plays the same role as the $x$ in $f(x) = x^2 + 1$ or $\int_0^1 x^2 \, dx$, or the $n$ in $\sum_{n=1}^t(2n - 1)$. The symbol used is unimportant; it can be replaced by any other variable name which does not already appear in the expression without changing the meaning.

II.1.3.2. A variable in a statement $P$ which has been quantified is called a bound variable. A bound variable cannot be quantified again.

Logicians use the term “bound” to refer only to a variable quantified by an existential quantifier. In fact, universal quantifiers are normally not used at all in mathematical logic; they are often just understood, or if needed can be defined in terms of existential quantifiers and negations ($\neg$). But universal quantifiers are very useful in practice in mathematics, and it is good form to use them explicitly rather than to let them just be understood or implicit; statements with explicit universal quantifiers (where needed) are generally much clearer and more comprehensible.

Restricted Quantification

II.1.3.3. When quantifying variables, we often (even usually) want to place explicit restrictions on the values the variables are allowed to take. Thus we often want to write expressions like $(\forall n \in \mathbb{N})$ or $(\exists \delta > 0)$. With proper interpretation, such restricted quantification is allowable and causes no problems; a restricted quantification of a statement will just be an ordinary quantification of a related but more complicated statement.
II.1.3.4. In the usual case where we have a specified domain of discourse, e.g. the real numbers $\mathbb{R}$, the expressions $(\forall x \in \mathbb{R})$ or $(\exists x \in \mathbb{R})$ mean exactly the same thing as the simple expressions $(\forall x)$ and $(\exists x)$ respectively, and writing $(\forall x \in \mathbb{R})$ or $(\exists x \in \mathbb{R})$ simply adds to the clarity of the statement by specifying what type of variable $x$ is.

II.1.3.5. In the case where $A$ is a constant (previously specified) subset of the domain of discourse, we take $(\forall x \in A)(P)$ to mean
\[(\forall x)[(x \in A) \Rightarrow P]\]
and we take $(\exists x \in A)(P)$ to mean
\[(\exists x)[(x \in A) \land P].\]

II.1.3.6. Similarly, if $R$ is a relation on the domain of discourse, and we write $R(x)$ to denote $R$ with $x$ used as one of the arguments, along with other terms, then by $(\forall x R(x))(P)$ we mean
\[(\forall x)[R(x) \Rightarrow P]\]
and by $(\exists x R(x))$ we mean
\[(\exists x)[R(x) \land P].\]

For example, if the domain of discourse is $\mathbb{R}$, the expression $(\exists \delta > 0)(P)$ means
\[(\exists \delta)[(\delta > 0) \land P].\]

Order of Quantifiers

II.1.3.7. In a statement with more than one quantifier, the order of quantifiers is crucial; interchanging the order of a universal quantifier and an existential quantifier can drastically alter the meaning of a statement. For example, consider the following statement:

“For every natural number $n$ there is a real number $x$ with $x > n$.”

The corresponding formal statement is
\[(\forall n \in \mathbb{N})(\exists x \in \mathbb{R})(x > n).\]

This statement is true (e.g. can take $x = n + 1$).

But suppose the order of the quantifiers is reversed, i.e. we take the formal statement
\[(\exists x \in \mathbb{R})(\forall n \in \mathbb{N})(x > n)\]

which in English says

“There is a real number $x$ such that for every natural number $n$, $x > n$.”

This statement is false (). The point is that in the first statement, only the existence of an $x$ satisfying $x > n$ for one particular, though unspecified, $n$ is asserted (the $x$ can depend on the $n$), whereas in the second it is asserted that there exists one $x$ which satisfies the inequality simultaneously for all $n.$
II.1.3.8. There is an implication one way, however: for any statement $P$, and for any free variables $x$ and $y$ in $P$, we have

$$[(\exists x)(\forall y)(P)] \Rightarrow [(\forall y)(\exists x)(P)].$$

II.1.3.9. The order of two consecutive universal quantifiers, or two consecutive existential quantifiers, can be reversed without changing the meaning of the statement. If two consecutive quantifiers of the same type have identical restrictions, they can be combined into one abbreviated quantification expression. For example, the expressions $(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})$ and $(\forall y \in \mathbb{R})(\forall x \in \mathbb{R})$ have an identical meaning, and can be abbreviated as $(\forall x, y \in \mathbb{R})$. We can similarly write $(\exists x, y \in \mathbb{R})$ to mean $(\exists x \in \mathbb{R})(\exists y \in \mathbb{R})$ or $(\exists y \in \mathbb{R})(\exists x \in \mathbb{R})$.

Closed Statements

II.1.3.10. DEFINITION. A closed statement is a formal statement $P$ in which every variable is quantified (i.e. every variable in $P$ is bound, or $P$ contains no free variables).

Closed statements are the ones which are “self-contained” and to which a well-defined truth value can (potentially) be attached. Restricted quantification is allowable in closed statements; such a statement can be rewritten as a more complicated closed statement with only ordinary quantification.

II.1.3.11. In principle, the statement of any theorem is a closed statement, and it is good mathematical form to write the statement of every result in the form of, or in a form easily translatable into, a closed statement. In practice it may sometimes be overly pedantic to do this, and we sometimes abbreviate when no confusion is likely. We also sometimes for stylistic reasons write theorem statements in a form not exactly of a closed statement. For example, many theorems are phrased in the form of the next statement, or slight variations of it:

**Theorem.** Let $x$ and $y$ be real numbers. If $x < y$, then there is a rational number $r$ with $x < r < y$.

The corresponding closed formal statement is

$$(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})[((x < y) \Rightarrow ((\exists r \in \mathbb{Q})[(x < r) \land (r < y)]))]$$

which is a little awkward to express directly in English.

II.1.3.12. Sometimes the formal statement is even more different from the English one. For example:

**Theorem.** Every nonempty subset of $\mathbb{N}$ has a smallest element.

A corresponding formal statement would be:

$$(\forall S \subseteq \mathbb{N})[(\exists n)(n \in S) \Rightarrow ((\exists m)[(m \in S) \land (\forall n \in S)(m \leq n)])].$$

This statement explicitly incorporates the meaning of “nonempty” and “smallest element.” It also illustrates that variable objects which can remain unnamed in the English statement must often be given names in the formal statement.
II.1.3.13. One mistake often made by beginners (and sometimes by non-beginners too!) is to write a statement where the same letter is used to mean two or more different things in the expression, i.e. to stand for two or more different variables or constants. The statement in II.1.3.12. does not have this mistake since the two uses of \( n \) occur within disjoint quantifications (convince yourself of this!), but it would have been safer, and perhaps clearer, to use another letter in place of one of the uses of \( n \). It is always safest to use a separate letter or symbol for every different object in a statement. See II.1.7.7. for further discussion of a related point.

II.1.3.14. There is a related (mostly nonmathematical) issue concerning style. In a mathematical statement or argument, it is a good idea to use easily distinguishable symbols for unrelated objects; using the same letter in upper and lower case, or in different typefaces, can be confusing to a reader. It is often good practice to use related symbols for related objects: for example, it is usually good practice to use a lower-case letter to denote an element of a set whose name is the same upper-case letter (e.g. \( x \in X, n \in \mathbb{N} \)). Some extension of this is often also good practice, e.g.

"Let \( x_n \in X \) for all \( n \geq N \)."

(Here \( N \) is a previously specified number.) But this can be taken to ridiculous lengths. For example, I have seen the following statement in a book (which shall remain anonymous):

"Let \( A_a \in \mathfrak{A}_a \) for all \( a \in A \)."

Here \( A \) was an index set, \( a \) an index, and \( \mathfrak{A}_a \) a set for each \( a \). Although there is nothing mathematically wrong with such a statement, DON'T DO THIS! (I admit we have been forced by standard conventions to do something like this in Section XIII.4.3., but it is still to be avoided whenever possible.)

II.1.3.15. There is a wide choice of possible symbols available for use in mathematical statements. In principle any sort of symbol can be used, although in practice one should use symbols which can be easily recognized, distinguished, and remembered (and pronounced!) Besides standard mathematical symbols, one should generally stick to Latin or Greek letters, although other symbols are sometimes used. German \textit{fraktur} is common (more common in older references). Hebrew letters, notably \( \aleph \), are standard in certain contexts. I admit to occasionally using Cyrillic letters. And in a paper with a Danish coauthor, we used the Danish letter \( æ \) in a context where it was appropriate, although we refrained from also using the Danish letter \( ø \), as in

"Let \( \emptyset \) be a set, and \( \emptyset \in \emptyset \)."

I recommend exercising restraint in using uncommon letters or symbols.

II.1.3.16. A closed statement need not have any variables at all. Ones which do not tend to be simple formulas like elementary statements about arithmetic (e.g. \( 2 + 2 = 4 \)), but they are not necessarily so simple. One example is the following result, which caused a minor stir when it was discovered by F. Cole in 1903 since it provided a counterexample to an old conjecture of Mersenne in number theory (cf. [Bel87, p. 228]):

\textbf{Theorem.} \( 2^{67} - 1 = 193707721 \times 761838257287 \).
II.1.4. Negations

II.1.4.1. It is particularly important to be able to write the negation of a mathematical statement. In one sense, it is a triviality to do this: just insert the word “not” (and perhaps adjust the statement slightly to make it grammatically acceptable), i.e. the negation of the statement \( P \) is \( \neg P \). But this begs the question: in order to have a useful phrasing of the negation of a statement (e.g. in order to disprove the original statement, or to prove it by contradiction), we need a way of writing the negation affirmatively without the words “no” or “not”. (In fact, it is most useful to write every mathematical statement affirmatively without such words when possible, and correspondingly to write formal statements without using the symbol \( \neg \), and it turns out it is almost always possible to do this.)

It is often quite difficult to do this for statements written in English (try it for the two statements in the introduction of this section!) But there is a systematic way to affirmatively write the negation of a formal statement (although it involves what could be regarded as a bit of a subterfuge for atomic statements). So a procedure for finding the negation of a mathematical statement expressed in English would be:

1. Rewrite the statement as a formal statement.
2. Find the negation of the formal statement.
3. If desired, convert the negated formal statement into an English statement.

II.1.4.2. We should emphasize that this discussion of negations uses ordinary (or “classical”) logic. Negations behave rather differently if intuitionistic logic is used (cf. (1)).

Negations of Atomic Statements

II.1.4.3. It is quite simple to affirmatively write the negation of an atomic statement: just replace the relation in the statement with its complementary relation. (A relation is a subset of a Cartesian product, and the complementary relation is just the set-theoretic complement in the Cartesian product.) Sometimes there is a simple formula for the complementary relation: for example, if the domain of discourse is \( \mathbb{R} \), the negation of the statement \( (x > 3) \) is \( (x \leq 3) \). But it is not always so simple: for example, in a general domain of discourse there is no simple way to write the negation of the sentence \( (x = y) \) except as \( (x \neq y) \). (If the domain of discourse is \( \mathbb{R} \), the negation can be written \([ (x < y) \lor (x > y) ] \), but this is rather special to \( \mathbb{R} \).)

Negations of Compound Statements

II.1.4.4. The negation of \( \neg P \) is \( P \), i.e. \( P \) is logically equivalent to \( \neg \neg P \) for any statement \( P \).

II.1.4.5. The negation of \( P \lor Q \) is

\[
[(\neg P) \land (\neg Q)].
\]

To see this, note that \( P \lor Q \) is true if and only if \( P \) is true or \( Q \) is true or both. Thus \( P \lor Q \) is false if and only if both \( P \) and \( Q \) are false.

An almost identical argument (or an application of II.1.4.4.) shows that the negation of \( P \land Q \) is

\[
[(\neg P) \lor (\neg Q)].
\]

So if \( P \) and \( Q \) are sentences whose negations can be written without use of \( \neg \) (e.g. if \( P \) and \( Q \) are atomic sentences), then the negations of \( P \lor Q \) and \( P \land Q \) can be written also without use of \( \neg \).
II.1.4.6. Example. The negation of \([(x > 3) \lor (x \leq 1)]\) is \([x \leq 3] \land (x > 1)\], which can be abbreviated \((1 < x \leq 3)\).

II.1.4.7. Since \((P \Rightarrow Q)\) is logically equivalent to \(\neg[(\neg P) \lor Q]\), the negation of \((P \Rightarrow Q)\) is

\[
[P \land (\neg Q)]
\]

So if \(P\) and \(Q\) are sentences whose negations can be written without use of \(\neg\) (e.g. if \(P\) and \(Q\) are atomic sentences), then the negation of \(P \Rightarrow Q\) can be written also without use of \(\neg\).

II.1.4.8. Since we have \(\neg(P \lor Q)\) logically equivalent to \([\neg P] \land (\neg Q)]\), we have that \((P \lor Q)\) is logically equivalent to \(\neg[(\neg P) \land (\neg Q)]\). Thus \((P \lor Q)\) can be defined to mean \((\neg P) \land (\neg Q)\), i.e. \(\lor\) can be defined in terms of \(\neg\) and \(\land\). Similarly, we have that \((P \land Q)\) is logically equivalent to \(\neg[(\neg P) \lor (\neg Q)]\), i.e. \(\land\) can be defined in terms of \(\neg\) and \(\lor\). We will not want to do this, for two reasons: (1) both \(\lor\) and \(\land\) are natural and useful in writing comprehensible formal statements, and (2) we actually want to avoid using \(\neg\) as much as possible.

II.1.4.9. The negation of \((P \iff Q)\) can be found by writing it as \([P \lor (\neg Q)] \land [Q \lor (\neg P)]\), whose negation is \([(\neg P) \land Q] \lor [(\neg Q) \land P]\). \((P \iff Q)\) can also be rewritten as \([P \land Q] \lor [(\neg P) \land (\neg Q)]\), whose negation is \([P \lor Q] \land [(\neg P) \lor (\neg Q)]\).

Negations of Quantified Statements

II.1.4.10. The negation of \((\exists x)(P)\) is

\[
(\forall x)(\neg P)
\]

To see this, note that \((\exists x)(P)\) is true if and only if there is at least one \(x\) making \(P\) true. The negation is that every \(x\) makes \(\neg P\) false, i.e. every \(x\) makes \(\neg P\) true.

Similarly (or by an application of II.1.4.4.), the negation of \((\forall x)(P)\) is

\[
(\exists x)(\neg P)
\]

Thus to take the negation of a statement with a quantifier, just take the negation of the inside statement and change the quantifier from existential to universal or vice versa.

II.1.4.11. If a statement has two or more consecutive quantifiers, change each but leave the order of the quantifiers unchanged. Thus the negation of the statement \((\forall x)(\exists y)(P)\) is

\[
(\exists x)(\forall y)(\neg P)
\]

Negations of Statements with Restricted Quantifiers

It can be confusing to try to reason out the negation of a statement with restricted quantifiers. It is easy to make the mistake of changing too much!
II.1.4.12. Proposition. The negation of the statement \((\forall x \in A)(P)\) is
\[(\exists x \in A)(\neg P)\].

The negation of the statement \((\exists x \in A)(P)\) is
\[(\forall x \in A)(\neg P)\].

Proof: These can be reasoned out just as ordinary quantifiers, but we can also argue formally on the basis of the meaning of the restricted quantifiers. Since \((\forall x \in A)(P)\) means
\[(\forall x)[(x \in A) \land P]\]
we have that the negation is
\[(\exists x)[\neg(x \in A) \lor (\neg P)]\]
which by definition is the same as
\[(\exists x)[(x \in A) \Rightarrow (\neg P)]\]
and this is the meaning of \((\exists x \in A)(\neg P)\). The second half is the same argument run backwards, with \(P\) and \(\neg P\) interchanged.

By an almost identical argument, we obtain:

II.1.4.13. Proposition. If \(R\) is a relation, then the negation of \((\forall x R(x))(P)\) is
\[(\exists x R(x))(\neg P)\]
and the negation of \((\exists x R(x))(P)\) is
\[(\forall x R(x))(\neg P)\].

II.1.4.14. Thus, in taking the negation of a statement with restricted quantifiers, the inside statement is negated, universal and existential quantifiers are switched, but the restrictions on each quantifier stay the same. If there is more than one quantifier, the order of the quantifiers stays the same.

II.1.4.15. Example. Consider the statement
\[(\forall a \in \mathbb{R})(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in \mathbb{R})[|x - a| < \delta \Rightarrow (|f(x) - f(a)| < \epsilon)]\]
of II.1.1.1. Its negation is
\[(\exists a \in \mathbb{R})(\exists \epsilon > 0)(\forall \delta > 0)(\exists x \in \mathbb{R})[|x - a| < \delta \land (|f(x) - f(a)| \geq \epsilon)]\]
(using an application of II.1.4.7. on the statement inside the quantifiers).
II.1.4.16. By repeated applications of the techniques of this section, the negation of any formal statement \( P \) can be written. If \( P \) can be written without use of \( \neg \), so can its negation.

II.1.5. Uniqueness Statements

II.1.6. Tautologies

II.1.6.1. A tautology is a statement form which is always true no matter what truth values are assigned to the constituents. Some standard tautologies (for ordinary logic) are:

\[
\neg[P \land (\neg P)] \quad (\text{Consistency}) \\
P \lor (\neg P) \quad (\text{Law of the Excluded Middle}) \\
[P \Rightarrow Q] \iff [(\neg P) \lor Q] \quad (\text{Definition of } \Rightarrow)
\]

Some tautologies discussed in the section on negation are:

\[
P \iff \neg(\neg P) \quad (\text{Law of Double Negation}) \\
[\neg(P \lor Q)] \iff [(\neg P) \land (\neg Q)] \\
[\neg(P \land Q)] \iff [(\neg P) \lor (\neg Q)] \\
[\neg(P \Rightarrow Q)] \iff [P \land (\neg Q)] \\
[\neg[(\forall x)(P)]] \iff [(\exists x)(\neg P)] \\
[\neg[(\exists x)(P)]] \iff [(\forall x)(\neg P)]
\]

Some other tautologies which can be verified by using truth tables:

\[
[(P \lor Q) \lor S] \iff [P \lor (Q \lor S)] \quad (\text{Associativity of } \lor) \\
[(P \land Q) \land S] \iff [P \land (Q \land S)] \quad (\text{Associativity of } \land) \\
(P \lor Q) \iff (Q \lor P) \quad (\text{Commutativity of } \lor) \\
(P \land Q) \iff (Q \land P) \quad (\text{Commutativity of } \land) \\
[(P \lor Q) \land S] \iff [(P \land S) \lor (Q \land S)] \quad (\text{Distributivity of } \land) \\
[(P \land Q) \lor S] \iff [(P \lor S) \land (Q \lor S)] \quad (\text{Distributivity of } \lor) \\
[(P \Rightarrow Q) \land (Q \Rightarrow S)] \Rightarrow [P \Rightarrow S] \quad (\text{Transitivity of } \Rightarrow \text{ or Syllogism Law}) \\
[P \Rightarrow Q] \iff [(\neg Q) \Rightarrow (\neg P)] \quad (\text{Law of Contrapositive}) \\
[P \Rightarrow (Q \Rightarrow S)] \iff [(P \land Q) \Rightarrow S] \\
P \iff (P \land [Q \lor (\neg Q)]) \\
P \iff (P \lor [Q \land (\neg Q)])
\]
\[ P \leftrightarrow ([P \land Q] \lor [P \land (\neg Q)]) \]

\[ P \leftrightarrow ([P \lor Q] \land [P \lor (\neg Q)]) \]

\[ P \leftrightarrow ([Q \Rightarrow P] \land [(\neg Q) \Rightarrow P]) \text{ (Case Testing)} \]

\[ [(\exists x)(\forall y)(P)] \Rightarrow [(\forall y)(\exists x)(P)] \]

\[ P \Rightarrow P \text{ (note that this is by definition equivalent to the Law of the Excluded Middle, and thus is not a tautology in intuitionist logic!)} \]

\[ P \Rightarrow (P \lor Q) \]

\[ (P \land Q) \Rightarrow P \]

\[ (\neg P) \Rightarrow (P \Rightarrow Q) \]

\[ Q \Rightarrow (P \Rightarrow Q) \]

\[ [P \land (P \Rightarrow Q)] \Rightarrow Q \text{ (Modus Ponens)} \]

II.1.6.2. A tautology of the form \( A \iff B \), where \( A \) and \( B \) are statement forms involving \( P \)'s, \( Q \)'s, etc., is called a tautological equivalence. A tautological equivalence is a logical equivalence, i.e. any occurrence of a statement of form \( A \) can be replaced by the corresponding statement of form \( B \), or vice versa, without changing the logical meaning of the statement. (A logical equivalence need not be a tautological equivalence, however.) Such substitutions allow for simplification or modification of statements in arguments.

II.1.6.3. A tautology of form \( A \Rightarrow B \) can be used similarly, allowing replacement of a statement of form \( A \) by the corresponding one of form \( B \) (but not conversely). The converted statement is not (necessarily) logically equivalent to the original, but is a logical consequence. This type of substitution can also be used in arguments; however, such an argument is then not reversible.

Use of modus ponens is a particularly important example: if one shows that \( P \) is true and that \( (P \Rightarrow Q) \) is true, it can be concluded that \( Q \) is true. This logical argument is used constantly.

II.1.6.4. If intuitionist logic is used, not all of these tautologies hold. For example, intuitionists do not accept the Law of the Excluded Middle (and hence not \( P \Rightarrow P \) either!) They (mostly) accept \( P \Rightarrow [\neg(\neg P)] \), but not the converse. Several of the other tautologies listed above are not valid in intuitionist logic; we leave it to the reader to identify which ones.

II.1.7. Definitions

II.1.7.1. Definitions in mathematics are of two rather different sorts:

(1) A precise specification of the meaning of a word or phrase.

(2) A specification that a certain symbol (or compound symbol) is to stand for a particular mathematical object (i.e. the specification of a constant in the sense of II.1.3.).
II.1.7.2. **Examples.** (i) Examples of definitions of the first kind are the definitions of *partial order* (), *field* (), *metric space* (), *measure* (), *Riemann integral* (), · · · .

(ii) Examples of definitions of the second type are expressions of the form

(a) “Let $f(x) = x^2 + 1.$”

(b) “Let $x = \lim_{k \to \infty} x_k.$”

(c) “Let $B$ be the set of all upper bounds of the set $A$ [where $A$ has been previously defined].”

II.1.7.3. A definition of the first kind is simply the establishment of an “abbreviation” for a (generally more complicated) phrase. In principle, the defined term can always be replaced by the phrase it abbreviates without loss of utility, and thus from a strictly logical point of view such definitions are unnecessary. However, in practice it is very useful, often essential, to be able to make these definitions: otherwise, statements involving even moderately complex mathematical ideas would become unmanageably or even incomprehensibly complicated.

Definitions of the first kind are often of words which are standard terms used widely by mathematicians, and are frequently even the basic subject matter of whole parts of mathematics. But definitions of this kind can also be conveniently used to create a name for a concept of no obvious interest beyond the specific discussion at hand but which must be referred to multiple times, such as a concept which will shortly be shown to be equivalent to some other well-known concept.

II.1.7.4. Definitions of the second kind are another story. In principle, they too can be avoided, but in practice it is often hard to see how arguments can be reasonably phrased without their use.

II.1.7.5. Definitions of the second kind are normally made as part of a mathematical argument, and are operative only for the duration of the argument. Definitions of the first kind can be made in the same way as part of an argument, but are more commonly stand-alone statements.

Note that a definition is not a formal mathematical statement as described in this section, although it may resemble a formal statement in form.

II.1.7.6. Sometimes a definition of the second kind is dependent on other results for it to make sense. For example, in II.1.7.2 (iib), it must be previously (or subsequently) shown that $\lim_{k \to \infty} x_k$ exists, and that limits are unique.

A common instance where a definition must be justified is a situation where the object defined apparently or potentially depends on some choices. It must be shown that the definition always gives the same result no matter how the choices are made. In this situation, we say the resulting object is *well defined*. For example, we might make a definition (cf. (i))

“Let $f(x) = \lim_{k \to \infty} f(x_k)$, where $(x_k)$ is any sequence in $E$ converging to $x$.”

To show that $f(x)$ is well defined, it must be shown not only that $\lim_{k \to \infty} f(x_k)$ always exists for every such choice of $(x_k)$, but also that the same limit is always obtained no matter how the $x_k$ are chosen (so long as they are in $E$ and converge to $x$).

Although from a logical standpoint justifications should be put before the definition, the justification for the legitimacy of a definition is stylistically often put after the definition itself.
Ambiguity of Terminology

Ideally, every term used in mathematics should have a unique, unambiguous precise definition. But in practice there are many instances where the same name is used in mathematics to refer to quite different concepts in different contexts (this is probably inevitable since there is only a limited supply of suitable words in English to describe mathematical concepts, and is not even entirely undesirable since there is often some similarity between distinct concepts). Some common examples are closed, complete, free, independent, regular, normal, and separable. In different instances there may be a greater or lesser philosophical resemblance between the usages; but there is still a tendency for the mathematically uninitiated to make too much of the coincidence of terminology by taking the names too literally.

In most if not all languages where mathematical terms have been named, there are similar instances where the same name is used in different contexts. But the ambiguities are not always the same. For example, French mathematicians use the word fermé to refer to closed sets in topology, but algebriquement clos for algebraically closed fields. On the other hand, the French word noeud is used to mean both node and knot (and I have seen it translated incorrectly in English translations of French mathematics books!) Thus a French-speaking neophyte in mathematics may confuse different mathematics concepts than an English-speaking one.

Separated Quantification

II.1.7.7. There is a variant of type two definitions which is frequently employed, which is technically not a true definition at all. In mathematical arguments, we often write phrases such as

“Let $x$ be an element of $A$.”

(Here, $A$ is a previously defined set; in order to legitimately make such a statement, it must be known, or already shown, that $A$ is nonempty.) This is not really a definition, since it does not specify exactly what $x$ is. The $x$ remains a variable, but from that point on it is treated essentially like a constant in that it must always refer to the same (variable) object (this is related to the discussion in II.1.3.13.). If we then show that a statement $P(x)$ containing the variable $x$ is true, without further specifying what $x$ is or making assumptions about it beyond that it is an element of $A$, what we have shown is that the statement $(\forall x \in A)[P(x)]$ is true.

II.1.7.8. What is technically being done here is that quantification of a variable is stylistically moved to the beginning of a section of a discussion in which it is used repeatedly, which we call separated quantification. Separated quantification can be eliminated, but the price of elimination in many cases is a horribly complex compound quantified statement which is difficult to understand. So separated quantification is very important to use to make mathematical discussions readable and understandable.

II.1.7.9. Sometimes “definitions” of this sort are made implicitly. For example, if $(x_k)$ is known to be a bounded sequence, one can legitimately say:

“Since $(x_k)$ is bounded, there is an $M$ such that $x_k \leq M$ for all $k$.”

This implicitly defines $M$ to be a (variable) number satisfying the inequality for all $k$.

It is good mathematical form to phrase things in this way. A bad form mistake beginners often make is to write something like
“Since \((x_k)\) is bounded, \(x_k \leq M\) for all \(k\).”

where the \(M\) is a variable which has not yet been specified, i.e. the variable name \(M\) has not yet appeared in the discussion. (It is a more serious mistake if the name \(M\) has already been used to mean something else.) A reader seeing a statement like this must assume that the \(M\) refers to something which has already been defined, and wonders “What is \(M\)?” Some time and effort then must be wasted trying to find or figure out what the \(M\) refers to, and if it is concluded that the statement is intended to be an implicit “definition” of \(M\), the conclusion is only an educated guess, so there is a lack of precision and rigor.

II.1.7.10. There are subtle and potentially confusing instances where it is unclear whether an implicit separated definition is being made, for example in the common calculus statement

\[
\int x \, dx = \frac{x^2}{2} + C.
\]

Is the \(C\) now a quantified variable? We often write

\[
\int x \, dx = \frac{x^2}{2} + C, \quad \int x^2 \, dx = \frac{x^3}{3} + C.
\]

Are the \(C\)'s supposed to be the same? If we have these two statements, what about

\[
\int (x + x^2) \, dx?
\]

Is this \(\frac{x^2}{2} + \frac{x^3}{3} + C\), or \(\frac{x^2}{2} + \frac{x^3}{3} + 2C\)? It is standard to write the first, since by convention the \(C\) in each statement is not a specific constant, but simply denotes that any constant can be added. This convention, although universal, is a little unfortunate. Another instance of violation of usual ambiguity rules of mathematical language is in the big-\(O\) and little-\(o\) notation ().

II.1.8. Languages and Relational Systems

We should really have begun the entire discussion of formal statements by first specifying a language and a domain of discourse. We did not do this at first because we didn’t want to make the discussion too abstract. But there is merit in considering systems of formal statements based on various languages, or, from a different point of view, systems where the language and statements are restricted to be of a certain form.

II.1.8.1. Definition. A relational system is a triple \(S = (\Omega, \mathcal{R}, \mathcal{F})\), where \(\Omega\) is a set, \(\mathcal{R}\) is a set of relations on \(\Omega\), and \(\mathcal{F}\) a set of functions on \(\Omega\) (with codomain \(\Omega\)). The relations and functions may be \(n\)-ary for any \(n \geq 1\), which may vary from relation to relation, and may have domains which are proper subsets. The set \(\Omega\) is called the domain of discourse of the relational system.

Every function is just a special kind of relation, so it seems unnecessary to explicitly also include a set of functions in the definition; but the relations and functions are used in very different ways, so it is necessary to specify both \(\mathcal{R}\) and \(\mathcal{F}\) as part of the definition. Note that \(\mathcal{R}\) and/or \(\mathcal{F}\) need not contain all relations or functions on \(\Omega\).

The discussion of the previous part of this section concerned the relational system where \(\Omega\) was some universal set containing all objects under discussion (in many cases \(\Omega\) could be taken to be the real numbers \(\mathbb{R}\)), and \(\mathcal{R}\) and \(\mathcal{F}\) the sets of all relations and all functions on \(\Omega\).
II.1.8.2. Definition. Let $\mathcal{S}$ be a relational system. A language for the relational system $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$ is:

(i) A set of symbols

(ii) A set of terms

(iii) A set of formulas

where the set of symbols includes at least one constant symbol for each element of $\Omega$ and each relation in $\mathcal{R}$ and function in $\mathcal{F}$, called its name (an element of $\Omega$ can have more than one name); a set of variables; a set of logical symbols including $\land$, $\lor$, and $\forall$; and parentheses. The terms satisfy

1. Each constant or variable symbol is a term.

2. If $f \in \mathcal{F}$ is a function of $n$ arguments, and $t_1, \ldots, t_n$ are terms, then $f(t_1, \ldots, t_n)$ is a term.

A variable $x$ occurs in the term $f(t_1, \ldots, t_n)$ if it equals or occurs in one of the $t_k$.

The formulas are certain finite strings of symbols including all strings of the form

(a) $R(t_1, \ldots, t_n)$, where $R$ is an $n$-ary relation in $\mathcal{R}$ and each $t_k$ is a term involving only constants which are names of elements of $\Omega$ and variables.

(b) $\forall x(P)$, where $P$ is a formula and $x$ is a variable not quantified in $P$. The variable $x$ is then quantified in $\forall x(P)$.

(c) $P_1 \land \cdots \land P_n$, where $P_1, \ldots, P_n$ are formulas, and if $x$ is a variable which is quantified in some $P_k$, then $x$ does not appear unquantified in any other $P_j$.

(d) $P_1 \Rightarrow P_2$, where $P_1$ and $P_2$ are formulas, and if $x$ is a variable which is quantified in one $P_k$, then $x$ does not appear unquantified in the other $P_j$.

A statement, or closed statement for emphasis, is a formula in which all variables are quantified, i.e. in which there are no free variables (variables which occur and are not quantified).

If every term in the language is obtained by a finite number of applications of (1) and (2), and every formula built up from a finite number of applications of (a), (b), (c), and (d), the language is a simple language for $\mathcal{S}$. Each term and each formula then has a well-defined level, and is built up from terms or formulas of strictly lower level.

In some references the word “sentence” is used in place of “formula” and in some other references “sentence” is used in place of “statement.” Because of the nonuniformity of usage, we have avoided the word “sentence” entirely. In the earlier part of this section, the word “statement” was used a little informally; in many of the places it was used we should have instead said “formula.”

II.1.8.3. A language can be a language for more than one relational system. But it is very hard for a language to be a simple language for more than one relational system, although it can be if the names of elements and relations are chosen in a different way. Conversely, a relational system can have many associated languages, but a simple language for a relational system is essentially uniquely determined by the relational system.
II.1.8.4. Suppose $L$ is a language for a relational system $S = (\Omega, \mathcal{R}, \mathcal{F})$, which we will take to be a simple language. It makes sense to say whether or not closed statements in $L$ (and only closed statements!) are “true” in $S$:

II.1.8.5. **Definition.** Suppose $L$ is a simple language for a relational system $S = (\Omega, \mathcal{R}, \mathcal{F})$. If $t$ is a term in $L$, then

(i) If $t$ is a constant $c$ which is the name of an element $\omega \in \Omega$, then $t$ is **interpretable in** $S$ as $\omega$.

(ii) If $t = f(t_1, \ldots, t_n)$ and $t_1, \ldots, t_n$ are interpretable in $S$ as $\omega_1, \ldots, \omega_n$ respectively, and $(\omega_1, \ldots, \omega_n)$ is in the domain of $f$, then $t$ is **interpretable in** $S$ as $f(\omega_1, \ldots, \omega_n)$.

If $t$ is interpretable in $S$ via a finite string of applications of (i) and (ii), then $t$ is interpretable in $S$; otherwise $t$ is not interpretable in $S$.

II.1.8.6. **Definition.** Suppose $L$ is a simple language for a relational system $S = (\Omega, \mathcal{R}, \mathcal{F})$. Then

(i) If $P$ is the statement $R(t_1, \ldots, t_n)$ involving no variables, and $t_1, \ldots, t_n$ are interpretable in $S$ as $\omega_1, \ldots, \omega_n$ respectively, and $(\omega_1, \ldots, \omega_n) \in R$ (i.e. $R(\omega_1, \ldots, \omega_n)$ holds in $\Omega$), then $P$ holds (or is true) in $S$.

(ii) If $P$ is the formula $R(t_1, \ldots, t_n)$, with all variables occurring in $t_1, \ldots, t_n$ included among $x_1, \ldots, x_m$, $c_1, \ldots, c_m$ are names of elements of $\Omega$, and $P$ holds when, for each $k$, $c_k$ is substituted for each instance of $x_k$ in $P$, then $P$ holds (or is true) for $(c_1, \ldots, c_m)$ in $S$.

(iii) If $P$ is a formula containing only unquantified variables among $x_1, \ldots, x_m, x_{m+1}, \ldots, x_p$ and quantified variables among $x_{p+1}, \ldots, x_r$, $c_1, \ldots, c_m$ are names of elements of $\Omega$, and for every set of constants $c_{m+1}, \ldots, c_r$ in $L$ which are names of elements of $\Omega$, $P$ holds for $(c_1, \ldots, c_m)$, then

\[
\forall x_{m+1} \cdots \forall x_p(P)
\]

holds (or is true) for $(c_1, \ldots, c_m)$ in $S$. If $m = 0$, then the statement

\[
\forall x_1 \cdots \forall x_p(P)
\]

holds (or is true) in $S$.

(iv) If $P_1, \ldots, P_n$ are formulas containing only unquantified variables among $x_1, \ldots, x_m$, with no variable quantified in some $P_k$ and free in another $P_j$, $c_1, \ldots, c_m$ are names of elements of $\Omega$, and each sentence $P_j$ holds for $(c_1, \ldots, c_m)$, then

\[
P_1 \land \cdots \land P_n
\]

holds (or is true) for $(c_1, \ldots, c_m)$ in $S$. If $m = 0$, i.e. if $P_1, \ldots, P_n$ are statements, then the statement

\[
P_1 \land \cdots \land P_n
\]

holds (or is true) in $S$. 

119
(v) If $P_1$ and $P_2$ are formulas containing only the free variables $x_1, \ldots, x_m$, none of which are quantified in one and free in the other, $c_1, \ldots, c_m$ are names of elements of $\Omega$ for which $P_2$ holds in $S$ or $P_1$ does not hold in $S$, then

$$P_1 \Rightarrow P_2$$

holds (or is true) for $(c_1, \ldots, c_m)$ in $S$. If $m = 0$, i.e. if $P_1$ and $P_2$ are statements, then the statement

$$P_1 \Rightarrow P_2$$

holds (or is true) in $S$.

If $P$ is a statement in $L$, and $P$ is built up from a finite chain of formulas which hold in $S$, then $P$ holds in $S$; otherwise $P$ does not hold (or is false) in $S$.

II.1.8.7. There are two quite distinct ways a statement can fail to hold:

(1) At some point, a term is not interpretable.

(2) All terms are always interpretable, but a relation does not hold.

II.1.8.8. Unless the set $R$ of relations is empty (in which case $L$ has no formulas) there are statements which hold in $S$: if $P$ is any formula, and $x_1, \ldots, x_m$ are the unquantified variables in $P$, then the statement

$$\forall x_1 \cdots \forall x_m (P \Rightarrow P)$$

holds in $S$. It can happen that every statement in $L$ holds in $S$, but only if $R = \emptyset$, or if $R$ consists only of the full relation and, for every $n$, every $n$-ary function in $F$ has full domain $\Omega^n$. Such a relational system is called degenerate. In a nondegenerate relational system there is a statement $\Phi = R(t_1, \ldots, t_n)$ containing no variables which is false (and there are also statements which are true, e.g. $\Phi \Rightarrow \Phi$).

II.1.8.9. If we fix such a $\Phi$, then for any formula $P$ containing free variables among $x_1, \ldots, x_m$, and every set $c_1, \ldots, c_m$ of names of elements of $\Omega$, the formula $P \Rightarrow \Phi$ holds for $(c_1, \ldots, c_m)$ in $S$ if and only if $P$ does not hold for $(c_1, \ldots, c_m)$ in $S$. We can thus interpret this formula as $\neg P$, i.e. we can extend $L$ by adding the symbol $\neg$, and formulas $\neg (P)$, where the formula $\neg (P)$ in the extended language is translated to mean $P \Rightarrow \Phi$ in the simple language.

II.1.8.10. We can then further extend $L$ by adding the symbols $\forall$ and $\exists$, where

$$P_1 \lor \cdots \lor P_n$$

is translated to mean

$$\neg (\neg (P_1) \land \cdots \land \neg (P_n))$$

(this will be a formula of the extended language if no variable is quantified in one of the $P_k$ and unquantified in another $P_j$), and

$$\exists x_1 \cdots \exists x_m (P)$$

is translated to mean

$$\neg (\forall x_1 \cdots \forall x_m (\neg (P)))$$

(this will be a formula of the extended language if no $x_k$ is quantified in $P$). More complicated formulas of the extended language can be built up from repeated applications of these translations.
II.1.8.11. We may thus extend a simple language $\mathcal{L}$ for a nondegenerate relational structure $\mathcal{S}$ to an extended simple language $\mathcal{L}'$, which includes all the standard logical symbols with their usual meanings, for which every formula has a well-defined translation into a formula of $\mathcal{L}$, and it makes sense to say whether statements in $\mathcal{L}'$ are true or false in $\mathcal{S}$.

II.1.8.12. In the extended simple language, the usual rules for taking negations hold. We can define an additional logical symbol $\iff$ by translating $P \iff Q$ as $(P \implies Q) \land (Q \implies P)$, and we say $P$ and $Q$ are equivalent in $\mathcal{S}$ if $P \iff Q$ holds in $\mathcal{S}$. Then

$$\neg(\neg(P)) \iff P$$

for every statement $P$. A statement form like this which holds in $\mathcal{S}$ for every statement $P$ is called a tautology in $\mathcal{S}$. The other tautologies in II.1.6.1. are also tautologies in $\mathcal{S}$.

II.1.9. Mathematical Arguments and Formal Proofs

II.1.9.1. Informally, a mathematical argument consists of a derivation of a mathematical statement (conclusion) from a set of initial assumptions, using acceptable mathematical reasoning. The argument, if logically sound, constitutes a proof of the conclusion.

II.1.9.2. The initial assumptions can be divided into two kinds: general assumptions underlying an entire mathematical theory, called axioms, and assumptions particular to the individual statement being proved, called hypotheses. There is not always a clear dividing line between axioms and hypotheses, and anyway the distinction between them is unimportant for the purpose of making and understanding mathematical arguments.

II.1.9.3. In practice, mathematical arguments are usually written out in a combination of English (or another language) and formal symbolism.

A formal mathematical argument

II.1.10. Exercises

II.1.10.1. Let $P(x)$ and $Q(x)$ be statements with a free variable $x$.
(a) Show that the statement $\forall x(P(x) \land Q(x))$ is logically equivalent to $[\forall x(P(x))] \land [\forall x(Q(x))]$.
(b) Show that the statement $\forall x(P(x) \lor Q(x))$ is not logically equivalent to $[\forall x(P(x))] \lor [\forall x(Q(x))]$.
(c) Show that the statement $\exists x(P(x) \lor Q(x))$ is logically equivalent to $[\exists x(P(x))] \lor [\exists x(Q(x))]$.
(d) Show that the statement $\exists x(P(x) \land Q(x))$ is not logically equivalent to $[\exists x(P(x))] \land [\exists x(Q(x))]$.
(e) Which implications are valid in (b) and (d)?

II.1.10.2. (a) Show that $P \implies P$ and $P \implies (Q \implies Q)$ are tautologies.
(b) Are $(P \implies P) \implies Q$ and $P \implies (P \implies Q)$ tautologies?
(c) Is $\implies$ “associative,” i.e. is

$$[(P \implies Q) \implies R] \iff [P \implies (Q \implies R)]$$

a tautology?
II.1.10.3. If $S$ is nondegenerate, a statement $\Phi$ is false if and only if $\Phi \Rightarrow P$ holds for every statement $P$. Show that any two false statements are equivalent.

II.1.10.4. Give a careful justification for all the assertions in II.1.8.11. and II.1.8.12.

II.1.10.5. [Bar00] Is

$$(P \Rightarrow Q) \lor (Q \Rightarrow P)$$

a tautology? Does this mean that for any statements $P$ and $Q$, either $P$ implies $Q$ or $Q$ implies $P$? In particular, does it mean “If a number is prime, then it is odd, or if a number is odd, then it is prime”?
Theorems, Corollaries, Propositions, Lemmas

The terms *Theorem*, *Corollary*, *Proposition*, and *Lemma* all mean essentially the same thing, with nuances of difference. All refer to mathematical statements which must be, and are, proved.

The terms are largely interchangeable, and the same statements can be found in various books called by any of the names. But there are conventions which are generally followed, although they are subjective and subject to interpretation. So the following descriptions should be taken with a grain of salt, and are intended only as a guide:

II.1.10.6. A *Theorem* is a result which (i) is of major significance and/or (ii) is difficult to prove. Many have widely-used official names, e.g. “The Fundamental Theorem of Calculus.” It is good practice to use the term “Theorem” sparingly, to really call attention to the big results.

II.1.10.7. A *Corollary* is a result which follows immediately or easily from another result. Sometimes a Corollary of a Theorem is simply a crucial special case of the theorem. A result which is a combination of two or more Theorems can also be called a Corollary.

II.1.10.8. A *Proposition* is a statement which (i) is less significant than a Theorem and/or (ii) has an easy or obvious proof. Propositions are not necessarily insignificant: fundamental immediate consequences of definitions and other observations are often called Propositions, with the proof often omitted if it is thought obvious.

II.1.10.9. A *Lemma* is a statement which is not of particular interest in and of itself, but which is of use in proving something else of greater interest. Statements of lemmas are often rather technical, and proofs can be complicated (although neither is a necessary attribute). Lemmas are roughly used in two ways. It is sometimes convenient to break a long proof involving many steps into a sequence of Lemmas to make the overall argument easier to follow and understand. And sometimes a subsidiary result which must be used repeatedly in the main argument can be conveniently isolated as a Lemma.

The best description of a Lemma I have heard came when I was a student in a course on Automata Theory. It was a mathematics course, with theorems and proofs, but many of the students were graduate students in Computer Science, with a different outlook than mathematicians. One day the instructor wrote a Lemma on the board, and one of the students asked, “What’s a Lemma?” Another CS student immediately replied, “It’s a subroutine.”

There are statements commonly called Lemmas which are significant Theorems, e.g. “Urysohn’s Lemma” () (which is probably called a “Lemma” simply to distinguish it from “Urysohn’s Metrization Theorem”, although it is a crucial tool in the proof of the Metrization Theorem).

II.1.10.10. The terms are occasionally used in a somewhat different sense. For example, the name “Fermat’s Last Theorem” has been used for nearly two hundred years, even though it only became a theorem (or, perhaps more accurately, became known to be a theorem, i.e. was proved) quite recently. And “Zorn’s Lemma” () is not really a Lemma, but a useful equivalent restatement of the Axiom of Choice ()

II.1.10.11. Certain theorem-level results have common names using other terms, such as the “Product Rule,” “l’Hôpital’s Rule,” the “Sequential Criterion for Continuity,” the “Ratio Test,” the “Cauchy Integral Formula,” the “Principle of the Argument,” etc.
II.2. Set Theory

II.2.1. New Sets from Old

The Power Set of a Set

Ordered Pairs and Cartesian Products

The Union of a Set

The Constituent Set of a Set

II.2.1.1. Definition. Let $X$ be a set. Inductively set

$$C_1(X) = \bigcup X = \{y : \exists x \in X \text{ with } y \in x\}$$

$$C_{n+1}(X) = C_1(C_n(X))$$

$$C(X) = \bigcup_{n=1}^{\infty} C_n(X).$$

The set $C(X)$ is called the constituent set of $X$.

Intuitively, $X \cup C(X)$ is the collection of all things $X$ is “made up of.” The constituents of $X$, i.e. the elements of $C(X)$, are the things that the elements of $X$ are made of. (It is a matter of convention not to assume that $X \subseteq C(X)$, i.e. to define $C(X)$ as we have instead of as $X \cup (\bigcup_{n=1}^{\infty} C_n(X))$, but our definition turns out to be more notationally useful.)

II.2.1.2. $C(X)$ is a set by the Axiom of Union and the Axiom of Infinity. However, the cardinality of $C(X)$ is quite independent of $\text{card}(X)$, as the following examples show (but see II.2.1.6.).

II.2.1.3. Examples. (i) $C_1(\emptyset) = C(\emptyset) = \emptyset$. $C_1(\{\emptyset\}) = C(\emptyset) = \emptyset$. $C_1(\{\{\emptyset\}\}) = C(\emptyset) = \emptyset$. $C_1(\emptyset) = C(\emptyset) = \emptyset$. $C_1(\{\emptyset\}) = C(\emptyset) = \emptyset$. $C_1(\emptyset) = C(\emptyset) = \emptyset$.

(ii) If $\sigma$ is a successor ordinal (II.8.2.8.), then $C_1(\sigma) = C(\sigma) = \sigma - 1$. If $\sigma$ is a limit ordinal, then $C_1(\sigma) = C(\sigma) = \sigma$.

(iii) If $x$ is a set of cardinality $\kappa$, and $X = \{x\}$, then $\text{card}(X) = 1$ but $\text{card}(C_1(X)) = \kappa$ since $C_1(X) = x$. The cardinality of $C(X)$ may be even larger.

(iv) If the elements of $X$ are not sets (i.e. are urelements ($\emptyset$)), then $C(X) = \emptyset$ no matter how large $\text{card}(X)$ is.

II.2.1.4. If $x \in X$, then $C_n(x) \subseteq C_{n+1}(X)$ for each $n$, and thus $C(x) \subseteq C(X)$. More generally, if $x \in C(X)$, then $C(x) \subseteq C(X)$. If $Y \subseteq X$, then $C_n(Y) \subseteq C_n(X)$ for every $n$, so $C(Y) \subseteq C(X)$. Similarly, if $Y \subseteq C(X)$, then $C(Y) \subseteq C(X)$. In particular, we have $C(C(X)) \subseteq C(X)$, but they need not be equal (II.2.1.3.(i)).

II.2.1.5. Let $X$ be a set. Since $C_1(\mathcal{P}(X)) = X$, $C(\mathcal{P}(X)) = X \cup C(X)$.

II.2.1.6. If $X$ contains no urelements, then $X \subseteq \mathcal{P}(C_1(X))$, and hence

$$\text{card}(X) \leq 2^{\text{card}(C_1(X))} \leq 2^{\text{card}(C(X))}$$

so $\text{card}(C(X))$ cannot be “too much smaller” than $\text{card}(X)$.
II.3. Relations and Functions

The idea of a relation is a crucial one, both within and outside of mathematics. Mathematicians have formalized the intuitive idea of a relation, and special kinds of relations abound throughout mathematics. The three most important kinds of relations in mathematics are functions, equivalence relations, and partial orderings. The first two will be discussed in this section, and partial orderings described in the next section.

II.3.1. General Relations

To formalize the notion of a relation between elements of a set $X$ and a (possibly different) set $Y$, we can think of the relation as consisting of certain ordered pairs $(x; y)$. Thus the formal definition is:

II.3.1.1. Definition. A relation from a set $X$ to a set $Y$ is a subset $R$ of $X \times Y$. If $(x; y) \in X \times Y$, write $x R y$ if $(x; y) \in R$. If $X = Y$, $R$ is called a relation on $X$.

The set $\{x \in X : x R y \text{ for some } y \in Y\}$ is called the domain of $R$, denoted $\text{Dom}(R)$; and the set $\{y \in Y : x R y \text{ for some } x \in X\}$ is called the range of $R$, denoted $\text{Range}(R)$.

II.3.1.2. Examples.

(i) The relation $\leq$ can be regarded as a relation on $\mathbb{R}$: it is technically the set $\{(x, y) \in \mathbb{R}^2 : x \leq y\} \subseteq \mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

(ii) Let $S$ be the set of students at Ivory State University, and $C$ the set of classes offered this semester. Define a relation $E$ from $S$ to $C$ by letting $s E c$ if student $s$ is enrolled in class $c$. (There are obviously an enormous variety of variations on this example.) The domain is probably not quite all of $S$, since there may be students not taking any classes this semester; the range may or may not be all of $C$, depending on whether classes with zero enrollment are counted.

(iii) Let $X$ and $Y$ be any sets. Then $\emptyset$ and $X \times Y$ are relations from $X$ to $Y$, called the empty relation and full relation respectively.

(iv) If $X$ is a set, the identity relation on $X$ is the relation $I$ (often written $I_X$) defined by $x I y$ if and only if $x = y$. The domain and range of $I$ are all of $X$.

II.3.1.3. Definition. Let $R$ and $S$ be relations from $X$ to $Y$. $S$ is an extension of $R$ if $R \subseteq S$, i.e. $x R y$ implies $x S y$.

If $A \subseteq X$, the restriction of $R$ to $A$ is

$$R|_A = R \cap (A \times Y) = \{(x, y) : x \in A, \ y \in Y, \ x R y\}.$$

Note that $R$ is an extension of $R|_A$. However, not every extension comes from a restriction in this way.
II.3.1.4. **Definition.** Let $R$ be a relation from $X$ to $Y$. The *inverse relation* $R^{-1}$ is the relation from $Y$ to $X$ defined by

\[ R^{-1} = \{(y, x) : (x, y) \in R\} \]

(in other words, $yR^{-1}x$ if and only if $xRy$.) The domain of $R^{-1}$ is the range of $R$, and vice versa; and $(R^{-1})^{-1} = R$.

As a relation, $R^{-1}$ appears to be only superficially or notationally different from $R$; however, it can have rather different properties, as we will see later.

**Composition**

Relations can be composed:

II.3.1.5. **Definition.** Let $R$ be a relation from a set $X$ to a set $Y$, and $S$ a relation from $Y$ to a set $Z$. The *composition* of $S$ and $R$, written $S \circ R$, is the relation from $X$ to $Z$ defined by

\[ S \circ R = \{(x, z) \in X \times Z : xRy \text{ and } ySz \text{ for some } y \in Y\} \]

(Note the order in which $R$ and $S$ are written.)

II.3.1.6. **Examples.**

(i) Let $E$ be the relation from $S$ to $C$ defined in II.3.1.2.(ii), and let $L$ be the relation from $C$ to the set $B$ of buildings at Ivory State with $cLb$ if class $c$ meets at least once a week in building $b$. Then $L \circ E$ is the relation $R$ from $S$ to $B$ with $sRb$ if and only if student $s$ has at least one class a week in building $b$.

(ii) If $I$ is the identity relation on a set $X$, and $R$ is any relation from $X$ to $Y$, then $R \circ I = R$.

(iii) If $F$ is the full relation from $X$ to $Y$, and $G$ the full relation from $Y$ to $Z$, then $G \circ F$ is the full relation from $X$ to $Z$ (provided $Y \neq \emptyset$).

(iv) If $<$ is the usual strict inequality relation on $\mathbb{R}$, then $(< \circ <) = <$. However, the relation $< \circ <$ on $\mathbb{N}$ is not the same as $<$ on $\mathbb{N}$: $m(< \circ <)n$ if and only if $n - m \geq 2$. Thus the sets on which a relation is defined are very important to specify carefully.

(v) The composition of $<$ and $>$ on $\mathbb{R}$ is the full relation on $\mathbb{R}$.

(vi) Let $P$ be the set of all people. Define relations $B$ and $S$ on $P$ by saying $xBy$ if $x$ is the brother of $y$, and $xSy$ if $x$ is the sister of $y$. Then (disregarding modern surgical procedures) $B \circ (S^{-1})$ is the empty relation on $P$.

II.3.1.7. The domain of $S \circ R$ is a subset of the domain of $R$, and the range is a subset of the range of $S$. Either or both containments can be proper. See II.3.4.1.
Images and Preimages

II.3.1.8. Definition. Let $R$ be a relation from $X$ to $Y$, and let $A \subseteq X$. Then the image of $A$ under $R$ is

$$R(A) = \{ y \in Y : xRy \text{ for some } x \in A \} .$$

If $B \subseteq Y$, the preimage (or inverse image) of $B$ under $R$ is

$$R^{-1}(B) = \{ x \in X : xRy \text{ for some } y \in B \} .$$

II.3.1.9. The notation is consistent: $R^{-1}(B)$ is the image of $B$ under the relation $R^{-1}$. Thus images and preimages can be considered on the same basis. We prefer the term “preimage” to “inverse image,” especially in the case of functions, where “inverse image” can mistakenly seem to imply the existence of an inverse function.

Images (and hence also preimages) behave nicely with respect to unions, but not quite so nicely under intersections:

II.3.1.10. Proposition. Let $R$ be a relation from $X$ to $Y$, and $\{A_i : i \in I\}$ an indexed collection of subsets of $X$. Then

(i) $R(\bigsqcup_i A_i) = \bigsqcup_i R(A_i)$.

(ii) $R(\bigcap_i A_i) \subseteq \bigcap_i R(A_i)$.

The proof is left as an exercise (Exercise {}).

II.3.1.11. Simple examples show that equality does not hold in (ii) in general. For example, if $F$ is the full relation from $X$ to $Y$, and $A$ is any nonempty subset of $X$, then $F(A) = Y$. If $A_1$ and $A_2$ are nonempty but $A_1 \cap A_2 = \emptyset$, then

$$\emptyset = F(\emptyset) = F(A_1 \cap A_2) \neq F(A_1) \cap F(A_2) = Y .$$

II.3.2. Functions

Functions are officially relations of a special kind, although they are not usually thought of in this way.

II.3.2.1. Definition. Let $X$ and $Y$ be sets. A function from $X$ to $Y$ is a relation $f$ in which no two ordered pairs have the same first coordinate, i.e., if $x \in X$ and $xfy_1$ and $xfy_2$ for some $y_1, y_2 \in Y$, then $y_1 = y_2$.

Thus, if $f$ is a function, then for every $x \in \text{Dom}(f)$ there is a unique $y \in Y$ with $xfy$. We normally write $f(x) = y$ instead of $xfy$, and call $y$ the value of $f$ at $x$ or the image of $x$ under $f$.

The vast majority of functions considered in analysis are ones taking numerical values, but there is no reason to restrict attention only to numerical functions.
II.3.2.2. Examples.

(i) If $X$ and $Y$ are sets, the empty relation from $X$ to $Y$ and the identity relation on $X$ are functions. The identity relation on $X$, when regarded as a function, is written $\iota_X$. The full relation from $X$ to $Y$ is not a function unless $Y$ is a singleton (or $\emptyset$).

(ii) Let $X$ and $Y$ be sets, and $y_0$ a fixed element of $Y$. Set $f(x) = y_0$ for all $x \in X$. Then $f$ is a function, called the constant function $y_0$. By slight abuse of notation, we will often use the symbol $y_0$ as the name of this constant function, especially if $Y$ is a set of numbers. Thus, for example, we will often use a phrase like “the constant function 5”, meaning the function with $f(x) = 5$ for all $x$.

(iii) The relations $E$ of II.3.1.2. (ii) and $B$ and $S$ of II.3.1.6. (vi) are not functions since a student cannot take more than one class and a person can have more than one brother or sister. The relation $L$ of II.3.1.6. (i) is also probably not a function, since a class can meet in different buildings on different days (for example, it could have a lab or discussion session in another building.)

(iv) The set $\{(x, y) \in \mathbb{R}^2 : y = x^2\}$ is a typical function from $\mathbb{R}$ to $\mathbb{R}$ defined by a formula.

II.3.2.3. Terminology: It is especially important with functions to carefully specify the domain; thus we should always specify something like “$f$ is a function from $X$ to $Y$ with domain $A$.” Without further qualification, the phrase “$f$ is a function from $X$ to $Y$” will be taken to mean that $\text{Dom}(f)$ is all of $X$. In this case, we often write $f : X \to Y$. When $f$ is defined by a formula or rule, without explicit specification to the contrary the domain will always be taken to be all elements of $X$ for which the formula or rule makes sense. Thus, when we write $f : X \to Y$, the domain $X$ is determined by $f$. The set $Y$ is called the codomain of $f$ when written in this way. It is a subtle matter whether the $Y$ should be regarded as determined by $f$, since it is not the range of $f$ in general. See II.3.2.12.

Mappings

Functions are usually thought of as mappings:

II.3.2.4. Definition. A mapping $f$ from a set $X$ to a set $Y$ is an assignment to each $x \in X$ of a unique element $f(x)$ of $Y$. A mapping $f$ is often represented pictorially:

II.3.2.5. This definition of a mapping is not quite precise enough mathematically, hence the formal definition of function given above. But there is a natural correspondence: a function can obviously be pictured as a mapping, and a mapping $f$ from $X$ to $Y$ gives a function

$$\{(x, y) \in X \times Y : f(x) = y\}.$$  

This set is often called the graph of the mapping $f$; formally, a function is its own graph. Thinking of functions and graphs as the same thing can be confusing, but is technically useful.
II.3.2.6. We are in the habit of defining functions, especially numerical functions, by algebraic formulas; we typically write something like “let \( f(x) = \frac{\cos x}{x-2} \)” to define a function, where we really mean

\[
f = \left\{ (x, y) \in \mathbb{R}^2 : y = \frac{\cos x}{x-2} \right\}.
\]

By our convention, the domain of this function would be \( \{x \in \mathbb{R} : x \neq 2\} \).

Images and Preimages

Since functions are relations, the definitions of image and preimage (II.3.1.8.) make sense. We rephrase them in function notation:

II.3.2.7. Definition. Let \( f : X \to Y \) be a function.

(i) If \( A \subseteq X \), the image of \( A \) under \( f \) is

\[
f(A) = \{ f(x) : x \in A \} \subseteq Y.
\]

(ii) If \( B \subseteq Y \), the preimage (or inverse image) of \( B \) under \( f \) is

\[
f^{-1}(B) = \{ x \in X : f(x) \in B \} \subseteq X.
\]

II.3.2.8. We prefer the term “preimage” to “inverse image,” since there is not necessarily an inverse function \( f^{-1} : Y \to X \) (there is an inverse relation, which is not a function in general). The notation \( f^{-1}(B) \) can also be confusing, but is standard; note that this notation does not mean there is an inverse function! Preimages are defined for arbitrary functions and arbitrary subsets of the codomain. In the special case that there is an inverse function, the preimage of a set is indeed just the image under the inverse function, so the notation is consistent.

II.3.2.9. If \( f : X \to Y \) and \( a \in X \), there is a subtle difference between \( f(a) \) and \( f(\{a\}) \): \( f(a) \) is an element of \( Y \), while \( f(\{a\}) \) is a subset of \( Y \). In fact, \( f(\{a\}) = \{f(a)\} \). The distinction is more pronounced for preimages: if \( b \in Y \), then \( f^{-1}(\{b\}) \) is a subset of \( X \) which can be a singleton, but can also be empty (if \( b \) is not in the range of \( f \)) or it can have many elements. The notation \( f^{-1}(b) \) does not make sense unless \( f \) has an inverse function \( f^{-1} \).

II.3.2.10. If \( f : X \to Y \) is a function, then \( f(\emptyset) = \emptyset \), and if \( A \subseteq X \) and \( f(A) = \emptyset \), then \( A = \emptyset \). Preimages are more complicated: \( f^{-1}(\emptyset) = \emptyset \), but if \( B \subseteq Y \), then \( f^{-1}(B) = \emptyset \) does not imply that \( B = \emptyset \) – it only means the intersection of \( B \) with the range of \( f \) is empty.

We have \( f^{-1}(Y) = X \), but not necessarily \( f(X) = Y \): \( f(X) \) is the range of \( f \). If \( A \subseteq X \) and \( B \subseteq Y \), we have \( A \subseteq f^{-1}(f(A)) \) and \( f(f^{-1}(B)) \subseteq B \), but we do not have equality in general.

Preimages respect unions, intersections, and complements, but images only respect unions in general (cf. II.3.4.3.):
II.3.2.11.  **Proposition.** Let $f : X \to Y$ be a function.

(i) If $\{A_i : i \in I\}$ is a collection of subsets of $X$, then $f(\bigcup_{i \in I} A_i) = \bigcup_{i \in I} f(A_i)$ and $f(\bigcap_{i \in I} A_i) \subseteq \bigcap_{i \in I} f(A_i)$.

(ii) If $\{B_i : i \in I\}$ is a collection of subsets of $Y$, then $f^{-1}(\bigcup_{i \in I} B_i) = \bigcup_{i \in I} f^{-1}(B_i)$ and $f^{-1}(\bigcap_{i \in I} B_i) = \bigcap_{i \in I} f^{-1}(B_i)$.

(iii) If $B \subseteq Y$, then $f^{-1}(B^c) = [f^{-1}(B)]^c$ (i.e. $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$).

**Proof:** We prove as a sample the second statement of (ii). The other proofs are very similar and are left as an exercise. Note that in all cases except the second statement of (i), two containments must be shown.

Let $x \in f^{-1}(\bigcap_{i \in I} B_i)$. Then $f(x) \in \bigcap_{i \in I} B_i$, so for each $i \in I$, $f(x) \in B_i$ and so $x \in f^{-1}(B_i)$. Thus $x \in \bigcap_{i \in I} f^{-1}(B_i)$.

For the reverse containment, let $x \in \bigcap_{i \in I} f^{-1}(B_i)$. Then, for each $i \in I$, $x \in f^{-1}(B_i)$, so $f(x) \in B_i$. Thus $f(x) \in \bigcap_{i \in I} B_i$, so $x \in f^{-1}(\bigcap_{i \in I} B_i)$.

---

**Injective and Surjective Functions**

**Is the Codomain Uniquely Determined?**

II.3.2.12.  There is a subtle problem with codomains of functions which is overlooked in many texts, which sometimes even make contradictory statements as a result. If we have a function $f : X \to Y$, by our convention (II.3.2.3.) $X$ is the domain of $f$ and is thus uniquely determined by $f$. The $Y$ is called the codomain of $f$, but it is only a set containing the range of $f$ in general. Should we regard the codomain as being intrinsic to $f$, i.e. uniquely specified by $f$? This is a trickier question than would first appear: no matter which convention we take, it will be impossible to cleanly answer both of the following questions:

(i) When are two functions equal?

(ii) When is a function surjective?

II.3.2.13.  It is stated in many books that two functions $f : X \to Y$ and $g : Z \to W$ are equal if and only if $X = Z$, $Y = W$, and $f(x) = g(x)$ for all $x \in X = Z$. But suppose $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$, and $g : \mathbb{R} \to [0, +\infty)$ is defined by $g(x) = x^2$. Are $f$ and $g$ equal functions, or are they different functions? (And does “equal functions” mean “the same function”?)

On the other hand, if $f$ and $g$ are the same function, does it really make sense to say that $g$ is surjective and $f$ is not surjective? How can a function be both surjective and not surjective?

II.3.2.14.  If a function is defined to be a set of ordered pairs, as is the usual definition in set theory, then two functions with the same domain taking the same values at all points of the domain, such as the $f$ and $g$, consist of exactly the same ordered pairs, and then must be equal as sets by the Extension Axiom of set theory, whether or not the specified codomains are the same.

One can proceed differently: in Category Theory, every function (morphism) has a specified domain and codomain which are part of the definition of a morphism. Thus from a category theory point of view the above $f$ and $g$ are different. This point of view gives a clean meaning to surjectivity (technically called an *epimorphism*, although the abstract category definition must be done differently, cf. ())).
II.3.2.15. Since we have used the ordered pair definition of function, we ought to, and will, adopt the
convention that the codomain of a function is not intrinsic to the function, and thus two functions are equal
if they have the same domain and take the same values at all points of the domain; thus we will regard the
above \( f \) and \( g \) to be the same function. This is the most common convention in Analysis, and is a useful
one: the codomain can be taken to just be any convenient target set containing the range. For example,
the most common types of functions considered in basic analysis are real-valued functions, and it is most
convenient to simply take the codomain to be \( \mathbb{R} \) in all cases. The reader should be warned, however, that
surjectivity of a function depends crucially on what codomain is specified.

II.3.2.16. The same logical question arises with more general relations: if \( \mathcal{R} \) is a relation from \( X \) to \( Y \),
are the \( X \) and \( Y \) intrinsic to \( \mathcal{R} \)? In this case we do not conventionally assume even that \( X \) is the domain of
the relation, so using the ordered pair definition we should not consider either \( X \) or \( Y \) to be intrinsic to \( \mathcal{R} \).
However, this problem will rarely if ever arise in this book.

Composition of Functions

Sequences

Indexed Collections of Sets

General Cartesian Products

II.3.3. Equivalence Relations

The second important class of relations are equivalence relations. There are three standard points of view
for these relations which can, and will, be used interchangeably.

II.3.3.1. Definition. A relation on a set is an \textit{equivalence relation} if it is reflexive, symmetric, and
transitive.

It is common to use the symbol \( \sim \) for an equivalence relation, although other similar symbols are sometimes
used. Thus a relation \( \sim \) on a set \( X \) is an equivalence relation if it has the following properties:

(i) \( x \sim x \) for all \( x \in X \).

(ii) If \( x, y \in X \) and \( x \sim y \), then \( y \sim x \).

(iii) If \( x, y, z \in X \), \( x \sim y \), and \( y \sim z \), then \( x \sim z \).

Note that by (i), the domain and range of an equivalence relation on \( X \) are all of \( X \).

II.3.3.2. Examples. (i) If \( X \) is any set, the identity relation (II.3.1.2.(iv)) is a (not very interesting)
equivalence relation on \( X \). The full relation (II.3.1.2.(iii)) on \( X \) is another equivalence relation (also not very
interesting). The identity relation and the full relation are the “smallest” and “largest” equivalence relations
on \( X \).

(ii) On \( \mathbb{R}^2 \), set \( (x_1, y_1) \sim (x_2, y_2) \) if and only if \( x_1 = x_2 \). Then \( \sim \) is an equivalence relation on \( \mathbb{R}^2 \).

(iii) Fix \( n \in \mathbb{N} \). On \( \mathbb{Z} \), set \( x \sim y \) if and only if \( x - y \) is divisible by \( n \). Then \( \sim \) is an equivalence relation on
\( \mathbb{Z} \), called \textit{congruence modulo} \( n \). We usually write \( x \equiv y \mod n \) if \( x \sim y \) for this relation.

131
(iv) For a generic example including (i) and (ii) (and, in slight disguise, (iii)), let $X$ and $Y$ be sets, and $f : X \to Y$ a function. If $x_1, x_2 \in X$, set $x_1 \sim_f x_2$ if and only if $f(x_1) = f(x_2)$. Then $\sim_f$ is an equivalence relation on $X$. We will see that every equivalence relation on $X$ is of this form for some $Y$ and $f$. By replacing $Y$ by the range of $f$, there is no loss of generality in assuming that $f$ is surjective, in which case the $Y$ and $f$ are uniquely determined (up to bijective equivalence) by the relation $\sim_f$ (II.3.3.9.).

Equivalence Classes

The crucial notion connecting equivalence relations with the other points of view is equivalence classes:

II.3.3.3. **Definition.** Let $\sim$ be an equivalence relation on a set $X$, and $x \in X$. Then the equivalence class (or $\sim$-equivalence class) of $x$ is

$$[x] = \{ y \in X : x \sim y \} \subseteq X.$$  

Note that $x \in [x]$ since $\sim$ is reflexive.

II.3.3.4. **Proposition.** If $\sim$ is an equivalence relation on a set $X$, and $x, y \in X$, then either $[x] = [y]$ or $[x]$ and $[y]$ are disjoint. We have $[x] = [y]$ if and only if $x \sim y$.

**Proof:** Suppose $x \sim y$. If $z \in [y]$, i.e. $y \sim z$, then $x \sim z$ by transitivity, so $z \in [x]$ and $[y] \subseteq [x]$. Since also $y \sim x$ by symmetry, by interchanging $x$ and $y$ in the argument we obtain $[x] = [y]$. Now suppose $[x]$ and $[y]$ are not disjoint, and let $z \in [x] \cap [y]$. Then $x \sim z$ and $z \sim y$, so $x \sim y$ and $[x] = [y]$.

II.3.3.5. Thus the set $X$ partitions into a disjoint union of equivalence classes. The equivalence classes completely determine the equivalence relation: two elements of $X$ are equivalent if and only if they are in the same equivalence class.

II.3.3.6. **Examples.** (i) If $\sim$ is the identity relation on $X$, then each equivalence class is a singleton (and conversely), i.e. $[x] = \{ x \}$ for each $x \in X$. If $\sim$ is the full relation on $X$, there is only one equivalence class: $[x] = X$ for every $x \in X$.

(ii) In the equivalence relation of II.3.3.2.(ii), we have

$$[\langle x_1, y_1 \rangle] = \{ \langle x_1, y \rangle : y \in \mathbb{R} \}$$

so the equivalence classes are the vertical lines in $\mathbb{R}^2$.

(iii) In the equivalence relation of II.3.3.2.(iii), we have

$$[x] = \{ x + kn : k \in \mathbb{Z} \}.$$ 

If $n = 2$, the equivalence class of any even integer is the set of all even integers, and the equivalence class of any odd integer is the set of all odd integers. For general $n$, there are exactly $n$ equivalence classes: $[0], [1], \ldots, [n-1]$.
(iv) In the equivalence relation of II.3.3.2 (iv), we have \([x] = f^{-1}(f(x))\) for any \(x \in X\), i.e. the equivalence classes are just the (nonempty) preimages of elements of \(Y\) (more correctly, singleton subsets of \(Y\)). Thus if \(f\) is surjective, there is a natural parametrization of the equivalence classes by elements of \(Y\).

II.3.3.7. Conversely, suppose \(\{X_i : i \in I\}\) is a partition of a set \(X\), i.e. the \(X_i\) are disjoint and nonempty and \(\bigcup_{i \in I} X_i = X\). Define a relation \(\sim\) on \(X\) by setting \(x \sim y\) if and only if \(x\) and \(y\) are in \(X_i\) for the same \(i\). Then \(\sim\) is an equivalence relation on \(X\) whose equivalence classes are precisely the \(X_i\). Thus there is a natural one-one correspondence between equivalence relations on \(X\) and partitions of \(X\). Viewing equivalence relations as partitions is the second standard point of view.

II.3.3.8. Suppose \(\{X_i : i \in I\}\) is a partition of a set \(X\). Define a function \(f : X \to I\) by setting \(f(x) = i\) if \(x \in X_i\). Then \(f\) is a well-defined function from \(X\) to \(I\), which is surjective since all the \(X_i\) are nonempty, and the equivalence relation defined by the partition is exactly \(\sim_f\) as defined in II.3.3.2 (iv). Conversely, if \(g\) is a function from \(X\) onto a set \(Y\), the equivalence classes of the relation \(\sim_g\) are indexed by the elements of \(Y\) as in II.3.3.6 (iv). Thus the third point of view of equivalence relations on a set \(X\) is via surjective functions from \(X\) to other sets. The target set corresponding to an equivalence relation can be identified with the set of equivalence classes of the relation.

We summarize the three points of view:

II.3.3.9. Theorem. Let \(X\) be a set. There is a natural one-one correspondence between

(i) Equivalence relations on \(X\).

(ii) Partitions of \(X\).

(iii) Surjective functions from \(X\) to other sets (up to bijective equivalence).

In (iii), “up to bijective equivalence” means: if \(f : X \to Y\) and \(g : X \to Z\) are surjective functions, then \(f\) and \(g\) are bijectively equivalent if there is a bijection \(\phi : Y \to Z\) with \(g = \phi \circ f\) (so \(f = \phi^{-1} \circ g\)), i.e. the functions are the “same” up to renaming the elements of the target set.

II.3.4. Exercises

II.3.4.1. Let \(R\) be a relation from \(X\) to \(Y\), and \(S\) a relation from \(Y\) to \(Z\). Show that \(\text{Dom}(S \circ R) = R^{-1}(\text{Range}(R) \cap \text{Dom}(S))\) and \(\text{Range}(S \circ R) = S(\text{Range}(R) \cap \text{Dom}(S))\).

II.3.4.2. Here is a common but fallacious argument that a relation which is symmetric and transitive must also be reflexive, hence an equivalence relation. “Let \(R\) be a relation on a set \(X\) which is symmetric and transitive, and let \(x \in X\). Let \(y\) be any element of \(X\) with \(xRy\). Then \(yRx\) by symmetry, and hence \(xRx\) by transitivity. Thus \(R\) is reflexive.”

(a) What is wrong with this argument?
(b) Give an example of a relation on a set which is symmetric and transitive, but not reflexive.
II.3.4.3. (a) What is wrong with the following argument that the image of an intersection under a function is the intersection of the images?

Let \( f : X \rightarrow Y \) be a function, and \( \{ A_i : i \in I \} \) an indexed collection of subsets of \( X \). Let \( y \in \bigcap_{i \in I} f(A_i) \). Then, for each \( i \in I \), \( y \in f(A_i) \), so there is an \( x \in A_i \) with \( f(x) = y \). Then \( x \in \bigcap_{i \in I} A_i \), so \( y \in f(\bigcap_{i \in I} A_i) \).

(b) Give an example of an \( f \) and two subsets \( A_1, A_2 \) of \( X \) and a \( y \in f(A_1) \cap f(A_2) \) for which \( y \notin f(A_1 \cap A_2) \). [Consider a constant function.]

(c) Use your example in (b) to show that \( f(A^c) \neq [f(A)]^c \) in general.

(d) Under what conditions on \( f \) can we assert that the image under \( f \) of an intersection is always the intersection of the images? Under what conditions is the image of a complement always the complement of the image?
II.4. Ordered Sets and Induction

II.4.1. Partially Ordered Sets

Partially ordered sets arise throughout mathematics. A particularly important use in analysis and topology is to use partially ordered index sets to describe convergence.

II.4.1.1. DEFINITION. A transitive, antisymmetric relation on a set \( X \) is called a partial order on \( X \). A partially ordered set is a set with a specified partial order. Elements \( x \) and \( y \) (\( x \neq y \)) in a partially ordered set \( (X \preceq) \) are comparable if \( x \preceq y \) or \( y \preceq x \); otherwise they are incomparable. A partial order \( \preceq \) on a set \( X \) is a total order if any two elements of \( X \) are comparable, i.e. if, for every \( x, y \in X \), at least one of \( x \preceq y \), \( x = y \), or \( y \preceq x \) holds.

The symbols \( \preceq, <, \preceq, \preceq, \ll, \) and their reversals (\( \geq, \), etc.) are typically used for partial orders, although other similar symbols can be used.

II.4.1.2. A partial order need not be reflexive; in fact, it can happen that \( x \not\preceq x \) for any \( x \), for example with \( < \) on \( \mathbb{R} \). Such a partial order is called a strict partial order. If \( \preceq \) is a partial order on a set \( X \), there is an associated reflexive partial order \( \preceq \) defined by \( x \preceq y \) if and only if \( x \preceq y \) and \( x \neq y \), and similarly there is an associated reflexive partial order \( \preceq \) defined by \( x \leq y \) if and only if \( x \preceq y \) or \( x = y \). We will usually consider reflexive partial orders for definiteness; there is no essential loss of generality in doing so. We will sometimes (slightly carelessly) not distinguish between partial orders whose associated reflexive partial orders coincide (and hence whose associated strict partial orders also coincide); such partial orders will be called equivalent.

II.4.1.3. If \( Y \) is a subset of \( X \), then the restriction to \( Y \) of any partial order on \( X \) is a partial order on \( Y \), called the induced ordering.

II.4.1.4. Examples.

(i) \( <, \leq, >, \geq \) are total orders on \( \mathbb{R} \), or any subset of \( \mathbb{R} \).

(ii) For any set \( X \), \( \subseteq \) and \( \supseteq \) are reflexive partial orders on \( \mathcal{P}(X) \). These are not total orders if \( X \) has more than one element: if \( x \neq y \), then \( \{x\} \) and \( \{y\} \) are incomparable.

(iii) On \( \mathbb{R}^2 \) (or any subset), define \( (x_1, y_1) \preceq (x_2, y_2) \) if and only if \( x_1 \leq x_2 \) and \( y_1 \leq y_2 \). This is a reflexive partial order which is not a total order [(1,0) and (0,1) are incomparable], called the ordinary ordering on \( \mathbb{R}^2 \).

(iv) On \( \mathbb{R}^2 \) (or any subset), define \( (x_1, y_1) \preceq (x_2, y_2) \) if and only if \( (x_1, y_1) = (x_2, y_2) \), or \( x_1 < x_2 \) and \( y_1 < y_2 \). This is a different reflexive partial order which is also not a total order [(1,0) and (1,1) are incomparable], called the strict ordering on \( \mathbb{R}^2 \) (the name is well established but somewhat unfortunate, since this is not a strict partial order.)

(v) On \( \mathbb{R}^2 \) (or any subset), define \( (x_1, y_1) \preceq (x_2, y_2) \) if and only if \( x_1 < x_2 \), or \( x_1 = x_2 \) and \( y_1 \leq y_2 \). This is a reflexive total order, called the lexicographic ordering on \( \mathbb{R}^2 \).
Examples (iii)-(v) can be generalized to arbitrary Cartesian products of partially ordered sets (for (v), the index set must be ordered; see Exercise ()). The term “lexicographic ordering” is used because this ordering is an analog of the alphabetical ordering of words, regarding the alphabet as a totally ordered set. An important generalization of (iii) is:

(vi) Let $X$ be a set, and $\mathbb{R}^X$ the set of all functions from $X$ to $\mathbb{R}$. If $f, g \in \mathbb{R}^X$, set $f \leq g$ if $f(x) \leq g(x)$ for all $x \in X$. This is the ordinary ordering on $\mathbb{R}^X$. It is not a total order if $X$ has more than one element.

A more complicated example, which is typical of a class of examples arising in analysis, is:

(vii) Let $X$ and $Y$ be sets, and $\mathcal{F}(X,Y)$ the set of functions from subsets of $X$ to $Y$. Write $f \preceq g$ if $g$ extends $f$, i.e. the domain of $g$ contains the domain of $f$ and $g = f$ on the domain of $f$. Then $\preceq$ is a partial order on $\mathcal{F}(X,Y)$ which is not a total order if $Y$ has more than one element. In fact, in general, “almost all” pairs of functions are incomparable.

There is a lot of terminology in the theory of partially ordered sets which we need to introduce, mostly related to upper and lower bounds of subsets.

II.4.1.5. Definition. Let $(X, \preceq)$ be a partially ordered set, and $A \subseteq X$. An element $x \in X$ is an upper bound for $A$ if $a \preceq x$ for every $a \in A$. The element $x$ is the supremum, or least upper bound, of $A$, written $x = \sup(A)$, if $x$ is an upper bound for $A$ and $x \preceq y$ for every upper bound $y$ of $A$. Lower bounds and greatest lower bounds are defined similarly; the greatest lower bound of $A$ is also called the infimum of $A$, written $\inf(A)$. $A$ is bounded above if it has an upper bound, and bounded below if it has a lower bound; $A$ is bounded if it is both bounded above and bounded below.

II.4.1.6. It is obvious from antisymmetry that the supremum or infimum of a set, if it exists, is unique. Note that an upper or lower bound for a subset $A$ is not a member of $A$ in general. For example, if $X = \mathbb{R}$ and $A = (0, 1)$, then any $x \geq 1$ is an upper bound for $A$, and $\sup(A) = 1$; similarly, any negative number is a lower bound for $A$ and $\inf(A) = 0$.

In a general partially ordered set, a subset which is bounded above need not have a supremum; examples are given in the exercises.

II.4.1.7. Definition. Let $(X, \preceq)$ be a partially ordered set. Then $(X, \preceq)$ is complete if every bounded subset of $X$ has a supremum and infimum.

II.4.1.8. One of the most important properties of $\mathbb{R}$ is the Completeness Axiom:

Completeness Axiom. $(\mathbb{R}, \leq)$ is complete.

There are many other properties of $\mathbb{R}$ which are equivalent to the completeness axiom, e.g. (), (), (), . . . . The completeness axiom is the main feature which distinguishes $\mathbb{R}$ from other number systems such as $\mathbb{Q}$. See Exercise ()

136
II.4.1.9. There are other important complete partial orders. Among the examples of II.4.1.4., (i), (ii), (iii), (vi), and (vii) are complete (in (ii), the supremum of a collection of subsets is their union; see Exercise () for (vii)), while (iv) and (v) are not complete (Exercises ().)

II.4.1.10. There are two finite versions of completeness which are important. A partially ordered set \((X;\preceq)\) is directed (technically, upward directed) if every pair of elements has an upper bound, and is a lattice if every pair of elements has a supremum and an infimum. The supremum and infimum of \(x\) and \(y\) are usually denoted \(x \lor y\) and \(x \land y\) respectively. Every totally ordered set is a lattice, and a lattice-ordered set is directed. Of the examples, (i), (ii), (iii), (iv), and (vi) are lattices, (iv) is directed but not a lattice, and (vii) is not directed (see Exercises ().) In a directed set \([\text{lattice}]\), every finite subset has an upper bound [supremum and infimum] (Exercise ()).

II.4.1.11. Definition. Let \((X;\preceq)\) be a partially ordered set. An element \(x \in X\) is a maximal element if \(y \in X\), \(x \preceq y\) implies \(y = x\). The element \(x \in X\) is the largest element if \(y \preceq x\) for all \(y \in X\). Minimal elements and smallest elements are defined similarly.

II.4.1.12. A largest element, if it exists, is obviously unique, and is a maximal element. In a totally ordered set, the two things are the same; but in a general partially ordered set a maximal element need not be a largest element, and there can be more than one maximal element. For example, in \(\mathcal{F}(X,Y)\) (II.4.1.4.(vii)), a function is a maximal element if and only if its domain is all of \(X\), and there is no largest element (if \(Y\) has more than one element.) The situation is the same for smallest or minimal elements.

There is also a notion of morphism for partially ordered sets:

II.4.1.13. Definition. Let \((X;\preceq)\) and \((Y;\leq)\) be partially ordered sets. A function \(f : X \to Y\) is order-preserving if \(x_1 \preceq x_2\) implies \(f(x_1) \leq f(x_2)\). The function \(f\) is an order-isomorphism if it is a bijection and both \(f\) and \(f^{-1}\) are order-preserving, i.e. \(x_1 \preceq x_2\) if and only if \(f(x_1) \leq f(x_2)\).

II.4.1.14. Caution: an order-isomorphism is more than an order-preserving bijection: for example, the identity map from \(\mathbb{R}^2\) with the strict ordering () to \(\mathbb{R}^2\) with the ordinary ordering () is an order-preserving bijection, but not an order-isomorphism.

II.4.1.15. A composition of order-preserving functions is order-preserving, so the collection of partially ordered sets and order-preserving functions forms a category.

II.4.1.16. We also slightly carelessly say that \((X;\preceq)\) and \((Y;\leq)\) are order-isomorphic if there is a bijection from \(X\) to \(Y\) which is an order-isomorphism between \(\preceq\) and a partial order on \(Y\) equivalent to \(\leq\). Thus, for example, we say that \((\mathbb{R},\leq)\) and \((\mathbb{R},<)\) are order-isomorphic.

II.4.2. Well-Ordered Sets and Transfinite Induction

There is a very special type of ordered set which is important in some applications:
II.4.2.1. Definition. A partially ordered set \((I, \preceq)\) is well-ordered if every nonempty subset has a smallest element.

We will generally use \(I\) and \(J\) as names of well-ordered sets since well-ordered sets are primarily used in analysis as index sets.

A well-ordered set \(I\) is totally ordered (if \(i, j \in I\), then \(\{i, j\}\) has a smallest element), and \(I\) itself has a smallest element. Every subset of a well-ordered set is well-ordered in the induced ordering.

The ordering on a well-ordered set is called a well ordering. While the term “well-ordered” is grammatically correct English, the term “well ordering” is an abuse of language (the word “well” used as an adjective is the opposite of “sick”), but the term is unfortunately “well established.”

II.4.2.2. Examples.

(i) Any finite totally ordered set is well-ordered.

(ii) \(\mathbb{N}\) is well-ordered.

(iii) \(\mathbb{N}^2\) with lexicographic ordering is well-ordered.

(iv) \(\mathbb{Z}, \mathbb{Q}, \mathbb{R}\) are not well-ordered since they have no smallest element. \([0, \infty)\) is not well-ordered either: it has a smallest element, but the subset \((0, \infty)\) does not.

II.4.2.3. Actually, well-ordered sets are quite unusual. While there are many of them, they all look somewhat the same except for length: if \(I\) and \(J\) are well-ordered, then one of them is order-isomorphic to an initial segment of the other, in a unique way (II.4.2.19.). Well-ordered sets are described in more detail in ()

II.4.2.4. Every element of a well-ordered set, except the largest element if there is one, has an immediate successor: for any \(i\), the set \(\{j : i \prec j\}\), if nonempty, has a smallest element. An element need not have an immediate predecessor, however; for example, in \(\mathbb{N}^2\) with lexicographic order, \((2, 1)\) has no immediate predecessor. An element with no immediate predecessor is called a limit element.

II.4.2.5. A well ordering is a complete ordering: if \((I, \preceq)\) is well-ordered, and \(A \subseteq I\) is nonempty, then \(A\) has a smallest element which is \(\inf(A)\); and the set of upper bounds for \(A\), if nonempty, has a smallest element which is \(\sup(A)\).

Transfinite Induction

A proof procedure generalizing ordinary induction can be based on any well-ordered set.

II.4.2.6. Theorem. [Principle of Transfinite Induction] Let \((I, \preceq)\) be a well-ordered set, and suppose \(P(i)\) is a statement for each \(i \in I\). Suppose, for each \(i \in I\), the following inductive implication holds:

“If \(P(j)\) is true for all \(j < i\), then \(P(i)\) is true.”
Then $P(i)$ is true for all $i \in I$.

We have not been precise about the meaning of the term “statement”, since to be so would require an extensive excursion into set theory and mathematical logic. Many books on mathematical logic such as [?] contain a careful discussion of this point. But any type of mathematical statements likely to be made by a working mathematician come within the scope of the principle.

II.4.2.7. We outline the proof (although we cannot be completely precise because of the imprecision in the term “statement.”) Let $J$ be the set of $j \in I$ such that $P(j)$ is true. If $J \neq I$, since $I$ is well-ordered there is a smallest element $i \in I \setminus J$. But then $P(j)$ is true for all $j < i$, and therefore $P(i)$ is true by the inductive implication; this is a contradiction, so $I = J$.

II.4.2.8. The term “transfinite induction” is usually used in the case where $I$ properly contains an initial segment order-isomorphic to $\mathbb{N}$, and especially in the case where $I$ is uncountable.

One of the most frequently used consequences of the principle of transfinite induction is:

II.4.2.9. Corollary. Let $(I, \leq)$ be a well-ordered set, and $J \subseteq I$. If $J$ has the property that, for every $i \in I$, $\{j \in J \text{ for all } j < i\}$ implies $i \in J$, then $J = I$.

Proof: Let $P(i)$ be the statement “$i \in J$”, and apply II.4.2.6.

II.4.2.10. Transfinite induction in the special case $I = \mathbb{N}$ is usually called complete induction. Complete induction is not quite the same as ordinary induction, where statements $P(n)$ indexed by $\mathbb{N}$ are proved for all $n$ by proving $P(1)$ and $[P(n) \Rightarrow P(n+1)]$ for all $n$. Instead, in complete induction, we assume $P(k)$ is true for all $k \in \mathbb{N}, k < n$, and deduce $P(n)$. Ordinary induction is a special case of complete induction; on the other hand, complete induction can be easily deduced from ordinary induction (Exercise ()).

Complete induction is useful where there is no simple relationship between $P(n)$ and $P(n+1)$, but there is a relationship between $P(n)$ and $P(k)$ for certain $k$’s less than $n$. The next simple example is a good illustration.

II.4.2.11. Proposition. Every natural number can be written as a finite product of prime numbers. [We define the product of an empty set of primes to be 1, and a product of one prime $p$ to be $p$.]

Proof: For each $n \in \mathbb{N}$, let $P(n)$ be the statement “The integer $n$ can be written as a finite product of prime numbers.” Fix $n \in \mathbb{N}$, and assume $P(k)$ is true for all $k < n$; we must prove that $P(n)$ is true. There are three cases:

1. $n = 1$. Then $n$ is an empty product of prime numbers.
2. $n > 1$ is prime. Then $n$ is a product of one prime number.
3. $n > 1$ is not prime. Then $n = n_1n_2$ for some $n_1, n_2 \in \mathbb{N}, n_1, n_2 > 1$. Then $n_1, n_2 < n$, so by assumption $n_1$ and $n_2$ can each be written as a finite product of prime numbers. Hence so can $n_1n_2 = n$. 

139
II.4.2.12. Note that in complete induction, it is not necessary in general to separately prove \( P(1) \), since it is included in the general case. In the case \( n = 1 \), the assumption that \( P(k) \) is true for all \( k < n \) is automatically satisfied since there are no such \( k \)'s (in this case, the hypothesis is said to be vacuous.) Of course, it is allowed to prove \( P(1) \) as a special case, as was done in the proof of II.4.2.11.

Definitions by Transfinite Induction or Recursion

Transfinite induction (or, technically, transfinite recursion) can also be used to define functions whose domain is a well-ordered set. We will again not be completely precise in defining terms; see, for example, [? p. 249-250] for a careful treatment.

II.4.2.13. Let \( (I, \leq) \) be a well-ordered set, and \( X \) any set. Suppose, for each \( i \in I \), we have a definite procedure or rule \( R_i \) for choosing an \( f(i) = x_i \in X \), which may depend on the choice of the \( f(j) \) for \( j < i \). Then

II.4.2.14. Theorem. There is a unique \( f : I \to X \) such that, for each \( i \in I \), \( f(i) \) is determined by the specified procedure \( R_i \) applied to \( \{f(j) : j < i\} \).

Roughly speaking, if \( f \) can be defined “one step at a time” by the rules \( R_i \), then \( f \) can be defined “all at once.”

II.4.2.15. We outline an argument “proving” the theorem, overlooking the imprecision in the term “procedure”. We first prove uniqueness of \( f \) by transfinite induction: if \( f \) and \( g \) have \( f(i) \) determined for each \( i \) by applying \( R_i \) to \( \{f(j) : j < i\} \), and \( g(i) \) determined by applying \( R_i \) to \( \{g(j) : j < i\} \), let \( J \) be the set of \( i \) for which \( f(i) = g(i) \). If \( i \in I \) and \( j \in J \) (i.e. \( f(j) = g(j) \)) for all \( j < i \), then applying \( R_i \) we obtain \( f(i) = g(i) \), so \( i \in J \). Thus \( J = I \) by II.4.2.9.

For existence, let \( J \) be the set of all \( i \) such that there is a function \( f_i \) on \( \{k : k \leq i\} \) with \( f_i(k) \) determined by \( R_k \) applied to \( \{f_i(j) : j < k\} \) for each \( k \leq i \). If \( i_1, i_2 \in J \) and \( k \leq i = \min(i_1, i_2) \), then \( f_{i_1}(k) = f_{i_2}(k) \) by the uniqueness of the first part of the proof, applied to \( \{k : k \leq i\} \). Thus, if \( i \in I \) and \( j \in J \) for all \( j < i \), there is a (unique) function \( f \) on \( \{j : j < i\} \) with \( f(j) \) determined by \( R_j \) applied to \( \{f(k) : k < j\} \) for all \( j < i \). Apply \( R_i \) to \( \{f(j) : j < i\} \) to define \( f(i) \) and extend \( f \) to \( \{j : j \leq i\} \). Thus \( i \in J \). By II.4.2.9, \( J = I \).

This “proof” is actually fallacious; see Exercise (). The actual proof (cf. []) is an adaptation of the proof of the following special case, which is the most important one for applications.

II.4.2.16. Theorem. Let \( X \) be a set, and for each \( n \in \mathbb{N} \) let \( f_n \) be a function from \( X^n \) to \( X \). If \( x_1 \) is any element of \( X \), then there is a unique sequence \( (x_n) \) in \( X \) with \( x_{n+1} = f_n(x_1, \ldots, x_n) \) for all \( n \in \mathbb{N} \).

Proof: Let \( S \) be the smallest subset of \( \mathbb{N} \times X \) containing \( (1, x_1) \), with the property that if, for any \( n \in \mathbb{N} \), whenever \( \{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S \) for \( (y_1, \ldots, y_n) \in X^n \), then \( (n+1, f_n(y_1, \ldots, y_n)) \in S \). \( S \) is the intersection of all subsets of \( \mathbb{N} \times X \) with these two properties; there is at least one such subset, \( \mathbb{N} \times X \) itself.) Then let

\[ A = \{n \in \mathbb{N} : (k, x_k) \in S \text{ for exactly one } x_k \in X \text{ for all } k \in \mathbb{N}, k \leq n\} \, . \]

First we show \( 1 \in A \). Set

\[ S' = \{(1, x_1)\} \cup \{(n, x) \in S : n > 1\} \]

140
and note that \( S' \subseteq S' \), \((1, x_1) \in S'\), and \( \{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S'\) implies \((n + 1, f_n(y_1, \ldots, y_n)) \in S'\); hence \( S' = S \) and \( 1 \in A \). Now suppose \( n \in A \), i.e. there is a unique \( x_k \in X \) with \((k, x_k) \in S\) for all \( k \in \mathbb{N} \), \( k \leq n \). Set

\[
S'' = \{(k, x_k) : k \in \mathbb{N}, k \leq n\} \cup \{(n + 1, f_n(x_1, \ldots, x_n))\} \cup \{(m, x) \in S : m > n + 1\}
\]

and note that \( S'' \subseteq S', (1, x_1) \in S'' \), and \( \{(k, y_k) : k \in \mathbb{N}, k \leq m\} \subseteq S'' \) implies \((m + 1, f_m(y_1, \ldots, y_m)) \in S''\). Thus \( S'' = S \) and \( n + 1 \in A \). So \( A = \mathbb{N} \).

For each \( n \in \mathbb{N} \), let \( x_n \) be the unique element of \( X \) for which \((n, x_n) \in S \). Let \( B = \{ n \in \mathbb{N} : x_{n+1} = f_n(x_1, \ldots, x_n) \}\). We have \( 1 \in B \) since \((2, f_1(x_1)) \in S\); and if \( n \in B \), then \((n + 2, f_{n+1}(x_1, \ldots, x_{n+1})) \in S\), so \( n + 1 \in B \). Thus \( B = \mathbb{N} \).

Note that no form of the Axiom of Choice is needed in this proof; in fact, the AC is not needed even for the transfinite result II.4.2.14.

II.4.2.17. In other words, if a definite rule is specified for each \( n \) for determining the \((n + 1)\)'st term of the sequence from the first \( n \) terms, there is a unique sequence generated beginning with any first term. The rule \( f_n \) can vary with \( n \), but all the \( f_n \) must be specified in advance. The sequence is said to be defined by induction (or, more correctly, defined by recursion) from \( x_1 \) by the functions \((f_n)\).

II.4.2.18. Caution: In definitions by transfinite induction, the rule \( R_i \) can be quite general, but it must be a definite rule giving a specific element as a function of \( \{f(j) : j < i\} \). Thus, for example, a “rule” like “Let \( f(i) \) be any element of \( X \setminus \{f(j) : j < i\} \)” is not permissible, even if it is proved that this set is always nonempty (say, by a cardinality argument.) But if \( X \) is well-ordered, it is permissible to let \( R_i \) be the rule “Let \( f(i) \) be the first element of \( X \setminus \{f(j) : j < i\} \)” (provided this set is always nonempty.) In fact, this rule can be used to give a proof of the “uniqueness up to length” of well orderings:

II.4.2.19. Theorem. Let \((I, \leq)\) and \((J, \leq)\) be well-ordered sets. Then there is a unique order-isomorphism from one of \( I \), \( J \) onto an initial segment of the other.

Proof: The idea is that we pair up the first element of \( I \) with the first element of \( J \), and each succeeding element of \( I \) with the next remaining element of \( J \), until we “run out of elements” in either \( I \) or \( J \). This is clearly the only possible way to obtain an order-isomorphism of the desired type, giving uniqueness.

More precisely, we try to define \( f : I \rightarrow J \) as follows. For each \( i \in I \) let \( f(i) \) be the first element of \( J \setminus \{f(k) : k \prec i\} \). This inductively defines \( f \) on the set \( A = \{i \in I : Y \setminus \{f(k) : k \prec i\} \neq \emptyset\} \). This set \( A \) is either all of \( I \) or is of the form \( \{k \in I : k \prec i\} \) for some \( i \in I \), which is a proper initial segment of \( I \); and \( f \) is an order-isomorphism from \( A \) onto \( f(A) \). In the case \( A \neq I \), \( f \) maps \( A \) bijectively to \( J \). If \( A = I \), then either \( f \) maps \( I \) onto \( J \), or there is a smallest element \( j \) of \( J \setminus f(I) \), in which case \( f(I) = \{k \in J : k < j\} \), which is an initial segment of \( J \), since we can never have \( j \leq f(i) \) for any \( i \in I \) by the way \( f \) is defined.

An alternate proof not using transfinite induction is outlined in Exercises ()

141
II.4.3. Exercises

II.4.3.1. Consider the partially ordered set \((\mathbb{R}^2, \preceq)\) of II.4.1.4.(iv) (strict ordering).
(a) Show that \((1,0), (0,1), \text{ and } (1,1)\) are mutually incomparable.
(b) Show that \(\{(1,0),(0,1)\}\) has no supremum or infimum.
(c) Show that \(\{(x,y) : i \in I\}\) is bounded in \((\mathbb{R}^2, \preceq)\) if and only if \(\{x_i\}\) and \(\{y_i\}\) are bounded in \(\mathbb{R}\). Conclude that \((\mathbb{R}^2, \preceq)\) is directed.
(d) Define an analog of the strict ordering on \(\mathbb{R}^X\) for a nonempty set \(X\), and prove that \((\mathbb{R}^X, \preceq)\) is directed but not a lattice (unless \(X\) is a singleton.)

II.4.3.2. Consider the partially ordered set \((\mathbb{R}^2, \preceq)\) of II.4.1.4.(v) (lexicographic ordering).
(a) Show that a subset \(\{(x_i,y_i) : i \in I\}\) is bounded in \((\mathbb{R}^2, \preceq)\) if and only if \(\{x_i\}\) is bounded in \(\mathbb{R}\).
(b) Show that \(\{(x,y) : 0 < x < 1\}\) and \(\{(x,y) : 0 \leq x \leq 1\}\) are bounded but have no supremum or infimum.

II.4.3.3. Consider the partially ordered set \(F(X,Y)\) of II.4.1.4.(vii).
(a) Show that if \(f\) and \(g\) have a common domain point at which they disagree, then \(\{f,g\}\) has no upper bound. Hence \(F(X,Y)\) is not (upward) directed.
(b) If \(A = \{f_i : i \in I\}\) is a nonempty subset of \(F(X,Y)\) which is bounded above by a function \(g\), let \(h\) be the restriction of \(g\) to the union of the domains of the \(f_i\). Then \(h = \sup(A)\).
(c) Every subset of \(F(X,Y)\) is bounded below; in fact, \(F(X,Y)\) has a smallest element, the function whose domain is \(\emptyset\).
(d) If \(A = \{f_i : i \in I\}\) is a nonempty subset of \(F(X,Y)\), set
\[
Z = \{x \in X : x \in \text{Dom}(f_i) \text{ for all } i \in I, \ f_i(x) = f_j(x) \text{ for all } i,j \in I\}
\]
and let \(g\) be the restriction of any \(f_i\) to \(Z\). Then \(g = \inf(A)\). Thus \(F(X,Y)\) is complete.

II.4.3.4. Suppose \(\{X_i : i \in I\}\) is an indexed collection of partially ordered sets, and \(I\) also has a well ordering. We will put a lexicographic ordering on \(\prod_{i \in I} X_i\). Use the symbol \(\preceq\) for all the partial orders on \(I\) and the \(X_i\)'s, and write \(x < y\) if \(x \preceq y\) and \(x \neq y\).
(a) If \(x = (x_i)\) and \(y = (y_i)\) are elements of \(\prod X_i\), and \(x \neq y\), then \(\{i \in I : x_i \neq y_i\}\) is nonempty, hence has a smallest element \(i_0\). Let \(x \preceq y\) if \(x_{i_0} \preceq y_{i_0}\). Show that \(\preceq\) is a partial order on \(\prod X_i\).
(b) \(\prod X_i\) is totally ordered if and only if each \(X_i\) is totally ordered.
(c) If each \(X_i\) is well-ordered, is \(\prod X_i\) well-ordered?
(d) Discuss the difficulties in trying to define a lexicographic ordering if \(I\) is not well-ordered.

II.4.3.5. Explain why ordinary induction is a special case of complete induction.

The next three problems give an alternate proof of Theorem II.4.2.19 not using transfinite induction.

II.4.3.6. Let \((I, \preceq)\) be a well-ordered set, and \(f : I \to I\) an injective order-preserving function. Prove directly from the definition of a well-ordered set that \(f(i) \geq i\) for all \(i \in I\). [If not, let \(i_0\) be the smallest \(i\) for which \(f(i) < i\), and set \(j = f(i_0)\). Show that \(f(j) < j\).]

142
II.4.3.7. (a) Let \((I, \leq)\) be a well-ordered set. Then the only order-isomorphism from \(I\) onto an initial segment of \(I\) is the identity map on \(I\). [If \(f\) is such an order-isomorphism, then \(f(i) \geq i\) for all \(i\) by II.8.4.2.]

Let \(i_0\) be the smallest \(i\) with \(f(i) > i\), and let \(j \in I\) with \(f(j) = i_0\). Show \(j < i_0\) and \(f(j) > j\].

(b) Let \((I, \preceq)\) and \((J, \leq)\) be well-ordered sets. Then there is at most one order-isomorphism from \(I\) to \(J\).

II.4.3.8. Let \((I, \preceq)\) and \((J, \leq)\) be well-ordered sets. If \(u \in I\), write \(I_u = \{i \in I : i < u\}\). Also, if \(X\) and \(Y\) are partially ordered sets, write \(X \sim Y\) if \(X\) and \(Y\) are order-isomorphic. Define

\[
\Gamma = \{(u, v) \in I \times J : I_u \sim J_v\}.
\]

Show that \(\Gamma\) (or \(\Gamma^{-1}\)) is the graph of an order-isomorphism from an initial segment of \(I\) onto \(J\) (or an order-isomorphism from an initial segment of \(J\) onto \(I\)), as follows:

(a) Let \(I_0 = \{u \in I : \exists v \in J : (u, v) \in \Gamma\}\), \(J_0 = \{v \in J : \exists u \in I : (u, v) \in \Gamma\}\)

(i.e. \(I_0\) and \(J_0\) are the domain and range of the relation \(\Gamma\)) Show that \(I_0\) and \(J_0\) are initial segments of \(I\) and \(J\) respectively. [If \((u_0, v_0) \in \Gamma\) and \(f : I_{u_0} \to J_{v_0}\) is an order-isomorphism, and \(u \leq u_0\), \(v \leq v_0\), restrict \(f\) and \(f^{-1}\) to \(I_u\) and \(J_v\).]

(b) Show that \(\Gamma\) is the graph of a one-to-one function \(\phi\) from \(I_0\) to \(J_0\). [Suppose \(u \in I\), \(v_1, v_2 \in J\), \(v_1 \leq v_2\), and \((u, v_1), (u, v_2) \in \Gamma\). Let \(f : I_u \to J_{v_1}\) and \(g : I_u \to J_{v_2}\) be order-isomorphisms. Then \(f \circ g^{-1}\) is an order-isomorphism from the well-ordered set \(J_{v_2}\) onto its initial segment \(J_{v_1}\), so \(v_1 = v_2\) (and \(f = g\) by II.4.3.7..]

(c) Show that \(\phi\) is an order-isomorphism from \(I_0\) onto \(J_0\). [Suppose \((u_1, v_1)\) and \((u_2, v_2)\) are in \(\Gamma\) and \(u_1 < u_2\), and \(f : I_{u_1} \to J_{v_1}\) and \(g : I_{u_2} \to J_{v_2}\) are order-isomorphisms. If \(v_1 > v_2\), then \(h = g \circ f^{-1}\) is an injective order-preserving function from \(J_{v_1}\) to \(J_{v_2}\) \(\subseteq J_{v_1}\) with \(h(v_2) < v_2\), contradicting II.8.4.2..]

(d) Show that either \(I_0 = I\) or \(J_0 = J\). [If not, let \(u_0\) be the smallest element of \(I \setminus I_0\) and \(v_0\) the smallest element of \(J \setminus J_0\). Then \(I_0 = I_{u_0}\), \(J_0 = J_{v_0}\), and \(\phi\) gives an order-isomorphism from \(I_{u_0}\) to \(J_{v_0}\), so \((u_0, v_0) \in \Gamma\).]

(e) The function \(\phi\) gives the unique order-isomorphism from \(I_0\) to \(J_0\) by II.4.3.7..
II.4.4. Directed Sets

There is another kind of “ordered” set which is important, especially in topology. In many instances, one only needs to have a notion of “sufficiently large” elements of an “ordered” set. This notion is made precise in a useful way in directed sets.

II.4.4.1. Definition. Let \( X \) be a nonempty set. A direction on \( X \) is a transitive relation (preorder) \( \leq \) with the property that any two elements of \( X \) have an upper bound, i.e. for any \( a, b \in X \) there is a \( c \in X \) with \( a \leq c \) and \( b \leq c \). A directed set is a pair \((X, \leq)\) where \( X \) is a nonempty set and \( \leq \) is a direction on \( X \). A direction \( \leq \) on a set \( X \) is proper if for every \( a \in X \) there is a \( c \in X \) with \( a \leq c \) and \( c \not\leq a \).

II.4.4.2. Note that a direction on a set does not have to be a partial order, i.e. a direction does not have to be antisymmetric; however, many important directions are partial orders, and there is no essential loss of generality in requiring that the preorder be a partial order (II.4.4.1.), although there is a loss of utility in applications.

A direction is also often required to be reflexive, but this is unnecessary (and can be easily arranged anyway; see Exercise II.4.5.1.). In the definition, \( a \) and \( b \) do not have to be distinct, i.e. in a directed set \( X \), for any \( a \in X \) there is a \( c \in X \) with \( a \leq c \). A directed set in the sense of II.4.4.1. is sometimes called upward directed. Downward directed sets can also be defined, but we will not work with them in this book.

II.4.4.3. Examples. (i) Any totally ordered set is a directed set (provided the “largest” element \( a \), if there is one, satisfies \( a \leq a \)). The ordering is a proper direction if and only if there is no largest element. In particular, \( \mathbb{N} \) is properly directed in the usual order.

(ii) More generally, any lattice is a directed set. In particular, the power set of any set \( X \) is directed by inclusion, as is the collection of finite subsets of \( X \); the latter is a properly directed set if \( X \) is infinite.

(iii) Any partially ordered set with a largest element is directed but not properly directed. These sets are not very interesting as directed sets; the interesting directed sets are ones with no maximal elements, since the most important feature of a directed set is its “tails.” A partially ordered and directed set is properly directed if and only if it has no maximal elements.

(iv) A set with the full relation (\( x \leq y \) for all \( x \) and \( y \)) is a directed set which is not partially ordered.

(v) The equality relation on a set with more than one element is a partial order which is not a direction.

(vi) The partially ordered set of (\( \emptyset \)) is not directed.

(vii) Here is a very important example from topology. Let \( X \) be a topological space and \( p \in X \). Let \( \mathcal{N}_p \) be the set of all neighborhoods of \( p \). If \( U, V \in \mathcal{N}_p \), let \( V \subseteq U \) if \( U \subseteq V \). Then \( \mathcal{N}_p \) is a directed set with this ordering. \( \mathcal{N}_p \) can be replaced here by the set of open neighborhoods of \( p \), or by any local base for the topology at \( p \). There are other variations, such as setting \( V \leq U \) if the closure of \( U \) is contained in \( V \) (if \( X \) is a regular space).

(viii) As a variation of (vii), let \( X \) be a set and \( \mathcal{F} \) be a filter (\( \emptyset \)) of subsets of \( X \). If \( A, B \in \mathcal{F} \), set \( A \leq B \) if \( B \subseteq A \). Then \((\mathcal{F}, \leq)\) is a directed set.

(ix) If \((X, \leq)\) is a directed set, a subset \( Y \) of \( X \) is cofinal in \( X \) if for every \( x \in X \) there is a \( y \in Y \) with \( x \leq y \). A cofinal subset of a directed set is directed in the induced preorder.
II.4.4.4. **Proposition.** Let \((X, \leq)\) be a directed set. Then any finite subset of \(X\) has an upper bound, i.e. for any finite subset \(\{a_1, \ldots, a_n\}\) of \(X\), there is a \(c \in X\) with \(a_k \leq c\) for all \(1 \leq k \leq n\).

**Proof:** The \(c\) (or, more accurately, a \(c\) that works – the \(c\) will not be unique in general) can be found iteratively: let \(c_2\) be an upper bound for \(a_1\) and \(a_2\), \(c_3\) an upper bound for \(c_2\) and \(a_3\), \ldots, \(c\) an upper bound for \(c_{n-1}\) and \(a_n\). (This argument should really be done more carefully by induction.)

II.4.4.5. **Proposition.** Let \(\leq\) be a direction on a set \(X\). Then \(\leq\) is proper if and only if there is no largest element (an element \(a\) is a largest element if \(x \leq a\) for all \(x \in X\)).

**Proof:** If there is a largest element, then \(\leq\) is obviously not proper. Suppose there is no largest element. If \(a \in X\), then there is a \(b \in X\) such that \(b \not\leq a\). Let \(c\) be an upper bound for \(\{a, b\}\). Then \(a \leq c\). If \(c \leq a\), then since \(b \leq c\) we get \(b \leq a\) by transitivity, a contradiction. Thus \(c \not\leq a\) and \(\leq\) is proper.

We also need the notion of a direction-preserving function between directed sets:

II.4.4.6. **Definition.** Let \((X, \leq)\) and \((Y, \preceq)\) be directed sets. A function \(f : X \to Y\) is directed if, for every \(b \in Y\), there is an \(a \in X\) such that \(b \preceq f(a)\) for all \(a \in X\) with \(a \leq x\).

The condition in the definition is sometimes called the cofinal condition.

II.4.4.7. **Note** that for \(f\) to be directed, it is neither necessary nor sufficient that \(f\) be order-preserving. For example, the function from \(\mathbb{N}\) to \(\mathbb{N}\) which interchanges \(2n - 1\) and \(2n\) for each \(n\) is directed but not order-preserving. In fact, any bijection from \(\mathbb{N}\) to \(\mathbb{N}\) is directed (II.4.5.2.). But a constant function from \(\mathbb{N}\) to \(\mathbb{N}\) is not directed, even though it is order-preserving. There are even injective order-preserving functions which are not directed: the inclusion of \(\mathbb{N}\) into \(\mathbb{N} \cup \{+\infty\}\) is an example. Directed functions are “almost order-preserving” (Exercise II.4.5.3.).

Directed functions preserve “tails” or “cofinality” and thus are the appropriate maps between directed sets. In particular, the inclusion of a subset \(Y\) of \(X\) into \(X\) is a directed function if and only if \(Y\) is cofinal (II.4.3.(ix)) in \(X\).

The next result is easily verified:

II.4.4.8. **Proposition.** A composition of directed functions is directed.

II.4.4.9. **Definition.** Let \((X, \leq)\) and \((Y, \preceq)\) be directed sets. A function \(f : X \to Y\) is a directed equivalence if \(f\) is a bijection and \(f\) and \(f^{-1}\) are directed. \((X, \leq)\) and \((Y, \preceq)\) are directed equivalent if there is a directed equivalence from \(X\) to \(Y\). Two directions \(\leq\) and \(\preceq\) on a set \(X\) are equivalent if the identity map from \((X, \leq)\) to \((X, \preceq)\) is a directed equivalence.
II.4.4.10. The directed sets and directed functions form a category (\(\mathcal{D}\)), and the directed equivalences are just the isomorphisms in this category.

A largest element in a partially ordered set, if there is one, is unique; but a largest element in a directed set is not necessarily unique (if \(X\) has the full relation, then every element of \(X\) is a largest element.) The directions equivalent to partially ordered directions are precisely the ones for which this does not happen:

II.4.4.11. Theorem. Let \(\leq\) be a direction on a set \(X\). Then there is an equivalent direction \(\preceq\) on \(X\) which is a reflexive partial order if and only if there is at most one largest element for \(\preceq\). In particular, if \(\leq\) is proper, there is an equivalent proper direction which is a reflexive partial order.

Proof: Suppose \(b\) and \(c\) are largest elements of \(X\), and suppose there is a direction \(\leq\) equivalent to \(\preceq\) which is a partial order. Let \(a\in X\) such that \(a\leq x\) implies \(b\leq x\). Applying this with \(x=c\), we obtain \(b\leq c\). A similar argument shows that \(c\leq b\), so \(b=c\) since \(\leq\) is a partial order. Thus \((X,\leq)\) cannot have more than one largest element.

Suppose \(\leq\) has exactly one largest element \(c\). Define \(\preceq\) by \(x\preceq y\) if \(x=y\) or \(y=c\). Then \(\preceq\) is a direction on \(X\) which is a reflexive partial order. To show that \(\preceq\) is equivalent to \(\leq\), we need that for every \(b\in X\) there is an \(a\in X\) such that \(a\leq x\) implies \(b\leq x\), and for any \(a\in X\) there is a \(b\in X\) such that \(b\leq x\) implies \(a\leq x\). Just take \(a=c\) in the first case and \(b=c\) in the second.

Now suppose \(\leq\) is proper, i.e. there are no largest elements. Define \(\preceq\) by setting \(x\preceq y\) if \(x=y\) or if \(x\neq y\), \(x\leq\), and \(y\neq x\). It is obvious that \(\preceq\) is reflexive and antisymmetric. To see that \(\preceq\) is transitive, suppose \(x\leq y\) and \(y\leq z\). If \(x=z\), then \(x\leq z\). Suppose \(x\neq z\). If \(x=y\) or \(y=z\), it is trivial that \(x\leq z\), so we may assume \(x, y, z\) are all distinct. Then \(x\leq y\) and \(y\leq z\), so \(x\leq z\). If \(z\leq x\), then \(z\leq y\) by transitivity; but \(y\neq z\) and \(y\leq z\), so \(z\neq y\), a contradiction. Thus \(z\neq x\) and \(x\leq z\).

To show that \(\preceq\) is equivalent to \(\leq\), let \(a, b\in X\). Take \(c\in X\) with \(a\leq c\) and \(b\leq c\), and take \(d\in X\) with \(c\leq d\) and \(d\not\leq c\). We have \(d\neq a\) and \(d\neq b\), and \(a\leq d\) and \(b\leq d\) by transitivity. If \(d\leq a\), we would have \(d\leq c\) by transitivity, a contradiction; thus \(d\not\leq a\), i.e. \(a\leq d\). Similarly \(b\leq d\).

Finally, to show that \(\preceq\) is equivalent to \(\leq\), let \(b\in X\). We must find \(a\in X\) such that \(a\leq x\) implies \(b\leq x\). Let \(a\in X\) with \(b\leq a\) and \(a\not\leq b\). If \(a\leq x\), then either \(a=x\) or \(a\leq x\); in either case \(b\leq x\). Conversely, let \(a\in X\). We must find \(b\in X\) such that \(b\leq x\) implies \(a\leq x\). Take \(b\in X\) with \(a\leq b\) and \(b\not\leq a\). If \(x\in X\) and \(b\leq x\), then \(x\neq a\) and \(a\leq x\) by transitivity. If \(x\leq a\), then \(b\leq a\) by transitivity, a contradiction, so \(x\not\leq a\) and \(a\leq x\).

Note that no form of Choice is needed for this result (rephrase the last two paragraphs as proofs by contradiction; a contradiction can be obtained by making only finitely many choices).

II.4.5. Exercises

II.4.5.1. Let \(\preceq\) be a direction on a set \(X\). Define \(\preceq\) on \(X\) by \(a\preceq b\) if \(a\preceq b\) or \(a=b\). Show that \(\preceq\) is a reflexive direction on \(X\) which is equivalent to \(\preceq\).

II.4.5.2. (a) Show that a function \(f: \mathbb{N} \to \mathbb{N}\) is directed (for the usual order) if and only if \(f^{-1}\{\{n\}\}\) is finite for every \(n\in \mathbb{N}\).

(b) Show that every bijection from \(\mathbb{N}\) to \(\mathbb{N}\) is a directed equivalence.
II.4.5.3. Let \((X, \leq)\) and \((Y, \preceq)\) be directed sets, and \(f : X \to Y\) a directed function. Show that there is a direction \(\leq\) on \(X\) which is equivalent to \(\leq\), for which \(f\) is order-preserving. [Set \(a \leq b\) if \(a \leq b\) and \(f(a) \preceq f(b)\).]

II.4.5.4. Let \((X, \leq)\) and \((Y, \preceq)\) be directed sets. On \(X \times Y\), define \((x_1, y_1) \leq (x_2, y_2)\) if \(x_1 \leq x_2\) and \(y_1 \preceq y_2\). Show that \(\leq\) is a direction on \(X \times Y\), and that the coordinate projections are directed functions. Generalize to arbitrary Cartesian products of directed sets (the Axiom of Choice is needed in the general case).

II.4.5.5. There are interesting countable directed sets other than \(\mathbb{N}\) with the usual ordering. Define a relation \(\lessdot\) on \(\mathbb{N}\) by setting \(n \lessdot m\) if \(n|m\).

(a) Show that \(\lessdot\) is a partial order and a proper direction on \(\mathbb{N}\).
(b) Show that \((n!)\) is a subsequence of \((\mathbb{N}, \lessdot)\), i.e. is cofinal.
(c) Show that the sequence of prime numbers is not a subsequence of \((\mathbb{N}, \lessdot)\).
(d) Show that \(\lessdot\) is not equivalent to the ordinary order \(\leq\) on \(\mathbb{N}\).
(e) Can \((\mathbb{N}, \lessdot)\) be written as an infinite product (cf. II.4.5.4.) of copies of \((\mathbb{N}, \leq)\)?
II.5. Cardinality of Sets

“The infinite! No other question has ever moved so profoundly the spirit of man; no other idea has so fruitfully stimulated his intellect; yet no other concept stands in greater need of clarification than that of the infinite.”

David Hilbert

“The rather colorless idea of a collection of elements that had lurked in the background of mathematical thought since prehistory might have remained there to this day if Cantor had not had the audacity to assume that they could be infinite. This was the bold and revolutionary hypothesis that launched modern mathematics; it should be seen as nothing less.”

Penelope Maddy

One of the great mathematical advances of the late nineteenth century was Cantor’s discovery, as part of his pioneering systematic study of infinite sets, that there are different “sizes” of infinite sets. Cantor’s ideas were only gradually accepted by mathematicians, but eventually became one of the foundational cornerstones of modern analysis.

II.5.1. Cardinality

The term cardinality is used to denote the “size” of sets. This notion is most interesting and important for infinite sets; for finite sets it reduces to the “number of elements” in the set. While it is not essential for an analyst to have a detailed knowledge of the theory of cardinality, it is crucial to understand the most basic aspects of the theory: the distinction between finite and infinite sets, and between countable and uncountable sets. In this section, we will develop only enough of the general theory to establish what we need in analysis; a more complete treatment of the theory is in ( ).

The underlying idea is quite straightforward: two sets are the “same size” if there is a one-to-one correspondence (bijective function) between them.

II.5.1.1. Definition. Let $X$ and $Y$ be sets. $X$ and $Y$ have the same cardinality, or are equipotent (or equipollent), written $\text{card}(X) = \text{card}(Y)$ or $|X| = |Y|$, if there is a bijective function from $X$ to $Y$. We say $\text{card}(X) \leq \text{card}(Y)$ if $X$ is equipotent with a subset of $Y$, i.e. if there is an injective function from $X$ to $Y$. Write $\text{card}(X) < \text{card}(Y)$ if $\text{card}(X) \leq \text{card}(Y)$ and $\text{card}(X) \neq \text{card}(Y)$.

Note that we do not actually say what $\text{card}(X)$ is; we only use the term in comparing two sets. A definition of a cardinality (cardinal number) is given in ( ). Later in this section we will use particular symbols to denote the cardinality of certain standard sets, but even then no intrinsic meaning need be assigned to the cardinalities involved.

“Mathematicians are not interested in the nature of their objects ... or in their metaphysical character. What matters are the relations and operations applying to the objects ... just as chess players are not interested in what a bishop 'is' but in how it performs.”

A. Fraenkel
II.5.2. Finite and Infinite Sets

“The notion of infinity is our greatest friend; it is also the greatest enemy of our peace of mind.”

James Pierpont

“The infinite in mathematics is always unruly unless it is properly treated.”

Edward Kasner and James Newman

Infinite sets have some unusual properties which could be considered paradoxical, but with sufficient familiarity one can work with them comfortably; an analyst must develop a comfort level with them. One important feature of infinite sets is that they contain proper subsets which are as “large” as the whole set. For example, \( \mathbb{N} \) can be written as the disjoint union of the even and odd natural numbers. Each of these sets is intuitively only “half” of \( \mathbb{N} \), but on the other hand they are both actually equipotent with \( \mathbb{N} \): the functions \( f(n) = 2n \) and \( g(n) = 2n - 1 \) give bijections from \( \mathbb{N} \) onto the even and odd numbers, respectively. (The oldest surviving printed record of this observation is due to Galileo, although it would be a little surprising if it was not known in ancient Greece.)

This phenomenon cannot occur in a finite set. Intuitively, a finite set is one which can be “counted.” There are thus two standard ways of defining finite sets:

II.5.2.1. Definition. Let \( X \) be a set.

(i) \( X \) is finite if \( X = \emptyset \) or \( \text{card}(X) = \text{card}(\{m \in \mathbb{N} : 1 \leq m \leq n\}) \) for some \( n \in \mathbb{N} \). In the latter case, we say \( \text{card}(X) = n \); and \( \text{card}(\emptyset) = 0 \).

(ii) \( X \) is Dedekind-finite if it is not equipotent with any proper subset of itself.

II.5.2.2. We can show (II.5.2.3.) that the two conditions are equivalent (technically, we must assume some version of the Axiom of Choice; the Countable AC suffices). A set which is not finite (in sense (i)) is called infinite. A set which is equipotent with a finite [infinite] set is finite [infinite]. It is intuitively obvious that any subset of a finite set is finite (see Exercise () for a proof); and hence any set which has an infinite subset is infinite.

We now carefully state the equivalence between the two definitions of a finite set. It is convenient to phrase the equivalence for infinite sets rather than for finite sets; the statement can be easily converted into the equivalence for finite sets.

II.5.2.3. Proposition. Let \( X \) be a set. Then the following are equivalent:

(i) \( X \) is infinite.

(ii) There is an injective function from \( X \) to \( X \) which is not surjective (i.e. \( X \) is not Dedekind-finite).

(iii) There is a surjective function from \( X \) to \( X \) which is not injective.

The proof, which uses II.5.2.5. below, is left to the exercises. The implication (ii) \( \Rightarrow \) (i) is Exercise ()(c), (iii) \( \Rightarrow \) (i) is Exercise ()(e), and (i) \( \Rightarrow \) (ii) and (i) \( \Rightarrow \) (iii) are Exercise ()


149
II.5.2.4. The most obvious and important example of an infinite set is \( \mathbb{N} \). (It is a fundamental property of \( \mathbb{N} \) that it is not equipotent with \( \{1, \ldots, n\} \) for any \( n \). Although this seems obvious, it must in fact be proved; see Exercise (.) \( \mathbb{N} \) is the “smallest” infinite set in the following sense:

II.5.2.5. **Proposition.** Let \( X \) be a set. Then \( X \) is infinite if and only if \( X \) has a subset equipotent with \( \mathbb{N} \), i.e. if and only if there is an injective function from \( \mathbb{N} \) to \( X \) (or, in other words, a sequence of distinct elements in \( X \)).

**Proof:** If there is an injective function from \( \mathbb{N} \) to \( X \), then \( X \) is infinite. Conversely, suppose \( X \) is infinite. We will use an inductive procedure to define an injective function \( f : \mathbb{N} \to X \). Let \( f(1) \) be any element of \( X \) (\( X \neq \emptyset \) since \( X \) is infinite.) Suppose \( f(1), \ldots, f(n) \) have been defined to all be distinct. \( X \setminus \{f(1), \ldots, f(n)\} \neq \emptyset \), since otherwise \( X \) would be equipotent with \( \{1, \ldots, n\} \) and thus finite. So choose \( f(n+1) \) to be any element of \( X \setminus \{f(1), \ldots, f(n)\} \) (the Axiom of Choice, or at least the Axiom of Dependent Choice, is needed to do this properly; in fact, the result depends on the Countable AC – see Exercises II.5.6.14. and II.6.4.9.) \( \Box \)

Another Approach

We describe an alternate approach to finite sets which does not presuppose the availability of \( \mathbb{N} \). This approach is perhaps less satisfying since it does not really specify what finite sets are, but it avoids certain circularities in definitions. It can be shown (II.5.6.16.) that the sets which are finite according to this definition are precisely the sets which are finite according to Definition II.5.2.1. This approach was first taken by Sierpiński [].

The following definition is quite reminiscent of Peano’s axiom P5 (III.2.2.1.).

II.5.2.6. **Definition.** If \( X \) is a set, the **finite subsets** of \( X \) are the elements of the smallest subset \( \mathcal{F}(X) \) of \( \mathcal{P}(X) \) satisfying:

(i) \( \emptyset \in \mathcal{F}(X) \).

(ii) \( \{x\} \in \mathcal{F}(X) \) for all \( x \in X \).

(iii) If \( A, B \in \mathcal{F}(X) \), then \( A \cup B \in \mathcal{F}(X) \).

\( X \) is a **finite set** if \( X \in \mathcal{F}(X) \).

For any \( X \), there is such a smallest subset \( \mathcal{F}(X) \) of \( \mathcal{P}(X) \), the intersection of all subsets of \( \mathcal{P}(X) \) satisfying (i)–(iii) (this is well defined within ZF set theory).

II.5.2.7. **Proposition.** Let \( Y \subseteq X \). Then \( \mathcal{F}(Y) = \{A \cap Y : A \in \mathcal{F}(X)\} \).

**Proof:** Let \( \mathcal{S} = \{A \cap Y : A \in \mathcal{F}(X)\} \subseteq \mathcal{P}(Y) \). Then it is easily checked that \( \mathcal{S} \) satisfies (i)–(iii) (with \( X \) replaced by \( Y \)), and hence \( \mathcal{F}(Y) \subseteq \mathcal{S} \). On the other hand, set

\[
\mathcal{T} = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\} \subseteq \mathcal{P}(X) .
\]

It is again easily checked that \( \mathcal{T} \) satisfies (i)–(iii), and hence \( \mathcal{F}(X) \subseteq \mathcal{T} \). If \( C \in \mathcal{P}(X) \), then \( C \) can be uniquely written as \( A \cup B \) with \( A \in \mathcal{P}(Y) \) and \( B \in \mathcal{P}(X \setminus Y) \), namely \( A = C \cap Y \) and \( B = C \cap (X \setminus Y) \); if \( C \in \mathcal{F}(X) \), then \( C \in \mathcal{T} \), and hence \( A = C \cap Y \in \mathcal{F}(Y) \). Thus \( \mathcal{S} \subseteq \mathcal{F}(Y) \). \( \Box \)
II.5.2.8. Proposition. If $A \in \mathcal{F}(X)$ and $B \subseteq A$, then $B \in \mathcal{F}(X)$. In particular, every subset of a finite set is finite.

Proof: Let $\mathcal{S} = \{A \in \mathcal{P}(X) : B \in \mathcal{F}(X) \text{ for all } B \subseteq A\}$. Then $\mathcal{S} \subseteq \mathcal{F}(X)$. It is easily checked that $\mathcal{S}$ satisfies (i)–(iii), so $\mathcal{F}(X) \subseteq \mathcal{S}$. 

II.5.2.9. Corollary. If $Y \subseteq X$, then $\mathcal{F}(Y) \subseteq \mathcal{F}(X)$, i.e. every subset of $Y$ which is a finite subset of $Y$ is a finite subset of $X$.

Proof: Let $B \in \mathcal{F}(Y)$. Then $B = Y \cap A$ for some $A \in \mathcal{F}(X)$. Thus $B \subseteq A$, so $B \in \mathcal{F}(X)$. 

II.5.2.10. Corollary. If $Y \subseteq X$, then $\mathcal{F}(X) = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\}$.

Proof: Set $\mathcal{T} = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\}$. The proof of II.5.2.7. shows that $\mathcal{F}(X) \subseteq \mathcal{T}$. Conversely, $\mathcal{F}(Y)$ and $\mathcal{F}(X \setminus Y)$ are contained in $\mathcal{F}(X)$ by II.5.2.9., and hence $\mathcal{T} \subseteq \mathcal{F}(X)$ by property (iii).

II.5.2.11. Proposition. Let $X$ and $Y$ be sets and $f : X \to Y$ a function. If $A \in \mathcal{F}(X)$, then $f(A) \in \mathcal{F}(Y)$, i.e. the image of a finite set is finite.

Proof: Let $\mathcal{S} = \{B \subseteq X : f(B) \in \mathcal{F}(Y)\}$. Then it is easily checked that $\mathcal{S}$ satisfies (i)–(iii); hence $\mathcal{F}(X) \subseteq \mathcal{S}$. 

II.5.2.12. Proposition. Let $X$ be a totally ordered set. Then every nonempty finite subset of $X$ has a largest element and a smallest element.

Proof: Let $\mathcal{S}$ be the collection of subsets of $X$ consisting of $\emptyset$ and all nonempty subsets of $X$ which have a largest element and a smallest element. It is easily checked that $\mathcal{S}$ satisfies (i)–(iii). Thus $\mathcal{F}(X) \subseteq \mathcal{S}$. 

II.5.3. Countable and Uncountable Sets

Until the work of Cantor, it was commonly believed by mathematicians that all infinite sets (if such sets exist at all; some mathematicians, both before and after Cantor, rejected infinite sets entirely) were essentially the same (although even in ancient Greece a distinction was made between the “infinite” and the “infinitesimal”). But in fact there are sets which are not only infinite, but “very infinite” in the sense of having cardinality greater than $\text{card}(\mathbb{N})$. We explain this surprising fact in this section.
II.5.3.1. **Definition.** A set $X$ is *countable* if $\text{card}(X) \leq \text{card}(\mathbb{N})$. $X$ is *countably infinite* if $\text{card}(X) = \text{card}(\mathbb{N})$. A (necessarily infinite) set $X$ is *uncountable* if $\text{card}(X) > \text{card}(\mathbb{N})$.

A set $X$ is uncountable if there are “too many” elements in $X$ to list in a sequence, or in other words if any sequence in $X$ necessarily leaves out some elements of $X$. Any subset of a countable set is countable; thus if a set $X$ has an uncountable subset, then $X$ itself is uncountable.

II.5.3.2. The symbol $\aleph_0$, pronounced “aleph zero”, “aleph naught”, or “aleph null”, is commonly used to denote the cardinality of $\mathbb{N}$. ($\aleph$, “aleph”, is the first letter of the Hebrew alphabet.) Thus $X$ is countable if $X$ is finite or $\text{card}(X) = \aleph_0$.

It is not obvious that uncountable sets exist, but the following famous theorem of **Cantor** shows that they do exist in abundance:

II.5.3.3. **Theorem.** Let $X$ be any set. Then $\text{card}(\mathcal{P}(X)) > \text{card}(X)$.

**Proof:** It is easy to see that $\text{card}(X) \leq \text{card}(\mathcal{P}(X))$, since there is an obvious one-one correspondence between $X$ and the set of singleton subsets of $X$. So we need to show that $X$ and $\mathcal{P}(X)$ are not equipotent. Suppose $f$ is a function from $X$ to $\mathcal{P}(X)$; we will show that $f$ cannot be surjective, which will prove the theorem. If $x \in X$, then $f(x)$ is a subset of $X$, so it makes sense to ask whether $x \in f(x)$. Define

$$Y = \{x \in X : x \notin f(x)\} \subseteq X.$$ 

We claim there can be no $y \in X$ with $Y = f(y)$. Suppose there is; is $y \in Y$? If $y \in Y$, then $y \in f(y)$ since $Y = f(y)$ by assumption; but then $y \notin Y$ by definition of $Y$, a contradiction. Similarly, if $y \notin Y$, then $y \in f(y)$, so $y \in Y$, another contradiction. Thus the assumption that $Y = f(y)$ for some $y \in X$ necessarily leads to a contradiction, so there is no such $y$.

II.5.3.4. This proof was adapted by **Bertrand Russell** to establish his Paradox (II.5.6.13.) (Zermelo actually discovered this paradox before Russell, using a similar argument). A simple version of this paradox based on II.5.3.3. is the following: if there is a set $X$ of all sets, then every subset of $X$ is a set, hence a member of $X$; so $\mathcal{P}(X) \subseteq X$. But it is then easy to construct a surjective function from $X$ to $\mathcal{P}(X)$ (cf. II.5.4.1.), contradicting II.5.3.3.

A similar, but more sophisticated, type of “self-referential” argument was used by **Kurt Gödel** in his famous *Incompleteness Theorem* (see [?]) or [Ruc82] for mathematically accurate popular accounts of this theorem; for the full story see [Smm92].

II.5.3.5. **Corollary.** $\mathcal{P}(\mathbb{N})$ is an uncountable set.

Since $\text{card}(\mathcal{P}(\mathbb{N})) = \text{card}([0,1]^{\mathbb{N}})$, we obtain that there are uncountably many sequences of 0’s and 1’s. Because of this, the symbol $2^{\mathbb{N}}$ is often used for the cardinality of $\mathcal{P}(\mathbb{N})$ (the notation is explained in ). From this we can obtain a proof (see Exercise ) of the following theorem, which is one of the most important facts in analysis. We instead give a somewhat different proof, using a “diagonalization argument” of Cantor (the proof is not really different; see II.5.6.9. For another version of the same proof, see II.5.6.10.)
II.5.3.6. Theorem. \( \mathbb{R} \) is uncountable. In fact, every nonempty open interval in \( \mathbb{R} \) is uncountable.

Proof: It suffices to show that \([0,1)\) is uncountable, since every open interval contains a half-open subinterval, which is equipotent with \([0,1)\) via a linear function. Every number in \([0,1)\) has a unique decimal expansion

\[ a_1a_2a_3 \cdots \]

where \(0 \leq a_n \leq 9\) for each \(n\), which does not end in an infinite string of 9's.

Suppose we consider a sequence \((x_n)\) in \([0,1)\). We will construct an \(x \in [0,1)\) which is not equal to any \(x_n\); this will prove that \([0,1)\) is uncountable since no sequence can exhaust the set. Write \(x_n\) in its decimal expansion

\[ x_n = a_{n1}a_{n2}a_{n3} \cdots \]

not ending in a string of 9's, where \(a_{nm}\) is the \(m\)'th decimal digit of \(x_n\). Now for each \(n\) let \(b_n\) be a digit which is not 9 and which is different from \(a_{nn}\) (to be precise, we can let \(b_n = 0\) if \(a_{nn} \neq 0\) and \(b_n = 1\) if \(a_{nn} = 0\)), and set \(x = .b_1b_2b_3 \cdots\). Then, for any \(n\), \(x \neq x_n\) since it differs from \(x_n\) in at least the \(n\)'th decimal place.

II.5.3.7. The cardinality of \(\mathbb{R}\) is often called the cardinality of the continuum, denoted \(\mathfrak{c}\). In fact, \(\text{card}(\mathbb{R}) = \text{card}(\mathcal{P}(\mathbb{N}))\) (for \(\mathbb{N}\) being the set of natural numbers), so \(\mathfrak{c} = 2^{\aleph_0}\). Every nontrivial interval in \(\mathbb{R}\) also has cardinality \(\mathfrak{c}\) by Exercise ().

II.5.3.8. From Theorem II.5.3.3., it follows that not only are there uncountable sets, but there are uncountable sets of different sizes: for example, \(\text{card}(\mathcal{P}(\mathbb{R})) = \text{card}(\mathcal{P}(\mathcal{P}(\mathbb{N}))) > \text{card}(\mathcal{P}(\mathbb{N})) = 2^{\aleph_0}\) (i.e. \(\mathcal{P}(\mathbb{R})\) is "very uncountable.") The symbol \(2^\mathfrak{c}\) or \(2^{\aleph_0}\) is used for \(\text{card}(\mathcal{P}(\mathbb{R}))\). In fact, there is a never-ending hierarchy of infinite cardinalities:

\[ \text{card}(\mathbb{N}) < \text{card}(\mathbb{R}) < \text{card}(\mathcal{P}(\mathbb{R})) < \text{card}(\mathcal{P}(\mathcal{P}(\mathbb{R}))) < \text{card}(\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{R})))) < \cdots \]

and there is no "largest cardinality." (There are even cardinalities much larger than all the ones on this list!)

II.5.3.9. Note that II.5.3.3. is valid also in the case where \(X\) is a finite set. In this case, it amounts to the statement that \(2^n > n\) for every \(n \geq 0\).

II.5.4. Properties of Countable Sets

We will now analyze countable sets in more detail. We first establish a useful criterion for countability:

II.5.4.1. Proposition. Let \(X\) be a nonempty set. Then \(X\) is countable if and only if there is a surjective function from \(\mathbb{N}\) to \(X\).

Proof: First assume that \(X\) is countable. By definition, there is an injective function \(f : X \to \mathbb{N}\). Fix an \(x_0 \in X\), and let \(S = f(X) \subseteq \mathbb{N}\). Define a function \(g : \mathbb{N} \to X\) by \(g(n) = x\) if \(n \in S\) and \(f(x) = n\) (this is well defined since \(f\) is injective), and set \(g(n) = x_0\) if \(n \notin S\). Then \(g\) is a surjective function from \(\mathbb{N}\) to \(X\).

Conversely, suppose \(g : \mathbb{N} \to X\) is a surjective function. For each \(x \in X\), let \(f(x)\) be the smallest number in \(g^{-1}(\{x\})\). Then \(f\) is an injective function from \(X\) to \(\mathbb{N}\), so \(\text{card}(X) \leq \text{card}(\mathbb{N})\).
II.5.4.2. Corollary. If $X$ and $Y$ are sets and there is a surjective function $f : X \to Y$, and $X$ is countable, then $Y$ is countable.

It seems obvious that if $X$ is countable, then either $X$ is finite or $\text{card}(X) = \text{card}(\mathbb{N})$, and hence the term “countably infinite” means precisely “countable and infinite”; however, this takes a little proof:

II.5.4.3. Proposition. Let $X$ be a countable set. Then either $X$ is finite, or $\text{card}(X) = \text{card}(\mathbb{N})$.

This would seem to follow from II.5.2.5., which says that $\text{card}(\mathbb{N}) \leq \text{card}(X)$ if $X$ is infinite, and $\text{card}(X) \leq \text{card}(\mathbb{N})$ by definition if $X$ is countable. However, it is not obvious that these two inequalities together imply that $\text{card}(X) = \text{card}(\mathbb{N})$ (that is, that the “ordering” on cardinalities is antisymmetric.) This antisymmetry is true in general by the Schröder-Bernstein Theorem, but we can and will take a simpler approach in the countable case.

Proof: If $X = \emptyset$, the conclusion is true since $\emptyset$ is finite by definition. Assume $X$ is nonempty. By II.5.4.1., there is a surjective function $g : \mathbb{N} \to X$. Inductively define a function $f : X \to \mathbb{N}$ as follows. Let $x_1 = g(1)$, and set $f(x_1) = 1$. If $x_1, \ldots, x_n$ have been defined with $f(x_k) = k$ for $1 \leq k \leq n$, then either $X = \{x_1, \ldots, x_n\}$, in which case $X$ is finite, or $Y_n = X \setminus \{x_1, \ldots, x_n\} \neq \emptyset$. In the latter case, let $k$ be the smallest element of $\mathbb{N}$ with $g(k) \in Y_n$ (there is such a $k$ since $g$ is surjective.) Set $x_{n+1} = g(k)$ and $f(x_{n+1}) = k$. If this process terminates, then $X$ is finite; and if it never terminates, then $f$ is an injective function from $X$ onto $\mathbb{N}$. ✖️

Another useful fact is that $\mathbb{N}^2 = \mathbb{N} \times \mathbb{N}$ is equipotent with $\mathbb{N}$. This seems a little surprising, but is easily proved:

II.5.4.4. Proposition. $\mathbb{N}^2$ is equipotent with $\mathbb{N}$, and hence is countable (countably infinite.)

Proof: The function $(n, m) \mapsto 2^{n-1}(2m - 1)$ is a bijection from $\mathbb{N}^2$ to $\mathbb{N}$, as is easily checked using the unique factorization in $\mathbb{N}$. ✖️

II.5.4.5. Corollary. A Cartesian product of a finite number of countable sets is countable.

II.5.4.6. It is not true that a Cartesian product of a countable number of countable sets is countable; for example, $\{0, 1\}^\mathbb{N}$ is uncountable .

There are some other very important consequences of II.5.4.4.:

II.5.4.7. Corollary. $\mathbb{Z}$ and $\mathbb{Q}$ are countable.

Proof: To show that $\mathbb{Z}$ is countable, define a function $f : \mathbb{N}^2 \to \mathbb{Z}$ by $f(n, m) = (-1)^n(m - 1)$; then $f$ is surjective, so $\mathbb{Z}$ is countable by II.5.4.4. and II.5.4.2.. Thus $\mathbb{Z} \times \mathbb{N}$ is countable. Define $g : \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ by $g(n, m) = n/m$, and apply II.5.4.2. again to conclude that $\mathbb{Q}$ is countable. ✖️
II.5.4.8. **Corollary.** A countable union of countable sets is countable. More precisely, let \( \{ X_i : i \in I \} \)
be an indexed collection of sets. If each \( X_i \) is countable and \( I \) is countable, then \( \bigcup_{i \in I} X_i \) is countable.

**Proof:** We may assume each \( X_i \) is nonempty by deleting the empty ones. For each \( i \) let \( f_i \) be a surjective function from \( \mathbb{N} \) onto \( X_i \). Then \( (i, n) \mapsto f_i(n) \) is a surjective function from \( I \times \mathbb{N} \) to \( \bigcup_{i \in I} X_i \). Apply II.5.4.5. and II.5.4.2. (This argument requires the Countable Axiom of Choice.)

There is also a version of these facts for the continuum:

II.5.4.9. **Proposition.**

(i) If \( X \) and \( Y \) have cardinality \( \leq 2^{\aleph_0} \), then \( \text{card}(X \times Y) \leq 2^{\aleph_0} \).

(ii) Let \( \{ X_i : i \in I \} \) be an indexed collection of sets. If \( \text{card}(X_i) \leq 2^{\aleph_0} \) for all \( i \) and \( \text{card}(I) \leq 2^{\aleph_0} \), then \( \text{card}(\bigcup_{i \in I} X_i) \leq 2^{\aleph_0} \).

**Proof:** It suffices to prove (i), since then (ii) follows as in the proof of II.5.4.8. (the full Axiom of Choice is needed). To prove (i), we need only show that \( \text{card}(\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})) = 2^{\aleph_0} \). But there is a bijective function from \( \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N}) \) to \( \mathcal{P}(\mathbb{N}) \) given by \( (A, B) \mapsto f(A) \cup g(B) \), where \( f(n) = 2n \) and \( g(n) = 2n - 1 \) are bijections from \( \mathbb{N} \) onto the even and odd numbers.

This result is actually true for any infinite cardinal (\( \) (but depends on the Axiom of Choice).

**Countable Ordered Sets**

II.5.4.10. The ordered set \( \mathbb{Q} \) can be characterized by a few simple properties. \( \mathbb{Q} \) satisfies:

(i) \( \mathbb{Q} \) is countable.

(ii) \( \mathbb{Q} \) is totally ordered.

(iii) \( \mathbb{Q} \) has no largest or smallest element.

(iv) \( \mathbb{Q} \) has no consecutive elements: if \( a, b \in \mathbb{Q} \), \( a < b \), then there is a \( c \in \mathbb{Q} \) with \( a < c < b \).

Property (iv) is sometimes phrased by saying that the ordering on \( \mathbb{Q} \) is dense. This condition implies, by repeated application, that between any two elements of \( \mathbb{Q} \) there are actually infinitely many other elements.

It turns out that these conditions completely characterize \( \mathbb{Q} \) as an ordered set:
II.5.4.11. Theorem. Let \((X, \preceq)\) be a countable totally ordered set with no consecutive elements and no first or last element. Then \((X, \preceq)\) is order-isomorphic to \(Q\).

Proof: Let \((x_n : n \in N)\) be an enumeration of \(X\), and \((y_n : n \in N)\) an enumeration of \(Q\). We will inductively construct an order-isomorphism \(f\) of \(X\) onto \(Q\). Begin by defining \(f(x_1) = y_1\). Now suppose \(f\) has been defined on \(x_1, \ldots, x_n\) and possibly finitely many other \(x_i\), and \(y_1, \ldots, y_n\) are in the range of the part of \(f\) already defined. We will define \(f\) on \(x_{n+1}\) and possibly on another \(x_i\), so that \(y_{n+1}\) is in the range. Let \(X_n\) be the set of \(x_i\) for which \(f\) has already been defined, and \(Y_n = f(X_n)\).

First define \(f\) on \(x_{n+1}\). If \(x_{n+1} \in X_n\), \(f(x_{n+1})\) is already defined and we skip this step. Otherwise, note that either \(x_{n+1} \prec x_i\) for all \(x_i \in X_n\), or there is a largest \(x_i \in X_n\) with \(x_i \prec x_{n+1}\); let \(a\) be this \(x_i\). Similarly, either \(x_j \prec x_{n+1}\) for all \(x_j \in X_n\), or there is a smallest \(x_j\) with \(x_{n+1} \prec x_j\); let \(b\) be this \(x_j\). Let \(k\) be the smallest index for which \(y_k\) satisfies \(f(a) < y_k < f(b)\) (or \(y_k < f(b)\) if \(a\) is undefined, \(f(a) < y_k\) if \(b\) is undefined.) We automatically have \(y_k \notin Y_n\). Set \(f(x_{n+1}) = y_k\).

Then work backwards in a similar way for the other part of the inductive step. If \(y_{n+1}\) is already in the range of \(f\), skip this step. Otherwise, let \(c\) and \(d\) be the largest and smallest elements of the range of \(f\) (as defined so far) with \(c < y_{n+1} < d\) (either \(c\) or \(d\) may be undefined.) Then \(c = f(x_i)\) and \(d = f(x_j)\) for some \(i, j\). Let \(k\) be the smallest index for which \(x_i \prec x_k \prec x_j\); then \(f(x_k)\) is undefined and so we can define \(f(x_k) = y_{n+1}\).

It is easy to check that the \(f\) thus defined is a bijection and an order-isomorphism.

II.5.4.12. A similar argument shows that if \((X, \preceq)\) is a countable totally ordered set with no consecutive elements, but with a first or last element, then \((X, \preceq)\) is order-isomorphic to \(Q \cap [0, 1)\), \(Q \cap (0, 1]\), or \(Q \cap [0, 1]\).

II.5.4.13. Corollary. Every countable totally ordered set is order-isomorphic to a closed subset of \(Q\).

Proof: Let \((X, \preceq)\) be a countable totally ordered set. If \(x \in X\), define \(Y_x \subseteq Q\) by

- \(Y_x = \{0\}\) if \(x\) has no immediate predecessor or successor.
- \(Y_x = Q \cap (-\infty, 0]\) if \(x\) has an immediate predecessor but no immediate successor.
- \(Y_x = Q \cap [0, \infty)\) if \(x\) has an immediate successor but no immediate predecessor.
- \(Y_x = Q\) if \(x\) has an immediate predecessor and an immediate successor.

Let \(Z = \{(x, y) : y \in Y_x\} \subseteq X \times Q\). Give \(X \times Q\) the lexicographic ordering, and give \(Z\) the relative ordering. It is easily checked that \(Z\) with this ordering is countable, totally ordered, and has no consecutive elements or largest or smallest element. Thus \(Z\) is order-isomorphic to \(Q\). \((X, \preceq)\) is order-isomorphic to the subset \(X \times \{0\}\) of \(Z\). It is easy to check that if \((x_n)\) is an increasing sequence in \(X\) with supremum \(x\), then \((x, 0)\) is the supremum of \(((x_n, 0))\) in \(Z\) and conversely, and similarly for decreasing sequences, so \(X \times \{0\}\) is closed in \(Z \cong Q\) by (A).

Note that we could have simplified the argument to just consider \((X, \preceq) \cong X \times \{0\}\) as a subset of \(X \times Q\) with lexicographic ordering if we did not want to require that the image of \(X\) in \(Q\) is closed.
II.5.5. Additional Facts about Cardinality

There are three rather natural questions about cardinality which turn out to have interesting answers:

II.5.5.1. Questions.

(1) If $X$ and $Y$ are sets with $\text{card}(X) \leq \text{card}(Y)$ and $\text{card}(Y) \leq \text{card}(X)$, is $\text{card}(X) = \text{card}(Y)$, i.e. if each of $X$ and $Y$ can be put in bijective correspondence with a subset of the other, are $X$ and $Y$ equipotent? (Is the “ordering” on cardinalities a true partial order, i.e. antisymmetric?)

(2) If $X$ and $Y$ are sets, is it always true that $\text{card}(X) \leq \text{card}(Y)$ or $\text{card}(Y) \leq \text{card}(X)$, i.e. are any two sets comparable in size? (Is the partial ordering on cardinalities a total ordering?)

(3) Is $\mathfrak{c} = 2^{\aleph_0}$ the smallest cardinality larger than $\aleph_0$, i.e. if $X \subseteq \mathbb{R}$, is it true that either $X$ is countable or $\text{card}(X) = \text{card}(\mathbb{R})$? (The i.e. statement is only equivalent to the first statement if the answer to question (2) is yes for subsets of $\mathbb{R}$.)

It turns out that only question 1 has a definitive answer, given by the Schröder-Bernstein Theorem:

II.5.5.2. Theorem. Let $X$ and $Y$ be sets. If $\text{card}(X) \leq \text{card}(Y)$ and $\text{card}(Y) \leq \text{card}(X)$, then $\text{card}(X) = \text{card}(Y)$.

A proof is outlined in Exercise II.5.6.11. Actually, the commonly-used name “Schröder-Bernstein Theorem” is not really accurate, since the proof published by SCHRÖDER in 1898 was incorrect; the proof published by F. BERNE Stein the same year (he was 19 years old when he did it) was the first and only correct proof at that time and is the one still commonly used. The theorem is sometimes called the “Cantor-Bernstein Theorem” since CANTOR stated it in his book without proof; apparently CANTOR tried rather hard to prove it, without success. There are indications that DEDEKIND found a proof in the 1880’s, but never published it.

II.5.5.3. Question 2 turns out to be dependent on the Axiom of Choice, which is independent of the other (ZF) axioms of set theory. If the Axiom of Choice is assumed, as is usual in analysis, then a positive answer to question 2 can be obtained; see II.9.7.1.

II.5.5.4. The conclusion of Question 3, that $\mathfrak{c}$ is the smallest uncountable cardinality, is called the Continuum Hypothesis. This is also known to be independent of the other axioms of set theory, including the Axiom of Choice (ZFC), through work of KURT GÖDEL and PAUL COHEN. There is also a Generalized Continuum Hypothesis which is independent of ZFC.

II.5.6. Exercises

II.5.6.1. Let $f$ be an injective function from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Inductively define a function $g$ from an initial segment of $\mathbb{N}$ to $\{1, \ldots, n\}$ as follows. Let $k_1$ be the smallest element of $S = f(\{1, \ldots, n\})$, and set $g(1) = f^{-1}(k_1)$. Inductively let $k_{r+1}$ be the smallest element of $S \setminus \{k_1, \ldots, k_r\}$ if this set is nonempty, and $g(r+1) = f^{-1}(k_{r+1})$. 157
(a) Show that \( g \) is defined precisely on \( \{1, \ldots, n\} \), and that \( g \) maps \( \{1, \ldots, n\} \) bijectively to itself.
(b) Let \( h = f \circ g \). Show that \( h \) is an injective function from \( \{1, \ldots, n\} \) onto \( S \), and prove inductively that \( h \)
is increasing and \( h(k) \geq k \) for every \( k \in \{1, \ldots, n\} \).
(c) On the other hand, \( h(n) \leq n \) since \( S \subseteq \{1, \ldots, n\} \). Conclude that \( h(k) = k \) for every \( k \), and thus that \( h \)
is surjective, i.e. that \( f \) is surjective. [Show inductively that if \( h(k_0) > k_0 \) for some \( k_0 \), then \( h(k) > k \) for all \( k \).
(d) Use this to show that an injective function from a finite set to itself is surjective.
(c) A real or complex number is algebraic if it is a root of a nonzero polynomial with rational coefficients (which can be assumed to be integers by clearing denominators.) Show that the set of algebraic numbers is countable.

II.5.6.5. Let \( S \) be a subset of \( \{1, \ldots, n\} \). Define a function \( f \) from \( S \) to \( \mathbb{N} \) inductively by letting \( k_1 \) be the smallest element of \( S \) and setting \( f(k_1) = 1 \), and inductively letting \( k_{r+1} \) being the smallest element of \( S \setminus \{k_1, \ldots, k_r\} \) if this set is nonempty, and \( f(k_{r+1}) = r + 1 \).

(a) Prove by induction that \( f(k) \leq k \) for every \( k \in S \).
(b) Conclude that \( f(S) = \{1, \ldots, m\} \) for some \( m \leq n \), and hence that \( S \) is finite.
(c) Show that a subset of a finite set is finite.
(d) Show that if \( X \) and \( Y \) are finite sets with \( card(X) = n \) and \( card(Y) = m \) for some \( n, m \in \mathbb{N} \cup \{0\} \), then \( card(Y) \leq card(X) \) if and only if \( m \leq n \), so the “ordering” on cardinalities of finite sets is consistent with the ordering on \( \mathbb{N} \cup \{0\} \). In particular, if \( Y = X \), then \( m = n \), i.e. the cardinality of a finite set is a well-defined nonnegative integer.

II.5.6.6. Let \( X \) be an infinite set, and \( Y \) a countable set. Prove that \( card(X \cup Y) = card(X) \). [Use II.5.6.5.]

II.5.6.7. Show that the set of sequences in \( \mathbb{Q} \) with only finitely many nonzero terms is countable. [Use II.5.4.5. and II.5.4.8.]

II.5.6.8. (a) Prove that the set of polynomials with rational coefficients is countable. [Use Exercise II.5.6.5.]
(b) A real or complex number is algebraic if it is a root of a nonzero polynomial with rational coefficients (which can be assumed to be integers by clearing denominators.) Show that the set of algebraic numbers is countable.
(c) A real or complex number is transcendental if it is not algebraic. Prove that there exists a transcendental real number (and, in fact, the cardinality of the set of transcendental real numbers is \( \aleph_0 \) by Exercise II.5.6.4.)
II.5.6.7.  (a) Prove that any two bounded open intervals \((a,b)\) and \((c,d)\) in \(\mathbb{R}\) are equipotent by finding a bijection [consider linear functions.]
(b) Use the tangent function to show that \((-\frac{\pi}{2}, \frac{\pi}{2})\) is equipotent with \(\mathbb{R}\).
(c) Give two proofs that \(\text{card}([0,1]) = \text{card}([0,1]) = \text{card}([0,1])\), one using the Schröder-Bernstein Theorem and the other by directly finding bijections [use II.5.2.5.]
(d) Show that all intervals in \(\mathbb{R}\) have the same cardinality.

II.5.6.8.  (a) Define a function \(f : \{0,1\}^\mathbb{N} \to [0,1]\) by \(f(a_1,a_2,a_3,\ldots) = .a_1a_2a_3\ldots\), where the last expression is the decimal expansion of a number in \([0,1]\). Show that \(f\) is injective, and hence \(\text{card}([0,1]) \geq \text{card}(\{0,1\}^\mathbb{N}) = 2^{\aleph_0}\), so \([0,1]\) is uncountable.
(b) Define a function \(g : \mathbb{R} \to \{0,1\}^\mathbb{N}\) by \(g(x) = (a_1,a_2,\ldots)\), where \(x = .a_1a_2\ldots\) is the binary expansion of \(x\) not ending in a string of 1’s. Show that \(g\) is injective, and hence \(\text{card}([0,1]) \leq \text{card}([0,1]^\mathbb{N}) = 2^{\aleph_0}\).
(c) Apply the Schröder-Bernstein Theorem to conclude that \(\text{card}([0,1]) = 2^{\aleph_0}\).
(d) As an alternative to (c), \([0,1]^\mathbb{N}\setminus g([0,1])\) consists precisely of all sequences which are eventually 1. Show that this set is countable (see Exercise II.5.6.5.), and hence \(\text{card}([0,1]^\mathbb{N}) = \text{card}(g([0,1]) = \text{card}([0,1]) = \text{card}(\mathbb{R})\) by Exercises II.5.6.4. and II.5.6.7.

II.5.6.9.  Show that the proof of II.5.3.3. is really a diagonalization argument along the lines of the proof of II.5.3.6., as follows.
(a) Let \(X\) be a set, and \(f\) a function from \(X\) to \(\mathcal{P}(X)\). Define a subset \(Y\) of \(X\) as follows: for \(x \in X\), let \(x \in Y\) if and only if \(x \notin f(x)\).
(b) Show that \(Y\) is not in the range of \(f\), i.e. \(Y\) is not in the “list” \(\{f(x) : x \in X\}\) of subsets of \(X\) defined by \(f\), since for each \(x \in X\) the set \(Y\) differs from \(f(x)\) at least with regard to membership of \(x\) (i.e. for each \(x \in X\), \(x\) belongs to the symmetric difference \(Y \triangle f(x)\)).

II.5.6.10.  This exercise outlines a different proof that \(\mathbb{R}\) is uncountable, based on the Nested Intervals Theorem (). This was essentially CANTOR’s original proof. We show that \([0,1]\) is uncountable.
(a) Let \((x_n)\) be a sequence in \([0,1]\). We will show that there is an \(x \in [0,1]\) with \(x \neq x_n\) for all \(n\).
(b) Let \(I_0 = [0,1]\). Inductively define \(I_n\) as follows. Suppose \(I_n\) has been defined. Subdivide \(I_n\) into three closed subintervals \(I_{n,1}, I_{n,2}, I_{n,3}\) from left to right, of equal length \(3^{-(n-1)}\), meeting only at endpoints. Then there is at least one \(k\) with \(x_{n+1} \notin J_{n,k}\); let \(I_{n+1}\) be the leftmost such \(J_{n,k}\).
(c) Set \(S = \cap_{n=1}^{\infty} I_n\). Then \(S \neq \emptyset\) by the Nested Intervals Theorem. If \(x \in S\), then \(x \neq x_n\) for all \(n\). (Actually \(S = \{x\}\).)

II.5.6.11.  This exercise outlines a proof of the Schröder-Bernstein Theorem. Suppose \(X\) and \(Y\) are sets, and there are injective functions \(f : X \to Y\) and \(g : Y \to X\). We will construct a bijection \(h\) from \(X\) to \(Y\).
Let \(X_1 = X \setminus g(Y)\) and \(Y_1 = Y \setminus f(X)\), and inductively define \(X_{n+1} = g(Y_n)\) and \(Y_{n+1} = f(X_n)\).
(a) Prove by induction on \(n\) that the \(X_n\) are disjoint subsets of \(X\) and the \(Y_n\) are disjoint subsets of \(Y\).
(b) Let \(X_\infty = X \setminus (\cup_{n=1}^{\infty} X_n)\) and \(Y_\infty = Y \setminus (\cup_{n=1}^{\infty} Y_n)\). Show that \(f\) maps \(X_\infty\) onto \(Y_\infty\).
(c) Define \(h : X \to Y\) as follows. If \(x \in X_n\), \(n\) odd, or if \(x \in X_\infty\), let \(h(x) = f(x)\). If \(x \in X_n\), \(n\) even, let \(h(x) = g^{-1}(x)\). Show that \(h\) is a bijection from \(X\) to \(Y\).
II.5.6.12. Let $X$ be a set of cardinality $\kappa$. Assume that $\kappa^2 = \kappa \geq \aleph_0$ and that $X$ can be totally ordered. Fix a total order on $X$ and a one-one correspondence $\phi : X \to X^2$, and a one-to-one map $\psi : \mathbb{N} \to X$.

(a) Show that for $n > 2$, iteration of $\phi$ defines a one-one correspondence $\phi_n$ from $X$ to $X^n$. Let $\phi_2 = \phi$ and $\phi_1 : X \to X$ be the identity.
(b) Show that the $\phi_n$ define a one-one correspondence between $\mathbb{N} \times X$ and the disjoint union of the $X^n$.
(c) Use the fixed ordering on $X$ to identify the set of $n$-element subsets of $X$ with a subset of $X^n$. From (b) and (c) obtain a one-one correspondence between the set of finite subsets of $X$ and a subset of $\mathbb{N} \times X$. The function $\psi$ then gives a one-one correspondence between $\phi$ and a subset of $X^2$. Conclude that $|\Phi| \leq |X^2| = |X|$, so that $|\Phi| = |X|$ by the Schröder-Bernstein Theorem.

II.5.6.13. (Russell’s Paradox) This argument shows that there can be no such thing as the “set of all sets.”

(a) Say that a set $A$ is ordinary if $A$ is not a member of $A$. (Almost every set normally encountered in mathematics is an ordinary set by this definition.)
(b) Suppose $X$ is the set of all sets. Let $Y$ be the collection of all ordinary sets. Then $Y$ is a subset of $X$, hence a set. Is $Y$ an ordinary set?

II.5.6.14. Show without using any form of the Axiom of Choice that a set is Dedekind-finite if and only if it does not have any countably infinite subset. [One direction follows from II.5.2.3. Conversely, if there is an $f : X \to X$ which is injective but not surjective, fix $x \in X$ not in the range of $f$, and consider $x, f(x), f(f(x)), \ldots$. See Exercise II.6.4.9.]

II.5.6.15. (a) Show (without any form of AC) that if $X$ is any set, then there is a set $z$ with $z \notin X$.
[Use II.5.3.3. to show that $X \cap \mathcal{P}(X)$ is a proper subset of $\mathcal{P}(X)$.]
(b) Use (a) to give an alternate proof that there is no “set of all sets.”
(c) Show (without any form of AC) that if $X$ and $Y$ are sets, then there is a set $Y'$ which is equipotent with $Y$ and disjoint from $X$. [Let $z$ be a set which is not in $\mathcal{C}(X)$, and let $Y' = \{(z, y) : y \in Y\} = \{\{z\}, \{z, y\} : y \in Y\}$. If $x \in Y'$, then $z \in \mathcal{C}(x)$, but then $x \notin X$ by II.2.1.4.]

II.5.6.16. Show that a nonempty set $X$ is finite in the sense of II.5.2.6. if and only if $X$ is equipotent to $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$. [Show that the set $\mathcal{S}$ of $\emptyset$ and all sets equipotent to $\{1, \ldots, n\}$ for some $n$ satisfies (i)–(iii). For the converse, see III.2.6.10.]
II.6. The Axiom of Choice and Zorn’s Lemma

If \( \{X_i : i \in I\} \) is an indexed collection of nonempty sets, it would seem obvious that we could pick an element out of each set; after all, isn’t this what is meant by saying that each \( X_i \) is nonempty? However, with more thought, although we can “obviously” choose an element out of any individual \( X_i \) (but see II.6.1.6.), it becomes less obvious that we could simultaneously choose an element out of each \( X_i \). In fact, it can be neither proved nor disproved in ordinary (ZF) set theory that this can always be done! Since it is very useful in analysis (and elsewhere in mathematics) to be able to make such selections, an additional axiom is usually added to set theory, called the Axiom of Choice; the expanded theory is called ZFC set theory.

II.6.1. The Axiom of Choice

II.6.1.1. Definition. Let \( \{X_i : i \in I\} \) be an indexed collection of (nonempty) sets. A choice function for \( \{X_i : i \in I\} \) is a function \( c : I \to \bigcup_{i \in I} X_i \) such that \( c(i) \in X_i \) for each \( i \).

If there is a choice function for \( \{X_i : i \in I\} \), then each \( X_i \) is obviously nonempty. The Axiom of Choice asserts the converse:

II.6.1.2. Axiom of Choice: (Indexed Form) If \( \{X_i : i \in I\} \) is an indexed collection (set) of nonempty sets, then there is a choice function for \( \{X_i : i \in I\} \).

The Axiom of Choice is often stated in a slightly different but equivalent form:

II.6.1.3. Axiom of Choice: (Power Set Form) If \( X \) is a nonempty set, there is a function (choice function) \( c : \mathcal{P}(X) \setminus \{\emptyset\} \to X \) with \( c(A) \in A \) for every nonempty \( A \subseteq X \).

II.6.1.4. To obtain the Power Set Form from the Indexed Form, set \( I = \mathcal{P}(X) \setminus \{\emptyset\} \), and set \( X_A = A \). Conversely, to obtain the Indexed Form from the Power Set Form, let \( \{X_i : i \in I\} \) be an indexed collection of nonempty sets, and set \( X = \bigcup_{i \in I} X_i \). Let \( c : \mathcal{P}(X) \setminus \{\emptyset\} \to X \) be a choice function for \( X \), and let \( h : I \to \mathcal{P}(X) \setminus \{\emptyset\} \) be defined by \( h(i) = X_i \). Then \( c \circ h \) is a choice function for \( \{X_i : i \in I\} \).

II.6.1.5. The Axiom of Choice (abbreviated AC) is often slightly misunderstood, and it is important to understand what it does not say. The AC is frequently stated somewhat informally as:

“If \( \{X_i\} \) is a collection of nonempty sets, you can choose (or simultaneously choose) an element out of each of the sets.”

This statement of the AC is not incorrect if properly interpreted, but it can be misleading. What the AC actually asserts is simply the existence of a choice function; it does not give any procedure for finding or describing one, or even assert that there is an explicit description of one (in fact, the AC is typically invoked precisely in situations where an explicit description of a choice function is not possible.)

---

8This statement is only true if ZF set theory is assumed to be consistent; consistency of ZF set theory remains unproved and is perhaps unprovable (it is unprovable within ZF.)
The distinction comes into sharpest focus when just one nonempty set $X$ is considered. One can ask two very different questions about $X$:

1. Does there exist an element (equivalently, a singleton subset) of $X$?

2. Can one find or identify an explicit element of $X$?

The answer to Question 1 is trivially yes: this is the very definition of the word “nonempty.” The Axiom of Choice in this setting simply asserts a positive answer to Question 1; thus the AC is trivially true for a collection consisting of one nonempty set. (It is also true, only slightly less trivially, for a collection of a finite number of nonempty sets.) But Question 2 is definitely nontrivial, and is more of a philosophical than a mathematical question. For example, suppose one proves that $\mathbb{R}$ is uncountable (II.5.3.6.) and that the set of algebraic numbers is countable (II.5.6.6.). These results together prove that the set of transcendental numbers is nonempty (and, in fact, uncountable). However, just knowing these results one would be very hard pressed to exhibit a transcendental number. (Actually, in this instance, the standard proofs of the assertions could in principle be used to give an extremely cumbersome construction of an explicit transcendental number. There are, of course, much better ways of identifying explicit transcendental numbers, e.g. III.8.2.35., III.9.3.1., III.9.3.2.)

As pointed out by J. Hadamard [Moo82, ], the situation is very similar to the question of whether we allow consideration of functions from $\mathbb{R}$ to $\mathbb{R}$ for which we have no formula or effective way of calculating the values. Indeed, until Riemann in the mid-19th century, the term “function” was restricted to functions with a closed-form formula, or at least a power-series representation (i.e. analytic functions). The modern definition of function as simply a set of ordered pairs without necessarily a specified rule or method for calculation of its values would have been much too broad to be acceptable 200 years ago.

The Axiom of Choice is a pure existence assertion, not a constructive one.

Equivalent Formulations

The Axiom of Choice has many other equivalent reformulations, some of which seem intuitively plausible and others intuitively implausible. This situation reflects the unexpectedly subtle nature of the Axiom of Choice; it must be kept in mind that it is meaningless (within ZF) to ask whether any of these versions of the Axiom are “true”.

An obvious reformulation which seems plausible concerns Cartesian products. A choice function for $\{X_i : i \in I\}$ is just an element of $\prod_{i \in I} X_i$. Thus the Axiom of Choice is equivalent to:

**Multiplicative Axiom:** If $\{X_i : i \in I\}$ is an indexed collection of nonempty sets, then $\prod_{i \in I} X_i$ is not empty.

Bertrand Russell, who first formulated the Multiplicative Axiom version (he phrased it somewhat differently since infinite Cartesian products were not defined at the time), regularly called the Axiom of Choice by this name.

One potential way of generating a choice function is by transfinite induction. Thus the Axiom of Choice appears to be related to well ordering. In fact, the Axiom of Choice is equivalent to the Well-Ordering Principle:

**Well-Ordering Principle:** Every set can be well ordered.
II.6.2.3. To see that the Well-Ordering Principle implies the Axiom of Choice, suppose \( \{X_i : i \in I\} \) is an indexed collection of nonempty sets, and fix a well ordering \( \preceq \) on \( I \) and a well ordering \( \preceq_i \) on \( X = \cup_{i \in I} X_i \). For \( i \in I \), let \( R_i \) be the rule “Let \( c(i) \) be the first element of \( (X, \preceq_i) \) which is in \( X_i \)” Then a choice function \( c \) is defined by transfinite induction (\( \cdot \)).

The proof that the Axiom of Choice implies the Well-Ordering Principle uses the notion of ordinals, and will be discussed in (\( \cdot \)). In fact, the Axiom of Choice was first formulated by E. Zermelo specifically to prove the Well-Ordering Principle.

**Zorn’s Lemma and the Hausdorff Maximal Principle**

There are two similar statements about partially ordered sets which are also equivalent to the Axiom of Choice. To describe them simply, we need a definition:

II.6.2.4. **Definition.** Let \( (X, \preceq) \) be a partially ordered set. A **chain** in \( (X, \preceq) \) is a subset of \( X \) which is totally ordered in the induced ordering. A **maximal chain** is a chain which is not properly contained in any larger chain.

II.6.2.5. **Hausdorff Maximal Principle:** Let \( (X, \preceq) \) be a partially ordered set. Then \( (X, \preceq) \) contains a maximal chain.

II.6.2.6. **Zorn’s Lemma:** Let \( (X, \preceq) \) be a partially ordered set. If each chain in \( (X, \preceq) \) has an upper bound in \( X \), then \( X \) contains a maximal element.

Zorn’s Lemma is called a “Lemma” for historical reasons, but it is in fact equivalent to the Axiom of Choice. A name like **Zorn’s Principle** might be more suitable from a logical standpoint. The Hausdorff Maximal Principle is also sometimes called **Kuratowski’s Lemma** (HAUSDORFF and KURATOWSKI independently studied essentially this condition, with nuances of differences; the Hausdorff Maximal Principle was originally stated only for sets ordered by inclusion.)

II.6.2.7. It is fairly easy to see that the Hausdorff Maximal Principle and Zorn’s Lemma are equivalent. To see that the Hausdorff Maximal Principle implies Zorn’s Lemma, let \( (X, \preceq) \) be a partially ordered set in which every chain has an upper bound. Assume the Hausdorff Maximal Principle and let \( C \) be a maximal chain in \( X \). Then \( C \) has an upper bound \( x \), and we must have \( x \in C \) since otherwise \( C \cup \{x\} \) would be a chain properly containing \( C \). Thus \( x \) is the largest element of \( C \). Then \( x \) must be a maximal element of \( X \), since if there were a \( y \in X \) with \( x \prec y \), then \( C \cup \{y\} \) would be a chain properly containing \( C \).

II.6.2.8. The proof that Zorn’s Lemma implies the Hausdorff Maximal Principle is a good example of how Zorn’s Lemma is used in practice. Let \( (X, \preceq) \) be a partially ordered set. Let \( C \) be the set of all chains in \( (X, \preceq) \); then \( C \) is partially ordered by inclusion. If \( \{C_i : i \in I\} \) is a chain in \( (C, \subseteq) \), i.e. each \( C_i \) is a chain in \( (X, \preceq) \) and for any \( i, j \in I \), one of \( C_i \) and \( C_j \) is contained in the other, then it is easy to verify that \( \cup_{i \in I} C_i \) is a chain in \( X \): if \( x, y \in \cup_{i \in I} C_i \), then \( x \in C_i, y \in C_j \) for some \( i, j \in I \); but either \( C_i \subseteq C_j \) or vice versa, so both \( x \) and \( y \) are in either \( C_i \) or \( C_j \), so they are comparable since both \( C_i \) and \( C_j \) are chains. Also, \( \cup_{i \in I} C_i \) contains each of the \( C_i \); so \( \cup_{i \in I} C_i \) is an upper bound for \( \{C_i : i \in I\} \) in \( (C, \subseteq) \). Thus by Zorn’s Lemma \( (C, \subseteq) \) has a maximal element \( C \), which is a maximal chain in \( (X, \preceq) \).
II.6.2.9. In applications of Zorn’s Lemma, as here, the Lemma is frequently applied to a judiciously constructed new partially ordered set. Another relevant example is the proof that Zorn’s Lemma implies the Well-Ordering Principle, which is left as an Exercise (Exercise (.) See Exercises (,), () for other applications of Zorn’s Lemma.

II.6.2.10. The proof that the Axiom of Choice, the Well-Ordering Principle, the Hausdorff Maximal Principle, and Zorn’s Lemma are equivalent is completed by showing that the Axiom of Choice implies Zorn’s Lemma. This argument uses ordinals, and will be discussed in ()

**Tukey’s Lemma**

Another variation of AC similar to Zorn’s Lemma and the Hausdorff Maximal Principle is Tukey’s Lemma, often (and properly) called the *Teichmüller-Tukey Lemma* since it was independently formulated by O. Teichmüller; but I don’t feel guilty slighting Teichmüller in the name since he was a Nazi.

II.6.2.11. **Definition.** A collection \( F \) of sets is of finite character if a set \( X \) is in \( F \) if and only if every finite subset of \( X \) is in \( F \).

A good example of a collection of finite character is the collection of all linearly independent subsets of a vector space \( V \), since a set of vectors in \( V \) is linearly independent if and only if every finite subset is linearly independent (since a linear combination involves only finitely many vectors).

II.6.2.12. **Tukey’s Lemma:** If \( F \) is a collection of sets of finite character, then \( F \) has a maximal element under inclusion, i.e. there is an \( X \in F \) not properly contained in any other set in \( F \).

II.6.2.13. Tukey’s Lemma is an easy consequence of Zorn’s Lemma, since if \( F \) is a collection of sets of finite character, regarded as a partially ordered set under inclusion, and \( C = \{C_i : i \in I\} \) is a chain in \( F \), then \( C = \cup_{i \in I} C_i \) is in \( F \) since every finite subset of \( C \) is contained in some \( C_i \), and \( C \) is an upper bound for \( C \).

II.6.2.14. It is also not difficult to show that Tukey’s Lemma implies the Axiom of Choice (Exercise II.6.4.11.).

II.6.2.15. The Axiom of Choice is also equivalent to Tikhonov’s Theorem: an arbitrary product of compact topological spaces is compact in the product topology. This theorem will be discussed in ().

II.6.2.16. It must be pointed out that the Axiom of Choice is not universally accepted by mathematicians, even by analysts. This was particularly true in the early 20th century when the foundations of mathematics were being axiomatized; see [Sie65] and [Moo82] for interesting discussions of the points of view of leading mathematicians of the time. Today most analysts are more comfortable with the Axiom of Choice. But there is a branch of analysis called *Constructive Analysis*, where the full Axiom of Choice is not and cannot be assumed. There are results of both theoretical and practical importance in Constructive Analysis, but this theory is a separate subject from the one treated in this book. See e.g. [BB85] and [BV06]. There is also an axiom called the *Axiom of Determinacy* which is sometimes considered, which is inconsistent with the Axiom of Choice. A weaker version of this axiom, called the *Axiom of Projective Determinacy*, not inconsistent with the Axiom of Choice, plays a role in Descriptive Set Theory ()

164
II.6.3. Weaker Forms of Choice

II.6.3.1. Weaker versions of the Axiom of Choice, which are somewhat more palatable to some mathematicians (see [Pot04, p. 164-165, 240] for some relevant philosophical discussion), are sometimes assumed. One is the Countable Axiom of Choice (or Axiom of Countable Choice), written Countable AC or sometimes ACω, which is also not provable or disprovable in ZF:

Countable Axiom of Choice: (Indexed Form) If \( \{ X_n : n \in \mathbb{N} \} \) is a sequence of nonempty sets, then there is a choice function for \( \{ X_n : n \in \mathbb{N} \} \).

Note that it is the index set, not the \( X_n \), which is required to be countable. It is a quite different matter to weaken the AC by requiring the \( X_i \) to be countable, or finite, with no restriction on the index set; see II.6.3.13. There is no implication either way between the Countable AC and these forms of AC.

II.6.3.2. Analysis in its usual form is impossible without at least the Countable AC. The following statements, which can be readily proved using Countable AC, cannot be proved in ZF [Jec73]:

(i) A countable union of countable sets is countable (Exercise II.6.4.8.).

(ii) \( \mathbb{R} \) is not a countable union of countable sets.

(iii) Every infinite set contains a countably infinite subset (Exercise II.6.4.9.).

(iv) In \( \mathbb{R} \), or in metric spaces, the sequential criterion for limits and continuity is equivalent to the \( \epsilon - \delta \) definition ( ).

(v) In \( \mathbb{R} \), or in metric spaces, a point \( p \) is in the closure of a subset \( A \) if and only if there is a sequence of points of \( A \) converging to \( p \) ( ).

(vi) Every second countable topological space is separable ( ).

II.6.3.3. It turns out that a slightly stronger axiom called the Axiom of Dependent Choice is often more useful:

Axiom of Dependent Choice: Let \( R \) be a relation on a set \( X \). If, for every \( x \in X \), there is a \( y \in X \) with \( xRy \), then there is a sequence \( (x_n) \) in \( X \) with \( x_nRx_{n+1} \) for all \( n \).

Roughly speaking, the Axiom of Dependent Choice (abbreviated DC) asserts that a sequence of choices can be made, each depending on the previous choice(s).

II.6.3.4. The Axiom of Choice implies the Axiom of Dependent Choice, which implies the Countable Axiom of Choice (Exercise ( )); neither of the implications can be reversed. See [?] for a discussion of the Axiom of Dependent Choice. In some of the applications of the Axiom of Choice in this book, one can get away with the Axiom of Dependent Choice or even the Countable AC; we will try to identify most of these (see Exercise ( ) for one example.)

II.6.3.5. Note that the Axiom of Dependent Choice is not needed (in fact no form of Choice is needed) to prove the existence of sequences defined by recursion (III.1.5.15.).

165
The Boolean Prime Ideal Theorem

Another AC-type statement which has several important and useful formulations is the Boolean Prime Ideal Theorem, abbreviated BPI. We give several equivalent formulations of this statement, without, however, explaining all the terms involved (see, for example, [Jec73]). For part of the proof, see XI.11.9.19.

II.6.3.6. Boolean Prime Ideal Theorem (BPI): Every ideal in a Boolean algebra is contained in a prime ideal.

II.6.3.7. Compactness Theorem of Model Theory: Let $\Sigma$ be a set of sentences in a first-order language. If every finite subset of $\Sigma$ has a model, then $\Sigma$ has a model.

II.6.3.8. Ultrafilter Property: Every filter on a set is contained in an ultrafilter (XII.1.6.15.).

II.6.3.9. Tikhonov’s Theorem for Hausdorff Spaces: An arbitrary product of compact Hausdorff spaces is compact in the product topology (XI.11.6.6.).

II.6.3.10. Stone-Čech Compactification Theorem: Every completely regular topological space has a Stone-Čech compactification (XI.11.9.13.).

II.6.3.11. Theorem. The following are equivalent in ZF set theory:

(i) The Boolean Prime Ideal Theorem.
(ii) The Compactness Theorem of Model Theory.
(iii) The Ultrafilter Property.
(iv) Tikhonov’s Theorem for Hausdorff spaces.

II.6.3.12. It is easy to deduce the Ultrafilter Property from Zorn’s Lemma (it is also not difficult to obtain the BPI directly from Zorn’s Lemma, and it is trivial to obtain the Hausdorff Tikhonov Theorem from the general one), and hence the AC implies the BPI. The converse is false, i.e. the BPI is a strictly weaker assumption than AC. In fact:

II.6.3.13. Theorem. The following statements are successively strictly logically weaker in ZF set theory, i.e. each can be proved from the previous one in ZF, and none follows in ZF from the succeeding ones (and even (v) is not a theorem of ZF for $n \geq 2$):

(i) The Axiom of Choice.
(ii) The Boolean Prime Ideal Theorem.
(iii) The Ordering Principle: Every set can be totally ordered.
(iv) The Axiom of Choice for Finite Sets: Every collection of nonempty finite sets has a choice function.

(v) For each fixed \( n \in \mathbb{N} \), the Axiom of Choice for \( n \)-element sets.

See Exercise II.6.4.10. for (iii) \( \Rightarrow \) (iv), and [Jec73] for a complete discussion including the failure of the converse implications, and other variations of AC. Here are some selected additional facts, discussed in [Jec73]:

**II.6.3.14.** DC does not imply (v), hence neither DC nor Countable AC implies any of the stronger statements. There is no implication the other way between any of (ii)–(v) and either the Countable AC or DC. (This last question is not discussed in [Jec73], but see [HR98].)

**II.6.3.15.** The statement that there exists a free ultrafilter on every infinite set is strictly weaker than the Ultrafilter Property. The existence of a free ultrafilter on \( \mathbb{N} \) (an assumption underlying nonstandard analysis) is weaker still.

**II.6.3.16.** There is another pair of equivalent statements which are important in analysis, which are strictly weaker than the BPI: the Hahn-Banach Theorem () and the Boolean Measure Theorem (). These statements are independent of Countable AC and DC and of II.6.3.13.(iv)–(v), and apparently independent of II.6.3.13.(iii).

See [RR85] for a compilation of many other statements equivalent to the AC or other versions of Choice, and [HR98] for an enormous variety of Choice statements of mostly strictly varying logical strength. The fact that there are so many models of set theory in which some of these statements are true and others false tends to seriously undercut the credibility of set theory in its present form as a satisfactory foundation for all of mathematics.

The Axiom of Choice has been controversial, and remains somewhat so; however, most modern mathematicians, especially analysts, have no serious qualms about accepting and using it, often without comment. The attitude of R. Boas [Boa96, p. xi] is widespread:

“[A]fter Gödel's results, the assumption of the axiom of choice can do no mathematical harm that has not already been done.”

In this book we will feel free to use the AC, but we will try to be careful to point out what version of Choice is used and really needed whenever it is employed. Analysis as we will develop it would not be the same, or even very similar, without it.

“Rarely have the practitioners of mathematics, a discipline known for the certainty of its conclusions, differed so vehemently over one of its central premises as they have done over the Axiom of Choice. Yet without the Axiom, mathematics today would be quite different. The very nature of modern mathematics would be altered and, if the Axiom's most severe constructivist critics prevailed, mathematics would be reduced to a collection of algorithms. Indeed, the Axiom epitomizes the fundamental changes – mathematical, philosophical, and psychological – that took place when mathematicians seriously began to study infinite collections of sets.”

*Gregory H. Moore*\(^9\)

---

\(^9\)[Moo82, p. 1]

II.6.4. Exercises

II.6.4.1. Show that a totally ordered set is well ordered if and only if it contains no strictly decreasing sequences. (The Axiom of Dependent Choice is needed for one direction.)

II.6.4.2. Let $X$ be a set. Let $W$ be the set of all pairs $(A, \leq)$, where $A \subseteq X$ and $\leq$ is a well ordering on $A$. Define a partial order $\preceq$ on $W$ by setting $(A, \leq_A) \preceq (B, \leq_B)$ if $A \subseteq B$, $A$ is an initial segment of $B$ under $\leq_B$, and the restriction of $\leq_B$ to $A$ is $\leq_A$. (Note that $W \neq \emptyset$ if $X \neq \emptyset$ since every singleton subset of $X$ has an obvious well ordering.)

(a) Show that if $\{(A_i, \leq_i) : i \in I\}$ is a chain in $W$, then $(A, \leq)$ is an upper bound, where $A = \bigcup_i A_i$ and $\leq = \leq_i$ on $A_i$. [It must be shown that $\leq$ is well defined on $A$ and is a well ordering, and that each $A_i$ is an initial segment.]

(b) If $(A, \leq)$ is any element of $W$, and $x \in X \setminus A$, extend $\leq$ to $A \cup \{x\}$ by setting $a \leq x$ for all $a \in A$. Show that $(A \cup \{x\}, \leq) \in W$.

(c) If $(A, \leq)$ is a maximal element of $W$, then $A = X$.

(d) Conclude that Zorn’s Lemma implies the Well-Ordering Principle.

II.6.4.3. (a) Prove that the Axiom of Choice implies the Axiom of Dependent Choice. [If $R$ is a relation on $X$, let $c : P(X) \setminus \{\emptyset\} \to X$ be a choice function, and set $x_{n+1} = c(\{y \in X : x_n R y\})$.] (b) Prove that the Axiom of Dependent Choice implies the Countable Axiom of Choice. [If $\{X_n : n \in \mathbb{N}\}$ is an indexed collection of nonempty sets, let $Z$ be the set of all ordered pairs $(n, x)$, where $n \in \mathbb{N}$ and $x \in X_n$. Let $(n, x) R (m, y)$ if and only if $m = n + 1$.]

II.6.4.4. Using the Axiom of Dependent Choice, prove that every infinite set $X$ contains a sequence $(x_n)$ of distinct elements (), i.e. there is an injective function from $\mathbb{N}$ to $X$. [Let $Z$ be the set of ordered pairs $(n, A)$, where $n \in \mathbb{N}$ and $A$ is an infinite subset of $X$. Define $(n, A) R (m, B)$ if and only if $m = n + 1$, $B \subseteq A$, and $A \setminus B$ contains exactly one element.]

II.6.4.5. The proof of Tikhonov’s theorem () uses the Axiom of Choice. Conversely, the Axiom of Choice can be proved from Tikhonov’s Theorem for $T_1$ spaces, showing that the two are equivalent.

(a) Let $X$ be a set. Define the $T_1$-compactification of $\mathbb{N}$ by $X$ to be the disjoint union $\mathbb{N} \cup X$, topologized by taking as the open sets any subset with finite complement, and any subset of $\mathbb{N}$. Show that $\mathbb{N} \cup X$ is compact if $X$ is nonempty, always $T_1$, but not Hausdorff if $X$ has more than one element.

(b) Let $\{X_i : i \in I\}$ be an indexed collection of nonempty sets, and let $X = \prod_{i \in I} X_i$. Show using Tikhonov’s Theorem that $X$ is nonempty, as follows. For each $i$, let $T_i$ be the $T_1$ compactification of $\mathbb{N}$ by $X_i$, and let $Y = \prod_{i \in I} Y_i$. Then $X \subseteq Y$, and $Y$ is compact by Tikhonov’s Theorem.

(c) $Y \neq \emptyset$: for each $n \in \mathbb{N}$ let $y_n \in Y$ be the point with i’th coordinate $n$ for all $i$. By compactness, the sequence $(y_n)$ has a cluster point $p$ in $Y$. Show that for each $i$, the $i$’th coordinate $p_i$ of $p$ is in $X_i$, so $p \in X$. It is known that Tikhonov’s Theorem for Hausdorff spaces is strictly weaker than the Axiom of Choice (it is equivalent to the BPI). Tikhonov’s Theorem for Hausdorff spaces is equivalent to the statement that any product of copies of $[0, 1]$ is compact. See II.6.3.11. for a discussion of this and related matters.
II.6.4.6. Identify how the proof of II.5.4.8. uses the Countable Axiom of Choice. [In fact, II.5.4.8. is not a theorem of ZF [], i.e. some type of Choice Axiom is necessary.]

II.6.4.7. ([Bla77]; cf. [Oxt71, p. 95]) Show that the Axiom of Dependent Choice is equivalent to the statement that if \((X_n)\) is a sequence of nonempty sets, then \(\prod_n X_n\) is a Baire space, where each \(X_n\) is given the discrete topology, as follows.

(a) Show that the product topology on \(\prod_n X_n\) is given by a complete metric (cf. ()).
(b) Let \(R\) be a relation on a nonempty set \(X\), such that for every \(x \in X\) there is a \(y \in X\) with \(x R y\). For each \(n\) let

\[
U_n = \{ f \in X^N : f(n) R f(m) \text{ for some } m > n \} .
\]

Show that \(U_n\) is a dense open subset of \(X^N\). [Write

\[
U_n = \bigcup_{m > n} \bigcup_{(a,b) \in R} \{ f \in X^N : f(n) = a, f(m) = b \} .
\]

(c) Let \(g \in \cap_{n=1}^\infty U_n\). Let \(k_1 = 1, x_1 = g(1)\), and inductively let \(k_{n+1}\) be the smallest \(m > k_n\) for which \(g(k_n) R g(m)\), and \(x_{n+1} = g(k_{n+1})\).

(d) Conclude that the Baire Category Theorem (BC) implies the Axiom of Dependent Choice (DC). Since the proof of (BC) uses only (DC) (along with ZF), ZF+(BC) is logically equivalent to ZF+(DC).

(e) Show that under the hypotheses of DC, the sequence \((x_n)\) can be chosen with \(x_1\) arbitrary. [\(\cap_{n=1}^\infty A_n\) is dense in \(X^N\) by (BC).]

(f) Show that the Countable Axiom of Choice is equivalent to the statement that \(\prod_n X_n\) is nonempty for any sequence \((X_n)\) of nonempty sets.

II.6.4.8. Show how the countable AC is used in the proof of ().

II.6.4.9. Using Countable AC, show that every infinite set contains a countably infinite subset. [A set \(X\) is infinite if and only if, for each \(n \in \mathbb{N}\), the collection \(\mathcal{P}_n(X)\) of \(n\)-element subsets of \(X\) is nonempty. Choose such a set for each \(n\) and consider their union.]

II.6.4.10. Show that the Ordering Principle implies the Axiom of Choice for Finite Sets. [If \(X\) is totally ordered, every finite subset of \(X\) has a smallest element.]

II.6.4.11. Let \(\{X_i : i \in I\}\) be an indexed collection of nonempty sets, and \(\mathcal{F}\) the set of all choice functions for subcollections of the form \(\{X_i : i \in J\}\) for \(J \subseteq I\). Each element of \(\mathcal{F}\) is a set of ordered pairs of the form \((i, x)\), where \(i \in I\) and \(x \in X_i\). Regard \(\mathcal{F}\) as a partially ordered collection of sets under inclusion.

(a) Show that \(\mathcal{F}\) is nonempty. [Consider choice functions on finite subsets of \(I\).]

(b) Show that \(\mathcal{F}\) is a collection of sets of finite character.

(c) Show that if \(f\) is a maximal element of \(\mathcal{F}\), then the domain of \(f\) is all of \(I\). [Show that a choice function defined on a proper subset of \(I\) can be extended to one more index.]

(d) If Tukey’s Lemma is assumed, conclude that \(\{X_i : i \in I\}\) has a choice function, i.e. that AC holds.
II.6.4.12. (cf. III.3.7.4.) Let $R$ be a unital ring (III.3.6.3.). A nonempty subset $I$ of $R$ is a \textit{(two-sided)} ideal if $I$ is closed under addition and $x \in I$, $y \in R$ imply $xy, yx \in I$. An ideal is \textit{proper} if it is a proper subset of $R$. A \textit{maximal ideal} is a proper ideal which is not contained in any strictly larger proper ideal.

(a) Show that $\{0\}$ is a (proper) ideal in $R$. More generally, if $S \subseteq R$, then the set of finite sums of elements of $R$ of the form $xzy$, where $z \in S$ and $x, y \in R$, is an ideal of $R$ (not necessarily proper), and is the smallest ideal of $R$ containing $S$, called the \textit{ideal generated by $S$}, usually denoted $(S)$.

(b) Show that if $I$ is an ideal in $R$, then $0 \in I$, and $x \in I$ implies $-x \in I$. [Use $-1 \in R$.]

(c) Show that an ideal $I$ in $R$ is proper if and only if $1 \notin I$.

(d) Show that any intersection of ideals in $R$ is an ideal in $R$.

(e) Show that a union of ideals is not an ideal in general. But the union of a \textit{chain} of ideals (under inclusion) is an ideal. If $I$ and $J$ are ideals, the ideal generated by $I \cup J$ is

$$I + J = \{x + y : x \in I, y \in J\}.$$ 

(f) Use Zorn’s Lemma to prove that every proper ideal in a unital ring $R$ is contained in a maximal ideal.

(g) Show that if $I$ and $J$ are ideals in $R$, then the set $IJ$ of finite sums of elements of the form $xy$, where $x \in I$ and $y \in J$, is an ideal of $R$ contained in $I \cap J$.

(h) A proper ideal $P$ of $R$ is \textit{prime} if, whenever $I$ and $J$ are ideals of $R$ with $IJ \subseteq P$, then $I \subseteq P$ or $J \subseteq P$. Show that every maximal ideal is a prime ideal. [If $M$ is a maximal ideal and $I$ and $J$ are ideals with $IJ \subseteq M$ and $I \not\subseteq M$, then $I + M = R$, so $1 \in I + M$.]

Ideals can be defined also in nonunital rings (although (b) is not automatic and must be included as part of the definition of ideal; cf. III.3.7.4.). However, a nonunital ring need not contain any maximal ideals, e.g. there can be an increasing sequence of proper ideals whose union is the whole ring.

The term “ideal” arose since ideals were first considered in rings of algebraic numbers (). Every element $x$ in a ring $R$ generates an ideal $(x) = \{x\}$; such ideals are called \textit{principal ideals}. In many rings such as $\mathbb{Z}$ or a polynomial ring in one variable over a field, every ideal is principal, so there is no real distinction between ideals and elements (or equivalence classes of elements obtained by multiplication by units, e.g. $(n) = (-n)$ in $\mathbb{Z}$). But in some rings of algebraic integers, not every ideal is principal, and in these rings arithmetic is more complicated. \textsc{Kummer} had the idea to add “ideal numbers” (i.e. nonprincipal ideals), and to consider the arithmetic of ideals, which turns out to be considerably simpler than arithmetic of numbers, e.g. every ideal in a ring of algebraic integers factors uniquely as a product of prime ideals. See books on Abstract Algebra or Algebraic Number Theory for details.
II.7. Axiomatic Set Theory

The first set of axioms for modern set theory was proposed by Ernst Zermelo in 1908, although closely related work had been done previously by Cantor, Russell, and several others. The axioms were refined by Abraham Fraenkel (in part using ideas of Thoralf Skolem), and are today known as the Zermelo-Fraenkel Axioms and the theory is called ZF set theory. An additional axiom proposed by John von Neumann (the Axiom of Regularity) is usually included. Most of Zermelo’s axioms were not controversial, but one of them was: the Axiom of Choice, discussed in (); the Axiom of Choice is customarily not included among the ZF axioms, and the theory with the Axiom of Choice included is called ZFC set theory. Other axiom systems, most notably the ones developed by Kurt Gödel and Paul Bernays, and Nicolas Bourbaki, have been proposed, but the principal ones turn out to be equivalent to ZF theory in the sense that exactly the same theorems about sets can be proved. The Gödel-Bernays axioms are generally similar to the ZF axioms except that classes which are not sets appear explicitly.

The framework of set theory is two undefined concepts: set and membership. All objects of set theory are sets (or sometimes classes which are too large to be sets); the elements of a set (or class) are also sets. Thus, to encode the rest of mathematics into set theory, other mathematical objects must be defined as sets of some kind. For example, an ordered pair \((x, y)\) is defined to be \(\{\{x\}, \{x, y\}\}\); thus relations and functions can be defined as sets of ordered pairs, i.e. sets of sets, and therefore as sets. In the appendix, we describe how to define the natural numbers as sets, and then \(\mathbb{Z}\) and \(\mathbb{Q}\) can be described as equivalence classes of ordered pairs of natural numbers. A description of real numbers as sets of rational numbers is given in ()

The basic axioms of set theory assert that the usual operations give sets. The Axiom of Infinity asserts that the natural numbers are sets and that the collection of all natural numbers is a set. The axioms of Union and Power Set assert that the union of the elements of a set \(x\) is a set, written \(\cup x\) (remember that the elements of a set are themselves sets, so this union makes sense), and that \(\mathcal{P}(x)\) is a set. Other axioms assert that any subset of a set is a set, and that the range of a function is a set. The axioms are all phrased in formal logical language: for example, the Axiom of Union is written

\[ \forall x \exists y \forall a \left( a \in y \iff (\exists z \left[ a \in z \& z \in x \right]\right) \]

We will not try to carefully state the other axioms, which can be found in most books on mathematical logic or set theory. The axioms are inspired by the following somewhat imprecise principle, first espoused by Cantor:

**Basic Principle of Set Theory:** Every collection of objects [sets] is regarded as a set unless there is some logical contradiction in allowing it.

Thus, for example, Russell’s Paradox disqualifies allowing “the set of all sets.”

Von Neumann’s Axiom of Regularity had its origin in Russell’s Paradox, where one of the main problems was the existence of sets which are not ordinary (i.e. a set \(x\) with \(x \in x\)) A similar problem arises if there is a finite loop \(x_1, \ldots, x_n\) of sets with \(x_{k+1} \in x_k\) for \(1 \leq k < n\) and \(x_1 \in x_n\), so this possibility should also be eliminated. Zermelo’s axioms allowed such situations, although they (apparently) do not create a paradox in his system. But it is undesirable to allow such sets, and Von Neumann’s solution was to add the following axiom, which is also sometimes called the “Axiom of Foundation”:

10The Nicolas Bourbaki of mathematics was not an actual person, but the *nom de plume* of a gradually changing consortium of leading French mathematicians whose goal has been to systematize all of mathematics. See [7], [8], or [9] for a history of this interesting project.
Axiom of Regularity: For every nonempty set $x$ there is a $y \in x$ with $y \cap x = \emptyset$.

Such a $y$ is called a minimal element of $x$ (the terminology is explained in ()). The Axiom of Regularity prevents the existence of loops of the type described above; in fact, it is equivalent to the following statement:

Descending Chain Condition: There is no infinite sequence $(x_n)$ of sets with $x_{n+1} \subset x_n$ for all $n$.

[The Axiom of Choice is needed to prove one direction of the equivalence.]

The Axiom of Regularity seems reasonable if one thinks of sets as being constructed in “stages”, for then the elements of a set constructed at one stage must have been constructed at a previous stage; an element constructed at the earliest possible stage would be a minimal element. This can be made precise, and it turns out that the Axiom of Regularity limits the universe just to the sets which can be constructed out of $\emptyset$ by repeated union, power set, and subset (see Exercise ()); thus it seems to be in conflict with the Basic Principle above. The fact that it is a restrictive axiom is the source of the discomfort some mathematicians have with it. On the other hand, set theory with the Axiom of Regularity is large enough to encode all of standard mathematics in the usual way, and there are no important results which seem to be provable with a weakened version of the Axiom but not with the full version, so there is (at present) no practical reason to weaken it.

The situation with the Axiom of Choice is the opposite: it is an expansive axiom, allowing a certain infinite operation. Some mathematicians are uncomfortable with it because it seems to allow too much freedom in manipulating sets (although it could be argued that it is more in line with the Basic Principle than the Axiom of Regularity is.) There continue to be parts of mathematics where the Axiom of Choice is not assumed, and sometimes a contradictory axiom is substituted. One of the root causes of the controversy over the Axiom of Choice was identified by B. Rotman and G. T. Kneebone:

“[T]he axiom of choice has a curious ambivalence, since it can be regarded with equal justification as a mathematical axiom, which guarantees the existence of a set in certain stated circumstances, and as a logical principle, needed for the authentication of a pattern of inference that is indispensable in many branches of mathematics.”

In his very interesting book, *The Banach-Tarski Paradox*, STAN WAGON describes the role of the Axiom of Choice in connection with the Paradox, and in mathematics in general:

“Despite the Banach-Tarski Paradox and other objections to the Axiom of Choice, often focusing on its nonconstructive nature, the great majority of contemporary mathematicians fully accept the use of AC. It is generally understood that nonmeasurable sets (either of the Vitali type or of the sort that arise in the Banach-Tarski Paradox) lead to curious situations that contradict physical reality, but the mathematics that is brought to bear on physical problems is almost always mathematics that takes place entirely in the domain of measurable sets. And even in this restricted domain of sets, the Axiom of Choice is useful, and ZF + AC provides a more coherent foundation than does ZF alone. . . .

Despite the general acceptance of AC, it is recognized that this axiom has a different character than the others. Because of the nonconstructivity it introduces, AC is avoided when possible, and proofs in ZF are considered more basic (and are often more informative) than proofs using AC. A purist might argue that once accepted as an axiom, AC should be accorded the same

---

status as the other axioms: a statement that is proved in ZF + AC is just as mathematically true as one proved in ZF. But there are sometimes beneficial side effects of a discriminating attitude towards the use of AC. For instance, consider the Schröder-Bernstein Theorem for cardinality [()], which is much easier to prove in ZFC than in ZF. . . But the ZF-proof . . . is much more valuable because it generalizes in a way that the ZFC-proof does not. . .

The interplay between ZF and ZF + AC works the other way as well. For example, the ZF-proof that a square is not paradoxical . . . was found by closely analyzing the original ZFC-proof. If a more restrictive attitude toward AC had prevailed, the original ZFC-proof might not have been discovered so soon, and the proof (in ZF) that squares are not paradoxical might have been delayed."

II.8. Ordinals

Ordinals can be thought of as standard models for well ordered sets.

II.8.1. Basic Definitions

The formal theory of ordinals takes place in ZF set theory, where every element of a set is itself a set and there is no reason to make notational distinctions between sets and their elements (so we will frequently use lower-case letters to denote both sets and elements.) Thus the following strange-looking definition makes sense:

II.8.1.1. Definition. A set \( x \) is transitive if every element of \( x \) is a subset of \( x \). A set \( x \) is an ordinal if \( x \) is transitive and every element of \( x \) is transitive. Denote the class of all ordinals by \( \Omega \).

The term “transitive” is appropriate, since \( y \subseteq x \) and \( z \in y \) imply \( z \subseteq x \).

We will normally use the Greek letters \( \sigma, \tau, \rho \) to denote general ordinals, and other lower-case Greek letters to denote specific ordinals.

There is one ordinal to begin with, \( \emptyset \) (so \( \Omega \neq \emptyset \)). There are many other ordinals which can be manufactured from \( \emptyset \) by “transfinite induction,” using the next two constructions.

II.8.1.2. Definition. Let \( \sigma \) be an ordinal. Then \( S(\sigma) = \sigma \cup \{\sigma\} \) is also an ordinal, called the successor of \( \sigma \).

Note that \( \sigma \notin \sigma \) by the Regularity Axiom (), so \( S(\sigma) \neq \sigma \). The term “successor” makes sense in reference to a natural ordering on ordinals, described below.

Similarly, the next result is immediate from the definition:

II.8.1.3. Proposition. Let \( \{\sigma_i : i \in I\} \) be a set of ordinals. Then \( \bigcup_{i \in I} \sigma_i \) and \( \bigcap_{i \in I} \sigma_i \) are ordinals.

Together, these operations give a very large supply of ordinals. In fact:

II.8.1.4. Proposition. If \( X = \{\sigma_i : i \in I\} \) is any set of ordinals, there is an ordinal \( \sigma \) which is not in \( X \).

Proof: Set \( \sigma = S(\bigcup_{i \in I} \sigma_i) \).

173
II.8.1.5. COROLLARY. The class $\Omega$ of all ordinals is not a set.

This is the Burani-Forti Paradox, which like Russell’s Paradox implies that there is no “set of all sets.”

We will describe specific examples arising from these constructions in the next section.

II.8.2. Ordering on Ordinals

We now discuss the natural ordering on ordinals. First comes a simple observation:

II.8.2.1. PROPOSITION. If $\sigma$ is an ordinal, then every element of $\sigma$ is also an ordinal.

Indeed, if $\tau \in \sigma$, then $\tau$ is transitive; and since $\tau \subseteq \sigma$, every element of $\tau$ is an element of $\sigma$, hence transitive.

II.8.2.2. DEFINITION. If $\sigma$ and $\tau$ are ordinals, then $\tau < \sigma$ if $\tau \in \sigma$.

Thus, if $\sigma$ is an ordinal, then

$$\sigma = \{ \tau \in \Omega : \tau < \sigma \}.$$  

So $<$ defines an “ordering” on the set $\sigma$.

II.8.2.3. PROPOSITION. The relation $<$ is a strict “partial order” on $\Omega$.

[The “relation” $<$ is technically not a relation in the usual sense, since $\Omega$ is not a set. However, the meaning of the term “partial order” is still clear. In particular, $<$ is a true partial order on any set of ordinals, and in particular on any ordinal itself.]

PROOF: If $\tau < \sigma$ and $\sigma < \rho$, then $\tau \in \sigma$ and $\sigma \subseteq \rho$, so $\tau \in \rho$, i.e. $\tau < \rho$, so $<$ is transitive. Antisymmetry and strictness of the partial order follow from the Regularity Axiom, which implies that $\sigma \not\in \sigma$ for any $\sigma$. \(\Diamond\)

Some version of the Regularity Axiom is necessary here; otherwise there could be strange sets $x$ with the property that $x = \{x\}$ and any such set would be an ordinal.

The Regularity Axiom also implies:

II.8.2.4. PROPOSITION. Every nonempty set of ordinals contains a minimal element (with respect to $<$.)

In fact, if $X$ is a nonempty set of ordinals, then any minimal element of $X$ in the sense of () is also minimal with respect to $<$, as is easily verified (this is the origin and explanation of the term minimal in the Regularity Axiom.)

The existence of minimal elements also applies to classes of ordinals which are not sets: if $V$ is a nonempty class of ordinals, let $\sigma \in V$, and let $X = \{ \tau \in V : \tau \leq \sigma \}$. Then $X$ is a nonempty set of ordinals, and any minimal element of $X$ is a minimal element of $V$.

This fact and the next one are probably the most important facts about the ordering on $\Omega$. 174
II.8.2.5. **Proposition.** The ordering on $\Omega$ is a total ordering, i.e. any two ordinals are comparable.

**Proof:** Suppose there are ordinals which are not comparable. Let $\sigma$ be a minimal ordinal for which there is an ordinal not comparable to $\sigma$, and let $\tau$ be a minimal ordinal which is not comparable to $\sigma$. If $\rho \in \sigma$, then $\rho < \sigma$ so $\rho$ is comparable to any ordinal, and hence to $\tau$. We cannot have $\tau \leq \rho$ since then $\tau < \sigma$, so $\rho < \tau$, i.e. $\rho \in \tau$, so $\sigma \subseteq \tau$. Similarly, if $\rho < \tau$, then $\rho$ is comparable to $\sigma$ and $\sigma \not\subseteq \rho$, so $\rho < \sigma$ and $\tau \subseteq \sigma$, i.e. $\tau = \sigma$, a contradiction. So there are no such $\sigma$ and $\tau$.

Putting together the last two results, we obtain:

II.8.2.6. **Theorem.** Every nonempty class of ordinals has a smallest element, i.e. $\Omega$ is well ordered (more correctly, every set of ordinals is well ordered in the induced ordering.)

In particular, since every ordinal is itself a set of ordinals, each ordinal is well ordered in its natural ordering. If $\sigma$ is an ordinal, then $S(\sigma)$ is indeed the smallest ordinal larger than $\sigma$, justifying the term “successor”.

II.8.2.7. **Corollary.** Let $\sigma$ and $\tau$ be ordinals. If $\tau \subseteq \sigma$, then $\tau \leq \sigma$ (i.e. $\tau = \sigma$ or $\tau \in \sigma$.)

Thus, on $\Omega$, the relation $\leq$ is the same as $\subseteq$.

It follows easily that if $x = \{ \sigma_i : i \in I \}$ is a nonempty set of ordinals, then $\bigcap_{i \in I} \sigma_i$ is the smallest element of $x$. It is also easily seen that $\bigcup_{i \in I} \sigma_i$ is the smallest ordinal $\sigma$ such that $\sigma_i \leq \sigma$ for all $i \in I$; this element is often called $\text{lim}(x)$ [$x$ is well ordered in the induced ordering and can be regarded as a net, and this net converges to $\text{lim}(x)$ in the order topology (])]. Note that $\text{lim}(x)$ is also the supremum of $x$ in $\Omega$, although the notation $\text{sup}(x)$ is often used in ordinal theory to denote $S(\text{lim}(x))$, in conflict with the usual notation for ordered sets.

II.8.2.8. **Definition.** An ordinal $\sigma$ is a **successor ordinal** if there is a $\tau$ with $\sigma = S(\tau)$. Otherwise $\sigma$ is a **limit ordinal**.

An ordinal $\sigma$ is a successor ordinal if and only if it contains a maximal element $\tau$, in which case $\sigma = S(\tau)$; it follows that the $\tau$ for which $\sigma = S(\tau)$ is unique.

By this definition, $\emptyset$ is a limit ordinal; but this ordinal is special (it is the smallest or first ordinal) and it is often treated separately from all others.

II.8.3. **Transfinite Induction**

The Principle of Transfinite Induction applies to the class $\Omega$:

II.8.3.1. **Theorem.** [PRINCIPLE OF TRANSFINITE INDUCTION] Suppose $P(\sigma)$ is a statement for each $\sigma \in \Omega$. Suppose, for each $\sigma$, the following **inductive implication** holds:

“If $P(\tau)$ is true for all $\tau < \sigma$, then $P(\sigma)$ is true.”

Then $P(\sigma)$ is true for all $\sigma \in \Omega$.

The argument is identical to the one in II.4.2.6.; the fact that $\Omega$ is not a set causes no difficulty. Definitions by Transfinite Induction also work for $\Omega$.

175
II.8.3.2. Theorem. Let $X$ be a set. Suppose, for each $\sigma \in \Omega$, we have a definite procedure or rule $R_\sigma$ for choosing an $f(\sigma) = x_\sigma \in X$, which may depend on the choice of the $f(\tau)$ for $\tau < \sigma$. Then there is a unique $f : \Omega \to X$ such that, for each $\sigma \in \Omega$, $f(\sigma)$ is determined by the specified procedure $R_\sigma$ applied to $\{f(\tau) : \tau < \sigma\}$.

[We are still being vague about the precise meaning of “statement”, “rule”, and “procedure”. There is also a technical difficulty in saying what is meant by a “function” whose domain is not a set; but the meaning in this context can be easily clarified in an intuitively acceptable way.]

II.8.3.3. Proposition. If $X$ is a set and $f : \Omega \to X$ is a function defined inductively as in II.8.3.2., then there are ordinals $\sigma$ and $\tau$ with $\sigma \neq \tau$ and $f(\sigma) = f(\tau)$.

This is because there cannot be a one-one correspondence between a set and a class which is not a set. (This follows from the Replacement Axiom of ZF set theory.)

II.8.4. Ordinals as Universal Models

We can now show that ordinals are universal models for well ordered sets. Recall that each ordinal is well ordered in its natural ordering.

II.8.4.1. Proposition. No two distinct ordinals are order-isomorphic.

Proof: It suffices to show that if $\tau < \sigma$, then there is no order-isomorphism from $\sigma$ onto $\tau$. Suppose $f$ is such an order-isomorphism. The set $\{\rho \in \sigma : f(\rho) < \rho\}$ is nonempty, since $\tau \in \sigma$ and $f(\tau) < \tau$. Let $\rho$ be the smallest element of $x$. Then, for every $\pi < \rho$, $\pi \leq f(\pi) < f(\rho)$ since $f$ is an order-isomorphism. But $\rho$ is the smallest ordinal larger than all $\pi < \rho$, so $\rho \leq f(\rho)$, a contradiction.

II.8.4.2. Theorem. Let $(I, \preceq)$ be a well ordered set. Then there is a unique ordinal $\sigma$ such that $(I, \preceq)$ is order-isomorphic to $(\sigma, \leq)$.

Proof: First fix an $x \notin I$ (for example, $x = I$.) Now define a function $f : \Omega \to I \cup \{x\}$ by transfinite induction, by letting $f(\sigma)$ be the first element of $I \setminus \{f(\tau) : \tau < \sigma\}$ if this set is nonempty, and $f(\sigma) = x$ otherwise. Then $f$ is not injective by II.8.3.3.; but $f(\sigma) = f(\tau)$ for $\sigma \neq \tau$ can only happen if $f(\sigma) = f(\tau) = x$ by the way $f$ is defined. Thus there is a smallest $\sigma$ such that $f(\sigma) = x$, and $f$ is a bijection from $\sigma = \{\tau \in \Omega : \tau < \sigma\}$ onto $I$. It is easily checked that $f$ is an order-isomorphism. Uniqueness follows from II.8.4.1.

The device in this argument of adding a new element $x$ and applying II.8.3.3. is commonly used, for example in the proof in II.8.5.2. and in Exercise (). It can also be used to carefully justify arguments such as the proof of (), where a function is defined by transfinite induction until the definition no longer “works”.

II.8.5. The Axiom of Choice and Zorn’s Lemma

Using ordinals and an argument like the proof of II.8.4.2., we can finally prove that the Axiom of Choice implies Zorn’s Lemma. We begin with a lemma:
II.8.5.1. **Lemma.** Let \((X, \preceq)\) be a partially ordered set in which every chain has an upper bound. If \(X\) has no maximal elements, then every chain \(C\) in \(X\) has an upper bound in \(X\) which is not in \(C\).

**Proof:** Suppose \(C\) is a chain. We must show there is an upper bound for \(C\) which is not in \(C\). By assumption, \(C\) has an upper bound \(x\). If \(x \notin C\), we are done. If \(x \in C\), then \(x\) is the largest element of \(C\). By assumption, \(x\) is not a maximal element of \(X\), so there is a \(y \in X\) with \(x < y\). Then \(y\) is an upper bound for \(C\) which is not in \(C\).

II.8.5.2. Now assume the Axiom of Choice, and let \((X, \preceq)\) be a partially ordered set in which every chain has an upper bound. Let \(d : \mathcal{P}(X) \setminus \{\emptyset\} \to X\) be a choice function for \(X\). Suppose that \(X\) has no maximal elements. Then by the lemma, for any chain \(C\), the set \(U_C\) of upper bounds for \(C\) which are not in \(C\) is nonempty. Fix \(y \notin X\) (say \(y = X\)) and define \(f : \Omega \to X \cup \{y\}\) by transfinite induction as follows. Let \(f(\sigma) = d(U_{C_\sigma})\) if \(C_\sigma = \{f(\tau) : \tau < \sigma\}\) is a chain, and \(f(\sigma) = y\) otherwise. It is then easily proved by transfinite induction that \(C_\sigma\) is a chain for every \(\sigma\), so \(f(\sigma) = d(U_{C_\sigma})\) for all \(\sigma\); hence \(f : \Omega \to X\) is injective, contradicting **II.8.3.3.** This contradicts the assumption that \(X\) has no maximal element.

II.8.5.3. Of course, since Zorn’s Lemma implies the Well-Ordering Principle, which implies the Axiom of Choice, the three are thus equivalent. A direct proof along the same lines as **II.8.5.2.** (but somewhat simpler) can be given that the Axiom of Choice implies the Well-Ordering Principle (Exercise ()).

II.8.6. **Ordinal Arithmetic and Examples**

In this section, we explicitly describe some of the standard ordinals, and also define arithmetic operations on the ordinals.

**Finite Ordinals**

II.8.6.1. The finite ordinals are essentially the natural numbers (along with 0.) In fact, in ZF set theory the natural numbers are defined to be the (nonzero) finite ordinals.

The finite ordinals are just constructed from \(\emptyset\) by repeated application of the successor operation:

- The first ordinal, \(\emptyset\), is usually denoted 0.
- The next ordinal, usually denoted 1, is \(S(0) = \{0\} = \emptyset\) (note that \(\emptyset \neq \emptyset\))
- The next ordinal, usually denoted 2, is \(S(1) = 1 \cup \{1\} = \{0, 1\} = \emptyset, \{\emptyset\}\).

The next ordinal, usually denoted 3, is \(S(2) = 2 \cup \{2\} = \{0, 1, 2\} = \emptyset, \{\emptyset\}, \emptyset, \{\emptyset\}\).

This process continues to give \(n = \{0, 1, 2, \ldots, n - 1\}\) for each \(n \in \mathbb{N}\).
Infinite Ordinals

II.8.6.2. The Axiom of Infinity in ZF set theory says that the class of all finite ordinals is a set, which is obviously an ordinal. This class, which is familiar as \( \mathbb{N} \cup \{0\} \), is usually denoted \( \omega \) when regarded as an ordinal (\( \omega_0 \) might be better notation.) The ordinal \( \omega \) is the first infinite ordinal, and is the first limit ordinal (other than 0.) The ordinals larger than \( \omega \) are also called infinite ordinals.

One can continue generating infinite ordinals via the successor operation, such as \( S(\omega) \), \( S(S(\omega)) \), etc. These can be much more efficiently described using ordinal arithmetic, so we will defer further discussion of them.

II.8.6.3. The Well-Ordering Principle implies that there is an uncountable well ordered set, and hence an uncountable ordinal by II.8.4.2.. It is perhaps surprising that the existence of an uncountable ordinal can be proved without resorting to the Axiom of Choice \( (\) \). The first uncountable ordinal is usually denoted \( \omega_1 \) (some authors use \( \Omega \), which we reserve for the class of all ordinals.) The ordinal \( \omega_1 \) is the set of all countable ordinals.

Similarly, the Well-Ordering Principle implies that there are ordinals of arbitrary cardinality. We will discuss this point more completely in the next section.

Ordinal Arithmetic

We can define addition, multiplication, exponentiation, and other algebraic operations on the set of ordinals. We begin with addition.

II.8.6.4. If \( \rho \) and \( \sigma \) are ordinals, we informally define \( \rho + \sigma \) to be the ordinal of the well-ordered set obtained by putting a copy of the well-ordered set \( \sigma \) at the end of the well-ordered set \( \rho \). More precisely, if \( \rho \) is a fixed ordinal, we define \( \rho + \sigma \) to be \( f(\sigma) \), where \( f \) is defined by transfinite induction according to the following rule:

\[
\begin{align*}
\quad f(0) & = \rho. \\
\quad \text{If } \sigma = S(\tau) & \text{, then } f(\sigma) = S(f(\tau)). \\
\quad \text{If } \sigma \text{ is a nonzero limit ordinal, then } f(\sigma) & = \lim \{f(\tau) : \tau < \sigma\}
\end{align*}
\]

We have that \( \rho + 1 = S(\rho) \) for every \( \rho \). In particular, \( \{\omega + 1, \omega + 2, \ldots\} \) is the strictly increasing sequence of immediate successors of \( \omega \). The limit of this set is \( \omega + \omega \).

II.8.6.5. Addition of finite ordinals coincides with usual addition of numbers, and has the usual familiar algebraic properties. However, most of the properties break down for general ordinals. For example, \( \omega + 1 = S(\omega) \neq \omega \), but it is easily checked that \( 1 + \omega = \omega \); thus addition of general ordinals is not commutative, and the right cancellation law also fails (\( \rho + \tau = \sigma + \tau \) does not imply \( \rho = \sigma \).) It can be proved by transfinite induction that the associative law does hold in general: \( (\rho + \sigma) + \tau = \rho + (\sigma + \tau) \) for any \( \rho, \sigma, \tau \) (Exercise 6.)

II.8.6.6. Subtraction of ordinals can be defined. If \( \tau < \sigma \), then there is a unique \( \rho \) such that \( \sigma = \tau + \rho \); \( \rho \) has the order type of \( \{\pi : \tau \leq \pi < \sigma\} \).
II.8.6.7. We can then define multiplication of ordinals in terms of addition. Informally, the well ordered set $\rho \cdot \sigma$ is obtained by taking the well ordered set $\sigma$ and replacing each element by a copy of $\rho$. One can describe this to be the set $\sigma \times \rho$ (note the order!) with lexicographic ordering. There is also an inductive definition. If $\rho$ is a fixed ordinal, we define $\rho \cdot \sigma$ to be $f(\sigma)$, where $f$ is defined by transfinite induction according to the following rule:

\[
f(0) = 0.
\]

If $\sigma = S(\tau)$, then $f(\sigma) = f(\tau) + \rho$.

If $\sigma$ is a nonzero limit ordinal, then $f(\sigma) = \lim\{f(\tau) : \tau < \sigma\}$

Then, for any $\rho$, $\rho \cdot 1 = \rho$, $\rho \cdot 2 = \rho + \rho$, etc.

II.8.6.8. As with addition, multiplication of finite ordinals coincides with ordinary multiplication of numbers, with the usual rules. But the commutative and right cancellation laws fail in general: $\omega \cdot 2 = \omega + \omega$, but it is easily checked that $2 \cdot \omega = \omega$. Ordinal multiplication is associative (Exercise).)

II.8.6.9. The next step is exponentiation. If $\rho$ is a fixed nonzero ordinal, we define $\rho^\sigma$ to be $f(\sigma)$, where $f$ is defined by transfinite induction according to the following rule:

\[
f(0) = 1.
\]

If $\sigma = S(\tau)$, then $f(\sigma) = f(\tau) \cdot \rho$.

If $\sigma$ is a nonzero limit ordinal, then $f(\sigma) = \lim\{f(\tau) : \tau < \sigma\}$

We define $0^\sigma = 0$ for all $\sigma \neq 0$ ($0^0$ is not defined.)

As usual, ordinal exponentiation coincides with usual exponentiation on the finite ordinals. In fact, for any $\rho$, we have $\rho^2 = \rho \cdot \rho$, etc.

This hierarchy can be continued, as will be indicated in the next subsection.

Countable Ordinals

We now give a more complete description of at least some of the countable ordinals.

First of all, note that the set of countable ordinals is closed under operations of finite or countable character:

II.8.6.10. Proposition.

(i) If $\sigma$ is a countable ordinal, so is $S(\sigma) = \sigma + 1$.

(ii) If $x = \{\sigma_i : i \in I\}$ is a countable set of countable ordinals, then $\lim(x)$ is countable.

(iii) If $\rho$ and $\sigma$ are countable ordinals, then $\rho + \sigma$, $\rho \cdot \sigma$, and $\rho^\sigma$ are countable.

Proof: Parts (i) and (ii) follow immediately from (i) (the Countable AC is needed for (ii)), and (iii) is easily proved by transfinite induction using (i) and (ii).

Countable ordinals can be pictured as subsets of $\mathbb{R}$ (in fact, of $\mathbb{Q}$ or $\mathbb{Q} \cap [0,1]$):

179
II.8.6.11. **Proposition.** Any subset of $\mathbb{R}$ which contains no strictly decreasing sequences is order-isomorphic (with respect to the induced ordering) to a countable ordinal. Every countable ordinal is order-isomorphic to such a set.

**Proof:** Any such set $X$ is well ordered in the induced ordering, hence is order-isomorphic to an ordinal $\sigma$. To see that $X$ is countable, let $g: \sigma \rightarrow X$ be an order-isomorphism, and for each $\tau < \sigma$, let $f(\tau)$ be a rational number between $g(\tau)$ and $g(\tau + 1)$ (to make this precise and avoid using the Axiom of Choice, let $f(\tau)$ be the first element of $G \cap (g(\tau), g(\tau + 1))$ according to some fixed enumeration of $G$.) Then $f$ is an injective function from $\sigma$ to $G$. The last statement follows from II.5.4.13.

Note, however, that there is no *natural or standard* way of picturing a countable ordinal as a subset of $\mathbb{R}$.

II.8.6.12. Some of the countable ordinals described earlier can be pictured in $[0, \infty)$ as follows:

- $\omega$ as $\{1 - \frac{1}{n} : n \in \mathbb{N}\}$ (or as $\mathbb{N} \cup \{0\}$)
- $\omega + 1$ as $\{1 - \frac{1}{n} : n \in \mathbb{N}\} \cup \{1\}$
- $\omega + 2$ as $\{1 - \frac{1}{n} : n \in \mathbb{N}\} \cup \{1, 2\}$
- $\omega + n$ as $\{1 - \frac{1}{n} : n \in \mathbb{N}\} \cup \{1, \ldots, n\}$

$\omega \cdot 2 = \omega + \omega$ as $\{1 - \frac{1}{n}, 2 - \frac{1}{n} : n \in \mathbb{N}\}$ (or as $\{1 - \frac{1}{n} : n \in \mathbb{N}\} \cup \mathbb{N}$)

$\omega \cdot 2 + 1$ as $\{1 - \frac{1}{n}, 2 - \frac{1}{n} : n \in \mathbb{N}\} \cup \{2\}$

$\omega^2 = \omega \cdot \omega$ as $\{1 - \frac{1}{n}, 1 - \frac{1}{n} - \frac{1}{n^2} : m, n \in \mathbb{N}\}$

$\omega^2 + \omega$ as $\{1 - \frac{1}{n}, 2 - \frac{1}{n}, 1 - \frac{1}{n} - \frac{1}{n^2} : m, n \in \mathbb{N}\}$

$\omega^2 \cdot 2 = \omega^2 + \omega^2$ as $\{1 - \frac{1}{n}, 2 - \frac{1}{n}, 1 - \frac{1}{n} - \frac{1}{n^2}, 2 - \frac{1}{n} - \frac{1}{n^2} : m, n \in \mathbb{N}\}$

$\omega^3 = \omega^2 \cdot \omega$ as $\{1 - \frac{1}{n}, 1 - \frac{1}{n} - \frac{1}{n^2}, 1 - \frac{1}{n} - \frac{1}{n^2} - \frac{1}{n^3} : m, n, k \in \mathbb{N}\}$

The intermediate and succeeding ordinals can be drawn in a similar manner. Note that the blocks between the ordinals shown grow successively longer: between $\omega^2$ and $\omega^3$ are all the ordinals of the form $\omega^2 \cdot n + \omega \cdot m + k$, for any $m, n, k, n \in \mathbb{N} \cup \{0\}$, $n > 0$.

II.8.6.13. We can make much larger countable ordinals. For example, $\omega^\omega = \lim \{\omega^n : n \in \mathbb{N}\}$. The elements of $\omega^\omega$ (i.e. the set of ordinals less than $\omega^\omega$) can be pictured as a subset of $\mathbb{R}$ (try it!) There is another way to describe this as an ordered set: they can be identified with the set of all polynomials with coefficients in $\mathbb{N} \cup \{0\}$, thought of as functions from $\mathbb{N}$ to $\mathbb{N}$, with $f < g$ if $f(n) < g(n)$ for all sufficiently large $n$. (This ordering is called the Fréchet ordering.)
II.8.6.14. We can go much farther. We can form \( \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \) etc. [As usual, these are grouped from the top down, e.g. \( \omega^{\omega^\omega} = \omega^{(\omega^\omega)}. \)] There is a limit to this sequence, i.e. a smallest ordinal larger than all these, which is countable and usually called \( \epsilon_0. \) See II.8.7.2. for a continuation.

II.8.6.15. The vertical arrow notation of Donald Knuth [Knu76] is useful here: if we define \( \uparrow \uparrow \sigma = \rho^\sigma \) for any \( \rho \) and \( \sigma, \) we can define

\[
\rho \uparrow \uparrow n = \rho \uparrow (\rho \uparrow \cdots (\rho \uparrow \rho) \cdots)) = \rho^{\rho^{\cdots}}
\]

(\( n \) levels), and \( \rho \uparrow \uparrow \omega = \lim \{\rho \uparrow \uparrow n : n \in \mathbb{N}\}. \) More generally, \( \rho \uparrow \uparrow \sigma \) for any \( \rho, \sigma \) can be defined inductively by setting \( \rho \uparrow \uparrow S(\sigma) = \rho \uparrow (\rho \uparrow \uparrow \sigma) = \rho^{\rho^{\cdots \rho}} \) and \( \rho \uparrow \uparrow (\lim(\sigma_i)) = \lim(\rho \uparrow \uparrow \sigma_i). \) (This operation is called tetration, and \( \rho \uparrow \uparrow \sigma \) is sometimes written \( \^\sigma \rho. \)) With this notation, \( \epsilon_0 = \omega \uparrow \uparrow \omega. \)

II.8.6.16. We can go on and define \( \uparrow \uparrow \uparrow \sigma \) starting by

\[
\rho \uparrow \uparrow \uparrow n = \rho \uparrow \uparrow (\rho \uparrow \uparrow \cdots (\rho \uparrow \rho) \cdots))
\]

(\( n \) levels), and then \( \rho \uparrow \uparrow \uparrow \sigma \) inductively for all \( \sigma. \) Similarly, for any \( n \) we can define \( \rho \uparrow \uparrow \uparrow \downarrow \sigma, \) \( \rho \uparrow \uparrow \downarrow \sigma = \rho \uparrow \cdots \uparrow \sigma, \) etc. The sequence

\[
\{\omega \uparrow \omega, \omega \uparrow \uparrow \omega, \omega \uparrow \uparrow \uparrow \omega, \omega \uparrow \uparrow \cdots \omega, \ldots \}
\]

has a limit which is again a countable ordinal.

II.8.6.17. We can still go on. We can define \( \uparrow \uparrow \sigma \) for any \( \rho, \sigma, \tau \) (Exercise ()), and then \( \rho \uparrow \sigma = \rho \uparrow ^\sigma \rho, \)

\[
\rho \uparrow \uparrow n = \rho \uparrow (\rho \uparrow \cdots (\rho \uparrow \rho) \cdots))
\]

(\( n \) levels), \( \rho \uparrow \uparrow \sigma, \rho \uparrow \uparrow \uparrow \sigma \) etc., etc., \ldots (see also II.8.7.4.) And if we run out of operations, we can always take a (countable) limit, take the successor of it, and start over again!

Even at that we have barely scratched the surface of the set of all countable ordinals. An internet search of “countable ordinals” will turn up numerous articles, from popular to technical, on naming or describing countable ordinals, some of which go far beyond what we have done in depth and sophistication (but even so fall far short of exhausting the countable ordinals). As Rudy Rucker puts it in his book, *Infinity and the Mind* [Ruc82],

“\[\text{In trying to think of bigger and bigger ordinals, one sinks into a kind of endless morass. Any procedure you come up with for naming larger ordinals eventually peters out, and the ordinals keep on coming. Finally, your mind snaps, and maybe you get a momentary glimpse of what the Absolute infinite is about. Then you try to formalize your glimpse, and you end up with a new system for naming ordinals \ldots which eventually peters out \ldots}\]

And this comment was just made in reference to the countable ordinals!
II.8.6.18. The first uncountable ordinal $\omega_1$ cannot be pictured as a subset of $\mathbb{R}$. However, it is order-isomorphic to a set of functions from $\mathbb{N} \cup \{0\}$ to $\mathbb{N} \cup \{0\}$ with Fréchet ordering, although there is no natural way to do it. An order-embedding $f : \sigma \mapsto f_\sigma$ can be defined by transfinite induction as follows. Let $f_0$ be the constant function 0. If $\sigma = \tau + 1$, let $f_\sigma(n) = f_\tau(n) + 1$ for all $n$. Now suppose $\sigma$ is a limit ordinal. It follows easily from II.8.6.11. that there is a strictly increasing sequence $(\sigma_n)$ of ordinals with $\sigma = \text{lim}(\sigma_n)$. Choose such a sequence (some version of the Axiom of Choice is needed to do this rigorously). Set $k_1 = 1$, and for $n > 1$ fix $k_n \in \mathbb{N}$ with $f_{\sigma_n}(m) > f_{\sigma_{n-1}}(m)$ for all $m \geq k_n$. Then set $f_\sigma(m) = f_{\sigma_n}(m)$ if $k_n \leq m < k_{n+1}$.

It is easily checked that this $f$ is an order-embedding. The function $f$ can be chosen to use the polynomial embedding for $\omega^\omega$, and can be extended in a standard way to the ordinals previously considered using the functions $x \uparrow n$, $x \uparrow^* x$, $x \uparrow^* n$, etc. [These formulas all give functions from $\mathbb{N}$ to $\mathbb{N}$, although they grow unimaginably fast: for example, even $4 \uparrow \uparrow 4 = 4^{4^{32}}$ is a number of more than $10^{154}$ digits, much larger than a googolplex, and $3 \uparrow \uparrow \uparrow 3 = 3 \uparrow (3 \uparrow 3) = 3 \uparrow \uparrow 3^{27} = \omega^{33 \cdot 3}$ (327 levels) is far too large to write using conventional notation.]

II.8.7. Exercises

II.8.7.1. Give a direct proof along the lines of (8) that the Axiom of Choice implies the Well-Ordering Principle. [Hint: If $X$ is a set, fix $y \notin X$ and a choice function $c : \mathcal{P}(X) \setminus \{\emptyset\} \to X$; set $f(\sigma) = c(X \setminus \{f(\tau) : \tau < \sigma\})$ if this set is nonempty, and $f(\sigma) = y$ otherwise.]

II.8.7.2. An $\epsilon$-number (defined by CANTOR) is an ordinal $\epsilon$ such that $\omega^\epsilon = \epsilon$.
(a) Show that $\epsilon_0$ (II.8.6.14.) is the smallest $\epsilon$-number.
(b) Show that the next $\epsilon$-number is
$$\epsilon_1 = \epsilon_0 \uparrow \omega = \text{lim}(\epsilon_0, \epsilon_0^\omega, \ldots) = \text{lim}(\omega^{\epsilon_0+1}, \omega^{\epsilon_0+1}, \ldots).$$
(c) Show that the $\epsilon$-numbers, in increasing order, are indexed by ordinals, and that $\epsilon_\alpha$ is countable if (and only if) $\alpha$ is countable.
(d) The ordinal $\Gamma_0 = \text{lim}(\epsilon_0, \epsilon_{\epsilon_0}, \epsilon_{\epsilon_{\epsilon_0}}, \ldots)$ is the smallest ordinal $\alpha$ such that $\epsilon_\alpha = \alpha$. $\Gamma_0$ is countable.

II.8.7.3. (a) Give a careful inductive definition of $\rho \uparrow \uparrow \sigma$ as outlined in II.8.6.15.
(b) Give a careful inductive definition of $\rho \uparrow \uparrow \uparrow \sigma$ and $\rho \uparrow \uparrow n \sigma$ for each $n$.
(c) Use transfinite induction on $\sigma$ to define $\uparrow^n \sigma$ for all $\sigma$ in terms of $\uparrow^\pi \sigma$ for all $\pi < \sigma$, i.e. suppose $\rho \uparrow^\pi \sigma$ has been defined for all $\pi < \sigma$ and all $\rho, \tau$. In defining $\rho \uparrow^\sigma \tau$, one may also assume that $\rho \uparrow^\sigma \pi$ has been defined for all $\pi < \tau$ (this assumption must be justified; consider lexicographic ordering.)
(d) Give a careful inductive definition of $\rho \uparrow \uparrow \sigma$, $\rho \uparrow \uparrow \uparrow \sigma$, $\rho \uparrow^\sigma \tau$, etc. Continue the hierarchy as far as possible.
II.8.7.4. Define Conway’s “chained arrow notation” as follows. All Latin letters denote natural numbers.

(i) \( a \rightarrow b \) denotes \( a^b \).

(ii) \( a \cdots x \rightarrow y \rightarrow 1 \) denotes \( a \cdots x \rightarrow y \).

(iii) \( a \cdots x \rightarrow y \rightarrow (z + 1) \) denotes

\[
\begin{align*}
& a \cdots x \text{ if } y = 1 \\
& a \cdots x \rightarrow (a \cdots x) \rightarrow z \text{ if } y = 2 \\
& a \cdots x \rightarrow (a \cdots x \rightarrow (a \cdots x) \rightarrow z) \rightarrow z \text{ if } y = 3 \\
& \quad \text{ etc.}
\end{align*}
\]

(a) Show that \( a \rightarrow b \rightarrow c = a \uparrow^c b \) for any \( a, b, c \).

(b) Show how to define \( \alpha \rightarrow \beta \rightarrow \cdots \rightarrow \sigma \) for any ordinals \( \alpha, \beta, \ldots, \sigma \). Show that these are again countable ordinals if \( \alpha, \beta, \ldots, \sigma \) are countable.

(c) Define \( a \Rightarrow b \) to be \( a \rightarrow a \rightarrow \cdots \rightarrow a (b \text{ terms}) \). Define \( a \Rightarrow b \Rightarrow \cdots \Rightarrow x \) inductively as for \( a \rightarrow b \rightarrow \cdots \rightarrow x \). Extend this to show how to define \( \alpha \Rightarrow \beta \Rightarrow \cdots \Rightarrow \sigma \) for any ordinals \( \alpha, \beta, \ldots, \sigma \). Show that these are again countable ordinals if \( \alpha, \beta, \ldots, \sigma \) are countable.

II.8.7.5. Transitive Sets. (a) Show that every nonempty transitive set contains \( \emptyset \) as an element. [If \( x \) is transitive, consider a minimal element of \( x \).]

(b) If \( x \) is any set, show there is a smallest transitive set containing \( x \). [Let \( x_1 = x \) and \( x_{n+1} = \cup x_n = \cup \{ y : y \in x_n \} \); then \( \cup \{ x_n : n \in \mathbb{N} \} \) is transitive.]

II.8.7.6. The ZF Universe. Define sets \( V_\sigma, \sigma \in \Omega \), inductively by letting \( V_0 = \emptyset, V_{\sigma+1} = \mathcal{P}(V_\sigma) \), and \( V_\sigma = \cup \{ V_\tau : \tau < \sigma \} \) if \( \sigma \) is a limit ordinal. Let \( V = \cup_\sigma V_\sigma \). Then \( V \) is a class which is not a set.

(a) Show by induction that each \( V_\sigma \) is transitive, and \( V_\tau \subseteq V_\sigma \) and \( V_\tau \in V_\sigma \) if \( \tau < \sigma \).

(b) Show that if \( x \) is a set and \( x \subseteq V \), then \( x \in V \). [For each \( y \in x \), let \( \sigma_y \) be the smallest \( \tau \) such that \( y \in V_\tau \), and set \( \sigma = \lim \{ \sigma_y : y \in x \} \).]

(c) Show that \( V \) is the entire universe \( \mathcal{U} \). [Suppose \( x \) is a transitive set which is not in \( V \). Then \( x \not\subseteq V \) by (b), so \( y = \{ z \in x : z \notin V \} \neq \emptyset \). Let \( z \) be a minimal element of \( y \). Then \( z \notin V \), so \( z \notin V \) by (b). If \( w \in z \subseteq x, w \notin V \), then \( w \in y \), a contradiction. Apply Exercise ()(b).]

(d) Give an explicit description of the sets in \( V_\omega \). (Note that all these sets are finite, and \( \text{card}(V_\omega) = \aleph_0 \).]

Thus every set is constructed from \( \emptyset \) by (transfinitely) repeated operations of subset, union, and power set. The sets in \( V_\sigma \) are called the sets constructed by the \( \sigma \)-th stage, or the sets constructible in \( \sigma \) steps.

The class \( V \) is called the von Neumann universe. There is an apparently smaller universe \( L \), called the Gödel constructible universe, where only “constructible” subsets are added at each stage. The hypothesis \( V = L \) is called the Axiom of Constructibility.

II.8.7.7. An ordinal \( \tau \) is called even if it is a limit ordinal or of the form \( \rho + \sigma 2n \) for a limit ordinal \( \rho \); otherwise \( \tau \) is odd. If \( \sigma \) is an ordinal, let \( \sigma_{ev} \) and \( \sigma_{odd} \) be the set of even and odd ordinals less than \( \sigma \), respectively. Then \( \sigma_{ev} \) and \( \sigma_{odd} \) are disjoint, and \( \sigma_{ev} \cup \sigma_{odd} = \sigma \). Show that if \( \sigma \) is a limit ordinal, then the maps \( f(\rho + \sigma 2n) = \rho + n \) and \( f(\rho + \sigma 2n + 1) = \rho + n, \) as \( \rho \) runs over all limit ordinals less than \( \sigma \), give bijections from \( \sigma_{ev} \) and \( \sigma_{odd} \) onto \( \sigma \).
II.9. Cardinals

We now want to make precise the notion of the cardinality of a set. In the course of this section, we will define \( \text{card}(X) \) for any set \( X \). To avoid confusion, we will use \(|X|\) to denote the informal notion of cardinality developed in (:); so the statements \(|X| = |Y|\), \(|X| \leq |Y|\), \(|X| < |Y|\) make sense for sets \( X, Y \), although we avoid attaching any meaning to \( |X| \) itself.

The idea is that we want to have a standard “model set” \( \text{card}(X) \) for any set \( X \), with the properties that 

\[ |\text{card}(X)| = |X| \]

for any \( X \) and, for any \( X, Y \), \( \text{card}(X) = \text{card}(Y) \) if and only if \(|X| = |Y|\). This process only really works well in the presence of the Axiom of Choice. Indeed, what we want to do is a form of Choice, identifying one model set from each equipotence equivalence class. We will, however, not make a blanket assumption of the Axiom of Choice, since the delineation of what can and cannot be proved without it gives good insight into the nature and role of the Axiom of Choice. Results which are dependent on the Axiom of Choice will be denoted by \((\text{AC})\); some of these can be obtained using weakened Choice axioms such as Dependent Choice (:).

II.9.1. Cardinals as Ordinals

Cardinals are ordinals of a special type:

II.9.1.1. Definition. An ordinal \( \sigma \) is a cardinal, or cardinal number, if \( |\tau| < |\sigma| \) for every \( \tau < \sigma \). A cardinal is finite [infinite] if it is finite [infinite] as an ordinal.

The letters \( \kappa, \lambda, \chi \) are frequently used to denote general cardinals, and \( \aleph \) is also used to denote a general infinite cardinal (if the Axiom of Choice is not assumed, the symbol \( \aleph \) is reserved for only certain infinite cardinals: see (:).) The class of cardinals has a natural ordering induced from \( \Omega \), under which it is well ordered (in particular, totally ordered.)

II.9.1.2. Examples.

(i) Any finite ordinal is a cardinal.

(ii) \( \omega \) is a cardinal. No other countable infinite ordinal is a cardinal.

(iii) \( \omega_1 \) is a cardinal.

When regarded as cardinals, \( \omega \) and \( \omega_1 \) are usually denoted \( \aleph_0 \) and \( \aleph_1 \) respectively. \( \aleph_0 \) is the smallest infinite cardinal, and \( \aleph_1 \) is the smallest uncountable cardinal and is the smallest cardinal greater than \( \aleph_0 \).

II.9.1.3. Proposition. If \( \kappa \) and \( \lambda \) are cardinals, then \( |\kappa| < |\lambda| \) if and only if \( \kappa < \lambda \).

This is obvious from the definition.

So far, we have not needed the Axiom of Choice. But the Axiom of Choice is crucial in the next result, which is the key fact about cardinals.
II.9.1.4. **Theorem.** (AC) Let $X$ be any set. Then there is a unique cardinal $\kappa$ such that $|X| = |\kappa|$.

**Proof:** By the Well-Ordering Principle, there is a $\sigma \in \Omega$ with $|X| = |\sigma|$. Let $\kappa$ be the smallest such $\sigma$. Then $\kappa$ is clearly a cardinal. Uniqueness follows from II.9.1.3.

[This result is actually equivalent to the Well-Ordering Principle, as is easily seen.]

II.9.1.5. **Definition.** If $X$ is a set, then the **cardinality** of $X$, denoted $\text{card}(X)$, is the unique cardinal $\kappa$ with $|X| = |\kappa|$.

II.9.1.6. **Examples.**

(i) If $X$ is a finite set with $n$ elements, then $\text{card}(X) = n$.

(ii) If $X$ is countably infinite, then $\text{card}(X) = \aleph_0$.

(iii) If $\kappa$ is a cardinal, then $\text{card}(\kappa) = \kappa$.

II.9.1.7. If $Y \subseteq X$, then $\text{card}(Y) \leq \text{card}(X)$. We can have $\text{card}(Y) = \text{card}(X)$ even if $Y$ is a proper subset of $X$, provided $X$ and $Y$ are infinite. If $\sigma$ and $\tau$ are ordinals, and $\tau < \sigma$, then $\text{card}(\tau) \leq \text{card}(\sigma)$. We can have $\text{card}(\tau) = \text{card}(\sigma)$; for example, $\text{card}(\omega) = \text{card}(\omega + 1) = \aleph_0$. We have $\text{card}(\sigma) \leq \sigma$ for any ordinal $\sigma$.

II.9.1.8. If $\{\kappa_i : i \in I\}$ is a set of cardinals, there is a smallest cardinal $\kappa$ such that $\kappa_i \leq \kappa$ for all $i \in I$. This $\kappa$ is denoted $\lim\{\kappa_i : i \in I\}$ (it is the same as $\lim\{\kappa_i : i \in I\}$ in the ordinal sense.)

II.9.1.9. **Theorem.** If $\sigma$ is any ordinal, then there is a cardinal $\kappa$ with $\sigma < \kappa$.

This is easy to prove using the Axiom of Choice: we can take $\kappa = \text{card}(\mathcal{P}(\sigma))$. However, it can be proved without the Axiom of Choice. We first prove a lemma:

II.9.1.10. **Lemma.** Let $\sigma$ be an ordinal. Then the class $Z(\sigma) = \{\tau \in \Omega : |\tau| = |\sigma|\}$ is a set.

**Proof:** Let $s$ be a set with $|s| = |\sigma|$ (we could take $s = \sigma$.) If $\tau \in Z(\sigma)$, then there is a well-ordering on $s$ making $s$ order-isomorphic to $\tau$; thus the collection $\mathcal{R}_{\tau}$ of all well orderings on $s$ order-isomorphic to $\tau$ is nonempty. $\mathcal{R}_{\tau}$ is contained in the set of all relations on $s$, which can be identified with $\mathcal{P}(s \times s)$. If $\tau, \rho \in Z(\sigma)$, $\tau \neq \rho$, then $\mathcal{R}_{\tau}$ and $\mathcal{R}_{\rho}$ are disjoint and thus distinct. Thus $\tau \mapsto \mathcal{R}_{\tau}$ is an injective map from $Z(\sigma)$ to $\mathcal{P}(\mathcal{P}(s \times s))$, which is a set. 

185
II.9.1.11. **Corollary.** Let \( \sigma \) be an infinite ordinal, and \( \kappa = |Z(\sigma)| \). Then \( \kappa \) is a cardinal; \( |\sigma| < \kappa \), and \( \kappa \) is the smallest cardinal greater than \( \text{card}(\sigma) \). In fact, \( \kappa = \{ \tau \in \Omega : |\tau| \leq |\sigma| \} \).

**Proof:** The last formula is easy to verify, and the other statements follow immediately (note that \( \kappa \notin \kappa \)).

II.9.1.12. **Corollary.** The class of cardinals is not a set.

II.9.1.13. **Definition.** For each ordinal \( \sigma \), we (inductively) define \( \aleph_\sigma \) to be the smallest infinite cardinal (aleph) greater than \( \{ \aleph_\tau : \tau < \sigma \} \).

By II.9.1.11, \( \aleph_\sigma \) is defined for every ordinal \( \sigma \). This gives a one-to-one order-preserving correspondence \( \sigma \leftrightarrow \aleph_\sigma \) between the ordinals and the infinite cardinals (alephs).

II.9.1.14. **Definition.** The cardinal \( \aleph_\sigma \) is a **successor cardinal** if \( \sigma \) is a successor ordinal, and a **limit cardinal** if \( \sigma \) is a limit ordinal.

By this definition, \( \aleph_0 \) and \( \aleph_\omega \) are examples of limit cardinals, and \( \aleph_1 \) is a successor cardinal.

II.9.1.15. Since every cardinal is an ordinal, the function \( \sigma \mapsto \aleph_\sigma \) can be regarded as a map from \( \Omega \) to \( \Omega \). It is easy to prove by induction that \( \sigma \leq \aleph_\sigma \) for every \( \sigma \), and intuitively it seems that \( \aleph_\sigma \) is much bigger than \( \sigma \). Put another way: if \( \sigma \) is an ordinal, then it seems intuitively that \( \{ \kappa : \kappa \text{ a cardinal, } \kappa < \sigma \} \) should be quite a small subset of \( \sigma \), i.e. the cardinals should be sparsely distributed among the ordinals. However, it is not hard to show (Exercise II.9.7.10.) that there is a \( \sigma \) for which \( \aleph_\sigma = \aleph_\sigma \). (Such a \( \sigma \) is obviously a cardinal.) We discuss related matters in ().

II.9.2. **General Cardinals and Alephs**

II.9.2.1. We have so far only defined cardinals of sets which can be well ordered; these cardinals, if infinite, are called **alephs**. In the absence of the Axiom of Choice, there will be sets which cannot be well ordered and whose cardinal is as yet undefined. As described in the introduction, we want to pick a “model set” to serve as the cardinal of such a set. This can only be done rigorously in the presence of the Axiom of Regularity, and even then in a not entirely satisfactory way without the Axiom of Choice. However, even without the Axiom of Regularity it is possible to use the whole cardinal terminology as simply shorthand for statements about sets which would otherwise be cumbersome and inelegant. For example, II.9.7.11.(a) can be phrased without the language of cardinals as follows:

“Let \( X \) be a set containing a countably infinite subset. If \( Y \) is any set disjoint from \( X \) such that \( X \cup Y \) is equipotent with \( \mathcal{P}(X) \), then \( Y \) is equipotent with \( \mathcal{P}(X) \).”

We will thus be slightly imprecise and assume that \( \text{card}(X) \) has been defined for every set \( X \), so that we can write such statements as “Let \( \kappa \) be a cardinal.” The usual “ordering” gives a partial ordering on the class of all cardinals (the antisymmetric property follows from the Schröder-Bernstein Theorem.)
II.9.2.2. Infinite cardinals which are not alephs (if they exist) are not comparable with alephs in general. Although infinite cardinals, alephs are “small” in the sense that any infinite cardinal dominated by an aleph is again an aleph. However, there is no cardinal larger than all the alephs, since the class of alephs do not form a set (II.9.1.12.) Thus the statements in the previous section must be modified in the absence of the Axiom of Choice by interpreting “cardinal” to mean “cardinal of a well ordered set”, i.e. a natural number or aleph. For example, $\aleph_\sigma$ is the smallest aleph greater than $\{\aleph_\tau : \tau < \sigma\}$: there could be another cardinal $\kappa$ greater than all these alephs but incomparable with $\aleph_\sigma$. There could even be an infinite cardinal which is incomparable with $\aleph_0$! 

II.9.2.3. One different way “cardinality” of a set could be defined is to associate to a set $X$ the partially ordered set $S(X)$ of equipotence classes of subsets of $X$. This partially ordered set always has a largest element, and, if $X$ is infinite, contains a hereditary subset order-isomorphic to $\omega$. The set $S(X)$ is totally ordered if and only if $S$ can be well ordered; thus the AC is equivalent to the statement that $S(X)$ is always totally ordered. In this case, $S(X)$ is well ordered; if $X$ has cardinality $\aleph_\alpha$, then $S(X) \cong \omega + \alpha + 1$. However, in the absence of AC it is unclear whether $S(X) \cong S(Y)$ implies that $X$ and $Y$ are equipotent. It would be interesting to characterize which partially ordered sets occur as $S(X)$ for some $X$ (in the absence of AC).

II.9.3. Cardinal Arithmetic

Cardinal arithmetic is easy to define. However, it must be kept in mind that it is not the same as ordinal arithmetic for infinite cardinals (alephs), and one must always carefully distinguish whether cardinal or ordinal arithmetic is being used. The distinction is usually clear from the context; but if there is any possibility of confusion we will use $+_c$ and $+_o$ for cardinal and ordinal addition respectively, and similarly $\cdot_c$ and $\cdot_o$. We can also use the vertical arrow notation for exponentiation to distinguish between cardinal and ordinal exponentiation (where the distinction is more crucial), writing $\sigma \uparrow_o \tau$ for $\sigma^\tau$ in the ordinal sense, and $\kappa \uparrow_c \lambda$ for cardinal exponentiation.

II.9.3.1. Definition. Let $\kappa$ and $\lambda$ be cardinals, and let $X$ and $Y$ be disjoint sets (see Exercise II.5.6.15.) with $\text{card}(X) = \kappa$, $\text{card}(Y) = \lambda$. Then

- $\kappa + \lambda = \kappa +_c \lambda = \text{card}(X \cup Y)$.
- $\kappa \cdot \lambda = \kappa \cdot_o \lambda = \text{card}(X \times Y)$.
- $\kappa^\lambda = \kappa \uparrow_c \lambda = \text{card}(X^Y)$ (where $X^Y$ denotes the set of functions from $Y$ to $X$.)

It is clear that this definition does not depend on the choice of the sets $X$ and $Y$.

The definition of cardinal exponentiation gives the explanation for the notation $\text{card}(\mathcal{P}(\mathbb{N})) = \text{card}(\{0, 1\}^\mathbb{N}) = 2^{\aleph_0}$, since for set theorists $2 = \{0, 1\}$.

Cardinal addition and multiplication agree with ordinary addition and multiplication for finite cardinals. Cardinal addition and multiplication are not very interesting for infinite cardinals, at least for alephs:

II.9.3.2. Theorem. Let $\kappa$ and $\lambda$ be nonzero cardinals of well-ordered sets, not both finite. Then

$\kappa + \lambda = \kappa \cdot \lambda = \max(\kappa, \lambda)$.

Note that $()$ and $()$ are special cases.
II.9.3.3. Exponentiation of cardinals is very interesting, however, and is quite different from ordinal exponentiation (for example, \( \omega^\omega = \omega \uparrow \omega \) is a countable ordinal, hence \( \text{card}(\omega^\omega) = \aleph_0 \), but \( \aleph_0^{\aleph_0} = \aleph_0 \uparrow \aleph_0 \) is an uncountable cardinal.)

Cardinal addition, multiplication, and exponentiation satisfy the usual algebraic rules (Exercise ().) However, subtraction and division are another matter: they are usually not well defined.

II.9.3.4. The sum of an infinite set of cardinals can also be defined. If \( \{\kappa_i : i \in I\} \) is an indexed set of cardinals, and \( \{X_i : i \in I\} \) is a collection of disjoint sets (see Exercise ()) with \( \text{card}(X_i) = \kappa_i \) (e.g. \( X_i = \{i\} \times \kappa_i \)), then \( \sum_i \kappa_i = \text{card}(\bigcup_i X_i) \). If \( \kappa = \lim \{\kappa_i : i \in I\} \) is an aleph and \( \text{card}(I) \leq \kappa \) (in particular, if the \( \kappa_i \) are all distinct), then \( \sum_{i \in I} \kappa_i = \kappa \). Infinite products can be defined similarly.

II.9.4. The Role of the Axiom of Choice

In this subsection, we explore what can and cannot be proved about cardinals without assuming the Axiom of Choice (AC).

II.9.4.1. If \( \aleph \) is an aleph, then \( \aleph + \aleph = \aleph \) and \( \aleph^2 = \aleph \aleph = \aleph \). It cannot be proved without AC that \( \kappa + \kappa = \kappa \) or \( \kappa^2 = \kappa \) for every infinite cardinal \( \kappa \). In fact, the statement that \( \kappa^3 = \kappa \) for every infinite cardinal \( \kappa \) implies (hence is equivalent to) AC [?, XVI.2 Theorem 6]. The statement that \( \kappa + \kappa = \kappa \) for all infinite \( \kappa \) is strictly weaker than AC ([?]; cf. Exercise II.9.7.4.). It can be proved without AC that some infinite cardinals not provably alephs, such as \( \kappa = 2^{\aleph_0} \), satisfy \( \kappa + \kappa = \kappa^2 = \kappa \); see Exercise (). [?, XVI.2] contains several other similar statements equivalent to AC (see Exercise ().)

If AC is not assumed, the class of cardinals is not totally ordered by \( \leq \) (in fact, \( \leq \) being a total order is equivalent to AC, cf. II.9.7.1.). In the absence of AC, there is another interesting and potentially different “ordering” on cardinals:

II.9.4.2. Definition. Let \( \kappa = \text{card}(X) \) and \( \lambda = \text{card}(Y) \) be cardinals. Then \( \kappa \not\lesssim \lambda \) if there is a surjective function from \( Y \) to \( X \).

The relation \( \not\lesssim \) is clearly well defined on cardinals, and transitive. If \( 0 \neq \kappa \leq \lambda \), then \( \kappa \not\lesssim \lambda \) (Exercise ());

if AC is assumed, \( \leq \) and \( \not\lesssim \) coincide (except that \( 0 \not\lesssim \kappa \) if \( \kappa \neq 0 \).) In the absence of AC, it cannot be proved that \( \not\lesssim \) is antisymmetric (i.e. there is no analog of the Schröder-Bernstein Theorem; see Exercise ()(b).) It is provable without AC that \( \kappa \not\lesssim \lambda \) implies \( 2^\kappa \leq 2^\lambda \) (Exercise ()).

II.9.4.3. Proposition. Let \( \kappa = \text{card}(I) \) and \( \lambda = \text{card}(Y) \) be cardinals. Then \( \kappa \not\lesssim \lambda \) if and only if \( Y \) can be written as \( \bigcup_{i \in I} Y_i \), with the \( Y_i \) nonempty and disjoint.

Proof: If \( f : Y \to I \) is surjective, let \( Y_i = f^{-1}(\{i\}) \). Conversely, if \( Y = \bigcup Y_i \), define \( f : Y \to I \) by \( f(y) = i \) if \( y \in Y_i \).
II.9.4.4. If $\kappa$ is a cardinal, then it is not true that $\aleph_\sigma \leq \kappa$ for all $\sigma$. Thus there is a smallest aleph $\aleph$ such that $\aleph \leq \kappa$ is not true. This aleph is denoted $\aleph^+(\kappa)$. (Some authors such as [?] use the notation $\aleph(\kappa)$, but this is confusing since $\aleph^+(\aleph_\sigma) = \aleph_{\sigma+1}$, not $\aleph_\sigma$.) It is easily checked that, if $\kappa$ is infinite,

$$\aleph^+(\kappa) = \{ \sigma \in \Omega : \text{card}(\sigma) \leq \kappa \} .$$

II.9.4.5. Theorem. Let $\kappa$ be an infinite cardinal. Then $\aleph^+(\kappa) \leq 2^{2^\kappa}$, and hence $\aleph^+(\kappa) < 2^{\aleph^+(\kappa)} \leq 2^{2^{2^\kappa}}$.

For the proof, note that if $X$ is a set with $\text{card}(X) = \kappa$, then each subset of $\mathcal{P}(X)$ (element of $\mathcal{P}(\mathcal{P}(X))$) has a natural partial order induced by set inclusion.

II.9.4.6. Lemma. For each ordinal $\sigma < \aleph^+(\kappa)$, there is a subset of $\mathcal{P}(X)$ (element of $\mathcal{P}(\mathcal{P}(X))$) which is well ordered of order type $\sigma$.

Proof: Let $f$ be a bijection from $\sigma$ to a subset $Y$ of $X$. For $\rho \leq \sigma$, let $Y_\rho = \{ f(\tau) : \tau < \rho \}$. Then $\rho \to Y_\rho$ is an order-embedding of $\sigma$ into $\mathcal{P}(X)$.

Proof: We now prove Theorem II.9.4.5. Construct a function $f : \mathcal{P}(\mathcal{P}(X)) \to \aleph^+(\kappa)$ as follows. If $S \in \mathcal{P}(\mathcal{P}(X))$ is well ordered of order type $\tau$ for some $\sigma < \aleph^+(\kappa)$, set $f(S) = \sigma$; otherwise set $f(S) = 0$. By the lemma, $f$ is surjective. The last statement follows from (? and (?).

II.9.4.7. Remark. If we instead let $\Sigma^+(\kappa)$ be the set of all order types of well ordered elements of $\mathcal{P}(\mathcal{P}(X))$, then it is easily checked that $\Sigma^+(\kappa)$ is an ordinal $\geq \aleph^+(\kappa)$. But it is not obvious without AC that $\Sigma^+(\kappa)$ is a cardinal, or that $\text{card}(\Sigma^+(\kappa)) = \aleph^+(\kappa)$ (although the last statement is asserted without proof in [?, ]). Using the Axiom of Choice, if $f : \sigma \to A_\sigma$ is an order-preserving bijection, we can choose an element $x_\tau \in A_{\tau+1} \setminus A_\tau$ for each $\tau < \sigma$, and it follows easily that $\Sigma^+(\kappa) = \aleph^+(\kappa)$.

It cannot be proved without AC that $\aleph^+(\kappa) \leq 2^{2^\kappa}$ (but see Exercise (?).)

II.9.4.8. If $\kappa$ is finite, then $\aleph^+(\kappa) = \aleph_0$. Without AC, it is conceivable that there could be infinite $\kappa$ for which $\aleph^+(\kappa) = \aleph_0$, i.e. $\aleph_0 \not\leq \kappa$. It is even conceivable that there could be infinite $\kappa$ for which $\aleph_0 \not\leq \kappa$. Thus, without AC, there are three possible definitions of infinite sets or cardinals, which are successively more restrictive:

II.9.4.9. Definition. Let $\kappa = \text{card}(X)$ be a cardinal.

$X$ is an infinite set $(\kappa$ is an infinite cardinal) if $\kappa \neq n$ for every $n \in \aleph$.

$X$ or $\kappa$ is countably decomposable if $\aleph_0 \not\leq \kappa$, i.e. if $X$ can be decomposed as a disjoint union of a countably infinite number of disjoint nonempty subsets.

$X$ or $\kappa$ is infinite in the sense of Dedekind if $\aleph_0 \leq \kappa$, i.e. if $X$ contains a countably infinite subset.

Dedekind’s definition was actually that a set is infinite if it can be put in one-one correspondence with a proper subset of itself. It is easily seen (?) that this condition is equivalent to the presence of a countably infinite subset.
It can be easily shown by induction () that for any \( \kappa \), either \( \kappa = n \) for some \( n \in \mathbb{N} \) or else \( n < \kappa \) for all \( n \in \mathbb{N} \). Thus \( \kappa \) is infinite if and only if \( n < \kappa \) for all \( n \in \mathbb{N} \).

It is obvious that a countably decomposable set is infinite, since there is no surjective function from \( n \) to \( \mathbb{N} \) ().

II.9.4.10. **PROPOSITION.** Let \( X \) be a set, and \( \kappa = \text{card}(X) \). The following are equivalent:

(i) \( X \) (or \( \kappa \)) is infinite.

(ii) \( \aleph_0 \nless 2^\kappa \).

(iii) \( 2^{\aleph_0} \leq 2^{2^\kappa} \).

(iv) \( 2^{\aleph_0} < 2^{2^\kappa} \).

**Proof:** (i) \( \Rightarrow \) (ii) is Exercise (), and (ii) \( \Rightarrow \) (iii) is Exercise ()(b) (cf. Exercise ()(b).) (iv) \( \Rightarrow \) (iii) \( \Rightarrow \) (i) is obvious since (iii) implies that \( 2^{2^\kappa} \) is infinite. To show (iii) \( \Rightarrow \) (iv), it suffices to show that \( 2^{\aleph_0} \neq 2^{2^\kappa} \) for any \( \kappa \). If \( 2^{\aleph_0} = 2^{2^\kappa} \), we have \( \kappa < 2^\kappa < 2^{2^\kappa} = 2^{\aleph_0} \), contradicting Exercise ().

II.9.4.11. It can be shown (Exercise ()) that \( X \) is countably decomposable if and only if \( X \) can be written as a union of countably many distinct subsets (Exercise ()). It cannot be shown without some form of AC (Countable AC suffices) that every infinite cardinal is countably decomposable; but any infinite cardinal which is comparable with \( 2^{\aleph_0} \) is countably decomposable (Exercise ()). We have:

II.9.4.12. **PROPOSITION.** Let \( X \) be a set and \( \kappa = \text{card}(X) \). The following are equivalent:

(i) \( X \) (or \( \kappa \)) is countably decomposable.

(ii) \( \mathcal{P}(X) \) is infinite in the sense of Dedekind, i.e. \( \aleph_0 \leq 2^\kappa \).

(iii) \( 2^{\aleph_0} \leq |\mathcal{P}(X)| = 2^\kappa \).

**Proof:** (i) \( \Rightarrow \) (iii) is Exercise (), and (ii) \( \Rightarrow \) (i) is Exercise () since then there is a countably infinite set of distinct subsets of \( X \), whose union is \( X \) since \( X \) can be added to the collection of subsets. (iii) \( \Rightarrow \) (ii) is trivial since \( \aleph_0 < 2^{\aleph_0} \).

It cannot be shown without some form of AC (Countable AC suffices) that every countably decomposable set is infinite in the sense of Dedekind. The next proposition gives some alternate characterizations of infiniteness in the sense of Dedekind.

190
II.9.4.13. **Proposition.** Let \( X \) be a set, and \( \kappa = \text{card}(X) \). The following are equivalent:

(i) \( X \) is infinite in the sense of Dedekind.

(ii) \( \kappa = \kappa + \aleph_0 \).

(iii) \( \kappa = \kappa + 1 \).

(iv) If \( \kappa = \lambda + 1 \), then \( \lambda = \kappa \).

(v) If \( Y \) is any set obtained from \( X \) by adding or removing finitely many points, then \( \text{card}(Y) = \kappa \).

(vi) There is a proper subset \( Y \) of \( X \) which is equipotent with \( X \).

**Proof:** (i) \( \Rightarrow \) (ii), (i) \( \Rightarrow \) (iii), (i) \( \Rightarrow \) (iv), (i) \( \Rightarrow \) (v) are all proved the same way: if \( Z \) is a countably infinite subset of \( X \), and \( W \) is \( Z \) with a finite number of elements removed, or a finite or countable number of points added, then \( |Z| = |W| \) and \( |X| = |(X \setminus Z) \cup Z| = |(X \setminus Z) \cup W| \).

It is obvious that any of (ii), (iii), (iv), (v) imply (vi).

(vi) \( \Rightarrow \) (i): Let \( Y \) be a proper subset of \( X \) and \( f : X \to Y \) a bijection. Fix \( x \in X \setminus Y \), and let \( x_n = f^{(n)}(x) \), where \( f^{(n)} = f \circ f \circ \cdots \circ f \) (\( n \) times.) Then \( \{x_n : n \in \mathbb{N}\} \) is a countably infinite subset of \( X \).

II.9.5. The Continuum Hypothesis

If \( X \) is any set, and \( \text{card}(X) = \kappa \), then \( \text{card}(\mathcal{P}(X)) = \text{card}(\{0, 1\}^X) = 2^\kappa \). By Cantor’s Theorem (\( \dagger \)), \( \kappa < 2^\kappa \).

In aleph notation, \( \aleph_{\sigma + 1} \leq 2^{\aleph_{\sigma}} \) for every \( \sigma \). (Without AC, it can only be proved that \( \aleph_{\sigma + 1} \preceq 2^{\aleph_{\sigma}} \) (Exercise (\( \dagger \)).) In particular, \( \aleph_1 \leq 2^{\aleph_0} \). Do we have equality? This is equivalent (under AC) to the question whether every uncountable subset of \( \mathbb{R} \) has cardinality \( 2^{\aleph_0} \).

II.9.5.1. **Continuum Hypothesis (CH):** \( 2^{\aleph_0} = \aleph_1 \).

II.9.5.2. It was a big mathematical industry in the first part of the 20th century (actually beginning with Cantor in the 19th century) to try to prove (or, sometimes, to disprove) this hypothesis. Perhaps surprisingly, it turns out to be independent of ZFC, i.e. it cannot be either proved or disproved in ZFC!\(^{12}\)

Mathematicians to this day have had various opinions about whether the CH “should be” true, and it is much more customary than with AC to carefully note any use of the CH in proofs and to avoid using it whenever possible.

The Axiom of Choice is also independent of ZF+CH. However, the CH implies a weak version of AC: \( \mathbb{R} \) can be well ordered since it is equipotent with \( \omega_1 \). This in turn implies that there is a choice function for \( \mathcal{P}(\mathbb{R}) \) (\( \dagger \)). But the CH does not even imply the Countable Axiom of Choice.

If the AC is not assumed, then there is a weaker version of the Continuum Hypothesis which is not equivalent to the one above:

\(^{12}\)This statement is again under the assumption that ZF is consistent.
II.9.5.3. **Weak Continuum Hypothesis:** There is no cardinal strictly between $\aleph_0$ and $2^{\aleph_0}$, i.e. if $X \subseteq P(\mathbb{N})$ (or $X \subseteq \mathbb{R}$), then either $X$ is countable or $|X| = |P(\mathbb{N})| = |\mathbb{R}|$.

In fact there are many inequivalent ways to phrase the CH in the absence of the AC. See [http://math.stackexchange.com/questions/404807/how-to-formulate-continuum-hypothesis-without-the-axiom-of-choice?rq=1](http://math.stackexchange.com/questions/404807/how-to-formulate-continuum-hypothesis-without-the-axiom-of-choice?rq=1) for a discussion.

There is a generalized version of the Continuum Hypothesis:

II.9.5.4. **Generalized Continuum Hypothesis (Version 1):** $2^{\aleph_\sigma} = \aleph_{\sigma+1}$ for every $\sigma$.

This is also independent of ZFC (?). However, without the Axiom of Choice there is another version analogous to the Weak Continuum Hypothesis:

II.9.5.5. **Generalized Continuum Hypothesis (Version 2):** If $\kappa$ is any infinite cardinal, then there is no cardinal strictly between $\kappa$ and $2^\kappa$.

II.9.5.6. Note that Version 2 is not really weaker than Version 1, since it is an assertion about all sets, not just those which can be well ordered. In fact, it turns out that Version 2 implies the Axiom of Choice, and thus also Version 1 (Exercise ()).

The Continuum Hypothesis does not imply the Generalized Continuum Hypothesis Version 1. There are intermediate hypotheses, saying $2^{\aleph_\sigma} = \aleph_{\sigma+1}$ for every $\sigma$ up to some point but not beyond.

II.9.6. **Large Cardinals**

We will give only the briefest introduction to the bizarre world of large cardinals, emphasizing the measurable cardinals (as is appropriate in a book on real analysis.) In this section, “cardinal” means “cardinal of a well ordered set”; we may as well assume the Axiom of Choice.

II.9.6.1. **Definition.** Let $\kappa$ be a cardinal. Then $\kappa$ is regular if $\kappa$ is not equal to $\sum_{i \in I} \kappa_i$ for any set $\{\kappa_i : i \in I\}$ with $\text{card}(I) < \kappa$ and $\kappa_i < \kappa$. Otherwise $\kappa$ is singular.

Equivalently, $\kappa$ is regular if a set of cardinality $\kappa$ cannot be written as a union of fewer than $\kappa$ subsets, each of cardinality less than $\kappa$. Among the finite cardinals, only 0, 1, and 2 are regular. Among infinite cardinals, every successor cardinal is regular; $\aleph_0$ is regular, but “most” limit cardinals are singular, for example $\aleph_\omega = \sum_{n \in \mathbb{N}} \aleph_n$.

II.9.6.2. **Definition.** An infinite cardinal $\kappa$ is weakly inaccessible if it is an uncountable regular limit cardinal.

II.9.6.3. A weakly inaccessible cardinal cannot be reached from below by taking successor cardinals or sums of smaller sets of smaller cardinals. The set of cardinals less than a weakly inaccessible cardinal $\kappa$ has cardinality $\kappa$, the same as the set of ordinals less than $\kappa$ (i.e. $\kappa = \aleph_\kappa$), so the cardinals less than $\kappa$ are in this sense not so sparsely distributed among the ordinals. (But $\kappa = \aleph_\kappa$ does not imply that $\kappa$ is weakly inaccessible (Exercise ())).
II.9.6.4. There is one potential way to reach a weakly inaccessible cardinal: in the absence of the Continuum Hypothesis, even $2^\alpha$ might be so huge as to be weakly inaccessible, or exceed a weakly inaccessible cardinal. In fact, it appears to be consistent with ZFC to assume this. So we strengthen the definition of inaccessibility:

II.9.6.5. Definition. A cardinal $\kappa$ is inaccessible, or strongly inaccessible, if it is weakly inaccessible and $\lambda < \kappa$ implies $2^\lambda < \kappa$.

If the Generalized Continuum Hypothesis is assumed, there is no difference between weak and strong inaccessibility.

II.9.6.6. The question then becomes: do (weakly) inaccessible cardinals exist?

Inaccessible cardinals cannot be reached by any of the operations of set theory, and thus their existence cannot be proved in ZFC. For if there does exist an inaccessible cardinal in some model of ZFC, let $\kappa$ be the smallest inaccessible cardinal. Then the set $V_\kappa$ of Exercise () can be taken to be the universe of a model for ZFC in which there are no inaccessible cardinals. On the other hand, it appears to be consistent with ZFC to assume that there exists an inaccessible cardinal.

II.9.6.7. Note that the cardinals immediately larger than an inaccessible cardinal are not inaccessible. In fact, if there is a second inaccessible cardinal, it must be vastly larger than the first one. The same proof as in the previous paragraph shows that even if an inaccessible cardinal is assumed to exist, it cannot be proved in ZFC that a second one also exists. Thus the inaccessible cardinals are (seemingly) very sparsely distributed even among all cardinals, let alone among all ordinals.

II.9.6.8. Inaccessible cardinals are often called large cardinals. (Set theorists make a technical distinction between inaccessible cardinals and “large cardinals”, which have no precise widely-accepted definition, but inaccessible cardinals are large cardinals by any reasonable definition, and the classes of large cardinals which have been studied are inaccessible cardinals.) There is an enormous hierarchy of types of large cardinals which have been considered. For an analyst, the most interesting class is the class of measurable cardinals, discussed in the next section. See [Kan03] for a full discussion of all types of large cardinals, or [Ruc82] for a readable semitechnical description of some of them.

Large cardinals may seem far removed from the “real world”, but it turns out that existence axioms for certain kinds of large cardinals have close connections with natural questions about subsets of $\mathbb{R}$ and even $\mathbb{N}$. See () for a discussion of a few of these; [Kan03] has a full treatment.

II.9.7. Exercises

II.9.7.1. (a) Show that the AC (Well-Ordering Principle) implies that $\leq$ is a total order (in fact, well order) on the class of all cardinals.

(b) Conversely, suppose $\leq$ is a total order on the class of all cardinals. If $\kappa$ is a cardinal, show that $\kappa \leq \aleph^+ (\kappa)$ (II.9.4.4.), and hence $\kappa$ is an aleph. Conclude that AC holds.

Thus the AC is equivalent to the statement that the class of cardinals is totally ordered.
II.9.7.2. (a) If $\kappa$ is an infinite cardinal (aleph), then $\text{card}(\kappa_{\text{ev}}) = \text{card}(\kappa_{\text{odd}}) = \kappa$ (II.8.7.7). Thus $\kappa + \kappa = \kappa$.
(b) Conclude that if $\kappa$ and $\lambda$ are infinite cardinals (alephs), then $\kappa + \lambda = \max(\kappa, \lambda)$.
(c) Show that if $\kappa$ is an infinite cardinal (aleph), then $\kappa + c n = \kappa$ for each finite cardinal $n$.

II.9.7.3. Let $\kappa$ be an infinite cardinal (aleph). Define an ordering on $\kappa \times \kappa$ by setting $(\pi, \rho) < (\sigma, \tau)$ if $\max(\pi, \rho) < \max(\sigma, \tau)$, or if $\max(\pi, \rho) = \max(\sigma, \tau)$ and $\pi < \sigma$, or if $\max(\pi, \rho) < \max(\sigma, \tau)$, $\pi = \sigma$, and $\rho < \tau$.
(a) Show that this is a well ordering on $\kappa \times \kappa$. Draw a picture illustrating the ordering.
(b) Assume that $\lambda \cdot \lambda < \kappa$ for all cardinals $\lambda < \kappa$. Show that every element of $\kappa \times \kappa$ has fewer than $\kappa$ predecessors under this ordering, so if $\sigma$ is the order type of the well ordering, then $\sigma \leq \kappa$. Conclude that $\text{card}(\kappa \cdot \kappa) \leq \kappa$. (The opposite inequality also obviously holds.)
(c) Prove by transfinite induction on $\kappa$ that $\kappa \cdot \kappa = \kappa$ if $\kappa$ is an aleph.
(d) Conclude that if $\lambda \neq 0$ is finite or an aleph, then $\kappa \cdot \lambda = \max(\kappa, \lambda)$.

II.9.7.4. (a) Let $\kappa$ be a cardinal. If $\kappa + \kappa = \kappa$, show that $\aleph_0 \cdot \kappa = \kappa$. [Let $X$ be a set with $|X| = \kappa$ and $f$ and $g$ one-to-one functions from $X$ to $X$ with disjoint ranges. Consider the functions $g \circ g \circ \ldots \circ g \circ f$.]
(b) Show that $\text{AC}$ implies that $\kappa + \kappa = \kappa$ for every infinite cardinal $\kappa$. The converse implication does not hold [].
(c) Assume $\kappa + \kappa = \kappa$ for all $\kappa \geq \aleph_0$. Show (without $\text{AC}$) that every infinite cardinal is Dedekind-infinite, i.e. $\geq \aleph_0$ [].

II.9.7.5. Suppose $X$ is a set of cardinality $\chi$, and there is a purely nonatomic (countably additive) $\{0,1\}$-valued measure $\mu$ on $(X, \mathcal{P}(X))$ with $\mu(X) = 1$.
(a) Suppose, for some $\kappa < \chi$, there are sets $\{A_i : i \in I\}$ with each $A_i \subseteq X$ and $\text{card}(I) = \kappa$, such that $\mu(A_i) = 0$ for all $i$ but $\mu(\bigcup_{i \in I} A_i) = 1$, and suppose $\kappa$ is the smallest cardinal for which this happens. If $E \subseteq I$, define $\nu(E) = \mu(\bigcup_{i \in E} A_i)$. Then $\nu$ is a $\{0,1\}$-valued measure on $I$ with $\nu(\{i\}) = 0$ for all $i \in I$ and $\nu(I) = 1$. Also, $\nu$ is $\kappa$-additive by minimality of $\kappa$, so $\kappa$ is measurable.
(b) If no $\kappa < \chi$ is measurable, conclude that $\mu$ is $\chi$-additive, so $\chi$ is measurable.
(c) If $\chi$ is the first measurable cardinal, and $\kappa$ is a cardinal, use the argument of (a) to show that every $\{0,1\}$-valued measure (on $(Y, \mathcal{P}(Y))$ for some set $Y$) is $\kappa$-additive if and only if $\kappa < \chi$.

II.9.7.6. [?] Suppose $X$ is a set, and $\nu$ is a purely nonatomic probability measure on $(X, \mathcal{P}(X))$. Show that there is a purely nonatomic $\{0,1\}$-valued measure $\mu$ on $(X, \mathcal{P}(X))$ with $\mu(X) = 1$. Conclude that $\text{card}(X)$ dominates a measurable cardinal. (This requires that $2^{\aleph_0}$ does not dominate an inaccessible cardinal.)
II.9.7.7.  Let \( \kappa \) and \( \lambda \) be cardinals.

(a) Show (without AC) that if \( \kappa \leq \lambda \), then \( \kappa \preceq \lambda \). [Mimic the proof of II.5.4.1.]

(b) Show (without AC) that if \( \kappa \preceq \lambda \), then \( 2^\kappa \leq 2^\lambda \). [If \( \text{card}(X) = \kappa \) and \( \text{card}(Y) = \lambda \), and \( f : Y \to X \), define \( f^*: \mathcal{P}(X) \to \mathcal{P}(Y) \) by \( f^*(A) = f^{-1}(A) \). If \( f \) is surjective, then \( f^* \) is injective.]

(c) Use (b) to show that if \( \kappa \) is any cardinal, then \( 2^\kappa \npreceq \kappa \). This also follows directly from the proof of II.5.3.3.

II.9.7.8.  (a) Show (without AC) that if \( \kappa \) is any infinite cardinal, then \( \aleph_0 \npreceq 2^\kappa \). [If \( \text{card}(X) = \kappa \) and \( n \in \mathbb{N} \), the set of subsets of \( X \) of cardinality \( n \) is nonempty. Define \( f : \mathcal{P}(X) \to \{0\} \cup \mathbb{N} \cup \{\infty\} \) by \( f(A) = \text{card}(A) \) if \( A \subseteq X \) is finite, and \( f(A) = \infty \) otherwise.]

(b) Conclude from II.9.7.7. (b) that \( 2^{\aleph_0} \leq 2^\kappa \). Explicitly describe a subset of \( \mathcal{P}(\mathcal{P}(X)) \) of cardinality \( 2^{\aleph_0} \).

[For \( S \subseteq \mathbb{N} \), consider \( \{A \subseteq X : \text{card}(A) \in S\} \}.]

II.9.7.9.  (a) Show (without AC or CH) that \( \aleph_1 \npreceq 2^{\aleph_0} \). [Define \( f : \mathcal{P}(\mathbb{Q}) \to \aleph_1 = \omega_1 \) as follows. If \( A \subseteq \mathbb{Q} \), set \( f(A) = \sigma \) if \( A \) is well ordered (in the induced order from \( \mathbb{Q} \)) and order-isomorphic to \( \sigma \); set \( f(A) = 0 \) if \( A \) is not well ordered.]

(b) \[?] If \( \aleph_1 \) and \( 2^{\aleph_0} \) are not comparable (i.e. if \( \aleph_1 \npreceq 2^{\aleph_0} \)), then there is a cardinal \( \kappa \) satisfying \( \kappa > 2^{\aleph_0} \) but \( \kappa \npreceq 2^{\aleph_0} \). [Set \( \kappa = \aleph_1 + 2^{\aleph_0} \). Let \( f : [0,1] \to \aleph_1 \) be surjective, and \( g : \mathbb{R} \setminus [0,1] \to \mathcal{P}(\mathbb{N}) \) be a bijection. Then \( f, g \) give a surjective map from \( \mathbb{R} \) to \( \aleph_1 \cup \mathcal{P}(\mathbb{N}) \). Then \( 2^\kappa = 2^{2^{\aleph_0}} \). It cannot be proved in ZF that \( \aleph_1 \leq 2^{\aleph_0} \).]

II.9.7.10.  Set \( \kappa_0 = \aleph_0, \kappa_{n+1} = \kappa_{\kappa_n} \) for each \( n \), and \( \kappa = \lim \{\kappa_n\} \). Show that \( \kappa = \aleph_\sigma \) for some \( \sigma \). Since \( \tau \leq \aleph_\tau \) for all \( \tau \) by (\( \cdot \)), \( \sigma \leq \kappa \). If \( \sigma < \kappa \), then \( \sigma < \kappa_n \) for some \( n \), so \( \aleph_\sigma < \aleph_{\kappa_n} = \kappa_{n+1} \leq \kappa \).

II.9.7.11.  (a) Show (without AC) that if \( \kappa \) is a cardinal with \( \aleph_0 \leq \kappa \), and \( \lambda \) is a cardinal with \( \kappa + \lambda = 2^\kappa \), then \( \lambda = 2^\kappa \). [This is a hard problem; see [?], IX.5 for a solution.]

(b) Show (without AC) that if \( \kappa \) is a cardinal with \( \aleph_0 \leq \kappa \), then \( 2^\kappa + 2^\kappa = 2^\kappa \). [It follows from \( \aleph_0 \leq \kappa \) that \( \kappa = 1 + \kappa \), so \( 2^\kappa = 2^{1+\kappa} = 2 \cdot 2^\kappa \).]

(c) Using (a), show (without AC) that if the GCH2 (\( \cdot \)) is assumed, then every cardinal is dominated by an aleph, hence is an aleph, yielding the Axiom of Choice. [If \( \kappa_0 \) is an infinite cardinal, set \( \kappa = \kappa_0 + \aleph_0 \), and \( \lambda = \kappa + \aleph_\kappa (\cdot) \). Then \( 2^{2^\kappa} + \lambda \leq 2^{2^\kappa} \) by (\( \cdot \)) and (b). Conclude from (a) and the GCH2 that either \( \lambda = 2^{2^\kappa} \), in which case by (a), \( \aleph_\kappa (\cdot) = 2^{2^\kappa} \); or else \( \lambda \leq 2^{2^\kappa} \). If \( \lambda \leq 2^{2^\kappa} \), then \( 2^\kappa + \lambda \leq 2^{2^\kappa} \), so by the same argument either \( \aleph_\kappa (\cdot) = 2^{2^\kappa} \) or \( \lambda \leq 2^{2^\kappa} \); if the latter, apply (a) and the GCH2 again to conclude that \( 2^\kappa = \aleph_\kappa (\cdot) \).]

II.9.7.12.  Show (without AC) that for any infinite cardinal \( \kappa \), we have \( \aleph_\kappa (\cdot) \leq 2^{2^\kappa} \).[Mimic the proof of II.9.1.10.]

II.9.7.13.  [?], VIII.2 Ex. 2 Show (without AC) that if \( \kappa \) is infinite and \( \kappa \leq 2^{\aleph_0} \), then \( \aleph_0 \npreceq \kappa \) (and hence \( 2^{\aleph_0} \leq 2^\kappa \) by (\( \cdot \)), as follows. (This appears to be a curious result: it apparently says that if \( \kappa \) is infinite and “not too large”, then \( \kappa \) is “not too small”.)
(a) Fix an enumeration \((r_1, r_2, \ldots)\) of \(\mathbb{Q}\). If \(X\) is an infinite subset of \(\mathbb{R}\), let \(r_n\) be the first rational number \(r\) such that there are \(a, b \in X\) with \(a \leq r < b\). Then, if \(Y = X \cap \{x \in \mathbb{R} : x \leq r_n\}\) and \(Z = X \setminus Y\), \(Y\) and \(Z\) are disjoint and nonempty, and \(X = Y \cup Z\). At least one of \(Y\), \(Z\) is infinite.

(b) Use the procedure in (a) inductively, beginning with an infinite subset \(X\) of \(\mathbb{R}\): write \(X = X_1 \cup Y_1\) with \(X_1\) nonempty, \(Y_1\) infinite, and \(X_1, Y_1\) disjoint. Similarly decompose \(Y_1 = X_2 \cup Y_2\), \(\ldots\), \(Y_n = X_{n+1} \cup Y_{n+1}\).

(c) Set \(X_{\infty} = \bigcap_n Y_n = X \setminus \bigcup_n X_n\). If \(X_{\infty} \neq \emptyset\), then \(X\) is the disjoint union
\[
X = X_1 \cup X_2 \cup \ldots \cup X_{\infty}
\]
of \(\aleph_0\) nonempty sets, and otherwise \(X = \bigcup_n X_n\) is a disjoint union of \(\aleph_0\) nonempty sets.

(d) Conclude that if \(\kappa\) is a cardinal and \(2^\kappa < 2^{\aleph_0}\), then \(\kappa\) is finite.

It cannot be proved without some form of Choice (Countable AC suffices) that an infinite subset of \(\mathbb{R}\) contains a countably infinite subset \(\emptyset\).

II.9.7.14. \(\Box\) Show (without AC) that if \(\kappa\) and \(\lambda\) are cardinals \(\geq 2\), at least one of which is infinite, and \(\kappa \leq \lambda\), then \(\kappa^\lambda = \lambda^\max\). In particular, if \(\kappa\) is infinite, then \(\kappa^2 \leq 2^\kappa\). (We can have \(\kappa^\lambda = \lambda^\kappa\) even if \(\kappa < \lambda\) and both are infinite. For example, if \(\kappa = \text{lim}\{\aleph_0, 2^{\aleph_0}, 2^{2^{\aleph_0}}, \ldots\}\), then \(2^\kappa = \aleph_0^\aleph_0 = \kappa^{\aleph_0} = \kappa^\kappa\) [?, VIII.3].)

II.9.7.15. (a) Show (without AC) that cardinal addition and multiplication satisfy the associative, commutative, and distributive laws.

(b) Show (without AC) that cardinal exponentiation satisfies the usual rules of exponents: if \(\kappa, \lambda, \mu\) are cardinals, then \(\kappa^{\lambda \mu} = (\kappa^\lambda)^\mu\) and \(\kappa^{\mu \lambda} = (\kappa^\mu)^\lambda\).

(c) Show (without AC) that if \(\kappa_1 \leq \kappa_2\) and \(\lambda_1 \leq \lambda_2\), then \(\kappa_1 + \lambda_1 \leq \kappa_2 + \lambda_2\), \(\kappa_1 \lambda_1 \leq \kappa_2 \lambda_2\), and \(\kappa_1^{\lambda_1} \leq \kappa_2^{\lambda_2}\).

(d) If \(\kappa_1 < \kappa_2\) and \(\lambda_1 \leq \lambda_2\), it is not necessarily true that \(\kappa_1 + \lambda_1 < \kappa_2 + \lambda_2\) and \(\kappa_1 \lambda_1 < \kappa_2 \lambda_2\), even with AC: we have \(\aleph_0 + \aleph_0 = \aleph_1 + \aleph_1 = \aleph_1\) and \(\aleph_0 \aleph_1 = \aleph_1^\aleph_1 = \aleph_1\). If \(\kappa_1 < \kappa_2\) and \(\lambda_1 < \lambda_2\), and the Axiom of Choice is assumed, then \(\kappa_1 + \lambda_1 < \kappa_2 + \lambda_2\) and \(\kappa_1 \lambda_1 < \kappa_2 \lambda_2\). (This cannot be proved without AC; in fact, each of these is equivalent to AC [?, XVI.2]. [Let \(\kappa\) be an infinite cardinal and \(\lambda = \aleph_0 \kappa\); then \(2\lambda = \lambda\). If \(\lambda \leq \aleph^+(\lambda)\), then \(\lambda < \lambda + \aleph^+(\lambda)\) and \(\aleph^+(\lambda) < \lambda + \aleph^+(\lambda)\), so
\[
\lambda + \aleph^+(\lambda) < 2\lambda + 2\aleph^+(\lambda) = \lambda + \aleph^+(\lambda)
\]
which is a contradiction. Thus \(\kappa \leq \lambda \leq \aleph^+(\lambda)\) and \(\kappa\) is an aleph. The argument for the second inequality is similar, using \(\lambda = \aleph_0 \aleph_0\}.])

(e) Even if \(\kappa_1 < \kappa_2\) and \(\lambda_1 < \lambda_2\), and the Axiom of Choice is assumed, it is not necessarily true that \(\kappa_1^{\lambda_1} < \kappa_2^{\lambda_2}\): if \(\kappa\) is as in Exercise (), then \(\kappa^2 = \kappa\), and we have \((2^\kappa)^\kappa = 2^{\kappa^2} = 2^\kappa = \kappa^{\aleph_0}\) (A. Tarski; cf. [?, VIII.3]).]

II.9.7.16. (a) [?, VI.6 Ex. 3] Show (without AC) that if a set \(X\) is a union of a sequence of distinct subsets, then \(X\) is a union of a sequence of disjoint nonempty subsets and hence \(\aleph_0 < \text{card}(X)\). [If \(X = \bigcup X_n\), for each \(x \in X\) define a sequence \((s_n(x))\) of 0’s and 1’s by \(s_n(x) = 1\) if and only if \(x \in X_n\). There are infinitely many resulting sequences since the \(X_n\) are distinct. Divide this set of sequences into \(\aleph_0\) disjoint nonempty subsets using Exercise (.)]

(b) If \(\kappa\) is a cardinal and \(\aleph_0 < \kappa\), then \(\aleph_0 < \kappa\) and hence \(2^{\aleph_0} < 2^\kappa\). In particular, if \(\kappa\) is any infinite cardinal, then \(\aleph_0 < 2^\kappa\) and \(2^{\aleph_0} < 2^{2^\kappa}\) (Exercise (.)]
II.9.7.17. Suppose \( \{X_i : i \in I\} \) is an indexed collection of sets. Let \( X = \bigcup_{i \in I} X_i \), and for each \( i \) let \( Y_i = \{i\} \times X_i \subseteq I \times X \). Then \( \{Y_i : i \in I\} \) is an indexed collection of disjoint sets with \( |Y_i| = |X_i| \) for each \( i \).

II.9.7.18. Fix an ordinal \( \sigma \). Let \( \omega_\sigma \) be \( \aleph_\sigma \) regarded as an ordinal. Let \( Q_\sigma \) be the set of all sequences from \( \omega_\sigma \) which are eventually 0. Give \( Q_\sigma \) the lexicographic ordering.

(a) Show that \( Q_0 \) is order-isomorphic to \( \mathbb{Q} \cap [0, \infty) \).

(b) Show that \( \operatorname{card}(Q_\sigma) = \aleph_\sigma \).

(c) Show that \( Q_\sigma \) has a cofinal subset order-isomorphic to \( \omega_\sigma \), and an explicit order-isomorphism from \( Q_\sigma \) to each half-open segment between consecutive elements of this cofinal subset. [Consider the set of sequences which are 0 after the first term.]

(d) If \( \alpha \) is an ordinal with \( \operatorname{card}(\alpha) \leq \aleph_\sigma \), show that there is a cofinal subset of \( \alpha \) of order type \( \leq \omega_\sigma \), with the half-open segments between consecutive elements of the subset of order type \( < \alpha \). [If \( \operatorname{card}(\alpha) = \aleph_\sigma \), let \( \{\beta_\tau : \tau < \omega_\sigma\} \) be an enumeration of the elements of \( \alpha \). For each \( \rho < \omega_\sigma \), let \( \gamma_\rho = \beta_\tau \) for the first \( \tau \) for which \( \beta_\tau \) is larger than \( \gamma_\pi \) for all \( \pi < \rho \), if there is one.]

(e) Using (c) and (d), define by transfinite induction a specific order-embedding of \( \alpha \) into \( Q_\sigma \), for each ordinal \( \alpha \) with \( \operatorname{card}(\alpha) \leq \aleph_\sigma \).

(f) Show (without AC or CH), by the argument of Exercise () using \( Q_\sigma \) in place of \( Q_\sigma \), that \( 2^{\aleph_\sigma} \simeq \mathfrak{c} \) (A. Tarski []).

*****Does this actually work without AC? In (d), must choose a bijection between \( \alpha \) and \( \omega_\sigma \).

II.9.7.19. Let \( \chi \) be the first measurable cardinal. This problem will show that \( \chi \) is inaccessible.

(a) Use problem II.9.7.5.(c) to show that, if \( \{\kappa_i : i \in I\} \) is a set of cardinals with \( \kappa_i < \chi \) for all \( i \) and \( \operatorname{card}(I) < \chi \), then \( \kappa = \lim \{\kappa_i\} < \chi \). [If \( X \) has cardinality \( \kappa \), write \( X = \bigcup_{i \in I} X_i \), where \( \operatorname{card}(X_i) = \kappa_i \).]

(b) If \( \kappa < \chi \), show that \( 2^\kappa < \chi \) as follows. Let \( X \) be a set of cardinality \( \kappa \), and \( \mu \) a \( \{0, 1\}\)-valued measure on \( (\mathcal{P}(X), \mathcal{P}(\mathcal{P}(x))) \). For each \( x \in X \), let

\[
A_x = \{Y \subseteq X : x \in Y\}, \quad B_x = \{Y \subseteq X : x \notin Y\}.
\]

Then \( A_x \) and \( B_x \) are complementary subsets of \( \mathcal{P}(X) \), so either \( \mu(A_x) = 1 \) or \( \mu(B_x) = 1 \). Set

\[
S = \{x \in X : \mu(A_x) = 1\}
\]

\[
S = \bigcap_{x \in S} A_x \cap \bigcap_{x \notin S} B_x.
\]

(i) Show that \( \mu(S) = 1 \). [Consider the complement, and use the fact from Problem 8(c) that \( \mu \) is \( \kappa \)-additive.]

(ii) Show that \( S = \{S\} \).

(iii) Conclude that \( 2^\kappa < \chi \).

(c) If \( \kappa < \chi \), use (b) to conclude that the successor of \( \kappa \) is also less than \( \chi \). Conclude that \( \chi \) is inaccessible.
II.10. Measurable Cardinals

II.10.1. Introduction

In this section, we discuss two fascinating set-theoretic problems which originally arose in the development of Lebesgue measure. Throughout this section, we will need to assume the Axiom of Choice, which will be used without comment.

II.10.1.1. In Lebesgue’s original work, he asked whether every subset of \( \mathbb{R} \) is Lebesgue measurable, or if Lebesgue measure can at least be extended to a translation-invariant measure defined on all subsets of \( \mathbb{R} \). Vitali’s example of a nonmeasurable set (XIII.3.9.5.) gave a negative answer (if AC is assumed).

Banach weakened the question to just ask: is there a measure defined on all subsets of \( \mathbb{R} \), not translation-invariant, which extends Lebesgue measure? More generally, is there a finite nonzero measure on \([0,1]\), or any set \( X \), defined on all subsets of \( X \), in which every singleton has measure 0? It is obvious that the answer to this question for a set \( X \) depends only on the cardinality of \( X \).

II.10.1.2. Definition. Let \( X \) be a set. A measure \( \mu \) on \((X, \mathcal{P}(X))\) is a full real-valued measure on \( X \) if it is a nonzero finite measure in which \( \mu(\{x\}) = 0 \) for every \( x \in X \). A full \( \{0,1\}\)-valued measure is a full real-valued measure taking only the values 0 and 1.

In set theory, a full real-valued measure on a set \( X \) is usually just called a real-valued measure, but this term is ambiguous in a measure-theory context (the main point, of course, is that a full real-valued measure is defined for all subsets). A full real-valued measure \( \mu \) can be, and usually is, renormalized so that \( \mu(X) = 1 \) (a full real-valued probability measure).

We then can phrase Banach’s question as:

II.10.1.3. Question. Does \([0,1]\) (or \( \mathbb{R} \)) have a full real-valued measure? Does there exist a set \( X \) with a full real-valued measure on \( X \)?

It is obvious from countable additivity that a set with a full real-valued measure must be uncountable.

II.10.1.4. Definition. Let \( \kappa \) be a cardinal. A measure \( \mu \) on a set \((X, \mathcal{P}(X))\) is \( \kappa^+ \)-additive if, whenever \( \{A_i : i \in I\} \) is an indexed collection of pairwise disjoint subsets of \( X \), with \( |I| \leq \kappa \), then

\[
\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i) .
\]

The measure \( \mu \) is \( \kappa \)-additive if it is \( \lambda^+ \)-additive for all \( \lambda < \kappa \).

This terminology is a little confusing and somewhat counterintuitive, but is standard in set theory and the theory of large cardinals. Note that “\( \aleph_2^+ \)-additive” and “\( \aleph_1 \)-additive” both mean “countably additive”; “\( \aleph_0 \)-additive” is the same as “finitely additive.”

II.10.1.5. Proposition. Let \( \kappa \) be a cardinal, \( X \) a set of cardinality \( \kappa \), and \( \mu \) a finite (or \( \sigma \)-finite) measure on \((X, \mathcal{P}(X))\). Then the following are equivalent:

(i) Whenever \( \{A_i : i \in I\} \) is a collection of subsets of \( X \) with \( |I| \leq \kappa \) and \( \mu(A_i) = 0 \) for all \( i \), we have \( \mu(\bigcup_{i \in I} A_i) = 0 \).
(ii) Whenever \( \{ A_i : i \in I \} \) is a collection of pairwise disjoint subsets of \( X \) with \( |I| \leq \kappa \) and \( \mu(A_i) = 0 \) for all \( i \), we have \( \mu(\bigcup_{i \in I} A_i) = 0 \).

(iii) \( \mu \) is \( \kappa^+ \)-additive.

**Proof:** (i) \( \Rightarrow \) (ii) and (iii) \( \Rightarrow \) (ii) are trivial.

(ii) \( \Rightarrow \) (i): Suppose \( \mu \) satisfies (ii), and \( \mu(A_i) = 0 \) for \( i \in I \). Fix a well ordering on \( I \), and for each \( i \) let \( B_i = A_i \setminus \bigcup_{j < i} A_j \) (set \( B_{i_0} = A_{i_0} \) if \( i_0 \) is the first element of \( I \)). Then the \( B_i \) are pairwise disjoint, \( \mu(B_i) = 0 \) since \( B_i \subseteq A_i \), and
\[
\mu(\bigcup_{i \in I} A_i) = \mu(\bigcup_{i \in I} B_i) = \sum_{i \in I} \mu(B_i) = 0 .
\]

(ii) \( \Rightarrow \) (iii): Suppose \( \mu \) satisfies (ii), and that \( \{ A_i : i \in I \} \) is a collection of pairwise disjoint subsets of \( X \) with \( |I| \leq \kappa \). Then \( J = \{ i \in I : \mu(A_i) > 0 \} \) is countable by (i), and we have \( \mu(\bigcup_{i \in I \setminus J} A_i) = 0 \) by (ii), since \( |I \setminus J| \leq \kappa \). Thus
\[
\mu(\bigcup_{i \in I} A_i) = \mu(\bigcup_{i \in J} A_i) + \mu(\bigcup_{i \in I \setminus J} A_i) = \sum_{i \in J} \mu(A_i) + 0 = \sum_{i \in I} \mu(A_i) .
\]

So, if \( \mu \) is a full real-valued measure on \( X \), then \( \mu \) is \( \kappa \)-additive if and only if the union of fewer than \( \kappa \) sets of measure zero has measure zero.

A key observation made by Banach was

**II.10.1.6.** **Proposition.** Let \( \kappa \) be the smallest cardinal for which there exists a set of cardinality \( \kappa \) with a full real-valued measure, and let \( \mu \) be a full real-valued measure on a set \( X \) (not necessarily of cardinality \( \kappa \)). Then \( \mu \) is \( \kappa \)-additive.

**Proof:** Suppose that \( \{ A_i : i \in I \} \) is an indexed collection of pairwise disjoint subsets of \( X \) with \( \mu(A_i) = 0 \) for all \( i \) and \( |I| < \kappa \). Let \( A = \bigcup_{i \in I} A_i \). If \( \mu(A) > 0 \), for \( S \subseteq I \) let \( \nu(S) = \mu(\bigcup_{i \in S} A_i) \). Then it is easy to check that \( \nu \) is a full real-valued measure on \( I \), contradicting minimality of \( \kappa \). Thus \( \mu(A) = 0 \) and \( \mu \) is \( \kappa \)-additive.

**II.10.2. Real-Valued Measurable Cardinals**

In light of **II.10.1.6.**, we make the following definition:

**II.10.2.1. Definition.** A cardinal \( \kappa \) is real-valued measurable if there is a \( \kappa \)-additive full real-valued measure on a set of cardinality \( \kappa \).

Note that if \( \kappa \) is any cardinal larger than a real-valued measurable cardinal, then there is a full real-valued measure on a set of cardinality \( \kappa \) (supported on a subset of real-valued measurable cardinality), but the measure will not be \( \kappa \)-additive in general. But **II.10.1.6.** says that if \( \kappa \) is the smallest cardinal of a set supporting a full real-valued measure, then \( \kappa \) is a real-valued measurable cardinal. Thus there exists a set with a full real-valued measure if and only if there exists a real-valued measurable cardinal.
II.10.2.2. Proposition. Let $\kappa$ be a real-valued measurable cardinal. Then $\kappa$ is an uncountable regular cardinal.

Proof: We have already observed that $\kappa$ is uncountable.

Let $X$ be a set of cardinality $\kappa$ and $\mu$ a $\kappa$-additive full real-valued measure on $X$. Note that if $A$ is a subset of $X$ with $|A| < \kappa$, then $\mu(A) = 0$ by $\kappa$-additivity. So if $\{A_i : i \in I\}$ is an indexed collection of pairwise disjoint subsets of $X$ with $|A_i| < \kappa$ for all $i$ and $|I| < \kappa$, then $\mu(A_i) = 0$ for all $i$ and hence $\mu(\bigcup_{i \in I} A_i) = 0$, so we cannot have that $X = \bigcup_{i \in I} A_i$.

Banach and Kuratowski proved that if the Continuum Hypothesis is assumed, there is no full real-valued measure on $\mathbb{R}$ or $[0, 1]$. Then Banach extended the arguments to show that if the Generalized Continuum Hypothesis (GCH) is assumed, then any real-valued measurable cardinal is inaccessible.

The next major contributions were made by S. Ulam. In his Ph.D. dissertation, he showed that every real-valued measurable cardinal is weakly inaccessible without the use of the GCH. The device he used is now called an Ulam matrix. The construction uses cardinal and ordinal notation and terminology; see (i) for an explanation.

II.10.2.3. Lemma. [Ulam Matrix] Let $\lambda$ be a cardinal, and $\kappa$ the successor cardinal of $\lambda$. Regard $\lambda$ and $\kappa$ as sets of ordinals. Then there are subsets $\{A_{\alpha}^\gamma : \alpha < \kappa, \gamma < \lambda\}$ of $\kappa = \{\sigma : \sigma < \kappa\}$ ($\alpha$, $\beta$, $\gamma$, $\sigma$, $\tau$ denote ordinals) such that

(i) If $\alpha < \beta < \kappa$ and $\gamma < \lambda$, then $A_{\alpha}^\gamma \cap A_{\beta}^\gamma = \emptyset$.

(ii) For any $\alpha < \kappa$, $[\kappa \setminus \bigcup_{\gamma < \lambda} A_{\alpha}^\gamma] \subseteq \alpha$, and in particular $|\kappa \setminus \bigcup_{\gamma < \lambda} A_{\alpha}^\gamma| < \kappa$.

The $A_{\alpha}^\gamma$ can be thought of as an array of subsets with $\kappa$ rows and $\lambda$ columns, with $\alpha$ denoting the row and $\gamma$ the column. The conditions say that the sets in each column are pairwise disjoint, and the union of each row is all but a "small" subset of the whole set $\kappa$.

Proof: If $\sigma < \kappa$, then $|\sigma + 1| \leq \lambda$, so there is a surjective function $f_\sigma : \lambda \to \sigma + 1$. Fix such a function for each $\sigma$. If $\alpha < \kappa$ and $\gamma < \lambda$, let

$$A_{\alpha}^\gamma = \{\sigma < \kappa : f_\sigma(\gamma) = \alpha\}.$$  

Property (i) is obvious. For (ii), fix $\alpha < \kappa$. Suppose $\sigma < \kappa$ and $\sigma \notin \bigcup_{\gamma < \lambda} A_{\alpha}^\gamma$. Then there is no $\gamma$ such that $f_\sigma(\gamma) = \alpha$, i.e. $\alpha$ is not in the range of $f_\sigma$, which is $\sigma + 1 = \{\tau : \tau \leq \sigma\}$. Thus $\sigma < \alpha$.

II.10.2.4. Theorem. Let $\kappa$ be a real-valued measurable cardinal. Then $\kappa$ is weakly inaccessible.

Proof: It has already been shown (II.10.2.2) that $\kappa$ is uncountable and regular; it remains to show that it is a limit cardinal. If $\kappa$ were the successor of a cardinal $\lambda$, let $\{A_{\alpha}^\gamma\}$ be an Ulam matrix as in II.10.2.3. If $\mu$ is a $\kappa$-additive full real-valued measure on $\kappa$, set $S = \{(\alpha, \gamma) : \mu(A_{\alpha}^\gamma) > 0\} \subseteq \kappa \times \lambda$. For each $\alpha$, we have $\mu(\kappa \setminus \bigcup_{\gamma < \lambda} A_{\alpha}^\gamma) = 0$ by II.10.2.3.(ii) and $\kappa$-additivity, and hence $\mu(\bigcup_{\gamma < \lambda} A_{\alpha}^\gamma) > 0$; then by $\kappa$-additivity $\mu(A_{\alpha}^\gamma) > 0$ for some $\gamma$, i.e. for each $\alpha < \kappa$ there is at least one $\gamma$ with $(\alpha, \gamma) \in S$. Thus $|S| \geq \kappa$. On the other hand, for each $\gamma < \lambda$ there are at most countably many $\alpha$ such that $(\alpha, \gamma) \in S$ by II.10.2.3.(i) and (i), since $\mu$ is finite. Hence $|S| \leq \aleph_0 \cdot \lambda = \lambda$, a contradiction. Thus $\kappa$ is not a successor cardinal.
More recently, R. Solovay [Sol71] proved that if $\kappa$ is a real-valued measurable cardinal, then the cardinality of the set of weakly inaccessible cardinals less than $\kappa$ is $\kappa$ (Solovay’s result is actually considerably stronger: a real-valued measurable cardinal is weakly Mahlo).

Ulam’s Dichotomy

Ulam then observed that Question II.10.1.3. naturally divides into two rather distinct questions, both of considerable interest. Recall that a set $A \in \mathcal{A}$ in a measure space $(X, \mathcal{A}, \mu)$ is an atom if $\mu(A) > 0$ and for every $B \in \mathcal{A}$, $B \subseteq A$, either $\mu(B) = 0$ or $\mu(A \setminus B) = 0$, and that $(X, \mathcal{A}, \mu)$ is atomless if there are no atoms.

II.10.2.5. Proposition. Let $\mu$ be a full real-valued measure on a set $X$. If $(X, \mathcal{P}(X), \mu)$ has an atom, then there is a full $\{0, 1\}$-valued measure on $X$.

Proof: Let $A$ be an atom in $(X, \mathcal{P}(X), \mu)$. For $B \subseteq X$, define

$$\nu(B) = \frac{\mu(A \cap B)}{\mu(A)}.$$ 

It is easily checked that $\nu$ is a full $\{0, 1\}$-valued measure on $X$.

So Question II.10.1.3. splits into the next two questions:

II.10.2.6. Question. Does there exist an atomless full real-valued measure on a set $X$?

II.10.2.7. Question. Does there exist a full $\{0, 1\}$-valued measure on a set $X$?

II.10.3. Atomless Full Real-Valued Measures

We will discuss Questions II.10.2.6. and II.10.2.7. separately. We first analyze the properties of an atomless full real-valued measure (if one exists). Ulam proved the following theorem:

II.10.3.1. Theorem. Let $\mu$ be an atomless full real-valued measure on a set $X$. Then $\mu$ is not $(2^{\aleph_0})^+\text{-additive}.$

Before we prove this theorem, let us draw the most important consequence:

II.10.3.2. Corollary. Suppose there is a set with an atomless full real-valued measure. Let $\kappa$ be the smallest real-valued measurable cardinal. Then $\kappa \leq 2^{\aleph_0}$.

Proof: Let $X$ be a set with an atomless full real-valued measure $\mu$ on $X$. Then $\mu$ is not $(2^{\aleph_0})^+$-additive by II.10.3.1. But $\mu$ is $\kappa$-additive by II.10.1.6. Thus $\kappa \leq 2^{\aleph_0}$. ✷
This result, combined with II.10.2.4., puts severe set-theoretic limitations on the existence of atomless full real-valued measures: they can only exist if \(2^{\aleph_0}\) is at least as large as the first weakly inaccessible cardinal (in fact, by SOLOVAY’S result, the first weakly Mahlo cardinal). It is believed that this is consistent with ZFC, but it would be in dramatic contradiction to the Continuum Hypothesis and would imply that \(2^{\aleph_0}\) is very large.

The proof of II.10.3.1. mostly consists of the next three lemmas. These results hold for arbitrary atomless finite measures, not just full ones.

II.10.3.3. **Lemma.** Let \((X, \mathcal{A}, \mu)\) be an atomless measure space, and \(Y \in \mathcal{A}\) with \(0 < \mu(Y) < \infty\). For any \(\epsilon > 0\), there is an \(A \in \mathcal{A}\), \(A \subseteq Y\), with \((1 - \epsilon)\mu(Y) < \mu(A) < \mu(Y)\).

**Proof:** Since \(Y\) is not an atom, it decomposes as a disjoint union of sets \(A_1, B_1\) of strictly positive measure. We may assume that \(\mu(A_1) \geq \mu(B_1)\); then \(\mu(A_1) \geq \mu(Y)/2\). Then \(B_1\) decomposes as a disjoint union of \(A_2\) and \(B_2\), of strictly positive measure, with \(\mu(A_2) \geq \mu(B_1)/2\). Continue inductively; for any \(n\) we have
\[
(1 - 2^{-n})\mu(Y) < \mu(\bigcup_{k=1}^{n} A_k) < \mu(Y)
\]
and we can take \(A = \bigcup_{k=1}^{n} A_k\) for \(n\) large enough that \(2^{-n} < \epsilon\).

II.10.3.4. **Lemma.** Let \((X, \mathcal{A}, \mu)\) be an atomless measure space, and \(Y \in \mathcal{A}\) with \(0 < \mu(Y) < \infty\). If \(0 < t < \mu(Y)\), there is a subset \(A\) of \(Y\), \(A \in \mathcal{A}\), with \(\mu(A) = t\).

**Proof:** For countable ordinals \(\sigma\), define subsets \(A_\sigma \in \mathcal{A}\) by transfinite induction as far as possible by setting 
\(A_0 = Y\); if \(\mu(A_\sigma) > t\), let \(A_{\sigma+1}\) be a subset of \(A_\sigma\) with \(t < \mu(A_{\sigma+1}) < \mu(A_\sigma)\) (using II.10.3.3.); and if \(\sigma\) is a (countable) limit ordinal, set \(A_\sigma = \bigcap_{\tau < \sigma} A_\tau\). If \(\mu(A_\sigma) = t\) for some \(\sigma\), terminate the construction and we are done. In fact, this must happen at some countable (limit) ordinal, since the \(\mu(A_\sigma)\) strictly decrease, and if \(\mu(A_\sigma) > t\) for all \(\sigma < \omega_1\) we would have that \(\sigma \mapsto -\mu(A_\sigma)\) would be an order-embedding of \(\omega_1\) into \(\mathbb{R}\), which is impossible. (Alternatively, if for each \(\sigma\) we set \(B_\sigma = A_\sigma \setminus A_{\sigma+1}\), then the \(B_\sigma\) would be an uncountable family of pairwise disjoint subsets of \(Y\) of strictly positive measure.)

II.10.3.5. **Lemma.** Let \((X, \mathcal{A}, \mu)\) be an atomless measure space, and \(Y \in \mathcal{A}\) with \(0 < \mu(Y) < \infty\). Then there are subsets \(\{A_t : 0 \leq t \leq 1\} \subseteq \mathcal{A}\) of \(Y\) such that

(i) \(A_s \subseteq A_t\) for \(s < t\).
(ii) \(A_t = \bigcap_{s > t} A_s\) for \(0 \leq t < 1\) and \(A_1 = Y\).
(iii) \(\mu(A_t) = t \cdot \mu(Y)\) for \(0 \leq t \leq 1\).

**Proof:** Let \(\{r_0, r_1, r_2, \ldots\}\) be an enumeration of \(\mathbb{Q} \cap [0, 1]\) with \(r_0 = 0\) and \(r_1 = 1\). Set \(B_0 = \emptyset\) and \(B_1 = Y\). Inductively define \(B_r\) for \(n \geq 3\) so that \(B_{r_m} \subseteq B_{r_m}\) if \(r_m < r_n\) and \(\mu(B_{r_m}) = r_m \cdot \mu(Y)\), using II.10.3.4. [at stage \(n\), if \(r_p\) and \(r_q\) are the closest numbers with \(r_p < r_n < r_q\) and \(p, q < n\), apply II.10.3.4. to \(B_{r_q} \setminus B_{r_p}\).]
and \( t = (r_n - r_p)\mu(Y) \) and add \( B_{r_p} \). Set \( A_1 = B_1 = Y \) and for \( 0 \leq t < 1 \) set \( A_t = \cap_{r_n > t} B_{r_n} \). Then (i) is obvious, and if \( r_n < t < r_m \), we have \( B_{r_n} \subseteq A_t \subseteq B_{r_m} \), so (ii) and (iii) follow easily.

We now prove Theorem II.10.3.1.

**Proof:** Let \( \mu \) be an atomless full real-valued measure on a set \( X \). Let \( \{A_t : 0 \leq t \leq 1\} \) be a collection of subsets as in II.10.3.5, with \( Y = X \). For each \( t \), set \( C_t = A_t \setminus \cup_{s < t} A_s \). We have that \( \mu(C_t) = 0 \) for each \( t \). If \( x \in X \), let \( t = \inf(s : x \in A_s) \) (note that this set of \( s \) is nonempty since \( x \in A_1 = X \)). Then \( x \in A_t \) by II.10.3.5.(i)–(ii), but \( x \notin A_s \) if \( s < t \), so \( x \in C_t \). Thus \( \cup_{0 \leq t \leq 1} C_t = X \), so \( \mu(\cup_{0 \leq t \leq 1} C_t) > 0 \) and \( \mu \) is not \((2^{\aleph_0})^+\)-additive.

Almost the same proof shows one direction of the following:

**II.10.3.6. Theorem.** There is an extension of Lebesgue measure to a measure on \((\mathbb{R}, \mathcal{P}(\mathbb{R}))\) if and only if there is a set with an atomless full real-valued measure.

**Proof:** Suppose there is an atomless full real-valued probability measure \( \mu \) on a set \( X \). Define the sets \( A_t \) and \( C_t \) as in the proof of II.10.3.1., and define \( f : X \to [0, 1] \) by \( f(x) = t \) if \( x \in C_t \) (note that the \( C_t \) are pairwise disjoint). Set \( \nu = f_*(\mu) \); then \( \nu \) is a full real-valued measure on \([0, 1]\). To see that \( \nu \) extends Lebesgue measure \( \lambda \) on \([0, 1]\), let \( 0 \leq t_1 < t_2 \leq 1 \); then \( f^{-1}((t_1, t_2)) = A_{t_2} \setminus A_{t_1} \), so

\[
\nu((t_1, t_2)) = \mu(A_{t_2} \setminus A_{t_1}) = t_2 - t_1 = \lambda((t_1, t_2))
\]

and thus by () \( \nu \) agrees with \( \lambda \) on Borel subsets of \([0, 1]\). Since for every Lebesgue measurable subset \( E \) of \([0, 1]\) there exist Borel sets \( A \) and \( B \) with \( A \subseteq E \subseteq B \) and \( \lambda(A) = \lambda(E) = \lambda(B) \) (), \( \nu \) must agree with \( \lambda \) on all Lebesgue measurable subsets of \([0, 1]\). Then \( \nu \) can be defined by translation on \([n, n+1]\) for all \( n \) and hence on all of \( \mathbb{R} \), extending Lebesgue measure.

Conversely, suppose \( \mu \) is an extension of \( \lambda \) to all subsets of \( \mathbb{R} \). Then the restriction of \( \mu \) to \([0, 1]\) is a full real-valued measure on \([0, 1]\). It is not obvious that this measure is atomless; however, if it were not there would be a full \([0, 1]\)-valued measure on \([0, 1]\) by II.10.2.5., which is impossible by II.10.4.3..

**II.10.4. Measurable Cardinals**

Now we turn to Question II.10.2.7.. The proof of the next result is essentially identical to the proof of II.10.1.6., and is left to the reader.

**II.10.4.1. Proposition.** Let \( \kappa \) be the smallest cardinal for which there exists a set of cardinality \( \kappa \) with a full \([0, 1]\)-valued measure, and let \( \mu \) be a full \([0, 1]\)-valued measure on a set \( X \) (not necessarily of cardinality \( \kappa \)). Then \( \mu \) is \( \kappa \)-additive.
II.10.4.2. **Definition.** A cardinal is measurable if, on a set $X$ of cardinality $\kappa$, there is a $\kappa$-additive full $\{0,1\}$-valued measure on $X$.

By II.10.4.1., the condition of $\kappa$-additivity is not crucial for the question of existence of a measurable cardinal; but it does ensure that if there is a second measurable cardinal, it is tremendously larger than the first (if this condition were eliminated from the definition, any cardinal larger than a measurable cardinal would be measurable.) The $\kappa$-additivity is very important in the set-theoretic interpretations of measurability.

A measurable cardinal is, of course, a real-valued measurable cardinal. By Ulam’s dichotomy and II.10.3.1., a real-valued measurable cardinal is either measurable or $\leq 2^{\aleph_0}$. Measurable cardinals are larger than $2^{\aleph_0}$ (in fact “much larger”), because of the following fact:

**II.10.4.3. Theorem.** A measurable cardinal is inaccessible.

**Proof:** Let $\kappa$ be a measurable cardinal. Then $\kappa$ is weakly inaccessible by II.10.2.4., so it remains to show that if $\lambda < \kappa$, then $2^\lambda < \kappa$. Fix $\lambda < \kappa$, and suppose $2^\lambda \geq \kappa$. Then there is a $\kappa$-additive full $\{0,1\}$-valued measure $\mu$ on $\mathcal{P}(X)$ for a set $X$ of cardinality $\lambda$ ($\mu$ is supported on a subset of $\mathcal{P}(X)$ of cardinality $\kappa$). For each $x \in X$, let

$A_x = \{ Y \subseteq X : x \in Y \}$, \quad $B_x = \{ Y \subseteq X : x \notin Y \}$.

Then $A_x$ and $B_x$ are complementary subsets of $\mathcal{P}(X)$, so either $\mu(A_x) = 1$ or $\mu(B_x) = 1$. Set

$S = \{ x \in X : \mu(A_x) = 1 \}$

$S = \bigcap_{x \in S} A_x \cap \bigcap_{x \notin S} B_x$.

The complement of $S$ is

$S^c = \bigcup_{x \notin S} A_x \cup \bigcup_{x \in S} B_x$

which is a union of $\lambda$ sets of measure 0, hence $\mu(S^c) = 0$, $\mu(S) = 1$. But it is easily seen that $S = \{ S \}$, and $\mu(T) = 0$ for every singleton $T$, a contradiction.

It turns out that measurable cardinals, if they exist, are enormous:

**II.10.4.4. Theorem.** [Hanf-Tarski] If $\kappa$ is a measurable cardinal, then the cardinality of the set of inaccessible cardinals less than $\kappa$ is $\kappa$.

In fact, “inaccessible” can be replaced by other types of large cardinal conditions weaker than measurability. See II.10.4.11. and II.10.4.12. for stronger statements about the size of measurable cardinals.

In a set $X$ of measurable cardinal $\kappa$, a $\kappa$-additive full $\{0,1\}$-valued measure gives a way of dividing all subsets of $X$ into “large” ones (of measure 1) and “small” ones (of measure 0). If $A$ is any subset of $X$, exactly one of $A$, $X \setminus A$ is “large”. Any subset of cardinality less than $\kappa$ is “small”. 

204
Ultrafilters and Measurable Cardinals

A good alternative description of the principle of the previous paragraph uses the language of ultrafilters. Recall that a filter on a set $X$ is a nonempty subset $\mathcal{F}$ of $\mathcal{P}(X)$ such that

(i) $\mathcal{F}$ is closed under supersets, i.e. if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $B \in \mathcal{F}$.
(ii) $\mathcal{F}$ is closed under finite intersections.
(iii) $\emptyset \notin \mathcal{F}$.

An ultrafilter on $X$ is a maximal filter on $X$. It is easily seen that a filter $\mathcal{F}$ is an ultrafilter if and only if, for every $A \subseteq X$, either $A \in \mathcal{F}$ or $(X \setminus A) \in \mathcal{F}$. An ultrafilter $\mathcal{F}$ is fixed if $\bigcap_{A \in \mathcal{F}} A$ is nonempty (in this case, we must have $\bigcap_{A \in \mathcal{F}} A = \{x\}$ for some $x \in X$); $\mathcal{F}$ is free if it is not fixed, i.e. if $\bigcap_{A \in \mathcal{F}} A = \emptyset$. $\mathcal{F}$ is free if and only if the complement of any finite subset of $X$ is in $\mathcal{F}$. (Note that AC is needed to show that free ultrafilters exist on any infinite set; in fact, by Zorn’s Lemma, every filter is contained in an ultrafilter, and in particular the filter consisting of complements of finite sets is contained in a free ultrafilter.)

If $\kappa$ is a cardinal, a filter $\mathcal{F}$ on a set $X$ is called $\kappa$-complete if $\mathcal{F}$ is closed under taking the intersection of fewer than $\kappa$ sets. Every fixed ultrafilter is $\kappa$-complete for any $\kappa$.

II.10.4.5. Example. Let $\kappa$ be a measurable cardinal, and $\mu$ a $\kappa$-additive full $\{0,1\}$-valued measure on a set $X$ of cardinality $\kappa$ (e.g. $X = \kappa = \{\sigma : \sigma < \kappa\}$). Let $\mathcal{F} = \{A \subseteq X : \mu(A) = 1\}$. Then $\mathcal{F}$ is a $\kappa$-complete free ultrafilter on $X$.

Conversely, suppose $\mathcal{F}$ is a $\kappa$-complete free ultrafilter on a set $X$ of cardinality $\kappa$. Define $\mu : \mathcal{P}(X) \to \{0,1\}$ by $\mu(A) = 1$ if $A \in \mathcal{F}$ and $\mu(A) = 0$ if $A \notin \mathcal{F}$. It is easily checked that $\mu$ is a $\kappa$-additive full $\{0,1\}$-valued measure on $X$, so $\kappa$ is a measurable cardinal.

So we obtain a rephrasing of the definition of measurable cardinals in the language of ultrafilters:

II.10.4.6. Proposition. Let $\kappa$ be a cardinal. Then $\kappa$ is measurable if and only if $\kappa > \aleph_0$ and there is a $\kappa$-complete free ultrafilter on $\kappa$ (or any set of cardinality $\kappa$).

In fact, there is a natural one-one correspondence between full $\{0,1\}$-valued measures on a set $X$ and free $\aleph_1$-complete ultrafilters on $X$, as in II.10.4.5..

Normal Ultrafilters and Nodal Sets

If $\mathcal{F}$ is an ultrafilter on a set $X$ and $f : X \to X$ is a bijection, then $f(\mathcal{F}) = \{f(A) : A \in \mathcal{F}\}$ is also an ultrafilter on $X$, which is $\kappa$-complete if $\mathcal{F}$ is. Thus if $X$ is a set of measurable cardinality $\kappa$, $\mathcal{F}$ is a $\kappa$-complete free ultrafilter on $X$, $A \in \mathcal{F}$ with $|A| = |X \setminus A| = \kappa$, and $f : X \to X$ is a bijection interchanging $A$ and $X \setminus A$, then $f(\mathcal{F})$ is a $\kappa$-complete free ultrafilter on $X$ with $(X \setminus A) \in f(\mathcal{F})$. Thus the intersection of all $\kappa$-complete free ultrafilters on $X$ is the ($\kappa$-complete) filter of all subsets whose complement has cardinality less than $\kappa$. So there does not appear to be any way to systematically distinguish “size” for sets like $A$ using $\kappa$-complete free ultrafilters.

However, it turns out that if $X = \kappa$, the ordering on $\kappa$ does allow a distinction to be made on the “size” of certain sets $A \subseteq \kappa$ even if $|A| = |\kappa \setminus A| = \kappa$. This is done through the notion of a normal ultrafilter.

A subset of $\kappa$ is cobounded if it contains $\{\sigma : \alpha \leq \sigma < \kappa\}$ for some $\alpha < \kappa$. A filter is cobounded if it contains all cobounded sets; the collection $\mathcal{C}_\kappa$ of all cobounded sets is a filter which is the smallest cobounded filter. A $\kappa$-complete filter on $\kappa$ must be cobounded. Any cobounded filter is free.

205
II.10.4.7. Definition. Let $\kappa$ be a cardinal, and $\mathcal{F}$ a cobounded filter on $\kappa$. Then $\mathcal{F}$ is normal if, whenever $\{S_\alpha : \alpha < \kappa\}$ is an indexed collection of sets in $\mathcal{F}$, we have that $$\Delta_{\alpha < \kappa}S_\alpha = \{\sigma < \kappa : \sigma \in \bigcap_{\alpha < \sigma}S_\alpha\} \in \mathcal{F}.$$ $\Delta_{\alpha < \kappa}S_\alpha$ is called the diagonal intersection of the collection $\{S_\alpha\}$.

By definition, a normal filter is cobounded. It is a little hard to motivate this definition intuitively. It means that for any collection $\{S_\alpha\}$ of sets in $\mathcal{F}$, “most” $\sigma$’s in $\kappa$ belong to $X_\alpha$ for all $\alpha < \sigma$. Normal filters are rather rare. For example, for each $n \in \mathbb{N}_0$ let $S_n = \{m \in \mathbb{N}_0 : m > n + 1\}$. If $\mathcal{F}$ is any cobounded filter on $\mathbb{N}_0$, then this is an indexed set in $\mathcal{F}$; but $\Delta_{n < \mathbb{N}_0}S_n = \emptyset$. Thus there are no normal filters on $\mathbb{N}_0$.

II.10.4.8. Proposition. If there is a normal filter $\mathcal{F}$ on $\kappa$, then $\mathcal{F}$ is $\kappa$-complete.

Proof: Let $\{X_i : i \in I\}$ be an indexed collection of sets in $\mathcal{F}$, with $|I| = \lambda < \kappa$. Let $f : \lambda \to I$ be a bijection; set $S_\alpha = X_{f(\alpha)}$ for $\alpha < \lambda$, and $S_\alpha = \kappa$ for $\alpha \in B := \{\sigma : \lambda \leq \sigma < \kappa\}$. Then $\Delta_{\alpha < \kappa}S_\alpha \in \mathcal{F}$, so $$A := B \cap [\Delta_{\alpha < \kappa}S_\alpha] \in \mathcal{F}$$ since $\mathcal{F}$ is cobounded. But it is easy to see that $A = B \cap \cap_{i \in I}X_i \subseteq \cap_{i \in I}X_i$, so $\cap_{i \in I}X_i \in \mathcal{F}$. \hfill \Box

II.10.4.9. Corollary. If there is a normal ultrafilter on $\kappa$, then $\kappa$ is measurable.

The converse is true also (cf. [?], [Kan03, 5.12]):

II.10.4.10. Theorem. A measurable cardinal has a normal ultrafilter.

Actually, a measurable cardinal may have many normal ultrafilters: it is equiconsistent with the existence of measurable cardinals that there are $2^{2^\kappa}$ normal ultrafilters over each measurable cardinal $\kappa$ ([KP71]; cf. [Kan03, 17.8]) (it is also equiconsistent that the smallest measurable cardinal has only one normal ultrafilter [Kun70]). There are $2^{2^\kappa}$ free ultrafilters on $\kappa$ for any (infinite) cardinal $\kappa$.

Sets which are in the intersection of all normal ultrafilters on a measurable cardinal $\kappa$ are called nodal, and can be considered “large” in a set-theoretic sense. The set of cardinals less than $\kappa$ is intuitively a “small” subset of $\kappa$ (there are intuitively many more ordinals which are not cardinals); however, it turns out that the set of cardinals less than $\kappa$ is a nodal subset of $\kappa$. Much more is true:

II.10.4.11. Theorem. If $\kappa$ is a measurable cardinal, then the set of inaccessible cardinals less than $\kappa$ is a nodal subset of $\kappa$.

This is at first glance surprising, since the inaccessible cardinals are seemingly very sparsely distributed even among the cardinals. The Hanf-Tarski Theorem (II.10.4.4.) is an immediate corollary. There are several much stronger results in [Kan03], replacing “inaccessible” with stronger large cardinal properties. The best known result along this line is the following result of F. Rowbottom (cf. [Kan03, 7.19]):
II.10.4.12. \textbf{Theorem.} Let $\kappa$ be a measurable cardinal. Then the set of Ramsey cardinals less than $\kappa$ is a nodal subset of $\kappa$. Thus the cardinality of the set of Ramsey cardinals less than $\kappa$ is $\kappa$.

This result gives some idea of the enormity of measurable cardinals.

II.10.5. \textbf{Consistency and Equiconsistency Results}

Perhaps remarkably, it appears to be consistent with ZFC to assume the existence of a measurable cardinal. Even more remarkably, even though measurable cardinals seem to be far removed from any aspect of the “real world” (even the “real mathematical world”), the existence of a measurable cardinal has significant implications in mathematical logic about subsets of $\mathbb{N}$! (Specifically, there are subsets of $\mathbb{N}$ which are not “constructible.”)

We list without proof several relative consistency results. All are discussed in detail in [Kan03]. If $A$ and $B$ are theories, then $B$ is of greater consistency strength than $A$ if $A$ can be proved consistent assuming that $B$ is consistent, and $A$ and $B$ are equiconsistent if each is of greater consistency strength than the other. We also use the term strictly greater consistency strength, with the obvious meaning. For example, ZF, ZFC, and ZFC+CH are equiconsistent by the results of Gödel. (Recall the fundamental theorem of model theory: a theory is consistent if and only if it has a model. The proofs that $B$ is of greater consistency strength than $A$ generally proceed by constructing a model for $A$ from a model for $B$.)

II.10.5.1. ZFC is not known to be consistent, and this may not be provable using existing theories. ZFC is presumed to be consistent, though, and all of set theory needs to be drastically altered if it is not. In the rest of this section (and throughout this book), we assume ZFC is consistent.

"God exists since mathematics is consistent, and the Devil exists since we cannot prove it."

\textit{André Weil}\textsuperscript{13}

II.10.5.2. It cannot be proved in ZFC that an inaccessible cardinal exists, i.e.

\[ \text{ZFC+}(\text{There does not exist an inaccessible cardinal}) \]

is equiconsistent with ZFC. This is discussed in (). It cannot even be proved in ZFC that a weakly inaccessible cardinal exists. Moreover, it cannot be proved that existence of an inaccessible cardinal is relatively consistent with ZFC, since existence of such a proof would contradict Gödel’s Incompleteness Theorem (cf. [?]). Thus, unlike the situation with CH,

\[ \text{ZFC+}(\text{There exists an inaccessible cardinal}) \]

is of strictly greater consistency strength than ZFC.

There is no current reason to believe, however, that even the existence of a measurable cardinal is inconsistent with ZFC, or with ZFC+GCH.

II.10.5.3. R. Solovay [Sol70] proved that

\[ \text{ZFC+(There exists an inaccessible cardinal)} \]

is of greater consistency strength than

\[ \text{ZF+DC+(Every subset of } \mathbb{R} \text{ is Lebesgue measurable)} \]

(plus several other properties of } \mathbb{R} \text{). S. Shelah [She84] showed that

\[ \text{ZF+(Countable AC)+(Every subset of } \mathbb{R} \text{ is Lebesgue measurable)} \]

is of greater consistency strength than

\[ \text{ZFC+(There exists an inaccessible cardinal)} \]

and thus all of these are equiconsistent. So, roughly speaking, the necessity of using AC to construct a nonmeasurable subset of } \mathbb{R} \text{ is equivalent to the existence of an inaccessible cardinal.}

II.10.5.4. (Solovay [Sol71]) The following are equiconsistent:

(i) ZFC+(There exists a measurable cardinal).

(ii) ZFC+(\(2^{\aleph_0}\) is real-valued measurable).

II.10.5.5. (J. Silver [Sil71]) If ZFC+(There exists a measurable cardinal) is consistent, so is

\[ \text{ZFC+GCH+(There exists a measurable cardinal)} \]

and thus these two are equiconsistent.

II.10.5.6. The existence of a real-valued measurable cardinal \(\leq 2^{\aleph_0}\) is obviously (from II.10.2.4.) inconsistent with the Continuum Hypothesis. It is also inconsistent with Martin’s Axiom. One version of Martin’s Axiom is a variant of the Baire property:

\text{MARTIN’S AXIOM. If } X \text{ is a compact Hausdorff space with the Countable Chain Condition (every family of pairwise disjoint nonempty open subsets in } X \text{ is countable), then the intersection of fewer than } 2^{\aleph_0} \text{ dense open sets in } X \text{ is dense.}

Martin’s Axiom (MA) can be thought of as a weak version of CH which is more palatable to some set theorists, although it is far from universally accepted. Since CH implies MA, ZFC+MA is equiconsistent with ZFC. See [Fre84] for details about Martin’s Axiom and its consequences.

Incidentally, both real-valued measurability of \(2^{\aleph_0}\) and Martin’s Axiom imply the following curious result: for every cardinal \(\kappa\) with \(\aleph_0 \leq \kappa < 2^{\aleph_0}\), we have \(2^\kappa = 2^{\aleph_0}\).

II.10.5.7. (D. Scott [Sco61]) The existence of a measurable cardinal is inconsistent with Gödel’s constructibility hypothesis \(V = L\) (II.8.7.6.). This result can be interpreted two ways: it can be taken as evidence that measurable cardinals should not exist, or that Gödel’s hypothesis is not reasonable. Most set theorists seem to take the latter view. Thus the existence of a measurable cardinal implies that there is a nonconstructible subset of \(\mathbb{N}\). See [Dev84], [Dev77], and [Dev73] for more on constructibility.
II.11. Categories and Functors

Category theory is a “theory of mathematical theories.” It was originally intended to provide a uniform abstract language for mathematical theories and to formulate and prove properties common to many theories. As often happens in mathematics, abstract category theory not only fulfills this original purpose but also turns out to be applicable to many other settings in mathematics as diverse as group theory, the theory of partially ordered sets, and mathematical physics; there is even a way to base an entire formulation of the foundations of mathematics on category theory rather than set theory ()

Category theory is characterized by a high degree of abstraction, and it is fashionable in some circles to view it pejoratively and call it “abstract nonsense.” However, a knowledge of the bare rudiments (the barest of which we describe here) is very useful in understanding the general nature of mathematical theories and how different parts of mathematics can be viewed as parts of a common whole.

A value-neutral nutshell description of category theory is “the theory of arrow-theoretic results.” The standard reference for category theory is [ML98].

Mathematical Theories

A mathematical theory generally consists of two parts:

1. A collection of objects which are sets with additional structure (algebraic, topological, etc.)
2. Functions between the objects which preserve the structure.

There are many familiar examples, some of which are described in more detail in this book. Here are a few:

- Vector spaces (over a fixed field, e.g. \( \mathbb{R} \)) and linear transformations
- Groups and group homomorphisms
- Metric spaces and continuous functions
- Topological spaces and continuous functions
- Partially ordered sets and order-preserving functions

There is an all-inclusive theory: sets (with no additional structure) and arbitrary functions between them.

In order to really understand a mathematical theory such as group theory, one must understand not only the objects (groups) but also the structure-preserving functions between them (group-homomorphisms.)

Categories

A category \( \mathcal{C} \) consists of:

- A collection \( \text{Ob}(\mathcal{C}) \) of objects.
- For each pair \((A, B)\) of objects, a collection \( \text{Hom}(A, B) \) of morphisms.

For all objects \( A, B, C \), a rule for composing morphisms, i.e. a function from \( \text{Hom}(B, C) \times \text{Hom}(A, B) \) to \( \text{Hom}(A, C) \), written \((g, f) \mapsto g \circ f\). Composition must be associative, i.e. for any objects \( A, B, C, D \), and any \( f \in \text{Hom}(A, B) \), \( g \in \text{Hom}(B, C) \), \( h \in \text{Hom}(C, D) \), we must have \( h \circ (g \circ f) = (h \circ g) \circ f \).
For any object $A$, an identity morphism $i_A$ satisfying $f \circ i_A = f$ for all $B$ and all $f \in \text{Hom}(A, B)$, and $i_A \circ g = g$ for all $B$ and all $g \in \text{Hom}(B, A)$.

These requirements are reasonable in light of the examples to be modeled, and are the minimum amount of structure needed for a satisfactory theory. In practically all the applications we will consider in this book, the objects will be sets and the morphisms certain (allowable) functions between them, with “composition” the usual composition of functions. The requirements are then that a composition of allowable functions in the theory is again an allowable function, and that the identity function from any set in the theory to itself is always an allowable function. Associativity of composition is automatic in this case. Such a category is called a category of sets and functions.

The examples described earlier give standard categories:

**Set**: Sets and functions (i.e. the objects of Set are all sets, and $\text{Hom}(X, Y)$ is the set of all functions from $X$ to $Y$)

**Vect$_R$**: Real vector spaces and linear transformations

**Gp**: Groups and group homomorphisms

**Ab**: Abelian groups and group homomorphisms

**MSp**: Metric spaces and continuous functions

**Top**: Topological spaces and continuous functions

**POSet**: Partially ordered sets and order-preserving functions

There is, however, no requirement that the morphisms of a category be functions. Some interesting categories in which the morphisms are not functions are given in the exercises.

When categories are defined within standard set theory, the collection of objects need not be a set (although it is a class, i.e. the objects themselves are sets); but $\text{Hom}(A, B)$ must be a set for every $A, B$. A category $C$ in which $\text{Ob}(C)$ is a set is called a small category. None of the examples above are small categories; but one could construct small categories from them by limiting the size of the underlying set (by restricting the underlying set to be a subset of some large fixed set.)

**Functors**

**Natural Transformations**

**Adjoint Functors**

**The Category of Small Categories**

Functors may be regarded as “morphisms” between categories. We may thus try to go to one further level of abstraction by regarding the collection of all categories as one supercategory, whose objects are categories and for which $\text{Hom}(C, D)$ is the collection of (covariant) functors from $C$ to $D$.

When working within set theory, this construction almost works, but not quite. We run into a version of Russell’s Paradox if we try to take all categories as objects, and we also violate the condition that $\text{Hom}(C, D)$ be a set for any $C$ and $D$. But these logical difficulties disappear if we take as our objects only small categories. Thus we may define a category $\text{SCat}$ whose objects are all small categories, and for which $\text{Hom}(C, D)$ is the set(!) of all functors from $C$ to $D$. 

210
Higher Categories

One can instead think of the collection of all categories as a “second-order category.” There is more structure than just a category: the objects are collections of objects with morphisms between them, and there are morphisms (functors) between the second-level objects preserving the first-order structure. One can formalize the notion of a second-order category.

Why stop there? The second-order category of all categories has morphisms between functors: natural transformations. One can define third-order categories by abstracting this. The process can be repeated indefinitely, apparently even transfinitely.

Categories as a Foundation of Mathematics

Categories were first defined by Eilenberg and MacLane in 1945 to provide a setting for studying natural transformations and formulating parts of algebraic topology and, later, homological algebra. They began to be more than just a language with Grothendieck’s 1957 paper recasting algebraic geometry in category terms. Lawvere in his 1963 Ph.D. thesis began an attempt to reformulate the foundations of mathematics in terms of categories rather than sets.

Lawvere’s idea was that instead of focusing on membership, which is the basic concept of set theory, one should focus on structure as encoded in the transformations between objects. This represents a “top-down” approach to the foundations of mathematics where it is not relevant what the objects of study (e.g. numbers) “are”, but rather what properties they have, rather than the “bottom-up” approach of set theory, which attempts to build up all the objects of mathematics from fundamental building blocks. This is arguably more in line with the way most mathematicians who are not logicians or set-theorists approach mathematics, and it can even be interpreted as being in line with logic and set theory, e.g. we don’t normally worry about what sets are, but only about how to work with them structurally. A great deal of mainstream mathematics, if not all, can (apparently) be done from this point of view without ever involving set theory.

The notion of a topos (plural often written topos as if it were a Greek word even though it isn’t), invented by Grothendieck, can be used in a foundational way. A topos is a category with extra structure (limits, colimits, products, quotients, etc.) which is similar enough to the category of sets and functions that it can be used in a similar way to define and describe all the standard structures of mathematics. They can even encode alternative logics such as intuitionist logic. (Grothendieck viewed topos as generalizations of topological spaces in his original work, which is another valid point of view; the fact that topos so nicely generalize set theory and topology makes them even more appealing as foundational structures.)

There has been much work on recasting the foundations of mathematics on category theory rather than set theory. However, this work is incomplete: there are still substantial gaps in what might be called the “foundations of category theory,” including avoiding potential paradoxes similar to the paradoxes of naive set theory which caused so much consternation in the early twentieth century. The potential, though, is to give a better foundation to the full spectrum of mathematics as it is actually practiced.

II.11.1. Exercises

1. Let $R$ be a transitive, reflexive relation on a set $X$. Define a category $C$ with $\text{Ob}(C) = X$, and such that $\text{Hom}(x, y)$ has exactly one element if $xRy$, and $\text{Hom}(x, y) = \emptyset$ if $x$ is not related to $y$. Composition is defined in the only possible way: if $xRy$ and $yRz$, and $f$ and $g$ are the unique elements of $\text{Hom}(x, y)$ and $\text{Hom}(y, z)$, then $g \circ f$ is the unique element of $\text{Hom}(x, z)$.

   (a) Show that $C$ is a (small) category.
(b) Let $S$ be another transitive, reflexive relation on a set $Y$, and $D$ the corresponding category. Describe all functors from $C$ to $D$.

2. Let $S$ be the category associated with the partially ordered set $\{0,1\}$ with the ordinary ordering $()$. If $C$ is any category, a functor from $S$ to $C$ is called a commutative square in $C$. Describe commutative squares in words, and draw a picture explaining the terminology (begin with a picture of the category $S$ by placing the objects at vertices of a square and drawing arrows for the morphisms.)

3. (a) Let $G$ be a group. Define a (small) category $G$ with one object $\bullet$, with $\text{Hom}(\bullet, \bullet) = G$. Define composition of morphisms via the group operation.
   (b) Let $H$ be another group, and $H$ the corresponding category. Describe all functors from $G$ to $H$.
   (c) A (necessarily small) category with one object is called a monoid. Describe which monoids arise from groups.

4. Define a category $\text{Mat}_\mathbb{R}$ whose objects are $\mathbb{N}$, and for which the morphisms $\text{Hom}(n, m)$ are the $m \times n$ matrices with real entries. Composition is matrix product. Show that $\text{Mat}_\mathbb{R}$ is a (small) category.

5. Here is a type of algebraic object which is important in algebraic topology and homological algebra. A chain complex is a diagram of the form

$$\cdots \xrightarrow{d_{n+2}} C_{n+1} \xrightarrow{d_{n+1}} C_n \xrightarrow{d_n} C_{n-1} \xrightarrow{d_{n-1}} \cdots$$

where the $C_n$ are abelian groups and $d_n : C_n \to C_{n-1}$ is a group homomorphism, which satisfies $d_n \circ d_{n+1} = 0$ for all $n$, i.e. the image $B_n$ of $d_{n+1}$ is contained in the kernel $Z_n$ of $d_n$ for each $n$. Often, but not always, $C_n = \{0\}$ for $n < 0$. The chain complex can formally be defined as a pair $C = \{(C_n), (d_n)\}$ consisting of the sequence $(C_n)$ together with the sequence $(d_n)$ of maps.

The elements of $C_n$ are called $n$-chains; the elements of $Z_n$ are called $n$-cycles, and the elements of $B_n$ $n$-boundaries (these terms are motivated by standard constructions of chain complexes in topology). The quotient group $Z_n/B_n$ is called the $n$’th homology group of $C$, denoted $H_n(C)$.

(a) Let $C = \{(C_n), (d_n)\}$ and $D = \{(D_n), (d'_n)\}$ be chain complexes. A morphism from $C$ to $D$ is a sequence $f = (f_n)$, where $f_n : C_n \to D_n$ is a group homomorphism making the squares in the following diagram commutative:

$$\begin{array}{cccccc}
\cdots & C_{n+1} & \xrightarrow{d_{n+1}} & C_n & \xrightarrow{d_n} & C_{n-1} & \xrightarrow{d_{n-1}} & \cdots \\
& \downarrow{f_{n+1}} & \quad & \downarrow{f_n} & \quad & \downarrow{f_{n-1}} & \\
\cdots & D_{n+1} & \xrightarrow{d'_{n+1}} & D_n & \xrightarrow{d'_n} & D_{n-1} & \xrightarrow{d'_{n-1}} & \cdots
\end{array}$$

i.e. $d'_n \circ f_n = f_{n-1} \circ d_n$ for each $n$. Morphisms are composed in the obvious way by setting $(g \circ f)_n = g_n \circ f_n$ (check that this makes sense). Show that the chain complexes with these morphisms form a category $\text{Chc}$.

(b) Show that, for fixed $n$, $H_n$ is a functor from $\text{Chc}$ to $\text{Ab}$. [Here $H_n$ sends the chain complex $C$ to the group $H_n(C)$. If $f \in \text{Hom}(C, D)$, show that there is a natural group homomorphism $H_n(f) : H_n(C) \to H_n(D)$ with the right properties.]

There is also a theory of cochain complexes, where the horizontal arrows are reversed and the subscripts written as superscripts. In an abstract algebraic sense, there is only a notational difference between chain complexes and cochain complexes, but they are used in a somewhat different and complementary way in
topology. Chain and cochain complexes can be defined also over an arbitrary ring $R$ (as diagrams of left $R$-modules). The algebra involved in the study of chain complexes and related matters is quite extensive. See books on homological algebra such as [2], or on algebraic topology such as [3], for a full treatment.

6. Here is a “paradox” concerning Gödel’s Incompleteness Theorem along the lines of the Berry Paradox (8). We work with the following somewhat informal version of Gödel’s Theorem:

**Theorem.** If $T$ is a consistent theory which is finitely describable and contains Peano arithmetic, then there is a statement in $T$ which is neither provable nor disprovable in $T$.

Fix a model for Peano arithmetic, e.g. the standard model ([4]), and let $T$ be the theory whose language is the language of Peano arithmetic and whose axioms are all the statements in Peano arithmetic which are true in the given model. Then $T$ is complete since every statement in $T$ or its negation is an axiom of $T$; $T$ is consistent since it has a model; $T$ obviously contains Peano arithmetic; and $T$ is finitely describable since we just gave a finite description!

Give a resolution of this “paradox.” (Note that the standard model for Peano arithmetic is finitely describable in ZF set theory, which is itself a finitely describable theory in the technical sense used by Gödel, so the demolition of the paradox must lie elsewhere.)
Chapter III

The Numbers of Analysis

The most basic object of study in analysis is the set of real numbers (along with functions on this set). The various number sets culminating in the real numbers (and, one step farther, the complex numbers) are generally familiar, at least informally. In this chapter, we will be fairly careful in describing these number systems through sets of axioms which characterize them and which can be used to develop all their properties and prove the theorems of analysis, while at the same time trying to avoid getting bogged down in set-theoretic technicalities. In the first section we describe the real numbers axiomatically, and then take a constructive approach in later sections beginning with the natural numbers and successively constructing the integers, rational numbers, real numbers, and complex numbers. The axiomatic approach of the first section may be sufficient for many readers, at least temporarily.

III.1. The Real Numbers – Axiomatic Approach

“Mathematics is fundamentally an egalitarian subject. It should in principle be possible for anyone to check the correctness of any argument; nothing needs to be taken on trust. This is the philosophy. The reality is that some parts of mathematics are much harder to grasp than others.”

J. K. Truss

There are two common ways the set \( \mathbb{R} \) of real numbers is specified as a basis for analysis. In the more complete approach, the real numbers are constructed from “scratch” beginning with the natural numbers. This approach is taken in succeeding sections. While this approach is the “right” one from a logical standpoint, it is complicated and time-consuming and is probably not the best way for a beginning analysis student to proceed.

The other approach is to simply begin with a list of reasonable assumptions about \( \mathbb{R} \), called axioms, which reflect the properties of the real numbers one expects from the informal approach seen in calculus and precalculus, and which are comprehensive enough that all the results of analysis can be developed from them. It is pretty easy for a beginning student to accept on faith that there is a number system satisfying the axioms, and it can be shown without too much difficulty that there is only one such number system.

\[1\] [Tru97, p. v]
(in a sense which will be precisely specified). A student who has any doubts about such an acceptance can
look at later sections to whatever extent necessary to convince him/her/self that the number system can
be carefully constructed. While in principle no part of mathematics should be simply accepted on faith,
in practice it may be prudent to postpone working through the full justification of some complicated parts
of mathematics which appear reasonable. (The history of mathematics is full of such temporary leaps of
faith, which have sometimes lasted for centuries! There is, of course, a danger in doing this, especially at
the frontiers of knowledge: one can never be entirely sure the details will work out as expected until they
are actually done.)

In this section, we take an axiomatic approach to the real numbers. We will not attempt to justify the
reasonableness of most of the axioms, which we hope will be self-evident in light of previous experience, nor
will we try to motivate the specific choice of axioms from among all possible statements we would expect
to be true about \( \mathbb{R} \) – after the fact it will become clear that we can get anything we might want as a
logical consequence of the axioms we have chosen. The one axiom we will discuss and motivate is the crucial
Completeness Axiom, which is the most complicated and least obvious of the axioms (and, for analysis,
arguably the most important).

Throughout this section, the real numbers \( \mathbb{R} \) will be assumed to be a set with various algebraic operations
and an ordering, satisfying properties specified in the axioms.

### III.1.1. The Algebraic Axioms of \( \mathbb{R} \)

The axioms for the real numbers come in three groups: the algebraic axioms, the order axioms, and one
additional axiom called the Completeness Axiom. In this subsection, we list the algebraic axioms and derive
the usual rules of algebra from them.

#### III.1.1.1. There are four standard algebraic operations on \( \mathbb{R} \): addition, subtraction, multiplication,
and division (other algebraic operations such as exponentiation can be defined in terms of these). In fact,
subtraction and division can be defined in terms of addition and multiplication respectively, so addition and
multiplication are the operations which need to be specified in the axioms.

#### III.1.1.2. The Algebraic Axioms

(Op) There are two binary operations + (addition) and \( \cdot \) (multiplication) on \( \mathbb{R} \) associating to each
pair \((x, y)\) of elements of \( \mathbb{R} \) elements \( x + y \) and \( x \cdot y \) of \( \mathbb{R} \) (\( x \cdot y \) is usually written \( xy \)).

(A1) The binary operation + is associative: \( x + (y + z) = (x + y) + z \) for all \( x, y, z \in \mathbb{R} \).

(A2) The binary operation + is commutative: \( x + y = y + x \) for all \( x, y \in \mathbb{R} \).

(A3) There is an element \( 0 \in \mathbb{R} \) such that \( 0 + x = x \) for all \( x \in \mathbb{R} \).

(A4) For every \( x \in \mathbb{R} \) there is an element \( -x \in \mathbb{R} \) with \( x + (-x) = 0 \).

(M1) The binary operation \( \cdot \) is associative: \( x(yz) = (xy)z \) for all \( x, y, z \in \mathbb{R} \).

(M2) The binary operation \( \cdot \) is commutative: \( xy = yx \) for all \( x, y \in \mathbb{R} \).

(M3) There is an element \( 1 \in \mathbb{R} \), \( 1 \neq 0 \), such that \( 1 \cdot x = x \) for all \( x \in \mathbb{R} \).

(M4) For every \( x \in \mathbb{R} \), \( x \neq 0 \), there is an element \( x^{-1} \in \mathbb{R} \) with \( xx^{-1} = 1 \).
(D) The operation \( \cdot \) distributes over \(+\): \( x(y + z) = xy + xz = (xy) + (xz) \) for all \( x, y, z \in \mathbb{R} \).

Axiom (D) is called the distributive law or distributive property.

All of the usual rules of arithmetic and algebra can be proved from the algebraic axioms. The next proposition gives some of the most important and often-used ones. We write out parts of the proof in excruciating detail to illustrate the techniques.

**III.1.1.3. Proposition.**

(i) If \( x, y, z \in \mathbb{R} \) and \( x + z = y + z \), then \( x = y \). (Additive Cancellation)

(ii) If \( x, y, z \in \mathbb{R} \), \( z \neq 0 \), and \( xz = yz \), then \( x = y \). (Multiplicative Cancellation)

(iii) If \( x \in \mathbb{R} \), then \( 0 \cdot x = 0 \).

(iv) If \( x, y \in \mathbb{R} \), then \( (-x)y = -(xy) \).

(v) If \( x \in \mathbb{R} \), then \( (-1)x = -x \).

(vi) If \( x \in \mathbb{R} \), then \( -(x) = x \).

(vii) If \( x, y \in \mathbb{R} \), then \( (-x)(-y) = xy \). In particular, \( (-1)(-1) = 1 \).

(viii) If \( x, y \in \mathbb{R} \), \( x \neq 0 \), and \( y \neq 0 \), then \( xy \neq 0 \) and \( (xy)^{-1} = (x^{-1})(y^{-1}) \).

**Proof:** (i): If \( x + z = y + z \), then \( (x + z) + (-z) = (y + z) + (-z) \). But

\[
(x + z) + (-z) = x + (z + (-z)) \text{ by (A1)}
\]

\[
= x + 0 \text{ by (A4) (definition of } -z)
\]

\[
= 0 + x \text{ by (A2)}
\]

\[
= x \text{ by (A3)}.
\]

By a similar argument, \( (y + z) + (-z) = y \). Thus \( x = y \).

(ii): The proof is almost identical to the proof of (i), and is left to the reader (Exercise (i)).

(iii): We have

\[
0 \cdot x + x = 0 \cdot x + 1 \cdot x \text{ by (M3)}
\]

\[
= x \cdot 0 + x \cdot 1 \text{ by (M2)}
\]

\[
= x(0 + 1) \text{ by (D)}
\]

\[
= x \cdot 1 \text{ by (A3)}
\]

\[
= 1 \cdot x \text{ by (M2)}
\]

\[
= x \text{ by (M3)}
\]

\[
= 0 + x \text{ by (A3)}.
\]

Thus we conclude that \( 0 \cdot x = 0 \) by (i) (applied to \( (0 \cdot x, 0, x) \) in place of \( (x, y, z) \)).
(iv): We have
\[(\neg x) y + (xy) = y(\neg x) + (yx) \quad \text{by (M2)}\]
\[= y((\neg x) + x) \quad \text{by (D)}\]
\[= y(x + (\neg x)) \quad \text{by (A2)}\]
\[= y \cdot 0 \quad \text{by (A4) (definition of } -x\text{)}\]
\[= 0 \cdot y \quad \text{by (M2)}\]
\[= 0 \quad \text{by (iii)}\]
\[= (xy) + (-xy)) \quad \text{by (A4) (definition of } -(xy)\text{)}\]
\[= (-xy)) + (xy) \quad \text{by (A2)}\]

Therefore, we have \((-x)y = -(xy)\) by (i).

(v): By (iv), \((-1)x = -(1 \cdot x)\). But \(1 \cdot x = x\) by (M3).

(vi): We abbreviate the argument (as we will from now on). We have
\[
[-(\neg x)] + (-x) = 0 = x + (-x)
\]
and so \((-\neg x) = x\) by (i).

(vii): We have
\[
(-x)(-y) = -[x(-y)] = -[-(xy)] = xy \quad \text{by (vi)}.
\]

(viii): We have
\[
(xy)[(x^{-1})(y^{-1})] = [(xy)x^{-1}](y^{-1}) = [x(yx^{-1})](y^{-1}) = [x(x^{-1}y)](y^{-1}) = [(xx^{-1})y](y^{-1}) = yy^{-1} = 1.
\]

If \(xy = 0\), then \((xy)[(x^{-1})(y^{-1})] = 0 \cdot (x^{-1}y^{-1}) = 0\) by (iii); but \(1 \neq 0\) by (M3), a contradiction. Thus \(xy \neq 0\). We then have \((xy)[(x^{-1})(y^{-1})] = 1 = (xy)[(x^{-1})(y^{-1})]\), so \((xy)^{-1} = (x^{-1})(y^{-1})\) by (ii).

III.1.1.4. By repeated application of the associative laws, a sum or product of three or more factors is unambiguously defined, so we may unambiguously write \(x + y + z\), \(xyz\), etc., for sums and products of real numbers. In particular, if \(x \in \mathbb{R}\), the product \(x \cdot x \cdots x\) \((n\text{ factors})\) is unambiguously defined; write \(x^n\) for this product. By repeated application of III.1.1.3.(viii), \((x^n)^{-1} = (x^{-1})^n\) for any nonzero \(x \in \mathbb{R}\) and any \(n\); write \(x^{-n}\) for this real number. (The last two arguments are a little too informal; they can be done carefully by induction (Exercise III.1.13.2.).) The usual rules of exponents hold (III.1.13.2.).

III.1.1.5. All the usual rules of algebra (with the exception of existence of roots; cf. III.1.12.) can be similarly derived from the algebraic axioms of \(\mathbb{R}\). In the future, we will generally use standard operations of algebra without comment. However, the reader should be sure he/she is able to write out the complete tedious proof of any algebraic formula if required, along the lines of the proofs in III.1.1.3.

From now on, we will also use standard algebraic notation: if \(x, y \in \mathbb{R}\), we will write \(x - y\) for \(x + (-y)\) and \(\frac{x}{y}\) for \(xy^{-1}\) (if \(y \neq 0\)). We will also group multiplication before addition in the usual way: \(xy + z\) means \((xy) + z\), etc.
III.1.1.6. A set with two binary operations satisfying (A1)–(A4), (M1)–(M4), and (D) is called a field, and these axioms are often called the field axioms. There are many fields with quite varied properties; see III.4.3.1. and III.4.3.5.(c), or books on abstract algebra for more details. Since the algebraic properties of \( \mathbb{R} \) are proved just using the field axioms, they are valid in any field.

III.1.2. The Order Axioms for \( \mathbb{R} \)

In addition to the algebraic structure, there is an ordering on the real numbers which (among other things) allows \( \mathbb{R} \) to be pictured geometrically as a line.

III.1.2.1. There are two equivalent ways the ordering on \( \mathbb{R} \) can be axiomatized: either directly, or by describing the set of positive numbers. Each is useful in different contexts. We will begin with the direct approach:

\((\text{Ord})\) There is a relation (strict total order) \( < \) on \( \mathbb{R} \), such that

(O1) If \( x, y \in \mathbb{R} \), then exactly one of \( x < y \), \( y < x \), \( x = y \) holds. (Trichotomy)

(O2) If \( x, y, z \in \mathbb{R} \), \( x < y \), and \( y < z \), then \( x < z \). (Transitivity)

(O3) If \( x, y, z \in \mathbb{R} \) and \( x < y \), then \( x + z < y + z \). (Translation Invariance)

(O4) If \( x, y, z \in \mathbb{R} \), \( x < y \), and \( 0 < z \), then \( xz < yz \). (Homogeneity or Scaling Invariance)

III.1.2.2. We will freely use the usual extensions of the order notation. We write \( x \leq y \) to mean \( x < y \) or \( x = y \). Write \( x > y \) to mean \( y < x \), and \( x \geq y \) to mean \( y \leq x \), i.e. \( x > y \) or \( x = y \).

As with the algebraic axioms, there are many standard order results which can be proved from the axioms. Here are some of the basic ones:

III.1.2.3. Proposition.

(i) If \( x, y \in \mathbb{R} \) and \( x < y \), then \( -y < (-x) \).

(ii) If \( x, y \in \mathbb{R} \), \( x < y \), and \( z < 0 \), then \( yz < xz \).

(iii) \( 0 < 1 \).

(iv) If \( x \in \mathbb{R} \) and \( 0 < x \), then \( 0 < x^{-1} \).

(v) If \( x, y \in \mathbb{R} \), \( 0 < x \), and \( x < y \), then \( y^{-1} < x^{-1} \).

(vi) If \( x, y, z, w \in \mathbb{R} \), \( x < y \), and \( z < w \), then \( x + z < y + w \).

(vii) If \( x, y, z, w \in \mathbb{R} \), \( 0 < x < y \), and \( 0 < z < w \), then \( xz < yw \).

Proof: (i): If \( x < y \), by (O3) we have \(-y = x - x - y < y - x - y = -x\).

(ii): If \( x < y \) and \( z < 0 \), by (i) we have \( 0 < (-z) \), so \(-(xz) = x(-z) < y(-z) = -(yz) \), and hence \( yz < xz \).

by (i).
(iii): Since $1 \neq 0$ by (M3), either $0 < 1$ or $1 < 0$. If $1 < 0$, we have $0 < (-1)$ by (i), so $0 < (-1)(-1) = 1$ by (O4), a contradiction.

(iv): We have $x^{-1} \neq 0$. If $x^{-1} < 0$, then we have

$$1 = xx^{-1} < x0 = 0$$

by (O4), contradicting (iii).

(v): We have $0 < x^{-1}$ and $0 < y^{-1}$ by (iv), so $1 = x^{-1}x < x^{-1}y$ by (O4), and hence $y^{-1} = y^{-1}1 < x^{-1}yy^{-1} = x^{-1}$ again by (O4).

(vi): We have $x + z < y + z < y + w$ by two applications of (O3). The proof of (vii) is almost identical. 

III.1.2.4. Transitivity also holds with $\leq$, and with $>$ and $\geq$:

If $x \leq y$ and $y \leq z$, then $x \leq z$.

If $x < y$ and $y \leq z$, then $x < z$.

If $x \leq y$ and $y < z$, then $x < z$.

If $x > y$ and $y > z$, then $x > z$.

The results of III.1.2.3. with $<$ replaced by $\leq$ throughout also hold. The simple proofs, checking case-by-case, are left to the reader.

Note, however, that transitivity only holds for inequalities going the same direction: for example, if $x < y$ and $y > z$, then no relation between $x$ and $z$ can be concluded. Thus the direction of inequalities must always be carefully noted.

III.1.2.5. Instead of using (Ord) as the axiomatization of the order structure on $\mathbb{R}$, one can equivalently use an axiom describing the positive numbers:

(Pos) There is a nonempty subset $P$ of $\mathbb{R}$ such that

(P1) If $x, y \in P$, then $x + y \in P$.

(P2) If $x, y \in P$, then $xy \in P$.

(P3) $0 \notin P$.

(P4) If $x \in \mathbb{R}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$, but not both.

In the presence of the algebraic axioms, it is easy to derive the (Ord) axioms from (Pos), and vice versa. The arguments are good examples of how properties are proved from the axioms.
III.1.2.6. Suppose we assume the axiom (Pos), along with the algebraic axioms. We then define the relation < by saying 

\[ x < y \text{ if (and only if)} (y-x) \in P. \]

We verify the axioms (O1)–(O4) from (Pos):

(O1): If \( x, y \in \mathbb{R} \), then from (P3) and (P4) we conclude that exactly one of \( y - x = 0 \), \( (y-x) \in P \), \(-y = -(y-x) \in P \) holds. Thus exactly one of \( x = y \), \( x < y \), or \( y < x \) holds.

(O2): If \( x < y \) and \( y < z \), i.e. \((y-x) \in P \) and \((z-y) \in P \), then \((z-x) = (y-x) + (z-y) \in P \) by (P1).

(O3): \((y+z) - (x+z) = y-x.\)

(O4): If \( x < y \) and \( 0 < z \), i.e. \((y-x) \in P \) and \( z \in P \), then \( yz - xz = z(y-x) \in P \) by (P2).

III.1.2.7. Now assume (Ord), along with the algebraic axioms. Define \( P = \{ x \in \mathbb{R} : 0 < x \} \). We verify that axioms (P1)–(P4) hold for this \( P \):

(P1): If \( x, y \in P \), i.e. \( 0 < x \) and \( 0 < y \), then we have, using (O3) and the algebraic axioms,

\[ 0 = 0 + 0 < x + 0 < x + y \]

so \( x + y \in P \).

(P2): This is very similar to (P1). If \( x, y \in P \), i.e. \( 0 < x \) and \( 0 < y \), by (O4) and III.1.1.3.(iii) we have

\[ 0 = 0 \cdot y < x \cdot y. \]

(P3): By (O1), we do not have \( 0 < 0 \).

(P4): If \( x \neq 0 \), then by (O1) we have \( 0 < x \) or \( x < 0 \), but not both. If \( 0 < x \), then \( x \in P \). If \( x < 0 \), then \( 0 = x + (-x) < 0 + (-x) = -x \) by (O3) (and (A3)–(A4)), so \((-x) \in P \).

III.1.2.8. Thus we are free to assume either (Pos) or (Ord), along with the algebraic axioms; exactly the same results can be obtained in each case. We will sometimes use (Pos) and sometimes (Ord), depending on which is more convenient at the time.

III.1.2.9. A field additionally satisfying (Ord) (or, equivalently, (Pos)) is called an ordered field. Ordered fields look somewhat more like \( \mathbb{R} \) than general fields, but there are still many ordered fields besides \( \mathbb{R} \); see, for example, III.5.3.3.(d) and III.10.2.21.. The results of III.1.2.3. and III.1.2.4. hold in a general ordered field since they are just proved from the axioms.

The final axiom, the Completeness Axiom, is what distinguishes \( \mathbb{R} \) from all other ordered fields. But before discussing this axiom, we describe some standard sets of numbers and their properties.

III.1.3. Absolute Value

III.1.4. The Natural Numbers

III.1.4.1. The first set of numbers developed, going back to prehistoric times, is the natural numbers or counting numbers, also described as the positive integers. This set is usually called \( \mathbb{N} \). Thus, as it is often described or defined,

\[ \mathbb{N} = \{1, 2, 3, \ldots \}. \]

Unfortunately, this “description” is totally inadequate as a definition of \( \mathbb{N} \): it is entirely unclear what the “\( \ldots \)” means (if you think you know, try giving a precise formulation of the meaning!)

Instead, we make the following careful definition (assuming we already have \( \mathbb{R} \)):
III.1.4.2. \textbf{Definition}. \( \mathbb{N} \) is the smallest subset of \( \mathbb{R} \) containing 1 and which is closed under adding 1, i.e. if \( n \in \mathbb{N} \), then \( (n+1) \in \mathbb{N} \).

To see that there is such a smallest subset of \( \mathbb{R} \), we could alternately define \( \mathbb{N} \) to be the intersection of all subsets \( S \) of \( \mathbb{R} \) with the property that 1 \( \in S \) and, whenever \( n \in S \), \( n+1 \) is also in \( S \). There is at least one such subset \( S \), namely \( \mathbb{R} \) itself.

III.1.4.3. One mathematically minor point needs to be addressed. There is no general agreement among mathematicians about whether \( \mathbb{N} \) includes 0; some mathematicians include it, and others do not. We will not. The question of whether to work with \( \mathbb{N} \) or \( \mathbb{N} \cup \{0\} \) as the basic set is strictly a matter of taste, since it is mathematically trivial to go from either to the other. See III.2.1.2. for further comments.

We now derive some of the basic properties of \( \mathbb{N} \) directly from its definition.

III.1.4.4. \textbf{Proposition}. \( \mathbb{N} \) is closed under addition, i.e. if \( n, k \in \mathbb{N} \), then \( (n + k) \in \mathbb{N} \).

\textbf{Proof}: Fix \( k \in \mathbb{N} \), and let \( S_k = \{ n \in \mathbb{N} : n + k \in \mathbb{N} \} \). Then \( 1 \in S_k \) since \( k + 1 \in \mathbb{N} \). If \( n \in S_k \), then \( n \in \mathbb{N} \) and \( n + k \in \mathbb{N} \), so \( n + 1 \in \mathbb{N} \) and \( (n+1) + k = (n+k) + 1 \in \mathbb{N} \), and hence \( n + 1 \in S_k \). Thus \( S_k = \mathbb{N} \). \( \Diamond \)

III.1.4.5. \textbf{Proposition}. \( \mathbb{N} \) is closed under multiplication, i.e. if \( n, k \in \mathbb{N} \), then \( nk \in \mathbb{N} \).

\textbf{Proof}: Fix \( k \in \mathbb{N} \), and let \( S_k = \{ n \in \mathbb{N} : nk \in \mathbb{N} \} \). Then \( 1 \in S_k \) since \( 1 \cdot k = k \in \mathbb{N} \). If \( n \in S_k \), then \( n \in \mathbb{N} \) and \( nk \in \mathbb{N} \). Then \( n + 1 \in \mathbb{N} \), and \( (n+1)k = nk + n \in \mathbb{N} \) by III.1.4.4., so \( n + 1 \in S_k \). Thus \( S_k = \mathbb{N} \). \( \Diamond \)

III.1.4.6. \textbf{Proposition}. If \( n \in \mathbb{N} \), then either \( n = 1 \) or \( n - 1 \in \mathbb{N} \).

\textbf{Proof}: Let \( S = \{ n \in \mathbb{N} : n = 1 \) or \( n - 1 \in \mathbb{N} \} \). Then \( 1 \in S \), and if \( n \in S \), then \( n + 1 \in \mathbb{N} \) and \( (n+1) - 1 = n \in \mathbb{N} \), so \( n + 1 \in S \). Thus \( S = \mathbb{N} \). \( \Diamond \)

III.1.4.7. \textbf{Proposition}. If \( m, n \in \mathbb{N} \) and \( m < n \), then \( n - m \in \mathbb{N} \).

\textbf{Proof}: Let \( S \) be the set of all \( k \in \mathbb{N} \) such that \( n \in \mathbb{N} \), \( n > k \) implies \( n - k \in \mathbb{N} \). Then \( 1 \in S \) by III.1.4.6.. If \( k \in S \) and \( n > k + 1 \), then \( n > k \), so \( n - k \in \mathbb{N} \), and \( n - k > 1 \) since \( n > k + 1 \); thus \( (n-k)-1 = n - (k+1) \in \mathbb{N} \) by III.1.4.6.. Thus \( k + 1 \in S \). So \( S = \mathbb{N} \). \( \Diamond \)

III.1.4.8. \textbf{Proposition}. If \( n \in \mathbb{N} \), then \( n \geq 1 \).

\textbf{Proof}: Let \( S = \{ n \in \mathbb{N} : n \geq 1 \} \). Then \( 1 \in S \), and if \( n \in S \), then \( n + 1 \in S \); so \( S = \mathbb{N} \). \( \Diamond \)

221
III.1.4.9. Proposition. If \( m, n \in \mathbb{N} \) and \( m \neq n \), then \( |m - n| \geq 1 \).

Proof: We may assume \( m < n \). Then \( n - m \in \mathbb{N} \), so \( n - m \geq 1 \). ☐

III.1.4.10. Definition. Let \( n \in \mathbb{N} \). An \( m \in \mathbb{N} \) is a predecessor of \( n \) if \( m < n \).

III.1.4.11. Proposition. Every \( n \in \mathbb{N} \) has only finitely many predecessors in \( \mathbb{N} \). (To avoid circularity, we interpret this statement to mean that the set of predecessors of any \( n \in \mathbb{N} \) is a finite set in the sense of II.5.2.6.)

Proof: Let \( S \) be the set of all \( n \in \mathbb{N} \) with only finitely many predecessors in \( \mathbb{N} \). Then \( 1 \in S \) since 1 has no predecessors in \( \mathbb{N} \) (III.1.4.8). Suppose \( n \in S \). If \( m \in \mathbb{N} \) is a predecessor of \( n + 1 \), i.e. \( m < n + 1 \), then \( (n + 1) - m \in \mathbb{N} \), so \( (n + 1) - m \geq 1 \). If \( (n + 1) - m = 1 \), then \( m = n \); otherwise \( (n + 1) - m > 1 \) (III.1.4.8.) and \( n > m \), i.e. \( m \) is a predecessor of \( n \). Thus the set of predecessors of \( n + 1 \) is \( \{n\} \cup \{ \text{predecessors of } n \} \), which is a finite set, so \( n + 1 \in S \). Thus \( S = \mathbb{N} \). ☐

III.1.4.12. Proposition. Every nonempty subset of \( \mathbb{N} \) has a smallest element (\( \mathbb{N} \) is well ordered (\( ))).

Proof: Let \( S \) be a nonempty subset of \( \mathbb{N} \), and let \( n \in S \). The set \( A \) of all elements of \( S \) which are predecessors of \( n + 1 \) is finite by III.2.6.10., and nonempty since \( n \in A \). But any finite nonempty subset of \( \mathbb{R} \) has a smallest element (II.5.2.12.). ☐

III.1.5. Induction and Recursion

The properties of \( \mathbb{N} \) we have developed lead to an important method of proof of statements involving natural numbers:

III.1.5.1. [Principle of Induction] Suppose \( P(n) \) is a statement for each \( n \in \mathbb{N} \). Suppose the following two statements hold:

(i) \( P(1) \) is true.

(ii) For every \( n \in \mathbb{N} \), if \( P(n) \) is true, then \( P(n + 1) \) is true.

Then \( P(n) \) is true for all \( n \in \mathbb{N} \).

We could have called this a Theorem, but the statement is not precise enough to qualify. One reason is that we have not been precise about the meaning of the term “statement”, since to be so would require an extensive excursion into set theory and mathematical logic. We have also been imprecise about what it means to say a statement is “true”; this should be interpreted to mean provable, i.e. a Theorem in the theory being considered, or occasionally “true in a given model.” Many books on mathematical logic such as [?] contain careful discussions of these points. But any type of mathematical statement or interpretation of truth likely to be made by a working mathematician comes within the scope of the principle.
III.1.5.2. To see why the Principle of Induction holds, set

\[ S = \{ n \in \mathbb{N} : P(n) \text{ is true} \} . \]

We have that \( 1 \in S \) by (i), and if \( n \in S \) we have \( n + 1 \in S \) by (ii). Thus \( S = \mathbb{N} \).

III.1.5.3. Proofs by induction have been described as knocking over a sequence of dominoes. The “inductive step” (ii) in the procedure amounts to showing that, for each \( n \), if the \( n \)'th domino falls, it knocks over the \( (n + 1) \)'st also. But you still have to push over the first domino by showing (i).

III.1.5.4. Careful proofs of many algebraic formulas can, and should, be done by induction. One example, the subject of a famous (but possibly apocryphal) story, is a formula for the sum of the first \( n \) natural numbers.

According to the story, when C. F. GAUSS, one of the greatest mathematicians in history, was a boy in the late 18th century, his schoolteacher became annoyed with his class and told the children to add the (natural) numbers from 1 to 100. The teacher soon saw GAUSS sitting quietly, not working, and told him to get busy. He replied that he had already finished, and the teacher found that he had the correct answer. When asked how he did it so quickly, GAUSS replied that he had written the numbers in increasing order on one line, in decreasing order on the next line, and added them vertically first pairwise to obtain a sum with all terms the same, which can be computed by multiplication, giving twice the desired sum. Symbolically, he wrote

\[
\begin{align*}
1 & + 2 + \cdots + 99 + 100 = s \\
100 & + 99 + \cdots + 2 + 1 = s \\
101 & + 101 + \cdots + 101 + 101 = 2s
\end{align*}
\]

He then computed the sum in the last line to be \((100)(101) = 10100\), so \( s = 10100/2 = 5050 \).

This clever argument can be easily modified to develop a formula for the sum of the natural numbers from 1 to \( n \), for any \( n \):

\[
\begin{align*}
1 & + 2 + \cdots + (n-1) + n = s \\
(n+1) & + (n+1) + \cdots + (n+1) + (n+1) = 2s
\end{align*}
\]

and the last sum is \( n(n+1) \), yielding the formula

\[ s = \frac{n(n+1)}{2} . \]

However, although this argument is a good way to derive (or, more accurately, guess) the correct formula, it is not precise enough to constitute a rigorous proof, primarily because of the \( \cdots \) in each row. To give a careful proof, we proceed by induction, taking, for each \( n \),

\[ P(n) : \sum_{k=1}^{n} k = \frac{n(n+1)}{2} . \]
To carry out the proof, we must (i) verify $P(1)$, which is easy: both sides of the equation are 1. Then we must show (ii) that $P(n)$ implies $P(n+1)$ for any $n$. So let $n$ be a natural number, and assume that

\[ \sum_{k=1}^{n} k = \frac{n(n+1)}{2}. \]

We then need to show that

\[ \sum_{k=1}^{n+1} k = \frac{(n+1)(n+2)}{2}. \]

But we have

\[ \sum_{k=1}^{n+1} k = \left( \sum_{k=1}^{n} k \right) + (n+1) = \frac{n(n+1)}{2} + (n+1) \]

\[ = \frac{n(n+1) + 2(n+1)}{2} = \frac{(n+2)(n+1)}{2}, \]

which is what was to be shown. By the Principle of Induction, $P(n)$ is true for all $n$, i.e. the formula holds for all $n$.

Proof by induction is used in far more circumstances than just establishing algebraic formulas, however.

**Complete Induction**

A proof procedure which is a variation of ordinary induction is complete induction.

**III.1.5.5.** [Principle of Complete Induction] Suppose $P(n)$ is a statement for each $n \in \mathbb{N}$. Suppose, for each $n \in \mathbb{N}$, the following *inductive implication* holds:

“If $P(m)$ is true for all $m < n$, then $P(n)$ is true.”

Then $P(n)$ is true for all $n \in \mathbb{N}$.

We could have again called this a Theorem, but the statement is also not precise enough to qualify because of the imprecision in the meaning of the terms “statement” and “true.”

**III.1.5.6.** Complete induction is not quite the same as ordinary induction, where statements $P(n)$ indexed by $\mathbb{N}$ are proved for all $n$ by proving $P(1)$ and $[P(n) \Rightarrow P(n+1)]$ for all $n$. Instead, in complete induction, we assume $P(m)$ is true for all $m \in \mathbb{N}$, $m < n$, and deduce $P(n)$. Ordinary induction is a special case of complete induction [if we have that $P(n)$ implies $P(n+1)$, then we certainly have that $P(m)$ for all $m < n+1$ implies $P(n+1)$]; on the other hand, complete induction can be easily deduced from ordinary induction. Indeed, suppose the statements $P(n)$ satisfy the hypotheses of complete induction. For each $n$, set

$Q(n)$: “$P(m)$ is true for all $m < n$.”
Then, for each \( n \geq 2 \), the statement “\( Q(n) \) implies \( P(n) \)” is logically equivalent to “\( Q(n) \) implies \( Q(n+1) \).” We have that \( Q(1) \) is automatically true since there are no \( m < 1 \), so by ordinary induction we obtain that \( Q(n) \) is true for all \( n \), which implies that \( P(n) \) is true for all \( n \).

Complete induction is useful where there is no simple relationship between \( P(n) \) and \( P(n+1) \), but there is a relationship between \( P(n) \) and \( P(m) \) for certain \( m \)’s less than \( n \). The next simple example is a good illustration. (We are slightly ahead of ourselves here, since prime numbers are technically only defined in III.1.6.13., but the definition is well known.)

### III.1.5.7. Proposition

Every natural number can be written as a finite product of prime numbers. [We define the product of an empty set of primes to be 1, and a product of one prime \( p \) to be \( p \).]

**Proof:** For each \( n \in \mathbb{N} \), let \( P(n) \) be the statement “The integer \( n \) can be written as a finite product of prime numbers.” Fix \( n \in \mathbb{N} \), and assume \( P(m) \) is true for all \( m < n \); we must prove that \( P(n) \) is true. There are three cases:

1. \( n = 1 \). Then \( n \) is an empty product of prime numbers.
2. \( n > 1 \) is prime. Then \( n \) is a product of one prime number.
3. \( n > 1 \) is not prime. Then \( n = n_1n_2 \) for some \( n_1, n_2 \in \mathbb{N}, n_1, n_2 > 1 \). Then \( n_1, n_2 < n \), so by assumption \( n_1 \) and \( n_2 \) can each be written as a finite product of prime numbers. Hence so can \( n_1n_2 = n \). ✷

### III.1.5.8. Note that in complete induction, it is not necessary in general to separately prove \( P(1) \), since it is included in the general case. In the case \( n = 1 \), the assumption that \( P(m) \) is true for all \( m < n \) is automatically satisfied since there are no such \( m \)’s (in this case, the hypothesis is said to be *vacuous*.) Of course, it is *allowed* to prove \( P(1) \) as a special case, as was done in the proof of III.1.5.7.

### Infinite Regress

There is an alternate way to phrase a proof by induction or complete induction, based on the fact (III.1.4.12.) that every nonempty subset of \( \mathbb{N} \) has a smallest element.

### III.1.5.9. [Principle of Infinite Regress]

Suppose \( P(n) \) is a statement for each \( n \in \mathbb{N} \). Suppose we have:

“For each \( n \in \mathbb{N} \), if \( P(n) \) is false, then there is an \( m \in \mathbb{N}, m < n \), such that \( P(m) \) is false.”

Then \( P(n) \) is true for all \( n \).

### III.1.5.10. To prove this, let

\[
S = \{ n \in \mathbb{N} : P(n) \text{ is false} \} .
\]

If \( S \) is nonempty, then it has a smallest element \( n \). But if \( n \in S \), the hypothesis insures that there is an \( m \in S \) with \( m < n \), a contradiction. Thus \( S = \emptyset \).
III.1.5.11. Actually, the statement in the Principle of Infinite Regress is just the contrapositive of the statement of Complete Induction, so Infinite Regress is logically equivalent to Complete Induction (and hence to ordinary Induction).

Infinite Regress is sometimes used in a seemingly more perverse way: if for each \( n \) we have a statement \( Q(n) \), and if we can show that if \( Q(n) \) is true, then there is an \( m < n \) such that \( Q(m) \) is true, then it follows that \( Q(n) \) cannot be true for any \( n \)!

To show this, let \( P(n) \) be the negation of \( Q(n) \) for each \( n \), and apply infinite regress to the \( P(n) \).

We often phrase arguments using Infinite Regress in a slightly different way, effectively combining Complete Induction with a proof by contradiction: to prove that \( P(n) \) is true for all \( n \) we assume \( P(n) \) is false for some \( n \), let \( n \) be the smallest natural number for which \( P(n) \) is false, and obtain a contradiction by showing that \( P(m) \) is false for some \( m \in \mathbb{N}, m < n \). “True” and “false” can be interchanged in this statement.

III.1.5.12. As an example of the use of this principle, we redo the proof of III.1.5.7.. Suppose not every element of \( \mathbb{N} \) can be written as a product of prime numbers, and let \( n \) be the smallest natural number which cannot. Then \( n \neq 1 \), since 1 can be written as the empty product of primes, and \( n \) is not prime, since a prime number can be written as a product of length 1. Thus \( n \) can be written as \( n_1 n_2 \), where \( 1 < n_1, n_2 < n \). Then at least one of \( n_1, n_2 \) cannot be written as a product of primes, contradicting that \( n \) is the smallest such number.

This argument is not really simpler when phrased this way, but some arguments do become much simpler or more natural (e.g. III.1.12.6.). In principle, any argument done by induction or complete induction can be rephrased as an infinite regress argument, and conversely.

**Definitions by Induction or Recursion**

Induction (or, technically, recursion) can also be used to define sequences. We will again not be completely precise in defining terms; see, for example, [7, p. 249-250] for a careful treatment.

III.1.5.13. Let \( X \) be a set. Suppose, for each \( n \in \mathbb{N} \), we have a definite procedure or rule \( R_n \) for choosing an \( x_n \in X \); the \( x_n \) may depend on the previous choice of the \( x_m \) for \( m < n \) (but the rule \( R_n \) does not depend on the choice of the \( x_m \) for \( m < n \)). Then

III.1.5.14. **[Principle of Recursion]** There is a unique sequence \((x_n)\) in \( X \) such that, for each \( n \in \mathbb{N} \), \( x_n \) is determined by the specified procedure \( R_n \) applied to \( \{x_m : m < n\} \).

Roughly speaking, if the sequence can be defined “one step at a time” by the rules \( R_n \), then it can be defined “all at once.” The procedure is often used when all the rules \( R_n \) (except \( R_1 \), which just amounts to specification of the first term of the sequence) are the same.

Here is a precise version of the Principle of Recursion:

III.1.5.15. **Theorem.** Let \( X \) be a set, and for each \( n \in \mathbb{N} \) let \( f_n \) be a function from \( X^n \) to \( X \). If \( x_1 \) is any element of \( X \), then there is a unique sequence \((x_n)\) in \( X \) with \( x_{n+1} = f_n(x_1, \ldots, x_n) \) for all \( n \in \mathbb{N} \).

**Proof:** Let \( S \) be the smallest subset of \( \mathbb{N} \times X \) containing \((1, x_1)\), with the property that, for any \( n \in \mathbb{N} \), whenever \( \{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S \) for \((y_1, \ldots, y_n) \in X^n \), then \((n + 1, f_n(y_1, \ldots, y_n)) \in S \). \( (S \) is the
The property that if

As with

Thus

S

"number") is customarily used to denote the set of integers. Thus, very informally,

n

and note that

hence

and note that

The next set of numbers we will consider is the

III.1.6.1. Caution:

In definitions by induction or recursion, the rule

III.1.5.16.

In other words, if a definite rule is specified for each

First we show

Then let

intersection of all subsets of \( N \times X \) with these two properties; there is at least one such subset, \( N \times X \) itself.)

Thus

S

and note that

The letter

Therefore positive fractions predated the concept of 0 or negative integers). The letter

III.1.5.17.

Caution: In definitions by induction or recursion, the rule

Recall that

Note that no form of the Axiom of Choice is needed in this proof; in fact, the AC is not needed even for the transfinite version II.4.2.14..

Thus

Note that

For each \( n \in N \), let \( x_n \) be the unique element of \( X \) for which \( (n, x_n) \in S \). Let \( B = \{ n \in N : x_{n+1} = f_n(x_1, \ldots, x_n) \} \). We have \( 1 \in B \) since \( (2, f_1(x_1)) \in S \); and if \( n \in B \), then \( (n + 2, f_{n+1}(x_1, \ldots, x_{n+1})) \in S \), so \( n + 1 \in B \). Thus \( B = N \).

Note that no form of the Axiom of Choice is needed in this proof; in fact, the AC is not needed even for the transfinite version II.4.2.14..

In other words, if a definite rule is specified for each \( n \) for determining the \( (n+1) \)'st term of the sequence from the first \( n \) terms, there is a unique sequence generated beginning with any first term. The rule \( f_n \) can vary with \( n \), but all the \( f_n \) must be specified in advance. The most common situation is that there is one function \( f : X \to X \) such that \( f_n(x_1, \ldots, x_n) = f(x_n) \) for each \( n \), i.e. the next term in the sequence is determined just from the previous term by the same rule each time.

The sequence is said to be defined by induction (or, more correctly, defined by recursion) from \( x_1 \) by the functions \( (f_n) \).

Caution: In definitions by induction or recursion, the rule \( R_n \) can be quite general, but it must be a definite rule giving a specific element as a function of \( \{ x_m : m < n \} \). Thus, for example, a “rule” like “Let \( x_n \) be any element of \( X \setminus \{ x_m : m < n \} \)” is not permissible, even if it is proved that this set is always nonempty (say, by a cardinality argument.) But if \( X \) is well-ordered (), it is permissible to let \( R_n \) be the rule “Let \( x_n \) be the first element of \( X \setminus \{ x_m : m < n \} \)” (provided this set is always nonempty.)

III.1.6. The Integers

The next set of numbers we will consider is the integers, or whole numbers (although historically positive fractions predated the concept of 0 or negative integers). The letter \( \mathbb{Z} \) (from the German Zahl, “number”) is customarily used to denote the set of integers. Thus, very informally,

\[ Z = \{ 0, \pm 1, \pm 2, \ldots \} \]

As with \( \mathbb{N} \), this is too imprecise to be even an informal definition of \( \mathbb{Z} \).

Instead, we define \( \mathbb{Z} \) to be the smallest subset of \( \mathbb{R} \) containing \( \mathbb{N} \) and closed under subtraction, i.e. with the property that if \( m, n \in \mathbb{Z} \), then \( m - n \in \mathbb{Z} \).
We could describe \( \mathbb{Z} \) more precisely as

\[
\mathbb{Z} = \{ m - n : m, n \in \mathbb{N} \}.
\]

Indeed, this set contains \( \mathbb{N} \) and is closed under subtraction; on the other hand, every difference of natural numbers must be contained in \( \mathbb{Z} \).

There is a conceptual problem with defining \( \mathbb{Z} \) to be the set of differences of natural numbers, however. Although this description is technically correct, there is no uniqueness: an integer can be written as a difference of natural numbers in many different ways. Thus it is conceptually clearer to define \( \mathbb{Z} \) the way we did.

There is another standard way to describe \( \mathbb{Z} \): \( \mathbb{Z} \) is the disjoint union of \( \mathbb{N} \), \( \{ 0 \} \), and \( \{-n : n \in \mathbb{N} \} \). It is clear that the union of these three sets is contained in \( \mathbb{Z} \), and is closed under subtraction by checking case-by-case, so equals \( \mathbb{Z} \). To show the sets are disjoint, just note that if \( n \in \mathbb{N} \), then \( n > 0 \), so \( -n < 0 \) and \( -n \notin \mathbb{N} \). Thus we observe

\[
\mathbb{N} = \{ x \in \mathbb{Z} : x > 0 \}.
\]

Since \( \mathbb{Z} \) is closed under subtraction and taking negatives, it is also closed under addition. In fact, it is the smallest subset of \( \mathbb{R} \) containing \( \mathbb{N} \) and closed under addition and taking negatives. (It is not the smallest subset of \( \mathbb{R} \) containing \( \mathbb{N} \) and closed under taking negatives: that would be \( \mathbb{N} \cup (-\mathbb{N}) \).)

It also follows immediately from III.1.6.3. and III.1.1.3.(iv) and (vii) that \( \mathbb{Z} \) is closed under multiplication. Thus \( \mathbb{Z} \) satisfies all the field axioms (III.1.1.2.) except (M4). (In algebraic language, it is a commutative unital ring). In fact, (M4) fails spectacularly: the only elements of \( \mathbb{Z} \) with multiplicative inverses in \( \mathbb{Z} \) are 1 and \(-1\).

Number Theory

All the algebraic properties of \( \mathbb{N} \) have direct analogs in \( \mathbb{Z} \). In fact, except for complications involved with keeping track of signs, the properties are identical. In particular, the Division Algorithm and the Fundamental Theorem of Arithmetic hold in \( \mathbb{Z} \); the statements are obvious modifications of the ones in \( \mathbb{N} \). (In algebraic language, \( \mathbb{Z} \) is a Euclidean Domain.)

Theorem. [Division Algorithm] Let \( n, d \in \mathbb{Z}, d \neq 0 \). Then there are unique \( q, r \in \mathbb{Z} \) such that \( 0 \leq r < |d| \) and \( n = qd + r \).

See III.2.9.4. for the proof if \( n, d \in \mathbb{N} \). The other cases are nearly identical. We say \( d \) divides \( n \), written \( d|n \), if \( r = 0 \), i.e. if \( n = qd \) for some \( q \in \mathbb{Z} \); \( q \) is called the quotient of \( n \) by \( d \), denoted \( \frac{n}{d} \). If \( d \) divides \( n \), then \( d \) also divides \(-n\), and \(-d \) divides \( \pm n \). Every nonzero integer divides 0. If \( n, d \in \mathbb{N} \), then \( q \in \mathbb{N} \), and \( q, d \leq n \), with equality for one if and only if the other is 1. If \( m \) and \( n \) each divide the other, then \( m = \pm n \). If \( d \) divides both \( m \) and \( n \), say \( m = ad \) and \( n = bd \) for \( a, b \in \mathbb{Z} \), then \( m + n = ad + bd = (a + b)d \), so \( d \) divides \( m + n \). Conversely, if \( d \) divides \( m \) and \( m + n \), then \( d \) divides \( (m + n) - m = n \).

With \( \mathbb{Z} \) at our disposal, there is a more efficient approach to the theory of prime numbers and factorization than can be done staying strictly within \( \mathbb{N} \).
III.1.6.8. **Definition.** If \( m, n \in \mathbb{N} \), the number \( d \in \mathbb{N} \) is the *greatest common divisor* for \( m, n \) if \( d \mid m \), \( d \mid n \), and \( c \mid d \) for any \( c \in \mathbb{N} \) which divides both \( m \) and \( n \). Write \( d = \gcd(m, n) \) or \((q) = (m, n)\) (ideal notation).

This definition can be made for general nonzero elements of \( \mathbb{Z} \), but is cleaner if we stick to \( \mathbb{N} \). It is clear that if the greatest common divisor of \( m \) and \( n \) exists, it is unique and is indeed the largest natural number that divides both \( m \) and \( n \), but it is not so clear that the largest such divisor is a greatest common divisor in the sense of the definition, and hence it is unclear that a greatest common divisor exists. However, this turns out to be true:

III.1.6.9. **Theorem.** If \( m \) and \( n \) be natural numbers. Then \( m \) and \( n \) have a (unique) greatest common divisor, which is of the form \( d = am + bn \) for some (nonunique) \( a, b \in \mathbb{Z} \).

**Proof:** Set \( J = \{mx + ny : x, y \in \mathbb{Z}\} \) (\( J \) is an *ideal* in \( \mathbb{Z} \)). Then \( J \cap \mathbb{N} \neq \emptyset \) since \( m, n \in J \). Let \( d \) be the smallest element of \( J \cap \mathbb{N} \), and write \( d = am + bn \) for \( a, b \in \mathbb{Z} \).

We show that \( d \) divides \( m \) and \( n \). In fact, \( d \) divides any \( z \in J \): suppose \( z = mx + ny \in J \), and write \( z = qd + r \) by the Division Algorithm with \( q \in \mathbb{Z} \) and \( 0 \leq r < d \). Then

\[
mx + ny = q(ax + by) + r
\]

and by minimality of \( d \) we have that \( r = 0 \).

If \( c \) divides both \( m \) and \( n \), then \( c \) clearly divides \( am + bn = d \).

III.1.6.10. There is an effective and numerically efficient algorithm (the *Euclidean algorithm*) for finding \( \gcd(m, n) \) for natural numbers \( m \) and \( n \).

III.1.6.11. If \( m, n \in \mathbb{N} \), we say \( m \) and \( n \) are *relatively prime* if \( \gcd(m, n) = 1 \). (This definition, like the notion of greatest common divisor itself, can be extended to pairs of nonzero integers.)

III.1.6.12. **Proposition.** Let \( d, m, n \in \mathbb{N} \). If \( d \) divides \( mn \) and \( d \) and \( m \) are relatively prime, then \( d \) divides \( n \).

**Proof:** If \( d \) and \( m \) are relatively prime, then there are integers \( a \) and \( b \) with \( ad + bm = 1 \). Then

\[
n = 1 \cdot n = (ad + bm)n = adn + bmn
\]

so \( n \) is divisible by \( d \).
III.1.6.13. A $p \in \mathbb{N}$ is a prime number if $p > 1$ and the only natural numbers which divide $p$ are 1 and $p$, i.e. the only integers dividing $p$ are $\pm 1$ and $\pm p$. (To be consistent with the language of abstract algebra, we should also call $-p$ prime, and we should use the term irreducible instead of prime until III.1.6.14. is proved.)

If $p, n \in \mathbb{N}$ and $p$ is prime, then either $p$ divides $n$ or $p$ and $n$ are relatively prime. In particular, if $p$ and $q$ are distinct prime numbers, then $p$ and $q$ are relatively prime. We also have the following corollary of III.1.6.12. (the last part of the statement is proved by induction from the first part):

III.1.6.14. Corollary. Let $p, m, n \in \mathbb{N}$ with $p$ prime. If $p$ divides $mn$, then $p$ divides $m$ or $p$ divides $n$. If $n_1, \ldots, n_r \in \mathbb{N}$ and $p$ divides $n_1 n_2 \cdots n_r$, then $p$ divides at least one of the $n_k$.

We can then state the Fundamental Theorem of Arithmetic:

III.1.6.15. Theorem. [Fundamental Theorem of Arithmetic] Every natural number can be written as a (finite) product of prime numbers, uniquely up to the order of the factors. (The “empty product” 1 and products of length 1 are included. Repeated factors are, of course, allowed.)

Proof: It has already been shown (III.1.5.7.) that every natural number can be written as a product of prime numbers, so we only need to show uniqueness. This can be done by induction on the length of the product, but we phrase the argument alternately by infinite regress.

Suppose

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$$

is a nonunique factorization with $r$ the smallest possible length for one of the factorizations (hence $r \leq s$). Then $n > 1$, so both $r$ and $s$ are at least 1, and neither can be 1 since the $p_j$ and $q_k$ are prime. Thus $r, s > 1$. We have that $p_1$ divides $q_1 \cdots q_s$, so it divides $q_k$ for some $k$, and since $q_k$ is prime we have $p_1 = q_k$. By reordering the $q$’s we may assume $k = 1$. Thus

$$\frac{n}{p_1} = p_2 \cdots p_r = q_2 \cdots q_s$$

is a nonunique factorization with one product of length $r - 1$, contradicting minimality of $r$. Thus nonunique factorizations do not exist. \(\diamondsuit\)

We conclude our brief introduction to the algebraic theory of $\mathbb{N}$ by noting that there are an infinite number of primes. This observation and its slick proof are due to Euclid. (Actually Euclid never stated that there are infinitely many primes; the ancient Greeks had no notion of infinity.)

III.1.6.16. Theorem. There are infinitely many prime numbers. If $n \in \mathbb{N}$, there is a prime number $p$ with $p > n$.

Proof: Suppose $p_1, \ldots, p_r$ is a finite set of prime numbers. Set $m = 1 + p_1 p_2 \cdots p_r$. Then no $p_k$ divides $m$, so if $p$ is a prime dividing $m$, $p$ is not one of the $p_k$ and no finite list of primes is complete.

The last statement follows from the first and the fact that the set of predecessors of $n$ is finite. As an alternate proof, the number $n! + 1$ (cf. Exercise III.1.13.3.) is not divisible by any natural number $\leq n$ except 1, so any prime number dividing it is greater than $n$. \(\diamondsuit\)
III.1.7. The Rational Numbers

The next step historically in building the number system was the rational numbers, usually denoted \( \mathbb{Q} \) (for “quotient”). Actually at least some simple rational numbers probably predated negative numbers historically.

III.1.7.1. \( \mathbb{Q} \) is the smallest subset of \( \mathbb{R} \) containing \( \mathbb{Z} \) and closed under quotients of nonzero elements.

III.1.7.2. If \( a, b \in \mathbb{R} \) and \( b \neq 0 \), then \( \frac{a}{b} \) is defined to be \( ab^{-1} \). The formulas

\[
\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}; \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}; \quad \left( \frac{b}{d} \right)^{-1} = \frac{d}{b}; \quad \frac{a}{b} + \frac{c}{d} = \frac{ad}{bc}
\]

for \( a, b, c, d \in \mathbb{R}, b \neq 0, d \neq 0 \) (and, in the last, \( c \neq 0 \)), provable from the field axioms, show that

\[
\mathbb{Q} = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \neq 0 \right\}
\]

and that \( \mathbb{Q} \) is the smallest subset of \( \mathbb{R} \) containing \( \mathbb{Z} \) and closed under multiplication and reciprocals of nonzero elements; the first formula also shows that \( \mathbb{Q} \) is closed under addition. Since \( -\frac{m}{n} = \frac{-m}{n} = \frac{m}{-n} \) for any \( m, n \) \((n \neq 0)\), we also have that \( \mathbb{Q} \) is closed under subtraction, and that any element of \( \mathbb{Q} \) can be written as \( \frac{m}{n} \), where \( m \in \mathbb{Z}, n \in \mathbb{N} \).

III.1.7.3. Rational numbers are often described as “fractions,” but this is incorrect. Rational numbers are not the same thing as fractions: many fractions such as \( \frac{1}{2} \) and \( \frac{2}{4} \) can represent the same rational number. Instead, a rational number is a number which can be represented as a fraction. Although each rational number does have a standard representation as a fraction in “lowest terms” (III.1.7.4.), when a rational number appears in mathematics it often does not have any natural representation as a fraction (e.g. it could arise as a repeating decimal).

III.1.7.4. Since for \( n, r \neq 0 \) we have \( \frac{mr}{nr} = \frac{m}{n} \) \( r \) and by definition \( \frac{r}{r} = 1 \), we have \( \frac{mr}{nr} = \frac{m}{n} \). Thus, if \( m, n \in \mathbb{N} \) and \( d = gcd(m, n) \), and \( m = ad, n = bd \), we have \( \frac{m}{n} = \frac{a}{b} \), and \( a \) and \( b \) are relatively prime. By the Fundamental Theorem of Arithmetic, the representation of a positive rational number as a quotient of relatively prime natural numbers is unique. (Actually it just follows immediately from III.1.4.12. that any positive rational number has a unique representation as \( \frac{a}{b} \) with \( b \in \mathbb{N} \) minimal, and this must be in lowest terms, i.e. \( a \) and \( b \) have no common factors. The fact that this is the unique representation as \( \frac{m}{n} \) with \( m 

III.1.7.5. The set \( \mathbb{Q} \) of rational numbers with its algebraic operations, and the ordering induced as a subset of \( \mathbb{R} \), satisfies all the axioms of \( \mathbb{R} \) we have discussed so far, i.e. it is an ordered field. (In fact it is the “smallest” ordered field: any ordered field contains a canonical subfield isomorphic to \( \mathbb{Q} \) as an ordered field.) It is, in fact, not so easy to see (but not terribly difficult; see e.g. III.1.12.2. – III.1.12.6.) that \( \mathbb{Q} \) is not all of \( \mathbb{R} \), i.e. that there exist irrational numbers. The axiom which distinguishes \( \mathbb{R} \) from \( \mathbb{Q} \) (and all other ordered fields) is the Completeness Axiom.
III.1.8. The Number Line

The geometric representation of the real numbers as a line is familiar, at least informally, to anyone who has studied analytic geometry or even high-school algebra. The number line should be viewed merely as a convenient picture of the real numbers and not as a rigorous mathematical construction, since it is logically either baseless or circular depending on whether, or how, the relevant geometric objects are defined.

III.1.8.1. DESCARTES was the one to formalize the notion that points on a line can be identified with [real] numbers, but, at least in retrospect, the basic idea really goes back at least to EUCLID (and probably to predecessors such as EUDOXUS). One of the essential features of Euclidean geometry (the name in Greek refers to “measuring the earth”) is that any two points on a line, or in space, have a numerical distance between them, at least once a fixed unit distance has been specified. EUCLID actually refers to “magnitudes” of lines [line segments] instead of numerical distances between points, but discusses equality, addition, subtraction, multiplication, and division of magnitudes, so his magnitudes correspond naturally to [constructible] nonnegative real numbers; indeed, the principal conceptual advance of DESCARTES’ time over ancient Greece was the abstraction of the notion of “magnitude” or “quantity” into “number”. (However, the conception of a line as a collection of points is a subtle one, and was not widely accepted in ancient Greece, or indeed universally accepted even today.) The only thing missing from EUCLID then, besides the abstract concept of number (and perhaps the conception of a line as a collection of points), is the idea of arbitrarily fixing one point on the line and calling it 0, and a unit distance and positive direction, which can be simultaneously specified by giving one additional point on the line to be called +1; once this is done, identifying all other points on the line with numbers, and vice versa, via distance and direction from 0, is almost immediate. (Identifying points in a plane or three-dimensional space with ordered pairs or triples of numbers is not quite so straightforward: not only must an origin and unit distance be chosen, but also a set of perpendicular axes and an orientation.)

III.1.8.2. To do the identification carefully (or at least systematically), we fix a line \( L \) and arbitrarily fix a point on \( L \) which we call 0. Then fix another point on \( L \) which we call 1. We usually draw \( L \) as a horizontal line, with 1 to the right of 0. Successively mark off equally spaced points to the right of 1 (i.e. the same direction from 0 as 1) and call these 2, 3, \ldots. Do the same to the left (the opposite direction from 1) of 0, and call these numbers \(-1, -2, \ldots\). In this way, the numbers in \( \mathbb{Z} \) are identified with points on \( L \). See Figure ( ). (We often mark points on the number line \( L \) by short cross segments, but these are of course not part of \( L \).)

III.1.8.3. Now, for each \( m \in \mathbb{N} \), choose \( m - 1 \) equally spaced points between 0 and 1, and call them \( \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m} \) from left to right. Similarly, for each \( n \in \mathbb{Z} \) choose \( m - 1 \) equally spaced points between \( n \) and \( n + 1 \), and call them \( n, n + \frac{1}{m}, \ldots, n + \frac{m-1}{m} \) from left to right. Thus each rational number is associated with a point on \( L \), and the ordering on \( \mathbb{Q} \) is the same as the geometric ordering on \( L \) from left to right. We can thus picture \( \mathbb{Q} \) as a subset of \( L \). Although it is not immediately obvious whether every point of \( L \) corresponds to a rational number (we will see that actually “most” do not), the subset corresponding to \( \mathbb{Q} \) is dense in \( L \), i.e. every subinterval of \( L \), no matter how short, contains a point of \( \mathbb{Q} \), in fact infinitely many. See Figure ( ). (Actually, density of \( \mathbb{Q} \) in the number line, which seems geometrically obvious, takes some proof; see ()..)
III.1.9. Suprema, Infima, and the Completeness Axiom

III.1.9.1. One plausible assumption about a line is that it has no “holes”. A way of expressing this is that any time a line $L$ is split into two disjoint subsets $A$ and $B$, with every point in $A$ to the left of (or, in the usual order, less than) every point of $B$, then there should be a point of $L$ exactly at the boundary of $A$ and $B$, i.e. either $A$ should have a largest (rightmost) point or $B$ should have a smallest (leftmost) point. (Both things cannot happen simultaneously, for then the largest point of $A$ and the smallest point of $B$ would be consecutive points on $L$, which clearly is intuitively impossible.) See Figure (1). The term used for this property, called the completeness property, is that a line is connected, or a continuum. It should be emphasized that the completeness property of a line cannot be proved – it is just a reasonable assumption, although a rather subtle one.

To rephrase this property in terms of the real numbers, we need some terminology:

III.1.9.2. Definition. If $A$ is a nonempty subset of $\mathbb{R}$, a number $b \in \mathbb{R}$ is an upper bound for $A$ if $a \leq b$ for all $a \in A$. If $A$ has an upper bound, then $A$ is bounded above. A number $c \in \mathbb{R}$ is a lower bound for $A$ if $c \leq a$ for all $a \in A$. If $A$ has a lower bound, then $A$ is bounded below. $A$ is bounded if $A$ is bounded above and bounded below.

The real number $s$ is the least upper bound, or supremum, of the nonempty subset $A$, written $s = \sup(A)$ (or sometimes $s = \text{lub}(A)$), if $s$ is an upper bound for $A$ and $s \leq b$ for any upper bound $b$ of $A$. The greatest lower bound, or infimum, of $A$, written $\inf(A)$ or $\text{glb}(A)$, is defined analogously.

III.1.9.3. Note that an upper or lower bound for a set $A$ is not necessarily an element of $A$. Not every subset of $\mathbb{R}$ is bounded above or below: for example, $A = \mathbb{R}$ has no upper or lower bound in $\mathbb{R}$. If $A$ is bounded above, there is nothing unique about an upper bound for $A$: if $b$ is an upper bound for $A$, so is $b'$ for any $b' > b$. Similarly, lower bounds are not unique. But the supremum of $A$, if it exists, is unique; similarly, the infimum of $A$ is unique if it exists. But note that it is not obvious that a set which is bounded above has a supremum, or that a set which is bounded below has an infimum.

We do not define upper or lower bounds, or the supremum or infimum, of the empty set $\emptyset$. (Technically, if Definition III.1.9.2. were applied to $A = \emptyset$, every real number would be both an upper bound and a lower bound for $\emptyset$.)

III.1.9.4. Examples. (i) If a subset $A$ of $\mathbb{R}$ has a maximum $b$, then $b = \sup(A)$ and $A$ is bounded above. If $A$ has a minimum $c$, then $c = \inf(A)$ and $A$ is bounded below. In particular, every finite subset of $\mathbb{R}$ has a supremum and an infimum, which are its largest and smallest elements respectively.

(ii) Let $A$ be the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$. Then $A$ is bounded above, and $1 = \sup(A)$ (to see this, note that if $b < 1$, then there is an $x \in \mathbb{R}$ with $b < x < 1$); but $A$ has no maximum element, since $1 \notin A$. Similarly, $0 = \inf(A)$, but $A$ has no minimum.

(iii) It is intuitively clear that $0 = \inf\{\frac{1}{n} : n \in \mathbb{N}\}$, but this requires proof; cf. (1).

(iv) If $A \subseteq \mathbb{R}$, set $-A = \{-x : x \in A\}$. Then $-b$ is a lower bound for $-A$ if and only if $b$ is an upper bound for $A$. So $-A$ is bounded below if and only if $A$ is bounded above, and $-A$ has an infimum if and only if $A$ has a supremum, with $\inf(-A) = -\sup(A)$. Thus the upper and lower situations are completely symmetric under taking negatives.
III.1.9.5. Proposition. Let $A$ be a nonempty subset of $\mathbb{R}$, and $s \in \mathbb{R}$. The following are equivalent:

(i) $s = \sup(A)$.

(ii) $x \leq s$ for all $x \in A$, and for any $b \in \mathbb{R}$, $b < s$, there is an $x \in A$ with $x > b$.

(iii) $x \leq s$ for all $x \in A$, and for every $\epsilon \in \mathbb{R}$, $\epsilon > 0$, there is an $x \in A$ with $x > s - \epsilon$.

Proof: (i) $\Rightarrow$ (ii): If $s = \sup(A)$ and $b < s$, then $b$ is not an upper bound for $A$.

(ii) $\Rightarrow$ (i): Under (ii), $s$ is an upper bound for $A$, and no $b < s$ is an upper bound for $A$.

(ii) $\iff$ (iii): Given $b$ as in (ii), set $\epsilon = s - b$; given $\epsilon$ as in (iii), set $b = s - \epsilon$.

III.1.9.6. Proposition. Let $A$ be a nonempty subset of $\mathbb{R}$, and $c \in \mathbb{R}$. Set $c + A = \{c + x : x \in A\}$. Then $c + A$ is bounded above if and only if $A$ is bounded above, and $\sup(c + A) = c + \sup(A)$. Similarly, $c + A$ is bounded below if and only if $A$ is bounded below, and $\inf(c + A) = c + \inf(A)$.

Proof: If $b$ is an upper bound for $A$, then $c + b$ is an upper bound for $c + A$; conversely, if $d$ is an upper bound for $c + A$, then $d - c$ is an upper bound for $A$. If $s = \sup(A)$, then it is clear that $c + s$ is the least upper bound of $c + A$, and conversely.

We can then phrase the Completeness Axiom as follows. This turns out to be equivalent to the completeness property of the real line as described above, but is easier to work with.

III.1.9.7. Completeness Axiom: Every nonempty subset of $\mathbb{R}$ which is bounded above has a supremum.

We will show that the Completeness Axiom is equivalent to the completeness property of the real line, as described in III.1.9.1. We first make a more general observation:

III.1.9.8. Proposition. Let $A$ and $B$ be nonempty subsets of $\mathbb{R}$, with $a \leq b$ for all $a \in A$, $b \in B$. Then $A$ is bounded above, $B$ is bounded below, and $\sup(A) \leq \inf(B)$.

Proof: Every $b \in B$ is an upper bound for $A$. So $A$ is bounded above, and so has a supremum $\sup(A)$ which satisfies $\sup(A) \leq b$ for each $b \in B$ since $\sup(A)$ is the least upper bound of $A$. This means $\sup(A)$ is a lower bound for $B$ (as is each $a \in A$), so $B$ is bounded below, and $\inf(B)$ exists and satisfies $\inf(B) \geq \sup(A)$ since $\inf(B)$ is the greatest lower bound of $B$.

We can then show using the Completeness Axiom that the real line has the completeness property. This is sometimes (cf. [Lan51]) called Dedekind’s Fundamental Theorem, although in our approach it hardly merits Theorem status since it is almost immediate (in other approaches to the real numbers it is a real theorem).
**III.1.9.9. Theorem.** If $A$ and $B$ are nonempty disjoint subsets of $\mathbb{R}$ with $A \cup B = \mathbb{R}$ and $x < y$ for all $x \in A$ and $y \in B$, then either $A$ has a largest element or $B$ has a smallest element, but not both. Alternatively, there is a unique real number $a$ such that \{ $x \in \mathbb{R} : x < a$\} $\subseteq A$ and \{ $y \in \mathbb{R} : a < y$\} $\subseteq B$ ($a$ may be either the largest element of $A$ or the smallest element of $B$).

**Proof:** If such an $a$ exists, it is both an upper bound for $A$ and a lower bound for $B$, and conversely, a number $a$ which is both an upper bound for $A$ and a lower bound for $B$ satisfies the conclusions. So we need to show that there is a unique $a$ with these properties.

By III.1.9.8., $\sup(A)$ and $\inf(B)$ exist and satisfy $\sup(A) \leq \inf(B)$. We cannot have $\sup(A) < \inf(B)$, since then $c = (\sup(A) + \inf(B))/2$ would satisfy $\sup(A) < c < \inf(B)$ and $c$ could not be in either $A$ or $B$. Thus $\sup(A) = \inf(B)$; call this real number $a$. If $b$ is an upper bound for $A$, then $a \leq b$, and if $d$ is a lower bound for $B$, then $d \leq a$. Thus $a$ is the unique number with both properties.

**III.1.9.10.** Conversely, if the real line is assumed to have the completeness property of III.1.9.1., then the Completeness Axiom can be proved, and thus the Completeness Axiom is equivalent to the completeness property. For suppose $C$ is a nonempty subset of $\mathbb{R}$ which is bounded above. We wish to show from the completeness property that $C$ has a supremum. If $C$ has a largest element, the conclusion is obvious. If $C$ has no largest element, i.e. if no element of $C$ is an upper bound for $C$, let $B$ be the set of all upper bounds for $C$ in $\mathbb{R}$, and let $A$ be the set of real numbers which are not upper bounds for $C$. Then $\mathbb{R}$ is the disjoint union of $A$ and $B$. $B$ is nonempty by assumption, and $A$ is nonempty since $C \subseteq A$. We clearly have $x < y$ for all $x \in A$, $y \in B$. Let $a$ be the unique element of $\mathbb{R}$ given by the completeness property. If $a \in A$, then $a$ would be an upper bound for $A$ and hence for $C$, contradicting the definition of $A$; hence $a \in B$ and $a$ is an upper bound for $C$. If $x < a$, then $x \in A$, so $x$ is not an upper bound for $C$; hence $a$ is the least upper bound of $C$.

**III.1.9.11.** The Completeness Axiom is really the cornerstone of analysis. Most of the basic concepts of analysis, e.g. convergence of sequences, limits and continuity of functions, and derivatives, can be equally well defined in the same way in any ordered field such as $\mathbb{Q}$, and many of the standard results about them (e.g. algebraic combinations of limits, rules of differentiation) hold in this setting with the same proofs as in $\mathbb{R}$. But the objects have very different properties from the real case. For example, working on $\mathbb{Q}$, define

$$f(x) = \begin{cases} -1 & \text{ if } x^2 < 2 \\ 1 & \text{ if } x^2 > 2 \end{cases} .$$

Then $f$ is continuous and even differentiable everywhere, and $f'$ is identically 0, but $f$ is not constant and does not satisfy the Intermediate Value Theorem. If

$$g(x) = \begin{cases} x & \text{ if } x^2 < 2 \\ 0 & \text{ if } x^2 > 2 \end{cases}$$

then $g$ is continuous on $[-2, 2]$, but does not satisfy the Extreme Value Theorem. Thus these properties of continuous or differentiable functions on $\mathbb{R}$, which may seem “obvious,” are in fact rather subtle since they depend on the Completeness Axiom.

These points are argued persuasively in [Kör04]. The book begins by asking, “Why do we bother?” and then states:
It is surprising how many people think that analysis consists in the difficult proofs of obvious theorems. All we need know, they say, is what a limit is, the definition of continuity and the definition of the derivative. All the rest is ‘intuitively clear.’ ”

This notion is then well refuted. There is also an interesting discussion of the question, “Is the Intermediate Value Theorem obvious?” In that book, the working definition of analysis is “mathematics based on the Completeness Axiom” (called the fundamental axiom); everything else (of the mathematics covered there) is regarded as “mere algebra.”

**Intervals in \( \mathbb{R} \)**

Using the Completeness Axiom, we can give a clean description of intervals in \( \mathbb{R} \).

**III.1.9.12. Definition.** A subset \( J \) of \( \mathbb{R} \) is an interval if

(i) \( J \) contains more than one real number.

(ii) If \( x, y \in J, x < y \), and \( z \in \mathbb{R}, x < z < y \), then \( z \in J \).

Informally, an interval is a subset of \( \mathbb{R} \) containing all numbers between any two of its elements. We do not regard \( \emptyset \) or singleton sets to be intervals, although they technically satisfy (ii), but we sometimes call these sets *degenerate intervals*.

**III.1.9.13. Examples.** Here are standard examples of intervals, and the standard interval notation used:

Let \( a, b \in \mathbb{R}, a < b \). Set

\[ [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\} \]

\[ [a, b) = \{x \in \mathbb{R} : a \leq x < b\} \]

\[ (a, b] = \{x \in \mathbb{R} : a < x \leq b\} \]

\[ (a, b) = \{x \in \mathbb{R} : a < x < b\} \]

If \( a \in \mathbb{R}, \) set

\[ [a, \infty) = [a, +\infty) = \{x \in \mathbb{R} : x \geq a\} \]

\[ (a, \infty) = (a, +\infty) = \{x \in \mathbb{R} : x > a\} \]

If \( b \in \mathbb{R}, \) set

\[ (-\infty, b] = \{x \in \mathbb{R} : x \leq b\} \]

\[ (-\infty, b) = \{x \in \mathbb{R} : x < b\} \]

Finally, set \( (-\infty, \infty) = (-\infty, +\infty) = \mathbb{R} \).

Intervals of the form \([a, b]\), \([a, \infty)\), and \((-\infty, b]\) are called *closed intervals*, and intervals of the form \((a, b)\), \((a, \infty)\), and \((-\infty, b)\) are called *open intervals*. The interval \((-\infty, \infty)\) is regarded as both a closed interval and an open interval. An interval of the form \([a, b]\) or \((a, b)\) is called a *half-open interval*.  

236
III.1.9.14. PROPOSITION. Every interval $J$ in $\mathbb{R}$ is of one of the forms from III.1.9.13. More specifically, $J$ is of the form $[a,b]$, $[a,b)$, $(a,b)$, or $(a,b)$ if $J$ is bounded, of the form $[a,\infty)$ or $(a,\infty)$ if $J$ is bounded below but not bounded above, of the form $(-\infty,b]$ or $(-\infty,b)$ if $J$ is bounded above but not bounded below, and $J = (-\infty, \infty)$ if $J$ is not bounded above or below.

PROOF: Suppose $J$ is bounded. Set $a = \inf(J)$ and $b = \sup(J)$. Then $a < b$ since $J$ is not a singleton, and $J \subseteq [a,b]$. If $z \in \mathbb{R}$, $a < z < b$, then $z$ is neither an upper or lower bound for $J$, so there are $x, y \in J$ with $x < z < y$. But then $z \in J$ by (ii). So $(a,b) \subseteq J$. The argument if $J$ is unbounded is almost identical. ⬜

III.1.9.15. Without the Completeness Axiom, we could not assert that every interval is of such a simple form. For example, we can formally define intervals in $\mathbb{Q}$ the same way. But then

$$\{x \in \mathbb{Q} : x^2 < 2\}$$

can be easily shown to be an interval in $\mathbb{Q}$; but even though it is a nonempty proper subset of $\mathbb{Q}$, it has no endpoints (in $\mathbb{Q}$) and so cannot be described as either an open or closed interval in the usual way.

Lengths of Intervals

An important consequence of the characterization of intervals in $\mathbb{R}$, sometimes not properly appreciated, is the fact that each interval has a well-defined numerical length:

III.1.9.16. DEFINITION. Let $I$ be an interval in $\mathbb{R}$. If $I$ is not bounded, set $\ell(I) = +\infty$. If $I$ is bounded, then $I$ is of the form $[a,b]$, $[a,b)$, $(a,b)$, or $(a,b)$ for some $a, b \in \mathbb{R}$, $a < b$, and we set $\ell(I) = b - a$. Then $\ell(I)$, called the length of $I$, satisfies $0 < \ell(I) \leq +\infty$.

III.1.9.17. Without the characterization of intervals, it would not be possible to give a numerical length to intervals in general. For example, if our number system were $\mathbb{Q}$, the interval described in III.1.9.15. would not have a well-defined numerical length within our number system.

The following two properties of interval length seem like (and really are) almost no-brainers, but are surprisingly powerful, eventually making possible a nontrivial theory of Lebesgue measure on $\mathbb{R}$.

III.1.9.18. PROPOSITION. (i) Let $I$ and $J$ be intervals in $\mathbb{R}$. If $J \subseteq I$, then $\ell(J) \leq \ell(I)$.

(ii) Let $I$ and $J$ be intervals in $\mathbb{R}$. If $I \cup J$ is an interval (e.g. if $I \cap J \neq \emptyset$), then

$$\ell(I \cup J) \leq \ell(I) + \ell(J).$$

237
**III.1.9.19. Theorem.** [Fundamental Theorem of Interval Length] Let $I_1, \ldots, I_n, J$ be intervals in $\mathbb{R}$. If $J \subseteq I_1 \cup \cdots \cup I_n$, then

$$\ell(J) \leq \sum_{k=1}^{n} \ell(I_k).$$

**Proof:** By induction on $n$. The case $n = 1$ is III.1.9.18.(i). Suppose the statement is true for $n$, and let $J \subseteq I_1 \cup \cdots \cup I_{n+1}$. We may assume $I_1$ is bounded, $I_1 \cap J \neq \emptyset$, and $J \not\subseteq I_1$. Let $a_1$ and $b_1$ be the left and right endpoints of $I_1$. Then either $a_1$ or $b_1$ is in $J$. Suppose $a_1 \in J$. (1) If $a_1 \notin I_1$, then there is an $I_k$ with $a_1 \in I_k$ and $I_1 \cup I_k$ is an interval. (2) If $a_1 \in I_1$ and $J$ contains a $c < a_1$, then there is one of the intervals $I_k$ with endpoints $a_k, b_k$ with $a_k < a_1$ and $b_k \geq a_1$ since there are only finitely many $I_k$ [if not, $d = \max\{b_k : a_k < a_1\} < a_1$ and $x \in J \setminus I_k$ for $d < x < a_1$]; then also $I_1 \cup I_k$ is an interval. The case where $b_1 \in J$ is similar. We must be in either case (1) or (2) for at least one of $a_1$ and $b_1$ since $J \not\subseteq I_1$.

Thus there must be an $I_k$ such that $I_1 \cup I_k$ is an interval. By renumbering, we may assume $k = n + 1$. Set $I'_1 = I_1 \cup I_{n+1}$. Then $J \subseteq I'_1 \cup I_2 \cup \cdots \cup I_n$, and $\ell(I'_1) \leq \ell(I_1) + \ell(I_{n+1})$ by III.1.9.18.(ii). Thus, by the inductive hypothesis,

$$\ell(J) \leq \ell(I'_1) + \sum_{k=2}^{n} \ell(I_k) \leq \sum_{k=1}^{n+1} \ell(I_k).$$


**III.1.10. The Archimedean Property and Density of $\mathbb{Q}$**

Using the Completeness Axiom, we now show that the Archimedean property (cf. (i)) also holds in $\mathbb{R}$:

**III.1.10.1. Theorem.** Let $a, b \in \mathbb{R}$ with $a > 0$. Then there is an $n \in \mathbb{N}$ with $b < na$.

**Proof:** Suppose $na \leq b$ for all $n \in \mathbb{N}$, i.e. $b$ is an upper bound for $A = \{na : n \in \mathbb{N}\}$. Then $A$ is bounded above, so by the Completeness Axiom $A$ has a supremum. Set $s = \sup(A)$. Then $s - a$ is not an upper bound for $A$ since $a > 0$, so there is an $n \in \mathbb{N}$ with $na > s - a$. But then $(n + 1)a \in A$ and $(n + 1)a = na + a > (s - a) + a = s$, contradicting the fact that $s$ is an upper bound for $A$. Thus the assumption that $na \leq b$ for all $n \in \mathbb{N}$ is false.

**III.1.10.2. Corollary.** Let $b \in \mathbb{R}$. Then there is an $n \in \mathbb{N}$ with $b < n$. ($\mathbb{N}$ is not bounded above in $\mathbb{R}$.)

**Proof:** Apply III.1.10.1. with $a = 1$.

**III.1.10.3. Corollary.** $\inf \{\frac{1}{n} : n \in \mathbb{N}\} = 0$, i.e. if $\epsilon \in \mathbb{R}$, $\epsilon > 0$, there is an $n \in \mathbb{N}$ with $\frac{1}{n} < \epsilon$.

**Proof:** $0$ is a lower bound for $\{\frac{1}{n} : n \in \mathbb{N}\}$, so $\inf \{\frac{1}{n} : n \in \mathbb{N}\}$ exists. If $\epsilon \in \mathbb{R}$, $\epsilon > 0$, then there is an $n \in \mathbb{N}$ with $\frac{1}{n} < \epsilon$. For this $n$, $\frac{1}{n} < \epsilon$. Thus $\epsilon$ is not a lower bound for $\{\frac{1}{n} : n \in \mathbb{N}\}$.

There is a tightening of III.1.10.2. which is an important observation:

238
**III.1.10.4.** Proposition. Let $x \in \mathbb{R}$, then there is a unique $n \in \mathbb{Z}$ with $n \leq x < n + 1$. The integer $n$ is called the integer part of $x$, denoted $\lfloor x \rfloor$.

**Proof:** Suppose first $x \geq 0$. Then $\{n \in \mathbb{N} : x < n\}$ is nonempty; let $m$ be the smallest element of this set, and set $n = m - 1$. Then $n \leq x < n + 1$. If $x < 0$, there is an $r \in \mathbb{N}$ such that $r \leq -x < r + 1$, so $-(r + 1) < x \leq -r$. If $x = -r$, set $n = -r$; otherwise set $n = -(r + 1)$.

The $n$ is unique since if $n \leq x < n + 1$ and $m \leq x < m + 1$, we have $-(m + 1) < x \leq -m$, so $n - (m + 1) < 0 < (n + 1) - m$, $-1 < m - n < 1$, $m - n = 0$.

We then get the density of $\mathbb{Q}$ in $\mathbb{R}$:

**III.1.10.5.** Theorem. Let $a, b \in \mathbb{R}$, $a < b$. Then there is an $r \in \mathbb{Q}$ with $a < r < b$.

In fact, by applying the theorem repeatedly to the intervals $(a, b)$, $(a, r)$, $(r, b)$, etc., it is readily shown that there are infinitely many rational numbers in the interval $(a, b)$.

**Proof:** By III.1.10.3, there is an $n \in \mathbb{N}$ with $\frac{1}{n} < b - a$. For this $n$, there is an $m \in \mathbb{Z}$ with $m - 1 \leq na < m$. We have $a < \frac{m}{n}$. Also, $m - na \leq 1$, $m \leq na + 1$, so

\[
\frac{m}{n} \leq a + \frac{1}{n} < a + (b - a) = b .
\]

**III.1.10.6.** Corollary. Let $a \in \mathbb{R}$, and set $A = \{r \in \mathbb{Q} : r < a\}$. Then $\sup(A) = a$.

**Proof:** $A$ is nonempty ($-n \in A$ for some $n \in \mathbb{N}$), and $a$ is an upper bound for $A$, so $s = \sup(A)$ exists and satisfies $s \leq a$. If $s < a$, then there is an $r \in \mathbb{Q}$ with $s < r < a$. But then $r \in A$, contradicting that $s$ is an upper bound for $A$. Thus $s = a$.

**III.1.11. Uniqueness of $\mathbb{R}$**

We can now prove that the real numbers are uniquely determined by the axioms, i.e. that any two structures satisfying the axioms for $\mathbb{R}$ are the same up to a renaming of the elements. Specifically:

**III.1.11.1.** Theorem. Let $\mathbb{R}$ and $\mathbb{R}'$ be sets with addition, multiplication, and order satisfying the algebraic, order, and completeness axioms. Then there is a unique function $\phi : \mathbb{R} \to \mathbb{R}'$ which is one-to-one and onto, such that for all $x, y \in \mathbb{R}$,

\[
\phi(x + y) = \phi(x) + \phi(y) \\
\phi(xy) = \phi(x)\phi(y) \\
\phi(x) < \phi(y) \text{ if and only if } x < y
\]
Similarly, define \( \phi \) and \( \phi' \) of adding 1, rational numbers. Let \( \phi, \phi' \) of III.1.11.2. and hence \( \phi(0) = 0' \) by additive cancellation in \( \mathbb{R}' \). Similarly,
\[
\phi(1) \cdot \phi(1) = \phi(1 \cdot 1) = \phi(1) = 1' \cdot \phi(1)
\]
and, since \( \phi(1) \neq \phi(0) = 0' \) because \( \phi \) is one-to-one, we have \( \phi(1) = 1' \) by multiplicative cancellation.

III.1.11.2. Now note that \( \mathbb{R} \) and \( \mathbb{R}' \) (in fact, any ordered field) each contain a “copy” (model) of the rational numbers. Let \( \mathbb{N} \) be the smallest subset of \( \mathbb{R} \) containing its multiplicative identity 1 and closed under adding 1, \( \mathbb{Z} \) the smallest subset of \( \mathbb{R} \) containing \( \mathbb{N} \) and closed under subtraction, and \( \mathbb{Q} \) the smallest subset of \( \mathbb{R} \) containing \( \mathbb{Z} \) and closed under division by nonzero elements; similarly, let \( \mathbb{N}' \) be the smallest subset of \( \mathbb{R}' \) containing its multiplicative identity 1' and closed under adding 1', and \( \mathbb{Z}' \) and \( \mathbb{Q}' \) the corresponding subsets of \( \mathbb{R}' \) closed under subtraction and division.

III.1.11.3. We will define \( \phi : \mathbb{R} \rightarrow \mathbb{R}' \) and its inverse \( \psi : \mathbb{R}' \rightarrow \mathbb{R} \) in stages. We begin by defining \( \phi \) and \( \psi \) inductively (recursively) on \( \mathbb{N} \) and \( \mathbb{N}' \). Define \( \phi(1) = 1' \), and if \( \phi(n) \) has been defined, define \( \phi(n + 1) = \phi(n) + 1' \). (We should write \( +', ' \), \( <' \) for the algebraic operations and order relation in \( \mathbb{R}' \), but we will not do so to avoid becoming overprimed.) This defines \( \phi \) on all of \( \mathbb{N} \), with range contained in \( \mathbb{N}' \). Similarly, define \( \psi : \mathbb{N}' \rightarrow \mathbb{N} \) inductively by \( \psi(1') = 1 \) and \( \psi(n' + 1') = \psi(n') + 1 \).

III.1.11.4. Lemma. We have \( \psi(\phi(n)) = n \) and \( \phi(\psi(n')) = n' \) for all \( n \in \mathbb{N}, n' \in \mathbb{N}' \). Thus \( \phi \) and \( \psi \) are mutually inverse bijections between \( \mathbb{N} \) and \( \mathbb{N}' \).

Proof: This is an easy proof by induction. We have \( \psi(\phi(1)) = \psi(1') = 1 \). If \( \psi(\phi(n)) = n \), then
\[
\psi(\phi(n + 1)) = \psi(\phi(n) + 1') = \psi(\phi(n)) + 1 = n + 1
\]
so \( \psi(\phi(n)) = n \) for all \( n \in \mathbb{N} \). The other formula is proved similarly by induction on \( \mathbb{N}' \).

III.1.11.5. Lemma. If \( n, m \in \mathbb{N} \), then \( \phi(n + m) = \phi(n) + \phi(m) \) and \( \phi(nm) = \phi(n)\phi(m) \).

Proof: For a fixed \( n \in \mathbb{N} \), we prove the formula by induction on \( m \). We have \( \phi(n + 1) = \phi(n) + 1' = \phi(n) + \phi(1) \), so the formula holds for \( m = 1 \). If \( \phi(n + m) = \phi(n) + \phi(m) \), then
\[
\phi(n + (m + 1)) = \phi((n + m) + 1) = \phi(n + m) + 1' = (\phi(n) + \phi(m)) + 1' = \phi(n) + (\phi(m) + 1') = \phi(n) + \phi(m + 1)
\]
so the formula holds for all \( m \). The formula for multiplication is proved similarly.
III.1.11.6. Lemma. If \( n, m \in \mathbb{N} \), then \( \phi(n) < \phi(m) \) if and only if \( n < m \).

Proof: We have \( n < m \) if and only if there is a \( p \in \mathbb{N} \) with \( m = n + p \). Then
\[
\phi(m) = \phi(n + p) = \phi(n) + \phi(p)
\]
so \( \phi(n) < \phi(m) \). The argument also works for \( \psi \), giving the converse. \( \blacksquare \)

III.1.11.7. We now extend \( \phi \) and \( \psi \) to \( \mathbb{Z} \) and \( \mathbb{Z}' \). If \( x \in \mathbb{Z} \), write \( x = n - m \) for \( n, m \in \mathbb{N} \), and define \( \phi(x) = \phi(n) - \phi(m) \). This is well defined since if \( n - m = q - p \), i.e. \( n + p = m + q \), we have
\[
\phi(n) + \phi(p) = \phi(n + p) = \phi(m + q) = \phi(m) + \phi(q)
\]
and thus \( \phi(q) - \phi(p) = \phi(n) - \phi(m) \). Alternatively, write \( \mathbb{Z} \) as the disjoint union of \( \mathbb{N} \), \( \{0\} \) and \( -\mathbb{N} \), and \( \mathbb{Z}' \) as the disjoint union of \( \mathbb{N}' \), \( \{0'\} \), and \( -\mathbb{N}' \), and define \( \phi(0) = 0' \) and \( \phi(-n) = -\phi(n) \) for \( n \in \mathbb{N} \). Define \( \psi : \mathbb{Z}' \rightarrow \mathbb{Z} \) in the same manner. It is obvious, especially from the second definition, that \( \psi \) is the inverse for \( \phi \), so \( \phi \) and \( \psi \) are bijections between \( \mathbb{Z} \) and \( \mathbb{Z}' \).

The next result is just a simple but somewhat tedious calculation checking case-by-case:

III.1.11.8. Lemma. If \( x, y \in \mathbb{Z} \), then \( \phi(x+y) = \phi(x) + \phi(y) \) and \( \phi(xy) = \phi(x)\phi(y) \). We have \( \phi(x) < \phi(y) \) if and only if \( x < y \).

III.1.11.9. Now we extend \( \phi \) and \( \psi \) to \( \mathbb{Q} \) and \( \mathbb{Q}' \). If \( x \in \mathbb{Q} \), write \( x = \frac{m}{n} \) for \( m, n \in \mathbb{Z} \), \( n \neq 0 \), and set \( \phi(x) = \frac{\phi(m)}{\phi(n)} \in \mathbb{Q}' \). This makes sense since \( \phi(n) \neq 0 \). Then \( \phi \) is well defined since, if \( \frac{m}{n} = \frac{p}{q} \), i.e. \( mq = np \), we have
\[
\phi(m)\phi(q) = \phi(mq) = \phi(np) = \phi(n)\phi(p)
\]
so \( \frac{\phi(m)}{\phi(n)} = \frac{\phi(p)}{\phi(q)} \). Similarly define \( \psi : \mathbb{Q}' \rightarrow \mathbb{Q} \). It is clear that \( \psi \) is the inverse of \( \phi \), so \( \phi \) and \( \psi \) are bijections.

III.1.11.10. Lemma. If \( x, y \in \mathbb{Q} \), then \( \phi(x+y) = \phi(x) + \phi(y) \) and \( \phi(xy) = \phi(x)\phi(y) \). We have \( \phi(x) < \phi(y) \) if and only if \( x < y \).

Proof: Write \( x = \frac{m}{n}, y = \frac{p}{q} \) with \( m, p \in \mathbb{Z} \) and \( n, q \in \mathbb{N} \). Then
\[
\phi(x+y) = \phi \left( \frac{m}{n} + \frac{p}{q} \right) = \phi \left( \frac{mq + np}{nq} \right) = \frac{\phi(m)\phi(q) + \phi(n)\phi(p)}{\phi(n)\phi(q)} = \frac{\phi(m)}{\phi(n)} + \frac{\phi(p)}{\phi(q)} = \phi(x) + \phi(y)
\]
\[
\phi(xy) = \phi \left( \frac{m}{n} \cdot \frac{p}{q} \right) = \phi \left( \frac{mp}{nq} \right) = \frac{\phi(mp)}{\phi(nq)} = \frac{\phi(m)\phi(p)}{\phi(n)\phi(q)} = \frac{\phi(m)}{\phi(n)} \cdot \frac{\phi(p)}{\phi(q)} = \phi(x)\phi(y).
\]
If \( x < y \), then \( mq < np \), so
\[
\phi(m)\phi(q) = \phi(mq) < \phi(np) = \phi(n)\phi(p)
\]
so we have \( \phi(x) = \frac{\phi(m)}{\phi(n)} < \frac{\phi(p)}{\phi(q)} = \phi(y) \), and the argument is reversible. \( \blacksquare \)
III.1.11.11. It is clear from the requirement that $\phi(0) = 0'$ and $\phi(1) = 1'$ that the definition of $\phi$ and its inverse $\psi$ between $N$ and $N'$ can only be made in the manner of III.1.11.3., i.e. there are unique functions $\phi$ and $\psi$ between $N$ and $N'$ with the specified properties. The extensions to $\mathbb{Z}$ and $\mathbb{Q}$ are then forced by the properties of $\phi$ needed, i.e. the uniqueness statement applies to functions from $\mathbb{Q}$ to $\mathbb{Q}'$ also.

III.1.11.12. So far, we have only used the algebraic and order axioms, so the arguments work for any ordered fields: if $F$ and $F'$ are ordered fields, then the smallest subfields of $F$ and $F'$ are naturally isomorphic (and isomorphic to $\mathbb{Q}$). Thus we may naturally regard $\mathbb{Q}$ as a subfield of any ordered field.

Now we use the Completeness Axiom and density of $\mathbb{Q}$ in $\mathbb{R}$, and density of $\mathbb{Q}'$ in $\mathbb{R}'$, to extend $\phi$ and $\psi$ to $\mathbb{R}$ and $\mathbb{R}'$.

III.1.11.13. Definition. Let $x \in \mathbb{R}$. Let $\phi(x) = \sup\{\phi(r) : r \in \mathbb{Q}, r < x\}$. If $x' \in \mathbb{R}'$, define $\psi(x') = \sup\{\psi(r') : r' \in \mathbb{Q}', r' < x'\}$.

It follows immediately from III.1.10.6. applied in both $\mathbb{R}$ and $\mathbb{R}'$ that if $x \in \mathbb{Q}$, $x' \in \mathbb{Q}'$, this definition of $\phi(x)$ and $\psi(x')$ agrees with the previous one, and that $\psi$ is the inverse of $\phi$ in general, so $\phi$ and $\psi$ are mutually inverse bijections between $\mathbb{R}$ and $\mathbb{R}'$.

III.1.11.14. Lemma. If $x, y \in \mathbb{R}$, then $x < y$ if and only if $\phi(x) < \phi(y)$.

Proof: Suppose $x < y$. Let $t \in \mathbb{Q}$, $x < t < y$. Then $\phi(t) \leq \phi(y)$ by definition of $\phi(y)$, and $\phi(t) \neq \phi(y)$ since $\phi$ is injective, so $\phi(t) < \phi(y)$. Also, $\phi(r) < \phi(t)$ for every $r < x$, so $\phi(t)$ is an upper bound for $\{\phi(r) : r \in \mathbb{Q}, r < x\}$ and therefore $\phi(x) \leq \phi(t)$ (even $\phi(x) < \phi(t)$ since $\phi$ is injective). So $\phi(x) < \phi(y)$ by transitivity. The same argument works for $\psi$, giving the converse.

III.1.11.15. Corollary. If $x \in \mathbb{R}$, then $\phi(x) = \inf\{\phi(r) : r \in \mathbb{Q}, x < r\}$.

Proof: Set $b' = \inf\{\phi(r) : r \in \mathbb{Q}, x < r\}$. Since $\phi(x)$ is a lower bound for this set, $\phi(x) \leq b'$. If $\phi(x) < b'$, let $r' \in \mathbb{Q}'$ with $\phi(x) < r' < b'$. Then $r' = \phi(r)$ for some $r \in \mathbb{Q}$, $x < r$, contradicting that $b'$ is a lower bound. Thus $b' = \phi(x)$.

III.1.11.16. Corollary. If $x \in \mathbb{R}$, then $\phi(-x) = -\phi(x)$.

It remains to show that $\phi$ preserves addition and multiplication.
III.1.11.17.  **Lemma.** Let \( x, y \in \mathbb{R} \). Then \( \phi(x + y) = \phi(x) + \phi(y) \) and \( \phi(xy) = \phi(x)\phi(y) \).

**Proof:** Suppose \( a \in \mathbb{Q} \). Then, if \( r \in \mathbb{Q} \), we have \( a + r < a + y \) if and only if \( r < y \), so

\[
\phi(a + y) = \sup\{\phi(a + r) : r \in \mathbb{Q}, r < y\} = \sup\{\phi(a) + \phi(r) : r \in \mathbb{Q}, r < y\}.
\]

By III.1.9.6., the last supremum equals

\[
\phi(a) + \sup\{\phi(r) : r \in \mathbb{Q}, r < y\} = \phi(a) + \phi(y).
\]

Then we have

\[
\phi(a) + \phi(y) = \phi(a + y) < \phi(x + y)
\]

for every \( a \in \mathbb{Q} \), \( a < x \), and

\[
\phi(x + y) < \phi(b + y) = \phi(b) + \phi(y)
\]

for every \( b \in \mathbb{Q} \), \( x < b \). But by III.1.11.15. and III.1.9.6., we have

\[
\phi(x) + \phi(y) = \sup\{\phi(a) + \phi(y) : a \in \mathbb{Q}, a < x\} = \inf\{\phi(b) + \phi(y) : b \in \mathbb{Q}, x < b\}
\]

and so \( \phi(x + y) = \phi(x) + \phi(y) \).

The proof that \( \phi(xy) = \phi(x)\phi(y) \) is very similar if \( x \) and \( y \) are positive. The proof for general \( x, y \) then follows from this and III.1.11.16.

III.1.11.18. By the density of \( \mathbb{Q} \) in \( \mathbb{R} \) and of \( \mathbb{Q}' \) in \( \mathbb{R}' \), the values of \( \phi \) and \( \psi \) are completely determined by their values on \( \mathbb{Q} \) and \( \mathbb{Q}' \) respectively. Since the functions between \( \mathbb{Q} \) and \( \mathbb{Q}' \) are unique (III.1.11.1.), we obtain the uniqueness of the function \( \phi \) (and \( \psi \)) with the specified properties.

This completes the proof of Theorem III.1.11.1.

III.1.12.  **Existence of Roots**

Although we will later get more general results, we show here using the Completeness Axiom that every positive real number has arbitrary roots. We first need an estimate:

III.1.12.1.  **Proposition.** Let \( a, b \in \mathbb{R} \). Set \( c = \max(|a|, |b|) \). If \( n \in \mathbb{N} \), then

\[
|b^n - a^n| \leq nc^{n-1}|b - a|.
\]

**Proof:** This is easily proved using the Mean Value Theorem, but there is an elementary proof: it follows immediately from the identity (for \( n > 1 \))

\[
b^n - a^n = (b - a) \sum_{k=0}^{n-1} a^k b^{n-k-1}
\]
which is easily verified by an algebraic calculation. Then we have, by the triangle inequality,

$$|b^n - a^n| = |b - a|\left|\sum_{k=0}^{n-1} a^k b^{n-k-1}\right| \leq |b - a| \sum_{k=0}^{n-1} |a|^k |b|^{n-k-1} \leq |b - a| \cdot nc^{n-1}.$$  

\* III.1.12.2.  **Theorem.** Let $a \in \mathbb{R}$, $a \geq 0$, and $n \in \mathbb{N}$. Then there is a unique $x \in \mathbb{R}$, $x \geq 0$, with $x^n = a$.

**Proof:** If there is a solution, it is unique since by the identity in the proof of III.1.12.1., if $0 < x < y$, then $x^n < y^n$. If $a = 0$, the unique solution is $x = 0$. For existence, it suffices to consider the case $0 < a < 1$ since if $a > 1$, so $\frac{1}{a} < 1$, and $x^n = \frac{1}{a}$, then $\left(\frac{1}{x}\right)^n = a$, and the case $a = 1$ is obvious.

Set 

$$A = \{ r \in \mathbb{R} : r > 0, r^n < a \}.$$  

$A$ is nonempty since there is an $k \in \mathbb{N}$ with $\frac{1}{k} < a$, and for this $k$ we have $\frac{1}{k^n} \leq \frac{1}{k} < a$. Also, $A$ is bounded: 1 is an upper bound for $A$. Thus $x = \sup(A)$ exists, and $x \leq 1$.

We claim $x^n = a$. First suppose $x^n > a$, and choose $k \in \mathbb{N}$ with $\frac{1}{k} < x$ and $\frac{1}{k} < \frac{a^n - a}{x^n}$. Then, by III.1.12.1.,

$$x^n - \left(x - \frac{1}{k}\right)^n \leq nx^{n-1} \frac{1}{k} \leq n \frac{1}{k} < x^n - a$$

and so $(x - \frac{1}{k})^n > a$ and $x - \frac{1}{k}$ is an upper bound for $A$, contradicting that $x$ is the least upper bound of $A$. Thus $x^n \leq a$.

So $x < 1$ since $a < 1$. If $x^n < a$, choose $k \in \mathbb{N}$ so that $x + \frac{1}{k} < 1$ and $\frac{1}{k} < \frac{a - x^n}{a}$. Then, by III.1.12.1.,

$$\left(x + \frac{1}{k}\right)^n - x^n \leq n \left(x + \frac{1}{k}\right)^{n-1} \frac{1}{k} < n \frac{1}{k} < a - x^n$$

so $(x + \frac{1}{k})^n < a$, contradicting that $x$ is an upper bound for $A$. Thus $x^n = a$.

If $n$ is odd, then negative real numbers have $n$’th roots too:

III.1.12.3. **Corollary.** Let $n \in \mathbb{N}$ be odd, and $a \in \mathbb{R}$. Then there is a unique $x \in \mathbb{R}$ with $x^n = a$.

**Proof:** If $a \geq 0$, existence comes from III.1.12.2. If $a < 0$, then by III.1.12.2. there is an $x \in \mathbb{R}$ with $x^n = -a$. Then, since $n$ is odd, $(-x)^n = -x^n = a$. Uniqueness follows as in the beginning of the proof of III.1.12.2.

III.1.12.4. If $n$ is even and $a < 0$, there is no solution to $x^n = a$ in $\mathbb{R}$, since if $n = 2m$, then $x^n = (x^m)^2 \geq 0$ for any $x \in \mathbb{R}$ (. If $a > 0$, there are two solutions, one positive and one negative. As usual, we will denote the positive solution as $\sqrt[n]{a}$. If $n$ is odd, then for any $a$ we denote the unique solution by $\sqrt[n]{a}$. When $n = 2$ (square root), we just write $\sqrt{a}$ for the nonnegative root when $a \geq 0$. 

244
III.1.12.5. So every positive real number has a square root. In particular, there is a real number $\sqrt{2}$ whose square is 2. Since there is no rational number whose square is 2 (I.5.5.), $\sqrt{2}$ is a real number which is not rational (an irrational number). It turns out that “most” real numbers are irrational ()

The classical proof that $\sqrt{2}$ is irrational is given in I.5.5.. Here is a more general result; the special case $d = 2$ gives another pretty proof for $\sqrt{2}$ which is even simpler than the classic one.

III.1.12.6. Theorem. If $d \in \mathbb{N}$, then $\sqrt{d}$ is either a natural number or irrational.

Proof: Suppose $\sqrt{d}$ is not a natural number. Since $\sqrt{d} > 1$ (because $(\sqrt{d})^2 > 1^2$; cf. III.1.2.3.(vii)), there is a $q \in \mathbb{N}$ with $q < \sqrt{d} < q + 1$, i.e. $0 < \sqrt{d} - q < 1$ (III.1.10.4.). (If $d = 2$, then $q = 1$.)

Now suppose $\sqrt{d}$ is rational. Then there is a natural number $n$ such that $n\sqrt{d}$ is an integer, i.e.

$$A = \{ n \in \mathbb{N} : n\sqrt{d} \in \mathbb{Z} \}$$

is nonempty. By III.1.4.12., $A$ has a smallest element $n$. But then

$$m := (\sqrt{d} - q)n = n\sqrt{d} - nq \in \mathbb{Z}$$

and $m > 0$ since $\sqrt{d} - q > 0$; hence $m \in \mathbb{N}$. Also,

$$m\sqrt{d} = (d - q\sqrt{d})n = dn - q(n\sqrt{d}) \in \mathbb{Z}$$

so $m \in A$. But $m < n$ since $\sqrt{d} - q < 1$, contradicting that $n$ is the smallest element of $A$. 

This proof is due to Dedekind. See III.1.13.6. for another version of the proof, and III.9.5.4. for a generalization.

III.1.12.7. One can significantly extend III.1.12.3. using the notions of limits and continuity, and the Intermediate Value Theorem. For example, it can be easily shown (V.2.3.6.) that every polynomial of odd degree with real coefficients has a root in $\mathbb{R}$ (III.1.12.3. is the special case of the polynomial $x^n - a$).

III.1.12.8. An ordered field in which every positive element has a square root and every polynomial of odd degree has a root is called a real-closed field. There are many other equivalent characterizations of real-closed fields. The theory of real-closed fields turns out to be exactly the first-order theory of $\mathbb{R}$, i.e. any statement true in $\mathbb{R}$ which has a first-order formulation (not allowing quantification over subsets) is true in every real-closed field. There are many real-closed fields other than $\mathbb{R}$, including countable examples (actually, the field of real algebraic numbers is a countable example). Other examples include any model for the hyperreal numbers of nonstandard analysis (), and the Levi-Civita numbers. See () for a description of the theory of real-closed fields.

III.1.13. Exercises

III.1.13.1. Prove by induction on $n$ that if $\{x_k, y_k : 1 \leq k \leq n \}$ is a set of real numbers, then

$$\sum_{k=1}^{n}(x_k + y_k) = \sum_{k=1}^{n}x_k + \sum_{k=1}^{n}y_k .$$
III.1.13.2.  (a) Let \( x \in \mathbb{R} \). Show how to define \( x^n \) for each \( n \in \mathbb{N} \) by recursion.
(b) Prove by induction that if \( x, y \in \mathbb{R} \) and \( m, n \in \mathbb{N} \), then \((xy)^n = x^ny^n\), \( x^mx^n = x^{m+n} \), \((x^m)^n = x^{mn}\), and, if \( y \neq 0 \), \((y^n)^{-1} = (y^{-1})^n\), \((\frac{x}{y})^n = x^n y^{-n}\) and \(\frac{y^m}{y^n} = y^{m-n}\). Extend these rules of exponents to arbitrary integer powers, with appropriate hypotheses.

III.1.13.3.  Show how to carefully define \( n! = 1 \cdot 2 \cdot 3 \cdots n \) for all \( n \in \mathbb{N} \) by recursion.

III.1.13.4.  Define the Fibonacci sequence \( (x_n) \) by \( x_1 = 1 \), \( x_2 = 1 \), \( x_n = x_{n-1} + x_{n-2} \) for \( n > 2 \).
(a) Show that this definition can be phrased as a proper recursive definition under III.1.5.15.
(b) Prove by induction that any two consecutive terms in the sequence are relatively prime.
(c) Rephrase the argument of (b) as an Infinite Regress argument. Which way is simpler?

III.1.13.5.  Carefully prove the assertions in III.1.7.2. and III.1.7.4.

(a) Suppose \( \sqrt{d} \) is not an integer, but is rational. Write \( \sqrt{d} = \frac{a}{b} \) in lowest terms (e.g. with \( b \) minimal in \( \mathbb{N} \)), with \( a, b \in \mathbb{N} \). Then also \( \sqrt{d} = \frac{d}{\sqrt{d}} = \frac{db}{a} \).
(b) Since \( \frac{db}{a} = \frac{a}{b} \), the fractional parts (which are nonzero since \( \sqrt{d} \notin \mathbb{N} \)) are equal. Thus there are \( p, q \in \mathbb{N} \) with \( p < a \) and \( q < b \) and \( \frac{p}{a} = \frac{q}{b} \).
(c) From \( \frac{p}{a} = \frac{q}{b} \) we get \( \frac{p}{q} = \frac{a}{b} = \sqrt{d} \), contradicting that \( \frac{a}{b} \) is in lowest terms.
(d) If \( r \) is a positive rational number and \( r = \frac{a}{b} \) with \( b \) minimal in \( \mathbb{N} \), then the fraction \( \frac{a^n}{b^n} \) is also the representation of \( a^{n-1}r \) with minimal denominator in \( \mathbb{N} \) (cf. III.1.7.4.).
(e) Using (d), prove that if \( d, n \in \mathbb{N} \), then \( \sqrt{d} \) is either an integer or irrational (cf. III.9.5.4.). [Use that \( \sqrt{d} = \frac{d}{(\sqrt{d})^{n-1}r} \).]
III.2. The Natural Numbers

We now begin the construction of the number systems from scratch, beginning with the natural numbers.

III.2.1. Introduction

III.2.1.1. The first set of numbers developed, going back to prehistoric times, is the natural numbers or counting numbers, also described as the positive integers. This set is usually called \( \mathbb{N} \). Thus, as it is often described or defined,

\[ \mathbb{N} = \{1, 2, 3, \ldots \} \]

Unfortunately, this “description” is totally inadequate as a definition of \( \mathbb{N} \), even an informal one, for two reasons. Most importantly, it is entirely unclear what the “…” means (if you think you know, try giving a precise formulation of the meaning!) And even if this point is specified, this description says nothing about the operations of arithmetic which are an essential part of the structure.

III.2.1.2. One mathematically minor point needs to be addressed. There is no general agreement among mathematicians about whether \( \mathbb{N} \) includes 0; some mathematicians include it, and others do not. We will not. Set theorists generally include it, since from the point of view of the set-theoretic construction of \( \mathbb{N} \) it is natural to do so. However, in my opinion 0 should not be called a “natural number.” The positive integers historically far predate the emergence of the concept of 0, and there are important societies, and even some modern-day societies, that never developed 0; the notion of 0 is far more abstract and subtle than that of numbers such as 1 and 2. (On the other hand, I acknowledge that it is also historically questionable whether “large” positive integers should be regarded as “natural numbers,” since there is a practical limit to the size of numbers which occur “naturally,” and in addition the development of notation which allows one to even contemplate numbers such as \( 10^{10^{10}} \) is relatively recent compared even to the concept of 0.) However, it is useful to have a notation for the set \( \mathbb{N} \cup \{0\} \) of nonnegative integers; we will use the notation \( \mathbb{N}_0 \) for this set. The question of whether to work with \( \mathbb{N} \) or \( \mathbb{N}_0 \) as the basic set is strictly a matter of taste, since it is mathematically trivial to go from either to the other.

III.2.2. Axiomatic Characterization of \( \mathbb{N} \)

III.2.2.1. Rather than trying to define what \( \mathbb{N} \) is, it is more useful to list a set of basic properties that \( \mathbb{N} \) has, from which all the usual structure can be derived, and then to show there is an essentially unique mathematical structure having these properties. The basic axiom scheme was first developed by G. Peano in 1889, although the actual construction of \( \mathbb{N} \) along these lines which motivated the axiomatization had already been done by R. Dedekind.

**Peano Axioms for \( \mathbb{N} \):**

(P1) There is a natural number 1.

(P2) Every natural number \( n \) has a successor \( S(n) \).

(P3) 1 is not the successor of any natural number.

(P4) If \( m, n \in \mathbb{N} \) and \( S(m) = S(n) \), then \( m = n \).

(P5) (Axiom of Induction) If \( A \subseteq \mathbb{N} \), 1 \( \in \) \( A \), and \( n \in A \) implies \( S(n) \in A \), then \( A = \mathbb{N} \).
The Axiom of Induction can be equivalently rephrased: “The smallest subset of \( \mathbb{N} \) containing 1 and closed under taking successors is all of \( \mathbb{N} \).” To make precise the notion of “smallest,” note that if \( \{ A_i \} \) is a collection of subsets of \( \mathbb{N} \) each containing 1 and closed under taking successors, then \( \bigcap_i A_i \) also contains 1 and is closed under taking successors. Thus the intersection of all subsets which contain 1 and are closed under taking successors is clearly the smallest such set.

**III.2.2.2.** This theory is technically a second-order theory since in the Axiom of Induction, there is quantification over subsets of \( \mathbb{N} \). There is a corresponding first-order theory, but it is not as strong and has infinitely many axioms replacing the Axiom of Induction. See books on mathematical logic and model theory, such as (), for an explanation of first-order Peano arithmetic and the difference between first-order and second-order theories.

Here are some simple observations. The first is a “converse” to axiom (P3):

**III.2.2.3.** Proposition. If \( n \in \mathbb{N} \), \( n \neq 1 \), then there is an \( m \in \mathbb{N} \) with \( n = S(m) \).

Proof: Let \( A \) be the set of all \( n \in \mathbb{N} \) such that either \( n = 1 \) or there is an \( m \in \mathbb{N} \) with \( n = S(m) \). We trivially have \( 1 \in A \). If \( n \in A \) (actually, if just \( n \in \mathbb{N} \)), then obviously \( S(n) \in A \). By axiom (P5), \( A = \mathbb{N} \). \( \square \)

**III.2.2.4.** Proposition. \( \mathbb{N} \) is infinite (in the sense of Dedekind (II.5.2.1.).)

Proof: The function \( S : \mathbb{N} \to \mathbb{N} \) is injective by axiom (P4), but not surjective by axiom (P3). \( \square \)

Sets and Models

“The definitions of number are very numerous and of great variety . . . We must not be surprised that there are so many. If any one of them was satisfactory we should not get any new ones.”

*H. Poincaré*

There are some logical subtleties in regarding \( \mathbb{N} \) as a “set”, which we now discuss briefly.

**III.2.2.5.** Definition. A *model* for Peano arithmetic is a triple \( (\mathbb{N}, e, s) \), where \( \mathbb{N} \) is a set, \( e \in \mathbb{N} \), and \( s \) is a function from \( \mathbb{N} \) to \( \mathbb{N} \) satisfying

(P3) \( s(x) \neq e \) for every \( x \in \mathbb{N} \).

(P4) If \( x, y \in \mathbb{N} \) and \( s(x) = s(y) \), then \( x = y \) (i.e. \( s \) is injective).

(P5) If \( A \subseteq \mathbb{N} \), \( e \in A \), and \( x \in A \) implies \( s(x) \in A \), then \( A = \mathbb{N} \).

(Of course, Peano axioms (P1) and (P2), suitably renamed, also hold tautologically in \( (\mathbb{N}, e, s) \).)
There is a “standard” model for Peano arithmetic in ZF set theory or the equivalent, the set $\omega$ of finite ordinals. Ordinals are discussed more carefully in ( ), but here we briefly outline the construction of $\omega$, which is originally due to J. von Neumann. Begin with $\emptyset$, which plays the role of 1 (or, more commonly, 0 in $\mathbb{N}_0$; set theorists often call this set 0). Define the successor of $\alpha$ to be $S(\alpha) = \alpha \cup \{\alpha\}$. Thus the successor of $\emptyset$ is $S(\emptyset) = \{\emptyset\} = \{0\}$ (note that $\{\emptyset\} \neq \emptyset$) Set theorists usually call this set 1. Then $S(1) = S(S(\emptyset)) = \{\emptyset, \{\emptyset\}\} = \{0, 1\}$ (called 2 in set theory). Continuing,

$$S(2) = S(S(S(\emptyset))) = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0, 1, 2\}$$

(called 3 in set theory), etc. The Axiom of Infinity then gives the set $\omega$.

There is a logical subtlety in this model: it is open to question whether second-order Peano arithmetic can be constructed in ZF set theory, but it can. (There is a further logical subtlety in that it apparently cannot be proved that ZF set theory is consistent; cf. ( ).) We do not discuss these questions further, but simply refer the reader to books on mathematical logic such as ( ), and we agree not to be bothered by them from now on and accept $\omega$ as a valid model for $\mathbb{N}$.

Set theorists generally define $\mathbb{N}$ (or $\mathbb{N}_0$) to be $\omega$. However, this leads to logical difficulties: technically, the natural number 1, the integer 1, the rational number 1, the real number 1, and the complex number 1 are all different objects (sets) as defined in set theory, and they must somehow, and somewhat awkwardly, be identified together (cf. III.2.2.9).

“If numbers are sets, then they must be particular sets, for each set is some particular set. But if the number 3 is really one set rather than another, it must be possible to give some cogent reason for thinking so; for the position that this is an unknowable truth is hardly tenable.”

Paul Benacerraf

It is hard to conceive that the principles of logic entail that we must inevitably conclude that the number 3 (whether regarded as natural, integer, rational, real, or complex) must be the set $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, or any other specific set.

Another approach to defining natural numbers is frequently considered, reflecting their use as counting numbers (cardinal numbers): the number $n$ is defined to be the class of all sets with exactly $n$ elements (this can be carefully phrased in a non-circular manner, e.g. by fixing a model for $\mathbb{N}$ such as $\omega$ and taking $n$ to be the class of all sets equipotent with the corresponding element of the model). Thus, for example, the number 1 is defined to be the class of all singleton sets. This approach has logical difficulties, since these objects cannot themselves be sets (e.g. there is an obvious one-one correspondence between the class of all singleton sets and the class of all sets), and anyway does not avoid the problem that there is no obvious way to extend the definition coherently to negative integers or rational numbers, let alone irrational numbers (attempts to do this logically, such as in [?], become unmanageably complicated and convoluted).

It is simpler, although perhaps less satisfying logically, not to precisely define $\mathbb{N}$ as a set. This approach will carry through also when the number system is extended, i.e. we will not define individual numbers to be any particular sets. This will not cause any difficulties in the actual use of the number systems in analysis. Anyone who is uncomfortable with this approach can be comforted by the knowledge that the number system can, in a somewhat arbitrary and awkward way, be precisely constructed within set theory.

“Questions about the true ‘nature’ of mathematical objects – such as whether \( \sqrt{2} \) is a set – are always irrelevant to mathematical practice. In mathematics one is exclusively concerned with abstract relationships rather than with the ‘internal constitution’ of objects.”

\[ \text{Martin Davis}^4 \]

### III.2.2.9. Technically, what we can do is assume that within set theory we have a sufficient supply of atoms (\( \varepsilon \)), and whenever each new collection of numbers is defined as a set, a collection of additional atoms is designated to name those numbers not already named. The operations of set theory and arithmetic can then be transferred to this set of atoms. Thus we will regard \( \mathbb{N} \) (and, later, \( \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \) and \( \mathbb{C} \)) as sets, with \( \mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C} \), but we will not regard individual numbers as sets.

The (second-order) Peano axioms uniquely determine the structure of \( \mathbb{N} \), up to a simple renaming of its elements, in the following sense:

### III.2.2.10. Theorem. Let \((N,e,s)\) and \((N',e',s')\) be models of the Peano axioms. Then there is a unique bijection \( \phi : N' \to N \) with \( \phi(e') = e \) and \( \phi(s'(x)) = s(\phi(x)) \) for all \( x \in N' \). (Such a \( \phi \) is called an isomorphism of \((N',e',s')\) and \((N,e,s)\).

\[ \text{Proof:} \] It suffices to assume that \((N',e',s') = (\omega,0,S)\), where \( S \) is the ordinal successor function \( S(\alpha) = \alpha \cup \{ \alpha \} \). If \( \phi : \omega \to N \) and \( \psi : \omega \to N' \) are successor-preserving bijections, then so is \( \phi \circ \psi^{-1} : N' \to N \).

Inductively define a function \( \phi \) from \( \omega \) to \( N \) by \( \phi(0) = e \) and \( \phi(S(\alpha)) = s(\phi(\alpha)) \). This function \( \phi \) preserves successors by construction. If \( R \subseteq N \) is the range of \( \phi \), then \( e \in R \), and if \( x = \phi(\alpha) \in R \), then so is \( s(x) = \phi(S(\alpha)) \). Thus \( R = N \) by axiom (P5), and \( \phi \) is surjective.

To show that \( \phi \) is injective, let \( A \) be the set of all \( x \in N \) such that \( x \) has a unique preimage. If \( \phi(\alpha) = e \), then \( \alpha = 0 \) since otherwise \( \alpha = S(\beta) \) for some \( \beta \) (III.2.2.3.), and \( e = \phi(\alpha) = \phi(S(\beta)) = s(\phi(\beta)) \), contradicting axiom (P3). Thus \( e \in A \). Now suppose \( x \in A \), and suppose \( \alpha, \beta \in \omega \) with \( \phi(\alpha) = \phi(\beta) = s(x) \). Then \( \alpha = S(\sigma) \) and \( \beta = S(\tau) \) for some \( \sigma, \tau \in \omega \). We have

\[
\phi(\sigma) = \phi(S(\sigma)) = \phi(S(\alpha)) = s(s(x)) = s(\phi(\tau))
\]

and so \( \phi(\sigma) = s(\phi(\tau)) \) by axiom (P4). Thus \( \sigma = \tau \) since \( x \) has unique preimage, and hence \( \alpha = \beta \), i.e. \( s(x) \in A \). Thus \( A = N \) by axiom (P5) and \( \phi \) is injective.

There is no such theorem for first-order Peano arithmetic; in fact, by general facts about first-order logic (the Lowenheim-Skolem Theorem), there are models of first-order Peano arithmetic of arbitrary infinite cardinality.

### III.2.2.11. In particular, every model for Peano arithmetic is isomorphic to \( \omega \). Thus, by slight abuse of language, we can talk about the set of natural numbers even though we technically do not define \( \mathbb{N} \) precisely within set theory (i.e. to be any specific model).

Since we have not defined \( \mathbb{N} \) as a particular set, we must technically identify \( \mathbb{N} \) with a specific model to define functions in or out; such definitions are independent of the choice of model in an evident way, so we will freely use \( \mathbb{N} \) as a domain or range for functions as though it were an honest set (this is logically justifiable by the procedure of III.2.2.9.).
III.2.3. Induction and Recursion

We will not reiterate here the full content of (indeed we cannot yet). Note that with the careful construction of \( \mathbb{N} \), we can give a more direct and complete justification for the principles of induction and recursion. We do not yet have addition on \( \mathbb{N} \), so we cannot phrase things using “\( n + 1 \)” and we must instead use the successor function. We also do not yet have the ordering on \( \mathbb{N} \) so we cannot phrase Complete Induction or Infinite Regress.

III.2.3.1. [Principle of Induction] Suppose \( P(n) \) is a statement for each \( n \in \mathbb{N} \). Suppose the following two statements hold:

(i) \( P(1) \) is true.

(ii) For every \( n \in \mathbb{N} \), if \( P(n) \) is true, then \( P(S(n)) \) is true.

Then \( P(n) \) is true for all \( n \in \mathbb{N} \).

III.2.3.2. To see why the Principle of Induction holds, set

\[ A = \{ n \in \mathbb{N} : P(n) \text{ is true} \} . \]

We have that \( 1 \in A \) by (i), and if \( n \in A \) we have \( S(n) \in A \) by (ii). Thus \( A = \mathbb{N} \).

Definitions by Induction or Recursion

We will write sequences, which are really functions on \( \mathbb{N} \), using functional notation. The following theorem is not quite as general as III.1.5.15., but is the version we will use at present. The proof is similar to the proof of III.1.5.15., but somewhat more complicated since we have less machinery available.

III.2.3.3. Theorem. Let \( X \) be a set, and let \( g \) be a function from \( X \) to \( X \). If \( x_1 \) is any element of \( X \), then there is a unique function \( f : \mathbb{N} \rightarrow X \) with \( f(1) = x_1 \) and \( f(S(n)) = g(f(n)) \) for all \( n \in \mathbb{N} \).

Proof: Let \( B \) be the smallest subset of \( \mathbb{N} \times X \) containing \( (1, x_1) \), with the property that, whenever \( (n, x) \in B \), then \( (S(n), g(x)) \in B \). (\( B \) is the intersection of all subsets of \( \mathbb{N} \times X \) with these two properties; there is at least one such subset, \( \mathbb{N} \times X \) itself.)

We first show that for every \( n \in \mathbb{N} \), there is at least one \( x \in X \) with \( (n, x) \in B \). Set

\[ A = \{ n \in \mathbb{N} : \exists x \in X \text{ with } (n, x) \in B \} . \]

Then \( 1 \in A \) since \( (1, x_1) \in B \). If \( n \in A \), and \( (n, x) \in B \), then \( (S(n), g(x)) \in B \) so \( S(n) \in A \). Thus \( A = \mathbb{N} \).

Next we show that \( B \) is a function, i.e. if \( (n, x) \) and \( (n, y) \) are in \( B \), then \( x = y \). Let

\[ A' = \{ n \in \mathbb{N} : (n, x) \in B \text{, } (n, y) \in B \text{ imply } x = y \} . \]

First we show \( 1 \in A' \). Set

\[ B' = \{(1, x_1)\} \cup \{(n, x) \in B : n \neq 1\} \]

251
and note that $B' \subseteq B$, $(1, x_1) \in B'$, and $(n, x) \in B'$ implies $(S(n), g(x)) \in B'$; hence $B' = B$ and $1 \in A'$. Now suppose $n \in A'$, i.e. there is a unique $x_n \in X$ with $(n, x_n) \in B$. Set

$$B'' = \{(S(n), g(x_n)) \cup \{(m, x) \in B : m \neq S(n)\}$$

and note that $B'' \subseteq B$, $(1, x_1) \in B''$, and $(m, x) \in B''$ implies $(S(m), g(x)) \in B''$ [if $m \neq n$ we have $S(m) \neq S(n)$ so the result is obvious; if $m = n$ then $x = x_n$ so the result holds too]. Thus $B'' = B$ and $S(n) \in A'$. So $A' = \mathbb{N}$ and $B$ is a function, i.e. for each $n \in \mathbb{N}$ we can set $f(n) = x_n$, where $x_n$ is the unique element of $X$ such that $(n, x_n) \in B$.

We need to show that $f$ is unique. If $f'$ is another function from $\mathbb{N}$ to $X$ with $f'(1) = x_1$ and $f'(S(n)) = g(f'(n))$ for all $n \in \mathbb{N}$, set

$$A'' = \{n \in \mathbb{N} : f(n) = f'(n)\}.$$

We have $1 \in A''$ since $f(1) = f'(1) = x_1$. Suppose $n \in A''$. Then

$$f(S(n)) = g(f(n)) = g(f'(n)) = f'(S(n))$$

so $S(n) \in A''$, and thus $A'' = \mathbb{N}$, $f = f'$.

A version of this theorem is used in defining algebraic operations on $\mathbb{N}$:

**III.2.3.4. Corollary.** Let $g$ be a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$, and $h$ a function from $\mathbb{N}$ to $\mathbb{N}$. Then there is a unique function $f$ from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ satisfying $f(m, 1) = h(m)$ and $f(m, S(n)) = g(m, f(m, n))$ for all $m, n \in \mathbb{N}$.

**Proof:** Fix $m \in \mathbb{N}$, and let $g_m$ be the function from $\mathbb{N}$ to $\mathbb{N}$ given by $g_m(n) = g(m, n)$. Define a function $f_m : \mathbb{N} \rightarrow \mathbb{N}$ recursively by $f_m(1) = h(m)$ and $f_m(S(n)) = g_m(f_m(n))$. Set $f(m, n) = f_m(n)$ for each $n$. This gives a well-defined function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, and we have $f(m, 1) = f_m(1) = h(m)$ and

$$f(m, S(n)) = f_m(S(n)) = g_m(f_m(n)) = g(m, f(m, n))$$

for $m, n \in \mathbb{N}$. Uniqueness is also obvious from the uniqueness of the $f_m$.

**III.2.4. Algebraic Structure of $\mathbb{N}$**

There is no explicit mention of algebraic operations on $\mathbb{N}$ in the Peano axioms. But addition and multiplication can be defined inductively in terms of the successor function:

**III.2.4.1. Definition.** Let $(\mathbb{N}, 1, S)$ be the natural numbers (or, technically, a model for the natural numbers). Define a function $+ : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ inductively (recursively) as follows:

(a) $+(m, 1) = S(m)$.

(b) If $m \in \mathbb{N}$, then $+(m, S(n)) = S(+(m, n))$.

This is done from III.2.3.4. by setting $h(m) = S(m)$ and $g(m, n) = S(n)$ for all $m, n$.

Then define a function $\times : \mathbb{N} \rightarrow \mathbb{N}$ by

252
(c) If $m \in \mathbb{N}$, then $\times (m, 1) = m$.

(d) If $m \in \mathbb{N}$, then $\times (m, S(n)) = +(m, \times (m, n))$ for all $y \in \mathbb{N}$.

This is done from III.2.4. by setting $h(m) = m$ and $g(m, n) = +(m, n)$ for all $m, n$.

We normally write $m + n$ for $+(m, n)$, and $mn$ instead of $\times (m, n)$. We will suppress the notations $+(m, n)$ and $\times (m, n)$ from now on, and use $m + n$ and $mn$ instead. We also adopt the standard convention that multiplication is grouped before addition, i.e. that $mn + r$ means $(mn) + r$, etc.

III.2.4.2. Proposition. The functions $+$ and $\times$ are well defined on all of $\mathbb{N} \times \mathbb{N}$.

PROOF: This is an immediate consequence of III.2.3.4. \hfill \Box

We can then prove the associative and commutative laws for addition.

III.2.4.3. Theorem. The binary operation $+$ is associative and commutative on $\mathbb{N}$, i.e.

(i) [Associative Law] We have $(x + y) + z = x + (y + z)$ for all $x, y, z \in \mathbb{N}$.

(ii) [Commutative Law] We have $x + y = y + x$ for all $x, y \in \mathbb{N}$.

PROOF: (i): Fix $m, n \in \mathbb{N}$. Let $A_{m,n}$ be the set of all $r \in \mathbb{N}$ such that $(m + n) + r = m + (n + r)$. By (a) and (b),

$$(m + n) + 1 = S(m + n) = m + S(n) = m + (n + 1)$$

so $1 \in A_{m,n}$. Now suppose $r \in A_{m,n}$. We have, by (b),

$$(m + n) + S(r) = S((m + n) + r) = S(m + (n + r)) = m + S(n + r) = m + (n + S(r))$$

and thus $S(r) \in A_{m,n}$. So $A_{m,n} = \mathbb{N}$, i.e. $(m + n) + r = m + (n + r)$ for all $r$. Since $m$ and $n$ are arbitrary, (i) is proved.

(ii): Let $C$ be the set of $n \in \mathbb{N}$ such that $n + 1 = 1 + n$. Trivially $1 \in C$. If $n \in C$, then by (a) and (b)

$$1 + S(n) = S(1 + n) = S(n + 1) = S(n) + 1$$

so $S(n) \in C$. Thus $C = \mathbb{N}$. Now fix $m \in \mathbb{N}$, and let $D_m$ be the set of $n \in \mathbb{N}$ such that $m + n = n + m$. We have $1 \in D_m$ by the first part of the proof. Now suppose $n \in D_m$. Then

$$m + S(n) = S(m + n) = (m + n) + 1 = 1 + (m + n) = 1 + (n + m) = (1 + n) + m = (n + 1) + m = S(n) + m$$

by (i) and the first part of the proof, so $S(n) \in D_m$. Thus $D_m = \mathbb{N}$, i.e. $m + n = n + m$ for all $n$. Since $m$ is arbitrary, this is true for all $m, n \in \mathbb{N}$ \hfill \Box

We now show that multiplication is associative, commutative, and distributive over addition.
III.2.4.4. Theorem.

(i) [Associative Law] We have \((mn)r = m(nr)\) for all \(m, n, r \in \mathbb{N}\).

(ii) [Commutative Law] We have \(mn = nm\) for all \(m, n \in \mathbb{N}\).

(iii) [Distributive Law] We have \(m(n + r) = mn + mr\) (and \((m + n)r = mr + nr\)) for all \(m, n, r \in \mathbb{N}\).

Proof: The proof is similar in form to the previous ones. We first prove the first (left) distributive law (iii).

First note that if \(m, n \in \mathbb{N}\), we have, by (c), (d), and commutativity of addition,

\[ m(n + 1) = mS(n) = m + mn = mn + m = mn + m1. \]

Fix \(m, n \in \mathbb{N}\), and let \(A_{m,n}\) be the set of all \(r \in \mathbb{N}\) such that \(m(n + r) = mn + mr\). We have \(1 \in A_{m,n}\) by the first part of the proof. Now let \(r \in A_{m,n}\). Then

\[ m(n + S(r)) = mS(n + r) = m + m(n + r) = m + (mn + mr) \]

\[ = (mn + mr) + m = mn + (mr + m1) = mn + m(r + 1) = mn + mS(n) \]

by (c), (d), associativity and commutativity of addition, and the first part of the proof applied to \(m\) and \(r\).

Thus \(A_{m,n} = \mathbb{N}\) and the first distributive law is proved.

The right distributive law \((m + n)r = mr + nr\) for all \(m, n, r \in \mathbb{N}\) will follow from the left distributive law and commutativity of multiplication. But we will need a special case of the right distributive law to prove commutativity of multiplication: \((1 + n)r = 1r + nr\) for all \(n, r \in \mathbb{N}\). To prove this, fix \(n \in \mathbb{N}\) and let \(B_n\) be the set of all \(r \in \mathbb{N}\) such that the formula holds.

\[ (1 + n)1 = 1 + n = 11 + n1 \]

by (c), \(1 \in B_n\). If \(r \in B_n\), then

\[ (1 + n)S(r) = (1 + n) + (1 + n)r = (1 + n) + (1r + nr) = (1r + 1) + (n + nr) \]

\[ = (1r + 1) + (n + nr) = 1(r + 1) + nS(r) = 1S(r) + nS(r) \]

by (d), the associative and commutative laws of addition, and the left distributive law. Thus \(S(r) \in B_n\), \(B_n = \mathbb{N}\), and \((1 + n)r = 1r + nr\) for all \(n, r \in \mathbb{N}\).

(i): To show that \(\times\) is associative, fix \(m, n \in \mathbb{N}\) and let \(C_{m,n}\) be the set of all \(r \in \mathbb{N}\) such that \((mn)r = m(nr)\). Since \((mn)1 = mn = m(n1)\) by (c), \(1 \in B_{m,n}\). If \(r \in A_{m,n}\), then

\[ (mn)S(r) = mn + (mn)r = mn + m(nr) = m(n + nr) = m(nS(r)) \]

by (d) and the (left) distributive law. So \(S(r) \in C_{m,n}\). Thus \(C_{m,n} = \mathbb{N}\) and the associative law is proved.

(ii): For the commutative law, let \(D\) be the set of \(n\) for which \(1n = n\). We have \(1 \in D\) by (c). If \(n \in D\), then

\[ 1S(n) = 1 + 1n = 1 + n = n + 1 = S(n) \]

by (d), (a), and commutativity of addition. So \(S(n) \in D\) and \(D = \mathbb{N}\), i.e. \(1n = n\) for all \(n \in \mathbb{N}\).

Now fix \(m \in \mathbb{N}\), and let \(E_m\) be the set of all \(n \in \mathbb{N}\) such that \(mn = nm\). By (c) and the previous argument, \(1 \in E_m\). If \(n \in E_m\), then

\[ mS(n) = m + mn = m + nm = 1m + nm = (1 + n)m = (n + 1)m = S(n)m \]

254
by (d), the previous calculation, and the special case of the right distributive law. Thus $E_m = \mathbb{N}$ and the commutative law is proved.

Among the many algebraic properties of $\mathbb{N}$ which can be proved from the Peano axioms, here are two which are particularly important:

### III.2.4.5. Proposition. There do not exist $m, n \in \mathbb{N}$ with $m + n = m$.

**Proof:** Let $A$ be the set of all $m \in \mathbb{N}$ such that there is no $n \in \mathbb{N}$ with $m + n = n$. We have $1 \in A$, since $1 + n = 1 = S(n)$, contradicting axiom (P3). Let $m \in A$, and suppose there is $n \in \mathbb{N}$ with $S(m) + n = S(m)$. Then

\[ S(m + n) = S(n + m) = n + S(m) = S(m) + n = S(m) \]

by III.2.4.1.(b) and commutativity of addition. Thus $m + n = m$ by Axiom (P4), contradicting that $m \in A$. So $S(m) \in A$, $A = \mathbb{N}$.

### III.2.4.6. Proposition. [Cancellation] Let $m, n, r \in \mathbb{N}$. If $m + r = n + r$, then $m = n$.

**Proof:** The proof follows the usual scheme. Let $A$ be the set of all $r \in \mathbb{N}$ such that $m + r = n + r$ implies $m = n$ for all $m, n \in \mathbb{N}$. We have $1 \in A$ by axiom (P4). If $r \in A$, let $m, n \in \mathbb{N}$ with $m + S(r) = n + S(r)$. Then

\[ S(m + r) = m + S(r) = n + S(r) = S(n + r) \]

by III.2.4.1.(b). Thus $m + r = n + r$ by axiom (P4), hence $m = n$ since $r \in A$; so $S(r) \in A$ and $A = \mathbb{N}$.

There is also cancellation for multiplication, but it takes more work to prove (III.2.6.7.).

### III.2.5. Algebraic Axiomatization of $\mathbb{N}$

#### III.2.5.1. An alternate axiomatization of $\mathbb{N}$ can be given which is much closer to the way $\mathbb{N}$ is thought of and used. We have shown that $\mathbb{N}$ satisfies the following set of axioms:

- (N1) $\mathbb{N}$ contains an element 1.
- (N2) $\mathbb{N}$ has an associative binary operation +.
- (N3) There is no $n \in \mathbb{N}$ with $n + 1 = 1$.
- (N4) For any $m, n \in \mathbb{N}$, $m + 1 = n + 1$ implies $m = n$.
- (N5) If $A \subset \mathbb{N}$, $1 \in A$, and $A$ is closed under adding 1 (i.e. if $n \in A$ implies $n + 1 \in A$), then $A = \mathbb{N}$. 255
III.2.5.2. Theorem. Let \((N, e, +)\) be a model for axioms \((N1)\)–\((N5)\), i.e. \(e \in N\), + is a binary operation on \(N\), there is no \(x \in N\) with \(x + e = e\), \(x + e = y + e\) implies \(x = y\), and a subset \(A\) of \(N\) containing \(e\) and closed under adding \(e\) is all of \(N\). For \(x \in N\), define \(s(x) = x + e\). Then \((N, e, s)\) is a model for Peano arithmetic.

Proof: This is quite obvious, since axioms \((N3)\), \((N4)\), and \((N5)\) are simple rephrasings of axioms \((P3)\), \((P4)\), and \((P5)\) respectively.

If we start with a model \((N, e, S)\) for Peano arithmetic, define + on \(N\) as in III.2.4.1., and then define \(s\) as in III.2.5.2. from +, it is obvious that \(s\) coincides with \(S\). One can go the other way too:

III.2.5.3. Proposition. Let \((N, e, +)\) be a model for axioms \((N1)\)–\((N5)\). Define \(s\) as in III.2.5.2., and define +′ on \(N\) from \(s\) as in III.2.4.1.. Then +′ coincides with +.

Proof: We need only check that \(e + e = s(e)\), \(s(x) + e = s(x + e)\) for all \(x \in N\), and \(x + s(y) = s(x + y)\) for all \(x, y \in N\), since these formulas with + replaced by +′ characterize +′ by III.2.4.1.. We have \(e + e = s(e)\) by definition of \(s\). For any \(x \in N\), we have

\[
s(x) + e = (x + e) + e = s(x + e)
\]

also by definition of \(s\). Finally, if \(x, y \in N\), we have

\[
x + s(y) = x + (y + e) = (x + y) + e = s(x + y).
\]

Putting together III.2.5.3., III.2.5.2., and III.2.2.10., we obtain:

III.2.5.4. Corollary. If \((N, e, +)\) and \((N′, e′, +′)\) are models for \((N1)\)–\((N5)\), then there is a bijection \(\phi : N \to N′\) such that \(\phi(e) = e′\) and \(\phi(x + y) = \phi(x) +′ \phi(y)\) for all \(x, y \in N\).

Proof: Let \(\phi\) be as in III.2.2.10. for the successor operations \(s\) and \(s′\). Let \(A\) be the set of all \(y \in N\) such that \(\phi(x + y) = \phi(x) +′ \phi(y)\) for all \(x \in N\). We have that \(e \in A\) by III.2.2.10.. If \(y \in A\), then

\[
\phi(x + s(y)) = \phi(s(x + y)) = s′(\phi(x + y)) = s′(\phi(x) +′ \phi(y)) = \phi(x) +′ s′(\phi(y)) = \phi(x) +′ \phi(s(y))
\]

and so \(s(y) \in A\), \(A = N\).

III.2.5.5. Corollary. If \((N, e, +)\) is a model for axioms \((N1)\)–\((N5)\), then the binary operation + is commutative.
III.2.5.6. Remarks. (i) Note that multiplication does not explicitly enter into this algebraic characterization of \( \mathbb{N} \). This is because multiplication is completely determined by addition and property (d) of III.2.4.1. In fact, if \( * \) is an associative binary operation on \( \mathbb{N} \) which distributes over addition (i.e. \( x \ast (y + z) = (x \ast y) + (x \ast z) \) and \( (x + y) \ast z = (x \ast z) + (y \ast z) \)) and for which \( 1 \ast 1 = 1 \), then \( * \) must agree with multiplication since it must satisfy III.2.4.1.(a)–(c).

(ii) An examination of the proof of III.2.5.2. shows that associativity of \( + \) is not used. Thus, if we modify \( (\mathbb{N}^2) \) to \( (\mathbb{N}^2)' \) by deleting the word “associative,” and take \( (\mathbb{N}1)' \), \( (\mathbb{N}3)' \), \( (\mathbb{N}4)' \), \( (\mathbb{N}5)' \) to be identical to \( (\mathbb{N}1) \), \( (\mathbb{N}3) \), \( (\mathbb{N}4) \), \( (\mathbb{N}5) \) respectively, and if \( (\mathbb{N}, e, +) \) is a model for \( (\mathbb{N}1)'–(\mathbb{N}5)' \) with \( + \) nonassociative, we can still define \( s \) as in III.2.5.2. and obtain a model of Peano arithmetic. But the conclusion of III.2.5.3. will fail for this model, and hence the corollaries will also fail. There are such models with nonassociative \( + \): see Exercise (i).

III.2.5.7. All of the theorems of elementary number theory (e.g. unique factorization of natural numbers into products of prime numbers) can be proved directly from the Peano axioms. See books on number theory, e.g. (i).

III.2.6. Order Structure of \( \mathbb{N} \)

There is also a natural order structure on \( \mathbb{N} \). This order structure can either be transferred from the set containment order on \( \omega \) (i), or defined directly from the addition. We take the latter approach since it is more elementary.

III.2.6.1. Definition. Let \( m, n \in \mathbb{N} \). Set \( m < n \) if there is an \( r \in \mathbb{N} \) with \( m + r = n \).

III.2.6.2. Proposition. Let \( m, n, r \in \mathbb{N} \). If \( m < n \) and \( n < r \), then \( m < r \) (the relation \( < \) is transitive).

Proof: We have \( m + s = n \) and \( n + t = r \) for some \( s, t \in \mathbb{N} \). Then \( m + (s + t) = r \).

The next simple fact is essentially a rephrasing of III.2.2.3.:  

III.2.6.3. Proposition. For any \( n \in \mathbb{N}, n \neq 1 \), we have \( 1 < n \).

Proof: If \( n \in \mathbb{N}, n \neq 1 \), then by III.2.2.3. \( n = S(m) \) for some \( m \in \mathbb{N}, \) so \( 1 + m = n \) and \( 1 < n \).

The next result is the basic fact about <:
III.2.6.4.  **Theorem.** If \( m, n \in \mathbb{N} \), then exactly one of the following is true: \( m < n \), \( m = n \), \( n < m \).

**Proof:** If \( m = n \), we cannot have \( m < n \) by III.2.4.5. Similarly, if \( m < n \) and \( n < m \), we have \( m + r = n \) and \( n + s = m \) for some \( r, s \in \mathbb{N} \); then \( m + (r + s) = m \), contradicting III.2.4.5. Thus, for any \( m \) and \( n \), at most one of \( m < n \), \( m = n \), \( n < m \) is true.

To show that at least one is true, let \( A \) be the set of \( m \in \mathbb{N} \) such that, for every \( n \in \mathbb{N} \), either \( m < n \), \( m = n \), or \( n < m \) is true. We have \( 1 \in A \) by III.2.6.3.. If \( m \in A \), and \( n \in \mathbb{N} \), then either \( n < m \), \( n = m \), or \( m < n \). If \( n = m \), then \( S(m) = n + 1 \) so \( n < S(m) \). Similarly, if \( n < m \), i.e. \( n + r = m \) for some \( r \in \mathbb{N} \), then \( S(m) = n + (r + 1) \), so \( n < S(m) \). If \( m < n \), then \( n = m + r \) for some \( r \in \mathbb{N} \). If \( r = 1 \), then \( n = S(m) \); otherwise \( r = 1 + s \) for some \( s \in \mathbb{N} \), in which case \( n = (m + 1) + s = S(m) + s \), so \( S(m) < n \). Thus \( S(m) \in A \), \( A = \mathbb{N} \).

III.2.6.5.  Combining III.2.6.4. and III.2.6.2., we obtain that the relation \( < \) is a **strict total order** \( () \) on \( \mathbb{N} \). As is customary, we will write \( m \leq n \) if \( m < n \) or \( m = n \). We will also write \( n > m \) if \( m < n \) and \( n \geq m \) if \( m \leq n \).

Another important property of \( < \) is translation-invariance:

III.2.6.6.  **Theorem.** Let \( m, n \in \mathbb{N} \).
(i) If \( m < n \), then \( m + r < n + r \) for all \( r \in \mathbb{N} \). If \( m + r < n + r \) for some \( r \in \mathbb{N} \), then \( m < n \).
(ii) If \( m < n \), then \( mr < nr \) for all \( r \in \mathbb{N} \). If \( mr < nr \) for some \( r \in \mathbb{N} \), then \( m < n \).

**Proof:** (i): If \( m < n \), i.e. there is an \( s \in \mathbb{N} \) with \( m + s = n \), then for any \( r \in \mathbb{N} \) we have \( (m + r) + s = (m + s) + r = n + r \), i.e. \( m + r < n + r \) for some \( s \in \mathbb{N} \). Conversely, suppose \( m + r < n + r \) for some \( r \), i.e. \( (m + r) + s = n + r \) for some \( s \in \mathbb{N} \). Then \( m + s = n \) by cancellation (III.2.4.6.), i.e. \( m < n \).

(ii): If \( m < n \), write \( n = m + s \) for some \( s \in \mathbb{N} \). Then for any \( r \in \mathbb{N} \) we have \( mr + sr = (m + s)r = nr \), i.e. \( mr < nr \). For the converse, we cannot use multiplicative cancellation which is not yet proved, so we need a different argument (this argument could also have been used for addition). Suppose \( mr < nr \) for some \( r \). We have either \( m < n \), \( m = n \), or \( m > n \). If \( m = n \), we have \( mr = nr \), contradicting III.2.6.4., and if \( n < m \) we would have \( nr < mr \) by the first part of the proof, again contradicting III.2.6.4.. Thus we must have \( m < n \).

As a corollary, we get cancellation for multiplication:

III.2.6.7.  **Corollary.** Let \( m, n, r \in \mathbb{N} \). If \( mr = nr \), then \( m = n \).

**Proof:** We have either \( m < n \), \( m = n \), or \( n < m \). If \( m < n \), we would have \( mr < nr \) by III.2.6.6., contradicting III.2.6.4.. Similarly, we cannot have \( n < m \).
III.2.6.8. Because of cancellation, we can replace $<$ by $\leq$ throughout III.2.6.6.

We also obtain:

III.2.6.9. Corollary. If $m, n \in \mathbb{N}$ and $mn = 1$, then $m = n = 1$.

Proof: Suppose $n \neq 1$. Then $n > 1$, so $mn \geq m1 = m \geq 1$, a contradiction. So $n = 1$, and hence $m = 1$ too.

There is one more important basic fact about the ordering on $\mathbb{N}$. If $m, n \in \mathbb{N}$, we say $m$ is a predecessor of $n$ if $m < n$.

III.2.6.10. Theorem. Every $n \in \mathbb{N}$ has only finitely many predecessors in $\mathbb{N}$. (To avoid circularity, we interpret this statement to mean that the set of predecessors of any $n \in \mathbb{N}$ is a finite set in the sense of II.5.2.6.)

Proof: Let $A$ be the set of all $n \in \mathbb{N}$ such that the set of predecessors of $n$ is finite. By III.2.6.3., the set of predecessors of $1$ is $\emptyset$, which is finite (), so $1 \in A$. Let $n \in A$. If $m$ is a predecessor of $S(n)$, then either $m$ is a predecessor of $n$ or $m = n$ [if $n < m$, i.e. $n + r = m$ for some $r$, we have $n + r < S(n) = n + 1$, so $r < 1$ by III.2.6.6., contradicting III.2.6.3.]. Thus, if $P$ is the set of predecessors of $n$, the set of predecessors of $S(n)$ is $P \cup \{n\}$, which is finite since $P$ is finite. Thus $S(n) \in A$, $A = \mathbb{N}$.

III.2.6.11. Corollary. If $A$ is a nonempty subset of $\mathbb{N}$, then $A$ has a smallest element ($\mathbb{N}$ is well-ordered ()).

Proof: Let $n$ be an element of $A$. If $n$ is the smallest element of $A$, we are done. Otherwise, the set of predecessors of $n$ which are in $A$ is a nonempty finite totally ordered set, hence has a smallest element (II.5.2.12.), which must be the smallest element of $A$ since $A$ is totally ordered.

III.2.7. Subtraction and Division in $\mathbb{N}$

III.2.7.1. Subtraction and division can be partially defined on $\mathbb{N}$. If $m, n \in \mathbb{N}$, we say $m - n$ is defined in $\mathbb{N}$ if there is an $r \in \mathbb{N}$ with $m = n + r$; if such a $r$ exists, it is unique by cancellation, and we set $m - n = r$.

We have that $m - n$ is defined in $\mathbb{N}$ if and only if $n < m$, and we have $m - n = (m + r) - (n + r)$ for any $r \in \mathbb{N}$. If $m, n, r, s \in \mathbb{N}$ and $m - n$ and $r - s$ are defined, then $(m + r) - (n + s)$ is defined, and $(m + r) - (n + s) = (m - n) + (r - s)$.

Similarly, if $m, n \in \mathbb{N}$, we say $m$ is divisible by $n$ in $\mathbb{N}$ if there is a (necessarily unique) $r \in \mathbb{N}$ with $m = nr$. We then say $r$ is the quotient of $m$ divided by $n$, written $r = \frac{m}{n}$. We have, for all $n \in \mathbb{N}$, that $m$ is divisible by 1 and $\frac{m}{n} = n$, and $n$ is divisible by $n$ and $\frac{n}{n} = 1$. If $m$ is divisible by $n$, we necessarily have $n \leq m$. If $m$ is divisible by $n$ and $r$ is divisible by $s$, then $m r$ is divisible by $n s$ and $\frac{m r}{n s} = \frac{m}{n} \frac{r}{s}$. In particular, if $m$ is divisible by $n$, then $m r$ is divisible by $m r$ for any $r \in \mathbb{N}$, and $\frac{m r}{m} = \frac{m}{n}$. 259
III.2.7.2. The fact that subtraction is only partially defined on \( \mathbb{N} \) leads to the expansion of \( \mathbb{N} \) to a larger number system \( \mathbb{Z} \) in which subtraction is everywhere defined. Similarly, the fact that division is only partially defined leads to the expansion of \( \mathbb{N} \) (or \( \mathbb{Z} \)) to the system of rational numbers \( \mathbb{Q} \). There is a choice of which order to do the extensions: in the following sections we will first expand to \( \mathbb{Z} \) to get subtraction defined everywhere, and then expand \( \mathbb{Z} \) to \( \mathbb{Q} \) to get division (except division by 0) everywhere defined. We could instead proceed by expanding \( \mathbb{N} \) to the positive rational numbers \( \mathbb{Q}_+ \) to get division defined everywhere, and then construct \( \mathbb{Q} \) from \( \mathbb{Q}_+ \) to make subtraction everywhere defined.

III.2.8. Axiomatization of \( \mathbb{N}_0 \)

III.2.8.1. We can also axiomatize \( \mathbb{N}_0 \). Actually, the Peano axioms (P1)–(P5) give an axiomatization of \( \mathbb{N}_0 \) if 1 is replaced by 0 throughout. But the addition (and multiplication) must be defined differently (Exercise ()). We can give an algebraic axiomatization of \( \mathbb{N}_0 \):

\[
\begin{align*}
(\mathbb{N}_01) & \quad \mathbb{N}_0 \text{ contains elements 0 and 1.} \\
(\mathbb{N}_02) & \quad \mathbb{N}_0 \text{ has an associative binary operation } +. \\
(\mathbb{N}_03) & \quad \text{If } n \in \mathbb{N}_0, \text{ then } n + 1 \neq 0. \\
(\mathbb{N}_04) & \quad m, n \in \mathbb{N}_0, \ n + 1 = n + 1 \text{ implies } m = n. \\
(\mathbb{N}_05) & \quad A \subset \mathbb{N}_0, \ 0 \in A, \text{ and } A \text{ is closed under adding } 1 \text{ (i.e. if } n \in A \text{ implies } n + 1 \in A), \text{ then } A = \mathbb{N}_0. \\
(\mathbb{N}_06) & \quad 0 \neq 1.
\end{align*}
\]

III.2.8.2. There is a model (in the obvious sense) for (\( \mathbb{N}_01 \))–(\( \mathbb{N}_06 \)): take \( \mathbb{N} \) (or, more precisely, any model for \( \mathbb{N} \)), let 0 be any object not in \( \mathbb{N} \), let \( \mathbb{N}_0 = \mathbb{N} \cup \{0\} \), and extend + from \( \mathbb{N} \) to \( \mathbb{N}_0 \) by setting \( 0 + x = x + 0 = 0 \) for all \( x \in \mathbb{N}_0 \). The axioms are easily checked.

Conversely, let \( (\mathbb{N}_0, 0, 1, +) \) be a model for (\( \mathbb{N}_01 \))–(\( \mathbb{N}_05 \)) (\( \mathbb{N}_06 \) is not needed here). If we define \( S(x) = x + 1 \) for \( x \in \mathbb{N}_0 \), then \( (\mathbb{N}_0, 0, S) \) is a model for Peano arithmetic.

III.2.8.3. Lemma. Let \( (\mathbb{N}_0, 0, 1, +) \) be a model for (\( \mathbb{N}_01 \))–(\( \mathbb{N}_06 \)). Then \( 0 + 1 = 1 \).

Proof: This proof is somewhat more complex than the others in this section. Define a sequence in \( \mathbb{N}_0 \) (function from \( \mathbb{N} \) to \( \mathbb{N}_0 \)) by setting \( x_1 = 0 + 1 \) and \( x_{S(n)} = x_n + 1 \) (we have avoided writing \( n + 1 \) for \( S(n) \) since in this proof + will denote the operation in \( \mathbb{N}_0 \)). We have that \( x_n \neq 0 \) for all \( n \) by (\( \mathbb{N}_03 \)). The set \( \{0\} \cup \{x_n : n \in \mathbb{N}\} \) contains 0 and is closed under adding 1, hence equals \( \mathbb{N}_0 \) by (\( \mathbb{N}_05 \)). Thus \( 1 = x_n \) for some \( n \in \mathbb{N} \) (since \( 1 \neq 0 \) by (\( \mathbb{N}_06 \))). Suppose \( n = S(m) \) for some \( m \in \mathbb{N} \). Then we have

\[
(0 + x_m) + 1 = 0 + (x_m + 1) = 0 + x_n = 0 + 1
\]

and hence \( 0 + x_m = 0 \) by (\( \mathbb{N}_04 \)). If \( m = S(r) \) for some \( r \in \mathbb{N} \), then

\[
0 = 0 + x_m = 0 + (x_r + 1) = (0 + x_r) + 1
\]

which contradicts (\( \mathbb{N}_03 \)); thus \( m = 1 \), i.e. \( 0 + x_1 = 0 \). But then

\[
(0 + 0) + 1 = 0 + (0 + 1) = 0 + x_1 = 0
\]

which again contradicts (\( \mathbb{N}_03 \)). So the assumption \( n = S(m) \) for some \( m \in \mathbb{N} \) is false, i.e. \( n = 1, 1 = x_1 = 0 + 1 \). \( \Box \)
III.2.8.4. We then get \(0 + 1 = 0 + (0 + 1) = (0 + 0) + 1\), so \(0 = 0 + 0\) by (N04). Then \((1 + 0) + 1 = 1 + (0 + 1) = 1 + 1\), so \(1 + 0 = 1\) by (N04). If \(A\) is the set of \(x \in N_0\) for which \(x + 0 = x\), we have \(0 \in A\), and if \(x \in A\), then \((x + 1) + 0 = x + (1 + 0) = x + 1\), so \(x + 1 \in A\); thus \(A = N_0\). Similarly, if \(B\) is the set of \(x \in N_0\) such that \(0 + x = x\), we have \(0 \in B\), and if \(x \in B\) then \(0 + (x + 1) = (0 + x) + 1 = x + 1\), so \(x + 1 \in B\); thus \(B = N_0\). So \(0 + x = x + 0 = x\) for all \(x \in N_0\).

It follows from (N03) by induction (in the usual way) that \(x + y \neq 0\) for any \(x, y \in N_0\), \(y \neq 0\). Thus \((N_0 \setminus \{0\}, 1, +)\) (we need \((N_01)\) to have \(1 \in N_0 \setminus \{0\}\)) is a model for (N1)–(N5), so this set may be identified with \(N\). From III.2.2.10. we then obtain:

III.2.8.5. Theorem. Let \((N, 0, 1, +)\) and \((N', 0', 1', +')\) be models for (N01)–(N06). Then there is a bijection \(\phi : N' \rightarrow N\) with \(\phi(0') = 0\), \(\phi(1') = 1\), and \(\phi(x +' y) = \phi(x) + \phi(y)\) for all \(x, y \in N'\).

III.2.8.6. So we may (again with slight abuse of language) talk about the structure \(N_0\), and regard \(N\) as a subset of \(N_0\). It follows that + is commutative on \(N_0\). We may also define multiplication on \(N_0\) by extending the multiplication on \(N\) by setting \(n0 = 0n = 0\) for all \(n \in N_0\). This definition is forced by the distributive law and cancellation: if \(n \in N\), then \(0 + n = n = 1n = (0 + 1)n = 0n + 1n = 0n + n\) for any \(n\), and the argument for \(n0\) is similar. It is easily checked that the extended multiplication is associative, commutative, and distributive over addition (\(\cdot\)), and \(mn = 0\) if and only if \(m = 0\) or \(n = 0\). We also have:

III.2.8.7. Proposition. Let \(m, n \in N_0\). Then there are unique \(r, s \in N_0\) such that \(m + r = n + s\) and at least one of \(r, s\) is 0.

III.2.8.8. We can extend the ordering on \(N\) to \(N_0\) by setting \(0 < n\) for all \(n \in N\). It is easily checked that III.2.6.4. and III.2.6.6.(i) continue to hold, and that III.2.6.6.(ii) holds with \(r \neq 0\).

III.2.8.9. It should be noted that we could have replaced axiom (N06) by

\[(N06')\quad 0 + 1 = 1.\]

Indeed, this axiom along with (N03) implies that \(0 \neq 1\). The proof of Theorem III.2.8.5. is streamlined using this axiom since Lemma III.2.8.3. is not needed. Note that either (N06) or (N06') (or some substitute) is necessary: \((N, 1, 1, +)\) is a model for (N01)–(N05).

III.2.9. Further Properties of \(N\)

Many more detailed properties of \(N\) can be proved. Much of the detailed algebraic theory of \(N\) is beyond the scope of an analysis text, but we will cover a few of the essentials here.

Complete Induction, Infinite Regress, and Recursion

III.2.9.1. First, now that we have the necessary algebraic and order structure on \(N\), we can formulate and justify the principles of Complete Induction, Infinite Regress, and Definition by Recursion as described in (). We will not repeat them here.
Divisibility and the Division Algorithm

If \( m, n, r \in \mathbb{N} \) and \( r \) divides both \( m \) and \( n \), then an application of the distributive law shows that \( r \) divides \( m + n \). There is a partial converse:

**III.2.9.2. Proposition.** Let \( m, n, r \in \mathbb{N} \). If \( r \) divides both \( m \) and \( m + n \), then \( r \) also divides \( n \).

**Proof:** Write \( m = ar \) and \( m + n = br \) for \( a, b \in \mathbb{N} \). We have \( ar = m < m + n = br \), so \( a < b \) by **III.2.6.6.(ii).** Thus \( b - a = c \) is defined in \( \mathbb{N} \). We have \( a + c = b \), so

\[
m + cr = ar + cr = br = m + n
\]

so \( cr = n \) by cancellation and \( r \) divides \( n \).

**III.2.9.3.** Now let \( n, d \in \mathbb{N} \). We want to “divide” \( n \) by \( d \) and get a “quotient” and a “remainder.” Since the numbers \( \{md : m \in \mathbb{N}\} \) are all distinct, only finitely many are less than or equal to \( n \), hence there is a largest such \( m \); call it \( q \). (There may not be any such \( m \), i.e. we could have \( d > n \); in this case set \( q = 0 \).) We then have \( q \in \mathbb{N}_0 \) and \( qd \leq n < (q + 1)d \). Then \( q \) is clearly the unique \( m \in \mathbb{N}_0 \) such that \( md \leq n < (m + 1)d \).

We call \( q \) the quotient of \( n \) by \( d \); if \( d \) divides \( n \) in the sense of **III.2.7.1.**, this definition of quotient agrees with the definition there.

We now define the remainder. If \( qd = n \), i.e. \( d \) divides \( n \), set \( r = 0 \). If \( qd < n \), then \( r = n - qd \) is defined in \( \mathbb{N} \). In either case we have \( n = qd + r \). We have \( 0 \leq r \) since \( r \in \mathbb{N}_0 \), and we have

\[
qd + r = n < (q + 1)d = qd + d
\]

so we obtain \( r < d \) by **III.2.6.6.(i).** Then \( r \) is called the remainder of \( n \) divided by \( d \). (Also, \( d \) is often called the divisor and \( n \) the dividend.)

We summarize the construction:

**III.2.9.4. Theorem.** [Division Algorithm] Let \( n, d \in \mathbb{N} \). Then there are unique \( q, r \in \mathbb{N}_0 \) such that \( 0 \leq r < d \) and \( n = qd + r \).

**Proof:** We only need to observe uniqueness. If we also have \( n = q'd + r' \) with \( 0 \leq r' < d \), then

\[
q'd \leq n = q'd + r' < q'd + d = (q' + 1)d
\]

and so \( q' = q \) by the uniqueness of \( q \) with this property, and hence \( r' = r \) by cancellation.
Prime Numbers and Factorization

III.2.9.5. A number $p \in \mathbb{N}$ is normally defined to be prime if $p > 1$ and $p$ cannot be written as a product $mn$ of natural numbers except by taking $m$ or $n$ to be 1.

Although this is the standard definition of a prime number, and we will use it, it should be noted that this definition is not the same as the usual definition in abstract algebra of a prime element of a ring (the abstract algebra definition of a prime element is one satisfying the conclusion of III.2.9.6.; an element which cannot be factored in a nontrivial way is called an irreducible element. In $\mathbb{N}$, or $\mathbb{Z}$ (or, more generally, a Unique Factorization Domain), irreducible elements and prime elements turn out to be the same thing, but this is false in general rings.)

By III.1.5.7., every natural number can be written as a product of prime numbers. We will eventually show (III.2.9.8.) that this product is unique.

III.2.9.6. Theorem. Let $p$ be a prime number, and $m, n \in \mathbb{N}$. If $p$ divides $mn$, then $p$ divides $m$ or $p$ divides $n$.

Proof: This can be proved somewhat more easily after $\mathbb{Z}$ is constructed (cf. III.1.6.12.), but we give a direct proof from the Peano axioms and the machinery we have already constructed from them. The proof is a multiple infinite regress.

Suppose there is a prime $p$ for which there are $m, n \in \mathbb{N}$ not divisible by $p$ for which $p$ divides $mn$. Let $p$ be the smallest such prime. For this $p$, let $m$ be the smallest natural number not divisible by $p$ for which there is an $n \in \mathbb{N}$, not divisible by $p$, such that $mn$ is divisible by $p$, and for this $m$ let $n$ be the smallest natural number satisfying these conditions.

Write $m = ap + r$, $n = bp + s$ with $a, b, r, s \in \mathbb{N}_0$ and $r, s < p$ by the Division Algorithm. We have $r \neq 0$ and $s \neq 0$ since $p$ does not divide $m$ or $n$; hence $r, s \in \mathbb{N}$. We have

$$mn = (ap + r)(bp + s) = abp^2 + aps + bpr + rs$$

and since $mn$ and $abp^2 + aps + bpr$ are divisible by $p$, we have that $rs$ is divisible by $p$ by (9). But $r$ and $s$ are not divisible by $p$ since they are less than $p$. Since $r \leq m$ and $s \leq n$, we have $r = m$ and thus $s = n$ by minimality of $m$ and $n$. Thus $m < p$ and $n < p$.

Write $mn = kp$ for some $k \in \mathbb{N}$. We have $k < p$ since $mn < p^2$. If $k = 1$, we have $p = mn$, contradicting that $p$ is prime. If $k > 1$, let $q$ be a prime number dividing $k$. Then $q < p$, and $q$ divides $mn$, so $q$ divides $m$ or $q$ divides $n$ by minimality of $p$. Suppose $q$ divides $m$, and write $m = qd$ and $k = qt$. Then $qdn = qtp$, so $dn = tp$ and $p$ divides $dn$. But $p$ does not divide $d$ since it does not divide $m$, and $d < m$, contradicting minimality of $m$. If $q$ divides $n$, a similar contradiction is obtained.

A simple induction proof then shows:

III.2.9.7. Corollary. Let $p$ be a prime number, and $n_1, \ldots, n_r \in \mathbb{N}$. If $p$ divides $n_1n_2\cdots n_r$, then $p$ divides at least one of the $n_k$.

We then obtain one of the most important facts about the natural numbers:

263
III.2.9.8. Theorem. [Fundamental Theorem of Arithmetic] Every natural number can be written as a (finite) product of prime numbers, uniquely up to the order of the factors. (The “empty product” 1 and products of length 1 are included. Repeated factors are, of course, allowed.)

Proof: It has already been shown (III.1.5.7.) that every natural number can be written as a product of prime numbers, so we only need to show uniqueness. This can be done by induction on the length of the product, but we phrase the argument alternately by infinite regress.

Suppose \( n \) is the smallest natural number for which there are two distinct factorizations

\[
 n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s .
\]

Then \( n > 1 \) by III.2.6.9., so both \( r \) and \( s \) are at least 1, and neither can be 1 since the \( p_j \) and \( q_k \) are prime. Thus \( r, s > 1 \). We have that \( p_1 \) divides \( q_1 \cdots q_s \), so it divides \( q_k \) for some \( k \), and since \( q_k \) is prime we have \( p_1 = q_k \). By reordering the \( q \)’s we may assume \( k = 1 \). Thus

\[
 \frac{n}{p_1} = p_2 \cdots p_r = q_2 \cdots q_s
\]

and \( \frac{n}{p_1} \) has two distinct factorizations. But \( \frac{n}{p_1} < n \), contradicting minimality of \( n \). Thus there is no such \( n \).

We conclude our brief introduction to the algebraic theory of \( \mathbb{N} \) by noting that there are an infinite number of primes. This observation and its slick proof are due to Euclid.

III.2.9.9. Theorem. There are infinitely many prime numbers. If \( n \in \mathbb{N} \), there is a prime number \( p \) with \( p > n \).

Proof: Suppose \( p_1, \ldots, p_r \) is a finite set of prime numbers. Set \( m = 1 + p_1 p_2 \cdots p_r \). Then no \( p_k \) divides \( m \), so if \( p \) is a prime dividing \( m \), \( p \) is not one of the \( p_k \) and no finite list of primes is complete.

The last statement follows from the first and the fact that the set of predecessors of \( n \) is finite. As an alternate proof, the number \( n! + 1 \) is not divisible by any number \( \leq n \) except 1, so any prime number dividing it is greater than \( n \).

III.2.10. Summary

III.2.10.1. The Peano axioms for the natural numbers, like all axiom systems in mathematics, cannot be proved. They are accepted as a set of basic assumptions mathematicians make about the natural numbers, for several reasons:

1. All of the axioms, with the arguable exception of P5, are statements which are “obviously” true about the natural numbers we learn about and use starting as children.

2. Assuming consistency of some appropriate form of set theory, e.g. ZF, it can be rigorously shown that there is a mathematical structure, and (up to an obvious renaming of elements) only one, which satisfies the axioms, and which furthermore has uniquely determined arithmetic operations, definable directly from the axioms, which capture all the properties and facts we would expect about the natural numbers and no obviously false ones.
All the standard properties of the natural numbers can be systematically and rigorously (if somewhat tediously) proved from the axioms.

From the resulting mathematical structure, a rigorous construction can be made of larger number systems, \( \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \) and \( \mathbb{C} \), which again have all the properties one would expect in these number systems. The constructed system \( \mathbb{R} \) also allows a rigorous development of analysis. (Analysis is complicated enough that it cannot be said that there are no “surprises”; however, there are no known contradictions or paradoxes resulting from the development from the Peano axioms, and even the surprises make sense and become unsurprising when thoroughly thought through.)

(1)–(3) have been discussed in this section, and part of (4) will be treated in succeeding sections. The part of (4) about analysis, however, will occupy the rest of the book (and more)!

III.2.11. Exercises

III.2.11.1. Show that the ordering on \( \mathbb{N} \) defined in III.2.6.1. coincides with the set containment order on \( \omega \) (using the standard identification of \( \mathbb{N} \) with \( \omega \)).

III.3. The Integers

III.3.1. Introduction and Axiomatization of \( \mathbb{Z} \)

III.3.1.1. The next set of numbers we will consider is the integers, or whole numbers (although historically positive fractions predated the concept of 0 or negative integers). The letter \( \mathbb{Z} \) (from the German Zahl, “number”) is customarily used to denote the set of integers. Thus, very informally,

\[
\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots \}
\]

As with \( \mathbb{N} \), this is too imprecise to be even an informal definition of \( \mathbb{Z} \).

III.3.1.2. There are several equivalent ways to axiomatize \( \mathbb{Z} \), and unlike with \( \mathbb{N} \) there is no generally accepted standard axiomatization. We will describe a few of the possibilities here. One efficient way is the following set of axioms:

(A) \( \mathbb{Z} \) has an associative binary operation \(+\).

(B) There is an element \( 0 \in \mathbb{Z} \) such that \( x + 0 = 0 + x = x \) for all \( x \in \mathbb{Z} \).

(C) For every \( x \in \mathbb{Z} \) there is an element \( -x \in \mathbb{Z} \) with \( x + (-x) = (-x) + x = 0 \).

(D) There is an element \( 1 \in \mathbb{Z} \) such that, if \( A \subseteq \mathbb{Z}, 1 \in A, \) and \( x \in A \) implies \( x + 1 \in A \) and \( (-x) \in A \), then \( A = \mathbb{Z} \).

(E) If \( P \) is the smallest subset of \( \mathbb{Z} \) containing 1, such that \( x \in P \) implies \( x + 1 \in P \), then \( 0 \notin P \).

A model for (A)–(E) is called a group. A group satisfying (D) is called a cyclic group, and a cyclic group satisfying (E) is a free cyclic group. There is a vast theory of groups, which is a large portion of the modern subject of abstract algebra.
III.3.2. Models for $\mathbb{Z}$

There is a standard model for $\mathbb{Z}$, which can be constructed from either $\mathbb{N}$ or $\mathbb{N}_0$. The construction is slightly simpler notationally starting from $\mathbb{N}_0$, so we will do it this way. The motivation for the construction is that every integer can be written as a difference of natural numbers. However, the representation is not unique, so it takes some work to handle the nonuniqueness: the idea is that $x_1 - y_1 = x_2 - y_2$ if and only if $x_1 + y_2 = x_2 + y_1$.

III.3.2.1. Definition. Define a relation on $\mathbb{N}_0 \times \mathbb{N}_0$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 + y_2 = x_2 + y_1$.

III.3.2.2. Proposition. (i) The relation $\sim$ is an equivalence relation ($\sim$) on $\mathbb{N}_0 \times \mathbb{N}_0$.

(ii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 + z_1, y_1 + w_1) \sim (x_2 + z_2, y_2 + w_2)$.

Proof: (i): If $(x_1, y_1) \sim (x_2, y_2)$, it is trivial that $(x_2, y_2) \sim (x_1, y_1)$. It is also trivial that $(x, y) \sim (x, y)$ for any $x, y$. If $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$, we have

$$(x_1 + y_3) + (x_2 + y_2) = (x_1 + y_2) + (x_2 + y_3) = (x_2 + y_1) + (x_3 + y_2) = (x_3 + y_1) + (x_2 + y_2)$$

by the associative and commutative laws, and hence $x_1 + y_3 = x_3 + y_1$ by cancellation. Thus $(x_1, y_1) \sim (x_3, y_3)$.

(ii): By the associative and commutative laws, we have

$$(x_1 + z_1) + (y_2 + w_2) = (x_1 + y_2) + (z_1 + w_2) = (x_2 + y_1) + (z_2 + w_1) = (x_2 + z_2) + (y_1 + w_1).$$

We will write $[(x, y)]$ for the equivalence class of the pair $(x, y)$. We will think of $[(x, y)]$ as representing the integer $x - y$; the equivalence relation $\sim$ is designed so that every representative of the equivalence class represents the same integer.

III.3.2.3. Definition. Let $\mathcal{Z}$ be the set of equivalence classes of ordered pairs. For $[(x, y)], [(z, w)] \in \mathcal{Z}$, define

$$[(x, y)] + [(z, w)] = [(x + z, y + w)].$$

Note that $+$ is well defined on $\mathcal{Z}$ by III.3.2.2.(ii).

III.3.2.4. Theorem. If $\mathcal{N}_0 = \{(x, 0) : x \in \mathbb{N}_0\}$, then $(\mathcal{N}_0, [(0, 0)], [(1, 0)], +)$ is a model for $\mathbb{N}_0$, and $(\mathcal{Z}, +, [(0, 0)], [(1, 0)])$ is a model for $\mathbb{Z}$. Addition in $\mathcal{Z}$ is commutative.

Proof: It is obvious that once we know that addition of equivalence classes is well defined, it inherits associativity (and also commutativity) from addition in $\mathbb{N}_0$. It is also obvious that $\mathcal{N}_0$ is closed under addition, and the map $\phi : \mathbb{N}_0 \to \mathcal{N}_0$ defined by $\phi(x) = [(x, 0)]$ satisfies $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in \mathbb{N}_0$. Also, $\phi(x) = \phi(y)$ if and only if $(x, 0) \sim (y, 0)$, i.e. if and only if $x = y$. Thus $\mathcal{N}_0$ is a model for $\mathbb{N}_0$, with 0-element $\phi(0) = [(0, 0)]$ and 1-element $\phi(1) = [(1, 0)]$.

We now check the $\mathbb{Z}$-axioms for $\mathcal{Z}$. (G1) is already checked. We have $[(0, 0)] + [(x, y)] = [(0 + x, 0 + y)] = [(x, y)]$ for all $x, y \in \mathbb{N}_0$, and similarly $[(x, y)] + [(0, 0)] = [(x, y)]$ for all $x, y$; thus (G2) holds. For $x, y \in \mathbb{N}_0$, 266
In the first case.

Proof:

(iii) If III.3.3.3. (G3). We will not prove this (cf. ()) since we will deduce commutativity of addition in another way.

There are noncommutative groups (). Commutativity of addition does follow from (Cyc), along with (G1)–(G2).

III.3.3.2. Here are some basic consequences of the axioms (G1)–(G3).

Z

We could have done the whole construction using N instead of N₀. In this case, the 0 element is [(1, 1)], and the 1-element is [(1 + 1, 1)]. The natural number n is represented by [(n + 1, 1)]. The approach of III.3.2.5. can also be done using N; details are left to the reader.

III.3.3.1. Here are some basic consequences of the axioms (G1)–(G3).

(i) [CANCELLATION] If x, y, z ∈ Z and x + z = y + z, then x = y: we have

\[ x = x + 0 = x + (z + (-z)) = (x + z) + (-z) = (y + z) + (-z) = y + (z + (-z)) = y + 0 = y. \]

Similarly, if z + x = z + y, then x = y. Since 0 + 0 = 0 = 0 + (−0) by (G2), we have −0 = 0.

(ii) If x, y ∈ Z and x + y = 0, then y = −x. In particular, −(−x) = x.

(iii) If x, y ∈ Z, then −(x + y) = −(y) + (−x): streamlining manipulations with the associative law, we have

\[ (x + y) + [(-y) + (-x)] = x + (y + (-y)) + (-x) = x + 0 + (-x) = x + (-x) = 0. \]

III.3.3.2. One property of Z which does not follow just from (G1)–(G3) is commutativity of addition: there are noncommutative groups (). Commutativity of addition does follow from (Cyc), along with (G1)–(G3). We will not prove this (cf. (i)) since we will deduce commutativity of addition in another way.

III.3.3.3. THEOREM. Let (Z, +, 0, 1) be a model for Z, and let P ⊆ Z be the set defined in (Fr). Then

(i) P is closed under addition, i.e. if x, y ∈ P, then x + y ∈ P.

(ii) (P, 1, +) is a model for N. Thus + is commutative on P.

(iii) If −P = {−x : x ∈ P}, then Z is the disjoint union of P, −P, and {0}.

Proof: (i): Let A be the set of all y ∈ P such that x + y ∈ P for all x ∈ P. Since x ∈ P implies x + 1 ∈ P, we have 1 ∈ A. If y ∈ A, then for any x ∈ P, we have x + (y + 1) = (x + y) + 1 ∈ P, so y + 1 ∈ A. So A is
a subset of \( Z \) containing 1 and closed under adding 1; thus \( P \subseteq A \) since \( P \) is the smallest such set. Since \( A \subseteq P \), we have \( A = P \).

(ii) Axiom (N1) holds by (i), and axiom (N2) is trivial. Since \( 0 + 1 = 1 \) and \( 0 \notin P \), we have \( x + 1 \neq 1 \) for all \( x \in P \) by cancellation, so axiom (N3) holds. Axiom (N4) holds by cancellation. Axiom (N5) also holds by an argument similar to the proof of (i). It follows from III.2.5.5. that \( + \) is commutative on \( P \).

(iii) We have \( \{0\} \cap P = \emptyset \) by (Fr), and similarly \( \{0\} \cap (\neg P) = \emptyset \) since \( -0 = 0 \). If \( x \in P \cap (\neg P) \), so \( x = -y \) for some \( y \in P \), then \( 0 = x + y \in P \) by (i), a contradiction; thus \( P \cap (\neg P) = \emptyset \). If \( x, y \in P \), then by (ii) and III.2.6.4. either \( x = y \), or there is a \( z \in P \) with \( x + z = y \), or there is a \( w \in P \) with \( x = y + w \). Now let \( A = P \cup (\neg P) \cup \{0\} \). We have \( 0 \in A \), and if \( x \in A \), then \( \neg x \in A \). If \( x \in A \), we show \( x + 1 \in A \). If \( x = 0 \) or \( x \in P \), this is obvious. If \( x \in (\neg P) \), then \( x = (\neg y) \) for some \( y \in P \). If \( y = 1 \), then \( x + 1 = 0 \in A \); if \( y = z + 1 = 1 + z \) for some \( z \in P \), then \( x = (\neg y) = (\neg z) + (\neg 1) \), so \( x + 1 = (\neg z) \in A \). Thus \( A = Z \) by (Cyc).

### III.3.4. Proposition. If \((Z, +, 0, 1)\) is a model for \( Z \), then \(+\) is commutative.

**Proof:** Addition is commutative on \( P \) by III.3.3.(ii), and we have \( 0 + x = x + 0 \) for any \( x \in Z \) by axiom (G2). It remains to show that \( x + (\neg y) = (\neg y) + x \) for \( x, y \in P \). Fix \( x, y \in P \). We have

\[
y + [x + (\neg y)] = (y + x) + (\neg y) = (x + y) + (\neg y) = x + [y + (\neg y)] = x = [y + (\neg y)] + x = y + [(\neg y) + x]
\]

and so \( x + (\neg y) = (\neg y) + x \) by cancellation. \( \diamondsuit \)

### III.3.5. Theorem. If \((Z, +, 0, 1)\) and \((Z', +', 0', 1')\) are models for \( Z \), then there is a bijection \( \phi : Z \to Z' \) such that \( \phi(0) = 0', \phi(1) = 1' \), and \( \phi(x + y) = \phi(x) +' \phi(y) \) for all \( x, y \in Z \).

**Proof:** Let \( P \) and \( P' \) be the subsets of \( Z \) and \( Z' \) respectively defined in (Fr). Since \( P \) and \( P' \) are models for Peano arithmetic, there is a bijection \( \phi : P \to P' \) with \( \phi(1) = 1' \) and \( \phi(x + y) = \phi(x) +' \phi(y) \) for all \( x, y \in P \) (III.2.5.4.). Extend \( \phi \) by setting \( \phi(0) = 0' \) and \( \phi(-x) = -\phi(x) \) for \( x \in P \). By III.3.3.3., \( \phi \) is well defined and bijective. We need only check that \( \phi \) preserves addition. Fix \( x, y \in P \). We know that \( \phi(x + y) = \phi(x) +' \phi(y) \). We also have

\[
\phi(-x) + (\neg y) = \phi(-y + x) = -\phi(y + x) = -((\neg y) +' \phi(x)) = [-\phi(x)] + [-\phi(y)].
\]

Finally, we either have \( x = y \), or \( x + z = y \) for some \( z \in P \), or \( x = y + w \) for some \( w \in P \). If \( x = y \), we have

\[
\phi(x + (\neg y)) = \phi(0) = \phi(x) +' [-\phi(y)] = \phi(x) +' \phi(-y).
\]

If \( x = y + w \), we have \( x + (-y) = w \) (using commutativity of addition). Then

\[
\phi(x) +' \phi(-y) = \phi(x) +' [-\phi(y)] = \phi(y + w) +' [-\phi(y)] = \phi(y) +' \phi(w) +' [-\phi(y)] = \phi(w) = \phi(x + (\neg y)).
\]

The argument if \( x + z = y \) is almost identical. \( \diamondsuit \)
III.3.3.6. So, as with \( \mathbb{N}_0 \) and \( \mathbb{N} \), we may talk about the set (or structure) \( \mathbb{Z} \) without defining it precisely as a set. We will identify \( \mathbb{N} \) with the subset \( P \), called the set of positive integers, and hence identify \( \mathbb{N}_0 \) with the set \( \mathbb{N} \cup \{0\} \) of nonnegative integers.

III.3.3.7. Note that the element 0 in \( \mathbb{Z} \) is uniquely determined by the algebraic structure: since \( x+0=x \) for any \( x \), by cancellation 0 is the unique \( x \in \mathbb{Z} \) with the property that \( x+x=x \). However, 1 is not completely determined by the algebraic (additive) structure of \( \mathbb{Z} \): if \( (\mathbb{Z}, +, 0, 1) \) is a model for \( \mathbb{Z} \), then \( (\mathbb{Z}, +, 0, -1) \) is also a model for \( \mathbb{Z} \).

In group-theoretic language, a free cyclic group has two generators. The axioms for \( \mathbb{Z} \) determine its algebraic (additive) structure: since any element \( x \) can be uniquely written in the form \( (x-x) + x \) where \( x-x \) is an integer, the rest of the structure then follows uniquely.

III.3.4. Ordering and Subtraction in \( \mathbb{Z} \)

III.3.4.1. We may also define a total ordering on \( \mathbb{Z} \) extending the order on \( \mathbb{N}_0 \) by setting \((-n) < 0 \) for \( n \in \mathbb{N} \), and \((-n) < (-m) \) if \( n, m \in \mathbb{N} \) and \( m < n \). It is easily checked that this is a total order and that \( x < y \) implies \( x+z < y+z \) for \( x, y, z \in \mathbb{Z} \). We will also use the symbols \( >, \leq, \geq \) with their usual (obvious) meaning.

III.3.4.2. We define the absolute value of an \( x \in \mathbb{Z} \) by \( |x| = x \) if \( x \geq 0 \) and \( |x| = -x \) if \( x < 0 \), i.e. \( |0| = 0 \) and \( |x| = |x| \) for \( x \in \mathbb{P} \). We obviously then have \( |x| \geq 0 \) and \( |x| = |x| \) for all \( x \in \mathbb{Z} \).

We also define the signum \( \text{sgn}(x) \) to be 1 if \( x \in \mathbb{P} \) and \(-1 \) if \( x \in \mathbb{N}_0 \); \( \text{sgn}(0) \) is not defined.

III.3.4.3. Proposition. [Triangle Inequality] If \( x, y \in \mathbb{Z} \), then \( |x| - |y| \leq |x+y| \leq |x| + |y| \).

The proof is a straightforward but tedious case-by-case check. The result is called the “triangle inequality” in analogy with the vector version (1).

III.3.4.4. We may define subtraction in \( \mathbb{Z} \) by \( x-y = x+(-y) \) for \( x, y \in \mathbb{Z} \). This operation extends the partially defined subtraction in \( \mathbb{N} \). Subtraction is a binary operation on \( \mathbb{Z} \), but is not associative or commutative. We have \( (x+z) - (y+w) = (x-y) + (z-w) \) (and in particular \( x-y = (x+z) - (y+z) \)) and \( y-x = -(x-y) \) for \( x, y, z, w \in \mathbb{Z} \). We then have \( |x-y| = |y-x| \) for all \( x, y \in \mathbb{Z} \).

III.3.5. Multiplication and Division in \( \mathbb{Z} \)

III.3.5.1. We now discuss multiplication in \( \mathbb{Z} \). We define multiplication on \( \mathbb{N}_0 \) as in III.2.8.6. There is only one way to extend this multiplication to \( \mathbb{Z} \) so that the distributive laws hold: since

\[
0 \cdot 0 = 0 = 0(x+(-x)) = 0x + 0(-x) = 0 + 0(-x) = 0(-x)
\]

for \( x \in \mathbb{N}_0 \), we must have \( 0y = 0 \) for all \( y \in \mathbb{Z} \). Similarly, \( y0 = 0 \) for all \( y \in \mathbb{Z} \). Then, if \( x, y \in \mathbb{Z} \),

\[
0 = x0 = x(y + (-y)) = xy + x(-y)
\]

so we must have \( x(-y) = -(xy) \) for all \( x, y \). Similarly, \( (-x)y = -(xy) \) for all \( x, y \in \mathbb{Z} \). Thus the following definition of multiplication is forced by the distributive laws:
III.3.5.2. **Definition.** If \( x, y \in \mathbb{N} \), define \( x(-y) = (-x)y = -(xy) \), \( (-x)(-y) = xy \), and \( x0 = 0x = 0 \).

It is obvious that this gives a well-defined binary operation on \( \mathbb{Z} \). Despite the use of the distributive law in motivating the definition of multiplication, the associative, commutative, and distributive laws must still be checked. The commutative law is obvious, and the proof of the distributive law is essentially the reversal of the argument in III.3.5.1. The associative law is a simple but tedious calculation checking case-by-case, using the associative law in \( \mathbb{N} \). Thus we obtain:

III.3.5.3. **Theorem.** Multiplication in \( \mathbb{Z} \) is associative, commutative, and distributes over addition, and is uniquely determined by these properties and the requirement that \( 11 = 1 \).

The last statement follows from III.3.5.1 and III.2.5.6(i).

The next results are easy consequences of the definitions, and the proofs are left to the reader.

III.3.5.4. **Proposition.** If \( x, y, z \in \mathbb{Z} \), \( z \neq 0 \), and \( xz = yz \), then \( x = y \). If \( xy = 0 \), then \( x = 0 \) or \( y = 0 \).

III.3.5.5. **Proposition.** If \( x, y \in \mathbb{Z} \), then \( |xy| = |x||y| \). If \( x, y \neq 0 \), then \( \text{sgn}(xy) = \text{sgn}(x)\text{sgn}(y) \).

III.3.5.6. **Proposition.** Let \( x, y, z \in \mathbb{Z} \) with \( x < y \). If \( z > 0 \), then \( xz < yz \). If \( z < 0 \), then \( xz > yz \).

III.3.5.7. We may define division in \( \mathbb{Z} \) as in \( \mathbb{N} \): if \( x, y \in \mathbb{Z} \), \( y \neq 0 \), we say \( x \) is divisible by \( y \) in \( \mathbb{Z} \) if there is a (necessarily unique) \( z \in \mathbb{Z} \) with \( x = yz \). We then say \( z \) is the quotient of \( x \) divided by \( y \), written \( z = \frac{x}{y} \). If \( x, y \in \mathbb{N} \), this notion coincides with divisibility and quotient in \( \mathbb{N} \). We have, for all \( x \in \mathbb{Z} \), that \( x \) is divisible by 1 and \( \frac{x}{1} = x \), and if \( x \neq 0 \), then \( x \) is divisible by \( x \) and \( \frac{x}{x} = 1 \). If \( x \) is divisible by \( y \) and \( z \) is divisible by \( w \), then \( xz \) is divisible by \( yw \) and \( \frac{xz}{yw} = \frac{x}{y} \frac{z}{w} \). In particular, if \( x \) is divisible by \( y \), then \( xz \) is divisible by \( yz \) for any \( z \in \mathbb{Z} \), \( z \neq 0 \), and \( \frac{xz}{yz} = \frac{x}{y} \). If \( x \neq 0 \), then no meaning can be attached to \( \frac{x}{0} \) since there is no \( z \) with \( z0 = x \). We also can give no usable definition of \( \frac{0}{v} \).

Quite apart from the impossibility of dividing by 0, which is insurmountable, division is still only partially defined on \( \mathbb{Z} \). It is thus desirable to expand \( \mathbb{Z} \) to a larger number system \( \mathbb{Q} \) in which division (except division by 0) is everywhere defined.

III.3.6. **Ring Axiomatization of \( \mathbb{Z} \)**

We give here an alternate axiomatization of \( \mathbb{Z} \). This axiomatization is not as “efficient” as the previous one, since it postulates existence of multiplication and its properties which can be deduced from the previous axioms, and also postulates commutativity of addition. However, this axiom scheme is commonly used, and natural subsets of these axioms define structures which are important in abstract algebra.
III.3.6.1. Ring Axioms for $\mathbb{Z}$:

(Op) There are two binary operations $+$ and $\cdot$ on $\mathbb{Z}$.

(A1) The binary operation $+$ is associative.

(A2) The binary operation $+$ is commutative.

(A3) There is an element $0 \in \mathbb{Z}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Z}$.

(A4) For every $x \in \mathbb{Z}$ there is an element $(-x) \in \mathbb{Z}$ with $x + (-x) = (-x) + x = 0$.

(M1) The binary operation $\cdot$ is associative.

(M2) The binary operation $\cdot$ is commutative.

(M3) There is an element $1 \in \mathbb{Z}$ such that $x \cdot 1 = 1 \cdot x = x$ for all $x \in \mathbb{Z}$.

(D) The operation $\cdot$ distributes over addition.

(Cyc) If $A \subseteq \mathbb{Z}$, $1 \in A$, and $x \in A$ implies $x + 1 \in A$ and $(-x) \in A$, then $A = \mathbb{Z}$.

(Fr) If $P$ is the smallest subset of $\mathbb{Z}$ containing $1$, such that $x \in P$ implies $x + 1 \in P$, then $0 \notin P$.

We have seen that a set $Z$ with a binary operation $+$ satisfying (A1), (A3), (A4), and with an element $1$ satisfying (Cyc) and (Fr) is a model for $\mathbb{Z}$, hence also satisfies (A2), and there is a unique multiplication on $Z$ for which the rest of the axioms are satisfied. Thus from III.3.3.5. we obtain:

III.3.6.2. Theorem. Let $(\mathbb{Z}, +, \cdot, 0, 1)$ be a model for the ring axioms of $\mathbb{Z}$. then there is a bijection $\phi : \mathbb{Z} \to \mathbb{Z}$ with $\phi(0) = 0, \phi(1) = 1, \phi(x + y) = \phi(x) + \phi(y)$, and $\phi(x \cdot y) = \phi(x)\phi(y)$, where $\mathbb{Z}$ has its previously defined addition, multiplication, 0, and 1.

III.3.6.3. Definition. A set $R$ with two binary operations $+$ and $\cdot$ satisfying (A1)–(A4) [with $\mathbb{Z}$ replaced by $R$ in the statements], (M1), and (D) is called a ring. A ring is commutative if it also satisfies (M2). A ring is unital if it also satisfies (M3) (an explicit assumption is usually made that $1 \neq 0$ in a unital ring; it is unnecessary to explicitly include this in our axiomatization of $\mathbb{Z}$ since it is an automatic consequence of (Fr)).

Several examples of rings are given in the Exercises, including commutative nonunital rings, noncommutative unital rings, and noncommutative nonunital rings. Note that the term commutative, when applied to a ring, refers to multiplication; addition is always commutative in a ring. Ring theory is an extensive and important part of abstract algebra, with many applications.

We will not develop the general theory of rings here, but we need one basic proposition:

III.3.6.4. Proposition. Let $R$ be a ring. Then

(i) $R$ has cancellation for addition.

(ii) If $x \in R$, then $x0 = 0x = 0$.

(iii) If $R$ is unital and contains more than one element, then $1 \neq 0$.

Proof: The proof of (i) is the same as for $\mathbb{Z}$ (III.3.3.1.(i)), which uses only (G1)–(G3), which are the same as (A1), (A3), (A4). (Actually cancellation holds in any group by this proof.)

For (ii), for any $x \in R$ we have $0 + 0x = 0x = (0 + 0)x = 0x + 0x$, so $0x = 0$. Similarly, $x0 = 0$.

(iii) follows immediately from (ii): if $x \in R$, $x \neq 0$, then $0x = 0 \neq x = 1x$. 🌟
III.3.6.5. The axiom (Cyc) is often replaced by the following axiom:

(Pos) There is a nonempty subset $P'$ of $\mathbb{Z}$, closed under addition and multiplication, such that $\mathbb{Z}$ is the disjoint union of $P'$, $\{0\}$, and $-P' = \{-x : x \in P'\}$.

It is obvious that (Pos) follows from the ring axioms for $\mathbb{Z}$ (taking $P' = P$). To show that (Pos) $\Rightarrow$ (Cyc) in the presence of the other axioms, we first need:

III.3.6.6. Proposition. Let $(\mathbb{Z}, +, \cdot, 0, 1)$ be a model for the ring axioms of $\mathbb{Z}$ with (Cyc) replaced by (Pos). Then $1 \in P'$.

Proof: Since $\mathbb{Z}$ is a unital ring and $P' \neq \emptyset$, we have $0 \neq 1$ in $\mathbb{Z}$ by III.3.6.4.(iii). Thus either $1 \in P'$ or $(-1) \in P'$. If $(-1) \in P'$, then $1 = (-1)(-1) \in P'$. (In fact, $(-1) \notin P'$ since then $1 \in P' \cap (-P')$.)

Once we know that $1 \in P'$, it follows from the proof of III.3.3.3. that $P = P'$ and that (Cyc) holds. So we get that the following set of axioms characterize $\mathbb{Z}$:

III.3.6.7. Ordered Ring Axioms for $\mathbb{Z}$:

(Op) There are two binary operations $+$ and $\cdot$ on $\mathbb{Z}$.

(A1) The binary operation $+$ is associative.

(A2) The binary operation $+$ is commutative.

(A3) There is an element $0 \in \mathbb{Z}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Z}$.

(A4) For every $x \in \mathbb{Z}$ there is an element $(-x) \in \mathbb{Z}$ with $x + (-x) = (-x) + x = 0$.

(M1) The binary operation $\cdot$ is associative.

(M2) The binary operation $\cdot$ is commutative.

(M3) There is an element $1 \in \mathbb{Z}$ such that $x \cdot 1 = 1 \cdot x = x$ for all $x \in \mathbb{Z}$.

(D) The operation $\cdot$ distributes over addition.

(Pos) There is a nonempty subset $P$ of $\mathbb{Z}$ such that

(P1) If $x, y \in P$, then $x + y \in P$.

(P2) If $x, y \in P$, then $xy \in P$.

(P3) $0 \notin P$.

(P4) If $x \in \mathbb{Z}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$, but not both.

(Min$_\mathbb{Z}$) $P$ is the smallest subset $A$ of $\mathbb{Z}$ containing $1$, such that $x \in A$ implies $x + 1 \in A$.

III.3.6.8. The ordering on $\mathbb{Z}$ can be cleanly expressed in terms of these axioms: $x < y$ if and only if $(y - x) \in P$. The specification of the ordering via (Pos) can be done alternately by the axiom

(Ord) There is a relation (strict total order) $<$ on $\mathbb{Z}$, such that

(O1) If $x, y \in \mathbb{Z}$, then exactly one of $x < y$, $y < x$, $x = y$ holds. (Trichotomy)

(O2) If $x, y, z \in \mathbb{Q}$, $x < y$, and $y < z$, then $x < z$. (Transitivity)
(O3) If \( x, y, z \in \mathbb{Z} \) and \( x < y \), then \( x + z < y + z \). (Translation Invariance)

(O4) If \( x, y, z \in \mathbb{Z} \), \( x < y \), and \( 0 < z \), then \( xz < yz \). (Homogeneity or Scaling Invariance)

It is easily seen that in the presence of the other axioms \([\text{Min}_2]\) is not needed, \((\text{Pos})\) is equivalent to \((\text{Ord})\); if \((\text{Ord})\) is assumed, \( P \) is defined to be \( \{x \in \mathbb{Z} : 0 < x\} \). It is a matter of taste whether the ordering is axiomatized using \((\text{Pos})\) or \((\text{Ord})\); both are used in the literature.

A commutative ring satisfying \((\text{Pos})\) (or, equivalently, \((\text{Ord})\)) is called an ordered ring.

III.3.6.9. REMARK. The axioms \((\text{P}3)\) and \((\text{P}4)\) can be replaced by

\[(\text{P}3')\] If \( x \in P \), then \(-x \notin P\).

\[(\text{P}4')\] If \( x \in \mathbb{Z}, x \neq 0 \), then either \( x \in P \) or \(-x \in P\).

It is obvious that \((\text{P}3)\) and \((\text{P}4)\) imply \((\text{P}3')\) and \((\text{P}4')\), and that \((\text{P}3')\) and \((\text{P}4')\) imply \((\text{P}4)\). Note also that if \((\text{P}3')\) is assumed, then we cannot have \( 0 \in P \): if \( 0 \in P \), then \((-0) \notin P\) by \((\text{P}3')\), but \(-0 = 0\), a contradiction. Thus \((\text{P}3')\) implies \((\text{P}3)\).

III.3.7. Exercises

III.3.7.1. Show that the set of even integers, with its usual addition and multiplication, is a nonunital commutative ring.

III.3.7.2. Fix \( n \in \mathbb{N} \), and let \( M_n(\mathbb{R}) \) be the set of all \( n \times n \) matrices with real entries. Give \( M_n(\mathbb{R}) \) the usual matrix addition and multiplication. Show that \( M_n(\mathbb{R}) \), with this addition and multiplication, is a unital ring, which is not commutative if \( n > 1 \). Do the same for \( M_n(\mathbb{Z}), M_n(\mathbb{Q}), \) and \( M_n(\mathbb{C}) \).

III.3.7.3. Let \( R \) be the set of all sequences of \( 2 \times 2 \) matrices (with real entries) which are eventually 0. Add and multiply sequences coordinatewise. Show that \( R \) with these operations is a noncommutative nonunital ring.

III.3.7.4. Let \( R \) be a ring. A subset \( J \) of \( R \) is an ideal of \( R \) if \( J \) is closed under addition, \( 0 \in J \), \( x \in J \) implies \((-x) \in J \), and \( x \in J, y \in R \) imply that \( xy, yx \in J \) (cf. II.6.4.12).

(a) Fix \( n \in \mathbb{N} \), and let \( n\mathbb{Z} \) be the set of all integers divisible by \( n \). Show that \( n\mathbb{Z} \) is an ideal in \( \mathbb{Z} \). Show that every ideal of \( \mathbb{Z} \) other than \( \{0\} \) is of this form.

(b) Let \( J \) be an ideal in a ring \( R \). If \( x, y \in R \), set \( x \sim y \) if \((y - x) \in J \). Show that \( \sim \) is an equivalence relation.

(c) Let \( R/J \) be the set of \( \sim \)-equivalence classes \( \{[x] : x \in R\} \) (cf. (b)). Show that \([x] + [y] = [x + y] \) and \([x][y] = [xy] \) give well-defined binary operations on \( R/J \) making it into a ring, which is commutative [resp. unital] if \( R \) is commutative [resp. unital]. \( R/J \) is called the quotient ring of \( R \) by \( J \), pronounced “\( R \) mod \( J \)”.

III.3.7.5. Use the notation and terminology of Exercise III.3.7.4.

(a) Show that the quotient ring \( \mathbb{Z}/n\mathbb{Z} \) can be identified with the ring \( \mathbb{Z}_n \) of “integers mod \( n \),” whose elements are \( \{0, 1, 2, \ldots, n-1\} \) and where addition and multiplication are done modulo \( n \).

(b) Show that \((\mathbb{Z}_n, 1, +)\) is a model for \( \mathbb{N} \) with axiom \((\text{N}3)\) deleted.

(c) Show that \((\mathbb{Z}_n, +, 0, 1)\) is a model for \( \mathbb{Z} \) with axiom \((\text{Fr})\) deleted.

273
This problem gives an extension of Theorem III.3.6.2. Let $R$ be a unital ring. Define a function $\phi : \mathbb{Z} \to R$ by $\phi(0) = 0$, $\phi(1) = 1$, and $\phi(x + 1) = \phi(x) + 1$ and $\phi(-x) = -\phi(x)$ for $x \in \mathbb{N}$.

(a) Show that $\phi(x + y) = \phi(x) + \phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in \mathbb{Z}$ (\(\phi\) is a ring-homomorphism).

(b) Show that $\phi(\mathbb{Z}) = \{\phi(x) : x \in \mathbb{Z}\}$ is a subring of $R$ (a subset closed under addition and multiplication which is itself a ring), and is the smallest subring of $R$ containing 1.

(c) Show that $J = \{x \in \mathbb{Z} : \phi(x) = 0\}$ is an ideal of $\mathbb{Z}$, and that $\phi$ gives a ring-isomorphism (bijective ring-homomorphism) from $\mathbb{Z}/J$ to $\phi(\mathbb{Z})$. Thus, either $\phi$ is injective and $\phi(\mathbb{Z})$ is isomorphic to $Z$ as a ring, or $J = n\mathbb{Z}$ for some $n \in \mathbb{N}$, in which case $\phi(\mathbb{Z})$ is isomorphic to $Z_n$ as a ring.

III.3.7.7. The Grothendieck Group. Let $(S, \ast)$ be an abelian semigroup, a set $S$ with an associative, commutative binary operation $\ast$. There is a standard construction of an abelian group $G(S)$ which is virtually identical to the construction of $\mathbb{Z}$ from $\mathbb{N}$ or $\mathbb{N}_0$. The only change is that the equivalence relation must be modified to account for possible lack of cancellation in $(S, \ast)$.

(a) Define a relation $\sim$ on $S \times S$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if there is a $z \in S$ with

$$x_1 \ast y_2 \ast z = x_2 \ast y_1 \ast z.$$ 

Show that $\sim$ is an equivalence relation on $S \times S$.

(b) Let $G(S)$ be the set of $\sim$-equivalence classes $\{(x, y) : x, y \in S\}$. Show that $\mathbb{Z}$ is a group (a subset closed under addition and multiplication which is itself a ring), and is the smallest subgroup of $R$ containing 1.

(c) Show that $\{[(x, y)] : x, y \in S\}$ for any $x, y \in S$, and that this class gives an additive identity.

(d) If $[(x, y)] = [(y, x)]$, show that $G(S)$ is an abelian (commutative) group, i.e. satisfies axioms (A1)--(A4).

(e) Show that $[(x \ast y, y)] = [(x \ast z, z)]$ for any $x, y, z \in S$. Show that the map $\iota : S \to G(S)$ defined by $\iota(x) = [(x \ast x, x)]$ is a (group-)homomorphism, i.e. $\iota(x \ast y) = \iota(x) \ast \iota(y)$ for all $x, y \in S$. Show that $\iota$ is injective if and only if $S$ has cancellation.

(f) Show that this construction, applied to $(\mathbb{N}, \cdot)$, gives the positive rational numbers with multiplication.

(g) Show that the group $G(S)$ is universal in the sense that if $H$ is any abelian group and $\phi : S \to H$ is a homomorphism, there is a unique homomorphism $\psi : G(S) \to H$ such that $\phi = \psi \circ \iota$.

(h) Show that $S \to G(S)$ is a functor (\() from the category of abelian semigroups to the category of abelian groups. Part of showing this is to show that if $S$ and $S'$ are abelian semigroups, and $\phi : S \to S'$ is a homomorphism, there is a corresponding induced homomorphism from $G(S)$ to $G(S')$.

The term abelian, as a synonym for commutative, is used for groups and semigroups (and also for rings by some authors; abelian categories, functions, and varieties also use this name) in honor of the Norwegian mathematician Niels Henrik Abel, who made important contributions to both algebra and analysis before his death at age 26 from tuberculosis.

III.3.7.8. An element $x$ in a commutative ring $R$ is a zero divisor if $x \neq 0$ and there exists $y \in R$, $y \neq 0$, with $xy = 0$.

(a) Show that $xz = yz$ implies $x = y$ in a commutative ring $R$ if and only if $z$ is not a zero divisor. [Consider $(x - y)z$]

(b) A unital commutative ring is an integral domain if it contains no zero divisors. Show that every unital subring of a field is an integral domain. Show that the polynomial ring $\mathbb{R}[X]$ (Exercise ()) is an integral domain.
III.4. The Rational Numbers

III.4.1. Introduction and Axiomatization of \(\mathbb{Q}\)

The next step in building the number system is the construction of the rational numbers, usually denoted \(\mathbb{Q}\) (for “quotient”). We will follow the usual process of listing a set of axioms satisfied by \(\mathbb{Q}\), and then showing that there is a model for the axioms and that the model is essentially unique.

III.4.1.1. Rational numbers are often described as “fractions,” but this is incorrect. Rational numbers are not the same thing as fractions: many fractions such as \(\frac{1}{2}\) and \(\frac{2}{4}\) can represent the same rational number. It is more correct to say that a rational number is a number which can be represented as a fraction. (A similar distinction can be made for integers: an integer is a number that can be represented as a difference of natural numbers, but integers are not the same thing as differences of natural numbers.) Saying that rational numbers are fractions leads to other logical problems. For example, is \(\cos \frac{\pi}{2}\) a rational number? What about \(\sum_{k=0}^{\infty} \frac{1}{3^k}\)?

III.4.1.2. The axioms satisfied by \(\mathbb{Q}\) are very similar to the ones satisfied by \(\mathbb{Z}\), along with an additional axiom stating that every nonzero element of \(\mathbb{Q}\) has a multiplicative inverse. We state them in the ordered field form in analogy with III.3.6.7. (These are not the most efficient axioms, but are the ones most commonly used.)

**Ordered Field Axioms for \(\mathbb{Q}\)**

(Op) There are two binary operations \(+\) and \(\cdot\) on \(\mathbb{Q}\).

(A1) The binary operation \(+\) is associative.

(A2) The binary operation \(+\) is commutative.

(A3) There is an element \(0 \in \mathbb{Q}\) such that \(x + 0 = 0 + x = x\) for all \(x \in \mathbb{Q}\).

(A4) For every \(x \in \mathbb{Q}\) there is an element \((-x) \in \mathbb{Q}\) with \(x + (-x) = (-x) + x = 0\).

(M1) The binary operation \(\cdot\) is associative.

(M2) The binary operation \(\cdot\) is commutative.

(M3) There is an element \(1 \in \mathbb{Q}\) such that \(x \cdot 1 = 1 \cdot x = x\) for all \(x \in \mathbb{Q}\).

(M4) For every \(x \in \mathbb{Q}\), \(x \neq 0\), there is an element \(x^{-1} \in \mathbb{Q}\) with \(x^{-1} x = xx^{-1} = 1\).

(D) The operation \(\cdot\) distributes over addition.

(Pos) There is a nonempty subset \(P\) of \(\mathbb{Q}\) such that

(P1) If \(x, y \in P\), then \(x + y \in P\).

(P2) If \(x, y \in P\), then \(xy \in P\).

(P3) \(0 \notin P\).

(P4) If \(x \in \mathbb{Q}\), \(x \neq 0\), then either \(x \in P\) or \((-x) \in P\), but not both.

(Min\(\mathbb{Q}\)) \(P\) is the smallest subset \(A\) of \(\mathbb{Q}\) containing 1, such that \(x \in A\) implies \(x + 1 \in A\) and \(x \in A\) implies \(x^{-1} \in A\).

As with \(\mathbb{Z}\), the axiom (Ord) may be used in place of (Pos):
There is a nonempty relation (strict total order) \(<\) on \(\mathbb{Q}\), such that

(O1) If \(x, y \in \mathbb{Q}\), then exactly one of \(x < y\), \(y < x\), \(x = y\) holds.  \((\text{Trichotomy})\)

(O2) If \(x, y, z \in \mathbb{Q}\), \(x < y\), and \(y < z\), then \(x < z\).  \((\text{Transitivity})\)

(O3) If \(x, y, z \in \mathbb{Q}\) and \(x < y\), then \(x + z < y + z\).  \((\text{Translation Invariance})\)

(O4) If \(x, y, z \in \mathbb{Q}\), \(x < y\), and \(0 < z\), then \(xz < yz\).  \((\text{Homogeneity or Scaling Invariance})\)

III.4.1.3. A set with two binary operations satisfying (A1)–(A4), (M1)–(M4), and (D), and in which

\(0 \neq 1\), is called a field. A field satisfying (Pos) is an ordered field (note that an ordered field contains more

than one element by (Pos), so \(0 \neq 1\) automatically by III.3.6.4.(iii)).

A field is just a unital commutative ring in which every nonzero element has a multiplicative inverse. There are fields with finitely many elements (Exercise III.4.3.1.). An ordered field looks more like \(\mathbb{Q}\) (Exercise III.4.3.7.), but there are many other ordered fields such as \(\mathbb{R}\); see also Exercises III.5.3.1. and III.5.3.3.(d). See also Exercise III.6.3.2. for an interesting fieldlike ring which is not commutative.

III.4.2. The Standard Model of \(\mathbb{Q}\)

We can form a standard model for \(\mathbb{Q}\) in much the same manner as for \(\mathbb{Z}\), although the construction will be

slightly more complicated since we must explicitly extend both addition and multiplication. The idea is that rational numbers can be represented by equivalence classes of fractions \(\frac{x}{y}\), where \(x \in \mathbb{Z}\), \(y \in \mathbb{N}\), and \(\frac{x}{y} = \frac{z}{w}\) if and only if \(xw = yz\).

III.4.2.1. Definition. Define a relation \(\sim\) on \(\mathbb{Z} \times \mathbb{N}\) by \((x_1, y_1) \sim (x_2, y_2)\) if \(x_1y_2 = x_2y_1\).

We think of \((x, y)\) as representing the rational number \(\frac{x}{y}\). We could just as well work on \(\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})\).

III.4.2.2. Proposition. (i) The relation \(\sim\) is an equivalence relation on \(\mathbb{Z} \times \mathbb{N}\).

(ii) If \((x_1, y_1) \sim (x_2, y_2)\) and \((z_1, w_1) \sim (z_2, w_2)\), then \((x_1w_1 + y_1z_1, y_1w_1) \sim (x_2w_2 + y_2z_2, y_2w_2)\) and \((x_1z_1, y_1w_1) \sim (x_2z_2, y_2w_2)\).

Proof: This is a straightforward calculation and is left to the reader. ⊥

The operations in (ii) are motivated by the informal (at this point) relations

\[
\frac{x}{y} + \frac{z}{w} = \frac{xw + yz}{yw} \quad \text{and} \quad \frac{x}{y} \frac{z}{w} = \frac{xz}{yw}.
\]

We use these relations to define addition and multiplication:

III.4.2.3. Definition. Let \(\mathbb{Q}\) be the set of equivalence classes for \(\sim\). If \([x, y], [z, w] \in \mathbb{Q}\), define

\([x, y] + [z, w] = [(xw + yz, yw)]\) \quad \text{and} \quad \([x, y] [z, w] = [(xz, yw)]\).
III.4.2.4. **Proposition.** Addition and multiplication are well-defined binary operations on $\mathbb{Q}$ making $\mathbb{Q}$ into a field.

**Proof:** Addition and multiplication are well defined by III.4.2.2.(ii). The associative law for multiplication, and the commutative laws for addition and multiplication, are obvious. The associative law for addition and the distributive law are straightforward calculations from the ring axioms of $\mathbb{Z}$, and are left to the reader.

The 0 element is $[(0,1)]$; it is obvious that $[(0,1)] + [(x,y)] = [(x,y)] + [(0,1)] = [(x,y)]$ for all $x,y$. Similarly, the negative of $[(x,y)]$ is $[(-x,y)]$, as is easily verified.

The 1 element is $[(1,1)]$. If $x \in \mathbb{N}$ (and of course $y \in \mathbb{N}$), then $[(x,y)]^{-1} = [(y,x)]$; if $(-x) \in \mathbb{N}$, then $[(x,y)]^{-1} = [(-y,-x)]$. These formulas can be easily verified by the reader. $lacksquare$

### III.4.3. Exercises

**III.4.3.1.** Let $\mathbb{Z}_n$ be as in Exercise III.3.7.5. Show that $\mathbb{Z}_n$ is a field if and only if $n$ is prime. [If $n > 1$ is not prime, there are $x,y \in \mathbb{Z}_n$ with $x \neq 0$ and $y \neq 0$ but $xy = 0$. If $n$ is prime, show that, for any $x \in \mathbb{Z}_n$, $x \neq 0$, the function $\phi_x : \mathbb{Z}_n \to \mathbb{Z}_n$ given by $\phi_x(y) = xy$ is injective. Use finiteness of $\mathbb{Z}_n$ to conclude that $\phi_x$ is also surjective.]

In particular, $\mathbb{Z}_2$ is a field with only two elements.

**III.4.3.2.** Let $F$ be a field. Let $\phi$ be the ring-homomorphism from $\mathbb{Z}$ to $F$ defined in Exercise III.3.7.6. (a) If the ideal $J$ of III.3.7.6.(c) is $\{0\}$, show that the injective ring-homomorphism $\phi$ extends to an injective ring-homomorphism from $\mathbb{Q}$ to $F$, also denoted $\phi$. Show that $\phi(\mathbb{Q})$ is the smallest subfield of $F$.

(b) If the ideal $J$ is $n\mathbb{Z}$, show that $n$ is prime, and $\phi(\mathbb{Z}) \cong \mathbb{Z}_n$ is the smallest subfield of $F$. [Use Exercises III.4.3.1. and III.3.7.6.(c).]

The number $n$ in case (b) is called the characteristic of $F$. In case (a), $F$ is said to have characteristic zero. Thus, for example, $\mathbb{Z}_p$ has characteristic $p$ and the fields of analysis, including $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, have characteristic zero.

**III.4.3.3.** (a) Show that $\mathbb{Z} + \mathbb{Z}\sqrt{2} = \{x + y\sqrt{2} : x,y \in \mathbb{Z}\}$ is a subring of $\mathbb{R}$.

(b) Show that $\mathbb{Q} + \mathbb{Q}\sqrt{2} = \{x + y\sqrt{2} : x,y \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$.

(c) Do the same for the subsets $\mathbb{Z} + \mathbb{Z}i$ and $\mathbb{Q} + \mathbb{Q}i$ of $\mathbb{C}$. Generalize. What if $\sqrt{2}$ is replaced by $\sqrt{2}$? By $\pi$?

(d) If $\mathbb{Q} + \mathbb{Q}\sqrt{2}$ is regarded as a subset of $\mathbb{R}$, the ordering on $\mathbb{R}$ induces an ordering on $\mathbb{Q} + \mathbb{Q}\sqrt{2}$ making it into an ordered field. Can $\mathbb{Q} + \mathbb{Q}i$ be made into an ordered field? [If $P$ is a positive cone in $\mathbb{Q} + \mathbb{Q}i$, then either $i \in P$ or $(-i) \in P$; in either case $(-1) \in P$.]

**III.4.3.4.** **Rings of Fractions.** Here is a construction which generalizes the construction of $\mathbb{Q}$ from $\mathbb{Z}$. Let $R$ be a commutative ring, and $D$ a nonempty subset of $R$ which is closed under multiplication and which does not contain 0 or any zero divisor (the construction can be done more generally, but then becomes rather artificial and degenerate). We will embed $R$ into a larger unital commutative ring $D^{-1}R$, in which the elements of $D$ have multiplicative inverses.

(a) Put a relation $\sim$ on $R \times D$ by setting $(x_1,y_1) \sim (x_2,y_2)$ if $x_1y_2 = x_2y_1$. Show that $\sim$ is an equivalence relation. [Use Exercise 14.]
Let $D^{-1}R$ be the set of $\sim$-equivalence classes. Show that $[(x, y)] + [(z, w)] = [(xw + yz, yw)]$ and $[(x, y)][(z, w)] = [(xz, yw)]$ give well-defined binary operations making $D^{-1}R$ into a unital commutative ring. [Think of $[(x, y)]$ as representing $\frac{x}{y}$.]

(c) Show that $\iota : R \to D^{-1}R$ given by $\iota(x) = [(xy, y)]$ for $y \in D$ is a well-defined injective ring-homomorphism.

(d) Show that every element of $D$ has a multiplicative inverse in $D^{-1}R$.

The ring $D^{-1}R$ is called the ring of fractions of $R$ with denominator $D$. In the special case where $R$ is an integral domain (Exercise III.3.7.8.) and $D$ is the set of nonzero elements of $R$, $D^{-1}R$ is a field, called the quotient field, or field of fractions, of $R$.

III.4.3.5. Use the notation and terminology of Exercise III.4.3.4.

(a) Describe $D^{-1}Z$, where $D$ is the set of powers of 2 in $\mathbb{N}$. This ring is called the ring of dyadic rationals.

(b) Describe $D^{-1}Z$, where $D$ is the set of odd numbers in $\mathbb{N}$. This ring is called the ring of integers localized at 2.

(c) Describe the quotient field of the ring $\mathbb{R}[X]$ of Exercise III.6.3.3. (cf. Exercise ()). This quotient field, usually denoted $\mathbb{R}(X)$, is, by slight abuse of language, called the field of rational functions in one variable with real coefficients. Explain why it is a slight abuse of language to call the elements of this field “functions.” [Consider the “functions” $f(X) = X + 1$ and $g(X) = \frac{X^2 + 1}{X - 1}$.]

III.4.3.6. Show that a field $F$ can be made into an ordered field if and only if 0 is not a sum of nonzero squares in $F$ (such an $F$ is said to be formally real). [The set $S$ of all quotients of sums of nonzero squares is closed under addition, multiplication, and reciprocals. Using Zorn’s Lemma, embed $S$ into a maximal set closed under addition, multiplication, and reciprocals, and not containing 0.] This is a version of the Artin-Schreier Theorem: every formally real field can be embedded in a real-closed field.

III.4.3.7. If $F$ is an ordered field, show that the homomorphism $\phi : Z \to F$ of Exercise III.3.7.6. is injective, and hence by Exercise III.4.3.2.(a) extends to an embedding of $\mathbb{Q}$ into $F$. Show that the induced ordering on $\mathbb{Q}$ is the ordinary one.
III.5. The Real Numbers

III.5.1. Construction of $\mathbb{R}$: the Dedekind Cut Approach

There are two standard approaches to the construction of the real numbers: via Dedekind cuts or via Cauchy sequences. (There is another nice approach using nonstandard analysis (III.10.8.12.).) Both are worthwhile, and each has uses and leads to insights about $\mathbb{R}$ not as easily obtainable from the other. The Dedekind cut approach (originally due, not surprisingly, to R. Dedekind in ()) is arguably slightly simpler, so we describe it first, in this section.

III.5.1.1. The idea is that if $a$ is a real number, then $a$ "cuts" the set of rational numbers into two sets $L_a = \{ x \in \mathbb{Q} : x < a \}$ and $U_a = \{ x \in \mathbb{Q} : x \geq a \}$ (if $a$ is irrational, then $U_a = \{ x \in \mathbb{Q} : x > a \}$; if $a$ is rational, we arbitrarily put it in $U_a$ instead of $L_a$). Every element of $L_a$ is smaller than every element of $U_a$, and we have $\sup(L_a) = \inf(U_a) = a$ (). Conversely, if $(L, U)$ is a division of $\mathbb{Q}$ into two nonempty subsets with each element of $L$ less than every element of $U$, by completeness there will be a real number $a$, e.g. $a = \sup(L) = \inf(U)$, so that $L = L_a$ (provided $L$ has no largest element) and $U = U_a$. Thus there is a one-one correspondence between real numbers and such pairs of subsets of $\mathbb{Q}$, which are called Dedekind cuts.

We may turn this argument around to define real numbers to be Dedekind cuts. (Technically, this will give us a model for $\mathbb{R}$.) In fact, the entire algebraic and order structure on $\mathbb{R}$ can be described directly in terms of the cuts in a fairly simple manner.

It turns out to be technically much easier to work only with positive numbers at first, i.e. to construct a model $\mathcal{P}$ of the set $\mathbb{P}$ of positive real numbers using Dedekind cuts in the set $\mathbb{Q}^+$ of positive rational numbers. The model for the whole set of real numbers is then constructed from $\mathcal{P}$ in a manner similar to the construction of $\mathbb{Z}$ from $\mathbb{N}$. (This approach is used in [Lan51].)

III.5.1.2. Definition. (i) A subset $S$ of $\mathbb{Q}^+ = \{ x \in \mathbb{Q} : x > 0 \}$ is hereditary if $x \in S$, $y \in \mathbb{Q}^+$, $y < x$ imply $y \in S$.

(ii) A lower set in $\mathbb{Q}^+$ is a nonempty hereditary subset of $\mathbb{Q}^+$ which is bounded above and has no largest element.

(iii) An upper set in $\mathbb{Q}^+$ is a subset $U$ of $\mathbb{Q}^+$ such that $\mathbb{Q}^+ \setminus U$ is a lower set in $\mathbb{Q}^+$.

(iv) A Dedekind cut in $\mathbb{Q}^+$ is a pair $(L, U)$, where $L$ is a lower set in $\mathbb{Q}^+$ and $U = \mathbb{Q}^+ \setminus L$.

The next simple observation is obvious, but important.

III.5.1.3. Proposition. If $(L, U)$ is a Dedekind cut in $\mathbb{Q}^+$, $x \in L$, and $y \in U$, then $x < y$.

Proof: We must have $x < y$ since $L$ is hereditary.
III.5.1.4. **Examples.** (i) If \( a \in \mathbb{Q}_+ \), set \( L_a = \{ x \in \mathbb{Q}_+ : x < a \} \) and \( U_a = \{ x \in \mathbb{Q}_+ : x \geq a \} = \mathbb{Q}_+ \setminus L_a \). \( L_a \) is nonempty since \( \frac{1}{n} < a \) for some \( n \in \mathbb{N} \) by (i), and \( L_a \) is bounded above by \( a \). \( L_a \) has no largest element since if \( x \in L_a \), then \( y = (a + x)/2 \in L_a \) and \( x < y \). \( L_a \) is hereditary by transitivity of \(<\). Thus \( L_a \) is a lower set, \( U_a \) is an upper set, and \((L_a, U_a)\) is a Dedekind cut in \( \mathbb{Q}_+ \). The set \( L_a \) is often denoted \( a^* \).

(ii) It is not *a priori* obvious that there are Dedekind cuts in \( \mathbb{Q}_+ \) not of the form \((L_a, U_a)\) for \( a \in \mathbb{Q}_+ \), but it turns out that there are very many which are not. Here is an example which is at least not obviously of this form. Let

\[
L = \{ x \in \mathbb{Q}_+ : x^2 < 2 \}
\]

and \( U = \mathbb{Q}_+ \setminus L \). \( L \) is hereditary since if \( 0 < y < x \), then \( y^2 < x^2 \). \( 1 \in L \), so \( L \) is nonempty, and \( L \) is bounded above since if \( x \in L \), we have \( x < 2 \) since \( x^2 < 2^2 \).

We show that \( L \) has no largest element. If \( x \in L \), set \( z = \frac{4 + \sqrt{2x}}{3 + 2x} \). Then

\[
z - x = \frac{4 - 2x^2}{3 + 2x} > 0 \quad \text{and} \quad 2 - z^2 = \frac{2 - x^2}{(3 + 2x)^2} > 0
\]

by a simple computation; thus \( x < z \) and \( z \in L \).

Although there is no obvious \( a \in \mathbb{Q}_+ \) for which \( L = L_a \), it takes a little work to show that in fact there is no such \( a \) (I.5.5.).

(iii) The only hereditary subset of \( \mathbb{Q}_+ \) which is not bounded above is all of \( \mathbb{Q}_+ \). For if \( S \) is such a set, and \( y \in \mathbb{Q}_+ \), then \( y \) is not an upper bound for \( S \), so there is an \( x \in S \) with \( y < x \). But then \( y \in S \) since \( S \) is hereditary, so \( S = \mathbb{Q}_+ \). Therefore a nonempty proper subset of \( \mathbb{Q}_+ \) which is hereditary and has no largest element is automatically a lower set.

(iv) An arbitrary union or intersection of hereditary subsets of \( \mathbb{Q}_+ \) is hereditary. An arbitrary union of lower sets is either \( \mathbb{Q}_+ \) or a lower set. A finite intersection of lower sets is a lower set. However, an arbitrary intersection of lower sets is not necessarily a lower set: it can be empty (e.g. \( \cap_n L_{1/n} = \emptyset \)) or it may have a largest element (e.g. \( \cap_n L_{1+1/n} \)).

(v) If \( S \) is a (necessarily proper) hereditary subset of \( \mathbb{Q}_+ \) which has a largest element \( a \), let \( L = S \setminus \{a\} \). Then \( L \) is nonempty (\( \frac{1}{n} \in L \) for sufficiently large \( n \)), bounded above (by \( a \)), and hereditary. \( L \) also has no largest element since if \( x \in L \), then \( y = (x + a)/2 \) is in \( L \) and \( x < y \). Thus \( L \) is a lower set. In fact, \( L \) is exactly \( L_a \) of (i).

Here is a useful technical fact about Dedekind cuts:

III.5.1.5. **Theorem.** Let \((L, U)\) be a Dedekind cut in \( \mathbb{Q}_+ \), and \( \epsilon \in \mathbb{Q}_+ \). Then there exist \( x \in L \) and \( y \in U \), \( y \) not minimal in \( U \), with \( y - x = \epsilon \).

**Proof:** Fix \( a \in L \) and \( b \in U \). The set \( \{a + n\epsilon : n \in \mathbb{N}\} \) is unbounded in \( \mathbb{Q}_+ \) by (), hence there is an \( n \) with \( a + n\epsilon > b \), and \( a + n\epsilon \in U \) for this \( n \). Thus there is a smallest \( n \) for which \( a + n\epsilon \in U \). For this smallest \( n \), set \( x = a + (n - 1)\epsilon \) and \( y = n\epsilon \). To insures that \( y \) is not the smallest element of \( U \), replace \( x \) and \( y \) by \( x + \frac{1}{m} \) and \( y + \frac{1}{m} \), where \( m \in \mathbb{N} \) is large enough that \( x + \frac{1}{m} \in L \). \( \square \)

280
We now make our main definition:

III.5.1.6. Definition. \( \mathcal{P} \) denotes the set of all lower subsets of \( \mathbb{Q}^+ \).

Ordering on \( \mathcal{P} \)

III.5.1.7. \( \mathcal{P} \) is naturally ordered by inclusion. If \( L, L' \in \mathcal{P} \), write \( L \leq L' \) if \( L \subseteq L' \), and \( L < L' \) if \( L \subseteq L' \) and \( L \neq L' \), i.e. if \( L \) is a proper subset of \( L' \). This ordering is antisymmetric: if \( L \leq L' \) and \( L' \leq L \), then \( L = L' \), or alternatively, at most one of \( L = L' \), \( L < L' \) and \( L' < L \) holds.

Transitivity of this ordering is obvious.
As usual, we will also write \( L' \geq L \) if \( L \leq L' \), and \( L' > L \) if \( L < L' \).

III.5.1.8. Proposition. The ordering \( \leq \) is a total ordering on \( \mathcal{P} \), i.e. if \( L, L' \in \mathcal{P} \), then either \( L \leq L' \) or \( L' \leq L \). Alternatively, exactly one of \( L < L' \), \( L = L' \), \( L' < L \) holds.

Proof: If \( L = L' \), there is nothing to prove. Suppose there is an \( a \in L' \) with \( a \notin L \). Thus \( a \in \mathbb{Q}^+ \setminus L \), so \( x < a \) for all \( x \in L \) by III.5.1.3.. But any such \( x \) is in \( L' \) since \( L' \) is hereditary, so \( L \subseteq L' \).

Completeness of this ordering is deceptively simple:

III.5.1.9. Proposition. If \( \mathcal{L} = \{ L_i : i \in I \} \) is a nonempty subset of \( \mathcal{P} \) which is bounded above, then \( \mathcal{L} \) has a supremum (least upper bound) in \( \mathcal{P} \).

Proof: \( \mathcal{L} \) is bounded above if (and only if) there is an \( L' \in \mathcal{P} \) with \( L_i \subseteq L' \) for all \( i \in I \). Fix such an \( L' \). Let \( L = \bigcup_{i \in I} L_i \). Since \( L \subseteq L' \), \( L \) is a proper subset of \( \mathbb{Q}^+ \), hence is a lower set by III.5.1.4.(iii)–(iv). We have \( L_i \leq L \) for all \( i \), and if \( L'' \in \mathcal{P} \) with \( L_i \leq L'' \) for all \( i \), then \( L \leq L'' \), so \( L \) is the least upper bound of \( \mathcal{L} \) in \( \mathcal{P} \).

Infima can be done similarly, but there is a slight technical glitch:

III.5.1.10. Proposition. If \( \mathcal{L} = \{ L_i : i \in I \} \) is a nonempty subset of \( \mathcal{P} \) which is bounded below, then \( \mathcal{L} \) has an infimum (greatest lower bound) in \( \mathcal{P} \).

Proof: \( \mathcal{L} \) is bounded below if (and only if) there is an \( L' \in \mathcal{P} \) with \( L' \subseteq L_i \) for all \( i \in I \). Fix such an \( L' \). Let \( S = \bigcap_{i \in I} L_i \). Since \( L' \subseteq S \), \( S \) is nonempty, and is a proper hereditary subset of \( \mathbb{Q}^+ \). But \( S \) could have a largest element. If \( S \) has no largest element, set \( L = S \), and if \( S \) has a largest element \( a \), set \( L = S \setminus \{ a \} \). In either case \( L \) is a lower set (cf. III.5.1.4.(v)). We have \( L \leq L_i \) for all \( i \), and if \( L'' \in \mathcal{P} \) with \( L'' \leq L_i \) for all \( i \), then \( L'' \leq L \), so \( L \) is the greatest lower bound of \( \mathcal{L} \) in \( \mathcal{P} \).

III.5.1.11. So far, we could have equally well worked with Dedekind cuts on all of \( \mathbb{Q} \), and defined \( \mathbb{R} \) as an ordered set; restriction to positive numbers only has no advantages to this point. But defining the algebraic operations on \( \mathbb{R} \) is a little messy, and it is much cleaner and easier to work with \( \mathcal{P} \) instead.
Algebraic Operations on $\mathcal{P}$

We first make a general definition of sums and products of subsets of $\mathbb{Q}_+$:

III.5.1.12. Definition. Let $A$ and $B$ be subsets of $\mathbb{Q}_+$. Define

\[
A + B = \{x + y : x \in A, y \in B\}
\]
\[
AB = \{xy : x \in A, y \in B\}.
\]

Note that $A + B$ and $AB$ are subsets of $\mathbb{Q}_+$ since $\mathbb{Q}_+$ is closed under addition and multiplication.

III.5.1.13. Caution: Some authors, especially in older references or in the context of Boolean algebras, write $A + B$ for $A \cup B$ and $AB$ for $A \cap B$. We will never use these meanings of $A + B$ and $AB$. Note that $A + B$ (as we have defined it) and $A \cup B$ are quite different in general, as are $AB$ and $A \cap B$.

III.5.1.14. Proposition. Let $L$ and $L'$ be lower sets in $\mathbb{Q}_+$. Then $L + L'$ and $LL'$ are lower sets in $\mathbb{Q}_+$.

Proof: $L + L'$ and $LL'$ are clearly nonempty. If $m$ and $n$ are upper bounds for $L$ and $L'$ respectively, then $m + n$ and $mn$ are upper bounds for $L + L'$ and $LL'$ respectively by ().

We show $L + L'$ is hereditary. If $z \in L + L'$ and $w \in \mathbb{Q}_+$, $w < z$, write $z = x + y$ for $x \in L, y \in L'$. If $w \leq y$, there is an $n \in \mathbb{N}$ such that $\frac{1}{n} < w$ and $\frac{1}{n} < x$; then $\frac{1}{n} \in L$ and $0 < w - \frac{1}{n} < y$, so $y - \frac{1}{n} \in L'$ and $w = \frac{1}{n} + (w - \frac{1}{n}) \in L + L'$. On the other hand, if $y < w$, we have $0 < w - y < z - y = x$, so $(w - y) \in L$, and $w = (w - y) + y \in L + L'$ again.

The proof that $LL'$ is hereditary is easier. If $z \in LL'$ and $w \in \mathbb{Q}_+$, $w < z$, write $z = xy$ for $x \in L, y \in L'$. Then $w < xy$, so $0 < \frac{w}{y} < x$ and hence $\frac{w}{y} \in L$ and $w = \frac{w}{y}y \in LL'$.

Finally, we show that $L + L'$ and $LL'$ have no maximal elements. Let $z \in L + L'$, and write $z = x + y$ for $x \in L, y \in L'$. Since $L$ has no largest element, there is an $r \in L$ with $x < r$. Then $r + y \in L + L'$, and $z < r + y$. The proof for $LL'$ is similar.

It is now clear how to define addition and multiplication in $\mathcal{P}$:

III.5.1.15. Definition. Define binary operations $+$ and $\cdot$ on $\mathcal{P}$ to be sum and product as defined in III.5.1.12.

As usual, we will write the product of $L$ and $L'$ either as $L \cdot L'$ or $LL'$. We will use the usual convention of grouping multiplication before addition, e.g. $LL' + L'' = (LL') + L''$.

The usual algebraic properties of addition and multiplication in $\mathcal{P}$ follow easily from those in $\mathbb{Q}$:
III.5.1.16. Proposition. Let \( L, L', L'' \in \mathcal{P} \). Then

(i) \((L + L') + L'' = L + (L' + L'')\) and \((LL')L'' = L(LL'')\). (Associative Law)

(ii) \(L + L' = L' + L\) and \(LL' = L'L\). (Commutative Law)

(iii) \(L(L' + L'') = LL' + LL''\). (Distributive Law)

Proof: The proofs are quite easy, and all are essentially identical. We prove the associative law for addition, and leave the others to the reader. We have

\[
(L + L') + L'' = \{x + y + z : x \in L, y \in L', z \in L''\}
\]

\[
L + (L' + L'') = \{x + (y + z) : x \in L, y \in L', z \in L''\}
\]

and by the associative law in \(\mathbb{Q}\) (A1), these two subsets of \(\mathbb{Q}_+\) are identical.

III.5.1.17. Proposition. The element \(1^* = L_1 = \{x \in \mathbb{Q}_+ : x < 1\}\) is a multiplicative identity for \(\mathcal{P}\), i.e. \(1^* \cdot L = L\) for any \(L \in \mathcal{P}\).

Proof: Let \(L \in \mathcal{P}\). Since \(L\) is hereditary, it is clear that \(1^* \cdot L \leq L\). On the other hand, if \(x \in L\), choose \(y \in L\) with \(x < y\) (since \(L\) has no largest element); then \(\frac{x}{y} < 1\), so \(\frac{x}{y} \in 1^*\), and \(x = \frac{x}{y} \in 1^* \cdot L\).

On the lower sets coming from elements of \(\mathbb{Q}_+\) as in III.5.1.4(i), addition and multiplication in \(\mathcal{P}\) agree with addition and multiplication in \(\mathbb{Q}_+\):

III.5.1.18. Proposition. Let \(a, b \in \mathbb{Q}_+\), and \(L_a = \{x \in \mathbb{Q}_+ : x < a\}\) and \(L_b = \{x \in \mathbb{Q}_+ : x < b\}\) the corresponding elements of \(\mathcal{P}\). Then \(L_a + L_b = L_{a+b}\) and \(L_aL_b = L_{ab}\).

Proof: It is obvious from (i) that \(L_a + L_b \leq L_{a+b}\). By \(L_a + L_b\), i.e. \(z < a + b\), set \(d = a + b - z\). Then \(x = a - \frac{d}{2} \in L_a\) and \(y = b - \frac{d}{2} \in L_b\), and \(z = x + y \in L_{a+b}\).

For the product, it follows from (i) that \(L_aL_b \leq L_{ab}\). If \(z \in L_{ab}\), we have \(\frac{z}{2} < a\). Choose \(x \in \mathbb{Q}_+\) with \(\frac{z}{2} < x < a\), so \(x \in L_a\). Then \(\frac{z}{2} < b\), so \(y = \frac{z}{2} \in L_b\), and \(z = xy \in L_{ab}\).

Multiplicative Inverses

III.5.1.19. Every element of \(\mathcal{P}\) has a multiplicative inverse, but the argument has a slight subtlety. Suppose \(L \in \mathcal{P}\), and \(U = \mathbb{Q}_+ \setminus L\) is the corresponding upper set. Since inversion reverses order (i), we should have

\[
L^{-1} = \{x^{-1} : x \in U\}.
\]

However, this may not be a lower set since it might have a largest element. In fact, this will happen if and only if \(L = L_a\) for some \(a \in \mathbb{Q}_+\) (we then obtain the set \(\{x \in \mathbb{Q}_+ : x \leq a^{-1}\}\)). In this case, we take \(L^{-1} = \{x \in \mathbb{Q}_+ : x < a^{-1}\} = L_{a^{-1}}\). Thus we define \(L^{-1}\) differently depending on whether or not \(L = L_a\) for some \(a \in \mathbb{Q}_+\).
III.5.1.20. \textbf{Proposition.} If $L \in \mathcal{P}$, and $L^{-1}$ is defined as above, then $L^{-1}$ is a lower set and $LL^{-1} = 1$.

\textbf{Proof:} Clearly $L^{-1}$ is nonempty. Suppose $x \in L^{-1}$, $y \in \mathbb{Q}_+$, and $y < x$. Then $x^{-1} < y^{-1}$ and since $x \notin L$, we have $y \notin L$ also since $L$ is hereditary. Thus $y \in U$ and $y^{-1} \in L^{-1}$. So $L^{-1}$ is hereditary. If $x \in L$, then $y^{-1} < x^{-1}$ for all $y \in U$ by (b) and III.5.1.3; hence $x^{-1}$ is an upper bound for $L^{-1}$. If $S = \{x^{-1} : x \in U\}$ has a largest element $b$, then $L^{-1} = S \setminus \{b\}$. If $x \in L^{-1}$, i.e. $x^{-1} \in U$ and $x < b$, and $y \in \mathbb{Q}_+$ with $x < y < b$ (e.g. $y = (x + b)/2$), then $y^{-1} > b^{-1}$, so $y \in S$, and $y \neq b$, so $y \in L^{-1}$; thus $L^{-1}$ has no largest element in any case. So $L^{-1} \in \mathcal{P}$.

If $x \in L$ and $y \in L^{-1}$, then $y^{-1} \in U$, so $x < y^{-1}$ and therefore $y < x^{-1}$, so $xy < xx^{-1} = 1$. Thus $LL^{-1} \leq 1$.

The opposite inequality is trickier. We use III.5.1.5. Fix $a \in L$. If $z < 1$, choose $x \in L$ and $y \in U$, $y$ not minimal in $U$, with $y - x = a(1 - z)$. Then $\frac{z}{y} \in LL^{-1}$, and

$$1 - \frac{x}{y} = \frac{a}{y}(1 - z) < 1 - z$$

since $a < y$. Thus $z < \frac{z}{y}$ and $z \in LL^{-1}$.

There are no additive inverses in $\mathcal{P}$. There is not even an additive identity:

III.5.1.21. \textbf{Proposition.} Let $L, L' \in \mathcal{P}$. Then $L < L + L'$.

\textbf{Proof:} If $x \in L$ and $y \in L'$, then $x < x + y \in L + L'$, so $x \in L + L'$ and $L \leq L + L'$. Set $U = \mathbb{Q}_+ \setminus L$. Fix $z \in L'$, and by III.5.1.5. choose $x \in L$, $y \in U$ with $y - x = z$. Then $y = x + z \in L + L'$, but $y \notin L$.

\textbf{Cancellation}

Cancellation for multiplication follows immediately from the existence of inverses:

III.5.1.22. \textbf{Proposition.} If $L, L', L'' \in \mathcal{P}$ and $LL' = LL''$, then $L' = L''$.

\textbf{Proof:} We have $L^{-1}(LL') = L^{-1}(LL'')$. But we have

$$L^{-1}(LL') = (L^{-1}L)L' = 1^* \cdot L' = L'$$

and similarly $L^{-1}(LL'') = L''$, so $L' = L''$.

Additive cancellation takes more work since there are no additive inverses. We first show that $<$ is translation-invariant:
III.5.1.23. **Proposition.** Let \( L, L', L'' \in P \). If \( L < L' \), then \( L + L'' < L' + L'' \).

**Proof:** It is obvious that \( L + L'' \leq L' + L'' \). Fix \( a \in L' \setminus L \), and \( b \in L' \) with \( a < b \). Let \( U'' = \mathbb{Q}_+ \setminus L'' \), and \( x \in L'' \), \( y \in U'' \) with \( y - x = b - a \) (III.5.1.5.). Then \( z = b + x \in L' + L'' \), but if \( z = s + t \) for \( s \in L \), \( t \in L'' \), we have \( s < a \), so \( b - s > b - a = y - x \), so \( t = z - s = (b - s) + x > (y - x) + x = y \), a contradiction. Thus \( z \in (L' + L'') \setminus (L + L'') \).

Additive cancellation then follows:

III.5.1.24. **Corollary.** Let \( L, L', L'' \in P \). If \( L + L'' < L' + L'' \), then \( L < L' \). If \( L + L'' = L' + L'' \), then \( L = L' \).

**Proof:** Either \( L < L' \), \( L = L' \), or \( L > L' \). In these cases \( L + L'' < L' + L'' \), \( L + L'' = L' + L'' \), and \( L + L' > L' + L'' \) respectively.

**Subtraction**

We also get subtraction defined to the extent possible in light of III.5.1.21.:

III.5.1.25. **Proposition.** Let \( L, L' \in P \). If \( L < L' \), then there is a unique \( L'' \in P \) such that \( L + L'' = L' \).

**Proof:** Uniqueness follows immediately from III.5.1.24. We show existence. Set \( U = \mathbb{Q}_+ \setminus L \), and let

\[
L'' = \{ x - y : x \in L', \ y \in U, \ y < x \} .
\]

We show \( L'' \) is a lower set. If \( y \in L' \setminus L \) and \( x \in L' \), \( y < x \), then \( x - y \in L'' \), so \( L'' \) is nonempty. If \( x \in L' \), \( y \in U \), and \( z \notin L' \), then \( x - y < x < z \), so \( z \notin L'' \) and \( L'' \) is bounded. If \( t \in L'' \) and \( 0 < s < t \), write \( t = x - y \) for \( x \in L' \), \( y \in U \), \( y < x \); then \( s + y < t + y \in L' \), and \( s = (s + y) - y \), so \( s \in L'' \). Thus \( L'' \) is hereditary. Finally, if \( t \in L'' \) with \( t = x - y \) for \( x \in L'' \), \( y \in U \), \( y < x \), let \( w \in L' \), \( x < w \); then \( s = w - y \in L'' \) and \( t < s \), so \( L'' \) has no largest element. Thus \( L'' \in P \).

If \( z \in L + L'' \), write \( z = w + (x - y) \) with \( w \in L \), \( x \in L' \), \( y \in U \), \( y < x \). Then \( w < y \), so \( z < w + (x - w) = x \in L' \), so \( z \in L' \). Thus \( L + L'' \leq L' \).

If \( z \in L' \), let \( b \in L' \), \( z < b \), and by III.5.1.5. choose \( x \in L \) and \( y \in U \) with \( y - x = b - z \). Then \( z - x = b - y \in L'' \), and \( z = x + (z - x) \in L + L'' \), \( L' \leq L + L'' \).

**Notation**

We now streamline our notation for elements of \( P \), eliminating the clumsy \( L, L' \), etc.

III.5.1.26. By III.5.1.18., the map \( \phi : \mathbb{Q}_+ \to P \) given by \( \phi(a) = L_a \) preserves addition and multiplication.

It is obvious that \( \phi \) is also order-preserving, i.e. if \( a < b \), then \( L_a < L_b \). We will identify \( \mathbb{Q}_+ \) with its image in \( P \), and just write \( a \) instead of \( \phi(a) \), \( L_a \), or \( a^* \), unless there is possibility of confusion.
III.5.1.27. Proposition. The image of $\mathbb{Q}_+$ is order-dense in $\mathcal{P}$, i.e. if $L, L' \in \mathcal{P}$ with $L < L'$, then there is an $a \in \mathbb{Q}_+$ with $L < a < L'$.

Proof: Let $L < L'$. Then $L$ is a proper subset of $L'$. Let $b \in L' \setminus L$, and let $a \in L'$ with $b < a$. Then $b \in L_a \setminus L$, so $L < L_a$, and $a \in L' \setminus L_a$, so $L_a < L'$.

III.5.1.28. With this identification, we can say that if $L \in \mathcal{P}$, then $L = \{x \in \mathbb{Q}_+: x < L\}$. (This statement seems strangely circular, but is not actually logically circular since it is not used as a definition of $L$.)

III.5.1.29. From now on, we will use lower-case letters to denote elements of $\mathcal{P}$ instead of $L$'s. Thus we will write $a, b, x, y \in \mathcal{P}$, etc.; the $a, b, x, y$ may denote elements of $\mathbb{Q}_+$, thought of as a subset of $\mathcal{P}$, or of $\mathbb{Q}_+ = \mathcal{P} \setminus \mathbb{Q}_+$, called the set of positive irrational numbers. (We have not yet shown that this set is nonempty.) So, if $a \in \mathcal{P}$, we officially have

$$a = \{x \in \mathbb{Q}_+: x < a\}$$

whether or not $a$ itself is rational. (This statement is technically logically inconsistent if $a$ is rational, since $a$ has previously been defined to be something else, but we will gloss over this technicality since we are not regarding individual numbers as being any specific set-theoretic objects.)

Construction of the Model for $\mathbb{R}$ from $\mathcal{P}$

We now essentially copy the construction of $\mathbb{Z}$ from $\mathbb{N}$ to obtain a model for $\mathbb{R}$.

III.5.1.30. The construction is notationally easier if we adjoin the rational number 0 (which is not in $\mathcal{P}$) to $\mathcal{P}$. Write $\mathcal{P}_0$ for $\mathcal{P} \cup \{0\}$. We define $0 + x = x + 0 = x$ and $0 \cdot x = x \cdot 0 = 0$ for all $x \in \mathcal{P}_0$. This extends the addition and multiplication on $\mathbb{Q}_+ \cup \{0\}$, and the associative, commutative, and distributive laws continue to hold in $\mathcal{P}_0$. We also extend the ordering on $\mathcal{P}$ to $\mathcal{P}_0$ by setting $0 < x$ for all $x \in \mathcal{P}$.

III.5.1.31. Definition. Define a relation on $\mathcal{P}_0 \times \mathcal{P}_0$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 + y_2 = x_2 + y_1$.

III.5.1.32. Proposition. (i) The relation $\sim$ is an equivalence relation on $\mathcal{P}_0 \times \mathcal{P}_0$.

(ii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 + z_1, y_1 + w_1) \sim (x_2 + z_2, y_2 + w_2)$.

(iii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 z_1 + y_1 w_1, x_1 w_1 + y_1 z_1) \sim (x_2 z_2 + y_2 w_2, x_2 w_2 + y_2 z_2)$.

(iv) $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, and $x_1 + w_1 < z_1 + y_1$, then $x_2 + w_2 < z_2 + y_2$.

Proof: (i): If $(x_1, y_1) \sim (x_2, y_2)$, it is trivial that $(x_2, y_2) \sim (x_1, y_1)$. It is also trivial that $(x, y) \sim (x, y)$ for any $x, y$. If $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$, we have

$$(x_1 + y_3) + (x_2 + y_2) = (x_1 + y_2) + (x_2 + y_3) = (x_2 + y_1) + (x_3 + y_2) = (x_3 + y_1) + (x_2 + y_2)$$

by the associative and commutative laws, and hence $x_1 + y_3 = x_3 + y_1$ by cancellation. Thus $(x_1, y_1) \sim (x_3, y_3)$.

(ii): By the associative and commutative laws, we have

$$(x_1 + z_1) + (y_2 + w_2) = (x_1 + y_2) + (z_1 + w_2) = (x_2 + y_1) + (z_2 + w_1) = (x_2 + z_2) + (y_1 + w_1).$$
(iii): We have
\[(x_1 z_1 + y_1 w_1, x_1 w_1 - y_1 z_1) \sim (x_1 z_1 + y_2 z_1 + x_2 w_1 + y_1 w_1, x_1 w_1 + x_2 w_1 + y_2 z_1 + y_1 z_1)\]
\[= ((x_1 + y_2) z_1 + (x_2 + y_1) w_1, x_1 w_1 + x_2 w_1 + y_2 z_1 + y_1 z_1) = ((x_2 + y_1) z_1 + (x_1 + y_2) w_1, x_1 w_1 + x_2 w_1 + y_2 z_1 + y_1 z_1)\]
\[= (x_2 z_1 + y_1 z_1 + x_1 w_1 + y_2 w_1, x_1 w_1 + x_2 w_1 + y_2 z_1 + y_1 z_1) \sim (x_2 z_1 + y_2 w_1, x_2 w_1 + y_2 z_1) .\]

An almost identical argument shows \((x_2 z_2 + y_2 w_1, x_2 w_1 + y_2 z_2) \sim (x_2 z_2 + y_2 w_2, x_2 w_2 + y_2 z_2)\).

(iv): We have
\[x + y = (x_1 + y_1) + (z_1 + w_1) = (x_1 + z_1) + (y_1 + w_1) \]
\[< (z_1 + y_1) + (y_1 + z_1) = (z_2 + z_2) + (y_1 + z_1) \]
so \(x + w < z + y\) by cancellation.

We will write \([x, y]\) for the equivalence class of the pair \((x, y)\). We will think of \([x, y]\) as representing the real number \(x - y\); the equivalence relation \(\sim\) is designed so that every representative of the equivalence class represents the same real number.

**III.5.1.33.** **Definition.** Let \(\mathcal{R}\) be the set of equivalence classes of ordered pairs. For \([x, y], [z, w] \in \mathcal{R}\), define
\[[x, y] + [z, w] = [(x + z, y + w)]\]
\[[x, y] \cdot [z, w] = [(xz + yw, xw + yz)]\]
\[[x, y] < [z, w] \text{ if } x + w < z + y .\]

Note that +, ·, and < are well defined on \(\mathcal{R}\) by **III.5.1.32**.

**III.5.1.34.** **Theorem.** (\(\mathcal{R}, +, [(0, 0)], [(1, 0)]\)) is a model for \(\mathbb{R}\).

**Proof:** It is obvious that once we know that addition and multiplication of equivalence classes is well defined, \(\mathcal{R}\) inherits associativity, commutativity, and the distributive law from \(\mathcal{P}_0\).

We have \([0, 0] + [x, y] = [(0 + x, 0 + y)] = [(x, y)]\) for all \(x, y \in \mathcal{P}_0\), and similarly \([x, y] + [(0, 0)] = [(x, y)]\) for all \(x, y\); thus (A3) holds with 0-element \((0, 0)\). For \(x, y \in \mathcal{P}_0\), we have \([x, y] + [(y, x)] = [(x + y, y + x)] = [(0, 0)]\), and similarly \([y, x] + [(x, y)] = [(0, 0)]\), so (A4) holds with \(-[(x, y)] = [(y, x)]\).

Suppose \(A \subseteq \mathcal{Z}\) contains \([(0, 0)]\), is closed under adding \([(1, 0)]\) and under negations. Then \(A \cap \mathcal{N}_0 = \mathcal{N}_0\) since \(\mathcal{N}_0\) is a model for \(\mathbb{N}_0\), so \([(x, 0)] \in A\) for all \(x \in \mathbb{N}_0\). Then also \(-[(x, 0)] = [(0, x)] \in A\) for all \(x \in \mathbb{N}_0\). If \(x, y \in \mathbb{N}_0\), either \(x = y\), or there is a \(z \in \mathbb{N}_0\) with \(x + z = y\), or there is a \(w \in \mathbb{N}_0\) with \(x = y + w\). In the first case \([(x, y)] = [(0, 0)] \in A\). If \(x + z = y\), then \([(x, y)] = [(0, z)] \in A\). And if \(x = y + w\), then \([(x, y)] = [(w, 0)] \in A\). Thus (Cyc) holds.

It is obvious that (Fr) holds, since we have shown that \([(x, 0)] \neq [(0, 0)]\) for any \(x \in \mathbb{N}_0, x \neq 0\). \(\blacksquare\)

**III.5.1.35.** Note that each equivalence class in \(\mathcal{R}\) contains exactly one pair \((x, y)\), where either \(x\) or \(y\) (or both) is 0. Thus we could have taken this set of ordered pairs as our model of \(\mathbb{R}\). This model has the advantage that the elements are more concrete; they are specific ordered pairs rather than equivalence classes. But defining addition and proving associativity are somewhat messy in this approach.
III.5.1.36. We could have done the whole construction using $P$ instead of $P_0$. In this case, the 0 element is $[(1, 1)]$, and the 1-element is $[(1 + 1, 1)]$. The positive real number $x$ is represented by $[(x + 1, 1)]$. The approach of III.5.1.35. can also be done using $P$; details are left to the reader.

III.5.2. Construction of $\mathbb{R}$: the Cauchy Sequence Approach

The other standard construction of $\mathbb{R}$ is via equivalence classes of Cauchy sequences of rational numbers. This approach was pioneered by Cantor.

III.5.2.1. The idea behind this construction is that every real number is the limit of a convergent sequence of rational numbers. Cauchy found a criterion for convergence of a sequence which involves only the terms of the sequence, with no explicit reference to the limit; a sequence satisfying this criterion is called a Cauchy sequence. Many different sequences can converge to the same limit, and Cauchy also found a criterion involving only the terms of the sequence for when this happens; Cauchy sequences satisfying this criterion are called equivalent. There is thus a one-one correspondence between real numbers and equivalence classes of Cauchy sequences of rational numbers.

The argument can be turned around to define real numbers to be equivalence classes of Cauchy sequences of rational numbers (this was Cantor’s idea). The algebraic and order properties of real numbers can also be cleanly expressed in terms of these equivalence classes.

We first recall the definition of a convergent sequence. We will phrase it in terms of rational numbers, since we do not assume we have yet constructed anything beyond $\mathbb{Q}$. For a fuller discussion and motivation of the definition, see (). Recall that $\mathbb{Q}_+ = \{x \in \mathbb{Q} : x > 0\}$.

III.5.2.2. Definition. Let $(x_n)$ be a sequence in $\mathbb{Q}$, and $x \in \mathbb{Q}$. Then $(x_n)$ converges to $x$, or $x$ is the limit of the sequence $(x_n)$, written $x_n \to x$ or $x = \lim_{n \to \infty} x_n$, if for every $\epsilon \in \mathbb{Q}_+$ there is a $K \in \mathbb{Q}$ such that $|x_n - x| < \epsilon$ for all $n \in \mathbb{N}, n > K$.

A general sequence in $\mathbb{Q}$ need not converge; in fact, “most” sequences do not converge. It can be readily proved () that a sequence cannot have more than one limit, justifying reference to the limit of a convergent sequence.

Cauchy’s observation is that convergent sequences have an intrinsic property expressible involving only the terms of the sequence:

III.5.2.3. Definition. Let $(x_n)$ be a sequence in $\mathbb{Q}$. Then $(x_n)$ is a Cauchy sequence if, for every $\epsilon \in \mathbb{Q}_+$, there is a $K \in \mathbb{Q}$ such that $|x_n - x_m| < \epsilon$ for all $n, m > K$.

III.5.2.4. Proposition. Let $(x_n)$ be a sequence in $\mathbb{Q}$. If $(x_n)$ converges, then $(x_n)$ is a Cauchy sequence.

Proof: Suppose $x_n \to x$. Let $\epsilon \in \mathbb{Q}_+$. Then there is a $K \in \mathbb{Q}$ such that $|x_n - x| < \frac{\epsilon}{2}$ for all $n > K$. If $n, m > K$, then by the triangle inequality we have

$$|x_n - x_m| \leq |x_n - x| + |x - x_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$
III.5.2.5. Cauchy’s idea was that, conversely, every Cauchy sequence should converge. This turns out to be true in \( \mathbb{R} \) by completeness, but is false in \( \mathbb{Q} \) (although this is not entirely obvious). Thus the idea is to define real numbers to be limits of Cauchy sequences of rational numbers.

However, we also need to specify when two sequences \((x_n)\) and \((y_n)\) should converge to the same limit. This should happen if and only if \(x_n\) and \(y_n\) become arbitrarily close to each other as \(n\) becomes large. So we define:

III.5.2.6. Definition. Let \((x_n)\) and \((y_n)\) be sequences in \(\mathbb{Q}\) (not necessarily Cauchy sequences). Then \((x_n)\) and \((y_n)\) are equivalent, written \((x_n) \sim (y_n)\), if for every \(\epsilon \in \mathbb{Q}_+\) there is a \(K \in \mathbb{Q}\) such that \(|x_n - y_n| < \epsilon\) for all \(n > K\).

The term “equivalent” is justified:

III.5.2.7. Proposition. The relation \(\sim\) is an equivalence relation on the set of all sequences in \(\mathbb{Q}\).

Proof: It is trivial that \(\sim\) is reflexive and symmetric. To prove transitivity, suppose \((x_n), (y_n), \) and \((z_n)\) are sequences in \(\mathbb{Q}\) with \((x_n) \sim (y_n)\) and \((y_n) \sim (z_n)\). Let \(\epsilon \in \mathbb{Q}_+\). There is a \(K_1 \in \mathbb{Q}\) such that \(|x_n - y_n| < \frac{\epsilon}{2}\) for all \(n > K_1\), and a \(K_2 \in \mathbb{Q}\) with \(|y_n - z_n| < \frac{\epsilon}{2}\) for all \(n > K_2\). Set \(K = \max(K_1, K_2)\). If \(n > K\), then by the triangle inequality

\[
|x_n - z_n| \leq |x_n - y_n| + |y_n - z_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]

If \((x_n)\) and \((y_n)\) are sequences and \(x_n = y_n\) for all but finitely many \(n\), then \((x_n) \sim (y_n)\): take \(K = \max\{n \in \mathbb{N} : x_n \neq y_n\}\).

Equivalent sequences converge to the same limit, and conversely:

III.5.2.8. Proposition. Let \((x_n)\) and \((y_n)\) be sequences in \(\mathbb{Q}\). If \((x_n) \sim (y_n)\) and \(x_n \to x\), then \(y_n \to x\). Conversely, if \(x_n \to x\) and \(y_n \to x\), then \((x_n) \sim (y_n)\).

Proof: Let \(\epsilon \in \mathbb{Q}_+\). There is a \(K_1 \in \mathbb{Q}\) such that \(|x_n - x| < \frac{\epsilon}{2}\) for all \(n > K_1\), and there is a \(K_2 \in \mathbb{Q}\) such that \(|x_n - y_n| < \frac{\epsilon}{2}\) for all \(n > K_2\). Set \(K = \max(K_1, K_2)\). If \(n > K\), then

\[
|y_n - x| \leq |y_n - x_n| + |x_n - x| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]

Conversely, suppose \(y_n \to x\). Let \(\epsilon \in \mathbb{Q}_+\), and let \(K_1\) be as above. There is also a \(K_2\) such that \(|y_n - x| < \frac{\epsilon}{2}\) for all \(n > K_3\). Set \(K' = \max(K_1, K_3)\). If \(n > K'\), then

\[
|x_n - y_n| \leq |x_n - x| + |x - y_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]

Since a constant sequence clearly converges to its constant value \(\), we obtain:
III.5.2.9. **Corollary.** Let \((x_n)\) be a sequence of rational numbers. Then \((x_n)\) converges to the rational number \(r\) if and only if \((x_n)\) is equivalent to the constant sequence \((r)\).

A variation of III.5.2.8. for Cauchy sequences is useful:

III.5.2.10. **Proposition.** Let \((x_n)\) and \((y_n)\) be sequences in \(\mathbb{Q}\). If \((x_n) \sim (y_n)\) and \((x_n)\) is a Cauchy sequence, then \((y_n)\) is also a Cauchy sequence.

**Proof:** Let \(\epsilon \in \mathbb{Q}_+\). There is a \(K_1 \in \mathbb{Q}\) such that \(|x_n - x_m| < \frac{\epsilon}{3}\) for all \(n, m > K_1\), and there is a \(K_2 \in \mathbb{Q}\) such that \(|x_n - y_n| < \frac{\epsilon}{3}\) for all \(n > K_2\). Set \(K = \max(K_1, K_2)\). If \(n, m > K\), then
\[
|y_n - y_m| \leq |y_n - x_n| + |x_n - x_m| + |x_m - y_m| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.
\]

We now make our main definition:

III.5.2.11. **Definition.** Let \(\mathbb{R}\) be the set of equivalence classes of Cauchy sequences of rational numbers. Denote the class of a Cauchy sequence \((x_n)\) by \([ (x_n) ]\). Denote by \(\mathbb{Q}\) the set of equivalence classes of constant sequences.

There is an obvious one-one correspondence between \(\mathbb{Q}\) and \(\mathbb{Q}\).

**Order Structure of \(\mathbb{R}\)**

The key to the order structure of \(\mathbb{R}\) is the next result:

III.5.2.12. **Proposition.** Let \((x_n)\) and \((y_n)\) be Cauchy sequences of rational numbers. If \((x_n)\) and \((y_n)\) are not equivalent, then there are \(r, s \in \mathbb{Q}\) with \(r < s\), and \(K \in \mathbb{Q}\), such that either \(x_n < r\) and \(y_n > s\) for all \(n > K\) or \(y_n < r\) and \(x_n > s\) for all \(n > K\).

**Proof:** If \((x_n)\) and \((y_n)\) are not equivalent, then there is an \(\epsilon \in \mathbb{Q}_+\) such that, for every \(K \in \mathbb{Q}\), there is an \(n > K\) with \(|x_n - y_n| \geq \epsilon\). For this \(\epsilon\), fix \(K\) such that \(|x_n - x_m| < \frac{\epsilon}{3}\) and \(|y_n - y_m| < \frac{\epsilon}{3}\) for all \(n, m > K\). Fix \(m > K\) with \(|x_m - y_m| \geq \epsilon\). Then either \(x_m - y_m \geq \epsilon\) or \(y_m - x_m \geq \epsilon\). If \(y_m - x_m \geq \epsilon\), set \(r = x_m + \frac{\epsilon}{3}\) and \(s = y_m - \frac{\epsilon}{3}\). Then \(r < s\), and for any \(n > K\) we have \(x_n < r\) and \(y_n > s\). A similar argument shows that if \(x_m - y_m \geq \epsilon\), and \(r = y_m + \frac{\epsilon}{3}\), \(s = x_m - \frac{\epsilon}{3}\), then \(y_n < r < s < x_n\) for all \(n > K\).

III.5.2.13. **Proposition.** Let \((x_n)\) and \((y_n)\) be sequences in \(\mathbb{Q}\), and \(r, s \in \mathbb{Q}\), \(K \in \mathbb{Q}\) with \(x_n < r < s < y_n\) for all \(n > K\). If \((z_n)\) and \((w_n)\) are sequences in \(\mathbb{Q}\) with \((z_n) \sim (x_n)\) and \((w_n) \sim (y_n)\), and \(r', s' \in \mathbb{Q}\) with \(r < r' < s' < s\), then there is a \(K' \in \mathbb{Q}\) such that \(z_n < r'\) and \(w_n > s'\) for all \(n > K'\).

**Proof:** There is a \(K_1 \in \mathbb{Q}\) such that \(|x_n - z_n| < r' - r\) for all \(n > K_1\), and \(K_2 \in \mathbb{Q}\) such that \(|y_n - w_n| < s - s'\) for all \(n > K_2\). Set \(K' = \max(K_1, K_2, K)\). If \(n > K'\), we have \(z_n = x_n + (z_n - x_n) < r + (r' - r) = r'\), and similarly \(w_n = y_n - (y_n - w_n) > s - (s - s') = s'\).

We can now define the order on \(\mathbb{R}\):
III.5.2.14. Definition. Let \([x_n], [y_n] \in \mathbb{R}\). Then \([x_n]\) < \([y_n]\) if there are \(r, s \in \mathbb{Q}\), \(r < s\), and \(K \in \mathbb{Q}\) such that \(x_n < r\) and \(y_n > s\) for all \(n > K\).

By III.5.2.13, this order is well defined on equivalence classes. It is clearly transitive and antisymmetric. By III.5.2.12, it is a total order, i.e. trichotomy holds. As usual, write \([x_n] \leq [y_n]\) if \([x_n] < [y_n]\) or \([x_n] = [y_n]\), and define > and ≥ accordingly. The induced ordering on \(\mathbb{Q}\) is just the usual ordering on \(\mathbb{Q}\), i.e. if \(r, s \in \mathbb{Q}\), then \([r]\) < \([s]\) if and only if \(r < s\).

III.5.2.15. Proposition. Let \((x_n), (y_n)\) be Cauchy sequences in \(\mathbb{Q}\). If there is a \(K \in \mathbb{Q}\) with \(x_n \leq y_n\) for all \(n > K\), then \([x_n]\) ≤ \([y_n]\).

Proof: Under the hypotheses, \([x_n]\) > \([y_n]\) is impossible!

III.5.2.16. The converse is also true by definition of the order. In fact, if \([x_n]\) < \([y_n]\), then there are Cauchy sequences \((z_n), (w_n)\) in \(\mathbb{Q}\), with \((x_n) \sim (z_n)\) and \((y_n) \sim (w_n)\), such that \(z_n < w_n\) for all \(n\): just take \(z_n = x_n\) and \(w_n = y_n\) if \(x_n < y_n\), and \(z_n = 0, w_n = 1\) if \(x_n \geq y_n\). Then \(z_n = x_n\) and \(w_n = y_n\) for all but finitely many \(n\). The statement also clearly holds with < replaced by ≤ throughout.

We then obtain density of \(\mathbb{Q}\) in \(\mathbb{R}\):

III.5.2.17. Proposition. \(\mathbb{Q}\) is order-dense in \(\mathbb{R}\), i.e. if \([x_n]\) < \([y_n]\), then there is a \(t \in \mathbb{Q}\) with \([x_n]\) < \([t]\) < \([y_n]\).

Proof: Let \(r, s \in \mathbb{Q}\) with \(r < s\), and \(K \in \mathbb{Q}\), such that \(x_n < r\) and \(y_n > s\) for all \(n > K\). Then \([x_n]\) ≤ \([r]\) < \([s]\) ≤ \([y_n]\) by III.5.2.15. If \(t = (r + s)/2\), we have

\[\[x_n]\] ≤ \([r]\) < \([t]\) < \([s]\) ≤ \([y_n]\) .

Algebraic Structure of \(\mathbb{R}\)

We would like to define addition and multiplication of equivalence classes of Cauchy sequences coordinatewise. The next proposition shows that addition is well defined in this way:

III.5.2.18. Proposition. Let \((x_n), (y_n), (z_n), (w_n)\) be sequences in \(\mathbb{Q}\).

(i) If \((x_n)\) and \((y_n)\) are Cauchy sequences, so is \((x_n + y_n)\).

(ii) If \((x_n) \sim (z_n)\) and \((y_n) \sim (w_n)\), then \((x_n + y_n) \sim (z_n + w_n)\).

Proof: The proof is similar to previous ones.
(i): Let $\epsilon > 0$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \frac{\epsilon}{2}$ for all $n, m > K_1$ and $K_2 \in \mathbb{Q}$ such that $|y_n - y_m| < \frac{\epsilon}{2}$ for all $n, m > K_2$. Set $K = \max(K_1, K_2)$. If $n, m > K$, then
\[|(x_n + y_n) - (x_m + y_m)| = |(x_n - x_m) + (y_n - y_m)| \leq |x_n - x_m| + |y_n - y_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon .\]

(ii): Let $\epsilon > 0$. There is a $K_3 \in \mathbb{Q}$ such that $|x_n - z_n| < \frac{\epsilon}{2}$ for all $n > K_3$ and $K_4 \in \mathbb{Q}$ such that $|y_n - w_n| < \frac{\epsilon}{2}$ for all $n > K_4$. Set $K' = \max(K_3, K_4)$. If $n > K'$, then
\[|(x_n + y_n) - (z_n + w_n)| = |(x_n - z_n) + (y_n - w_n)| \leq |x_n - z_n| + |y_n - w_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon .\]

\[\square\]

Dealing with products is a little harder. We first need a result of general interest. A sequence $(x_n)$ in $\mathbb{Q}$ is bounded if there is an $M \in \mathbb{Q}$ with $|x_n| \leq M$ for all $n \in \mathbb{N}$.

III.5.2.19.  **Proposition.** Every Cauchy sequence in $\mathbb{Q}$ is bounded.

**Proof:** Let $(x_n)$ be a Cauchy sequence. Fix $K \in \mathbb{Q}$ such that $|x_n - x_m| < 1$ for all $n, m > K$, and fix $n > K$. Then, if $m > n$, we have $|x_m| \leq |x_n| + |x_m - x_n| < |x_n| + 1$. The set $\{|x_k|: 1 \leq k < n\}$ is finite, hence has a maximum $M'$. Set $M = \max(M', |x_n| + 1)$. Then $|x_k| \leq M$ for all $k \in \mathbb{N}$.

\[\square\]

III.5.2.20.  **Proposition.** Let $(x_n), (y_n), (z_n), (w_n)$ be sequences in $\mathbb{Q}$.

(i) If $(x_n)$ and $(y_n)$ are Cauchy sequences, so is $(x_n y_n)$.

(ii) If $(x_n) \sim (z_n)$ and $(y_n) \sim (w_n)$, then $(x_n y_n) \sim (z_n w_n)$.

**Proof:** By III.5.2.19, there are $M, M', M'' \in \mathbb{Q}$ such that $|x_n| \leq M$, $|y_n| \leq M'$, $|z_n| \leq M''$ for all $n$. By increasing $M$, $M'$, and $M''$ if necessary, we may assume $M, M', M'' > 0$.

(i): Let $\epsilon > 0$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \frac{\epsilon}{2M''}$ for all $n, m > K_1$ and $K_2 \in \mathbb{Q}$ such that $|y_n - y_m| < \frac{\epsilon}{2M''}$ for all $n, m > K_2$. Set $K = \max(K_1, K_2)$. If $n, m > K$, then
\[|x_n y_n - x_m y_m| = |x_n y_n - x_m y_n + x_m y_n - x_m y_m| \leq |x_n y_n - x_m y_n| + |x_m y_n - x_m y_m| \]
\[= |y_n||x_n - x_m| + |x_m||y_n - y_m| < M' \frac{\epsilon}{2M''} + M' \frac{\epsilon}{2M''} = \epsilon .\]

(ii): Let $\epsilon > 0$. There is a $K_3 \in \mathbb{Q}$ such that $|x_n - z_n| < \frac{\epsilon}{2M''}$ for all $n > K_3$ and $K_4 \in \mathbb{Q}$ such that $|y_n - w_n| < \frac{\epsilon}{2M''}$ for all $n > K_4$. Set $K' = \max(K_3, K_4)$. If $n > K'$, then
\[|x_n y_n - z_n w_n| = |x_n y_n - z_n y_n + z_n y_n - z_n w_n| \leq |x_n y_n - z_n y_n| + |z_n y_n - z_n w_n| \]
\[= |y_n||x_n - z_n| + |z_n||y_n - w_n| < M' \frac{\epsilon}{2M''} + M'' \frac{\epsilon}{2M''} = \epsilon .\]

\[\square\]

As a result, we may make the following definition:
III.5.2.21. Definition. Define binary operations $+$ and $\cdot$ on $\mathbb{R}$ by $[(x_n)] + [(y_n)] = [(x_n + y_n)]$ and $[(x_n)] \cdot [(y_n)] = [(x_n y_n)]$.

III.5.2.22. Associativity of addition on $\mathbb{R}$ is immediate, since if $(x_n)$, $(y_n)$, and $(z_n)$ are Cauchy sequences in $\mathbb{Q}$, the sequences $(x_n + y_n + z_n)$ and $(x_n + [y_n + z_n])$ are not only equivalent but even identical. A similar observation shows that associativity of multiplication, commutativity of addition and multiplication, and the distributive law hold in $\mathbb{R}$.

We also have that $[(0)]$ is an additive identity and $[(1)]$ is a multiplicative identity. If $(x_n)$ is a Cauchy sequence, then so is $(-x_n)$, and $[(-x_n)]$ is an additive inverse for $[(x_n)]$.

It is obvious that $\mathbb{Q}$ is closed under $+$ and $\cdot$, and the restrictions are the usual operations from $\mathbb{Q}$, i.e. we have $[(r)] + [(s)] = [(r + s)]$ and $[(r)] \cdot [(s)] = [(rs)]$ for $r, s \in \mathbb{Q}$.

There are also multiplicative inverses for nonzero elements of $\mathbb{R}$. First note the next immediate corollary of III.5.2.12:

III.5.2.23. Corollary. Let $(x_n)$ be a Cauchy sequence in $\mathbb{Q}$. If $[(x_n)] \neq [(0)]$, then there are $\alpha \in \mathbb{Q}_+$ and $K \in \mathbb{Q}$ with $|x_n| \geq \alpha$ for all $n > K$.

III.5.2.24. Proposition. Every nonzero element of $\mathbb{R}$ has a multiplicative inverse.

Proof: Let $(x_n)$ be a Cauchy sequence in $\mathbb{Q}$ with $[(x_n)] \neq [(0)]$, and fix $\alpha \in \mathbb{Q}_+$ and $K \in \mathbb{Q}$ such that $|x_n| \geq \alpha$ for all $n \in \mathbb{K}$. Set

$$y_n = \begin{cases} x_n^{-1} & \text{if } x_n \neq 0 \\ 1 & \text{if } x_n = 0 \end{cases}.$$

We show $(y_n)$ is a Cauchy sequence. Let $\epsilon \in \mathbb{Q}_+$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \alpha^2 \epsilon$ for all $n > K_1$. If $K' = \max(K, K_1)$, and $n, m > K'$, then $|x_n|, |x_m| \geq \alpha$, so $x_n, x_m \neq 0$ and $y_n = \frac{1}{x_n}$, $y_m = \frac{1}{x_m}$, and

$$|y_n - y_m| = \left| \frac{1}{x_n} - \frac{1}{x_m} \right| = \frac{|x_m - x_n|}{|x_m x_n|} \leq \frac{|x_m - x_n|}{\alpha^2} < \epsilon.$$

So $\mathbb{R}$ is a field.

III.5.2.25. We now check that $\mathbb{R}$ is an ordered field. It is easiest to check (Pos). Let $\mathbf{P} = \{(x_n) \in \mathbb{R} : [(x_n)] > [(0)]\}$. We have (P3) and (P4), i.e. that $\mathbb{R}$ is the disjoint union of $\mathbf{P}$, $\{(0)\}$, and $-\mathbf{P}$, so we need only check that $\mathbf{P}$ is closed under addition and multiplication. Let $[(x_n)], [(y_n)] \in \mathbf{P}$, then there are $\alpha, \beta \in \mathbb{Q}_+$ and $K_1, K_2 \in \mathbb{Q}$ such that $x_n \geq \alpha$ for all $n > K_1$ and $y_n \geq \beta$ for all $n > K_2$. Set $K = \max(K_1, K_2)$. If $n > K$, then $x_n + y_n \geq \alpha + \beta \in \mathbb{Q}_+$ and $x_n y_n \geq \alpha \beta \in \mathbb{Q}_+$. So $[(x_n + y_n)]$ and $[(x_n y_n)]$ are in $\mathbf{P}$.

Summarizing, we have

III.5.2.26. Theorem. $\mathbb{R}$ is an ordered field under $+$, $\cdot$, and $<$, with additive identity $[(0)]$ and multiplicative identity $[(1)]$. $\mathbb{Q}$ is an order-dense subfield order-isomorphic to $\mathbb{Q}$.
The Completeness Axiom

Showing that \( \mathbb{R} \) satisfies the Completeness Axiom takes some work. Note that the order-theoretic completeness of the Completeness Axiom is not the same thing as metric completeness, which asserts that every Cauchy sequence converges. For metric completeness of \( \mathbb{R} \), see (a).

III.5.2.27. Theorem. \( (\mathbb{R}, <) \) satisfies the Completeness Axiom, i.e. every nonempty subset of \( \mathbb{R} \) which is bounded above has a supremum.

Proof: Let \( S \) be a nonempty bounded subset of \( \mathbb{R} \). We will construct two equivalent Cauchy sequences in \( \mathbb{Q} \) whose equivalence class is \( \sup(S) \) by a bisection process. First fix \( a, b \in \mathbb{Q} \) such that \( b \) (actually \( [(b)] \)) is an upper bound for \( S \), and \( a \) is not an upper bound. (For \( r \in \mathbb{Q} \) we will write \( r \) for \( \lfloor r \rfloor \) to streamline notation, except when there is possibility of confusion.)

Set \( c_1 = (a + b)/2 \). Either \( c_1 \) is an upper bound for \( S \), or it is not. If \( c_1 \) is an upper bound for \( S \), set \( y_1 = a \) and \( z_1 = c_1 \); otherwise set \( y_1 = c_1 \) and \( z_1 = b \). Then \( y_1 \) is not an upper bound for \( S \), \( z_1 \) is, and \( z_1 - y_1 = 2^{-1}(b - a) \). Repeat the process by setting \( c_2 = (y_1 + z_1)/2 \) and setting \( y_2 = y_1, z_2 = c_2 \) if \( c_2 \) is an upper bound for \( S \) and \( y_2 = c_2, z_2 = z_1 \) otherwise. Iterating the process, we obtain sequences \( (y_n), (z_n) \) in \( \mathbb{Q} \) such that \( y_n \leq y_{n+1} < z_{n+1} \leq z_n \) and \( z_n - y_n = 2^{-n}(b - a) \).

To show that \( (y_n) \) and \( (z_n) \) are Cauchy sequences, and \( (y_n) \sim (z_n) \), let \( \epsilon > 0 \), and fix \( K \in \mathbb{N} \) such that \( 2^{-K}(b - a) \leq \epsilon \). Then, if \( n, m > K, n < m \), we have \( y_K \leq y_n \leq y_m < z_m \leq z_n \leq z_K \), with at least one of the outside inequalities strict. Thus \( |y_n - y_m|, |z_n - z_m|, |y_n - z_n| \) are all less than \( z_K - y_K = 2^{-K}(b - a) \leq \epsilon \).

Thus \( [(y_n)] = [(z_n)] \in \mathbb{R} \). Call this element \( y \). We will show that \( y = \sup(S) \). First suppose \( [(x_n)] \in S \). If \( [(x_n)] > y = [(z_n)] \), then by III.5.2.12. there are \( r, s \in \mathbb{Q} \) with \( r < s \), and \( K_1 \in \mathbb{Q} \), such that \( z_n < r \) and \( x_n > s \) for all \( n > K_1 \). Fix \( m > K_1 \) and set \( u = z_m \). Since the \( z_n \) are decreasing, we have \( z_n \leq u \) for all \( n > m \), and thus \( y \leq [(u)] \). We also have \( x_n > s \) for all \( n > m \) and \( u < s \), so \( [(u)] < [(x_n)] \). But by construction \( u \) (actually \( [(u)] \)) is an upper bound for \( S \), a contradiction. Thus \( [(x_n)] \leq y \), and \( y \) is an upper bound for \( S \).

Now suppose \( [(w_n)] \in \mathbb{R} \) with \( [(w_n)] < y = [(y_n)] \). Then there are \( r, s \in \mathbb{Q}, r < s \), and \( K_2 \in \mathbb{Q} \), such that \( w_n < r \) and \( y_n > s \) for all \( n > K_2 \). Fix \( m > K_2 \), and set \( v = y_m \). By construction, \( v \) (actually \( [(v)] \)) is not an upper bound for \( S \), so there is an \( [(x_n)] \in S \) with \( [(v)] < [(x_n)] \), so there is a \( K_3 \in \mathbb{Q} \) with \( x_n > v \) for all \( n > K_3 \). Since \( v > s \), we have \( w_n < r \) and \( s < x_n \) for all \( n > \max(K_2, K_3) \), so \( [(w_n)] < [(x_n)] \), and \( [(w_n)] \) is not an upper bound for \( S \). Thus \( y \) is the least upper bound of \( S \).

The bottom line is:

III.5.2.28. Theorem. \( \mathbb{R} \), with its addition, multiplication, and order, is a model for \( \mathbb{R} \).

III.5.3. Exercises

III.5.3.1. Let \( \mathbb{R}(X) \) be the field of rational “functions” of Exercise III.4.3.5.(c). If \( 0 \neq f(X) = \frac{p(X)}{q(X)} \in \mathbb{R}(X) \), we say \( f(X) \) is positive if the leading coefficients (coefficients of the highest powers of \( X \)) in \( p(X) \) and \( q(X) \) have the same sign. Let \( P \) be the set of positive elements.

(a) Show that the property of \( f(X) \) being positive is independent of the way \( f(X) \) is represented as a fraction \( \frac{p(X)}{q(X)} \). If \( f \) is regarded as a function from \( \mathbb{R} \setminus \{ x : q(x) = 0 \} \) to \( \mathbb{R} \), show that \( f \) is positive if and only if
f(x) > 0 for all sufficiently large \( x \in \mathbb{R} \).
Show that a sum, product, or quotient of positive elements is positive.

(b) Show that \( \mathbb{R}(X) \) is the disjoint union of \( P, \{0\}, \) and \( -P; \) thus \( (\mathbb{R}(X), P) \) is an ordered field. (This ordering is called the Fréchet ordering on \( \mathbb{R}(X) \).)

(c) If \( \mathbb{R} \) is regarded as the subfield of \( \mathbb{R}(X) \) of constant functions, the order on \( \mathbb{R}(X) \) induces the usual order on \( \mathbb{R} \).

(d) The element \( f(X) = X \) is a positive element which is infinitely large, i.e. \( f > r \) for every \( r \in \mathbb{R} \). The element \( g(X) = \frac{1}{x} \) is a positive element which is infinitesimal, i.e. \( 0 < g < \epsilon \) for all \( \epsilon \in \mathbb{R}, \epsilon > 0 \). Thus \( \mathbb{R}(X) \) is a nonarchimedean ordered field.

III.5.3.2. Formal Power Series. Let \( F \) be a field, e.g. \( F = \mathbb{Q}, \mathbb{R}, \) or \( \mathbb{C} \) (in fact the construction works over any ring).

(a) A formal power series over \( F \) is an expression of the form \( \sum_{k=0}^{\infty} a_k X^k \), where the \( a_k \) are in \( F \) and \( X \) is an indeterminate symbol. Define addition and multiplication of power series by

\[
\left( \sum_{k=0}^{\infty} a_k X^k \right) + \left( \sum_{k=0}^{\infty} b_k X^k \right) = \sum_{k=0}^{\infty} \left( a_k + b_k \right) X^k \\
\left( \sum_{k=0}^{\infty} a_k X^k \right) \cdot \left( \sum_{k=0}^{\infty} b_k X^k \right) = \sum_{k=0}^{\infty} \left( \sum_{j=0}^{k} c_k X^k \right), \text{ where } c_k = \sum_{j=0}^{k} a_j b_{k-j}.
\]

Let \( F[[X]] \) be the set of all formal power series over \( F \). Show that \( F[[X]] \) is a commutative ring under these operations, and that the polynomial ring \( F[X] \) is a subring. In particular, \( F \) can be identified with the set of “constant” power series with \( a_k = 0 \) for \( k > 0 \).

(b) Show that a power series \( \sum_{k=0}^{\infty} a_k X^k \) has a multiplicative inverse in \( F[[X]] \) if and only if \( a_0 \neq 0 \). [Generate the coefficients of the multiplicative inverse series \( \sum_{k=0}^{\infty} b_k X^k \) by successively solving a sequence of linear equations in \( F \).]

The multiplication formula for formal power series can be motivated by writing the series out in informal notation

\[
\sum_{k=0}^{\infty} a_k X^k = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots
\]

and similarly for \( \sum_{k=0}^{\infty} b_k X^k \), and then formally multiplying out

\[
(a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots)(b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \cdots)
\]

and collecting together like powers of \( X \).

The elements of \( F[[X]] \) are called “formal power series” since there is no notion of convergence implied (in fact, over a general field, the idea of convergence of a power series does not even make sense). So formal power series are just algebraic expressions and should not be thought of as corresponding to “functions”. For example, the series \( \sum_{k=0}^{\infty} k! X^k \) is a perfectly good formal power series over \( \mathbb{Q} \) (or \( \mathbb{R}, \mathbb{C} \)), even though it diverges when any nonzero complex number is substituted for \( X \), and hence cannot be thought of as giving a function on \( \mathbb{C} \) (\( \mathbb{R}, \mathbb{Q} \)). Later, we will consider rings of convergent power series, where ideas from analysis come into the picture in an essential way.
III.5.3.3.  **Formal Laurent Series.** This problem is somewhat of a continuation of Problem III.5.3.2.
Let $F$ be a field. A **formal Laurent series** over $F$ is a sum of the form $\sum_{k=m}^{\infty} a_k X^k$ with the $a_k$ in $F$, where $m \in \mathbb{Z}$ depends on the series, and may be positive, negative, or zero. We will often implicitly assume when writing such an expression that $a_m \neq 0$, although this is not necessary. We will informally write such a series as a formal sum of monomials

$$\sum_{k=m}^{\infty} a_k X^k = a_m X^m + a_{m+1} X^{m+1} + \cdots$$

where only finitely many negative powers of $X$ (perhaps none) occur.

(a) Define addition and multiplication of formal Laurent series as in Problem III.5.3.2. [The formula for multiplication is more complicated, but can be found by considering the informal term-by-term product]

$$(a_m X^m + a_{m+1} X^{m+1} + \ldots)(b_n X^n + b_{n+1} X^{n+1} + \cdots).$$

Show that the set $F((X))$ of formal Laurent series is a commutative ring with this addition and multiplication, and the ring of formal power series $F[[X]]$ is a subring.

(b) Show that every nonzero element of $F((X))$ has a multiplicative inverse. [Use the method of problem III.5.3.2.(b). If the series is $\sum_{k=m}^{\infty} a_k X^k$ with $a_m \neq 0$, generate a series $\sum_{k=-m}^{\infty} b_k X^k$.] Thus $F((X))$ is a field.

(c) Show that $F((X))$ is naturally isomorphic to the quotient field ($()$ of $F[[X]]$. In particular, there is a canonical embedding of the field $F(X)$ of rational functions into $F((X))$, although the formula for the embedding is not obvious: the multiplicative inverse of a polynomial is usually not a polynomial or even a Laurent polynomial (Laurent series with only finitely many nonzero terms).

(d) Suppose $F$ is an ordered field. Declare a nonzero power series $\sum_{k=m}^{\infty} a_k X^k$ with $a_m \neq 0$ to be **positive** if $a_m > 0$. Let $P$ be the set of all positive series. Show that $P$ satisfies (Pos) and hence makes $F((X))$ into an ordered field. Show that $F((X))$ is nonarchimedean: we have $X > 0$, but $nX < 1$ for all $n \in \mathbb{N}$ (identified with a subset of $F$ as in ()), so $X$ is an **infinitesimal** positive element of $F((X))$.

(e) The ordering on $F(X)$ induced by its embedding in $F((X))$ from (c) is not the same as the ordering from Problem III.5.3.1. But show it can be obtained from this ordering by the substitution $X \mapsto \frac{1}{X}$ in $F(X)$.

III.5.3.4.  **Puiseux Series.** This and the next two problems are a continuation of Problem III.5.3.3.; we will define three successively larger fields containing the field $\mathbb{R}((X))$. We will work over $\mathbb{R}$, although the construction can be done over any field.

A **Puiseux Series** is a formal expression of the form $\sum_{k=m}^{\infty} a_k/n X^{k/n}$ for some $m \in \mathbb{Z}$, $n \in \mathbb{N}$, where $a_k/n \in \mathbb{R}$ for all $k$. In other words, a Puiseux series is a formal Laurent series in $X^{1/n}$ for some $n$ (which may vary from series to series). Any formal Laurent series over $\mathbb{R}$ is a Puiseux series with $n = 1$. We identify $X^{k/n}$ and $X^{kr/nr}$ for any $r \in \mathbb{N}$; thus any two (or any finite number of) Puiseux series can be written as formal Laurent series in $X^{1/n}$ for a common $n$.

(a) Define addition and multiplication of Puiseux series in analogy with the definition for formal Laurent series. Show that the set $P$ of Puiseux series is a field containing $\mathbb{R}((X))$ as a subfield. In particular, $P$ contains $\mathbb{R}$ as the subfield of “constant” series, where $a_{k/n} = 0$ for all $k \neq 0$.

(b) Define an ordering on $P$ by defining the positive cone as in Exercise III.5.3.3.(d): If $x = \sum_{k=m}^{\infty} a_k/n X^{k/n}$ is a nonzero element of $P$, with $m/n$ the smallest $k/n$ for which $a_{k/n} \neq 0$ ($m/n$ is called the **degree** of the
series), set \( x > 0 \) if \( a_{m/n} > 0 \). This ordering agrees with the previous one from Exercise III.5.3.3.(d) on \( \mathbb{R}((X)) \).

(c) Show by direct computation that every nonnegative element of \( P \) has a square root. [If \( x \in P \) with degree \( m/n \) with \( m \geq 0 \), write \( x = a_{m/n}X^{m/n}(1 + y) \) where \( a_{m/n} > 0 \) and \( y \) has degree \( > 0 \), and use the binomial series for \((1 + y)^{1/2}\).

In fact, \( P \) is a real-closed field, so is a first-order model for \( \mathbb{R} \) (and is the “smallest” such model containing an infinitesimal). Puiseux series were first formally studied by Puiseux in about 1850, although many of the basic ideas go back to Newton. See [BPR06, Chapter 2] for more information.

III.5.3.5. The Levi-Civita Numbers. This is a continuation of Problem III.5.3.4. We will work over \( \mathbb{R} \), although the construction can be done over any field. A Levi-Civita number is a formal expression of the form \( \sum_{q \in \mathbb{Q}} a_q X^q \), where \( a_q \in \mathbb{R} \) for all \( q \), and for any such sum the set \( S = \{q : a_q \neq 0\} \) is a left-finite set, i.e. for any \( r \in \mathbb{Q} \) the set \( \{q \in S : q < r\} \) is finite. In other words, either \( S \) is finite or the elements of \( S \) can be arranged in a strictly increasing sequence converging to \(+\infty\). (The set \( S \) may vary from series to series.) Any Puiseux series over \( \mathbb{R} \) is a Levi-Civita number.

(a) Define addition and multiplication of Levi-Civita numbers in analogy with the definition for Puiseux series. Show that the set \( \mathcal{R} \) of Levi-Civita numbers is a field containing \( P \) as a subfield.

(b) Define an ordering on \( \mathcal{R} \) by defining the positive cone as in Exercise III.5.3.3.(d): If \( x = \sum_{q \in \mathbb{Q}} a_q X^q \) is a nonzero element of \( \mathcal{R} \), with \( q_0 \) the smallest \( q \) for which \( a_q \neq 0 \) (\( q_0 \) is called the degree of the series), set \( x > 0 \) if \( a_{q_0} > 0 \). This ordering agrees with the previous one from Exercise III.5.3.4. on \( P \).

(c) Show by direct computation that every nonnegative element of \( \mathcal{R} \) has a square root. [The argument is similar to the one in III.5.3.4.(c).]

(d) Show that \( \mathcal{R} \) is complete in the following sense:

In fact, \( \mathcal{R} \) is a real-closed field, so is a first-order model for \( \mathbb{R} \) (and is the “smallest” such model which is complete in the sense of (d) and contains an infinitesimal). The Levi-Civita numbers are important in computational analysis. See [?], [SB10], [Rib92], or [DW96, Theorem 2.15] for more information.

III.5.3.6. Hahn Series. This is a further continuation of Problem III.5.3.5. We will work over \( \mathbb{R} \), although the construction can be done over any field. Let \( G \) be a totally ordered abelian group, e.g. \( \mathbb{Q} \) or \( \mathbb{R} \). A Hahn Series over \( G \) is a formal expression of the form \( \sum_{q \in G} a_q X^q \), where \( a_q \in \mathbb{R} \) for all \( q \), and for any such sum the set \( S = \{q : a_q \neq 0\} \) is a well-ordered set, i.e. every subset of \( S \) has a smallest element (equivalently, \( S \) contains no strictly decreasing sequences). The set \( S \) may vary from series to series. Every Levi-Civita number is a Hahn series over \( \mathbb{Q} \) (but there are Hahn series over \( \mathbb{Q} \) which are not Levi-Civita numbers): a formal Laurent series is a Hahn series over \( \mathbb{Z} \).

(a) Define addition and multiplication of Hahn series in analogy with the definition for Levi-Civita numbers. Show that the set \( \mathcal{S}_G \) of Hahn series over \( G \) is a field.

(b) Define an ordering on \( \mathcal{S}_G \) by defining the positive cone as in Exercise III.5.3.3.(d): If \( x = \sum_{q \in G} a_q X^q \) is a nonzero element of \( \mathcal{S}_G \), with \( q_0 \) the smallest \( q \) for which \( a_q \neq 0 \) (\( q_0 \) is called the degree of the series), set \( x > 0 \) if \( a_{q_0} > 0 \). This ordering agrees with the previous one from Exercise III.5.3.3.(d) on \( \mathbb{R}((X)) \).

(c) Show by direct computation that if \( G \) is 2-divisible, every nonnegative element of \( \mathcal{S}_G \) has a square root. [The argument is similar to the one in III.5.3.4.(c).]

(d) Show that \( \mathcal{S} \) is complete in the sense of III.5.3.5.(d).

In fact, \( \mathcal{S}_G \) is also a real-closed field if \( G \) is divisible. See [DW96] for more information.
This problem is a continuation of III.5.3.2. A formal power series $\sum_{k=0}^{\infty} a_k X^k$ has order $m$ if $a_m \neq 0$ and $a_k = 0$ for $k < m$.

We will suggestively use letters $f, g, h, \ldots$ for elements of $F[[X]]$, or even more suggestively $f(X)$, etc., even though formal power series are not functions in any literal sense.

(a) Let $(f_n)$ be a sequence of formal power series over $F$. Show that if for every $m$ there are only finitely many $f_n$ of order $\leq m$, then the infinite sum $\sum_{n=1}^{\infty} f_n$ is well defined in $F[[X]]$. [Add the coefficients of each power of $X$.]

(b) Give another proof of III.5.3.2. (b) by using the formal infinite series

$$(1 - x)^{-1} = \sum_{n=0}^{\infty} x^n.$$ 

(c) If $f = \sum_{k=0}^{\infty} a_k X^k$ and $g = \sum_{k=0}^{\infty} b_k X^k$ are formal power series over $F$, and $g$ has order $\geq 1$ (i.e. $b_0 = 0$), then the infinite sum

$$\sum_{k=0}^{\infty} a_k g^k = \sum_{k=0}^{\infty} a_k [g(X)]^k$$

is a well-defined element of $F[[X]]$ which can be denoted $f(g(X))$ or $f \circ g$. [The order of $g^n$ is at least $n$.] There is not a well-defined composition if $b_0 \neq 0$.

This can be motivated by regarding formal power series as defining “function germs” around $0 \in F$. It makes sense that composition can only be defined if the inside function sends $0$ to $0$.

(d) Prove that composition, when defined, is associative: if $g$ and $h$ have order $\geq 1$, then $g \circ h$ has order $\geq 1$, and for every $f$ we have $(f \circ g) \circ h = f \circ (g \circ h)$.

(e) Prove the following “Formal Inverse Function Theorem” (where $\iota$ is the “identity function” $X$, i.e. $\iota = \sum_{k=0}^{\infty} a_k X^k$, where $a_1 = 1$ and $a_k = 0$ for $k \neq 1$) by successively solving a sequence of linear equations for the coefficients of $g$:

**Theorem.** Let $f = \sum_{k=0}^{\infty} a_k X^k \in F[[X]]$. Then there exists $g \in F[[X]]$ with $f \circ g = \iota$ if and only if $a_0 = 0$ and $a_1 \neq 0$. If such $g$ exists, it is unique, and also $g \circ f = \iota$.

If $f = \sum_{k=0}^{\infty} a_k X^k \in F[[X]]$, we may write $f(0) = a_0$ and $f'(0) = a_1$ (there is a difficulty carrying this over to higher derivatives unless $F$ has characteristic 0 since a factorial should occur in the formula). Then, if the $g$ exists in the theorem, we have $g(0) = 0$ and $g'(0) = [f'(0)]^{-1}$.

This problem is a continuation of III.5.3.2. and III.5.3.7. A formal power series in two variables $X$ and $Y$ over a field $F$ is a formal sum

$$\sum_{k,j=0}^{\infty} a_{k,j} X^k Y^j$$

where the $a_{k,j}$ are elements of $F$. There is an obvious way of adding and multiplying two such series, making the set $F[[X,Y]]$ of such formal power series into a ring.

(a) Show that there are natural isomorphisms

$$F[[X,Y]] \cong (F[[X]])[[Y]] \cong (F[[Y]])[[X]].$$
In particular, $F[[X]]$ is a subring of $F[[X,Y]]$ consisting of all sums for which $a_{k,j} = 0$ for $j > 0$. Similarly, $F[[Y]]$ is a subring of $F[[X,Y]]$. Write $f$ or $f(X,Y)$, etc., for a general element of $F[[X,Y]]$; elements of the subrings $F[[X]]$ and $F[[Y]]$ can be written $f(X)$, $g(Y)$, etc.

If $f(X,Y) = \sum_{k,j=0}^{\infty} a_{k,j} X^k Y^j$ is an element of $F[[X,Y]]$, we may write $f(0,0) = a_{0,0}$, $f_X(0,0) = \frac{\partial f}{\partial X}(0,0) = a_{1,0}$, and $f_Y(0,0) = \frac{\partial f}{\partial Y}(0,0) = a_{0,1}$.

(b) Describe the invertible elements of $F[[X,Y]]$.

c) If $f, g, h \in F[[X,Y]]$, with $g(0,0) = h(0,0) = 0$, show that there is a well-defined element $f(g(X,Y), h(X,Y)) \in F[[X,Y]]$. If $g, h \in F[[X]]$, then $f(g(X), h(X)) \in F[[X]]$ also.

d) Prove the following “Formal Implicit Function Theorem” (cf. VIII.6.3.10.):

**Theorem.** Let $f(X,Y) \in F[[X,Y]]$ with $f(0,0) = 0$ and $f_Y(0,0) \neq 0$. Then there is a unique $g \in F[[X]]$ such that $f(X, g(X)) = 0$ (where 0 denotes the zero element of $F[[X,Y]]$). We have $g(0) = 0$ and $g'(0) = -\frac{f_X(0,0)}{f_Y(0,0)}$.

e) Discuss Laurent series in two variables and the quotient field of $F[[X,Y]]$.

(f) Generalize to formal power series rings $F[[X_1, \ldots, X_n]]$. Formulate and prove a higher-dimensional version of the Formal Implicit Function Theorem analogous to VIII.6.1.7..

g) To what extent can Puiseux series, Levi-Civita Numbers, and Hahn Series be generalized to several variables?

References: [?, 4.4.7], [Jz75], [Sok11], [KP02a].

**III.5.3.9. The $p$-adic Numbers.** The real numbers are not the only field formed by completing $\mathbb{Q}$. This problem describes the completion of $\mathbb{Q}$ with respect to a different metric. There is actually one of these for each prime number $p$. These fields are important in number theory.

Fix a prime number $p$. If $x \in \mathbb{Q}$, $x \neq 0$, then $x$ can be written $p^n \cdot \frac{a}{b}$, where $n \in \mathbb{Z}$ and $a, b \in \mathbb{Z}$ are not divisible by $p$. The number $n$ is uniquely determined by $x$; write $\nu_p(x) = n$ and $|x|_p = p^{-n} = p^{-\nu_p(x)}$. Set $\nu_p(0) = +\infty$ and $|0|_p = 0$.

(a) Show that if $x, y \in \mathbb{Q}$, $\nu_p(x + y) \geq \min(\nu_p(x), \nu_p(y))$ and $\nu_p(xy) = \nu_p(x) + \nu_p(y)$.

(b) Show that if $x, y \in \mathbb{Q}$, then $|x + y|_p \leq \max(|x|_p, |y|_p) \leq |x|_p + |y|_p$ and $|xy|_p = |x|_p |y|_p$.

(c) Show that $d_p$ defined by $d_p(x, y) = |x - y|_p$ is a metric on $\mathbb{Q}$.

(d) Let $\mathbb{Q}_p$ be the set of equivalence classes of Cauchy sequences of rational numbers with respect to $d_p$, where equivalence is defined as in (). Mimic the proof of () to define addition and multiplication on $\mathbb{Q}_p$, and show that $\mathbb{Q}_p$ is a field. Extend $| \cdot |_p$ and $d_p$ to $\mathbb{Q}_p$, and show that $\mathbb{Q}_p$ is complete with respect to $d_p$. (Note that the topology defined by $d_p$ is quite different from the usual topology on $\mathbb{Q}$: $p^n \to 0$ in the $d_p$-topology, for example.)

(e) Show that if $x \in \mathbb{Z}$, then $|x|_p \leq 1$, and the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$ is

$$\mathbb{Z}_p = \{ x \in \mathbb{Q}_p : |x|_p \leq 1 \}.$$  

This is a subring of $\mathbb{Q}_p$, called the ring of $p$-adic integers. (The notation $\mathbb{Z}_p$ is inconsistent with (); this meaning of the notation $\mathbb{Z}_p$ will be used only in this problem.) Note that $\mathbb{Z}_p \cap \mathbb{Q}$ is larger than $\mathbb{Z}$; it includes all rational numbers $\frac{a}{b}$ where $p$ does not divide $b$. 

299
(f) Show that $\mathbb{Q}_p$ is locally compact with its metric topology, and $\mathbb{Z}_p$ is a compact open subset.

(g) Show that $\mathbb{Q}_p$ is totally disconnected.

(h) Show that the ring $\mathbb{Z}_p$ has only one maximal ideal $J$, the set of multiples of $p$, and $\mathbb{Z}_p/J \cong \mathbb{Z}/p\mathbb{Z}$.
Every nonzero ideal in $\mathbb{Z}_p$ is a principal ideal generated by $p^n$ (the set of all multiples of $p^n$) for some $n \in \mathbb{N}$.

(i) Show that the quotient field (of $\mathbb{Z}_p$) is $\mathbb{Q}_p$.

### III.6. The Complex Numbers

Although they are not as familiar as the real numbers, the complex numbers $\mathbb{C}$ are the natural culmination of the expansion of the number system. Although we will make only very limited use of them in this book, they are also extremely important in analysis, and many parts of analysis, e.g. the theory of analytic functions and Fourier series, are naturally done on $\mathbb{C}$ instead of $\mathbb{R}$. Complex numbers and analysis done on them (complex analysis) are among the most beautiful parts of mathematics, and the most useful in applications.

From a purely algebraic standpoint, the real numbers have almost all the properties one might want in a number system. The only obvious shortcoming is that negative numbers do not have square roots in $\mathbb{R}$. The idea of making the complex numbers is to expand $\mathbb{R}$ to a larger system of numbers in which square roots of negative real numbers (and, as it turns out, square roots of arbitrary numbers in the expanded system) can be taken. The resulting system actually yields a lot more than just square roots.

#### III.6.1. Axiomatization of $\mathbb{C}$

**III.6.1.1.** Axioms for $\mathbb{C}$ are not standardized, but here is a set which can be used:

- (C1) $\mathbb{C}$ has two binary operations making it into a field.
- (C2) $\mathbb{C}$ contains $\mathbb{R}$ as a subfield (more precisely, $\mathbb{C}$ contains a specified subfield $\mathbb{R}$ with a specified isomorphism between $\mathbb{R}$ and $\mathbb{R}$).
- (C3) $\mathbb{C}$ contains an element $i$ with $i^2 = -1$.
- (C4) Every element of $\mathbb{C}$ can be written as $a + bi$ for $a, b \in \mathbb{R}$ (or $\mathbb{R}$).

If $\mathcal{C}$ is any model for $\mathbb{C}$, then since $\mathcal{C}$ is a field, all the standard algebraic operations and identities hold in $\mathcal{C}$, and we are justified in using standard algebraic notation and manipulation in $\mathcal{C}$ without comment.

**III.6.1.2.** **Proposition.** Let $\mathcal{C}$ be a model for $\mathbb{C}$. If $z \in \mathcal{C}$, then $z$ has a unique representation $z = x + yi$ where $x$ and $y$ are in the specified subfield $\mathcal{R}$ which is a model for $\mathbb{R}$.

**Proof:** There is such a representation for $z$ by (C4). To show uniqueness, if $z = x_1 + y_1 i = x_2 + y_2 i$, then $0 = (x_1 - x_2) + (y_1 - y_2) i$, so it suffices to show that if $0 = a + bi$ with $a, b \in \mathcal{R}$, then $a = b = 0$. But

$$0 = 0 \cdot (a - bi) = (a + bi)(a - bi) = a^2 - b^2 i^2 = a^2 + b^2$$

and thus $a = b = 0$.

We can then quickly show that any two models for $\mathbb{C}$ are essentially identical (isomorphic):
III.6.1.3. **Theorem.** Let \((C, \mathcal{R}, i)\) and \((C', \mathcal{R}', i')\) be models for \(\mathbb{C}\), and let \(\psi: \mathcal{R} \to \mathcal{R}'\) be the isomorphism induced by the specified isomorphisms of \(\mathcal{R}\) and \(\mathcal{R}'\) with \(\mathbb{R}\). Define \(\phi: C \to C'\) by

\[
\phi(a + bi) = \psi(a) + \psi(b)i'
\]

(this is well defined by III.6.1.2). Then \(\phi\) is a bijection, and for all \(z, w \in C\),

\[
\phi(z + w) = \phi(z) + \phi(w), \quad \phi(zw) = \phi(z)\phi(w).
\]

Thus the two models are isomorphic.

**Proof:** It is clear that \(\phi\) is a bijection from III.6.1.2, and the fact that \(\psi\) is a bijection. We need check only the algebraic formulas. If \(z = x + yi\) and \(w = u + vi\), we have

\[
zw = (x + yi)(u + vi) = xu + xvi + yui + yvi^2 = (xu - yv) + (xv + yu)i
\]

\[
\phi(z) + \phi(w) = [\psi(x) + \psi(y)i'] + [\psi(u) + \psi(v)i'] = [\psi(x) + \psi(u)] + [\psi(y) + \psi(v)]i' = \phi(x + u) + \phi(y + v)i' = \phi(z + w)
\]

\[
\phi(z)\phi(w) = [\psi(x) + \psi(y)i'][\psi(u) + \psi(v)i']
\]

\[
= \psi(x)\psi(u) + [\psi(x)\psi(v) + \psi(y)\psi(u)]i' + \psi(y)\psi(v)^2 = [\psi(x)\psi(u) - \psi(y)\psi(v)] + [\psi(x)\psi(v) + \psi(y)\psi(u)]i'
\]

\[
= \phi(xu - yv) + \phi(xv + yu)i' = \phi(xu - yv) + [xv + yu]i' = \phi(zw).
\]

Thus, once we have shown that there is a model for \(\mathbb{C}\), as usual we can talk about the complex numbers \(\mathbb{C}\), and regard \(\mathbb{R}\) as a subset of \(\mathbb{C}\).

III.6.1.4. If \((C, \mathcal{R}, i)\) is a model for \(\mathbb{C}\), and \(z = a + bi\) for \(a, b \in \mathcal{R}\), the numbers \(a\) and \(b\) are uniquely determined by \(z\) and are called the real and imaginary parts of \(z\), denoted \(a = \Re(z), b = \Im(z)\) (the notation \(a = \Re(z)\) and \(b = \Im(z)\) is also used). The element \(a - bi\) of \(\mathbb{C}\) is called the complex conjugate of \(z\), denoted \(\overline{z}\), and the real number \(\sqrt{a^2 + b^2}\) is called the absolute value (or modulus, magnitude, length, or norm) of \(z\), denoted \(|z|\). We have \(|i| = 1\), and \(|z|\) on \(\mathcal{R}\) coincides with the usual absolute value on \(\mathbb{R}\) under the specified identification.

The proofs in the next proposition are simple algebraic calculations and are left to the reader.
III.6.1.6. **Proposition.** Let \((C, R, i)\) be a model for \(\mathbb{C}\), and let \(z = x + yi\) and \(w = u + vi\) be elements of \(\mathbb{C}\). Then

(i) \(\bar{z} + \bar{w} = \bar{z + w}\) and \(\bar{z}w = \bar{z}\bar{w}\).

(ii) \(\bar{z} = \bar{z}\).

(iii) \(z\bar{z} = |z|^2\).

(iv) \(|zw| = |z||w|\). In particular, \(|-z| = |z|\) and \(|zi| = |z|\).

(v) \(|Re(z)| \leq |z|\) and \(|Im(z)| \leq |z|\).

It is also true that \(|z + w| \leq |z| + |w|\), but we defer the proof to the next section.

III.6.1.7. It is impossible to make \(\mathbb{C}\) (any model of \(\mathbb{C}\)) into an ordered field. For if \(P\) is a positive cone in \(\mathbb{C}\) satisfying (Pos), then either \(i\) or \(-i\) is in \(P\); in either case \(i^2 = (-i)^2 = -1\) would be in \(P\), a contradiction.

III.6.2. The Standard Model for \(\mathbb{C}\): the Complex Plane

There is a universally-used standard model for \(\mathbb{C}\), the complex plane. In fact, this model for \(\mathbb{C}\) is so standard that \(\mathbb{C}\) is generally defined to be, or at least thought of as, this model. There is no real harm in this, and even some advantage, since the geometric properties of the complex plane are of considerable use in complex analysis. But see III.6.2.24. and I.6.1.11. for some cautionary comments.

III.6.2.1. By considering real and imaginary parts, complex numbers can be naturally identified with ordered pairs of real numbers. So in our standard model, the elements will be ordered pairs of real numbers. Addition of complex numbers just corresponds to addition of the real parts and of the imaginary parts, i.e. coordinatewise addition, which is the usual addition of ordered pairs of numbers. The first calculation in the proof of III.6.1.3. gives a strong clue to how multiplication of ordered pairs should be defined. Thus we are led to the following definition:

III.6.2.2. **Definition.** Let \(\mathbb{C}\) be the set \(\mathbb{R}^2\) of ordered pairs of real numbers. Define addition and multiplication in \(\mathbb{C}\) by

\[
(x, y) + (u, v) = (x + u, y + v)
\]

\[
(x, y)(u, v) = (xu - yv, xv + yu)
\]

III.6.2.3. **Proposition.** The associative, commutative, and distributive laws hold for addition and multiplication in \(\mathbb{C}\).

**Proof:** The commutative laws obviously hold. The associative law for addition is a simple calculation:

\[
[(x, y) + (u, v)] + (s, t) = [(x + u) + s, (y + v) + t] = (x + [u + s], y + [v + t]) = (x, y) + [(u, v) + (s, t)].
\]
The associative law for multiplication and the distributive law are straightforward but tedious calculations which should probably just be left to the interested, skeptical, or masochistic reader, but we give them for completeness:

\[
[(x;y)(u;v)](s;t) = (xu - yv, xv + yu)(s, t) = (xus - yvs - [xv + yu])u + yv + xvs + yus)
= (xus - xvt - [yvs + yvt], xuT + xvs + yus - yv) = (x, y)(us - vt, ut + vs) = (x, y)[(u, v)(s, t)] .
\]

\[
(x, y)[(u, v) + (s, t)] = (x, y)(u + s, v + t) = (xu + xs - [yv + yt], xv + xt + yu + ys)
= ([xu - yv] + [xv + yu], [xv + yu] + [xt + ys]) = (xu - yv, xv + yu) + (xs - yt, xt + ys) = (x, y)(u, v) + (x, y)(s, t) .
\]

**III.6.2.4.** The pair \((0, 0)\) is an additive identity, and \((-x, -y)\) is an additive inverse for \((x, y)\). We also have \((1, 0)(x, y) = (x, y)\) for any \((x, y)\), so \((1, 0)\) is a multiplicative identity. There are also multiplicative inverses:

**III.6.2.5.** Proposition. If \((x, y) \neq (0, 0)\), then \(x^2 + y^2 > 0\) and \(\left(\frac{x}{x^2 + y^2}, -\frac{y}{x^2 + y^2}\right)\) is a multiplicative inverse for \((x, y)\).

**Proof:** The first statement follows from (). We have

\[
(x, y)\left(\frac{x}{x^2 + y^2}, -\frac{y}{x^2 + y^2}\right) = x\left(\frac{x}{x^2 + y^2} - y \left[-\frac{y}{x^2 + y^2}\right]\right) + y\left(\frac{x}{x^2 + y^2} + y \left[-\frac{x}{x^2 + y^2}\right]\right)
= \left(\frac{x^2}{x^2 + y^2} + \frac{y^2}{x^2 + y^2}, -\frac{x}{x^2 + y^2} + \frac{x}{x^2 + y^2}\right) = (1, 0) .
\]

We summarize these results:

**III.6.2.6.** Theorem. \(C\) is a field with respect to the operations + and \(\cdot\) of III.6.2.2., with zero element \((0, 0)\) and multiplicative identity \((1, 0)\).

**III.6.2.7.** Proposition. Let \(a, b, u, v \in \mathbb{R}\). Then in \(C\), \((a, 0)(u, v) = (au, av)\). In particular, \((a, 0)(b, 0) = (ab, 0)\). So, since \((a, 0) + (b, 0) = (a + b, 0)\), the map \(\psi : \mathbb{R} \to C\) with \(\psi(x) = (x, 0)\) is an isomorphism from \(\mathbb{R}\) onto the subfield \(\mathcal{R} = \{(x, 0) : x \in \mathbb{R}\}\) of \(C\).

The proof is a simple calculation left to the reader.

Finally, we need \(i\). The proof of the next proposition is also a simple calculation.

303
III.6.2.8. Proposition. In $C$, $(0, 1)^2 = (-1, 0)$. If $(x, y) \in C$, then $(x, y) = (x, 0) + (y, 0)(0, 1)$.

III.6.2.9. Thus, if we call $(0, 1)$ $i$, we have that every complex number $(x, y)$ can be written $a + bi$ for $a, b \in R$, namely $a = (x, 0), b = (y, 0)$. Under the identification of $R$ with $R$ via $\psi$, we have $i^2 = -1, x = Re((x, y)), y = Im((x, y)), (x, y) = (x, -y)$. We usually write $x + yi$ instead of $(x, y)$ for elements of $C$.

The definition of the complex conjugate, along with III.6.1.6., motivates the calculation in III.6.2.5.: we have $z^{-1} = \frac{\overline{z}}{|z|^2}$ if $z \neq 0$.

Summarizing again, we get our final result:

III.6.2.10. Theorem. $(C, R, (0, 1))$ is a model for $C$. Every model for $C$ is isomorphic to this one.

Geometry of $C$

We thus normally represent $x + yi \in C$ geometrically as $(x, y) \in C = R^2$. The real number $x$ is represented as $(x, 0)$, and $i$ is represented as $(0, 1)$. The $x$-axis in $R^2$ is $R$, called the real axis; the $y$-axis is the set of real multiples of $i$, called the imaginary axis. See Figure ().

III.6.2.11. We also have that $|x + yi| = \sqrt{x^2 + y^2}$. This is the same as $\|(x, y)\|$, where $\| \cdot \|$ is the standard norm in $R^2$ (). Thus $|x + yi|$ can be interpreted as the Euclidean distance between $x + yi = (x, y)$ and $0 = (0, 0)$. The next result then follows immediately from ():

III.6.2.12. Proposition. The triangle inequality holds in $C$: if $z, w \in C$, then $|z + w| \leq |z| + |w|$. In particular, $|z| \leq |Re(z)| + |Im(z)|$.

III.6.2.13. There are nice geometric interpretations of addition and multiplication in $C$ (regarded as $C$). Since addition in $C$ coincides with vector addition in $R^2$, it is given by the parallelogram law (); see Figure ()

Multiplication also has a simple interpretation using polar coordinates. We first describe the polar coordinates of an element of $C$. Some of this discussion must be regarded as informal until the trigonometric functions are properly defined (cf. (); note that no complex numbers are used in the definition of the trigonometric functions or development of their properties).

III.6.2.14. Definition. Let $z = x + yi \in C, z \neq 0$. If $(r, \theta)$ are the polar coordinates () of the point $(x, y) \in R^2$, then $r = \sqrt{x^2 + y^2} = |x + yi|$ is called the modulus, magnitude, length, or norm of $z$, and $\theta$ is called the argument of $z$, denoted $Arg(z)$. We have $x = r \cos \theta$ and $y = r \sin \theta$; if $x \neq 0$, we have $\tan \theta = \frac{y}{x}$. We write $e^{i\theta}$ for the complex number $\cos \theta + i \sin \theta$.

Conversely, if $(r, \theta)$ is a pair of real numbers with $r \neq 0$, then there is a unique $z \in C$ with $r = |z|$ and $\theta = Arg(z)$. We do not define $Arg(0)$. The argument of a positive real number is 0, and the argument of a negative real number is $\pi$.
Note that the argument of a complex number is only determined up to addition of a multiple of $2\pi$. This ambiguity leads to considerable difficulties, but also a certain richness, in complex analysis; cf. III.6.2.21. In particular, the argument cannot be defined continuously for all $z \in \mathbb{C} \setminus \{0\}$, although it can be continuously defined locally (except around 0). Note that even though $\theta$ is not uniquely determined by $z$, the complex number $e^{i\theta}$ is uniquely determined.

We will consider $e^{i\theta}$ to simply be a convenient notational abbreviation for $\cos \theta + i \sin \theta$, but it is actually a special case of the complex exponential function (X.1.1.8 (iii)). This notation retains at least one important property of the usual exponential function:

III.6.2.15. **Proposition.** If $\theta$ and $\phi$ are real numbers, then $e^{i(\theta+\phi)} = e^{i\theta} e^{i\phi}$.

**Proof:** This is just an application of basic trigonometric identities. We have

$$e^{i(\theta+\phi)} = \cos(\theta + \phi) + i \sin(\theta + \phi) = [\cos \theta \cos \phi - \sin \theta \sin \phi] + i[\sin \theta \cos \phi + \cos \theta \sin \phi]$$

$$= (\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi) = e^{i\theta} e^{i\phi}.$$

The next observation is simple, but fundamental:

III.6.2.16. **Proposition.** If $z \in \mathbb{C}$, $z \neq 0$, and $r = |z|$, $\theta = \text{Arg}(z)$, then $z = re^{i\theta}$. So $e^{i\theta} = z/|z|$.

As a corollary of the last two propositions, we get a geometric interpretation of multiplication:

III.6.2.17. **Corollary.** Let $z$ and $w$ be nonzero complex numbers. Then $|zw| = |z||w|$ and $\text{Arg}(zw) = \text{Arg}(z) + \text{Arg}(w)$ (where the last formula is interpreted modulo $2\pi$). Alternatively, if $z = re^{i\theta}$ and $w = se^{i\phi}$ with $r, s, \theta, \phi \in \mathbb{R}$ and $r, s > 0$, then $zw = rse^{i(\theta+\phi)}$.

III.6.2.18. Multiplication by a fixed complex number $z$ can thus be interpreted geometrically as scaling by a factor of $|z|$ and rotating counterclockwise around the origin by an angle $\text{Arg}(z)$. See Figure ( ). This interpretation is in line with the geometric interpretation of multiplication of vectors by real numbers (scalars): multiplication by a positive real number just scales length, and multiplication by a negative real number scales length and reverses direction (rotates by an angle $\pi$).

III.6.2.19. The polar coordinate representation provides an easy way to show that complex numbers have arbitrary roots. Suppose $z \in \mathbb{C}$ and $n \in \mathbb{N}$; we want to show there is an $x \in \mathbb{C}$ with $x^n = z$. If $z = 0$, this is obvious (and the only solution is $x = 0$). Suppose $z \neq 0$. Set $s = |z|^{1/n}$ and $\phi = \text{Arg}(z)/n$. Then $w = se^{i\phi}$ satisfies $w^n = z$. Any multiple of $\frac{2\pi}{n}$ can be added to $\phi$ with the same result; thus there are exactly $n$ distinct solutions in $\mathbb{C}$ to $x^n = z$. Summarizing, we obtain:

III.6.2.20. **Proposition.** Let $z \in \mathbb{C}$ and $n \in \mathbb{N}$. Then the equation $x^n = z$ has a solution in $\mathbb{C}$. If $z = 0$, the unique solution is $x = 0$. If $z \neq 0$, there are exactly $n$ solutions, distributed uniformly on the circle of radius $|z|^{1/n}$ around 0 at angles $\frac{\text{Arg}(z)}{n} + \frac{2k\pi}{n}$, $0 \leq k \leq n - 1$. 

305
III.6.2.21. In particular, every nonzero complex number $z$ has exactly two square roots, which are negatives of each other on the line through the origin of angle $\frac{\text{Arg}(z)}{2}$ with the positive real axis. However, as $z$ varies there is no way to choose one of the two square roots of $z$ continuously on the whole complex plane, although a continuous choice can be made locally (except around 0). Thus there is no continuous square root function from all of $\mathbb{C}$ to $\mathbb{C}$.

There is a considerable generalization of III.6.2.20. which is one of the most important theorems of algebra, as its name suggests.

III.6.2.22. Theorem. [Fundamental Theorem of Algebra] Any nonconstant polynomial with complex coefficients has a root in $\mathbb{C}$. ($\mathbb{C}$ is algebraically closed.)

There are several quite different proofs of this theorem; see X.1.4.2. for a nice one. III.6.2.20. is the special case of the polynomial $x^n - z$ for any $z \in \mathbb{C}$.

III.6.2.23. As a result, there is no algebraic reason to try to expand the number system beyond $\mathbb{C}$, so $\mathbb{C}$ is a natural culmination of the progression from $\mathbb{N}$ through $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.

III.6.2.24. As mentioned earlier, although $\mathbb{C}$ is normally, and profitably, thought of as being the complex plane $\mathbb{C}$, it must be kept in the back of the mind that $\mathbb{C}$ is only a model for $\mathbb{C}$. For example, we want to regard $\mathbb{R}$ as a subfield of $\mathbb{C}$. But $\mathbb{R}$ is not a subset of $\mathbb{C}$: the subfield $\mathbb{R}$ of ordered pairs of real numbers with second coordinate 0 is not the same thing as $\mathbb{R}$, although there is an obvious and natural one-one correspondence between them. And there are other models of $\mathbb{C}$ which are useful in some contexts; see III.6.3.1., III.6.3.2.(e), III.6.3.3. for some examples.

III.6.3. Exercises

III.6.3.1. Consider the following subset of $M_2(\mathbb{R})$:

$$\mathcal{C} = \left\{ \begin{bmatrix} x & -y \\ y & x \end{bmatrix} : x, y \in \mathbb{R} \right\}.$$

(a) Show that $\mathcal{C}$ is a subring of $M_2(\mathbb{R})$.
(b) Show that $\mathcal{C}$ is unital and commutative.
(c) Show that every element $z$ of $\mathcal{C}$ can be written uniquely as $z = x_1 + yi$ for $x, y \in \mathbb{R}$, where

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$  

Define $\bar{z} = x_1 - yi$. Show that $zz = (x^2 + y^2)1 = (\text{det } z)1$. If we define $|z| = \sqrt{x^2 + y^2}$, then $|zw| = |z||w|$ for all $z, w \in \mathcal{C}$.
(d) Show that every nonzero element of $\mathcal{C}$ has a multiplicative inverse. Thus $\mathcal{C}$ is a field.
(e) Show that $\mathfrak{R} = \{x_1 : x \in \mathbb{R}\}$ is a subfield of $\mathcal{C}$ isomorphic to $\mathbb{R}$.
(f) Show that $(\mathcal{C}, \mathfrak{R}, i)$ is a model for $\mathbb{C}$.
This model for \( \mathbb{C} \) arises naturally from linear algebra. If \( z = a + bi \) is a fixed element of \( \mathbb{C} \), then multiplication by \( z \) is a real linear transformation \( M_z \) of \( \mathbb{C} = \mathbb{R}^2 \), and the matrix of this linear transformation with respect to the standard basis \( \{1, i\} \) of \( \mathbb{C} = \mathbb{R}^2 \) is \[
abla z = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.
\]

III.6.3.2. Consider the following subset of \( M_2(\mathbb{C}) \):
\[
\mathbb{H} = \left\{ \begin{bmatrix} z & -\bar{w} \\ w & \bar{z} \end{bmatrix} : z, w \in \mathbb{C} \right\}.
\]

(a) Show that \( \mathbb{H} \) is a subring of \( M_2(\mathbb{C}) \). (By considering \( \mathbb{C} \) as a subring of \( M_2(\mathbb{R}) \) as in (), \( \mathbb{H} \) can be rewritten as the subring
\[
\left\{ \begin{bmatrix} a & -c & -b & -d \\ c & a & d & -b \\ b & -d & a & c \\ d & b & -c & a \end{bmatrix} : a, b, c, d \in \mathbb{R} \right\}
\]

of \( M_4(\mathbb{R}) \).)

(b) Show that \( \mathbb{H} \) is unital and noncommutative. (Consider the following elements:
\[
1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad j = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad k = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.
\]

Compute \( ij, ji, jk, k j, ki, ik \).]

(c) Show that every element \( x \) of \( \mathbb{H} \) can be written uniquely as \( x = a1 + bi + cj + dk \) for \( a, b, c, d \in \mathbb{R} \). Define \( \bar{x} = a1 - bi - cj - dk \). Show that \( x\bar{x} = (a^2 + b^2 + c^2 + d^2)1 \). If we define \( |x| = \sqrt{a^2 + b^2 + c^2 + d^2} \), then \( |x y| = |x||y| \) for all \( x, y \in \mathbb{H} \).

(d) Show that every nonzero element of \( \mathbb{H} \) has a multiplicative inverse. Thus \( \mathbb{H} \) satisfies all the axioms for a field except (M2). Such a ring is called a division ring or skew field.

(e) Show that each of the sets \( \mathbb{R} + \mathbb{R}i, \mathbb{R} + \mathbb{R}j, \) and \( \mathbb{R} + \mathbb{R}k \) is a subring of \( \mathbb{H} \) isomorphic to \( \mathbb{C} \).

(f) If \( x = a1 + bi + cj + dk \in \mathbb{H} \), \( a \) is called the real part or scalar part of \( x \), and \( bi + cj + dk \) the imaginary part of \( x \). If \( i, j, k \) are the standard basis vectors in \( \mathbb{R}^3 \), then \( bi + cj + dk \) can be identified with the vector \( \mathbf{v} = bi + cj + dk \in \mathbb{R}^3 \), called the vector part of \( x \). Thus an element of \( \mathbb{H} \) may be regarded as a pair \( (a, \mathbf{v}) \), where \( a \in \mathbb{R} \) and \( \mathbf{v} \in \mathbb{R}^3 \). Show that the addition is coordinatewise, and the multiplication is given by
\[
(a, \mathbf{v})(b, \mathbf{w}) = (ab - \mathbf{v} \cdot \mathbf{w}, a\mathbf{w} + b\mathbf{v} + \mathbf{v} \times \mathbf{w})
\]
where \( \cdot \) and \( \times \) are the dot and cross product of vectors in \( \mathbb{R}^3 \). (Although \( \times \) is not associative, this multiplication is associative!)

The ring \( \mathbb{H} \) is called the ring of quaternions. The name refers to the fact that it is a four-dimensional vector space (algebra) over \( \mathbb{R} \). The letter \( \mathbb{H} \) is used in honor of the Irish mathematician William Rowan Hamilton, who first defined the quaternions in 1843 and studied their properties. Hamilton became convinced that quaternions were the fundamental structure which could be used to describe and explain most of mathematics and physics. While quaternions have beautiful and useful algebraic and geometric structure, they have not quite lived up to Hamilton’s expectations, although recently they have become an increasingly
important computational tool in such areas as computer graphics and control theory. Quaternions are also of some importance in quantum mechanics; the matrices $i, j, k$ are called the Pauli spin matrices. See (c) for details.

There is a way to make $\mathbb{R}^8$ into a somewhat similar algebraic structure called the Cayley numbers in which every nonzero element has a multiplicative inverse. Multiplication in the Cayley numbers is, however, not associative. It turns out that the only values of $n$ for which $\mathbb{R}^n$ can be made into such a structure are $n = 1, 2, 4,$ and $8$, and the only structures possible are $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, and the Cayley numbers. See (c) for details.

III.6.3.3. Let $\mathbb{R}[X]$ denote the set of all polynomials with real coefficients, i.e.

$$\mathbb{R}[X] = \{a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n : n \in \mathbb{N}_0, a_0, \ldots, a_n \in \mathbb{R}\}.$$ 

(a) Show that $\mathbb{R}[X]$ is a unital commutative ring with usual addition and multiplication of polynomials.

(b) Let $J$ be the set of all polynomials in $\mathbb{R}[X]$ which are divisible by $X^2 + 1$. Show that $J$ is an ideal (Exercise III.3.7.4.) in $\mathbb{R}[X]$.

(c) Show that the quotient ring $\mathbb{R}[X]/J$ is a field which is a model for $\mathbb{C}$, with the classes of the constant functions playing the role of $\mathbb{R}$, and the class of $X$ playing the role of $i$.

Levi-Civita numbers
III.7. Fundamental Theorems About $\mathbb{R}$

III.7.1. Partitions

In this section, we will discuss partitions and tagged partitions of intervals in $\mathbb{R}$. The primary use of partitions is in Riemann-type integration, but several important basic theorems in analysis can also be proved using tagged partitions.

**Definition.** Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$. A partition of $[a, b]$ is a division of $[a, b]$ into a finite number of subintervals $\{[x_{i-1}, x_i] : 1 \leq i \leq n\}$ with only endpoints in common. A partition $\mathcal{P}$ is usually described by its set of endpoints $\{x_0, x_1, \ldots, x_n\}$, where $a = x_0 < x_1 < \cdots < x_n = b$.

The norm or mesh of a partition $\mathcal{P} = \{x_0, x_1, \ldots, x_n\}$ is
\[ \|\mathcal{P}\| = \max\{x_i - x_{i-1} : 1 \leq i \leq n\}. \]

A tag in an interval $[c, d]$ is a specified number in $[c, d]$. A tagged partition is a partition $\{x_0, x_1, \ldots, x_n\}$ with a specified tag $t_i$ in $[x_{i-1}, x_i]$ for each $i$. A tagged partition is usually denoted by $\mathcal{P} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}$.

The norm or mesh of a partition is a measure of how fine the partition is. There is a more sophisticated and flexible notion of fineness of a partition based on gauges:

**Definition.** A gauge on an interval $[a, b]$ is a function $\delta$ from $[a, b]$ to $(0, \infty)$.

If $\delta$ is a gauge on $[a, b]$, a tagged partition $\mathcal{P} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}$ of $[a, b]$ is $\delta$-fine if $x_i - x_{i-1} < \delta(t_i)$ for all $1 \leq i \leq n$.

The word “tagged” is frequently omitted in referring to $\delta$-fine partitions, although it must be kept in mind that a $\delta$-fine partition is by definition tagged.

Note that a gauge is just a strictly positive real-valued function on $[a, b]$; no continuity or other restrictions on $\delta$ are assumed. Allowing discontinuous gauges, and ones not bounded below by a positive number, gives the theory of gauges and fine partitions its power.

The smaller a gauge $\delta$ is, the harder it is for a partition to be $\delta$-fine. The most common gauges are constant functions, e.g. $\delta(x) = \epsilon$ for all $x$. For such a $\delta$, a tagged partition $\mathcal{P}$ is $\delta$-fine if and only if $\|\mathcal{P}\| < \epsilon$.

If $\delta$ is a constant gauge, then it follows easily from the Archimedean property of $\mathbb{R}$ that there is a $\delta$-fine partition of any interval $[a, b]$. However, it is not at all obvious that a nonconstant gauge $\delta$ on $[a, b]$ which is not bounded below by a positive number has a $\delta$-fine partition. This turns out to be true, and is the main result of this section; this theorem is actually equivalent to the Completeness Property and easily gives other important equivalent properties of $\mathbb{R}$.

**Theorem.** [Fine Partition Theorem] Let $\delta$ be any gauge on an interval $[a, b]$ in $\mathbb{R}$. Then there is a $\delta$-fine partition of $[a, b]$.

**Proof:** Let $S = \{x \in [a, b] : \text{there is a } \delta\text{-fine partition of } [a, x]\}$. If $a < x < a + \delta(a)$, then $\mathcal{P} = \{a, x; a\}$ is a $\delta$-fine partition of $[a, x]$ (with one subinterval). So $S$ is nonempty. Let $c = \sup S$; then $c \geq a + \delta(a) > a$.

We will show that $c = b$. First note that there is a $\delta$-fine partition of $[a, c]$; if $x \in S$ with $c - \delta(c) < x < c$, and $\{a, x_1, \ldots, x_{n-1}, x; t_1, \ldots, t_n\}$ is a $\delta$-fine partition of $[a, x]$, then $\{a, x_1, \ldots, x_{n-1}, x, c; t_1, \ldots, t_n, c\}$ is a $\delta$-fine partition of $[a, c]$. Now suppose that $c < b$, and let $y \in [a, b]$, $c < y < c + \delta(c)$; then, using the previous notation, $\{a, x_1, \ldots, x_{n-1}, x, c; y; t_1, \ldots, t_n, c, c\}$ is a $\delta$-fine partition of $[a, y]$, contradicting the definition of $c$. Thus $c = b$ and the theorem is proved.
III.7.2. Compactness and Continuity

From the Fine Partition Theorem, we derive several fundamental results about the real numbers. The standard proofs of all these results resemble the proof of the Fine Partition Theorem; our derivation of the results from the Fine Partition Theorem follows R. A. Gordon [?].

A collection $\mathcal{U} = \{U_i : i \in \Omega\}$ of open sets in $\mathbb{R}$ (or in a general topological space) is an open cover of a set $S$ if $S \subseteq \cup U_i$. A subcover of $\mathcal{U}$ for $S$ is a subset of $\mathcal{U}$ which is also a cover of $S$. A subset $F$ of $\mathbb{R}$ is compact if every open cover of $F$ has a finite subcover.

**Theorem.** A closed bounded interval $[a, b]$ in $\mathbb{R}$ is compact.

**Proof:** Let $\mathcal{U} = \{U_i : i \in \Omega\}$ be an open cover of $[a, b]$. For each $x \in [a, b]$, choose an $i_x \in \Omega$ with $x \in U_{i_x}$, and choose $\delta(x) > 0$ such that $(x - \delta(x), x + \delta(x)) \subseteq U_{i_x}$. Then $\delta : [a, b] \to (0, \infty)$ is a gauge on $[a, b]$. Let $\{x_0, x_1, \ldots, x_n, t_1, \ldots, t_m\}$ be a $\delta$-fine partition of $[a, b]$. Then $[x_{k-1}, x_k] \subseteq (t_k - \delta(t_k), t_k + \delta(t_k)) \subseteq U_{i_{t_k}}$ for each $k$, so $\{U_{i_1}, \ldots, U_{i_{t_n}}\}$ is a finite subcover of $\mathcal{U}$.

**Corollary.** Every closed bounded subset of $\mathbb{R}$ is compact.

**Proof:** If $F$ is closed and contained in $[a, b]$, and $\mathcal{U}$ is an open cover of $F$, then $\mathcal{U} \cup \{F^c\}$ is an open cover of $[a, b]$, and this open cover has a finite subcover $\mathcal{V}$ for $[a, b]$. $\mathcal{V} \setminus \{F^c\} \subseteq \mathcal{U}$ is then a finite subcover of $\mathcal{U}$ for $F$.

Virtually the same proof shows that any closed subset of a compact set is compact. The converse is also true:

**Proposition.** A compact subset of $\mathbb{R}$ is closed and bounded.

**Proof:** Let $E$ be a compact subset of $\mathbb{R}$. If $U_m = (-m, m)$, then $\{U_m : m \in \mathbb{N}\}$ is an open cover of $E$, so there is a finite subcover $\{U_{m_1}, \ldots, U_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $E \subseteq (-M, M)$. To see that $E$ is closed, let $c \in E^c$, and for each $\epsilon > 0$ let $U_\epsilon = \{x \in \mathbb{R} : |x - c| > \epsilon\}$. Then $\{U_\epsilon : \epsilon > 0\}$ is an open cover of $E$, so there is a finite subcover $\{U_{\epsilon_1}, \ldots, U_{\epsilon_n}\}$. If $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$, then $(c - \epsilon, c + \epsilon) \subseteq E^c$, so $E^c$ is open.

Combining these two results, we obtain:

**Corollary.** [Heine-Borel Theorem] A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

The Heine-Borel Theorem gives a generalization of the Nested Interval Theorem:

**Theorem.** Let $(F_m)$ be a decreasing sequence of nonempty closed bounded subsets of $\mathbb{R}$. Then $\cap_{m} F_m$ is not empty.

**Proof:** Suppose $F_1 \subseteq [a, b]$. If $\cap F_m = \emptyset$, then $\{F_m^c : m \in \mathbb{N}\}$ is an open cover of $[a, b]$, which has a finite subcover $\{F_{m_1}, \ldots, F_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $F_M = F_{m_1} \cap \cdots \cap F_{m_n} = \emptyset$, a contradiction.

From this, we get the important Bolzano-Weierstrass Theorem:

**Theorem.** [Bolzano-Weierstrass Theorem] Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.
The definition of the formula for the function.

Proof: Suppose \( c \) is not one of the \( b_n \). For \( \epsilon > 0 \), choose \( \delta_1 > 0 \) such that \( c - \delta_1 < x < c \) and \( x \in E \) imply \( |f(x) - f(c)| < \epsilon \) (using the fact that \( f \) is continuous on \( E \)). Since \( c \) is not one of the \( b_n \), there is a

$$
\text{We then get some basic theorems about continuous functions as consequences:}
$$

**Theorem.** Let \( E \) be a compact subset of \( \mathbb{R} \), and let \( f : E \to \mathbb{R} \) be continuous. Then \( f(E) \) is compact.

**Proof:** We first show that \( f(E) \) is bounded. Suppose for each \( n \) there is an \( x_n \in E \) with \( |f(x_n)| \geq n \). Since \( E \) is bounded, \( (x_n) \) has a subsequence \( (x_{k_n}) \) converging to some \( y \in \mathbb{R} \), and \( y \in E \) since \( E \) is closed. Because \( f \) is continuous, the sequence \( (f(x_{k_n})) \) converges to \( f(y) \) by the Sequential Criterion. But this is impossible since \( (f(x_{k_n})) \) is unbounded by construction, a contradiction.

Now suppose that \( c \) is in the closure of \( f(E) \), and choose \( x_n \in E \) such that \( f(x_n) \to c \). Again by the Bolzano-Weierstrass Theorem, \( (x_n) \) has a subsequence \( (x_{k_n}) \) converging to some \( y \in E \). But then \( (f(x_{k_n})) \) converges to \( f(y) \) by the Sequential Criterion: since \( (f(x_{k_n})) \) converges to \( c \), we have \( c = f(y) \in f(E) \) and \( f(E) \) is closed.

**Corollary.** [Max-Min Theorem] Let \( E \) be a closed bounded subset of \( \mathbb{R} \), and \( f : E \to \mathbb{R} \) a continuous function. Then \( f \) is bounded on \( E \) and attains its maximum and minimum on \( E \), i.e. there are \( c, d \in E \) such that \( f(c) \leq f(x) \leq f(d) \) for all \( x \in E \).

For the proof, just note that \( f(E) \) is compact and that a compact subset of \( \mathbb{R} \) has a maximum and a minimum element.

See ( ) for a more complete discussion of compactness.

The next theorem is a variation of the Tietze Extension Theorem of topology ( ). It is certainly not true that a continuous function on a subset \( E \) of \( \mathbb{R} \) can be extended to a continuous function on all of \( \mathbb{R} \); there are many counterexamples, such as \( f(x) = \frac{1}{x} \) on \((0, \infty)\). But if \( E \) is closed, there is always an extension:

**Theorem.** [Tietze Extension Theorem for \( \mathbb{R} \)] Let \( E \) be a closed subset of \( \mathbb{R} \), and \( f : E \to \mathbb{R} \) a continuous function. Then \( f \) can be extended to a continuous function on all of \( \mathbb{R} \), i.e. there is a continuous function \( g : \mathbb{R} \to \mathbb{R} \) which agrees with \( f \) on \( E \).

**Proof:** The definition of \( g \) is very simple. Let \( U = E^c; \) then \( U \) is open, so it consists of a union of countably many disjoint open intervals \( I_n = (a_n, b_n) \) and possibly an unbounded interval of the form \((\infty, a)\) (if \( E \) has a smallest element \( a \)) and/or an unbounded interval \((b, \infty)\) (if \( E \) has a largest element \( b \)). Define \( g \) to be equal to \( f \) on \( E \), linear on \([a_n, b_n] \) for each \( n \), and if necessary constant on \((\infty, a)\) and/or \([b, \infty)\). If \( x \in I_n \), the formula for \( g(x) \) is

$$
g(x) = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n} (x - a_n).$$

It is geometrically clear that \( g \) is continuous everywhere, but takes a little work to write down a careful proof. Suppose \( c \in \mathbb{R} \); we will show \( g \) is continuous at \( c \). If \( c \notin E \), this is obvious since \( g \) is linear in a neighborhood of \( c \). Suppose \( c \in E \). We will show that \( \lim_{x \to c} f(x) = f(c) \). Let \( \epsilon > 0 \) be given. If \( c = b_n \) for some \( n \), let \( \delta = \min(b_n - a_n, 1/(b_n - f(a_n)) \) (if \( f(a_n) = f(b_n) \), set \( \delta = b_n - a_n \)). Then if \( c - \delta < x < c \), \( g \) is linear on \([x, c]\), so \( |f(x) - f(c)| < \epsilon \).

Now suppose that \( c \) is not one of the \( b_n \). For \( \epsilon > 0 \), choose \( \delta_1 > 0 \) such that \( c - \delta_1 < x < c \) and \( x \in E \) imply \( |f(x) - f(c)| < \epsilon \) (using the fact that \( f \) is continuous on \( E \)). Since \( c \) is not one of the \( b_n \), there is a
Let \( f : [a, b] \to \mathbb{R} \) be bounded, and continuous a.e. Then \( f \) is Riemann integrable on \([a, b] \). 

Theorem. Let \( F \) be a closed bounded set in \( \mathbb{R} \). Then any continuous function from \( F \) to \( \mathbb{R} \) is uniformly continuous.

Apply this theorem to \( F = [a, b] \setminus U \): there is thus a \( \delta > 0 \) such that, whenever \( x, t \in F \) and \( |x - t| < \delta \), \( |f(x) - f(t)| < \epsilon / 2(b - a) \).

Now choose a partition \( \mathcal{P} \) of \([a, b]\), containing the intervals \( I_1, \ldots, I_n \), and such that the other intervals in the partition all have length < \( \delta \). Let \( \phi \) and \( \psi \) be the step functions with values \( \inf_{x \in I} f(x) \) and \( \sup_{x \in I} f(x) \) on \( I \) respectively, for each subinterval \( I \) in \( \mathcal{P} \). (It doesn’t matter how \( \phi \) and \( \psi \) are defined on the endpoints of the subintervals so long as \( \phi \leq f \leq \psi \) there.) Then \( 0 \leq \psi - \phi \leq 2M \), and \( \psi - \phi \leq \epsilon / 2(b - a) \) on \( F \). So

\[
\int_{[a,b]} (\psi - \phi) \, d\lambda = \int_U (\psi - \phi) \, d\lambda + \int_F (\psi - \phi) \, d\lambda \leq 2M \lambda(U) + \frac{\epsilon}{2(b-a)} \lambda(F)
\]

\[
< 2M \frac{\epsilon}{4M} + \frac{\epsilon}{2(b-a)} (b-a) = \epsilon.
\]

Now consider the more difficult case where \( U \) is an infinite disjoint union of open intervals \( I_1, I_2, \ldots \). We may define \( F \) and \( \delta \) as before. We can choose a sequence \( \{\mathcal{P}_n\} \) of partitions of \([a, b]\) such that \( \mathcal{P}_n \) contains \( I_1, \ldots, I_n \) and such that the rest of the subintervals have length < \( \delta \), and define \( \phi_n, \psi_n \) as above. We may with some care arrange for the \( \mathcal{P}_n \) to become successively finer as \( n \) increases (i.e. each interval in \( \mathcal{P}_n \) is a union of intervals in \( \mathcal{P}_{n+1} \)): it is only necessary to insure that every interval \( I_k \) in \( U \) is entirely contained in a single subinterval in \( \mathcal{P}_1 \), which is easily arranged since there are only finitely many of the \( I_k \) with length \( \geq \delta / 2 \). If the \( \mathcal{P}_n \) are successively finer, we have

\[
\phi_1 \leq \phi_2 \leq \cdots \leq \phi \leq \cdots \leq \psi_2 \leq \psi_1.
\]

Let \( \phi = \sup \phi_n, \psi = \inf \psi_n \). Then the \( \phi_n \) and \( \psi_n \) are step functions, but \( \phi \) and \( \psi \) are not step functions in general. However, they are Lebesgue integrable. Also, we have \( \phi \leq f \leq \psi, 0 \leq \psi - \phi \leq 2M \), and \( \psi - \phi \leq \epsilon / 2(b-a) \) on \( F \), so as before we have \( \int_{[a,b]} (\psi - \phi) \, d\lambda < \epsilon \). But by the Monotone Convergence Theorem (or the Dominated Convergence Theorem),

\[
\int_{[a,b]} (\psi - \phi) \, d\lambda = \lim_{n \to \infty} \int_{[a,b]} (\psi_n - \phi_n) \, d\lambda
\]

so \( \int_{[a,b]} (\psi_n - \phi_n) \, d\lambda \leq \epsilon \) for sufficiently large \( n \).
III.7.3. The Creeping Lemma

An alternate approach to many fundamental theorems about \( \mathbb{R} \) was presented in [MR68], based on the following fact which depends crucially on the Completeness Property of \( \mathbb{R} \):

**Theorem. [Creeping Lemma]** Let \( I \) be an interval in \( \mathbb{R} \), and let \( R \) be a transitive relation on \( I \) with the following property: for every \( x \in I \), there is a \( \delta_x > 0 \) (perhaps depending on \( x \)) such that, whenever \( y \in I \), \( x - \delta_x < y \leq x \), then \( yRx \), and whenever \( y \in I \), \( x \leq y < x + \delta_x \), then \( xRy \). Then \( R \) extends \( \leq \) on \( I \), i.e. for every \( a, b \in I \), \( a \leq b \), we have \( aRb \).

**Proof:** \( R \) is automatically reflexive by the assumed condition. Fix \( a, b \in I \), \( a \leq b \). We may suppose \( a < b \). Let
\[
S = \{ x \in [a, b] : aRx \} .
\]
Then \( S \) is a bounded subset of \( \mathbb{R} \) which is nonempty since \( a \in S \). Set \( c = \sup(S) \). Then \( c > a \) since \( [a, a + \delta_a] \subseteq S \) for some \( \delta_a > 0 \), so \( c \geq a + \delta_a \).

There is a \( \delta_c > 0 \) such that if \( c - \delta_c < y \leq c \), then \( yRc \). Since \( c = \sup(S) \), there is an \( x \in S \), \( c - \delta_c < x \leq c \); then \( aRx \) and \( xRc \). Thus \( aRc \) by transitivity.

If \( c < b \), then if \( c < x < \min(c + \delta_c, b) \), we have \( cRx \); since \( aRc \), we have \( aRx \), so \( x \in S \), contradicting that \( c = \sup(S) \). Thus \( c = b \). \( \diamond \)

The next corollary is called the *Weak Creeping Lemma* in [MR68]. For a slick version of the proof, see ().

**Corollary.** Let \( I \) be an interval in \( \mathbb{R} \), and let \( R \) be an equivalence relation on \( I \) with the following property: for every \( x \in I \), there is a \( \delta_x > 0 \) (perhaps depending on \( x \)) such that, whenever \( y \in I \), \( |x - y| < \delta_x \), then \( xRy \). Then \( R \) is the full relation on \( I \), i.e. for every \( a, b \in I \), we have \( aRb \).

Here is a simple sample application. Only the weak version is needed.

**Proposition.** Let \( f \) be a continuous function on a closed bounded interval \([a, b] \). Then \( f \) is bounded on \([a, b]\).

**Proof:** Define an equivalence relation on \( I = [a, b] \) by \( x \sim y \) if \( f \) is bounded on the closed interval between \( x \) and \( y \). Fix \( x \in [a, b] \). There is a \( \delta_x \) such that \( |f(x) - f(y)| < 1 \) if \( y \in [a, b] \) and \( |x - y| < \delta_x \). Thus \( x \sim y \) if \( |x - y| < \delta_x \). By III.7.3.2., \( a \sim b \). \( \diamond \)

**Corollary.** [Extreme Value Theorem] Let \( f \) be a continuous function on a closed bounded interval \([a, b] \). Then \( f \) is bounded on \([a, b]\), and attains a maximum and minimum on \([a, b]\).

**Proof:** By III.7.3.3., \( f \) is bounded. Let \( c = \sup\{ f(x) : x \in [a, b] \} \). If \( f \) never takes the value \( c \) on \([a, b]\), set \( g(x) = \frac{1}{c - f(x)} \); then \( g \) is continuous on \([a, b]\), and positive everywhere. By III.7.3.3., \( g \) is bounded. If \( d = \sup\{ g(x) : x \in [a, b] \} \), then
\[
\frac{1}{c - f(x)} \leq d
\]

313
\[
\frac{1}{d} \leq c - f(x)
\]
\[
f(x) \leq c - \frac{1}{d}
\]
for all \(x \in [a, b]\), contradicting the definition of \(c\). Thus \(f\) takes the value \(c\) on \([a, b]\). The proof that \(f\) attains a minimum is nearly identical.

### III.7.4. Exercises

#### III.7.4.1.
Using the proof of III.7.1. as a model, give a direct proof of III.7.2..

#### III.7.4.2.
The Completeness Property of \(\mathbb{R}\) implies the Fine Partition Theorem, as the proof of III.7.1. shows. Show that the conclusion of the Fine Partition Theorem implies the Completeness Property as follows.

(a) Suppose \(S\) is a nonempty bounded subset of \(\mathbb{R}\) with no supremum. Let \(a \in S\), and \(b\) an upper bound for \(S\).

(b) Define a gauge \(\delta\) on \([a, b]\) as follows. Let \(x \in [a, b]\). If \(x\) is an upper bound for \(S\), let \(y\) be a smaller upper bound for \(S\), and define \(\delta(x) = x - y\). And if \(x\) is not an upper bound for \(S\), let \(y \in S\) with \(x < y\) and set \(\delta(x) = y - x\). (To avoid using the Axiom of Choice, if \(x\) is an upper bound for \(S\) set \(\delta(x) = \frac{1}{n}\), where \(n\) is the smallest natural number for which \(x - \frac{1}{n}\) is an upper bound for \(S\); if \(x\) is not an upper bound for \(S\), let \(\delta(x) = \frac{1}{n}\), where \(n\) is the smallest natural number for which there is a \(y \in S\) with \(x + \frac{1}{n} \leq y\).)

(c) Show that there is no \(\delta\)-fine partition of \([a, b]\).

#### III.7.4.3.
Explain why the name “Creeping Lemma” is appropriate. (See [Pét38].)
III.8. Representation of Real Numbers

III.8.1. Decimal Expansions

“It is India that gave us the ingenious method of expressing all numbers by means of ten symbols, each symbol receiving a value of position as well as an absolute value; a profound and important idea which appears so simple to us now that we ignore its true merit. But its very simplicity and the great ease which it has lent to computations put our arithmetic in the first rank of useful inventions; and we shall appreciate the grandeur of the achievement the more when we remember that it escaped the genius of Archimedes and Apollonius, two of the greatest men produced by antiquity.”

Pierre-Simon Laplace

The most familiar way to represent real numbers is by finite or infinite decimals. In fact, we are accustomed to thinking of real numbers as being decimals. For example, we are used to writing expressions like

\[
\frac{1}{4} = 0.25
\]

\[
\frac{1}{3} = 0.333333333\cdots
\]

\[
\pi = 3.1415926\cdots
\]

even though the precise meaning of the last expression is rather obscure (the first expression, and even the second, has a fairly evident meaning).

It is better from a logical standpoint to regard decimals as merely convenient representations of real numbers. Even natural numbers should not really be thought of as being decimal expressions, since it is sometimes convenient to represent them in other ways. For example, are

\[
2 , \quad 1.999999\cdots , \quad \sqrt{4} , \quad \int_0^\pi \sin x \, dx , \quad \sum_{k=0}^{\infty} \frac{1}{2k}
\]

all natural numbers? Are they the same number? And one runs into immediate logical difficulties in saying that rational numbers are the same thing as their decimal expansions.

In this section, we will carefully develop the theory of decimal expansions, and expansions to other bases, and show that there is a (near) one-to-one correspondence between real numbers and infinite decimals. The arguments are nice illustrations of the Completeness Axiom in various forms, particularly the Nested Intervals Property.

Bases

III.8.1.1. The first ingredient in decimal representations is to fix a base \( b \), which is a natural number greater than 1. The idea is then to write all real numbers in terms of powers of \( b \).
It is pretty universal in modern human societies to use as base the number we call ten. In principle, this base is no better or worse than any other choice of base; it almost certainly came into general use simply because most of us have ten fingers (the basic meaning of the word digit is “finger or toe”). Indeed, the word decimal technically only correctly applies to base ten expansions, although we will, by abuse of language, use it more generally for expansions with different bases.

Other bases have occasionally been used. Besides ten, at least the numbers five, twelve, twenty, forty, and sixty (and reportedly also four, six, fifteen, twenty-four, and twenty-seven) have been used at various times by different societies. There are vestiges of these systems in the names of certain numbers in some modern-day languages (including the word score in English, Russian kopok, Danish halvtreds, and French numbers like quatre-vingt-onze), and in our units of time and measurement of angles. The notorious English system of measurement (and, until recently, money), can also be regarded as a number system with different base (actually more than one!)

I remember as a boy being fascinated by a book advocating use of a base twelve (“duodecimal”) system. The main argument was that arithmetic is simpler in this base since twelve has more divisors than ten; the multiplication table, although somewhat larger, is more systematic and easier to remember, and many more rational numbers have terminating decimals. Although these claims are arguably true, the duodecimal movement never got anywhere.

With the advent of computers, it has become natural in certain settings to use binary (base 2) or hexadecimal (base 16) expansions since they match up nicely with computer architecture and hardware. Binary “decimal” digits are called bits, and groups of binary digits of fixed length (usually 8) are called bytes. Terms such as megabyte and gigabyte are now in common use in the language.

“There are only 10 kinds of people in the world: those who understand binary, and those who don’t.”

Seen on a T-Shirt

In describing decimal expansions, we will generally express everything in the usual base ten. The exposition can easily be adapted to any base b by just replacing 10 by b. We will use the usual base ten digit symbols 0 through 9; for another base b, if b < 10 the digits 0 through b − 1 can be used, and if b > 10 additional symbols must be added. For example, in binary only the digits 0 and 1 are used, and in hexadecimal the letters A through F are customarily used to denote the numbers ten through fifteen, which are single digits in hexadecimal.

It is theoretically possible to vary the base from place to place (e.g. to use base ten for the first decimal digit, base 7 for the second, etc.); it seems almost inconceivable that this would ever be useful in practice, but the English essentially did this for centuries with money, and we still do it with time, and, in the U.S., other measurements. (The theory of continued fractions gives a much more useful version of varying bases; there the “base” varies not only from place to place, but from number to number.) The interested reader can figure out how to adapt the arguments to this case.

There is also a fascinating generalization of decimal expansions where the base b is not required to be a natural number, but can be an arbitrary real or complex number with |b| > 1, and the “digits” are a finite set of real or complex numbers containing 0. Such expansions are intimately connected with the theory of fractal sets in the plane. See [Edg08, §1.6] for details.
Decimal Representations of Integers

Finite Decimals

III.8.1.6. We first review the meaning of a finite decimal. This is an expression of the form

$$\pm m.d_1d_2 \cdots d_n$$

where \(m\) is a nonnegative integer and \(d_1, \ldots, d_n\) are decimal digits between 0 and 9. This expression by definition represents the rational number

$$\pm \left[ m + \frac{d_1d_2 \cdots d_n}{10^n} \right]$$

where \(d_1d_2 \cdots d_n\) denotes a nonnegative integer in the usual way:

$$d_1d_2 \cdots d_n = 10^{n-1}d_1 + 10^{n-2}d_2 + \cdots + 10d_{n-1} + d_n.$$

Alternatively, we have

$$\pm m.d_1d_2 \cdots d_n = \pm \left[ m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n} \right] = \pm \left[ m + \sum_{k=1}^{n} \frac{d_k}{10^k} \right].$$

The \(m\) of course can also be written in decimal form in a similar manner.

III.8.1.7. Conversely, any rational number of the form

$$\frac{p}{10^q}$$

where \(p, q \in \mathbb{Z}\) has a finite decimal representation: suppose \(p > 0\) (the case \(p \leq 0\) is an obvious variation), and \(q > 0\), which can be arranged by multiplying top and bottom by a power of 10. Then we can uniquely write

$$p = 10^qm + r$$

where \(0 \leq r < 10^q\), and we can then expand

$$r = 10^{q-1}d_1 + 10^{q-2}d_2 + \cdots + 10d_{q-1} + d_q$$

for digits \(d_1, \ldots, d_q\). Thus the rational numbers whose denominator is a power of 10 (more precisely, the rational numbers which have such a fractional representation) are precisely the numbers with a finite decimal representation.

III.8.1.8. Definition. A rational number which can be written with a denominator which is a power of 10 is a decadent rational.

Any integer, or any rational number which can be written with a denominator divisible only by 2 and/or 5 is decadent. Every decadent rational is an integer times a power of 10 (and conversely). The set of decadent rational numbers is dense in \(\mathbb{R}\) (in the sense that every interval in \(\mathbb{R}\) contains such a number, in fact infinitely many).
There is one obvious sense in which the finite decimal expression for a decadent rational number is not unique, however: we have

\[ m.d_1d_2\cdots d_n = m.d_1d_2\cdots d_n00\cdots 0 \]

for any (finite) string of zeros. Thus any number of zeros can be appended or deleted at the end of the expression. Except for this, the finite decimal expansion of a decadent rational number is unique. (We can theoretically also add zeros in places to the left of the first digit in the integer part, but it is conventional not to write such zeros.)

The ordering on the decadent rational numbers is obtained easily from the decimal expansion. Working from left to right in the decimal expansions of (nonnegative) \( x \) and \( y \), just find the first place where the expansions differ; the larger digit in this place denotes the larger number. (This is an effective procedure since there are only finitely many nonzero digits in each expansion; the expansions may have to be extended by adding zeros on the left and/or right ends to allow comparison.)

The Decimal Representation of a Real Number

We will show that every real number \( x \) has a decimal expansion of the form

\[ \pm m.d_1d_2\cdots \]

where \( m \) is a nonnegative integer and the \( d_i \) are digits between 0 and 9, that the decimal expansion is unique except for decadent rational numbers which have two such expansions, and that every such decimal expression is the decimal expansion of a unique real number.

We will require that the infinite decimal expansion of a decadent rational number be compatible with the finite expansion, i.e. if

\[ x = \pm m.d_1d_2\cdots d_n \]

we should have that

\[ \pm m.d_1d_2\cdots d_n00\cdots \]

is an infinite decimal expansion for \( x \).

We also want the infinite decimal expansion to reflect the ordering on \( \mathbb{R} \) in the same way as the finite expansions of III.8.1.10. We must be careful, however: although the algorithm described in III.8.1.10 makes sense in this setting (i.e. it is a well-defined finite algorithm), it does not always give the correct answer due to the nonuniqueness of certain infinite decimal expansions (\( c \)). We can, however, say the following: suppose

\[ m.d_1d_2\cdots \]

is an infinite decimal expansion for a real number \( x \). Then, for any \( n \), we have

\[ m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n} \leq x \leq m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n + 1}{10^n} \]

If \( d_n \neq 9 \), this inequality can be concisely written

\[ m.d_1d_2\cdots d_n \leq x \leq m.d_1d_2\cdots d_n c_n \]

where \( c_n = d_n + 1 \). If \( d_n = 9 \), however, the right-hand inequality does not make sense, and it is notationally very complicated to write a valid version (one must “carry” back past any 9’s at the end of the expression).

As we will see, these two provisions completely determine decimal expansions and give a hint how to find them.
III.8.1.12. The sign in an infinite decimal expression is self-explanatory. From now on, we will only deal with nonnegative real numbers; the decimal expansion of a negative number $x$ will be the decimal expansion of $|x|$ preceded by a minus sign.

III.8.1.13. If $x$ is nonnegative, the $m$ is also clear: it is the unique nonnegative integer such that $m \leq x < m + 1$. This $m$ is called the integer part of $x$, denoted $\lfloor x \rfloor$.

III.8.1.14. By subtracting off the integer part, we may reduce the problem of defining (or computing) the decimal expansion of a number $x$ to the case where $0 \leq x \leq 1$. We will make this restriction from now on.

Generating the Decimal Expansion of a Real Number
Finding a Real Number with a Given Decimal Expansion
Decimal Expansions of Rational and Irrational Numbers
Defining Real Numbers as Decimals

III.8.1.15. Since there is essentially a one-one correspondence between real numbers and decimal expressions, it is possible to define the real numbers via infinite decimals, and this approach is taken in some references. But there are some serious difficulties with this approach.

III.8.1.16. Besides the general problem of defining the real numbers as being one particular model, be it decimal expansions, Dedekind cuts, equivalence classes of Cauchy sequences, or whatever, there are two additional problems in defining the real numbers via decimal expressions:

(i) The correspondence is not quite one-to-one.

(ii) It is not easy or obvious how to define or describe the algebraic operations on $\mathbb{R}$ using the decimal expressions.

III.8.1.17. The first difficulty is easily overcome. One way is just to exclude decimals ending in an infinite string of 9's.

III.8.1.18. The second difficulty is not easily overcome. In school we learn algorithms for adding, subtracting, multiplying, and dividing integers expressed in decimal notation, including memorization of the multiplication table for one-digit natural numbers; later, we also learn the simple modifications needed to use the same algorithms for arithmetic of finite (i.e. terminating) decimals. But there are no finite algorithms for adding or multiplying infinite decimals, since there is no place to “start”. For example, can you even determine the first decimal digit of

$$x = (0.2531486073\cdots) + (0.3468513926\cdots)$$

without further information? Could you even be sure this digit can be determined in a finite length of time with a finite amount of information about the decimal expansions? (The problem is even worse, and more obvious, for multiplication.)
III.8.1.19. The only reasonable way to proceed is to appeal to one of the other approaches, most obviously the Cauchy sequence approach. In the example of the last paragraph, although even the first decimal digit of \( x \) is not clearly determined by the information given, \( x \) itself is determined to a high degree of accuracy (within \( 2 \times 10^{-10} \)) by the ten decimal places of the given expressions. If one has the complete decimal expansions of the summands, or an algorithm for generating them, the number \( x \) can be determined to arbitrary accuracy, i.e. a sequence of rational numbers (terminating decimals) can be generated which converges to \( x \).

III.8.1.20. The problem is that real numbers which are very close together can have different expansions starting early in the expression if they happen to be on opposite sides of a number with a small terminating expansion (i.e. an integer times a small negative power of 10).

Another way of expressing the difficulty is that if we know that real numbers \( x \) and \( y \) satisfy \( |x - y| < 10^{-n} \), we do not necessarily know that \( x \) and \( y \) agree to \( n \), or even \( n - 1 \), decimal places, although they “usually” will; in unusual cases they may not even agree to one decimal place if one expansion has a long string of 0’s after the first place and the other a corresponding long string of 9’s. They will not even necessarily agree when rounded to one decimal place, e.g. if \( x = 0.24999999999999999 \) and \( y = 0.25000000000000001 \). In fact, for any \( n \), the functions which truncate or round a real number to \( n \) decimal places are step functions with jump discontinuities of height \( 10^{-n} \).

III.8.2. Continued Fractions

A somewhat more complicated, but in some ways more efficient, way to represent real numbers by sequences of natural numbers is by continued fractions. The theory of continued fractions is usually, and rightly, considered to be a part of number theory; however, the applications in analysis and probability, along with applications in closely related fields such as dynamical systems, are important enough that the basic theory of continued fractions should be a more integral part of the standard analysis curriculum than is currently the case.

While continued fractions give a much better way of representing real numbers than decimal representations for some purposes (such as finding good rational approximations), they are not nearly as good as decimal representations for the purposes of doing arithmetic, and this shortcoming is the principal reason their use has not become so widespread or popular. But mathematicians need to be familiar with both ways of representing numbers since they are useful for different purposes.

We will cover only a few important basic properties of the theory of continued fractions here which are most relevant in analysis and probability, and omit most of the many fascinating connections with number theory (some of these are described at the end of the section and in the Exercises). Books on number theory such as the classic [HW08] (there are also many others; perhaps my favorite is the little gem [Khi97]) cover continued fractions in more detail; [Sta78] has a thorough and geometrically appealing treatment of the theory.

The fundamental difference between decimal representations and continued fractions is that decimal representations require an initial choice of a base, and all representations are tied to that choice of base; some numbers have nice representations with respect to the chosen base, while others do not. But the continued fraction representation of a number is not tied to any initial choice of base; indeed, one way of thinking of the theory is that at each step in the expansion of a number, the perfect “base” is used for that number at that step.

As a result, the theory of continued fractions concerns the most efficient approximation of real numbers by rational numbers. Of course, the finite parts of the decimal expansion of a real number give rational
approximations, but one can do much better in general by a different approach. Any real number can be approximated arbitrarily closely by rational numbers, but close rational approximations to irrational numbers will necessarily have large denominators. By an “efficient” approximation we will mean an approximation which is close by some measure depending on the size of the denominator of the rational number when expressed in lowest terms. For example, we could ask:

III.8.2.1.  **Question.** For an irrational number \( x \), and a natural number \( n \), what is the rational number with denominator \( \leq n \) which most closely approximates \( x \)?

The theory of continued fractions gives a complete answer to this question.

III.8.2.2.  A **continued fraction** is an expression of the form

\[
x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_n}}}}}.
\]

This is just a formal expression, which may be finite or infinite. In interpreting the meaning of this expression, the \( x_k \) may stand for numbers, functions, indeterminates, \( \ldots \). For example, there has been a large mathematical industry in expanding functions in continued fractions; cf. [Wal48], [JT80], [LW08]. For us, the \( x_k \) will always stand for positive real numbers, usually natural numbers.

More general continued fraction expressions of this type are sometimes considered, where the numerators are not necessarily 1’s; in this case, continued fractions with 1’s in the numerators (the only kind we will consider) are called **simple continued fractions**.

III.8.2.3.  **Definition.** A **finite continued fraction** of depth \( n \in \mathbb{N} \) is an expression of the form

\[
x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_n}}}}}.
\]

The theory of continued fraction expansions of real numbers is based on the following “obvious” fact, which can be easily proved by induction (Exercise (i)).

III.8.2.4.  **Proposition.** (i) Let \( x_1, \ldots, x_n \) be positive real numbers, and \( x_0 \) a real number. Then the expression

\[
x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_n}}}}}.
\]

unambiguously defines a real number by the usual rules of algebra. If \( x_0, \ldots, x_n \) are rational, then the number defined by the expression is rational.
(ii) If \( x_1, \ldots, x_n, x'_n \) are positive real numbers, and \( x_0 \) is a real number, and
\[
\begin{align*}
x_0 + \frac{1}{x_1 + \frac{1}{x_2 + \frac{1}{x_3 + \cdots}}} &= x_0 + \frac{1}{x_1 + \frac{1}{x_2 + \frac{1}{x_3 + \cdots}}} \\
&= x_n \quad \text{for all } n.
\end{align*}
\]
then \( x_n = x'_n \).

Statement (i) can fail if the \( x_k \) \((k \geq 1)\) are not positive, since then a denominator of zero can arise during simplification.

It is notationally clumsy to write these complex fractions; we will use the common notation \([x_0; x_1, \ldots, x_n]\) for the fraction in (i). Variations of this notation, and many other notations for continued fractions, are used in the literature.

We now define a function which will be used repeatedly:

III.8.2.5. Definition. Let \( x \) be a real number. Then the \textit{integer part} of \( x \), denoted \([x]\), is the largest integer which is \( \leq x \). We can now inductively define the continued fraction expansion of a real number:

III.8.2.6. Definition. Let \( x \) be a real number. Define numbers as follows:

Set \( a_0 = x \) and \( r_0 = a_0 = \lfloor a_0 \rfloor \). Let \( p_0 = r_0 \) and \( q_0 = 1 \).

Set \( \beta_1 = a_0 - a_0 = x - r_0 \). Then \( 0 \leq \beta_1 < 1 \).

If \( \beta_1 = 0 \) (i.e. if \( x \in \mathbb{Z} \)), terminate the expansion. Otherwise set \( \alpha_1 = \frac{1}{\beta_1} \) and \( a_1 = \lfloor \alpha_1 \rfloor \). Then \( a_1 \in \mathbb{N} \).

Set \( r_1 = a_0 + \frac{1}{a_1} = \frac{a_0a_1 + 1}{a_1} \). Then \( r_1 \in \mathbb{Q} \), and \( r_0 < x \leq r_1 \). Set \( p_1 = a_0a_1 + 1 \), \( q_1 = a_1 \), so \( r_1 = \frac{p_1}{q_1} \), and define \( \beta_2 \) by \( \beta_2 = a_1 - a_1 \), i.e. \( x = a_0 + \frac{1}{a_1 + \beta_2} \). Then \( 0 \leq \beta_2 < 1 \).

If \( \beta_2 = 0 \) (i.e. if \( x = r_1 \)), terminate the expansion. Otherwise set \( \alpha_2 = \frac{1}{\beta_2} \) and \( a_2 = \lfloor \alpha_2 \rfloor \in \mathbb{N} \).

Set \( r_2 = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = \frac{p_2}{q_2} \). Then \( r_2 \leq x < r_1 \). Define \( \beta_3 = a_2 - a_2 \) (i.e. \( x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \beta_3}} \)). Then \( 0 \leq \beta_3 < 1 \).

Continue inductively defining \( \beta_n, \alpha_n, a_n, r_n = \frac{p_n}{q_n} \) for all \( n \) (or until the process terminates). The finite or infinite sequence \([a_0; a_1, a_2, \ldots]\) is called the \textit{continued fraction} (expression, expansion, decomposition) of \( x \). The \( a_n \) are called the \textit{terms} (or \textit{elements}) of the continued fraction.

The numbers \( a_n, r_n, p_n, q_n, \alpha_n, \beta_n \) depend on \( x \); we will write them \( a_n(x) \), etc., when it is necessary to specify the dependence on \( x \).

The name “continued fraction” for this construction is appropriate, since the idea is to write
\[
x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}}.
\]

We will discuss later how this expression can be precisely interpreted as the limit of the finite continued fractions \( r_n \).

It is easy to verify the next result (note that we have a recursive formula \( a_n = a_n + \frac{1}{n_{n+1}} \) for all \( n \)):
III.8.2.7. PROPOSITION. Let $x \in \mathbb{R}$, and $n \in \mathbb{N}$. Then
\[
x = [a_0(x); a_1(x), \ldots, a_{n-1}(x), a_n(x)]
\]
\[
r_n(x) = [a_0(x); a_1(x), \ldots, a_n(x)].
\]

We thus get the following “contraction” principle:

III.8.2.8. PROPOSITION. Let $x \in \mathbb{R}$, and $n \in \mathbb{N} \cup \{0\}$. If the continued fraction construction for $x$ lasts at least $n + 1$ steps (i.e. $a_n(x)$ is defined), then $a_n(x) = a_k(a_{n-k}(x))$ for $0 \leq k \leq n$.

We then get a recursive formula for $r_n$:

III.8.2.9. COROLLARY. Let $x \in \mathbb{R}$, and $n \in \mathbb{N} \cup \{0\}$. If the continued fraction construction for $x$ lasts at least $n + 1$ steps, then
\[
r_n(x) = a_0 + \frac{1}{r_{n-1}(\alpha_1(x))} = a_0 + \frac{1}{a_1 + \frac{1}{r_{n-2}(\alpha_2(x))}} = \cdots = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{r_{n-3}(\alpha_3(x))}}} = \cdots + \frac{1}{r_0(\alpha_n(x))}.
\]

We also have the following uniqueness statement:

III.8.2.10. PROPOSITION. Let $n \in \mathbb{N}$, $b_0 \in \mathbb{Z}$, $b_1, \ldots, b_n \in \mathbb{N}$. If $x \in \mathbb{R}$, and the continued fraction expansion of $x$ lasts at least $n + 2$ steps (i.e. $a_{n+1}(x)$ is defined), then $a_k(x) = b_k$ for $0 \leq k \leq n$ if and only if $x = [b_0; b_1, \ldots, b_n, \alpha]$ for some $\alpha > 1$, i.e. if and only if $x$ is in the open interval between $[b_0; b_1, \ldots, b_n]$ and $[b_0; b_1, \ldots, b_{n-1}, b_n + 1]$.

III.8.2.11. There is one slight ambiguity in continued fraction expansions, however: for any $b_0 \in \mathbb{Z}$, $b_1, \ldots, b_n \in \mathbb{N}$,
\[
[b_0; b_1, \ldots, b_n, 1] = [b_0; b_1, \ldots, b_n + 1]
\]
and the second expression, not the first, is the correct continued fraction expansion of the corresponding rational number. Thus a finite continued fraction expression ending in 1 can be “contracted”, and a terminating continued fraction expansion (of length at least 2) cannot end in 1.

Since the $r_n$ are rational, it is obvious that if $x$ is irrational, then the construction never terminates, so the continued fraction expression for $x$ is an infinite sequence. Conversely, if $x$ is rational, then the construction necessarily terminates, although this is not immediately obvious:
III.8.2.12. **Proposition.** If $r$ is rational, then the continued fraction expansion of $x$ terminates.

**Proof:** Write $x = \frac{p}{q}$. We may suppose $p, q \in \mathbb{N}$ for simplicity (this will happen after one step if $x < 0$, and the result is obvious if $x = 0$). If $p < q$, then $a_0 = 0$ and $\frac{s}{q}$ with $0 \leq s < q \leq p$. If $s = 0$, the process terminates. Otherwise, $a_1 = \frac{q}{s} = a_1 + \frac{1}{\alpha_1}$ with $0 \leq t < s < q$. If $t = 0$, the process terminates. Otherwise, $\alpha_2 = \frac{s}{t}$, and we have $s, t \in \mathbb{N}$, $s < p$, $t < q$, so after 2 steps the remaining procedure computes the continued fraction expansion of a fraction with strictly smaller numerator and denominator. Thus the process must terminate in no more than $2 + 1 \leq 2q + 1$ steps.

In fact, for $x = \frac{p}{q}$ the continued fraction construction is really the same as the Euclidean Algorithm for determining the greatest common divisor of $p$ and $q$. The termination estimate of $1 + 2q$ steps is rather crude: it is known (LAMÉ’s Theorem) that if $p, q > 0$, then the number of steps to termination is bounded by

$$2 + 5 \min(\log_{10}(p), \log_{10}(q))$$

which is roughly 5 times the number of decimal digits in the smaller of $p$ and $q$.

III.8.2.13. If $[b_0, b_1, b_2, \ldots]$ is any sequence with $b_0 \in \mathbb{Z}$ and $b_n \in \mathbb{N}$ for $n \geq 1$, we can define $r_n$ in the analogous manner, i.e.

$$r_n = [b_0, b_1, \ldots, b_n] = b_0 + \frac{1}{b_1 + \frac{1}{b_2 + \frac{1}{\ddots + \frac{1}{b_n}}}}.$$ 

By III.8.2.10., the $b_n$ can be recovered from the continued fraction expansions of the $r_n$.

III.8.2.14. In III.8.2.6., we were not completely precise about the definition of $p_n$ and $q_n$. These are defined by simply converting the compound fraction to a simple fraction and taking numerator and denominator. For example, we have

$$r_2 = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = a_0 + \frac{a_2}{a_1 a_2 + 1} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_1 a_2 + 1}$$

so

$$p_2 = a_0 a_1 a_2 + a_0 + a_2 = p_0 + a_2 p_1, \quad q_2 = a_1 a_2 + 1 = q_0 + a_2 q_1.$$ 

In fact, if $[a_0; a_1, a_2, \ldots]$ is any sequence with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$, and

$$r_n = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{\ddots + \frac{1}{a_n}}}}$$

and $p_n, q_n$ are defined from $r_n$ by clearing complex fractions, the same formulas hold for $n = 2$. 

324
III.8.2.15. Proposition. For any $n \geq 2$,

(i) $p_n = p_{n-2} + a_n p_{n-1}$ and $q_n = q_{n-2} + a_n q_{n-1}$.

(ii) $p_{n-1}q_{n-2} - p_{n-2}q_{n-1} = (-1)^n$.

In particular, $p_n$ and $q_n$ are relatively prime, and $q_n > q_{n-1}$.

In particular, these formulas hold for a sequence arising as the continued fraction of a real number. (We will see later (III.8.2.19.) that every sequence occurs as a continued fraction.)

Proof: We prove (i) by induction, using III.8.2.8. and III.8.2.9. The formula holds for $n = 2$ by the previous calculation. Suppose it holds for $n$ for all $x$. Fix $x$, and let $\alpha_1 = \alpha_1(x)$. Then, from III.8.2.9., $p_{k-1}(\alpha_1) = q_k(x)$ and $q_{k-1}(\alpha_1) = p_k(x) - a_0(x)q_k(x)$ for all $k$. Then

$$p_{n+1}(x) = a_0(x)p_n(\alpha_1) + q_n(\alpha_1).$$

By the inductive hypothesis applied to $\alpha_1$, this equals

$$a_0(x)[p_{n-2}(\alpha_1) + a_n(\alpha_1)p_{n-1}(\alpha_1)] + [q_{n-2}(\alpha_1) + a_n(\alpha_1)q_{n-1}(\alpha_1)]$$

$$= a_0(x)[q_{n-1}(x) + a_{n+1}(x)q_n(x)]$$

$$+ ([p_{n-1}(x) - a_0(x)q_{n-1}(x)] + a_{n+1}(x)[p_n(x) - a_0(x)q_n(x)])$$

$$= p_{n-1}(x) + a_{n+1}(x)p_n(x).$$

A similar calculation shows that

$$q_{n+1}(x) = q_{n-1}(x) + a_{n+1}(x)q_n(x).$$

We now prove (ii) by induction. It holds for $n = 2$ by direct calculation. If it holds for $n$, then by (i)

$$p_nq_{n-1} - p_{n-1}q_n = (p_{n-2} + a_n p_{n-1})q_{n-1} - p_{n-1}(q_{n-2} + a_n q_{n-1})$$

$$= p_{n-2}q_{n-1} - p_{n-2}q_{n-1} + a_n p_{n-1}q_{n-1} - a_n p_{n-1}q_{n-1}$$

$$= -(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = -(-1)^n = (-1)^{n+1}$$

so the formula holds for $n + 1$.

It follows from (ii) that $p_n$ and $q_n$ are relatively prime, and it is obvious from (i) that $q_n > q_{n-1}$ since $q_{n-2} \geq 1$ and $a_n \geq 1$. \(\Box\)

III.8.2.16. Corollary. (i) Let \(a_0; a_1, \ldots\) be a sequence with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$, and let $(r_n)$ be the corresponding sequence of finite continued fractions. Then

$$r_0 < r_2 < r_4 < \cdots < r_5 < r_3 < r_1$$

and $|r_n - r_{n-1}| = \frac{1}{q_n q_{n-1}}$ for $n \geq 1$.
(ii) Let \( x \in \mathbb{R} \) with nonterminating continued fraction expansion \([a_0; a_1, \ldots]\), and \((r_n)\) the corresponding sequence of finite continued fractions. Then
\[
\begin{align*}
r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1
\end{align*}
\]
and \( r_n \to x \). In fact, for \( n \geq 1 \),
\[
|x - r_n| < |r_{n+1} - r_n| = \frac{1}{q_n q_{n+1}} < \frac{1}{a_{n+1} q_n^2} \leq \frac{1}{q_n^2}
\]
(the extreme inequality also obviously holds for \( n = 0 \)).

Because of (ii), \( r_n \) is often called the \( n \)'th continued fraction convergent to \( x \), and we may justifiably write
\[
x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \ddots}}}
\]

An alternate way of phrasing the convergence is that the alternating series
\[
a_0 + \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{q_n q_{n-1}}
\]
converges to \( x \). (Note that this series is absolutely convergent since \( q_n \geq n \) for all \( n \).)

III.8.2.17. **Corollary.** If \( x \) and \( y \) are distinct real numbers, their continued fraction expansions are different.

III.8.2.16. almost shows that every sequence as in (i) is the continued fraction expansion of a (necessarily irrational) real number. For under the conditions of (i), by the Nested Intervals Property there is a unique \( x \in \mathbb{R} \) satisfying
\[
r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1
\]
and we need only show that the continued fraction expansion of this \( x \) is \([a_0; a_1, \ldots]\). This follows from the next proposition, which is essentially a rephrasing of III.8.2.10.: 

III.8.2.18. **Proposition.** Let \([b_0; b_1, \ldots]\) be a sequence with \( b_0 \in \mathbb{Z} \) and \( b_n \in \mathbb{N} \) for \( n \geq 1 \), and let \((r_n)\) be the corresponding sequence of finite continued fractions. If \( x \) is between \( r_n \) and \( r_{n+1} \), then \( a_k(x) = b_k \) for \( 0 \leq k \leq n \).

III.8.2.19. **Corollary.** Let \([a_0; a_1, \ldots]\) be a sequence with \( a_0 \in \mathbb{Z} \) and \( a_n \in \mathbb{N} \) for \( n \geq 1 \), and let \((r_n)\) be the corresponding sequence of finite continued fractions. Let \( x \) be the unique real number satisfying
\[
r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1
\]
Then the continued fraction expansion of \( x \) is \([a_0; a_1, \ldots]\).

We also have the following useful corollary about the continuity of continued fraction decompositions:
III.8.2.20. Corollary. Let \( x \in \mathbb{R} \) have continued fraction expansion of length at least \( n + 1 \), and \((r_k)\) the corresponding convergents. If \( y \) is between \( r_n \) and \( r_{n+1} \), then \( a_k(y) = a_k(x) \) for \( 0 \leq k \leq n \).

We have the following relations between \( x \) and the \( \alpha_k(x) \), which can be easily proved by induction using III.8.2.15. (i):

III.8.2.21. Proposition. Let \( x \in \mathbb{R} \), and \( n \in \mathbb{N} \). If the continued fraction construction for \( x \) lasts at least \( n + 1 \) steps, we have

\[
x = \frac{\alpha_1 a_0 + 1}{\alpha_1}
\]

and, if \( n \geq 2 \),

\[
x = \frac{\alpha_n p_{n-1} + p_{n-2}}{\alpha_n q_{n-1} + q_{n-2}}.
\]

This gives a refinement of III.8.2.16. (ii) with a lower bound:

III.8.2.22. Corollary. Let \( x \) be an irrational number, and \( r_n = \frac{p_n}{q_n} \) its \( n \)'th continued fraction convergent. Then

\[
\frac{1}{q_n(q_n + q_{n+1})} < |x - r_n| < \frac{1}{q_nq_{n+1}}.
\]

We combine and summarize our main results in a theorem:

III.8.2.23. Theorem. The continued fraction expansion establishes a one-one correspondence between the set \( \mathcal{J} \) of irrational numbers and the set of sequences \([a_0; a_1, a_2, \ldots]\) with \( a_0 \in \mathbb{Z} \) and \( a_n \in \mathbb{N} \) for \( n \geq 1 \). If \((x_m)\) is a sequence of irrational numbers and \( x \) an irrational number, then \( x_m \to x \) if and only if, for each \( n \), there is an \( m_0 \) such that \( a_n(x_m) = a_n(x) \) for all \( m \geq m_0 \).

Approximation by Continued Fractions

We conclude with a survey of some number-theoretic results about continued fractions, mostly stated without proof.

The continued fraction convergents to an irrational number \( x \) are the “best” rational approximants to \( x \) in the sense of Question III.8.2.1. We first make a couple of definitions:
III.8.2.24. Definition. Let \( x \) be an irrational number. A rational number \( \frac{p}{q} \) with \( p \in \mathbb{Z}, q \in \mathbb{N} \) is a good approximation to \( x \) if
\[
\left| x - \frac{p'}{q'} \right| > \left| x - \frac{p}{q} \right|
\]
for all \( p', q' \in \mathbb{Z}, 1 \leq q' \leq q, (p', q') \neq (p, q) \).
A rational number \( \frac{p}{q} \) with \( p \in \mathbb{Z}, q \in \mathbb{N} \) is a best approximation to \( x \) if
\[
|q'x - p'| > |qx - p|
\]
for all \( p', q' \in \mathbb{Z}, 1 \leq q' \leq q, (p', q') \neq (p, q) \).
A best approximation to \( x \) is obviously a good approximation to \( x \). Good approximations and best approximations are sometimes respectively called “best approximations of the first kind” and “best approximations of the second kind.”

III.8.2.25. Theorem. The continued fraction convergents \( \frac{p_n}{q_n} \) to an irrational number \( x \) for \( n \geq 1 \) are precisely the best approximations to \( x \), i.e. every continued fraction convergent is a best approximation, and every best approximation is a continued fraction convergent.

Not every good approximation is a continued fraction convergent, but all good approximations to \( x \) are secondary convergents, numbers of the form \( [a_0; a_1, \ldots, a_{n-1}, c_n] \), where \( [a_0; a_1, \ldots] \) is the continued fraction expansion of \( x \) and \( 1 \leq c_n \leq a_n \). If \( a_n > 1 \) and \( a_n/2 < c_n \leq a_n \), then the secondary convergent is a good approximation; if \( a_n \) is even and \( c_n = a_n/2 \), the secondary convergent is sometimes, but not always, a good approximation, cf. [].

These solve Question III.8.2.1.: since the solution to III.8.2.1. for a given \( n \) is obviously a good approximation, one only needs to find the good approximation with the largest denominator \( \leq n \). One only needs to know the continued fraction expansion to the point where \( q_k \geq n \).

The next theorem is an important result about continued fractions for number theory purposes, and further justifies the statement that the continued fraction convergents to a real number are the “best” rational approximants.

III.8.2.26. Theorem. Let \( x \) be an irrational number, and, for \( n \in \mathbb{N} \), let \( \frac{p_n}{q_n} \) be the \( n \)’th convergent to \( x \).
(i) If \( n \geq 2 \), then at least one of the following inequalities holds:
\[
|x - r_{n-1}| < \frac{1}{2q_n^2} \quad \text{or} \quad |x - r_n| < \frac{1}{2q_n^2}.
\]
(ii) Conversely, if \( r = \frac{p}{q} \ (p, q \in \mathbb{Z}) \) satisfies
\[
|x - r| < \frac{1}{2q^2}
\]
then \( r = r_n \) for some \( n \).

Proof: We prove only (i). Since \( x \) is between \( r_{n-1} \) and \( r_n \), we have
\[
|x - r_{n-1}| + |x - r_n| = |r_n - r_{n-1}| = \frac{1}{q_n q_{n-1} q_{n+1}}.
\]
If the statement is false, i.e.

\[ |x - r_{n-1}| \geq \frac{1}{2q_{n-1}^2} \quad \text{and} \quad |x - r_n| \geq \frac{1}{2q_n^2} \]

we would have

\[ \frac{1}{2q_{n-1}^2} + \frac{1}{q_n^2} \leq \frac{1}{q_{n-1}q_n} \]

and multiplying both sides by \(2q_{n-1}q_n^2\), we have

\[ q_n^2 + q_{n-1}^2 \leq 2q_nq_{n-1} \]

\[ (q_n - q_{n-1})^2 \leq 0 \]

which is a contradiction since \(q_n \neq q_{n-1}\).

Actually, one can do a little better:

**III.8.2.27.** **Theorem.** Let \(x\) be an irrational number, and, for \(n \in \mathbb{N}\), let \(r_n = \frac{p_n}{q_n}\) be the \(n\)’th convergent to \(x\). Then for infinitely many \(n\) (at least one of every three consecutive \(n\)),

\[ |x - r_n| < \frac{1}{\sqrt{5}q_n^2} . \]

The example of the golden ratio (Exercise (\ref{ex:golden_ratio})) shows that \(\sqrt{5}\) is the largest possible constant \(M\) for which

\[ |x - r_n| < \frac{1}{Mq_n} \]

for infinitely many \(n\) in general.

**III.8.2.28.** If the \(a_n(x)\) are unbounded, which is the case for almost all \(x\) (III.8.2.29.), we can do much better, since if \(a_{n+1}(x)\) is large, \(r_n(x)\) is a very good approximation to \(x\) by III.8.2.16.. For example, since \(a_4(\pi) = 292\), we have that \(r_3(\pi) = \frac{355}{113}\) satisfies

\[ \left| \pi - \frac{355}{113} \right| < \frac{1}{292 \cdot 113^2} \approx 2.7 \times 10^{-7} \]

so \(\frac{355}{113}\) agrees with \(\pi\) to (at least, in fact exactly) 6 decimal places. A theorem of **Lochs** (\ref{thm:lochs}) says that for almost all \(x \in \mathbb{J}\) (in the sense of Lebesgue measure), \(r_n(x)\) asymptotically agrees with \(x\) to about \(n\) decimal places.

**Asymptotic Behavior of Continued Fractions**

Here are some asymptotic results about the growth rate of the coefficients \(a_n(x)\) for “typical” irrational numbers:
III.8.2.29. **Theorem.** Let \( \phi : \mathbb{N} \to (1, \infty) \) be a function. Then

(i) If \( \sum_{n=1}^{\infty} \frac{1}{\phi(n)} \) diverges, then the set of \( x \in \mathbb{J} \) for which \( a_n(x) = O(\phi(n)) \) as \( n \to \infty \) has Lebesgue measure 0. In particular, the set of \( x \in \mathbb{J} \) for which \( a_n(x) \) is bounded has Lebesgue measure 0.

(ii) If \( \sum_{n=1}^{\infty} \frac{1}{\phi(n)} \) converges, then \( a_n(x) = o(\phi(n)) \) as \( n \to \infty \) for almost all \( x \in \mathbb{J} \), i.e. the set of \( x \) for which this fails has Lebesgue measure 0.

Thus, almost all irrational numbers have many \( a_n \) which are “large” but which do not grow too rapidly. Theorem III.8.2.29. is roughly a version of the Borel-Cantelli Lemma (it is not exactly the Borel-Cantelli Lemma since the random variables \( (a_n) \) are not quite independent; cf. ().) In contrast, Khinchin [Khi97] proved the following remarkable theorem, which is essentially a version of the Strong Law of Large Numbers ():

III.8.2.30. **Theorem.** For almost all \( x \in \mathbb{J} \), the geometric mean of the \( a_n \) approaches a constant

\[
\kappa = \prod_{k=1}^{\infty} \left[ 1 + \frac{1}{k(k+2)} \right]^{\log_2 k} = \prod_{k=1}^{\infty} k^{\log_2 \left[ 1 + \frac{1}{k(k+2)} \right]} = 2.685452 \cdots
\]

i.e.

\[
\lim_{n \to \infty} \left[ a_1(x)a_2(x) \cdots a_n(x) \right]^{1/n} = \kappa.
\]

So, for almost all irrational numbers \( x \), “most” of the \( a_n(x) \) are small. Note, however, that the arithmetic mean of the \( a_n \)’s almost always diverges: since \( \sum \frac{1}{n \log n} \) diverges, for almost all \( x \) we have \( a_n(x) > n \log n \) for infinitely many \( n \) by III.8.2.29.(i), and thus, for almost all \( x \),

\[
\frac{1}{n} \sum_{k=1}^{n} a_k(x) > \log n
\]

for infinitely many \( n \).

Khinchin’s result is more general:

III.8.2.31. **Theorem.** Let \( f : \mathbb{N} \to [0, \infty) \) be a function for which there exist positive constants \( C \) and \( \delta \) such that

\[
f(k) < Ck^{1/2-\delta}
\]

for all \( k \). Then, for almost all \( x \in \mathbb{J} \),

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(a_k(x)) = \sum_{k=1}^{\infty} f(k) \log_2 \left[ 1 + \frac{1}{k(k+2)} \right].
\]

III.8.2.30. is the case \( f(k) = \log k \). Another interesting case is to let \( f(k_0) = 1 \) for some \( k_0 \) and \( f(k) = 0 \) for \( k \neq k_0 \). This case shows that the number \( k_0 \) appears with the same asymptotic relative frequency in the continued fraction expansion of almost all numbers, and the frequency is

\[
\log_2 (k_0^2 + 2k_0 + 1) - \log_2 (k_0^2 + 2k_0).
\]

330
For example, for almost all \( x \), the relative frequency of 1 in the continued fraction expansion of \( x \) is
\[
\log_2 4 - \log_2 3 = .4150 \ldots
\]
i.e. asymptotically about 41.5% of the terms in the continued fraction expansion of \( x \) are 1’s. Similarly, about 17% are 2’s, 9.3% are 3’s, etc.

**Approximation of Algebraic and Transcendental Numbers**

**III.8.2.32. Definition.** A real (or complex) number \( x \) is **algebraic** if \( x \) is a root of a polynomial with integer coefficients. A real (or complex) number which is not algebraic is **transcendental**.

A rational number \( \frac{p}{q} \) is a root of the polynomial \( qx - p \), hence is algebraic. It is a standard theorem of abstract algebra that every algebraic number is a root of a unique (up to constant multiple) polynomial with integer coefficients which is irreducible over \( \mathbb{Z} \) (or equivalently over \( \mathbb{Q} \)). The degree of this polynomial is the degree of the algebraic number. Degree 1 algebraic numbers are precisely the rational numbers; degree 2 algebraic numbers are called **quadratic numbers**.

Our first result is a simple argument showing that irrational algebraic numbers cannot be approximated too closely by rational numbers. This result is due to Liouville in 1844.

**III.8.2.33. Theorem.** Let \( x \) be an algebraic number of degree \( n > 1 \). Then there is a constant \( C, 0 < C < 1/2 \), such that
\[
|x - r| > \frac{C}{q^n}
\]
for all rational numbers \( r = \frac{p}{q} \).

**Proof:** Let \( x \) be a root of an irreducible polynomial \( f \) of degree \( n \) with integer coefficients, and let \( M \) be the maximum of \( |f'| \) on the interval \( [x - \frac{1}{2}, x + \frac{1}{2}] \). Then \( M > 0 \) since \( f \) is not constant. By the Mean Value Theorem, if \( r = \frac{p}{q} \) satisfies \( |x - r| < \frac{1}{2} \), there is a \( c \) between \( x \) and \( r \) such that
\[
|q^n f(r)| = |q^n(f(r) - f(x))| = q^n |f'(c)||r - x| \leq Mq^n |x - r|.
\]
Note that \( q^n f(r) \in \mathbb{Z} \), and \( f(r) \neq 0 \) since \( f \) has no rational roots. Thus, if \( C < \frac{1}{M} \), then \( 1 \leq Mq^n |x - r| \), \( |x - r| > \frac{C}{q^n} \) if \( |x - r| < \frac{1}{2} \). Also, \( M > 2 \) since there is an \( s \in \mathbb{Z} \) with \( |x - s| < \frac{1}{2} \), and by the above argument
\[
1 \leq |f(s)| \leq M|x - s|
\]
so \( C < \frac{1}{2} \) and \( \left| x - \frac{p}{q} \right| > \frac{C}{q^n} \) also if \( \left| x - \frac{p}{q} \right| > \frac{1}{2} \).

This result shows that irrational numbers which can be approximated very closely by rational numbers must be transcendental.
III.8.2.34. Definition. An irrational number \( x \) is a Liouville number if, for every \( n \), there are \( p, q \in \mathbb{Z} \) with \( q \geq 2 \) and
\[
\left| x - \frac{p}{q} \right| < \frac{1}{q^n}.
\]

By III.8.2.33., every Liouville number is transcendental.

III.8.2.35. Example. Let \( x = \sum_{k=1}^{\infty} 10^{-k!} \). For each \( n \), we can take \( \frac{p}{q} = \sum_{k=1}^{n} 10^{-k!} \). This was Liouville’s original example, and the first explicit number to be proved transcendental.

The set \( L \) of Liouville numbers is quite an interesting subset of \( \mathbb{R} \). It is very “small” in a measure-theoretic sense, but “large” in a topological (category) sense:

III.8.2.36. Theorem. \( L \) is a dense \( G_\delta \); hence the complement of \( L \) is a meager set.

Proof: For each \( n \), let
\[
U_n = \bigcup_{q \in \mathbb{Q}} \left( \frac{p}{q} - \frac{1}{q^n}, \frac{p}{q} + \frac{1}{q^n} \right).
\]
Then \( U_n \) is an open set containing \( \mathbb{Q} \), hence dense; and \( L = \mathbb{J} \cap \cap_n U_n \). Since \( \mathbb{J} \) and \( \cap_n U_n \) are dense \( G_\delta \)'s, so is \( L \).

III.8.2.37. Theorem. The set \( L \) has \( s \)-dimensional Hausdorff measure 0, for every \( s > 0 \). So \( L \) has Hausdorff dimension 0 and Lebesgue measure 0.

Proof: It suffices to show that if \( 0 < s \leq 1 \) is fixed, the \( s \)-dimensional Hausdorff measure of \( L \cap (-m, m) \) is 0 for every \( m \in \mathbb{N} \), i.e. that for every \( \epsilon > 0 \) there is a sequence \( (I_k) \) of open intervals with \( \sum_k [\ell(I_k)]^s < \epsilon \) and \( L \cap (-m, m) \subseteq \cup I_k \). Fix \( m \in \mathbb{N} \) and \( \epsilon > 0 \). Then, for any \( n \),
\[
L \cap (-m, m) \subseteq U_n \cap (-m, m) = \bigcup_{q=2}^{mq} \bigcup_{p=-mq}^{m} \left( \frac{p}{q} - \frac{1}{q^n}, \frac{p}{q} + \frac{1}{q^n} \right).
\]
We have that
\[
\ell \left( \left( \frac{p}{q} - \frac{1}{q^n}, \frac{p}{q} + \frac{1}{q^n} \right) \right) = \frac{2}{q^n}
\]
so, for sufficiently large \( n \), we have the sum of the \( s \) powers of the lengths of the intervals is
\[
\sum_{q=2}^{\infty} \sum_{p=-mq}^{mq} \left( \frac{2}{q^n} \right)^s = \sum_{q=2}^{\infty} 2^s(2mq + 1) \frac{1}{q^{ns}} \leq \sum_{q=2}^{\infty} 2^s(2mq + q) \frac{1}{q^{ns}}
\]
\[
= 2^s(2m + 1) \sum_{q=2}^{\infty} \frac{1}{q^{ns}} \leq 2^s(2m + 1) \int_{1}^{\infty} \frac{1}{x^{ns}} \, dx = \frac{2^s(2m + 1)}{ns - 2} < \epsilon.
\]
It can be shown (Exercise (i)) that every Liouville number is in the complement of the set of III.8.2.29.(ii) for some φ, giving another proof that L has Lebesgue measure 0.

III.8.2.38. However, it is far from true that all transcendental numbers have rapidly converging continued fraction convergents (or, equivalently, rapidly growing continued fraction coefficients), as III.8.2.29.(ii) shows. In fact, there are transcendental numbers whose continued fraction convergents converge nearly as slowly as those of φ (Exercise III.8.4.2.), namely numbers whose continued fraction expansion is all 1’s except for some (infinitely many!) sparsely distributed 2’s (to see that there are transcendental numbers with such continued fraction expansions, it suffices to note that there are uncountably many such sequences of 1’s and 2’s no matter what your definition of “sparsely distributed” may be. Actually, it is conjectured that every algebraic number of degree > 2 has unbounded continued fraction expansion; if so, all numbers with bounded continued fraction expansion which is not eventually periodic (cf. Exercise III.8.4.4.) would be transcendental.) In fact, there appears to be no foolproof way to distinguish between algebraic and transcendental numbers by a statistical analysis of the pattern of the terms in the continued fraction expansion.

Liouville’s theorem has been improved by various authors, beginning with Thue in 1909. The best result known is due to Roth in 1955:

III.8.2.39. Theorem. Let x be an irrational algebraic number. Then for every ε > 0 there is a C > 0 such that

\[ |x - r| > \frac{C}{q^2 r^\epsilon} \]

for all \( r = \frac{p}{q} \in \mathbb{Q} \).

This can be used to construct other transcendental numbers (cf. Exercise (i)).

III.8.3. Measure Theory of Continued Fractions

In this section, we develop some measure-theoretic properties of continued fraction expansions, and the asymptotic behavior of the terms and convergents.

First, given \( n, k \in \mathbb{N} \), what is the probability that, for a randomly chosen \( x \), we have \( a_n(x) = k \)? To make this precise, let

\[ E_n^{(k)} = \{ x \in (0, 1) : a_n(x) = k \} \]

and let \( P_n^{(k)} \) be the Lebesgue measure of \( E_n^{(k)} \). Then \( P_n^{(k)} \) is a reasonable interpretation of the “probability” that \( a_n(x) = k \) for a randomly chosen \( x \).

III.8.3.40. The case \( n = 1 \) is the easiest to analyze. We have

\[ E_1^{(k)} = \left( \frac{1}{k+1}, \frac{1}{k} \right) \]

and so

\[ P_1^{(k)} = \frac{1}{k(k+1)} = \frac{1}{k^2 + k} \].

333
Similarly, if \(a_1(x)\) is small for a randomly chosen \(x\): in fact, the probability that \(a_1(x) \geq k\) is the Lebesgue measure of \((0, 1/k] = 1/k\).

It is not easy to give such a precise description of \(P_n^{(k)}\) for \(n > 1\). But in general, we have the following estimate:

\[
\frac{1}{3k^2} < P_n^{(k)} < \frac{2}{k^2}.
\]

III.8.3.41. **Theorem.** For any \(n\) and \(k\),

\[
\frac{1}{3k^2} < P_n^{(k)} < \frac{2}{k^2}.
\]

Note that the bounds are independent of \(n\).

**Proof:** The case \(n = 1\) follows from the previous calculation. Let \(n > 1\). Fix \(k_1, \ldots, k_{n-1} \in \mathbb{N}\), and let

\[
J(k_1, \ldots, k_{n-1}) = \{x \in (0, 1] : a_j(x) = k_j \text{ for } 1 \leq j \leq n-1\}.
\]

Then we have \(\lambda\) that \(J(k_1, \ldots, k_{n-1})\) is the interval with endpoints \(\frac{p_{n-1}}{q_{n-1}}\) and \(\frac{p_{n-1} + p_{n-2}}{q_{n-1} + q_{n-2}}\), where \(p_j\) and \(q_j\) are the numerator and denominator of the \(j\)th convergent to the finite continued fraction \([0; k_1, \ldots, k_{n-1}]\). Thus we have that the length of \(J(k_1, \ldots, k_{n-1})\) is

\[
\left| \frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-1} + p_{n-2}}{q_{n-1} + q_{n-2}} \right| = \frac{1}{q_{n-1}(q_{n-1} - q_{n-2})} = \frac{1}{q_{n-1}^2 \left(1 + \frac{q_{n-2}}{q_{n-1}}\right)}.
\]

Similarly, if \(k \in \mathbb{N}\) is fixed, the endpoints of \(J(k_1, \ldots, k_{n-1}, k)\) are \(\frac{p_{n-1}k + p_{n-2}}{q_{n-1}k + q_{n-2}}\) and \(\frac{p_{n-1}k + p_{n-2}}{q_{n-1}k + q_{n-2}}\), and its length is

\[
\left| \frac{p_{n-1}k + p_{n-2}}{q_{n-1}k + q_{n-2}} - \frac{p_{n-1}(k+1) + p_{n-2}}{q_{n-1}(k+1) + q_{n-2}} \right| = \frac{1}{q_{n-1}^2 k^2 \left(1 + \frac{q_{n-2}}{q_{n-1}k}\right)}\left(1 + \frac{q_{n-2}}{q_{n-1}k}\right)
\]

and so

\[
\lambda(J(k_1, \ldots, k_n, k)) = \frac{1}{k^2} \frac{1 + \frac{q_{n-2}}{q_{n-1}}}{1 + \frac{q_{n-2}}{q_{n-1}k}} \lambda(J(k_1, \ldots, k_n))
\]

and we have

\[
1 \leq \frac{1 + \frac{q_{n-2}}{q_{n-1}k}}{1 + \frac{q_{n-2}}{q_{n-1}k}} < 2
\]

\[
1 < \frac{1}{k} + \frac{q_{n-2}}{q_{n-1}k} < 3
\]

so we conclude that

\[
\frac{1}{3k^2} \lambda(J(k_1, \ldots, k_{n-1}, k)) < \lambda(J(k_1, \ldots, k_{n-1}, k)) < \frac{2}{k^2}
\]

\[
\frac{1}{3k^2} \lambda(J(k_1, \ldots, k_{n-1})) < \lambda(J(k_1, \ldots, k_{n-1})) < \frac{2}{k^2} \lambda(J(k_1, \ldots, k_{n-1}))
\]

334
We have that, as \((k_1, \ldots, k_{n-1})\) varies over \(\mathbb{N}^{n-1}\), \(E_n^{(k)}\) is the disjoint union of the \(J(k_1, \ldots, k_{n-1}, k)\) and \((0,1)\) is the disjoint union of the \(J(k_1, \ldots, k_{n-1})\); adding the inequalities over all \((k_1, \ldots, k_{n-1})\), we obtain the desired result.

It is interesting to note the following example. Let \(D\) be the set of real numbers \(x\) such that there exists a strictly increasing sequence \((r_n)\) in \(\mathbb{N}\) for which \(a_{r_n}(x)\) divides \(a_{r_{n+1}}(x)\) for all \(n\). Then \(D\) is a large subset of \(\mathbb{R}\) (it contains all \(x\) which have infinitely many 1's in its continued fraction expansion, and hence \(\mathbb{R} \setminus D\) has Lebesgue measure 0).

**III.8.3.42. Theorem.** \(D\) is not a Borel set in \(\mathbb{R}\).

In fact, \(D\) is an analytic subset of \(\mathbb{R}\). \(\mathbb{R} \setminus D\) is not analytic. These are perhaps the simplest such subsets of \(\mathbb{R}\) to explicitly describe.

**III.8.4. Exercises**

**III.8.4.1.** Prove III.8.2.4. by induction on \(n\). [Hint: apply the inductive hypothesis to the denominator of the fraction.]

**III.8.4.2.** Let \(\varphi\) be the “golden ratio” \(\frac{1 + \sqrt{5}}{2}\). Prove that the continued fraction expansion of \(\varphi\) is \([1; 1, 1, \ldots]\), and relate it to the Fibonacci sequence \((1, 1, 2, 3, 5, 8, \ldots)\), where each term is the sum of the two previous terms. This is the continued fraction expansion which converges most slowly.

The golden ratio is a number which has been of considerable mathematical and cultural significance. See [Liv02] for a cultural history of \(\varphi\).

**III.8.4.3.** (a) Let \(d \in \mathbb{N}\), not a perfect square. Find the continued fraction expansion of \(\sqrt{d}\). Relate the expansion to the integral solutions of Pell’s equation [the name is a misnomer, since Pell made little if any contribution to the theory of this equation; the basic work is due to Fermat]

\[x^2 - dy^2 = \pm 1.\]

(b) Let \(r_n\) be the \(n^{th}\) continued fraction approximation to \(\sqrt{2}\). If Newton’s approximation to the positive root of \(f(x) = x^2 - 2\) is used with initial approximation \(r_n\), i.e.

\[x_{m+1} = \frac{1}{2} \left( x_m + \frac{2}{x_m} \right)\]

with \(x_0 = r_n\), show that the next approximation \(x_1\) is \(r_{2n+1}\). Find the relation between continued fractions and Newton’s method for \(\sqrt{d}, d > 2\) [Duj01], [Mik54], [CMR95], [Ele97].

**III.8.4.4.** (a) Let \(\alpha\) be a quadratic number, i.e. an irrational root of a quadratic polynomial \(ax^2 + bx + c\) with \(a, b, c \in \mathbb{Z}\). Prove that the continued fraction expansion of \(\alpha\) is eventually periodic.

(b) Conversely, suppose \(\beta\) is an irrational number and the continued fraction expansion of \(\beta\) is eventually periodic. Prove that \(\beta\) is a quadratic number.

(Part (a) is an involved problem, and it is unreasonable for the reader to solve it without referring to number theory texts. Part (b) is fairly easy. The two parts were respectively theorems of Lagrange and Euler, two of the most outstanding mathematicians of the 18th century.)
III.8.4.5. Show that the continued fraction expansion of $e$ is
\[ [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, \ldots, 1, 2k, 1, \ldots] \,.
\]
[First find the continued fraction expansion of $\frac{e^4}{e^4 - 1}$.] This expansion was found by Euler.
In contrast, the first few terms of the continued fraction expansion of $\pi$ are
\[ [3; 7, 15, 1, 292, 1, 1, 2, 1, 3, 14, 2, 1, 1, 2, 2, 1, 2, 1, 4, 1, 1, 84, 2, 1, 15, 3, 13, 1, 4, 2, 6, 6, 99, 1, 2, 2, 6, 3, 5, 1, 1, 6, 8, 1, 7, 1, 2, 3, 7, 1, 2, 1, 12, 1, 1, 3, 1, 1, 8, 1, 1, 2, 1, 6, 1, 1, 5, 2, 2, 3, 1, 2, 4, 4, 16, 1, 161, 45, 1, 22, 1, 2, 2, 1, 4, 1, 2, \ldots] \]
and no pattern is known (although there are other regular nonsimple continued fraction representations of $\pi$). Millions of terms of this expansion can be found on the internet.

III.8.4.6. (a) Write out the details of the proof that every Liouville number is transcendental. (b) Prove that the number $\sum_{k=1}^{\infty} 10^{-k!}$ is a Liouville number. (c) Prove that the number with continued fraction expansion $[0; 10, 100, \ldots, 10^{n!}, \ldots]$ is a Liouville number.

III.8.4.7. Let $\varphi$ be the golden ratio, and $F_n$ the $n$th Fibonacci number (Exercise (i)). Since $\lim_{n \to \infty} \frac{F_{n+1}}{F_n} = \varphi$, the $F_n$ grow exponentially: if $b$ is any number with $0 < b < \varphi$, then $F_n > b^n$ for all sufficiently large $n$ (and if $c > \varphi$, $F_n < c^n$ for sufficiently large $n$). (a) If $x$ is any irrational number, show that $q_n(x) \geq F_n$ for all $n$. [Prove by induction, using III.8.2.15.(i),] Thus $q_n(x)$ grows at least exponentially. (The inequality $q_n(x) \geq F_n$ also applies for $x$ rational whenever $q_n(x)$ is defined.) (b) Show by induction on $k$ and III.8.2.15.(i) that if $x$ is irrational and $n \geq 2$, then, for all $k$,
\[ \frac{q_{n+k}(x)}{q_n(x)} > F_{k+1} \]
and that $F_{k+1}$ is the largest possible universal constant for the right-hand side. [Note that $q_n(x)$ can be arbitrarily large compared to $q_{n-1}(x)$ if $a_n(x)$ is large.]

III.8.4.8. (a) Prove Proposition III.8.2.21. by induction, using III.8.2.15.(ii) and the relation $a_n(x) = a_n(x) + \frac{1}{a_{n+1}(x)^2}$. (b) Prove the left-hand inequality in III.8.2.22., using $a_n \leq a_n < a_n + 1$.

III.8.4.9. (a) Using III.8.2.22., show that for an irrational number $x$, the following are equivalent:

(i) $x$ is a Liouville number.
(ii) For each $k \in \mathbb{N}$, $q_n(x) \geq q_n(x)^k$ for infinitely many $n$.
(iii) For each $k \in \mathbb{N}$, $a_{n+1}(x) \geq q_n(x)^k$ for infinitely many $n$.

(b) A Liouville number $x$ can have many small $a_n(x)$'s. In fact, we can have the lim inf of the geometric means of the $a_n(x)$ equal to 1 if many 1's are interspersed between the $a_{n+1}$'s of (iii). But (iii) implies that the lim sup of the geometric means of the $a_n(x)$ is infinite. Thus no Liouville numbers are in the set for which the conclusion of III.8.2.30. holds.
III.8.4.10. Champernowne’s constant is obtained by concatenating all positive integers into a single decimal:

\[ c = 0.1234567891011121314 \cdots \]

The continued fraction expansion is

\[ [0; 8, 9, 1, 1, 1] \]

where \( a_{18} \) is a 166 digit number. This is only the beginning: \( a_{40}(c) \) has 2504 digits, and \( a_{34062}(c) \) (the largest one calculated so far) has 7,311,111,092 digits! See http://mathworld.wolfram.com/ChampernowneConstant.html for more information.

Despite the apparent rapid growth of the \( a_n(c) \), \( c \) is not a Liouville number, although it is transcendental [?].

III.8.4.11. This problem gives an analog of (a) for the growth of the \( q_n \). Let \( \psi : \mathbb{N} \to [1, \infty) \) be a function.

(a) Show that if \( \sum_{q=1}^{\infty} \frac{1}{\psi(q)} \) converges, then for almost all \( x \) we have that the inequality

\[ \left| x - \frac{p}{q} \right| < \frac{1}{q \psi(q)} \]

has only finitely many solutions. [We may assume \( 0 < x < 1 \). Fix \( N \). For each \( p \) and \( q, q \geq N, 1 \leq p \leq q \), let \( I_{p,q} \) be an interval of length \( \frac{2}{q \psi(q)} \) centered at \( \frac{p}{q} \). Let \( S_N = \bigcup_{p,q} I_{p,q} \). Then \( \lambda(S_N) \leq \sum_{q=1}^{\infty} \frac{1}{\psi(q)} \). If the inequality has infinitely many solutions for \( x, x \) is in \( \bigcap_N S_N \).

(b) If \( \frac{\psi(q)}{q} \) increases and \( \sum_{q=1}^{\infty} \psi(q) \) diverges, then for almost all \( x \) the inequality

\[ \left| x - \frac{p}{q} \right| < \frac{1}{q \psi(q)} \]

has infinitely many solutions.

III.8.4.12. Consider \( a_n \) as a random variable on the interval \( (0, 1) \).

(a) Find the conditional probability \( P(a_2 = k | a_1 = j) \) [The set

\[ \{ x \in (0, 1) : a_1(x) = j, a_2(x) = k \} \]

is an interval.]

(b) Compute \( P_2^{(k)} = P(a_2 = k) \).

(c) Show that \( a_m \) and \( a_n \) are not independent for any \( m \) and \( n \).

(d) Using III.8.3.41., show that there are independent random variables \( (b_n) \) and positive constants \( \alpha \) and \( \beta \) such that

\[ \alpha < \frac{a_n(x)}{b_n(x)} < \beta \text{ a.e.} \]
III.8.4.13. ([?], [Wag93, p. 75-76]; cf. [?]) For each \( t > 0 \), define
\[
\nu_t = \sum_{n=1}^{\infty} 2^{\left\lfloor nt \right\rfloor} - 2^{n^2}.
\]
The \( \nu_t \) are called the von Neumann numbers.

(a) Show that the series converges for all \( t > 0 \).

(b) Show that each \( \nu_t \) is a Liouville number, and in particular transcendental.

(c) Show that \( \nu_t \) is not algebraic over the subfield \( F_{t} \) of \( \mathbb{R} \) generated by \( \{ \nu_s : s < t \} \), i.e. \( \nu_t \) is not a root of a nonzero polynomial with coefficients in \( F_{t} \). In other words, \( \{ \nu_t : t > 0 \} \) is an algebraically independent set of real numbers of cardinality \( 2^{80} \). [Show that \( \nu_{t} \) can be approximated much more closely by rationals than \( \nu_{s} \) for \( s < t \).] Note that the Axiom of Choice is not needed to construct this set of algebraically independent numbers or to prove its properties.

III.8.4.14. Let \( f \) be a strictly increasing function from \( \mathbb{N} \) to \( \mathbb{N} \), and let \( \mu = \limsup_{n \to \infty} \frac{f(n+1)}{f(n)} \). Let \( d \in \mathbb{N} \), \( d > 1 \), and set \( x = \sum_{n=1}^{\infty} \frac{1}{d^{f(n)}} \).

(a) Show that the equation
\[
|x - r| < \frac{C}{q^{\mu - \epsilon}}
\]
has infinitely many solutions in \( r = \frac{p}{q} \in \mathbb{Q} \) for any fixed \( C > 0 \) and \( \epsilon > 0 \).

(b) Use Roth’s Theorem (III.8.2.39.) to show that \( x \) is transcendental if \( \mu > 2 \).

III.8.4.15. (a) If \( x = \frac{p}{q} \) is a rational number, show that the decimal expansion of \( x \) is eventually periodic, with the length of the period not more than \( q - 1 \). [Divide \( q \) into \( p \) by long division.]

(b) Conversely, if \( x \in \mathbb{R} \) has an eventually periodic decimal expansion, then \( x \) is rational. [If the length of the period of the expansion is \( k \), show that \( 10^{k}x - x \) is a decadent rational.]

If \( 0 \leq x \leq 1 \) and the decimal expansion of \( x \) is exactly periodic, a fractional representation of \( x \) is easy to obtain. For example,
\[
x = .250425042504 \cdots = \frac{2504}{9999},
\]
since \( 10000x - x = 2504 \), and the same pattern works in general, with \( k \) 9’s in the denominator if the period is \( k \). (This representation will not be in lowest terms in general.)

(c) What is the exact relationship between the length of the period of the decimal expansion of \( \frac{p}{q} \) and the denominator \( q \)?

(d) Show that analogous results hold for “decimal” expansions with any base.

III.8.4.16. Show that if \( a \) and \( b \) are positive numbers (e.g. natural numbers), then
\[
\sqrt{a^2 + b} = a + \frac{b}{2a + \frac{b}{2a + \frac{b}{2a + \frac{b}{x + \frac{b}{x + \frac{b}{x + \ddots}}}}}}.
\]
III.9. Irrational and Transcendental Numbers

“It can be of no practical use to know that is irrational, but if we can know, it surely would be intolerable not to know.”

E. C. Titchmarsh

III.9.1. Introduction

III.9.1.1. Analytic methods can be used to prove that specific numbers such as \( e \) and \( \pi \) are irrational and even transcendental. In this section, we give a few proofs not requiring anything beyond basic analysis techniques. More sophisticated results can be obtained by combining such methods with ideas from algebraic number theory; we survey some of these results, mostly without proof.

III.9.1.2. It was known as a result of Euler’s work on continued fractions (III.8.4.5.) that \( e \) (and also \( e^2 \)) is irrational, and in fact, using Lagrange’s Theorem (III.8.4.4.), not algebraic of degree 2. Another proof based on the power series representation was given by Fourier in 1815 (cf. V.5.5.6.; this proof can be easily modified to also give the quadratic result.)

III.9.1.3. The first proof that \( \pi \) is irrational was given by Lambert in 1766, using continued fractions (his proof was not rigorous by modern standards). Legendre gave a rigorous proof along the same lines in 1794 for \( \pi \) and many other numbers. Hermite developed another method in 1873 giving the transcendence of \( e \), which was adapted by Lindemann to give broad transcendence results including transcendence of \( \pi \). Significant advances beginning with the same general method were obtained throughout the twentieth century. See, for example, [Niv56] or the introduction to [Shi89] for a more complete survey.

III.9.2. Irrationality of \( \pi \)

III.9.2.1. We first give a simple proof of the irrationality of \( \pi \) (and \( \pi^2 \)) along the lines of the arguments of Hermite and Lindemann, due to I. Niven [Niv47] and Y. Iwamoto [Iwa49] (cf. [Wac49], [Niv56, 2.6]), which will introduce the techniques used in later proofs of transcendence, even though the result will be subsumed in III.9.3.2. The idea in all these proofs is to assume that a certain rational or polynomial relation exists, and obtain a contradiction by constructing an integral whose value is an integer strictly between 0 and 1. The functions used in the arguments can seem to materialize out of thin air, but actually arise from considerations related to continued fraction expansions; see [Sie49, I, §3], for example, for motivation.

We first make the following simple observation:

III.9.2.2. Lemma. Let \( g(x) \) be a polynomial with integer coefficients, and \( n \in \mathbb{N} \). Set \( f(x) = \frac{x^n g(x)}{n!} \). Then the \( k \)’th derivative \( f^{(k)}(0) \) is an integer for every \( k \).

Proof: Let \( x^n g(x) = \sum_{k=0}^{n+d} c_k x^k \). Then \( c_k \in \mathbb{Z} \) for all \( k \), and \( c_k = 0 \) for \( k < n \). We have, for all \( k \),

\[
  f^{(k)}(0) = \frac{k!c_k}{n!} \in \mathbb{Z}.
\]

III.9.2.3. Theorem. The number $\pi^2$ is irrational. (Hence $\pi$ is also irrational.)

Proof: Let $n$ be a fixed positive integer (which we will choose later), and set

$$f(x) = \frac{x^n(1-x)^n}{n!}.$$ 

Note that if $0 < x < 1$, then $0 < f(x) < \frac{1}{n!}$. By III.9.2.2., $f^{(k)}(0) \in \mathbb{Z}$ for all $k$; and $f^{(k)}(1) \equiv 0$ for $k > 2n$. Since $f(1-x) = f(x)$ for all $x$, we also have $f^{(k)}(1) \in \mathbb{Z}$ for all $k$.

Suppose $\pi^2 = \frac{a}{b}$ for relatively prime $a, b \in \mathbb{N}$. Set

$$F(x) = b^n[\pi^{2n}f(x) - \pi^{2n-2}f^{(2)}(x) + \pi^{2n-4}f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)]$$

and note that $F(0), F(1) \in \mathbb{Z}$. Next,

$$\frac{d}{dx}(F'(x) \sin \pi x - \pi F(x) \cos \pi x) = (F''(x) + \pi^2 F(x)) \sin \pi x$$

$$= b^n \pi^{2n+2} f(x) \sin \pi x = \pi^2 a^n f(x) \sin \pi x$$

so we have

$$\pi a^n \int_0^1 f(x) \sin \pi x \, dx = \left[ \frac{F'(x) \sin \pi x}{\pi} - F(x) \cos \pi x \right]_0^1 = F(1) + F(0) \in \mathbb{Z}.$$ 

On the other hand, since $0 < f(x) \sin \pi x < \frac{1}{n!}$ for $0 < x < 1$,

$$0 < \pi a^n \int_0^1 f(x) \sin \pi x \, dx < \frac{\pi a^n}{n!} < 1$$

for sufficiently large $n$, a contradiction.

A nearly identical proof ([Kok49], [Wac49]) shows that $e^n$ is irrational for all $n$ (see III.9.3.1. for a better result). A modified version of this argument, somewhat more complicated but still elementary (cf. [Niv56, 2.5]), gives:

III.9.2.4. Theorem. If $\alpha$ is any nonzero rational number, then $\cos \alpha$ is irrational.

III.9.2.5. This theorem gives another proof that $\pi$ is irrational, since otherwise $\cos \pi = -1$ would be irrational. It also implies that $\sin^2 \alpha = \frac{1}{2}(1 - \cos 2\alpha)$ and hence $\sin \alpha$ and $\cos^2 \alpha = 1 - \sin^2 \alpha$ are irrational, for nonzero rational $\alpha$. In addition, for $\alpha$ nonzero rational, $\tan^2 \alpha$ is irrational, since otherwise $\cos^2 \alpha = \frac{1}{1 + \tan^2 \alpha}$ would be rational, and hence $\tan \alpha$ is irrational. Thus, if $\alpha$ is nonzero rational, $\arcsin \alpha$ and $\arctan \alpha$ are irrational. These results will be subsumed by the much more general result III.9.3.6.(iii).

340
III.9.3. Transcendence of $e$ and Related Numbers

A proof along the same general lines can be given of the transcendence of $e$, due to A. Hurwitz (1893):

**III.9.3.1. Theorem.** The number $e$ is transcendental.

**Proof:** Suppose $e$ is algebraic. Then there are integers $a_0, a_1, \ldots, a_m$ with $a_0 \neq 0$ and

$$a_0 + a_1 e + a_2 e^2 + \cdots + a_m e^m = 0.$$  

If $n \in \mathbb{N}$ and $f$ is any polynomial of degree $n$ with real coefficients, and we set

$$F(x) = f(x) + f'(x) + f''(x) + \cdots + f^{(n)}(x),$$  

then by repeated integration by parts we obtain, for any $b > 0$,

$$\int_0^b f(t)e^{-t} \, dt = f(0) - f(b)e^{-b} + \int_0^b f'(t)e^{-t} \, dt = \cdots = F(0) - F(b)e^{-b}$$  

called Hermite’s Identity.

We will apply Hermite’s Identity in the case where

$$f(x) = \frac{1}{(p-1)!}x^{p-1}(x-1)^p(x-2)^p\cdots(x-m)^p$$

where $m$ is the degree of the polynomial relation for $e$ given above and $p$ is a large prime number which will be chosen later (recall that there are infinitely many prime numbers, so $p$ can be chosen as large as desired).

Since $f$ is a polynomial of degree $n = p(m+1)-1$, we may apply Hermite’s Identity for $b = 1, \ldots, k, \ldots, m$, we obtain

$$F(0)e^k - F(k) = e^k \int_0^k f(t)e^{-t} \, dt.$$  

The equation also holds when $k = 0$ (both sides are 0). Multiplying by $a_k$ and adding, we get the fundamental relation

$$\sum_{k=0}^m a_k e^k F(0) - \sum_{k=0}^m a_k F(k) = \sum_{k=0}^m a_k e^k \int_0^k f(t)e^{-t} \, dt. \quad (III.1)$$  

The first sum is 0 by the assumed polynomial relation for $e$, so the left side becomes

$$-\sum_{k=0}^m a_k F(k) = -\sum_{k=0}^m \sum_{r=0}^n a_k f^{(r)}(k).$$

If $g$ is a polynomial with integer coefficients, then for any $r$ we have that all coefficients of $g^{(r)}$ are divisible by $r!$ (V.17.7.6.). So, if $r \geq p$, all the coefficients of $f^{(r)}$ are integers divisible by $p$, and hence for $r \geq p$ we have that $f^{(r)}(k)$ is an integer divisible by $p$ for any $k \in \mathbb{Z}$ (and, in particular, for $k = 0, 1, \ldots, m$). Also, for $k = 1, \ldots, m$, $f$ has a zero of order $p$ at $k$, so $f^{(r)}(k) = 0$ for $0 \leq r < p$. Thus $f^{(r)}(k)$ is an integer divisible by $p$ for all $r$. 

341
Since $f$ has a zero of order $p - 1$ at 0, we also have $f^{(r)}(0) = 0$ for $0 \leq r < p - 1$. We can calculate directly from the definition that
\[
f^{(p-1)}(0) = (-1)^p(-2)^p \cdots (-m)^p = (-1)^mp(m!)^p
\]
which is an integer not divisible by $p$ if $p > m$.

So every term on the left is an integer divisible by $p$ except the term $a_0 f^{(p-1)}(0)$, which is an integer not divisible by $p$ if $p > m$ and $p > |a_0|$ (recall that $a_0 \neq 0$). So if $p$ is sufficiently large, the left side of III.1 is a nonzero integer.

Now consider the right side of III.1. For $0 \leq x \leq m$, we have
\[
|f(x)| \leq \frac{m^p-1m^p \cdots m^p}{(p-1)!} = \frac{m^{mp+p-1}}{m(p-1)!} = \frac{(m^{m+1})^p}{m(p-1)!}.
\]
So, if we set
\[
c = e^m \sum_{k=0}^m |a_k| \quad \text{and} \quad d = m^{m+1}
\]
(note that $c$ and $d$ do not depend on $p$), we get for the right side
\[
\left| \sum_{k=0}^m a_k e^k \int_0^k f(t)e^{-t} dt \right| \leq \sum_{k=0}^m |a_k| e^k m \frac{(m^{m+1})^p}{m(p-1)!} \leq c \frac{d^p}{(p-1)!},
\]
which is strictly less than 1 for $p$ sufficiently large, contradicting that the left side is a nonzero integer. ☺

III.9.3.2. A proof that $\pi$ is transcendental can be given along the same lines, but it is slightly more complicated and uses some elementary results from number theory; see, for example, [Bak90, 1.3] or [Shi89, p. 47-50].

Using a refined version of the same argument, plus some facts about algebraic numbers, LINDEMAN in his 1882 paper outlined a proof of a much more general theorem, for which WEIERSTRASS later provided a complete and simplified proof (HERMITE gave a proof of III.9.3.5.):

III.9.3.3. Theorem. Let $\alpha_1, \ldots, \alpha_n$ be distinct algebraic numbers, and $\beta_1, \ldots, \beta_n$ nonzero algebraic numbers. Then
\[
\beta_1 e^{\alpha_1} + \cdots + \beta_n e^{\alpha_n} \neq 0.
\]

This result has the following equivalent formulation, which is often called Lindemann’s Theorem:
III.9.3.4. Theorem. Let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers which are linearly independent over $\mathbb{Q}$. Then $e^{\alpha_1}, \ldots, e^{\alpha_n}$ are algebraically independent over $\mathbb{Q}$, i.e. there does not exist a nonzero polynomial $f(x_1, \ldots, x_n)$ with rational coefficients (or even with algebraic coefficients) with $f(e^{\alpha_1}, \ldots, e^{\alpha_n}) = 0$.

It is a simple standard fact from field theory (cf. Exercise III.9.5.2.) that if $\gamma_1, \ldots, \gamma_n$ are complex numbers which are algebraically independent over $\mathbb{Q}$, then they are algebraically independent over the field of algebraic numbers (the algebraic closure of $\mathbb{Q}$).

To obtain III.9.3.4. from III.9.3.3., let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers which are linearly independent over $\mathbb{Q}$. If $f(x_1, \ldots, x_n)$ is a nonzero polynomial with rational coefficients, then $f$ is a sum of terms of the form $c_{(k_1, \ldots, k_n)} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$, where the $k_i$ are nonnegative integers. So $f(e^{\alpha_1}, \ldots, e^{\alpha_n})$ is a sum of terms of the form $c_{(k_1, \ldots, k_n)} e^{k_1 \alpha_1} e^{k_2 \alpha_2} \cdots e^{k_n \alpha_n} = c_{(k_1, \ldots, k_n)} e^{k_1 \alpha_1 + \cdots + k_n \alpha_n}$ and the exponents are distinct integer linear combinations of $\alpha_1, \ldots, \alpha_n$, hence distinct algebraic numbers by linear independence. Thus by III.9.3.3., $f(e^{\alpha_1}, \ldots, e^{\alpha_n}) \neq 0$.

Conversely, suppose III.9.3.4. holds, and let $\alpha_1, \ldots, \alpha_n$ be distinct algebraic numbers. If the rational span of $\{\alpha_1, \ldots, \alpha_n\}$ has dimension $m$ over $\mathbb{Q}$, then there are algebraic numbers $\gamma_1, \ldots, \gamma_m$ which are linearly independent over $\mathbb{Q}$ such that, for each $k$, $1 \leq k \leq n$,

$$\alpha_k = \sum_{j=1}^{m} a_{jk} \gamma_j$$

with $a_{jk} \in \mathbb{Q}$; we may assume each $a_{jk} \in \mathbb{Z}$ by replacing $\gamma_j$ by $\gamma_j/r_j$ for suitable $r_j \in \mathbb{Z}$. Now let $\beta_1, \ldots, \beta_n$ be nonzero algebraic numbers, and set

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{n} \beta_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_m^{a_{mk}}.$$

Then $f$ is a nonzero rational function with algebraic coefficients (the monomials in the terms are all different since the $\alpha_k$ are distinct), hence a quotient of polynomials with algebraic coefficients (the denominator can be taken to be a monomial), so $f(e^{e\gamma_1}, \ldots, e^{e\gamma_m}) \neq 0$ by III.9.3.4. But

$$f(e^{e\gamma_1}, \ldots, e^{e\gamma_m}) = \sum_{k=1}^{n} \beta_k e^{a_{1k} \gamma_1 + \cdots + a_{mk} \gamma_m} = \sum_{k=1}^{n} \beta_k e^{a_k} \neq 0.$$

The special case of III.9.3.4. with $n = 1$ is:

III.9.3.5. Corollary. If $\alpha$ is a nonzero algebraic number, then $e^\alpha$ is transcendental.

In these theorems, the $\alpha_k$ and $\beta_k$ do not have to be real numbers (i.e. they can be complex algebraic numbers).
III.9.3.6. Here are some consequences of III.9.3.5:

(i) This result also implies that $\pi$ is transcendental, since if $\pi$, and hence $\pi i$, were algebraic, then $e^{\pi i} = -1$ would be transcendental.

(ii) If $\alpha$ is a positive algebraic number, $\alpha \neq 1$, then $\log \alpha$ is transcendental, since if it were algebraic we would have $\alpha = e^{\log \alpha}$ transcendental.

(iii) If $\alpha$ is a nonzero algebraic number, then $e^{i\alpha} = \cos \alpha + i \sin \alpha$ is transcendental. If $\cos \alpha$ were algebraic, $\sin \alpha = \pm \sqrt{1 - \cos^2 \alpha}$ would also be algebraic, so $e^{i\alpha}$ would be algebraic. Thus $\cos \alpha$ and $\sin \alpha$ are transcendental. We then also have $\tan \alpha$, $\arcsin \alpha$, $\arctan \alpha$ transcendental by the argument of III.9.2.5.

III.9.4. Twentieth Century Results

III.9.4.1. In his famous 1900 list of problems (#7), HILBERT asked to prove that if $\alpha$ is an algebraic number other than 0 or 1, and $\beta$ is an irrational algebraic number, then $\alpha^\beta$ is transcendental. In a 1920 lecture he expressed the opinion that both the Riemann Hypothesis and Fermat’s Last Theorem would be solved first, and that this problem would not be solved within the lifetime of anyone in the room. Actually, the problem was solved independently in 1934 by GELFOND and SCHNEIDER, well within HILBERT’s own lifetime, using key ideas of GELFOND (1929) and SIEGEL (1929). (Fermat’s Last Theorem was proved in 1994, and the Riemann Hypothesis remains unproved, showing the difficulty in trying to predict the future of mathematics.)

III.9.4.2. Theorem. [Gelfond-Schneider] Let $\alpha$ and $\beta$ be algebraic complex numbers, with $\alpha$ not 0 or 1 and $\beta \notin \mathbb{Q}$. Then $\alpha^\beta$ is transcendental.

Actually, in general $\alpha^\beta$ has many possible values depending on the branch of logarithm chosen; the theorem asserts that all these possible values are transcendental. See [Niv56, Chapter 10] for a proof.

III.9.4.3. For example, numbers such as $2\sqrt{2}$ and $\sqrt[3]{2}$ are transcendental. Also, we obtain that $e^\pi$ is transcendental, since it is one of the values of $i^{-2i}$. Similarly, $e^{\pi \alpha}$ is transcendental for any nonzero real algebraic $\alpha$.

An especially interesting example is the “Ramanujan Constant” (actually discovered by HERMITE in 1859; the name stems from an April Fool joke by MARTIN GARDNER in 1975. See http://en.wikipedia.org/wiki/Ramanujan_constant#Almost_integers_and_Ramanujan.27s_constant or http://mathworld.wolfram.com/RamanujanConstant.html for more information.) Calculation shows that

$$e^{\pi \sqrt{163}} = 262537412640768744.00000000000$$

correct to eleven decimal places; nonetheless, this number is not an integer, and in fact is transcendental. Powers of this number are also remarkably close to integers. The underlying reason is that 163, which is prime, is the largest squarefree integer $d$ such that the ring of integers in the field $\mathbb{Q}(\sqrt{-d})$ has unique factorization.

We also have the following corollary, which was stated but not proved by EULER:
III.9.4.4. Corollary. Let $\alpha$ and $\beta$ be positive algebraic numbers, with $\alpha \neq 1$. If $\beta$ is not a rational power of $\alpha$ (i.e. if $\log_{\alpha} \beta$ is irrational; cf. III.9.5.3.), then $\log_{\alpha} \beta$ is transcendental.

Proof: If $\gamma = \log_{\alpha} \beta$ is irrational but algebraic, then $\beta = \alpha^\gamma$ is transcendental, a contradiction. 

A significant advance arose from the fundamental work of A. Baker in the late 1960’s. One of his theorems was:

III.9.4.5. Theorem. If $\alpha_1, \ldots, \alpha_n$ are nonzero algebraic numbers such that $\log \alpha_1, \ldots, \log \alpha_n$ (any possible values of these are allowed) are linearly independent over $\mathbb{Q}$, then $1, \log \alpha_1, \ldots, \log \alpha_n$ are linearly independent over the algebraic numbers.

III.9.4.6. The case $n = 2$ implies III.9.4.2: if $\alpha$ and $\beta$ are algebraic numbers with $\alpha$ not 0 or 1 (so $\log \alpha$ is defined and nonzero), and $\alpha^\beta$ is algebraic, set $\alpha_1 = \alpha$, $\beta_1 = \beta$, $\alpha_2 = \alpha^\beta$, and $\beta_2 = -1$. Then $\beta_1 \log \alpha_1 + \beta_2 \log \alpha_2 = 0$, so by III.9.4.5. $\{\log \alpha_1, \log \alpha_2\} = \{\log \alpha, \beta \log \alpha\}$ is linearly dependent over $\mathbb{Q}$, i.e. $\beta \in \mathbb{Q}$. (The converse implication is not quite true: III.9.4.2. implies that if $\alpha_1$ and $\alpha_2$ are algebraic numbers with $\{\log \alpha_1, \log \alpha_2\}$ linearly independent over $\mathbb{Q}$, then $\{\log \alpha_1, \log \alpha_2\}$ is linearly independent over the algebraic numbers, but it does not imply that $\{1, \log \alpha_1, \log \alpha_2\}$ is linearly independent over the algebraic numbers.)

III.9.4.7. Corollary. 

(i) If $\alpha_1, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$ are nonzero algebraic numbers, then

$$\beta_0 + \beta_1 \log \alpha_1 + \cdots + \beta_n \log \alpha_n \neq 0.$$ 

(ii) If $\alpha_1, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$ are nonzero algebraic numbers, then

$$e^{\beta_0 \alpha_1^\beta_1 \cdots \alpha_n^\beta_n}$$

is transcendental.

(iii) If $\alpha_1, \ldots, \alpha_n$ are algebraic numbers other than 0 or 1, and $\beta_1, \ldots, \beta_n$ are algebraic numbers with $\{1, \beta_1, \ldots, \beta_n\}$ linearly independent over $\mathbb{Q}$, then

$$\beta_1 \cdots \beta_n$$

is transcendental.

III.9.4.8. These results imply that $\pi + \log \alpha$ is transcendental for any nonzero algebraic number $\alpha$, and that $e^{\pi \log \alpha} = \pi + \log \alpha$ is transcendental for any algebraic $\alpha$ and $\beta$ with $\beta \neq 0$. 

345
III.9.4.9. **Siegell** (1929) began a general study of the values of certain transcendental functions satisfying linear differential equations, and he and subsequent authors obtained far-reaching generalizations of the results of III.9.3.6. See [Shi89] for the full story.

III.9.4.10. We have only scratched the surface of the extensive research done in this subject. Books such as () give a much more complete picture. But despite all the advances, there are still basic questions remaining open. For example, it has been difficult to get algebraic independence results for numbers of the form $\alpha^\beta$, where $\alpha$ and $\beta$ are algebraic, analogous to III.9.3.4. III.9.4.7.(ii)–(iii) are weak analogs, but a better result would be

III.9.4.11. **Conjecture.** Let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers not 0 or 1, let $\beta_1, \ldots, \beta_n$ be algebraic numbers with $\{1, \beta_1, \ldots, \beta_n\}$ linearly independent over $\mathbb{Q}$, and let $\gamma_1, \ldots, \gamma_m$ be algebraic numbers linearly independent over $\mathbb{Q}$. Then $\{\alpha_1^{\beta_1}, \ldots, \alpha_n^{\beta_n}, e^{\gamma_1}, \ldots, e^{\gamma_m}\}$ is algebraically independent over $\mathbb{Q}$.

The $\gamma_j$ and $e^{\gamma_j}$ can be eliminated from this conjecture. See [Rib00a, p. 323-327] for more related conjectures.

III.9.4.12. As an example of the incomplete state of the theory of transcendental numbers, it is not known whether $\pi$ and $e$ are algebraically independent (this is apparently universally believed to be true, but remains unproved). It is not even known whether $\pi + e$ and/or $\pi e$ is irrational! It is known from the following simple proposition that at least one of these numbers is transcendental, but nothing is known about either number individually. It is also not known whether $e^e$, $\pi^\pi$, $\pi^\pi$, or the Euler-Mascheroni constant $\gamma$ are rational or irrational, algebraic or transcendental.

III.9.4.13. **Proposition.** Let $\alpha$ and $\beta$ be numbers.

(i) If $\alpha$ is transcendental, then either $\alpha + \beta$ or $\alpha \beta$ (or both) is transcendental.

(ii) If $\alpha$ is not a quadratic number, then either $\alpha + \beta$ or $\alpha \beta$ (or both) is irrational.

**Proof:** (i) If $\alpha + \beta$ and $\alpha \beta$ are both algebraic, then $\alpha$ is a root of the quadratic polynomial

$$f(x) = (x - \alpha)(x - \beta) = x^2 - (\alpha + \beta)x + \alpha \beta$$

with algebraic coefficients, so $\alpha$ would be algebraic. The proof of (ii) is a slight variation left to the reader. \(\Box\)

III.9.5. **Exercises**

III.9.5.1. (http://math.stackexchange.com/questions/654232/if-cosa-pi-frac13-then-a-is-irrational?noredirect=1&lq=1)

(a) Let $p, q \in \mathbb{N}$. Then $e^{\pm \frac{\pi p}{q}}$ is a root of $X^{2q} - 1$, hence an algebraic integer. Thus $2\cos\left(\frac{\pi p}{q}\right) = e^{\frac{\pi p}{q}} + e^{-\frac{\pi p}{q}}$ is an algebraic integer.
(b) If \( r \in \mathbb{Q} \) and \( \cos \pi r \in \mathbb{Q} \), then \( 2 \cos \pi r \) is an integer (cf. III.9.5.5); hence \( |\cos \pi r| = 0, \frac{1}{2}, \) or 1. Thus if \( 0 < r < \frac{1}{2} \), then \( r = \frac{1}{3} \).

(c) If \( 0 < \alpha < \frac{\pi}{2} \) and \( \cos \alpha \) is rational, then either \( \alpha = \frac{\pi}{3} \) or \( \frac{\pi}{6} \) is irrational.

In light of the fact that \( \pi \) is irrational, even transcendental, this result is almost completely complementary to III.9.2.4. or III.9.3.6 (iii). The proof outlined is quite simple, but uses one nontrivial fact from abstract algebra or number theory: that the sum of two algebraic integers is an algebraic integer. A more elementary proof is given in [HD64] using only III.11.7.33. for \( n = 2 \) and some trigonometric identities; it is also noted there that the result implies that the acute angles in any Pythagorean triangle are irrational multiples of \( \pi \).

**III.9.5.2.** This exercise requires knowledge of the basics of field theory in abstract algebra. See books on abstract algebra such as [DF04] for details, and for an explanation of the standard notation and terminology used in this problem.

Suppose \( K \) is a field, \( F \) a subfield of \( K \), and \( E \) a subfield of \( K \) containing \( F \).

(a) If \( E \) is a finite extension of \( F \) and \( \alpha \in K \), then \( E(\alpha) \) is a finite extension of \( F(\alpha) \) and

\[ |E(\alpha) : F(\alpha)| \leq |E : F| . \]

(b) If \( E \) is an algebraic extension of \( F \) and \( \alpha \in K \), then \( E(\alpha) \) is an algebraic extension of \( F(\alpha) \).

(c) If \( E \) is an algebraic extension of \( F \) and \( \alpha \in K \) is algebraic over \( E \), then \( \alpha \) is algebraic over \( F \).

(d) If \( \gamma_1, \ldots, \gamma_n \in K \), then \( \{\gamma_1, \ldots, \gamma_n\} \) is algebraically independent over \( F \) if and only if \( \gamma_1 \) is transcendental over \( F \) and \( \gamma_{k+1} \) is transcendental over \( F(\gamma_1, \ldots, \gamma_k) \) for \( 1 \leq k < n \).

(e) Show inductively using (b), (c), and (d) that if \( \{\gamma_1, \ldots, \gamma_n\} \subseteq K \) is algebraically independent over \( F \) and \( E \) is algebraic over \( F \), then \( \{\gamma_1, \ldots, \gamma_n\} \) is algebraically independent over \( E \).

**III.9.5.3.**

(a) Prove that \( \log_2 3 \) is irrational. [\( 2^p \) is even and \( 3^q \) is odd for all \( p, q \in \mathbb{N} \).]

(b) Prove that if \( m, n \) are integers \( > 1 \), and \( \log_m n \) is rational, then \( m \) and \( n \) are both integer powers of the same natural number.

(c) Give a similar criterion for rationality of \( \log \alpha \beta \) for positive rational numbers \( \alpha, \beta \) with \( \alpha \neq 1 \).

**III.9.5.4.** This problem shows that if \( n, d \in \mathbb{N} \), then \( \sqrt[d]{n} \) is either a natural number or irrational.

(a) Suppose \( \sqrt[d]{a} = \frac{b}{d} \) for natural numbers \( a, b \). Then \( a^n = b^n d \). Write

\[ a = p_1^{r_1} p_2^{r_2} \cdots p_m^{r_m} \]
\[ b = p_1^{s_1} p_2^{s_2} \cdots p_m^{s_m} \]
\[ d = p_1^{t_1} p_2^{t_2} \cdots p_m^{t_m} \]

where \( p_1, \ldots, p_m \) are distinct prime numbers and the exponents are nonnegative integers (some may be zero).

(b) From \( a^n = b^n d \) obtain that

\[ p_1^{nr_1} p_2^{nr_2} \cdots p_m^{nr_m} = p_1^{ns_1+t_1} p_2^{ns_2+t_2} \cdots p_m^{ns_m+t_m} \]

and conclude that \( t_k \) is divisible by \( n \) for all \( k \), i.e. that \( d = c^n \) for some \( c \in \mathbb{N} \).

This argument uses the unique factorization of natural numbers into products of prime numbers (cf. ()), hence for \( n = 2 \) is not quite as elementary as the one in III.1.12.6. See III.1.13.6.(e) for an argument not requiring unique factorization.

347
III.9.5.5. The purpose of this problem is to give a purely algebraic proof of the following theorem:

**Theorem.** Let \( f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \) be a monic polynomial with integer coefficients. Then any rational number which is a root of \( f \) is an integer (i.e. a rational number which is an algebraic integer is an ordinary integer).

This theorem is a special case of *Gauss’s Lemma*. See books on abstract algebra such as [DF04] for the statement and proof of the general form of Gauss’s Lemma. (This is only one of several results commonly called “Gauss’s Lemma”; he produced a lot of important mathematics.)

If \( n, d \in \mathbb{N} \), applying this theorem to \( f(x) = x^n - d \), we obtain (see Exercise III.9.5.4. for another proof):

**Corollary.** If \( n, d \in \mathbb{N} \), then \( n/d \) is either a natural number or irrational.

These results are special cases of the more general result (also a special case of Gauss’s Lemma):

**Theorem.** If \( f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \) is a polynomial with integer coefficients and \( a_0 \neq 0 \), \( a_n \neq 0 \), and \( \alpha \) is a (nonzero) fraction in lowest terms which is a root, then \( b \) divides \( a_n \), \( c \) divides \( a_0 \), and \( f(x) = (bx - c)k(x) \) for a polynomial \( k \) with coefficients in \( \mathbb{Z} \).

(a) Show that if \( g(x) \) and \( h(x) \) are polynomials with integer coefficients, and \( p \) is a prime number that divides every coefficient of \( g(x)h(x) \), then either \( p \) divides every coefficient of \( g(x) \) or it divides every coefficient of \( h(x) \). [If not, let \( r \) be the largest number for which the coefficient of \( x^r \) in \( g \) is not divisible by \( p \), \( s \) the largest number for which the coefficient of \( x^s \) in \( h \) is not divisible by \( p \), and consider the coefficient of \( x^{r+s} \) in \( gh \).]

(b) If \( q = \frac{\alpha}{\beta} \in \mathbb{Q} \), show that \( f(x) = (x - q)g(x) \) for a polynomial \( g \) with rational coefficients.

(c) Show there is a \( d \in \mathbb{N} \) such that
\[
bd \cdot f(x) = (bx - c)h(x)
\]
where \( h \) is a polynomial with integer coefficients.

(d) If \( p \) is a prime number dividing \( bd \), use (a) to show that \( p \) divides every coefficient of \( h \).

(e) Dividing out prime factors of \( bd \) repeatedly, show that
\[
f(x) = (bx - c)k(x)
\]
for a polynomial \( k \) with integer coefficients. Multiply out to obtain the divisibility conclusion.

(f) In fact, \( \mathbb{Z}[X] \) is a unique factorization domain or *UFD*: any polynomial with integer coefficients can be written as a product of irreducible polynomials with integer coefficients, uniquely up to the order and signs of the factors.

III.9.5.6. Show that there is a unique real number \( a \) such that \( e^a = \frac{1}{a} \), and that this number is transcendental. (We have \( a = W(1) \approx 0.567 \), where \( W \) is the Lambert \( W \) function.)

III.9.5.7. Are the following formulas correct?
\[
\pi^4 + \pi^8 = e^6.
\]
\[
e^\pi - \pi = 20.
\]
III.9.5.8. [Sin13] Use a calculator or computer algebra program to calculate

\[ 12\sqrt{3987^{12} + 4365^{12}}. \]

Does this provide a counterexample to Fermat’s Last Theorem?
III.10. Hyperreal Numbers and Nonstandard Analysis

In the early years, calculus, especially in the Leibniz formulation, was based on the notion of an “infinitesimal” change; indeed, the name “calculus” is a shortened form of “infinitesimal calculus.” The logical foundations of calculus were vigorously, and sometimes rightly, challenged, and appeared to be quite shaky even to some of the developers and defenders of the theory, although no one could deny the apparent accuracy and usefulness of the theory in describing and explaining analytic and physical phenomena. (Leibniz himself regarded infinitesimals as a “fiction, although a well-founded one,” and considered them to be convenient symbolic shorthand for limit arguments, although his pronouncements on the subject were somewhat inconsistent. Newton felt much the same way. See [Jah03, p. 97-98].) When it became more important to mathematicians in the nineteenth century to carefully develop a rigorous theory of calculus and analysis, the intuitive approach via infinitesimals was deemed incapable of direct logical justification (Cantor was particularly firmly convinced that infinitesimals were a logical fiction, calling them the “cholera bacillus of mathematics”⁷), and was bypassed by the familiar $\varepsilon - \delta$ formulations of Cauchy and Weierstrass, which remain the standard approach to this day. See [Gra08, p. 328-346] for a lucid and fascinating discussion of the attitudes about measurement, infinitesimals, and the Archimedean axiom around the turn of the 20th century.

Intuitive arguments involving infinitesimals are far from dead, however, and are quite useful in many instances, particularly to scientists using calculus in their work. The dream among some mathematicians of a rigorous theory of infinitesimals never faded away.

Finally, in 1960, Abraham Robinson (partially anticipated by C. Schmieden and D. Laugwitz [SL58]) found a way to use fairly sophisticated ideas and results from mathematical logic to give a rigorous direct approach to analysis on $\mathbb{R}$ using infinitesimals. The subject growing out of his work is called nonstandard analysis (Robinson’s name).

Nonstandard analysis is one of the most controversial parts of mathematics. There are some “true believers” who think the nonstandard approach is the way to do analysis, and that eventually the mathematical world will see the light; on the other hand, nonstandard methods are definitely unfashionable among most mainstream analysts, although some aspects of probability and mathematical economics seem to have more natural nonstandard formulations (). It should be noted that the controversy has little if anything to do with the correctness or legitimacy of nonstandard analysis as a part of mathematics; no one today except perhaps for a few constructivists (who also reject substantial parts of “classical” analysis) questions the soundness of nonstandard analysis as a mathematical theory.

My own view of nonstandard analysis is that the views on both sides of the controversy are overblown. It is a remarkable theoretical achievement that the infinitesimal approach to analysis can be made completely rigorous. And this is of more than theoretical importance: users of calculus can be assured (if indeed they need such assurance) that arguments involving infinitesimals can be carefully justified logically. While it can be shown that (assuming some form of the Axiom of Choice) exactly the same theorems about analysis on $\mathbb{R}$ can be proved using nonstandard analysis as by traditional methods, it can happen (and in fact has happened) that nonstandard arguments are easier to find in some cases, so nonstandard analysis is a tool which should not be rejected or set aside as merely a curiosity. On the other hand, I think it is unlikely in the foreseeable future that nonstandard analysis will supplant the traditional approach. Proponents of nonstandard analysis often underestimate or underappreciate the degree of sophistication in mathematical logic needed to prove even basic theorems about limits via nonstandard analysis. (For example, the book [Gol98] is one of the most readable and appealing treatments of nonstandard analysis; nevertheless, it should be

⁷cf. [?, p. xiii].
noted that this book does not contain a single complete rigorous nonstandard analysis proof of any theorem of real analysis since the transfer principle is not proved in any form. This is not necessarily a criticism of the book, but does tend to compromise the credibility of the author’s forcefully held belief that nonstandard methods are the future of analysis.

The basic idea of nonstandard analysis is a two-part one. First, the real numbers \( \mathbb{R} \) are embedded in a larger ordered field *\( \mathbb{R} \) called the hyperreal numbers, with particular properties, notably including infinitesimal elements. Second, there is a systematic procedure, called the transfer principle, to convert any appropriately formulated statement *\( P \) about *\( \mathbb{R} \), with the property that *\( P \) is true in *\( \mathbb{R} \) if and only if \( P \) is true in \( \mathbb{R} \). One then proves \( P \) by proving *\( P \). Transfer is used in both directions since the transfer of statements already known to be true in \( \mathbb{R} \) can be used to prove *\( P \). It is also often possible to rephrase *\( P \) as a simpler or more natural statement than \( P \) (e.g. the transfer of some statements can be rephrased as a precise version of a classical statement involving infinitesimals). Stating and proving the validity of the transfer principle is the place where mathematical logic comes into nonstandard analysis. The transfer principle is one of Robinson’s fundamental theoretical contributions. (There is another, somewhat more esoteric, feature of more advanced nonstandard analysis, also introduced by Robinson: concurrence, which is discussed in ().)

To avoid the logical technicalities needed to define the hyperreal numbers and prove the transfer principle, the hyperreals and the transfer principle are sometimes presented axiomatically. It is argued that in many ordinary real analysis courses and texts, the real numbers are also presented axiomatically without a rigorous construction, and this approach to nonstandard analysis is no different. This argument is superficially valid, but on closer examination its persuasiveness begins to evaporate:

1. Rigorous mathematics can in fact be done in both settings, although the statements of the theorems proved are really different from the way they are usually expressed. For example, in an ordinary real analysis class in which \( \mathbb{R} \) is presented axiomatically, the theorems are properly of the form

   “If there is a model for \( \mathbb{R} \), then statement \( P \) is true in this model.”

   (Here \( P \) is a statement about real numbers, sets of real numbers, functions on \( \mathbb{R} \), . . .). One must then be convinced of the existence (and essential uniqueness) of a model for \( \mathbb{R} \) for such a theorem to assume its usual form.

   Similarly, theorems in a nonstandard analysis course in which *\( \mathbb{R} \) (as well as \( \mathbb{R} \)) is presented axiomatically will be of the form

   “If a model for \( \mathbb{R} \) exists, and a model for *\( \mathbb{R} \) with a valid transfer principle exists for this model of \( \mathbb{R} \), then statement \( P \) is true in this model of \( \mathbb{R} \).”

   This is quite a different theorem until one is convinced of the existence of models of the hyperreal numbers.

   I can write down any list of axioms I want and prove theorems from them, but until it is shown that there is a model for the axioms, or preferably that the axioms are satisfied in some natural setting with real applications, the theory is an empty one. (This is one of my principal objections to formalist mathematics, although far from my only one.)

2. By the time they begin calculus, most students are prepared, and even comfortable, to accept on faith the existence of a model for \( \mathbb{R} \) satisfying the usual axioms. But few students, even (one might even say especially) good ones, are prepared to accept on faith the existence of a model for the hyperreals. This point is elucidated at length in the introduction to [Gol98]. I generally agree with Goldblatt on this issue, but I would go farther in that I do not share his expectation that eventually beginning calculus students will be as willing to accept the hyperreals as they are to accept \( \mathbb{R} \) today, primarily because of the next point.

351
3. There are rigorous explicit models for $\mathbb{R}$, e.g. via Dedekind cuts, which are sufficiently simple and concrete that they can be followed and understood by almost any undergraduate mathematics major or even honors calculus student. But all (known) models for the hyperreals are at least an order of magnitude more difficult. Even the construction via ultrapowers, which is probably the simplest and most easily understood of the (known) models, is beyond the ability of most undergraduates to fully understand and appreciate. (To begin with, a free ultrafilter on $\mathbb{N}$ is needed. There is no explicit description of any such ultrafilter; indeed, some form of the Axiom of Choice is needed to even prove existence. And even after the construction of $^*\mathbb{R}$ some nontrivial mathematical logic is needed to justify the transfer principle.)

In this section, we will not take an axiomatic approach. We will first construct (a model for) the hyperreals via ultrapowers, and then discuss and justify the transfer principle.

Every serious mathematician needs to be aware of the basics of nonstandard analysis, and to have a tolerant if not enthusiastic attitude about its place in the world of mathematics.

### III.10.1. Ordered Fields Containing $\mathbb{R}$

The basic object of study in nonstandard analysis is an ordered field $^*\mathbb{R}$ of hyperreal numbers which contains $\mathbb{R}$ as a proper subfield. Let us first examine what such an ordered field must look like.

#### III.10.1.1. Definition.

Let $F$ be an ordered field containing $\mathbb{R}$, and let $x, y \in F$.

- $x$ is **limited** if $|x| < r$ for some $r \in \mathbb{R}$.
- $x$ is **unlimited** if it is not limited.
- $x$ is **infinitesimal** if $|x| < r$ for all $r \in \mathbb{R}, r > 0$.

The terms finite and infinite (or infinitely large) are sometimes used instead of limited and unlimited respectively, but these terms can be confusing in certain contexts. According to our definition, 0 is an infinitesimal; this is not entirely standard.

Every element of $\mathbb{R}$ is clearly limited; the only infinitesimal element of $\mathbb{R}$ is 0.

The field $F$ has the Archimedean property () if and only if $F$ has no unlimited or infinitesimal elements.

#### III.10.1.2. Proposition.

Let $F$ be an ordered field containing $\mathbb{R}$, and let $x, y \in F$.

(i) If $x$ and $y$ are limited, so are $x + y$ and $xy$.

(ii) If $x$ and $y$ are infinitesimal, then $x + y$ is infinitesimal.
(iii) If $x$ is infinitesimal and $y$ is limited, then $xy$ is infinitesimal. In particular, if $x$ is infinitesimal and $r \in \mathbb{R}$, then $rx$ is infinitesimal.

(iv) If $x \neq 0$, then $x$ is infinitesimal if and only if $x^{-1}$ is unlimited.

**III.10.1.4. Proposition.** Let $F$ be an ordered field properly containing $\mathbb{R}$. Then

(i) $F$ contains unlimited elements and nonzero infinitesimals.

(ii) If $x \in F$ is limited, then there is a unique $s \in \mathbb{R}$ such that $x - s$ is infinitesimal. (This $s$ is called the *shadow* of $x$, written $s = sh(x)$.)

**Proof:** (ii) Replacing $x$ by $-x$ if necessary, we may assume $x \geq 0$. Let $A = \{ r \in \mathbb{R} : x < r \}$. By assumption, $A$ is nonempty, and it is bounded below (e.g. by 0). Let $s = \inf(A)$ (infimum in $\mathbb{R}$). If $x = s$, we are done. If $x < s$, then $0 < s - x$. If $s - x > r$ for $r \in \mathbb{R}$, $r > 0$, then $x < s - r$, contradicting the definition of $s$. Thus $s - x$ is infinitesimal. If $x > s$, a similar proof shows that $x - s$ is infinitesimal. Uniqueness is obvious from III.10.1.3.(ii)–(iii) if $x - s$ and $x - t$ are infinitesimal, so is $s - t = (x - t) - (x - s)$.

(i): Let $x \in F \setminus \mathbb{R}$. If $x$ is unlimited, then $x^{-1}$ is infinitesimal. If $x$ is limited, let $s = sh(x)$. Then $x - s$ is a nonzero infinitesimal, and $(x - s)^{-1}$ is unlimited.

The next result is also a corollary of (i), but we give a direct proof.

**III.10.1.5. Corollary.** Let $F$ be an ordered field properly containing $\mathbb{R}$. Then $F$ is not order-complete.

**Proof:** Let $I$ be the set of infinitesimals in $F$. Then $I$ is nonempty and bounded (e.g. by 1). Suppose $I$ has a supremum $x$. Since $F$ contains a positive infinitesimal by III.10.1.4., $x > 0$. Also, every positive $r \in \mathbb{R}$ is an upper bound for $I$, so $x$ is infinitesimal. But then $2x$ is infinitesimal and $2x > x$, a contradiction.

**III.10.1.6.** It is obvious that if $x, y$ are limited, then $sh(x + y) = sh(x) + sh(y)$ and $sh(xy) = sh(x)sh(y)$, and that $x$ is infinitesimal if and only if $sh(x) = 0$; $sh(x) = sh(y)$ if and only if $x - y$ is infinitesimal.

**III.10.1.7. Definition.** If $F$ is an ordered field containing $\mathbb{R}$ and $x, y \in F$, then $x \simeq y$ if $x - y$ is infinitesimal. The *halo* of $x$ is

$$hal(x) = \{ y \in F : y \simeq x \} = \{ x + z : z \text{ infinitesimal} \}.$$ 

**III.10.1.8.** It is immediate from III.10.1.3.(ii)–(iii) that $\simeq$ is an equivalence relation; thus, if $x, y \in F$, then either $hal(x) = hal(y)$ or $hal(x)$ and $hal(y)$ are disjoint. If $x$ and $y$ are limited, then $hal(x) = hal(y)$ if and only if $sh(x) = sh(y)$. Each halo of limited elements contains exactly one real number ($hal(x) = hal(sh(x))$ if $x$ is limited).
III.10.1.9. One can also define a weaker equivalence relation: $x \sim y$ if $x - y$ is limited. The $\sim$-equivalence class of $x$ is called the galaxy of $x$. The galaxy of any limited element is the set of all limited elements.

III.10.2. The Hyperreal Numbers via Ultrafilters

It is not obvious that there exist ordered fields properly containing $\mathbb{R}$. A simple example (the simplest example) is given in Exercise III.5.3.1. In this subsection we describe a somewhat more complicated example with additional properties which will allow us to do nonstandard analysis.

III.10.2.10. Recall () that an ultrafilter on a set $X$ is a collection $\mathcal{U}$ of subsets of $X$ such that

(i) $\mathcal{U}$ is closed under finite intersections.

(ii) If $A \in \mathcal{U}$ and $A \subseteq B$, then $B \in \mathcal{U}$.

(iii) $\emptyset \notin \mathcal{U}$.

(iv) If $A \subseteq X$, then either $A \in \mathcal{U}$ or $A^c \in \mathcal{U}$.

The ultrafilter $\mathcal{U}$ is free if the intersection of the sets in $\mathcal{U}$ is $\emptyset$ (equivalently, if every set with finite complement is in $\mathcal{U}$).

We think of an ultrafilter as giving a coherent way of declaring which subsets of $X$ are “large” (those in the ultrafilter) and which are “negligibly small” (those not in the ultrafilter). Every subset is either large or negligibly small, but never both. If the ultrafilter is free, every finite subset is negligibly small.

III.10.2.11. Using the Axiom of Choice, it can be shown () that every infinite set has a free ultrafilter on it; in fact, there are $2^{2^{\aleph_0}}$ free ultrafilters on an infinite set $X$.

Throughout the rest of this subsection, $\mathcal{U}$ will denote a free ultrafilter on $\mathbb{N}$ which is specified once and for all. It does not matter for nonstandard analysis purposes which of the $2^{2^{\aleph_0}}$ free ultrafilters on $\mathbb{N}$ we choose. While we use $\mathbb{N}$ as an index set to make things a simple and concrete as possible, the construction works identically for any index set; and for some purposes a much larger index set is needed (III.10.8.13.), although not too large (III.10.8.14.).

III.10.2.12. Definition. Let $S$ be the set of all sequences of real numbers. Define an equivalence relation on $S$ as follows. If $x = (x_m)$ and $y = (y_m)$ are in $S$, set $x \sim y$ if

$$\{m \in \mathbb{N} : x_m = y_m\} \in \mathcal{U}.$$ Denote the equivalence class of $x$ by $[[x_m]]$, and let $^*\mathbb{R}$ be the set of equivalence classes. $^*\mathbb{R}$ is called the set of hyperreal numbers (with respect to $\mathcal{U}$).

It is easy to check that $\sim$ is an equivalence relation. We think of two sequences as being equivalent if they agree “almost everywhere” (i.e. except on a “negligibly small” set of indices).

III.10.2.13. In order to specify an element of $^*\mathbb{R}$, it is only necessary to specify a “large” set of coordinates (a set in $\mathcal{U}$). The other coordinates can be chosen arbitrarily and do not affect the element of $^*\mathbb{R}$ specified. This observation will be quite useful in avoiding the annoying and unnecessary step of extending definitions to all of $\mathbb{N}$.

354
III.10.2.14. $S$ has natural algebraic operations of addition, subtraction, and multiplication defined coordinatewise. These drop to well-defined operations on $^*\mathbb{R}$ by the following proposition:

**Proposition.** If $(x_m) \sim (z_m)$ and $(y_m) \sim (w_m)$, then $(x_m + y_m) \sim (z_m + w_m)$, $(x_m - y_m) \sim (z_m - w_m)$, and $(x_my_m) \sim (z_mw_m)$.

**Proof:** We have

$$\{m \in \mathbb{N} : x_m + y_m = z_m + w_m \} \supseteq \{m \in \mathbb{N} : x_m = z_m \} \cap \{m \in \mathbb{N} : y_m = w_m \}.$$  

By assumption, the sets on the right are in $U$, hence so is their intersection by III.10.2.10.(i); therefore the set on the left is also in $U$ by III.10.2.10.(ii). The proof in the other two cases is almost identical. \hspace{1cm} \&

III.10.2.15. **Proposition.** The coordinatewise operations on $S$ give well-defined operations of addition, subtraction, and multiplication on $^*\mathbb{R}$ making it into a unital commutative ring $(,)$. The embedding of $\mathbb{R}$ into $^*\mathbb{R}$ as equivalence classes of constant sequences respects addition, subtraction, and multiplication.

**Proof:** It is immediate from III.10.2.15. that addition, subtraction, and multiplication are well defined on $^*\mathbb{R}$, and then almost a triviality that addition and multiplication are associative and commutative, and that the distributive law holds. The additive identity is the equivalence class of the constant sequence $0$, and the multiplicative unit the equivalence class of the constant sequence $1$. The last statement is also obvious. \hspace{1cm} \&

III.10.2.16. **Proposition.** The coordinatewise operations on $S$ give well-defined operations of addition, subtraction, and multiplication on $^*\mathbb{R}$ making it into a unital commutative ring $(,)$. The embedding of $\mathbb{R}$ into $^*\mathbb{R}$ as equivalence classes of constant sequences respects addition, subtraction, and multiplication.

**Proof:** It is immediate from III.10.2.15. that addition, subtraction, and multiplication are well defined on $^*\mathbb{R}$, and then almost a triviality that addition and multiplication are associative and commutative, and that the distributive law holds. The additive identity is the equivalence class of the constant sequence $0$, and the multiplicative unit the equivalence class of the constant sequence $1$. The last statement is also obvious. \hspace{1cm} \&

If $r \in \mathbb{R}$, we will identify $r$ with the equivalence class of the constant sequence $(r)$, and hence regard $\mathbb{R}$ as a subset of $^*\mathbb{R}$; the operations on $^*\mathbb{R}$ extend the usual ones on $\mathbb{R}$.

III.10.2.17. Reciprocals and division can also be defined with only a slight bit of care. Suppose $x = \langle\langle x_m \rangle\rangle \in ^*\mathbb{R}$, $x \neq 0$. Then $\{m \in \mathbb{N} : x_m \neq 0 \} \in U$, so we can define $x^{-1} \in ^*\mathbb{R}$ as the equivalence class of $(x_m^{-1})$ (this sequence is only defined on a set of indices in $U$, but this is sufficient by the remark in III.10.2.13.). It is clear that $xx^{-1} = 1$, so $x^{-1}$ is a multiplicative inverse for $x$. Thus $^*\mathbb{R}$ is a field, and $\mathbb{R}$ is a subfield.

III.10.2.18. We can also define a total ordering on $^*\mathbb{R}$. If $x \in ^*\mathbb{R}$, $x = \langle\langle x_m \rangle\rangle$, we would like to say that $x > 0$ if $\{m \in \mathbb{N} : x_m > 0 \} \in U$. It is easily seen that this definition is independent of the representing sequence $(x_m)$. It gives a good definition due to the next proposition.

III.10.2.19. **Proposition.** If $(x_m) \in S$, then exactly one of the following sets is in $U$:

$$A = \{m \in \mathbb{N} : x_m > 0 \}, \quad B = \{m \in \mathbb{N} : x_m = 0 \}, \quad C = \{m \in \mathbb{N} : x_m < 0 \}.$$ 

**Proof:** By III.10.2.10.(iv), either $A$ or $A^c = \{m \in \mathbb{N} : x_m \leq 0 \}$ is in $U$, but not both. If $A \in U$, then $A^c \notin U$, so $B \notin U$ and $C \notin U$ by III.10.2.10.(ii) since $B, C \subseteq A^c$. Suppose $A^c \in U$; then $A \notin U$. Either $C \in U$ or $C^c \in U$ but not both. If $C \in U$, then $B \notin U$ as above. If $C^c \in U$, then $B = A^c \cap C^c \in U$ and $A, C \notin U$. \hspace{1cm} \&

We can then define the positive cone in $^*\mathbb{R}$. The construction is a special case of taking the nonstandard extension of a subset of $\mathbb{R}$.
**III.10.2.20.** Corollary. Let \( *P \) be the set of \( x = \lfloor (x_m) \rfloor \in *\mathbb{R} \) such that \( \{m \in \mathbb{N} : x_m > 0\} \in \mathcal{U} \). Then \( *P \) satisfies the positive cone axioms (1) and thus defines a total order on \( *\mathbb{R} \) extending the usual order on \( \mathbb{R} \).

**Proof:** By III.10.2.19., \( *\mathbb{R} \) is the disjoint union of \( *P \), \( \{0\} \), and \(-(*P)\). We only need to check that \( *P \) is closed under addition and multiplication, which follows from an argument almost identical to the proof of III.10.2.15., and that \((*P) \cap \mathbb{R}\) is the usual positive cone of \( \mathbb{R} \), which is obvious.

We summarize the construction in a theorem:

**III.10.2.21.** Theorem. \( *\mathbb{R} \) with the above algebraic operations and ordering is an ordered field containing \( \mathbb{R} \).

One can even extend the algebraic operations on \( \mathbb{R} \) such as exponentiation. This is also a special case of a general procedure for extending functions from \( \mathbb{R} \) to \( *\mathbb{R} \).

**III.10.2.22.** Definition. Let \( x = \lfloor (x_m) \rfloor \) and \( a = \lfloor (a_m) \rfloor \) be elements of \( *\mathbb{R} \), with \( x > 0 \). Define

\[
x^a = \lfloor (x_m^a) \rfloor
\]

(note that \( x_m^a \) is defined for almost all \( m \)).

**III.10.2.23.** Proposition. Exponentiation is well defined, and satisfies the usual rules of exponents: if \( x, y, a, b \in *\mathbb{R} \) with \( x, y \geq 0 \), then \( x^a x^b = x^{a+b} \), \( \frac{x^a}{x^b} = x^{a-b} \), \( (x^a)^b = x^{ab} \), \( (xy)^a = x^a y^a \). If \( x < y \) and \( a > 0 \), then \( x^a < y^a \); if \( a < 0 \), then \( x^a > y^a \). If \( x, a \in \mathbb{R} \), exponentiation in \( *\mathbb{R} \) agrees with usual exponentiation in \( \mathbb{R} \).

The proof is routine along the usual lines. See Exercise (1).

**III.10.2.24.** Corollary. Let \( x \in *\mathbb{R} \), \( x > 0 \), and \( a \in *\mathbb{R} \), \( a \neq 0 \). Then there is a unique \( y \in *\mathbb{R} \), \( y > 0 \), such that \( y^a = x \). In particular, if \( n \in \mathbb{N} \), then \( x \) has a unique nonnegative \( n \)th root in \( *\mathbb{R} \).

**III.10.2.25.** We can similarly define logarithms on \( *\mathbb{R} \). If \( 0 < a = \lfloor (a_m) \rfloor \in *\mathbb{R} \) and \( 0 < x = \lfloor (x_m) \rfloor \in *\mathbb{R} \), define

\[
\log_a x = \lfloor (\log_{a_m} x_m) \rfloor
\]

(this is well defined by the usual argument). We then have \( y = \log_a x \) if and only if \( a^y = x \).

**III.10.2.26.** It may not yet be obvious that \( *\mathbb{R} \) is larger than \( \mathbb{R} \). But it is in fact much larger. We can easily define an unlimited element of \( *\mathbb{R} \): set \( x_m = m \) and \( x = \lfloor (x_m) \rfloor \in *\mathbb{R} \). If \( r \in \mathbb{R} \), then \( \{m \in \mathbb{N} : x_m \leq r\} \) is finite, so \( x > r \). Similarly, if \( y_m = m^{-1} \) and \( y = \lfloor (y_m) \rfloor \in *\mathbb{R} \), then \( y \) is a positive infinitesimal in \( *\mathbb{R} \). Actually, if \( (x_m) \) is any sequence of distinct real numbers, then \( \lfloor (x_m) \rfloor \in *\mathbb{R} \setminus \mathbb{R} \).

Note, however, that \( *\mathbb{R} \) is not larger than \( \mathbb{R} \) in a cardinality sense. We have \( |\mathbb{S}| = (2^{|\mathbb{R}|})^{|\mathbb{R}|} = 2^{|\mathbb{R}|} \), so \( |*\mathbb{R}| \leq 2^{|\mathbb{R}|} \), and the opposite inequality is obvious, so \( |*\mathbb{R}| = 2^{|\mathbb{R}|} = |\mathbb{R}| \). (This is a result of taking an ultralimit on \( \mathbb{N} \) for the construction of \( *\mathbb{R} \). If a free ultrafilter is used on a set of large cardinality, the cardinality of \( *\mathbb{R} \) can be larger than \( 2^{|\mathbb{R}|} \). See Exercise III.10.8.10.)

356
III.10.2.27. Like any totally ordered set, \( *\mathbb{R} \) has an order topology \( (\cdot) \). However, this topology has pathological properties (Exercise III.10.8.9.), and is of little if any use in nonstandard analysis.

III.10.3. Internal and External Sets and Functions

Since \( *\mathbb{R} \) is considerably larger than \( \mathbb{R} \), general subsets of \( *\mathbb{R} \) and general functions from \( *\mathbb{R} \) to \( *\mathbb{R} \) are very complicated (and can depend on the specific construction of \( *\mathbb{R} \) used). Relatively few subsets of \( *\mathbb{R} \) or functions from \( *\mathbb{R} \) to \( *\mathbb{R} \) are sufficiently well describable in standard terms that they can be considered “transferable” from \( \mathbb{R} \) using a transfer principle (in a way independent of the construction used for \( *\mathbb{R} \)). In this subsection we describe the ones which can be used, called internal sets and functions.

The Nonstandard Extension of a Subset of \( \mathbb{R} \)

The simplest subsets of \( *\mathbb{R} \) for transfer purposes are ones directly definable from a subset of \( \mathbb{R} \).

III.10.3.1. Definition. Let \( A \subseteq \mathbb{R} \). Then the nonstandard extension of \( A \) is the set

\[
*A = \{ \llbracket (x_m) \rrbracket \in *\mathbb{R} : x_m \in A \text{ for almost all } m \}
\]

where “\( x_m \in A \) for almost all \( m \)” means “\( \{ m \in \mathbb{N} : x_m \in A \} \in \mathcal{U} \).”

Note that if \( \llbracket (x_m) \rrbracket = \llbracket (y_m) \rrbracket \), i.e. \( x_m = y_m \) for almost all \( m \), then \( x_m \in A \) for almost all \( m \) if and only if \( y_m \in A \) for almost all \( m \) (cf. the proof of III.10.2.15.), so membership in \( *A \) is well defined. \( *A \) could also be defined to be simply the set of equivalence classes of sequences in \( A \).

III.10.3.2. Proposition. Let \( A \subseteq \mathbb{R} \). Then

(i) \( *A \cap \mathbb{R} = A \) (in particular, \( A \subseteq *A \)).

(ii) \( *A = A \) if and only if \( A \) is finite.

(iii) \( (*\mathbb{R} \setminus A) = *\mathbb{R} \setminus *A \).

Proof: (i) is obvious: if \( r \in \mathbb{R} \), then \( \llbracket (r) \rrbracket \in *A \) if and only if \( r \in A \). For (ii), if \( (x_m) \) is a sequence of distinct elements of \( A \), then \( \llbracket (x_m) \rrbracket \in *A \setminus A \), so if \( A \) is infinite we have \( *A \neq A \). On the other hand, if \( A = \{ r^{(1)}, \ldots, r^{(n)} \} \), and \( \llbracket (x_m) \rrbracket \in *A \), set \( N_k = \{ m \in \mathbb{N} : x_m = r^{(k)} \} \) for \( 1 \leq k \leq n \). Then \( N_1 \cup \cdots \cup N_n \in \mathcal{U} \), so \( N_k \in \mathcal{U} \) for some (exactly one) \( k \), i.e. \( \llbracket (x_m) \rrbracket = r^{(k)} \). (Note also that \( *\emptyset = \emptyset \).) For (iii), note that if \( \llbracket (x_m) \rrbracket \in *\mathbb{R} \), then exactly one of the sets \( \{ m \in \mathbb{N} : x_m \in A \} \) and \( \{ m \in \mathbb{N} : x_m \notin A \} \) is in \( \mathcal{U} \).

III.10.3.3. Proposition. Let \( A, B \subseteq \mathbb{R} \). Then \( *(A \cup B) = *A \cup *B \), \( *(A \cap B) = *A \cap *B \), and \( *(A \setminus B) = *A \setminus *B \).

Proof: Note that if \( x = \llbracket (x_m) \rrbracket \in *\mathbb{R} \), then

\[
\{ m \in \mathbb{N} : x_m \in A \} \cap \{ m \in \mathbb{N} : x_m \in B \} = \{ m \in \mathbb{N} : x_m \in A \cap B \}
\]
and the set on the right is in $\mathcal{U}$ if and only if both sets on the left are in $\mathcal{U}$, so $x \in ^* (A \cap B)$ if and only if $x \in ^* A \cap ^* B$. The result for unions and symmetric differences follows from this and III.10.3.2.(iii) by taking complements.

III.10.3.4. Note that this statement is false in general for infinite unions and intersections, even countable ones. (Any infinite set is a union of finite sets; use III.10.3.2.(ii).)

The proof of the next proposition is routine along the same lines:

III.10.3.5. Proposition. Let $a, b \in \mathbb{R}$, $a < b$, and denote by $[a, b]$, $(a, b)$, $(a, b]$, $[a, b)$ the usual intervals in $\mathbb{R}$. Then

\begin{align*}
^*[a, b] &= \{ x \in ^\ast \mathbb{R} : a \leq x \leq b \} \\
^*(a, b) &= \{ x \in ^\ast \mathbb{R} : a < x < b \} \\
^*(a, b] &= \{ x \in ^\ast \mathbb{R} : a < x \leq b \} \\
^*[a, b) &= \{ x \in ^\ast \mathbb{R} : a \leq x < b \} .
\end{align*}

Similar statements hold for $^*(-\infty, a]$, $^*(a, +\infty)$, etc.

III.10.3.6. Proposition. Let $A \subseteq \mathbb{R}$. Then the following are equivalent:

(i) $^*A$ is bounded above in $^\ast \mathbb{R}$.

(ii) Every positive element of $^*A$ is limited.

(iii) $A$ is bounded above in $\mathbb{R}$.

If these conditions hold, any upper bound for $A$ in $\mathbb{R}$ is an upper bound for $^*A$ in $^\ast \mathbb{R}$.

Proof: (ii) $\Rightarrow$ (i) is trivial: any unlimited positive element of $^\ast \mathbb{R}$ would be an upper bound for $^*A$. For (iii) $\Rightarrow$ (ii), and for the last statement, if $b$ is an upper bound for $A$ in $\mathbb{R}$, then $^*A \subseteq ^*(-\infty, b]$, so by III.10.3.5. $b$ is an upper bound for $^*A$ and (ii) holds. To prove (i) $\Rightarrow$ (iii), suppose $A$ is not bounded above in $\mathbb{R}$. Let $x = \lfloor \{x \} \rfloor \in ^\ast \mathbb{R}$. For each $m$ let $y_m \in A$ with $y_m \geq x_m$. Then $y = \lfloor \{y \} \rfloor \in ^*A$ and $x \leq y$.

III.10.3.7. Perhaps the most important nonstandard extension set is $^*\mathbb{N}$, the hypernatural numbers, equivalence classes of sequences of natural numbers. We also have $^*\mathbb{Z}$, the hyperintegers, and $^*\mathbb{Q}$, the hyperrational numbers. We have that $^*\mathbb{N}$ is closed under addition and multiplication, $^*\mathbb{Z}$ is a subring of $^*\mathbb{R}$, and $^*\mathbb{Q}$ is a subfield of $^*\mathbb{R}$. Every element of $^*\mathbb{Z}$ is a difference of two elements of $^*\mathbb{N}$, and every element of $^*\mathbb{Q}$ is a quotient of an element of $^*\mathbb{Z}$ by an element of $^*\mathbb{N}$. All of these sets turn out to be uncountable ( ). See Exercises ( ) for some properties of $^*\mathbb{N}$.
Nonstandard Extensions of Functions

III.10.3.8. Let \( A \subseteq \mathbb{R} \), and let \( f \) be a function from \( A \) to \( \mathbb{R} \). We can extend \( f \) to a function \( *f \) from \( *A \) to \( *\mathbb{R} \) as follows. Let \( x = [((x_m))] \in *A \). Then \( x_m \in A \) for almost all \( m \), so the element \( [((f(x_m)))] \) is well defined as an element of \( *\mathbb{R} \). It is easily seen in the usual way that this element of \( *\mathbb{R} \) is independent of the choice of the representative \( (x_m) \) chosen for \( x \). Write \( *f(x) = [((f(x_m)))] \).

III.10.3.9. If \( f \) is a polynomial or rational function with real coefficients, then \( *f \) defined this way agrees with the usual algebraic extension of \( f \) to a function on \( \mathbb{R} \). More generally, if \( a \in \mathbb{R} \), and \( f(x) = x^a \) on \( [0, +\infty) \), then \( *f \) is the function on \( *[0, +\infty) \) defined by nonstandard exponentiation (III.10.2.22.). Similarly, if \( 0 < a \in \mathbb{R} \), the nonstandard extension of the exponential function \( f(x) = a^x \) is given by nonstandard exponentiation. If \( 0 < a \in \mathbb{R} \), then the nonstandard extension of the function \( f(x) = \log_a x \) is the nonstandard logarithm function of III.10.2.25..

III.10.3.10. Another common instance of nonstandard extension is for trigonometric functions. The functions \( *\sin x \) and \( *\cos x \) are defined for all nonstandard real numbers; the domain of \( *\tan x \) is the set

\[
* \left( \mathbb{R} \setminus \left\{ \frac{\pi}{2} + n\pi : n \in \mathbb{N} \right\} \right) = *\mathbb{R} \setminus \left\{ \frac{\pi}{2} + n\pi : n \in *\mathbb{N} \right\}
\]

and the other extended trig functions have similar domains.

III.10.3.11. It is usual to omit the * on the nonstandard extension of a function except in the relatively rare case (e.g. in III.10.3.14.–III.10.3.15.) where a distinction must be made between the original function and its extension. Thus, for example, we will usually just write \( \sin, \log_a, \) etc., for the extended functions.

The proof of the next result is routine by the usual arguments:

III.10.3.12. Proposition. All standard trigonometric identities hold for the nonstandard extensions. For example:

\[
\begin{align*}
\sin^2 x + \cos^2 x &= 1 \text{ for all } x \in *\mathbb{R}. \\
\sin(x + y) &= \sin x \cos y + \cos x \sin y \text{ for all } x, y \in *\mathbb{R}.
\end{align*}
\]

III.10.3.13. Inequalities valid for standard functions also hold for their extensions. For example:

\[
\begin{align*}
e^x &> 0 \text{ for all } x \in *\mathbb{R}; \ e^x < 1 \text{ for } x < 0 \text{ and } e^x > 1 \text{ for } x > 0. \\
\log x &< 0 \text{ for } x \in *\mathbb{R}, \ 0 < x < 1; \ \log x > 0 \text{ for } x \in *\mathbb{R}, \ x > 1. \\
-1 &\leq \sin x \leq 1 \text{ and } -1 \leq \cos x \leq 1 \text{ for all } x \in *\mathbb{R}.
\end{align*}
\]

If the domain of \( f \) is \( A \), then the domain of \( *f \) is by definition \( *A \). We also have:
III.10.3.14. Proposition. If \( f : A \to \mathbb{R} \) and \( B \subseteq A \), then \( *f(*B) = *[f(B)] \). In particular, the range of \( *f \) is \( *[f(A)] \).

Proof: If \( \mathbf{x} = [[[x_m]]] \in *B \), then \( x_m \in B \) for almost all \( m \), so \( f(x_m) \in f(B) \) for almost all \( m \) and \( [[[f(x_m)]]] \in *[f(B)] \). Conversely, if \( \mathbf{y} = [[[y_m]]] \in *[f(B)] \), then for almost all \( m \) we have \( y_m \in f(B) \), so there is an \( x_m \in B \) with \( y_m = f(x_m) \). The element \( \mathbf{x} = [[[x_m]]] \) is well defined and in \( *B \), and \( *f(\mathbf{x}) = \mathbf{y} \), so \( \mathbf{y} \in *f(*B) \).

III.10.3.15. If \( f : A \to \mathbb{R} \), \( g : B \to \mathbb{R} \), and \( f(A) \subseteq B \), then \( g \circ f : A \to \mathbb{R} \) is defined. It is routine to check that \( *(g \circ f) = *g \circ *f \) as a function from \( *A \) to \( *\mathbb{R} \). In particular, if \( f \) is one-to-one on \( A \), then \( *(f^{-1}) \circ *f \) is the identity on \( *A \) and \( *f \circ *(f^{-1}) \) is the identity on \( *[f(A)] \), so \( *f \) is a bijection from \( *A \) to \( *[f(A)] \) and \( *(f^{-1}) = *(f^{-1}) \).

Nonstandard Euclidean Space

III.10.3.16. For fixed \( n \in \mathbb{N} \), there is a natural one-one correspondence between \( \mathbb{S}^n \), the set of \( n \)-tuples of sequences of real numbers, and the set of sequences of \( n \)-tuples of real numbers:

\[
((x_{m}^{(1)}, x_{m}^{(2)}, \ldots, x_{m}^{(n)})) \leftrightarrow ((x_{m}^{(1)}, x_{m}^{(2)}, \ldots, x_{m}^{(n)}))
\]

and we have that

\[
(x_{m}^{(1)}, x_{m}^{(2)}, \ldots, x_{m}^{(n)}) = (y_{m}^{(1)}, y_{m}^{(2)}, \ldots, y_{m}^{(n)})
\]

for almost all \( m \) if and only if, for each \( k \), \( 1 \leq k \leq n \), we have \( x_{m}^{(k)} = y_{m}^{(k)} \) for almost all \( m \). Thus there is an unambiguous equivalence relation on the set of sequences of \( n \)-tuples of real numbers, and the set \( *(\mathbb{R}^n) \) of equivalence classes can be naturally identified with \( (*\mathbb{R})^n \), the set of \( n \)-tuples of hyperreal numbers. We will denote this set by \( *\mathbb{R}^n \) and freely describe it both ways. Write \( [[[x_{m}^{(1)}, \ldots, x_{m}^{(n)}]]] \) for the equivalence class of the sequence \( ((x_{m}^{(1)}, \ldots, x_{m}^{(n)})) \), thought of as a typical representative of an element of \( *\mathbb{R}^n \).

The Nonstandard Extension of a Relation

The nonstandard extensions of sets and functions can be subsumed in the general notion of the nonstandard extension of an \( n \)-ary relation.

III.10.3.17. Recall that an \( n \)-ary relation on \( \mathbb{R} \) is simply a subset of \( \mathbb{R}^n \). In particular, a usual (binary) relation is a subset of \( \mathbb{R}^2 \); recall that a function is officially a binary relation (its graph). A unary relation is just a subset of \( \mathbb{R} \).

III.10.3.18. Definition. If \( A \) is a subset of \( \mathbb{R}^n \), i.e. an \( n \)-ary relation on \( \mathbb{R} \), we define its nonstandard extension to be

\[
*A = \{ [[[x_{m}^{(1)}, \ldots, x_{m}^{(n)}]]] : (x_{m}^{(1)}, \ldots, x_{m}^{(n)}) \in A \text{ for almost all } m \} \subseteq *\mathbb{R}^n .
\]

If \( n = 1 \), this agrees with the previous definition of the nonstandard extension of a subset of \( \mathbb{R} \). It is also easy to verify that the nonstandard extension of a function on \( \mathbb{R} \) coincides with the nonstandard extension of the corresponding binary relation (the graph of the function).
The binary relation $<$ on $\mathbb{R}$ has a nonstandard extension $^*<$, which coincides with the ordering $<$ on $^*\mathbb{R}$ defined in III.10.2.20.. Even the binary relation $=$ on $\mathbb{R}$ has ordinary equality on $^*\mathbb{R}$ as its nonstandard extension.

A function from $\mathbb{R}^n$ to $\mathbb{R}^k$ can be regarded as an $(n+k)$-ary relation on $\mathbb{R}$, and thus has a nonstandard extension, which can be defined similarly to the nonstandard extension of a function from $\mathbb{R}$ to $\mathbb{R}$. For some simple examples, addition and multiplication may be regarded as functions from $\mathbb{R}^2$ to $\mathbb{R}$; their nonstandard extensions are nothing but addition and multiplication on $^*\mathbb{R}$ as we have defined them. Another example: $f(x, y) = \log_x y$ is a real-valued function defined on $(0, \infty)^2$, and its nonstandard extension is the logarithm function defined in III.10.2.25.

III.10.4. The Transfer Principle

By now we have shown that a great many statements about the real numbers have nonstandard analogs, and the proofs of the nonstandard statements from the standard ones all seem pretty much the same. A few classes of examples (phrased informally) are:

- $\mathbb{R}$ satisfies the ordered field axioms.
- The rules of exponents hold in $\mathbb{R}$.
- Trigonometric identities hold in $\mathbb{R}$.

We might suspect that there is a metaprinciple that every statement about $\mathbb{R}$ has an analog in $^*\mathbb{R}$ and that there is a single scheme for simultaneously deriving proofs of the nonstandard versions of all true standard statements. With due care, there is: this is the transfer principle.

However, we must be careful. Not every true statement about $\mathbb{R}$ can have a nonstandard version. It is easy to find counterexamples: a simple one is “$\mathbb{R}$ contains no nonzero infinitesimals.” A more important example is the Completeness Property of $\mathbb{R}$; for somewhat deeper reasons (III.10.1.5.) this statement cannot be true verbatim about $^*\mathbb{R}$.

So we must instead say that every appropriately formulated statement about $\mathbb{R}$ has a nonstandard analog, which is true if and only if the standard version is true. Precisely describing what “appropriately formulated” means is where mathematical logic comes into the picture.

The transfer principle has two levels: first-order transfer and higher-order transfer. First-order transfer is relatively simple to formulate and even not too hard to prove, and only requires a modest amount of mathematical logic; and first-order transfer is adequate for most elementary applications of nonstandard analysis. Higher-order transfer is more sophisticated logically, but also more powerful. We will concentrate on first-order transfer and only give a brief survey of higher-order transfer.

Simple or First-Order Transfer

The setup for first-order transfer will be to begin with a relational structure $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$ (II.1.8.1.), which will be called the standard structure. In almost all applications $\mathcal{R}$ and $\mathcal{F}$ will be the sets of all relations and functions on $\Omega$ respectively; in basic nonstandard analysis $\Omega$ will be $\mathbb{R}$ with $\mathcal{R}$ and $\mathcal{F}$ the sets of all relations and functions. At first the reader may want to just assume we are working with this structure.

A subset of $\Omega^n$ can be regarded as an $n$-ary relation on $\Omega$; for technical reasons we will assume that the domain of any function in $\mathcal{F}$ is a relation in $\mathcal{R}$.
In addition, we will fix an index set $I$ and an ultrafilter $\mathcal{U}$ on $I$. For basic nonstandard analysis we take $I = \mathbb{N}$ and $\mathcal{U}$ any free ultrafilter on $\mathbb{N}$.

III.10.4.2. The first step is to define a nonstandard version of $S$. We let $^\ast \Omega$ be the set of equivalence classes of functions from $I$ to $\Omega$, where two functions are equivalent if they agree “almost everywhere” (abbreviated a.e.), or “for almost all $i$,” i.e. on a set in $\mathcal{U}$. In the basic nonstandard analysis case where $\Omega = \mathbb{R}$ and $I = \mathbb{N}$ the set $^\ast \Omega$ is just $^\ast \mathbb{R}$ as defined in (i).

III.10.4.3. If $R$ is an $n$-ary relation in $\mathcal{R}$, we define a nonstandard version $^\ast R$, which will be an $n$-ary relation on $^\ast \Omega$, by letting $^\ast R$ be the set of all $n$-tuples $(\omega_1, \ldots, \omega_n)$ for which there is for each $k$ a representation $\omega_k = [[(\omega_{ki})]]$ such that $$(\omega_{i1}, \ldots, \omega_{in}) \in R$$ for almost all $i \in I$. A routine argument along the lines of () shows that if this condition is satisfied for one $n$-tuple of representatives of $\omega_1, \ldots, \omega_n$, it is satisfied for any other $n$-tuple of representatives. In the basic nonstandard analysis case, $^\ast R$ is just the relation defined in III.10.3.18.. Similarly, if $f \in F$ is a function from a subset of $\Omega^n$ to $\Omega$, we define $^\ast f$ to be the function from a subset of $(^\ast \Omega)^n$ to $^\ast \Omega$ given by $^\ast f(\omega_1, \ldots, \omega_n) = [[(f(\omega_{i1}, \ldots, \omega_{in}))]]$ if $\omega_k = [[(\omega_{ki})]]$ for each $k$ and $(\omega_{i1}, \ldots, \omega_{in})$ is in the domain of $f$ for almost all $i$. It is again routine to check that $^\ast f$ is well defined independent of the choice of representatives, and that the domain of $^\ast f$ is $^\ast (\text{dom } f)$ (recall that $\text{dom } f$ is a relation in $\mathcal{R}$, so $^\ast (\text{dom } f)$ is a defined subset of $(^\ast \Omega)^n$). In the basic nonstandard analysis case, $^\ast f$ is just the function defined in III.10.3.8..

III.10.4.4. Definition. The relational structure $^\ast S = (^\ast \Omega, ^\ast \mathcal{R}, ^\ast \mathcal{F})$, where $^\ast \mathcal{R} = \{^\ast R : R \in \mathcal{R}\}$ and $^\ast \mathcal{F} = \{^\ast f : f \in \mathcal{F}\}$, is called the (first-order) nonstandard structure for $S$ (with respect to $I$ and $\mathcal{U}$).

Although the notation does not reflect it, $^\ast S$ depends on $I$ and $\mathcal{U}$ as well as on $S$. However, for nonstandard analysis purposes the dependence on $I$ and $\mathcal{U}$ is mostly irrelevant and can be largely ignored (but see III.10.8.13. and III.10.8.14.). Note that even if $\mathcal{R}$ is the set of all relations on $\Omega$, $^\ast \mathcal{R}$ will be far from containing all relations on $^\ast \Omega$ in general, and similarly for $^\ast \mathcal{F}$. The relations and functions in $^\ast \mathcal{R}$ and $^\ast \mathcal{F}$ will be the relations and functions on $^\ast \Omega$ which are definable by transfer of definitions in $\Omega$.

III.10.4.5. Next we fix a simple language $\mathcal{L}$ for the relational structure $S$ (), and a language $^\ast \mathcal{L}$ for the relational structure $^\ast S$. We do not require that $^\ast \mathcal{L}$ be a simple language for $^\ast S$, e.g. it could be a simple language for a larger relational structure such as $(^\ast \Omega, ^\ast \mathcal{R}, ^\ast \mathcal{F})$, where $^\ast \mathcal{R}$ and $^\ast \mathcal{F}$ are the sets of all relations and all functions on $^\ast \Omega$. But the only formulas in $^\ast \mathcal{L}$ we will consider are ones from the simple sublanguage for $^\ast S$, so we may restrict to this sublanguage when convenient. We may assume that the variable names in $^\ast \mathcal{L}$ contain all the variable names in $\mathcal{L}$ (in fact, since we only ever use a countable set of variables, we could always use a common set of variable names in all languages).

We then define a “transfer” of each symbol, term, and formula of $\mathcal{L}$ into a corresponding symbol, term, or formula in $^\ast \mathcal{L}$:
III.10.4.6. Definition. (i) If \( c \) is the name of an element \( \omega \in \Omega \), then the transfer \( ^*c \) of \( c \) is a name of the element of \( ^*\Omega \) defined by the constant function \( \omega \) from \( I \) to \( \Omega \). (We will notationally identify the element \( c \in \mathcal{L} \) with \( \ ^*c \in \ ^*\mathcal{L} \), i.e. we will regard \( \Omega \subseteq ^*\Omega \), to simplify notation.)

(ii) If \( R \in \mathcal{R} \), the transfer of \( R \) is \( ^*R \in ^*\mathcal{R} \).

(iii) If \( f \in \mathcal{F} \), the transfer of \( f \) is \( ^*f \in ^*\mathcal{F} \).

(iv) The transfer of any variable or logical symbol \( (\forall, \Rightarrow) \) is itself.

(v) The transfer of any term is defined inductively: the transfer of a constant or variable has already been defined, and the transfer of a term \( f(t_1, \ldots, t_n) \) is \( ^*f(^*t_1, \ldots, ^*t_n) \), where \( ^*t_k \) is the transfer of \( t_k \) for each \( k \).

Transfer of formulas is then defined inductively:

III.10.4.7. Definition. (i) The transfer of the formula \( R(t_1, \ldots, t_n) \) is \( ^*R(^*t_1, \ldots, ^*t_n) \), where \( ^*t_k \) is the transfer of the term \( t_k \) for each \( k \).

(ii) The transfer of the formula \( \forall x_1 \cdots \forall x_n(P) \) is \( \forall x_1 \cdots \forall x_n(^*P) \), where \( ^*P \) is the transfer of \( P \).

(iii) The transfer of the formula \( P_1 \land \cdots \land P_n \) is \( (^*P_1) \land \cdots \land (^*P_n) \), where \( ^*P_k \) is the transfer of \( P_k \) for each \( k \).

(iv) The transfer of the formula \( P_1 \Rightarrow P_2 \) is \( (^*P_1) \Rightarrow (^*P_2) \), where \( ^*P_k \) is the transfer of \( P_k \) for each \( k \).

(v) The transfer of any formula of level \( n + 1 \) is defined by application of the appropriate one of (i)–(iv) to the constituent formulas of level \( n \).

A routine induction argument verifies:

III.10.4.8. Proposition. The transfer of any symbol, term, or formula in \( \mathcal{L} \) is a well-defined symbol, term, or formula in \( ^*\mathcal{L} \). For any formula \( P \), the variables appearing in \( ^*P \) are the same as the ones appearing in \( P \); and a variable is quantified in \( ^*P \) if and only if it is quantified in \( P \). In particular, the transfer of any statement in \( \mathcal{L} \) is a statement in \( ^*\mathcal{L} \).

The Transfer Principle is:

III.10.4.9. Theorem. [Simple or First-Order Transfer Principle] Let \( S \) be a relational structure, \( ^*S \) a first-order nonstandard structure for \( S \), \( \mathcal{L} \) a simple language for \( S \), and \( ^*\mathcal{L} \) a language for \( ^*S \). If \( P \) is any statement of \( \mathcal{L} \), with transfer \( ^*P \) as defined in III.10.4.7., then \( ^*P \) holds in \( ^*S \) if and only if \( P \) holds in \( S \).

III.10.4.10. This result is symmetric in the sense that transfer goes both ways due to the “if and only if.” But it is important to note that transfer is not quite symmetric in \( \mathcal{L} \) and \( ^*\mathcal{L} \): every statement in \( \mathcal{L} \) has a transfer in \( ^*\mathcal{L} \), but not every statement in \( ^*\mathcal{L} \) is the transfer of a statement in \( \mathcal{L} \) in general. If \( ^*\mathcal{L} \) is a simple language for \( ^*S \), then transfer is truly symmetric, since the only formulas of \( ^*\mathcal{L} \) will be transfers of formulas in \( \mathcal{L} \).
We can extend transfer a little to make it more useful. If the relational structure $\mathcal{S}$ is nondegenerate (II.1.8.8.), we can fix a false statement $\Phi$ in $\mathcal{L}$ containing no variables. Then $^*\Phi$ is also false in $^*\mathcal{S}$, and contains no variables. We can extend the simple language $\mathcal{L}$ to an extended simple language $\mathcal{L}'$ by adding the symbols $\neg, \lor, \exists, \iff$ as in II.1.8.10., with a standard translation as in II.1.8.10. using $\Phi$; similarly, if $^*\mathcal{L}$ is simple, or more generally does not contain the extra logical symbols, we can extend $^*\mathcal{L}$ to $^*\mathcal{L}'$ in the same manner, using $^*\Phi$. We then routinely check:

**III.10.4.12.** Proposition. Each formula $P$ of $\mathcal{L}$ has a well-defined transfer to a formula $^*P$ of $^*\mathcal{L}$. If $Q$ is the standard translation of $P$ to $\mathcal{L}$, then the standard translation of $^*P$ to $^*\mathcal{L}$ is $^*Q$. The transfer of a statement is a statement.

Using this, we can extend Transfer:

**III.10.4.13.** Theorem. [Extended Simple or First-Order Transfer Principle] Let $\mathcal{S}$ be a relational structure, $^*\mathcal{S}$ a first-order nonstandard structure for $\mathcal{S}$, $\mathcal{L}$ a simple language for $\mathcal{S}$, $^*\mathcal{L}$ a simple language for $^*\mathcal{S}$, and $^*\mathcal{L}'$ the corresponding extended languages. If $P$ is any statement of $^*\mathcal{L}$, with transfer $^*P$ as defined in III.10.4.12., then $^*P$ holds in $^*\mathcal{S}$ if and only if $P$ holds in $\mathcal{S}$.

Proof: Let $Q$ be the standard translation of $P$ in $\mathcal{L}$. Then

$$
P \text{ holds in } S \iff Q \text{ holds in } S \iff ^*Q \text{ holds in } ^*S \iff ^*P \text{ holds in } ^*S$$

where the first and third equivalences are by definition (III.10.4.12. is needed for the third) and the second equivalence is by III.10.4.9.. ⬅️

**Proof of the Simple Transfer Principle**

The proof of the Simple Transfer Principle, by induction on level, is actually quite straightforward and routine. It is a nearly immediate corollary of the simple version of Loś’s Theorem:

**III.10.4.14.** Theorem. [Loś’s Theorem, Simple or First-Order Version] Let $\mathcal{S}$ be a relational structure, $^*\mathcal{S}$ a first-order nonstandard structure for $\mathcal{S}$, $\mathcal{L}$ a simple language for $\mathcal{S}$, and $^*\mathcal{L}$ a language for $^*\mathcal{S}$. Let $P$ be any formula of $\mathcal{L}$, with transfer $^*P$ as defined in III.10.4.7., with unquantified variables among $x_1, \ldots, x_n$; let $c_1, \ldots, c_n$ be names in $^*\mathcal{L}$ for elements of $^*\Omega$ with $(c_ki)$ a representative for $c_k$ with $c_ki$ the name of an element of $\Omega$ for each $(k, i)$. Then $^*P$ holds for $(c_1, \ldots, c_n)$ in $^*\mathcal{S}$ if and only if, for almost all $i$, $P$ holds for $(c_{1i}, \ldots, c_{ni})$ in $\mathcal{S}$. 364
Superstructures

**III.10.4.15. Definition.** If $X$ is a set, we define

\[
V_0(X) = X \\
V_{n+1}(X) = V_n(X) \cup \mathcal{P}(V_n(X)) \\
V(X) = \bigcup_{n=0}^{\infty} V_n(X)
\]

$V(X)$ is called the *superstructure* of $X$.

Note that if $y \in V(X) \setminus X$, then $y \subseteq V(X)$. The converse is false; however, if $y \subseteq V_n(X)$ for some $n$, then $y \in V_{n+1}(X) \subseteq V(X)$.

**III.10.4.16.** There is a similarity between the construction of superstructures and II.8.7.6. However, $V(X)$ is not a model for set theory (for example, $\mathcal{P}(V(X))$ is not an element of $V(X)$); to get a model for set theory, the construction has to be continued transfinitely for all ordinals (and then the result is a class, not a set). For nonstandard analysis, however, it is adequate to just continue the construction to $\aleph_0$.

**III.10.4.17.** For technical reasons, we will want to assume that $\emptyset \notin X$ and that no element of $V(X)$ belongs to an element of $X$, i.e. that $x \cap V(X) = \emptyset$ for all $x \in X$. This is normally accomplished by assuming that the elements of $X$ are *atoms* (or *urelements*), i.e. if $x \in X$, then $x \neq \emptyset$ but that $x$ has no elements. For a general $X$ we can just replace $X$ by an equipotent set consisting of atoms. For basic nonstandard analysis, we simply assume that real numbers are atoms of set theory, which we can do (III.2.2.9). We will call the elements of $X$ the *elements* of the superstructure $V(X)$, and the elements of $V(X) \setminus X$ the *sets* of the superstructure $V(X)$. If $A$ is a set in $V(X)$, then all the elements of $A$ are either elements or sets in $V(X)$.

**III.10.4.18.** We can define ordered pairs in $V(X)$: if $a, b \in V_k(X)$, then \{a\} and \{a, b\} are in $V_{k+1}(X)$, so $(a, b) = \{\{a\}, \{a, b\}\} \in V_{k+2}(X)$. Similarly, ordered $n$-tuples of elements of $V(X)$ are again elements of $V(X)$. In particular, all ordered $n$-tuples of elements of $X$ are in $V_2(X)$, and thus $X^n \in V_3(X)$. More generally, the cartesian product of any finite number of subsets of $V_k(X)$ is again an element of $V_{k+3}(X) \subseteq V(X)$.

An $n$-ary relation on $V_k(X)$ is a subset of $V_k(X)^n$, hence an element of $V_{k+3}(X)$. Thus $V(X)$ contains all relations on $V_k(X)$ for any $k$. Any such relation is a relation on $V(X)$. But note that not all relations on $V(X)$ are of this form. An $n$-ary relation on $V(X)$ which is contained in $V_k(X)^n$ for some $k$ is called *limited*; these are precisely the relations on $V(X)$ which are elements of $V(X)$.

Similarly, any function from $V_k(X)^n$ to $V_k(X)^m$ is in $V_{k+4}(X)$, hence in $V(X)$. In particular, any function from $X^n$ to $X^m$ is in $V_4(X)$, hence in $V(X)$. As with relations, not every function on $V(X)$ is in $V(X)$, only ones whose domain and range is contained in $V_k(X)$ for some $k$ (called *limited functions*).

**III.10.4.19.** We can thus form a relational structure $(V(X), \mathcal{R}_f, \mathcal{F}_f)$, where $\mathcal{R}_f$ and $\mathcal{F}_f$ are the sets of limited relations and limited functions in $V(X)$. It is convenient to only consider limited relations and functions and not the unwieldy sets of all relations and functions on $V(X)$. However, there are a few basic relations and functions on $V(X)$ which are not limited, which it is convenient to include: $\in, \subseteq, \cup, \cap, \setminus$. These can be avoided at some notational cost: for example, $\in$ can be avoided by only working with the coherent family of limited relations $\in_k$, where $\in_k$ is the membership relation on $V_k(X)$. We will thus include
these five relations in $\mathcal{R}_\ell$ and $\cup$, $\cap$, and \setminus in $\mathcal{F}_\ell$, essentially to ease notation. (Actually any relations or functions on $V(X)$ can be thrown in with no cost by regarding them as coherent families of limited relations or functions in the same manner.)

**III.10.4.20.**  We can then fix a simple language (II.1.8.2.) $\mathcal{L}_X$ for the relational structure $\mathcal{X} = (V(X), \mathcal{R}_\ell, \mathcal{F}_\ell)$, which will be called the **language** of the superstructure of $X$. Since this relational structure is always nondegenerate (even if $X = \emptyset$!), the language $\mathcal{L}_X$ can be extended to an extended simple language $\bar{\mathcal{L}}_X$ adding $\neg$, $\lor$, $\exists \iff$ (II.1.8.10.), which will have their usual meanings. The language $\bar{\mathcal{L}}_X$ is rich enough that we can express any reasonable statement about $X$, subsets of $X$, relations and functions on $X$, etc., in a relatively straightforward way.

**III.10.4.21.**  Some references argue that inclusion of $\mathcal{F}_\ell$ in the relational structure is not necessary: in a relational structure $(\Omega, \mathcal{R}, \mathcal{F})$, the only use of the functions in $\mathcal{F}$ is to define terms; in the case of $(V(X), \mathcal{R}_\ell, \mathcal{F}_\ell)$ this is redundant since the terms defined this way are already constants in $V(X)$. Thus the use of terms, as opposed to constants, is unnecessary. This is technically correct, but terms in a simple language can include variables as well as constants, so inclusion of the functions in $\mathcal{F}_\ell$ serves to expand the simple language of the superstructure in a useful way without causing additional difficulties (sentences in the expanded language can be systematically translated into the simpler language).

**Nonstandard Superstructures**

**III.10.4.22.**  If we now fix and index set $I$ and an ultrafilter $\mathcal{U}$ on $I$, we can form the corresponding nonstandard relational structure $^*\mathcal{X} = (^*V(X), ^*\mathcal{R}_\ell, ^*\mathcal{F}_\ell)$. We would like to regard this relational structure as a (generally proper) substructure of $(V(^*X), \mathcal{R}_\ell^*, \mathcal{F}_\ell^*)$, where $\mathcal{R}_\ell^*$ and $\mathcal{F}_\ell^*$ are the sets of limited relations and functions on $V(^*X)$. However, there is a logical problem: $^*V(X)$ is not a subset of $V(^*X)$ (although $^*X$ is naturally a subset of both). For example, if $A \subseteq X$, then the object $^*A$ of $^*V_1(X)$ is not a subset of $^*X$. Put another way, $A$ may be regarded as a unary relation $R_A$ on $X$, i.e. as an element of $\mathcal{R}_\ell$, but the element $^*A$ of $^*V_1(X)$ is not the unary relation $^*R_A$ on $^*X$ (which is technically a subset of $^*X$, i.e. an element of $V(^*X)$).

This observation leads to the solution: identify $^*V_2(X) = X \cup \mathcal{P}(X)$ with a subset of $V_1(^*X) = X \cup \mathcal{P}(^*X)$ by identifying $^*x \in X$ with itself and $A \subseteq X$ with $^*R_A \in \mathcal{P}(^*X)$. The embedding is well defined since $X \cap \mathcal{P}(X) = \emptyset$ and hence $^*X \cap \mathcal{P}(X) = \emptyset$, and is indeed an embedding (injective) since $^*X \cap \mathcal{P}(^*X) = \emptyset$. Call the embedding $\mu_0$.

This procedure can be continued inductively, although with a slight complication. The embedding $\mu_1$ of $^*V_1(X)$ into $V_2(^*X)$ defines an embedding of $\mathcal{P}(^*V_1(X))$ into $\mathcal{P}(V_1(^*X))$, so we can map $^*V_2(X) = V_1(^*X) \cup \mathcal{P}(V_1(^*X))$ into $V_2(^*X) = V_1(^*X) \cup \mathcal{P}(V_1(^*X))$ as follows. The map $\mu_2$ is defined on $^*V_1(X)$ as $\mu_1$. If $B \subseteq V_1(X)$, so $^*B \in ^*\mathcal{P}(V_1(X))$, regard $B$ as a unary relation $R_B$ on $V_1(X)$; then $^*R_B$ is a unary relation on $^*V_1(X)$, hence defines an element of $\mathcal{P}(^*V_1(X))$ which in turn defines an element $C$ of $\mathcal{P}(V_1(^*X))$. Set $\mu_2(^*B) = C$. It is not entirely obvious that $\mu_2$ is well defined, since $V_1(X) \cap \mathcal{P}(V_1(X)) \neq \emptyset$; but in fact $V_1(X) \cap \mathcal{P}(V_1(X)) = \{\emptyset\}$ and $^*\emptyset = \emptyset$ unambiguously. It is unclear that $\mu_2$ is injective in general.

Iterating this construction, we get a well-defined map $\mu_n$ from $^*V_n(X)$ to $V_n(^*X)$ for each $n$ (the proof that the map is well defined becomes successively more complicated as $n$ increases), and hence a map $\mu : ^*V(X) \to V(^*X)$, called the **Mostowski collapsing map**. We also need to identify relations in $^*\mathcal{R}_\ell$ with relations in $\mathcal{R}_\ell^*$: since every relation in $^*\mathcal{R}_\ell$ is an element of
III.10.5. Other Nonstandard Structures

III.10.6. Proofs and Applications of Nonstandard Analysis

III.10.7. Loeb Measures and a Nonstandard Approach to Measure Theory

III.10.8. Exercises

III.10.8.1. (a) Show that \(*\mathbb{R}\) has no countable cofinal subset: if \((a^{(n)})\) is a sequence in \(*\mathbb{R}\), then there is a \(b \in \ast \mathbb{R}\) with \(a^{(n)} < b\) for all \(n\). [For each \(m \in \mathbb{N}\), set \(b_m = 1 + \max\{a^{(1)}_m, \ldots, a^{(m)}_m\}\).]
(b) Show that if \((a^{(n)})\) and \((b^{(n)})\) are sequences in \(*\mathbb{R}\) with \(a^{(n)} < b^{(r)}\) for all \(n\) and \(r\), there are \(c, d \in \ast \mathbb{R}\) with \(a^{(n)} < c < d < b^{(n)}\) for all \(n\). [For almost all \(m \in \mathbb{N}\), choose \(c_m, d_m \in \mathbb{R}\) with \(a^{(n)}_m < c_m < d_m < b^{(r)}_m\) for all \(n, r \leq m\).]

III.10.8.2. (a) Show that if \(F\) is any ordered field and \(a, b, c, d \in F\) with \(a < b\) and \(c < d\), the open intervals \((a, b)\) and \((c, d)\) are order-isomorphic. [Use a “linear” function.]
(b) Show that any two open intervals in \(*\mathbb{R}\) are order-isomorphic. [Use the extensions \((\cdot)\) of the tangent and arctangent functions to give an order-isomorphism between \(*\mathbb{R}\) and \((-\frac{\pi}{2}, \frac{\pi}{2})\).]

III.10.8.3. Let \(n \in \ast \mathbb{N}\setminus \mathbb{N}\). Show that \(n - 1 \in \ast \mathbb{N}\setminus \mathbb{N}\). Show that the intersection \(\gamma(n)\) of the galaxy of \(n\) with \(\ast \mathbb{N}\) is a copy of \(\mathbb{Z}\) (consisting of \(\{n + k : k \in \mathbb{Z}\}\)). Show that \(\ast \mathbb{N}\) consists of \(\mathbb{N}\) and uncountably many disjoint such \(\gamma(n)\), and between any two such \(\gamma(n)\), or between \(\mathbb{N}\) and any \(\gamma(n)\), there are uncountably many \(\gamma(k)\). In particular, \(\ast \mathbb{N}\) is uncountable. Give similar results for \(\ast \mathbb{Z}\) and \(\ast \mathbb{Q}\). [Argue as in III.10.8.1.]

III.10.8.4. If \(x, a \in \ast \mathbb{N}\), then the nonstandard exponential \(x^a\) is in \(\ast \mathbb{N}\).

III.10.8.5. Let \(n = [[(n_m)]] \in \ast \mathbb{N}\). Either \(\{m \in \mathbb{N} : n_m\text{ is odd}\}\) or \(\{m \in \mathbb{N} : n_m\text{ is even}\}\) is in \(\mathcal{U}\). Say \(n\) is odd or even in the two cases. Show that if \(n \in \ast \mathbb{N}\), then either \(n\) or \(n + 1\) is even, and \(n\) is even if and only if \(n/2 \in \ast \mathbb{N}\).

III.10.8.6. If \(k, n \in \ast \mathbb{N}\), say \(k\) divides \(n\) (written \(k|n\)) if \(n/k \in \ast \mathbb{N}\).
(a) If \(k = [[(k_m)]]\) and \(n = [[(n_m)]]\) with \(k_m, n_m \in \mathbb{N}\) for almost all \(m\), then \(k|n\) if and only if \(k_m|n_m\) for almost all \(m\).
(b) If \(k, n \in \ast \mathbb{N}\), show that \(k\) and \(n\) have a unique greatest common divisor in \(\ast \mathbb{N}\), an element \(d \in \ast \mathbb{N}\) such that \(d|k, d|n\), and for any \(c \in \ast \mathbb{N}\) such that \(c|k\) and \(c|n\), we have that \(c|d\). Similarly show that \(k\) and \(n\) have a least common multiple in \(\ast \mathbb{N}\).

III.10.8.7. Let \(n = [[(n_m)]] \in \ast \mathbb{N}\). Show that \(n\) is divisible by every \(k \in \mathbb{N}\).

III.10.8.8. Let \(p \in \ast \mathbb{N}\). Say \(p\) is prime if whenever \(p = kn\) with \(k, n \in \ast \mathbb{N}\), then either \(k\) or \(n\) is 1.
(a) If \(p = [[(p_m)]]\) with \(p_m \in \mathbb{N}\) for almost all \(m\), then \(p\) is prime in \(\ast \mathbb{N}\) if and only if \(p_m\) is prime in \(\mathbb{N}\) for almost all \(m\).
(b) If \(\Pi\) is the set of primes in \(\mathbb{N}\), then \(\ast \Pi\) is the set of primes in \(\ast \mathbb{N}\).
(c) If \(p\) is prime in \(\ast \mathbb{N}\), and \(k, n \in \ast \mathbb{N}\), then \(p|kn\) if and only if \(p|k\) or \(p|n\).
(d) If \( p \) is prime in \( {}^*\mathbb{N} \), and \( n \in {}^*\mathbb{N} \) with \( p \mid n \), then there is a unique \( r \in {}^*\mathbb{N} \) for which \( n = p^r k \) for some \( k \in {}^*\mathbb{N} \) such that \( p \) does not divide \( k \).

e Let \( n_1, n_2 \in {}^*\mathbb{N} \). Show that \( n_1 = n_2 \) if and only if, for every prime \( p \) in \( {}^*\mathbb{N} \), when we write \( n_1 = p^r k_1 \) and \( n_2 = p^s k_2 \) as in (d), we have \( r = s \).

(f) Can every \( n \in {}^*\mathbb{N} \) be written as a (finite) product of prime powers? [See III.10.8.5..]

**III.10.8.9.** Let \( T \) be the order topology () on \( {}^*\mathbb{R} \), i.e. the topology generated by the open intervals.

(a) Show that the halo and the galaxy of any element are both open and closed in \( {}^*\mathbb{R} \).

(b) Show that the relative topology from \( T \) on \( \mathbb{R} \) is the discrete topology.

(c) Show that \( ({}^*\mathbb{R}, T) \) is totally disconnected. [Use III.10.8.2..]

(d) Show that no point of \( {}^*\mathbb{R} \) has a countable local base for \( T \). [Use III.10.8.1..]

**III.10.8.10.** Let \( X \) be an uncountable set. Mimic the construction of \( {}^*\mathbb{R} \) using an ultrafilter on \( X \) containing all subsets of \( X \) with cardinality less than \( |X| \). Show that ordered fields containing \( \mathbb{R} \) of arbitrarily large cardinality can be constructed. (See [GJ76, 12E].) Are such fields a satisfactory model for the hyperreal numbers for the purpose of nonstandard analysis? See Exercise III.10.8.14.

**III.10.8.11.** (a) Repeat the construction of the hyperreal numbers using \( {}^*\mathbb{R} \) in place of \( \mathbb{R} \): let \( U \) be a free ultrafilter on \( \mathbb{N} \), and \( {}^*\mathbb{S} \) the set of sequences of elements of \( {}^*\mathbb{R} \). Define an equivalence relation on \( {}^*\mathbb{S} \) via \( \mathcal{F} \) as before, and let \( **\mathbb{R} \) be the set of equivalence classes. Show that \( **\mathbb{R} \) is an ordered field containing \( {}^*\mathbb{R} \). Is it essentially larger? Is it isomorphic to \( {}^*\mathbb{R} \)?

(b) Repeat the construction transfinitely, taking a “union” (direct limit) at limit ordinals, to construct ordered fields of arbitrary cardinality containing \( \mathbb{R} \). Is there a transfer principle for these fields?

**III.10.8.12.** [Dav77] A Nonstandard Construction of \( \mathbb{R} \). Construct the nonstandard rational numbers \( {}^*\mathbb{Q} \). Let \( L \) be the set of all limited elements of \( {}^*\mathbb{Q} \), and \( I \) the set of infinitesimals in \( {}^*\mathbb{Q} \).

(a) Show that \( L \) is a commutative unital ring, and that \( I \) is a maximal ideal in \( L \). Thus \( R = L/I \) is a field.

(b) Show that the ordering on \( {}^*\mathbb{Q} \) induces an ordering on \( R \) making it an ordered field. If \( \mathbb{Q} \) is regarded as a subring of \( L \) in the usual way, then the image of \( \mathbb{Q} \) in \( R \) is order-dense.

(c) If \( A \) and \( B \) are nonempty subsets of \( {}^*\mathbb{Q} \) with \( a < b \) for all \( a \in A \), \( b \in B \), then there is a \( c \in {}^*\mathbb{Q} \) with \( a < c < b \) for all \( a \in A \), \( b \in B \). [Use the argument of III.10.8.1.(b).]

(d) Conclude that \( R \) is a complete ordered field. [If \( A \) is a nonempty subset of \( R \) with no largest element which is bounded above, let \( B \) be the set of upper bounds for \( A \). Consider the inverse images in \( L \), and apply (c).]

This is a very efficient construction of \( \mathbb{R} \), but unlike the Dedekind cut or Cauchy sequence constructions it requires some form of AC (existence of a free ultrafilter on \( \mathbb{N} \)).

**III.10.8.13.** (a) Show that (assuming AC) using a countable index set (e.g. \( \mathbb{N} \)) in the construction of \( {}^*\mathbb{R} \) is not sufficient to obtain concurrence (); an index set \( I \) of cardinality at least \( 2^\aleph_0 \) is needed. [Consider a well ordering of \( \mathbb{R} \) according to the smallest ordinal of cardinality \( 2^\aleph_0 \), which is a concurrent relation on \( \mathbb{R} \). Every element of \( {}^*\mathbb{R} \) is determined by a function from \( I \) to \( \mathbb{R} \), hence by at most \( |I| \) elements of \( \mathbb{R} \).]
(b) Show similarly that (under AC) to obtain concurrence in a nonstandard model for a structure of cardinality \( \kappa \), an index set of cardinality at least \( \kappa \) is necessary. (Even then the ultrafilter needs to be carefully chosen.)

**III.10.8.14.**  
(a) Show that if an ultrafilter on some index set \( I \) is used for the construction of \( *\mathbb{R} \) which is not closed under countable intersections (such an ultrafilter is necessarily free), we obtain \( *\mathbb{R} \neq \mathbb{R} \) so that \( *\mathbb{R} \) is a true nonstandard model for \( \mathbb{R} \). [Show there is a sequence \( (V_n) \) of subsets of \( I \) with \( \cup_{n=1}^{\infty} V_n = I \) and \( V_n \notin \mathcal{U} \) for every \( n \). Use this sequence to construct an element of \( *\mathbb{N} \setminus \mathbb{N} \).]  
(b) Conversely, if an ultrafilter is used which is closed under countable intersections (this can only happen if the ultrafilter is fixed or if \( I \) is huge, of cardinality at least as large as a measurable cardinal), then \( *\mathbb{R} = \mathbb{R} \) so the construction is degenerate and not suitable for nonstandard analysis. [First show that if \( U_1, U_2, \ldots \) is a sequence of subsets of \( I \) and \( \cup_{n=1}^{\infty} U_n \in \mathcal{U} \), then some \( U_n \in \mathcal{U} \). Then show that \( *\mathbb{N} = \mathbb{N} \) and that \( *\mathbb{R} \) has no unlimited elements, and apply III.10.1.4.]

**III.10.8.15. Theorem.** The following theorems are logically equivalent:

(i) Tikhonov’s Theorem for Hausdorff spaces.

(ii) Tikhonov’s Theorem for copies of \([0,1]\).


**Proof:** (i) \( \Rightarrow \) (ii) and (iv) \( \Rightarrow \) (iii) are trivial. (ii) \( \Rightarrow \) (iv) is the standard construction of the Stone-Čech compactification \( \beta X \). So we need only prove (iii) \( \Rightarrow \) (i).

Let \( \{X_i : i \in I\} \) be an indexed collection of compact Hausdorff spaces, and let \( X = \prod_i X_i \) with the product topology. Let \( Y \) be the set \( X \) with the discrete topology, and \( \beta Y \) its Stone-Čech compactification. For each \( i \), the coordinate map \( \pi_i : X \to X_i \), regarded as a map from \( Y \) to \( X_i \), extends uniquely to a continuous function \( f_i \) from \( \beta Y \) to \( X_i \), by the universal property of \( \beta Y \). The \( f_i \) define a continuous function from \( \beta Y \) to \( X \) \( \), which is surjective since its restriction to \( Y \) is the identity map from \( Y \) to \( X \). Thus \( X \) is a continuous image of the compact space \( \beta Y \), hence compact by \( \beta \).

Note that (ii) \( \Rightarrow \) (i) is a simple consequence of the fact that every compact Hausdorff space can be embedded in a product of intervals, but we did not need to use this argument.

It is a direct consequence of (iii) that every filter on a set \( X \) is contained in an ultrafilter. Thus this statement, along with the ones from the theorem, is logically strictly weaker than the Axiom of Choice.

369
III.11. Linear Algebra

Linear algebra is one of the most important and fundamental parts of elementary (university-level) mathematics. Although linear algebra is not directly part of the subject of this book, it is an indispensable tool in analysis, particularly vector analysis and functional analysis, as well as throughout mathematics in general. We will not give a full treatment of the subject here (there are any number of good texts available, as well as some that are not so good), but simply review the parts of the subject most relevant to analysis, often from the point of view of the analysis applications.

The linear algebra done in most texts is finite-dimensional linear algebra. While finite-dimensional linear algebra is important in analysis, especially vector analysis, much of advanced analysis uses infinite-dimensional vector spaces, so we will place more emphasis than is usual in linear algebra on aspects of the subject relevant in the infinite-dimensional case.

Abstract linear algebra is done over an arbitrary (but fixed) field $\mathbb{F}$. In 75% of the analysis applications (at least in this book), $\mathbb{F}$ will be the real numbers $\mathbb{R}$, and in 99% of the applications in which $\mathbb{F}$ is not $\mathbb{R}$ it is the complex numbers $\mathbb{C}$. The reader may just take the symbol $\mathbb{F}$ to stand for either $\mathbb{R}$ or $\mathbb{C}$ without essential loss of generality for analysis purposes. (In the remaining applications, $\mathbb{F}$ could be the rational numbers $\mathbb{Q}$, the hyperreal numbers $^*\mathbb{R}$ (III.10.2.21.), a field of rational functions (III.4.3.5.(c)) or Laurent series (III.5.3.3.), the $p$-adic numbers (III.5.3.9.), or occasionally something else.)

III.11.1. Matrix Algebra

Matrices can be regarded simply as an extremely useful shorthand notation for certain complicated algebraic systems. They are so useful, however, that they are really more than just a shorthand, becoming important mathematical objects in their own right.

III.11.1.1. Definition. Let $m, n \in \mathbb{N}$. An $m$-by-$n$ ($m \times n$) matrix over $\mathbb{F}$ is an array

$$A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}$$

of elements $\{a_{ij}\} \in \mathbb{F}$, with $m$ rows and $n$ columns. The element $a_{ij}$ in the $i$'th row and $j$'th column is called the $(i, j)$ entry of the matrix $A$, sometimes written $A_{ij}$. $A$ is sometimes denoted $[a_{ij}]$ or $[a_{ij}]_{mn}$.

Actually, most of the theory of matrices works equally well over any ring $R$, i.e. the $a_{ij}$ are all elements of $R$. For some of the theory it is necessary to restrict to a commutative ring (we will note these instances).

Addition of Matrices

III.11.1.2. Only matrices of the same size can be added. Addition is done entrywise:

$$\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix} + \begin{bmatrix}
b_{11} & b_{12} & \cdots & b_{1n} \\
b_{21} & b_{22} & \cdots & b_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
b_{m1} & b_{m2} & \cdots & b_{mn}
\end{bmatrix} = \begin{bmatrix}
a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\
a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn}
\end{bmatrix}$$

The sum is another matrix of the same size.
III.11.1.3. Example. Let
\[ A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 1 & -2 \end{bmatrix}, B = \begin{bmatrix} -2 & 1 & 0 \\ 5 & -3 & 3 \end{bmatrix}, C = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}. \]

Then
\[ A + B = \begin{bmatrix} -1 & -1 & 4 \\ 5 & -2 & 1 \end{bmatrix} \]

and \( A + C \) and \( B + C \) are not defined since \( A \) and \( B \) are \( 2 \times 3 \) and \( C \) is \( 2 \times 2 \).

The next result is obvious:

III.11.1.4. Proposition. Let \( m, n \in \mathbb{N} \). Addition of \( m \times n \) matrices over \( \mathbb{F} \) is associative and commutative. There is an \( m \times n \) zero matrix \( 0 = 0_{m \times n} \) such that \( 0 + A = A + 0 = A \) for every \( m \times n \) matrix \( A \). For every such \( A \) there is a unique \( m \times n \) matrix \( -A \) such that \( A + (-A) = 0 \).

In fact, \( 0_{m \times n} \) is the \( m \times n \) matrix whose entries are all 0, and if \( A = [a_{ij}] \), then \( -A = [-a_{ij}] \).

III.11.1.5. We can also define scalar multiplication of matrices. If \( A = [a_{ij}] \) is an \( m \times n \) matrix over \( \mathbb{F} \) and \( r \in \mathbb{F} \), then define \( rA = [ra_{ij}] \). Scalar multiplication can be regarded as a special case of matrix multiplication (), and thus satisfies all the usual algebraic rules () (this is easily verified directly). We have \( -A = (-1)A \) for any \( A \).

Row and Column Vectors

III.11.1.6. A \( 1 \times n \) matrix over \( \mathbb{F} \) is of the form
\[ \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \]
and is usually called a row \( n \)-vector. The set of row \( n \)-vectors can be naturally identified with \( \mathbb{F}^n \).

An \( m \times 1 \) matrix over \( \mathbb{F} \) is of the form
\[ \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} \]
and is usually called a column \( m \)-vector. The set of column \( m \)-vectors can be naturally identified with \( \mathbb{F}^m \).

Under both identifications, matrix addition and scalar multiplication corresponds with usual vector addition and scalar multiplication.
Matrix Multiplication

III.11.1.7. The normal way to multiply matrices is more complicated and less obvious. Coordinatewise multiplication of matrices of the same size turns out not to be the useful product for most purposes (although this multiplication, called Schur product, does have some limited applications). The most useful product is hard to motivate until some further linear algebra is developed, so we will just give the definition and then see why it is important and has nice properties.

Multiplication of matrices is not necessarily defined for matrices of the same size. Instead, if $A$ is an $m \times n$ matrix and $B$ is a $q \times p$ matrix, then $AB$ is defined if and only if $n = q$, and the product is then an $m \times p$ matrix. Here is the formal definition:

III.11.1.8. Definition. Let $A = [a_{ij}]$ be an $m \times n$ matrix and $B = [b_{ij}]$ a $q \times p$ matrix. Then the matrix product $AB$ is defined to be the $m \times p$ matrix $C = [c_{ij}]$, where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

if $n = q$, and $AB$ is undefined if $n \neq q$.

The idea is that to get the $(i,j)$ entry of $AB$, the $i$'th row of $A$ is “multiplied” times the $j$'th column of $B$ (in a manner strongly reminiscent of the dot product in $\mathbb{R}^n$; in fact, this “multiplication” is exactly the dot product of the $i$’th row vector of $A$ and the $j$’th column vector of $B$ if $F = \mathbb{R}$.) The product of the matrices is then defined if and only if the row vectors of $A$ and the column vectors of $B$ have the same length.

If $A$ and $B$ are square matrices of the same size (i.e. $n \times n$ for some $n$), then $AB$ is defined and is again an $n \times n$ matrix.

III.11.1.9. Examples. Let

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 1 & -2 \end{bmatrix}, B = \begin{bmatrix} -2 & 1 \\ 0 & 5 \\ -3 & 3 \end{bmatrix}, C = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}.$$  

Then

$$AB = \begin{bmatrix} -14 & 3 \\ 6 & -1 \end{bmatrix}, BA = \begin{bmatrix} -2 & 5 & -10 \\ 0 & 5 & -10 \\ -3 & 9 & -6 \end{bmatrix}$$

$$CA = \begin{bmatrix} 2 & -1 & 2 \\ -1 & 4 & 0 \end{bmatrix}, BC = \begin{bmatrix} -5 & -4 \\ -5 & 10 \\ -9 & -3 \end{bmatrix}, C^2 = CC = \begin{bmatrix} 1 & 12 \\ -4 & 1 \end{bmatrix}$$

and $AC$, $CB$, $A^2$, and $B^2$ are undefined.

III.11.1.10. It is obvious from this example that $AB$ and $BA$ are usually not the same. In fact, one of these products can be defined and the other undefined, and even if both are defined they need not even be the same size. Even if $A$ and $B$ are square matrices of the same size, so $AB$ and $BA$ are both defined and the same size, they need not be equal (in almost any example one tries randomly they will be different). Thus matrix multiplication is noncommutative in a very fundamental sense.

It may then be surprising that the other usual rules of algebra hold for matrix multiplication:

372
III.11.1.11. **Proposition.** Matrix multiplication, when defined, is associative: if $A, B, C$ are matrices and $(AB)C$ is defined, then $A(BC)$ is also defined and $A(BC) = (AB)C$.

**Proof:** This is a straightforward calculation. The products are defined if and only if $A$ is $m \times n$, $B$ is $n \times p$, and $C$ is $p \times q$ for some $m, n, p, q \in \mathbb{N}$. In this case, for $1 \leq i \leq m$, $1 \leq j \leq q$ we have

$$[(AB)C]_{ij} = \sum_{r=1}^{p} (AB)_{ir}C_{rj} = \sum_{r=1}^{p} \left( \sum_{k=1}^{n} A_{ik}B_{kr} \right)C_{rj} = \sum_{r=1}^{p} \sum_{k=1}^{n} A_{ik}B_{kr}C_{rj}$$

$$= \sum_{k=1}^{n} \sum_{r=1}^{p} A_{ik}B_{kr}C_{rj} = \sum_{k=1}^{n} A_{ik} \left( \sum_{r=1}^{p} B_{kr}C_{rj} \right) = \sum_{k=1}^{n} A_{ik}(BC)_{kj} = [A(BC)]_{ij}.$$

III.11.1.12. **Proposition.** The distributive laws hold for matrix addition and multiplication:

(i) If $A, B, C$ are matrices with $B$ and $C$ the same size, and $AB$ (and hence also $AC$) is defined, then $A(B+C)$ is defined and $A(B+C) = AB + AC$.

(ii) If $A, B, C$ are matrices with $A$ and $B$ the same size, and $AC$ (and hence also $BC$) is defined, then $(A+B)C$ is defined and $(A+B)C = AC + BC$.

The proof is a straightforward calculation. Note that both distributive laws must be stated and proved because of the noncommutativity of matrix multiplication.

**The Identity Matrix**

III.11.1.13. For $n \in \mathbb{N}$, define the *identity matrix of size $n$* to be the (square) $n \times n$ matrix

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

i.e. the $(i, j)$ entry is 1 if $i = j$ and 0 if $i \neq j$.

A simple calculation shows that the identity matrices are “multiplicative identities”:

III.11.1.14. **Proposition.** Let $A$ be an $m \times n$ matrix. Then $I_mA = AI_n = A$.

**Rings of Square Matrices**

III.11.1.15. If $n \in \mathbb{N}$, the set $M_n(\mathbb{F})$ of $n \times n$ matrices over $\mathbb{F}$ is closed under addition and multiplication, and the ring axioms () are satisfied. Since $I_n \in M_n(\mathbb{F})$, $M_n(\mathbb{F})$ is a unital ring under this addition and multiplication; this ring is noncommutative if $n > 1$. More generally, if $R$ is any ring, then the set $M_n(R)$ of $n \times n$ matrices over $R$ is a ring, which is unital if $R$ is unital, and noncommutative if $R$ is unital and $n > 1$.
**Invertible Matrices**

**III.11.1.16.** If $A$ is an $m \times n$ matrix and $B$ is $n \times m$, and $AB = I_m$, then $A$ is a left inverse for $B$ and $B$ is a right inverse for $A$, and $A$ and $B$ are right and left invertible respectively. It turns out that this is possible only if $n \geq m$. If $n = m$, left and right inverses, if they exist, are unique and are actually two-sided inverses, i.e. if $A$ and $B$ are $n \times n$ matrices and $AB = I_n$, then automatically $BA = I_n$ also. (These results are valid only for matrices over commutative rings.) A square matrix $A$ with a (necessarily two-sided) inverse is called invertible, or nonsingular, and its inverse is denoted $A^{-1}$.

**III.11.1.17.** **Examples.** The identity matrix $I_n$ is invertible, with $I_n^{-1} = I_n$. The zero matrix is obviously not invertible.

Not every nonzero square matrix is invertible: a simple calculation shows, for example, that

\[
\begin{bmatrix}
1 & 0 \\
0 & 0
\end{bmatrix}
\]

is not invertible.

We will later give criteria for when a square matrix is invertible.

**Transpose**

**III.11.1.18.** If $A$ is an $m \times n$ matrix, the transpose of $A$, denoted $A^t$, is the $n \times m$ matrix obtained by interchanging the rows and columns of $A$, i.e. if $A = [a_{ij}]_{mn}$, then $A^t = [a_{ji}]_{nm}$. The transpose of a row vector is a column vector, and conversely. If $A$ is any matrix, then $[A^t]^t = A$. If $A$ and $B$ are matrices of the same size, then $(A + B)^t = A^t + B^t$.

Transpose reverses multiplication:

**III.11.1.19.** **Proposition.** Let $A$ be $m \times n$ and $B$ $n \times p$. Then $(AB)^t = B^t A^t$.

The proof is a simple calculation left to the reader. (This result is only valid over a commutative ring.)

Since $I_n^t = I_n$, we have:

**III.11.1.20.** **Corollary.** If $A$ is an $n \times n$ matrix, then $A^t$ is invertible if and only if $A$ is invertible, and $(A^t)^{-1} = (A^{-1})^t$.

**Multiplication by Blocks**

**III.11.1.21.** A very useful symbolic procedure for multiplying matrices, frequently overlooked in linear algebra texts, is by decomposition into blocks.

Let $m, n, m_1, m_2, n_1, n_2 \in \mathbb{N}$ with $m = m_1 + m_2, n = n_1 + n_2$. If $A$ is an $m \times n$ matrix over $\mathbb{F}$, then $A$ can be decomposed or partitioned symbolically as

\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}
\]
where $A_{ij}$ is an $m_i \times n_j$ matrix ($A_{11}$ and $A_{12}$ comprise the first $m_1$ rows of $A$, etc.) Sometimes we draw dotted lines in an array to denote a partition of the matrix. (We have described this in the case where the rows and columns are each partitioned into two blocks, but any finite numbers of blocks, including one, can be used independently for the rows and columns.)

The proof of the next key proposition is a straightforward but ugly calculation along the lines of the proof of III.11.1.11.

**III.11.1.22. Proposition.** Let $m, n, p, m_1, m_2, n_1, n_2, p_1, p_2 \in \mathbb{N}$ with $m = m_1 + m_2$, $n = n_1 + n_2$, $p = p_1 + p_2$. Let $A$ be an $m \times n$ matrix and $B$ an $n \times p$ matrix, and set $C = AB$. Partition $A, B, C$ according to the $m_i, n_i, p_i$. Then for any $i$ and $j$,

$$C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j}$$

($C_{ij}$ is an $m_i \times p_j$ matrix, and so are the two terms on the right). A similar result holds for partitions into other numbers of blocks (provided only that the columns of $A$ and the rows of $B$ are partitioned in the same way).

Symbolically, $AB$ can be computed by multiplying

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

according to the usual rule for multiplying two $2 \times 2$ matrices, i.e. the matrix product can be “computed by blocks.”

If the matrices are square and all blocks are the same size, as a special case we obtain:

**III.11.1.23. Corollary.** If $m, n \in \mathbb{N}$, then $M_m(M_n(\mathbb{F}))$ is naturally isomorphic to $M_{mn}(\mathbb{F})$.

This result holds over any ring. For fields (or commutative rings), it can also be proved by tensor products.

The technique of multiplication by blocks is especially useful if one of the matrices has a rectangular block of zeroes. Here is a sample application, which will be used in the proof of the Implicit Function Theorem of vector calculus:

**III.11.1.24. Proposition.** Let $n, n_1, n_2 \in \mathbb{N}$ with $n = n_1 + n_2$, $A$ an $n_1 \times n_1$ matrix, and $B$ an $n_1 \times n_2$ matrix. Then the $n \times n$ matrix

$$C = \begin{bmatrix} A & B \\ 0 & I \end{bmatrix}$$

(where the 0 is an $n_2 \times n_1$ block and $I$ is $n_2 \times n_2$ identity matrix) is invertible if and only if $A$ is invertible, and in this case

$$C^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}B \\ 0 & I \end{bmatrix}.$$
Proof: It is easily checked by multiplication by blocks that if \( A \) is invertible, then the formula gives an
inverse for \( C \). Conversely, suppose \( C \) is invertible and
\[
C^{-1} = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}
\]
with blocks of the same size as in \( C \). We then have
\[
\begin{bmatrix} A & B \\ 0 & I \end{bmatrix} \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} AX + BZ & AY + BW \\ Z & W \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}
\]
and comparing entries, we get \( Z = 0, W = I, X = A^{-1}, Y = -A^{-1}B \).

Systems of Linear Equations

As an example of the utility of using matrix notation, we briefly review the theory of systems of linear
equations.

III.11.1.25. A system of \( m \) linear equations in \( n \) unknowns over \( F \) is of the form
\[
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\]
where the \( a_{ij} \) and the \( b_j \) are given elements of \( F \), and \( x_1, \ldots, x_n \) are unknowns. The system is homogeneous
if all the \( b_j \) are 0. A solution is, of course, a set of values for the \( x_i \) in \( F \) satisfying the equations.

III.11.1.26. This system can be conveniently abbreviated in matrix form. If
\[
A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}
\]
we can then compactly represent the system as the matrix equation
\[
Ax = b
\]
where \( A \) is a given \( m \times n \) matrix, \( b \) a given column \( m \)-vector, and \( x \) an unknown column \( n \)-vector. This
clean formulation alone gives some justification for why matrix multiplication as we have defined it is useful.
But this is only the beginning of the utility.
III.11.1.27. We would expect to have a unique solution to the system if and only if \( m = n \); if \( m < n \) the system is “underdetermined” and we would expect many solutions, and if \( m > n \) we would expect that the system is “overdetermined” and that there is no solution in general. These answers are “usually” correct, but can fail if there are degeneracies in the system.

In the case \( m = n \), if the matrix \( A \) is invertible, the matrix form gives a way to find a solution (which turns out to be the unique solution for any \( b \)):

\[
x = A^{-1}b.
\]

We will examine the general theory after developing some linear algebra machinery ().

III.11.2. Vector Spaces
A vector space is simply an abstraction of Euclidean space \( \mathbb{R}^n \) with its essential algebraic structure. The two basic algebraic operations in \( \mathbb{R}^n \) are vector addition and scalar multiplication. Other algebraic operations such as the dot product will not be used at this point, but will be incorporated later.

Let \( \mathbb{F} \) be a field (usually \( \mathbb{R} \) or \( \mathbb{C} \) in analysis applications), called the field of scalars, which is fixed throughout the discussion. A vector space over \( \mathbb{F} \) will be a set equipped with an addition and a scalar multiplication satisfying a list of natural algebraic properties. Specifically:

III.11.2.1. Definition. A vector space over \( \mathbb{F} \) is a set \( V \) with two binary operations \( + : V \times V \to V \) (written \( (x, y) = x + y \)) and \( \cdot : \mathbb{F} \times V \to V \) (written \( (\alpha, x) = \alpha \cdot x \) or usually just \( \alpha x \)) satisfying:

- **(A1)** The binary operation \( + \) is associative: \( x + (y + z) = (x + y) + z \) for all \( x, y, z \in V \).
- **(A2)** The binary operation \( + \) is commutative: \( x + y = y + x \) for all \( x, y \in V \).
- **(A3)** There is an element \( 0 \in V \) such that \( 0 + x = x \) for all \( x \in V \).
- **(A4)** For every \( x \in V \) there is an element \( -x \in V \) with \( x + (-x) = 0 \).
- **(M1)** The binary operation \( \cdot \) is associative: \( \alpha (\beta x) = (\alpha \beta) x \) for all \( \alpha, \beta \in \mathbb{F}, x \in V \).
- **(M2)** The binary operation \( \cdot \) is distributive: \( \alpha (x + y) = \alpha x + \alpha y \) and \( (\alpha + \beta) x = \alpha x + \beta x \) for all \( \alpha, \beta \in \mathbb{F}, x, y \in V \).
- **(M3)** \( 1 \cdot x = x \) for all \( x \in V \).

If \( R \) is a ring, a set with two operations satisfying these axioms is called a left \( R \)-module. Thus a vector space is just a left module over a field.

III.11.2.2. Examples. (i) If \( n \in \mathbb{N} \), then the set \( \mathbb{F}^n \) of \( n \)-tuples of elements of \( \mathbb{F} \), with coordinatewise addition and scalar multiplication, is a vector space over \( \mathbb{F} \). This is the prototype example. It will turn out that any finite-dimensional vector space over \( \mathbb{F} \) is essentially identical to this example for a unique \( n \).

(ii) Let \( m, n \in \mathbb{N} \). The set \( M_{m,n}(\mathbb{F}) \) of \( m \times n \) matrices over \( \mathbb{F} \), with matrix addition and scalar multiplication, is a vector space over \( \mathbb{F} \). This vector space has a natural identification with \( \mathbb{F}^{mn} \) by writing the entries of the matrix in a single string in a fixed order.
(iii) Here is a general example including (i) and (ii) as a special case: let $X$ be a set and $V$ the set of all functions from $X$ to $\mathbb{F}$. Define addition and scalar multiplication pointwise: if $f, g \in V$ and $\alpha \in \mathbb{F}$, set $(f+g)(x) = f(x)+g(x)$ and $(\alpha f)(x) = \alpha f(x)$ (where the operations on the right-hand sides are the operations in $\mathbb{F}$). Example (i) is the case where $X = \{1, \ldots, n\}$ and example (ii) is the case $X = \{1, \ldots, m\} \times \{1, \ldots, n\}$.

(iv) Let $P$ be the set of all polynomials with coefficients in $\mathbb{F}$, expressions of the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where the $a_k \in \mathbb{F}$ and the $n$ may vary from polynomial to polynomial. Add polynomials formally in the usual way by collecting together terms with the same power of $x$, and multiply by scalars termwise. Then $P$ is a vector space over $\mathbb{F}$.

(v) Let $[a, b]$ be an interval in $\mathbb{R}$, and $C([a, b])$ the set of continuous functions from $[a, b]$ to $\mathbb{R}$. Add functions and multiply by scalars pointwise in the usual way, as in (iii) (a theorem is needed to show that $C([a, b])$ is closed under these operations). Then $C([a, b])$ is a vector space over $\mathbb{R}$.

Example (v) has many variations which are important in analysis.

III.11.2.3. In some of these examples, there is a natural multiplication defined on pairs of elements of the vector space which is compatible (bilinear) with scalar multiplication. This multiplication is not part of the vector space structure. A vector space over $\mathbb{F}$ with a compatible multiplication is called an algebra over $\mathbb{F}$. $M_n(\mathbb{F})$ and $P$ are typical examples of algebras over $\mathbb{F}$. $C([a, b])$ is an algebra over $\mathbb{R}$.

III.11.2.4. There are two important points which must be emphasized about the definition of a vector space. First, the operations of addition and scalar multiplication are part of the definition and must be specified; thus it is technically incorrect to make a statement such as “Let $V$ be a vector space over $\mathbb{F}$.” But we do this all the time when the operations in $V$ are understood, such as in these examples or variations of them.

Secondly, the field of scalars must be specified; it is incorrect to say “Let $V$ be a vector space.” (We often do this too, but the field of scalars must be understood.) There is a variation of this point which must be kept in mind: if $V$ is a vector space over $\mathbb{C}$, then since $\mathbb{R} \subseteq \mathbb{C}$ we can restrict scalar multiplication to $\mathbb{R}$ and obtain a real vector space (vector space over $\mathbb{R}$). Some references thus state:

“A complex vector space is a real vector space.”

From a logical standpoint, however, this statement is meaningless: real vector spaces and complex vector spaces are distinct mathematical structures and it does not make sense to mix them. A more careful and logically correct way of phrasing this statement is:

“A complex vector space has a unique underlying structure as a real vector space.”

The situation is similar to the corresponding relation between metric spaces and topological spaces (I). The same considerations apply to any other field and subfield (e.g. $\mathbb{Q} \subseteq \mathbb{R}$).
Elements of a vector space are typically called points or vectors. This terminology does not carry with it any necessary geometric interpretation. In fact, general vector spaces do not have any natural geometric interpretation. But the quasi-geometric language used is well established, and can be useful in motivating results, although care must be exercised in proofs not to use anything beyond what simply follows from the axioms.

The standard algebraic properties similar to those of \( \mathbb{R} \) are consequences of the axioms. Here are some of the most important ones:

**Proposition.** Let \( V \) be a vector space over \( F \). Then

(i) \( 0x = 0 \) for all \( x \in V \).

(ii) If \( \alpha \in F \) and \( x \in V \), then \( \alpha x = 0 \) if and only if \( \alpha = 0 \) or \( x = 0 \).

(iii) If \( x, y, z \in V \) and \( x + z = y + z \), then \( x = y \).

(iv) If \( \alpha \in F, \alpha \neq 0, \) and \( x, y \in V, \) and \( \alpha x = \alpha y, \) then \( x = y \).

(v) \( (-1)x = -x \) for all \( x \in V \).

**Isomorphism of Vector Spaces**

As is usual in mathematics, we need a precise specification of when two vector spaces are abstractly the “same.” The obvious criterion is:

**Definition.** Let \( V \) and \( W \) be vector spaces over \( F \). A function \( T : V \to W \) is an isomorphism (of vector spaces over \( F \)) if \( T \) is a bijection which is linear, i.e. respects addition and scalar multiplication:

\[
T(x + y) = T(x) + T(y)
\]

\[
T(\alpha x) = \alpha T(x)
\]

for all \( x, y \in V, \alpha \in F \). \( V \) and \( W \) are isomorphic (as vector spaces over \( F \)) if there is an isomorphism from \( V \) to \( W \).

**Definition.** Let \( V \) and \( W \) be vector spaces over \( F \). A subset \( W \) of \( V \) is a vector subspace or linear subspace (often just called a subspace) if \( W \) is itself a vector space under the same operations of \( V \).

If \( W \) is a vector subspace of \( V \), the definition entails that \( W \) is closed under addition and scalar multiplication, i.e. if \( x, y \in W \) and \( \alpha \in F \), then \( x + y \) and \( \alpha x \) are in \( W \). \( W \) is also nonempty since \( 0 \in W \). These are sufficient:
III.11.2.10. Proposition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W} \subseteq \mathcal{V}$. Then $\mathcal{W}$ is a vector subspace of $\mathcal{V}$ if and only if $\mathcal{W}$ is nonempty and closed under addition and scalar multiplication.

Proof: Necessity has been observed. Conversely, suppose $\mathcal{W}$ is nonempty and closed under addition and scalar multiplication. Axioms (A1), (A2), and (M1)–(M3) are automatic in $\mathcal{W}$ since they hold in $\mathcal{V}$. If $x \in \mathcal{W}$, then $0x = 0 \in \mathcal{W}$, so (A3) holds. And if $x \in \mathcal{W}$, then $-x = (-1)x \in \mathcal{W}$, so (A4) holds.

III.11.2.11. Examples. (i) If $\mathcal{V}$ is any vector space, then $\mathcal{V}$ and $\{0\}$ are vector subspaces of $\mathcal{V}$. These are called the trivial subspaces.

(ii) Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $x \in \mathcal{V}$. Then

$$\{\alpha x : \alpha \in \mathbb{F}\}$$

is a vector subspace of $\mathcal{V}$. If $x \neq 0$, this is called the one-dimensional subspace of $\mathcal{V}$ spanned by $x$. This example will be generalized in III.11.2.14.

(iii) Let $\mathcal{V}$ be the set of all functions from $[a,b]$ to $\mathbb{R}$. Then $\mathcal{V}$ is a real vector space as in III.11.2.2 (iii). Then $C([a,b])$ (III.11.2.2.(v)) is a vector subspace of $\mathcal{V}$.

Span of a Subset, Linear Combinations

III.11.2.12. It is obvious that an arbitrary intersection of vector subspaces of a vector space $\mathcal{V}$ over $\mathbb{F}$ is a vector subspace (the intersection is nonempty since every subspace contains $0$). Thus, if $S$ is any subset of $\mathcal{V}$, the intersection of all subspaces of $\mathcal{V}$ containing $S$ is a subspace of $\mathcal{V}$ which is the smallest subspace of $\mathcal{V}$ containing $S$. (Note that there is at least one subspace of $\mathcal{V}$ containing $S$, $\mathcal{V}$ itself.) This smallest subspace is called the span of $S$, denoted $span(S)$. We say $S$ spans $\mathcal{V}$ if $span(S) = \mathcal{V}$.

There is a concrete description of $span(S)$ using the notion of linear combination:

III.11.2.13. Definition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $x_1, \ldots, x_n \in \mathcal{V}$. A linear combination ($\mathbb{F}$-linear combination) of $\{x_1, \ldots, x_n\}$ is a vector of the form

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

where the $\alpha_k$ are in $\mathbb{F}$.

If $S$ is an arbitrary nonempty subset of $\mathcal{V}$, then a linear combination of $S$ is a vector which is a linear combination of a set of vectors in $S$. The set of linear combinations of $\emptyset$ is defined to be $\{0\}$.

Note that even if $S$ is infinite, a linear combination of $S$ is a linear combination of a finite subset of $S$, i.e. an individual linear combination involves only finitely many elements of $S$.

III.11.2.14. Proposition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $S \subseteq \mathcal{V}$. Then $span(S)$ is precisely the set $Lin(S)$ of linear combinations of $S$.

Proof: Any vector in $S$ is trivially in $Lin(S)$, and every linear combination of $S$ is in any vector subspace containing $S$. Thus $S \subseteq Lin(S) \subseteq span(S)$. So it suffices to note that $Lin(S)$ is a vector subspace of $\mathcal{V}$. It is obvious that $0 \in Lin(S)$, and that a sum or scalar multiple of linear combinations of $S$ is again a linear combination of $S$. 

$\Box$
III.11.2.15. Proposition. Let $V$ be a vector space over $F$, and $S \subseteq V$. If $x_0 \in S$ and $x_0$ can be written as a linear combination of other vectors in $S$, then $\text{span}(S \setminus \{x_0\}) = \text{span}(S)$.

Proof: Suppose $x_0 = \alpha_1 x_1 + \cdots + \alpha_n x_n$ with $\alpha_1, \ldots, \alpha_n \in F$ and $x_1, \ldots, x_n \in S \setminus \{x_0\}$. Let $x \in \text{span}(S)$. Then $x = \beta_1 y_1 + \cdots + \beta_m y_m$ for some $y_1, \ldots, y_m \in S$ and $\beta_1, \ldots, \beta_m \in F$. By collecting together terms, we may assume all the $y_i$ are distinct. If none of the $y_i$ are equal to $x_0$, then $x$ is already written as a linear combination of $S \setminus \{x_0\}$. If $y_i = x_0$ for some $i$, by reordering we may assume $i = 1$. Then

$$x = \beta_1 (\alpha_1 x_1 + \cdots + \alpha_n x_n) + \beta_2 y_2 + \cdots + \beta_m y_m$$

and $x$ is again in $\text{span}(S \setminus \{x_0\})$.

III.11.2.16. Definition. Let $V$ be a vector space over $F$. $V$ is finite-dimensional if there is a finite subset of $V$ which spans $V$.

III.11.2.17. Examples. (i) $F^n$ is spanned by $\{e_1, \ldots, e_n\}$, where $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in the $k$’th place. (This set of vectors is called the standard basis of $F^n$). Thus $F^n$ is finite-dimensional.

(ii) The vector space $\mathcal{P}$ of all polynomials over $F$ (III.11.2.2.(iv)) is not finite-dimensional, since a finite set of polynomials has a maximal degree and no linear combination of the finite set of polynomials can have greater degree. The countable set $\{1, x, x^2, \ldots\}$ does span $\mathcal{P}$.

It is a little harder to show that Example III.11.2.2.(v), and III.11.2.2.(iii) for infinite $X$, are not finite-dimensional (this follows from III.11.2.30.(iv) in both cases; cf. III.11.2.21.(iv)).

Linear Independence

III.11.2.18. Definition. Let $V$ be a vector space over $F$, and $S$ a subset of $V$. Then $S$ is linearly independent (over $F$) if no vector in $S$ is a linear combination of the other vectors in $S$.

In other words, $S$ is linearly independent if it has no “redundancy,” i.e. no proper subset of $S$ has the same span as $S$: every vector in $S$ makes an essential contribution to the span. Since the zero vector is in the span of $\emptyset$, no linearly independent set can contain the zero vector.

The empty set is vacuously linearly independent.

A convenient rephrasing is:

III.11.2.19. Proposition. Let $V$ be a vector space over $F$, and $S \subseteq V$. The following are equivalent:

(i) $S$ is linearly independent.

(ii) Whenever $x_1, \ldots, x_n \in S$ and $\alpha_1, \ldots, \alpha_n \in F$, and

$$\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$$

then $\alpha_1 = \cdots = \alpha_n = 0$. 381
(iii) Every nonzero vector in \( \text{span}(S) \) can be written as a linear combination of vectors in \( S \) with nonzero coefficients in a unique way.

The uniqueness in (iii) must be expressed carefully; additional terms with coefficients 0 may be added to the sum without changing the vector represented.

**Proof:** (i) \( \Rightarrow \) (ii): Suppose \( x_1, \ldots, x_n \in S \), \( \alpha_1, \ldots, \alpha_n \in \mathbb{F} \), and \( \alpha_1 x_1 + \cdots + \alpha_n x_n = 0 \) with not all \( \alpha_k = 0 \). By reordering, we may assume \( \alpha_1 \neq 0 \). Then we have

\[
x_1 = -\frac{\alpha_2}{\alpha_1} x_2 - \cdots - \frac{\alpha_n}{\alpha_1} x_n
\]

so \( S \) is not linearly independent.

(ii) \( \Rightarrow \) (iii): Suppose, for some \( x \in \text{span}(S) \),

\[
x = \alpha_1 x_1 + \cdots + \alpha_n x_n = \beta_1 x_1 + \cdots + \beta_n x_n
\]

with \( \alpha_k \neq \beta_k \) for some \( k \). Then

\[
(\alpha_1 - \beta_1) x_1 + \cdots + (\alpha_n - \beta_n) x_n = 0
\]

with \( \alpha_k - \beta_k \neq 0 \), so (ii) does not hold.

(iii) \( \Rightarrow \) (i): If \( x \in S \) were a linear combination \( x = 1x = \alpha_1 x_1 + \cdots + \alpha_n x_n \) with \( x_k \neq x \) for all \( k \), then \( x \) would be a linear combination of vectors in \( S \) in a nonunique way.

\( \blacklozenge \)

**III.11.2.20.** A sum \( \alpha_1 x_1 + \cdots + \alpha_n x_n = 0 \) as in (ii) with \( x_1, \ldots, x_n \in S \) and \( \alpha_1, \ldots, \alpha_n \in \mathbb{F} \) not all zero is called a linear dependence (or linear relation) in \( S \) (over \( \mathbb{F} \)), and a set with a linear dependence (i.e. a set which is not linearly independent) is called linearly dependent.

**III.11.2.21.** Examples. (i) A set \( \{x\} \) of one vector in a vector space \( V \) is linearly independent if and only if \( x \neq 0 \).

(ii) If \( \{x, y\} \) is a subset of \( V \) with \( x \neq y \), then \( \{x, y\} \) is linearly independent in \( V \) if and only if \( x \) is nonzero and \( y \) is not a scalar multiple of \( x \). (The condition is symmetric in \( x \) and \( y \).)

(iii) It is harder to tell by inspection whether a set of three or more vectors is linearly independent, since the set can be linearly dependent even if no vector in the set is a scalar multiple of another. For example, it is easily seen that

\[
\{(1,0), (0,1), (1,1)\}
\]

is linearly dependent in \( \mathbb{R}^2 \) since the third vector is the sum of the first two.

(iv) In Example III.11.2.2, (iv) or (v), let \( f_n(x) = x^n \). It is easy to verify that \( \{f_n : n \in \mathbb{N}\} \) is linearly independent in each case. In case (iv) this set spans the vector space, and in (v) it does not.
Bases

III.11.2.22. **DEFINITION.** Let $V$ be a vector space over $F$. A *basis* for $V$ (over $F$) is a linearly independent set of vectors in $V$ which spans $V$.

A basis is a set of vectors which is “just large enough” to span $V$ yet small enough to contain no redundancies. If $B$ is a basis for $V$, then no smaller subset of $V$ spans $V$ and no larger subset of $V$ is linearly independent, i.e. $B$ is a maximal linearly independent subset of $V$.

The plural of *basis* is *bases*. The plural of *base* is spelled the same way but pronounced differently.

III.11.2.23. Bases are especially important in finite-dimensional vector spaces, where they can be regarded as giving a “coordinate system.” They are of less crucial importance in infinite-dimensional vector spaces. Certain infinite-dimensional topological vector spaces have other basis-like sets (where “linear combinations” involve infinite series) which are a more useful substitute. In infinite-dimensional vector spaces a basis in the linear algebra sense (i.e. in the sense of III.11.2.22.) is often called a *Hamel basis* to distinguish the concept from the other one.

III.11.2.24. **PROPOSITION.** Let $V$ be a vector space over $F$. A subset $B$ of $V$ is a basis for $V$ if and only if $B$ is a maximal linearly independent subset.

**Proof:** It has already been observed that a basis is a maximal linearly independent subset. On the other hand, suppose $B$ is a maximal linearly independent set in $V$, and $x \in V$. If $x \notin \text{span}(B)$, then $S \cup \{x\}$ is linearly independent, contradicting maximality of $S$. Thus $\text{span}(B) = V$ and $B$ is a basis for $V$.

III.11.2.25. **EXAMPLES.** (i) Let $V = F^n$ as a vector space over $F$. Then the set of standard basis vectors $\{e_1, \ldots, e_n\}$ is a basis for $F^n$: if $x = (\alpha_1, \ldots, \alpha_n) \in F^n$, then

$$(\alpha_1, \ldots, \alpha_n) = \alpha_1 e_1 + \cdots + \alpha_n e_n$$

is the unique expression of $x$ as a linear combination of $\{e_1, \ldots, e_n\}$.

(ii) One can give completely different bases for $F^n$. For example, a simple calculation shows that

$$\begin{align*}
x_1 &= (1, -1, 0, \ldots, 0) \\
x_2 &= (0, 1, -1, \ldots, 0) \\
&\vdots \\
x_{n-1} &= (0, 0, \ldots, 1, -1) \\
x_n &= (-1, 0, 0, \ldots, 0, 1)
\end{align*}$$

forms a basis for $F^n$. (There are also *many* other bases, especially if $F$ is $\mathbb{R}$ or $\mathbb{C}$.)

(iii) The set $\{f_n : n \in \mathbb{N}\}$ (III.11.2.21.(iv)) is a basis for the vector space $P$ over $\mathbb{R}$ (III.11.2.2.(iv)).

383
\textbf{III.11.2.26.} Examples (i) and (ii) show that there is nothing unique about bases for a vector space; two bases may have no vectors in common. But it turns out (III.11.2.28.(iv)) that the number of vectors in any two bases for a vector space $V$ is always the same; this number is called the \textit{dimension} of $V$.

\textbf{III.11.2.27.} Elaborating further on Example III.11.2.25., if $V$ is a vector space over $\mathbb{F}$ with basis $B = \{x_1, \ldots, x_n\}$, then the (ordered) basis $B$ induces a natural one-one correspondence with $\mathbb{F}^n$:

$$(\alpha_1, \ldots, \alpha_n) \leftrightarrow \alpha_1 x_1 + \cdots + \alpha_n x_n$$

which respects vector addition and scalar multiplication (a \textit{vector space isomorphism}). Thus an ordered basis for $V$ can be regarded as a “coordinate system” in $V$. Indeed, a finite-dimensional abstract vector space over $\mathbb{F}$ can just be regarded as a copy of $\mathbb{F}^n$ for some $n$ with the “natural” coordinate system stripped away; choice of a basis specifies a coordinate system, with any such coordinate system as good as any other for most linear algebra purposes (if a linear operator is being considered, some coordinate systems may be more compatible with the operator than others; this will be discussed later).

\textbf{Fundamental Theorems of Linear Algebra}

It is not obvious that every vector space has a basis (in fact, in general, the Axiom of Choice is needed to prove this). The following blockbuster theorem shows that the situation is as nice as could be hoped for:

\textbf{III.11.2.28. \textsc{Theorem.}} Let $V$ be a vector space over $\mathbb{F}$. Then

(i) $V$ has a basis.

(ii) Every linearly independent subset of $V$ can be expanded to a basis for $V$.

(iii) Every spanning subset of $V$ contains a basis for $V$.

(iv) Any two bases for $V$ have the same cardinality (if $V$ is finite-dimensional, any two bases for $V$ have the same finite number of elements).

This theorem is elementary to prove if $V$ is finite-dimensional, but more delicate (requiring the AC) in the general case. We give the finite-dimensional proof here and defer the general proof until III.11.2.32. \textsc{Proof:} (Finite-dimensional case.) Clearly (ii) $\Rightarrow$ (i) (begin with $\emptyset$). To (partly) prove (ii), let $\{y_1, \ldots, y_m\}$ be a finite set which spans $V$, and let $\{x_1, \ldots, x_r\}$ be a finite linearly independent set in $V$. Set

$$S_1 = \{x_1, \ldots, x_r, y_1, \ldots, y_m\}$$

regarded as an ordered set. $S_1$ spans $V$, and is linearly dependent since each $x_i$ can be written as a linear combination of the $y_j$. Working from left to right, let $z_1$ be the first vector which can be written as a linear combination of the preceding ones. There must be such a $z_1$, and it must be one of the $y_j$ since the $x_i$ are linearly independent. Set $S_2 = S_1 \setminus \{z_1\}$. Then $\text{span}(S_2) = \text{span}(S_1) = V$ by III.11.2.15.. If $S_2$ is linearly independent, it is a basis containing $\{x_1, \ldots, x_r\}$; otherwise, repeat the process, letting $z_2$ be the first vector in the ordered set $S_2$ which can be written as a linear combination of the preceding vectors, and $S_3 = S_2 \setminus \{z_2\}$. Since $S_1$ is finite, we must reach a linearly independent set in a finite number of steps; this
set will be a basis for $V$ consisting of $x_1, \ldots, x_r$ and (in general) some, but not all, of the $y_j$. This argument proves (ii) for a finite linearly independent set, and thus the special case (i).

The argument for (iv) is similar. We will show a stronger statement: if $\{x_1, \ldots, y_n\}$ is a basis for $V$, and $\{x_1, \ldots, x_r\}$ is a linearly independent set in $V$, then $r \leq n$. This will show (iv) by symmetry, and also complete the proof of (ii) since it implies that any linearly independent set in $V$ is finite. Suppose $r > n$.

Set $S_1 = \{x_1, y_1, \ldots, y_n\}$, regarded as an ordered set. This set spans $V$ and is linearly dependent. Let $z_1$ be the first vector which can be written as a linear combination of the previous ones; then $z_1$ must be one of the $y_j$. Set

$$S_2 = \{x_2, x_1, y_1, \ldots, y_n\}$$

with $z_1 = y_j$ removed. Then $\text{span}(S_2) = \text{span}(S_1) = V$, and $S_2$ is linearly dependent since $x_2 \in \text{span}(S_1 \setminus \{z_1\}) = V$. Repeat the process. At each stage another $x_k$ is added and a $y_j$ is removed; thus after $n$ steps we have $S_n \setminus \{z_n\} = \{x_n, \ldots, x_1\}$, and this set spans $V$, contradicting that $x_{n+1}$ is not a linear combination of $\{x_1, \ldots, x_n\}$. Thus $r \leq n$.

Finally, we show (iii). Let $S$ be a set which spans $V$, and $\{y_1, \ldots, y_n\}$ a basis for $V$. Each $y_j$ can be written as a linear combination of a finite number of vectors in $S$; hence there is a finite subset of $S$ whose span includes all the $y_j$, hence is $V$. Thus we may assume $S$ is finite. List the elements of $S$ and successively eliminate any vector which can be written as a linear combination of the preceding ones. Each reduced set will still span $V$ by III.11.2.15. In a finite number of steps a linearly independent set will be reached, which is a basis for $V$.

\[\square\]

III.11.2.29. Definition. Let $V$ be a vector space over $F$. The (linear) dimension of $V$, denoted $\text{dim}(V)$, is the cardinality of a basis of $V$.

The dimension of $V$ is well defined by III.11.2.28.(iv), and is finite (a natural number) if and only if $V$ is finite-dimensional in the sense of III.11.2.16. We sometimes write $\text{dim}_F(V)$ to emphasize the field we are working over (cf. III.11.2.33).

We note the following corollary of the finite-dimensional version of III.11.2.28.

III.11.2.30. Corollary. Let $V$ be a finite-dimensional vector space over $F$, with $\text{dim}(V) = n$. Then

(i) If $\{x_1, \ldots, x_r\}$ is a linearly independent set of vectors in $V$, then $r \leq n$.

(ii) If $S = \{x_1, \ldots, x_n\}$ is a linearly independent set of vectors in $V$, then $S$ spans $V$ (i.e. is a basis for $V$).

(iii) If $\{x_1, \ldots, x_r\}$ spans $V$, then $r \geq n$.

(iv) If $S = \{x_1, \ldots, x_n\}$ is a set of vectors in $V$ which spans $V$, then $S$ is linearly independent (i.e. is a basis for $V$).

(v) If $W$ is a vector subspace of $V$, then $\text{dim}(W) \leq \text{dim}(V)$; $\text{dim}(W) = \text{dim}(V)$ if and only if $W = V$. In particular, any subspace of $V$ is finite-dimensional.

Proof: (i)–(iv) are obvious from III.11.2.28. For (v), suppose $\text{dim}(W) > n$ (or $W$ is not finite-dimensional). Then (without any form of AC) we can choose a set $S = \{x_1, \ldots, x_{n+1}\}$ in $W$ which is linearly independent [let
$x_1$ be any nonzero vector in $W$, and successively choose $x_{k+1}$ to be a vector not in the span of $\{x_1, \ldots, x_k\}$. Then $S$ is a linearly independent subset of $V$, contradicting (i). If $\dim(W) = n$, apply (ii) to a basis for $W$. 

**III.11.2.31. Theorem.** Two vector spaces over $\mathbb{F}$ are isomorphic if and only if they have the same dimension. In particular, any $n$-dimensional vector space over $\mathbb{F}$ is isomorphic to $\mathbb{F}^n$.

**Proof:** Let $V$ and $W$ be vector spaces over $\mathbb{F}$ of the same dimension, and let $B$ and $C$ be bases for $V$ and $W$ respectively. Let $\psi : B \to C$ be a bijection. Extend $\psi$ to a bijection $\phi : V \to W$ as follows. If $x \in V$, then $x$ can be written uniquely as

$$x = \alpha_1 x_1 + \cdots + \alpha_n x_n$$

with $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, all nonzero, and $x_1, \ldots, x_n \in B$. Define

$$\phi(x) = \alpha_1 \psi(x_1) + \cdots + \alpha_n \psi(x_n).$$

Then $\phi$ is well defined and one-to-one since $C$ is linearly independent. If $y \in W$, then

$$y = \alpha_1 y_1 + \cdots + \alpha_m y_m$$

for some $\alpha_1, \ldots, \alpha_m \in \mathbb{F}$ and $y_1, \ldots, y_m \in C$; if $y_j = \psi(x_j)$ for each $j$, then $y = \phi(x)$, where

$$x = \alpha_1 x_1 + \cdots + \alpha_m x_m.$$ 

So $\phi$ is surjective. It is routine to check that $\phi$ is linear.

**III.11.2.32.** We now prove the general version of Theorem III.11.2.28. This proof requires the Axiom of Choice (in fact, (i)–(iii) may each be equivalent to AC; (iv) only needs a weaker version of Choice).

**Proof:** We again begin by proving (ii), which clearly implies (i) as before. Let $S$ be a linearly independent subset of $V$. Let $\mathcal{X}$ be the collection of all linearly independent subsets of $V$ containing $S$, ordered by inclusion. If $\{X_i : i \in I\}$ is a chain in $\mathcal{X}$, then $X = \cup_i X_i$ is a linearly independent subset of $V$, since a linear dependence involves only finitely many vectors and any finite subset of $X$ is contained in some $X_i$. Thus $X \in \mathcal{X}$, and $X$ is an upper bound for $\{X_i : i \in I\}$. By Zorn’s Lemma, $\mathcal{X}$ contains a maximal element $B$. As in III.11.2.24., $B$ is a basis for $V$ containing $S$.

The proof of (iii) is similar. Let $S$ be a set which spans $V$. Let $\mathcal{X}$ be the collection of all linearly independent subsets of $S$, ordered by inclusion. As before, Zorn’s Lemma applies to give a maximal element $B$ of $\mathcal{X}$. As in the proof of III.11.2.24., if $\text{span}(B)$ is not all of $V$, it cannot contain all of $S$ since $S$ spans $V$; if $x \in S \setminus \text{span}(B)$, then $B \cup \{x\}$ is linearly independent, contradicting maximality of $B$. Thus $B$ is a basis for $V$.

The proof of (iv) is more complicated. Suppose $\{x_i : i \in I\}$ and $\{y_j : j \in J\}$ are bases for $V$. We may assume $I$ and $J$ are infinite since the finite case has already been proved. Let $\kappa = \text{card}(I)$ and $\lambda = \text{card}(J)$. Each $x_i$ is a linear combination of finitely many $y_j$, and every $y_j$ occurs in one of these finite sets, or otherwise all the $x_i$ would be contained in the proper subspace of $V$ spanned by a proper subset of the $y_j$. Thus $\{y_j : j \in J\}$ is a union of $\kappa$ finite subsets, so $\lambda \leq 8^n \kappa = \kappa$ since $\kappa$ is infinite (II.9.7.4.). Symmetrically, $\kappa \leq \lambda$. Thus $\kappa = \lambda$ by the Schröder-Bernstein Theorem.
III.11.2.33. When the same set is regarded as a vector space over more than one field, the notions of span, linear independence, basis, and dimension are dependent on the choice of base field. For example, \( \mathbb{C}^n \) may be regarded as an \((n\text{-dimensional})\) vector space over \( \mathbb{C} \), or as a vector space over \( \mathbb{R} \) by restriction of scalar multiplication. The standard basis \( \{e_1, \ldots, e_n\} \) is indeed a basis for \( \mathbb{C}^n \) over \( \mathbb{C} \). This set is linearly independent over \( \mathbb{R} \), but not a basis for \( \mathbb{C}^n \) over \( \mathbb{R} \); in fact, its real span is the real vector subspace \( \mathbb{R}^n \) of \( \mathbb{C}^n \). On the other hand, the set 
\[
\{e_1, \ldots, e_n, i e_1, \ldots, i e_n\}
\]
is linearly independent over \( \mathbb{R} \) but not over \( \mathbb{C} \). It is easily checked that this set is a basis for \( \mathbb{C}^n \) over \( \mathbb{R} \), and thus \( \dim_{\mathbb{R}}(\mathbb{C}^n) = 2n \).

Complementary Subspaces

III.11.2.34. Definition. Let \( \mathcal{V} \) be a vector space over \( \mathbb{F} \). Two subspaces \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \) of \( \mathcal{V} \) are complementary if \( \mathcal{W}_1 \cap \mathcal{W}_2 = \{0\} \) and \( \text{span}(\mathcal{W}_1 \cup \mathcal{W}_2) = \mathcal{V} \). \( \mathcal{W}_2 \) is complementary to \( \mathcal{W}_1 \) and vice versa.

This definition is symmetric in \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \). But we often talk about subspaces complementary to a given subspace.

III.11.2.35. There is nothing remotely unique about complementary subspaces to a given subspace. For example, if \( \mathcal{W} \) is any one-dimensional subspace of \( \mathbb{R}^2 \), then any one-dimensional subspace of \( \mathbb{R}^2 \) not equal to \( \mathcal{W} \) is complementary to \( \mathcal{W} \).

III.11.2.36. Proposition. Let \( \mathcal{V} \) be a vector space over \( \mathbb{F} \), and \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \) subspaces of \( \mathcal{V} \). Then \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \) are complementary subspaces of \( \mathcal{V} \) if and only if every vector \( x \in \mathcal{V} \) can be uniquely written as \( x = x_1 + x_2 \), with \( x_j \in \mathcal{W}_j \).

Proof: If \( x \in \text{span}(\mathcal{W}_1 \cup \mathcal{W}_2) \), then \( x \) can be written as
\[
x = \alpha_1 y_1 + \cdots + \alpha_n y_n + \beta_1 z_1 + \cdots + \beta_m z_m
\]
with \( y_1, \ldots, y_n \in \mathcal{W}_1 \) and \( z_1, \ldots, z_m \in \mathcal{W}_2 \). If \( x_1 = \alpha_1 y_1 + \cdots + \alpha_n y_n \) and \( x_2 = \beta_1 z_1 + \cdots + \beta_m z_m \), then \( x_j \in \mathcal{W}_j \) and \( x = x_1 + x_2 \). Thus every vector in \( \mathcal{V} \) can be so written if and only if \( \text{span}(\mathcal{W}_1 \cup \mathcal{W}_2) = \mathcal{V} \). Suppose \( \mathcal{W}_1 \cap \mathcal{W}_2 = \{0\} \), and \( x = x_1 + x_2 = y_1 + y_2 \) with \( x_j, y_j \in \mathcal{W}_j \). Then \( x_1 - y_1 = x_2 - y_2 \); the left side is in \( \mathcal{W}_1 \) and the right side is in \( \mathcal{W}_2 \), so both sides are in \( \mathcal{W}_1 \cap \mathcal{W}_2 \). Hence both sides are zero and the representation of \( x \) is unique. Conversely, if \( x \in \mathcal{W}_1 \cap \mathcal{W}_2 \), then \( x = x + 0 = 0 + x \) gives a nonunique decomposition unless \( x = 0 \).

III.11.2.37. Corollary. Let \( \mathcal{V} \) be a vector space over \( \mathbb{F} \), and \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \) be subspaces of \( \mathcal{V} \), \( B_j \) a basis for \( \mathcal{W}_j \), and \( B = B_1 \cup B_2 \). Then

(i) \( \mathcal{W}_1 \cap \mathcal{W}_2 = \{0\} \) if and only if \( B \) is linearly independent.

(ii) \( \mathcal{W}_1 \cup \mathcal{W}_2 \) spans \( \mathcal{V} \) if and only if \( B \) spans \( \mathcal{V} \).

(iii) \( \mathcal{W}_1 \) and \( \mathcal{W}_2 \) are complementary if and only if \( B \) is a basis for \( \mathcal{V} \).

387
In particular, if $W_1$ and $W_2$ are complementary, then $\dim(W_1) + \dim(W_2) = \dim(V)$.

This suggests a method of proof of the following theorem, which is the main result on this topic:

**III.11.2.38. Theorem.** Let $V$ be a vector space over $F$, and $W$ a subspace of $V$. Then there is a subspace of $V$ complementary to $W$.

**Proof:** This is an easy consequence of Theorem III.11.2.28. (thus it is elementary if $V$ is finite-dimensional, and requires AC in general). Let $S$ be a basis for $W$, and $B$ a basis for $V$ containing $S$. Then the span of $B \setminus S$ is a subspace of $V$ complementary to $W$, as is easily checked.

There is an alternate direct proof of this theorem using Zorn's Lemma (Exercise ()).

**III.11.2.39.** From a theoretical point of view, this theorem is a fundamental fact; it is the essential feature distinguishing the theory of vector spaces from the theory of modules over rings. In fact, it is precisely the failure of this result for modules over a ring which is not a field that makes module theory considerably more complicated (and more interesting!) than the theory of vector spaces.

Two ways to rephrase this result are:

- Every subspace of a vector space is a direct summand.
- Every short exact sequence of vector spaces splits.

See books on abstract algebra such as [DF04] for details.

**Quotient Spaces**

There is one remaining topic about vector spaces, which is somewhat more complicated conceptually than the results so far, and which is not usually treated in linear algebra texts: quotient spaces. This construction is used frequently in functional analysis.

**III.11.2.40.** Let $V$ be a vector space over $F$, and $W$ a subspace of $V$. Define a relation $\sim$ on $V$ by $x \sim y$ if $x - y \in W$.

**III.11.2.41. Proposition.**

(i) If $x, y, z \in V$, $x \sim y$, and $y \sim z$, then $x \sim z$ ($\sim$ is transitive).

(ii) If $x_1, x_2, y_1, y_2 \in V$, $x_1 \sim y_1$, and $x_2 \sim y_2$, then $x_1 + x_2 \sim y_1 + y_2$.

(iii) If $x, y \in V$, $\alpha \in F$, and $x \sim y$, then $\alpha x \sim \alpha y$. 

388
Proof: (i): we have $x - y \in W$ and $y - z \in W$, so
\[ x - z = (x - y) + (y - z) \in W. \]
(ii) is similar: $x_1 - y_1 \in W$ and $x_2 - y_2 \in W$, so
\[ (x_1 + x_2) - (y_1 + y_2) = (x_1 - y_1) + (x_2 - y_2) \in W. \]
The proof of (iii) is almost identical.

III.11.2.42. By (i), $\sim$ is an equivalence relation () since it is obviously reflexive and symmetric. Thus $V$ partitions into a disjoint union of equivalence classes (called cosets of $W$ in $V$). Denote by $V/W$ the set of equivalence classes, and for $x \in V$ let $[x]$ denote the equivalence class $\{y \in V : y \sim x\}$ of $X$.

By (ii), addition of equivalence classes by $[x] + [y] = [x + y]$ is well defined, and by (iii) scalar multiplication by $\alpha [x] = [\alpha x]$ is also well defined.

III.11.2.43. Proposition. With this addition and scalar multiplication, $V/W$ is a vector space over $F$.

Proof: To prove associativity of addition, if $x, y, z \in V$, we have
\[ ([x] + [y]) + [z] = [x + y] + [z] = [(x + y) + z] = [x + (y + z)] = [x] + [y + z] = [x] + ( [y] + [z]) . \]
The proofs of the other properties are similar.

III.11.2.44. Definition. The vector space $V/W$ is called the quotient vector space of $V$ by $W$. The dimension of $V/W$ is called the codimension of $W$ in $V$.

III.11.2.45. Examples. (i) Let $\mathbb{F} = \mathbb{R}, V = \mathbb{R}^2$, and $W$ the one-dimensional space spanned by $e_2 = (0, 1)$. If $x = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, then the equivalence class of $x$ is
\[ [x] = \{(\alpha_1, \beta) : \beta \in \mathbb{R}\} . \]
Thus the equivalence classes are parametrized by their first coordinates, and there is a natural one-one correspondence of $\mathbb{R}^2/W$ with $\mathbb{R}^1$; this correspondence respects addition and scalar multiplication, so $\mathbb{R}^2/W \cong \mathbb{R}^1$.

(ii) Here is a more typical example of how quotient spaces are used in analysis. Let $X$ be an infinite set, and $V$ the vector space of all functions from $X$ to $\mathbb{F}$ (III.11.2.2. (iii)). Let $W$ be the subspace of all functions which are 0 except on a finite subset of $X$ (which may vary from function to function). It is easy to check that $W$ is a subspace of $V$. If $f, g \in V$, then $f \sim g$ if and only if the functions $f$ and $g$ agree “almost everywhere” (except on a finite subset of $X$), i.e.
\[ \{x \in X : f(x) \neq g(x)\} \]
is a finite subset of $X$. We can form the quotient vector space $V/W$, whose elements are equivalence classes of functions which agree “almost everywhere”; we often by slight abuse of terminology regard elements of $V/W$ as functions from $X$ to $\mathbb{F}$, with functions agreeing almost everywhere identified together.
III.11.2.46. There is a systematization of the identification in (i) using complementary subspaces which is sometimes useful. Let $V$ be a vector space and $W$ a subspace of $V$. Let $W'$ be a complementary subspace to $W$. For $x \in W'$, define $\phi(x) = [x] \in V/W$.

III.11.2.47. **Proposition.** The map $\phi: W' \to V/W$ is an isomorphism.

**Proof:** If $x \in V$, then $x$ can be written as $y + z$, with $y \in W$ and $z \in W'$. Then $x - z = y \in W$, so $[x] = [z] = \phi(z)$ and $\phi$ is surjective. If $x, y \in W'$ and $\phi(x) = [x] = [y] = \phi(y)$, then $x - y \in W$; since $x - y \in W'$, we have $x - y \in W \cap W' = \{0\}$, so $x = y$ and $\phi$ is injective. The fact that $\phi$ is linear follows immediately from the way addition and scalar multiplication in $V/W$ are defined. 

III.11.2.48. **Corollary.** If $V$ is finite-dimensional and $W$ is a subspace of $V$, then $V/W$ is finite-dimensional and $\dim(V/W) = \dim(V) - \dim(W)$. Thus the dimension of $W$ plus the codimension of $W$ in $V$ equals the dimension of $V$.

III.11.2.49. Even if $V$ and $W$ are infinite-dimensional, the codimension of $W$ in $V$ is well defined. Thus the last statement of III.11.2.48. is true in general, and we obtain the nonobvious general fact that any two subspaces of $V$ complementary to $W$ have the same dimension.

III.11.2.50. The identification in III.11.2.47. is important in principle, but it does not help much in understanding an example like III.11.2.45.(ii), since in this case there is no obvious explicitly describable subspace of $V$ which is complementary to $W$ (in fact, in general the AC must be employed to obtain a complement in this example).

**Summary**

III.11.3. **Linear Transformations**

It follows from the results of the last section that the theory of vector spaces over a field $F$ (especially finite-dimensional ones) is quite simple and not very interesting: the only isomorphism invariant is the dimension, and there are simple standard examples of vector spaces of each dimension.

The primary subject matter of linear algebra is not vector spaces themselves, but linear transformations between them. These form a much more complicated, and interesting, theory, which we outline in this section.

III.11.3.1. **Definition.** Let $V$ and $W$ be vector spaces over $F$. A function $T : V \to W$ is a **linear transformation** (linear function, linear operator) if $T$ respects addition and scalar multiplication:

$$T(x + y) = T(x) + T(y)$$
$$T(\alpha x) = \alpha T(x)$$

for all $x, y \in V, \alpha \in F$. 

390
III.11.3.2. The terms *linear transformation*, *linear function*, and *linear operator* are synonymous and interchangeable. We will usually use the term *linear transformation*, but often the term *(linear) operator* when $V = W$. The “linear” is often left out and understood when the term operator is used; nonlinear operators are also sometimes considered, but we will always qualify such operators explicitly as “nonlinear.”

Note that the field $\mathbb{F}$ must be specified in the definition of linearity. If there is any potential ambiguity about the field we are working over, we use the more precise term $\mathbb{F}$-linear in place of linear. For example, if $V$ and $W$ are vector spaces over $\mathbb{C}$, we distinguish between real-linear and complex-linear transformations (any complex-linear transformation is real-linear, but not conversely); in this setting, “linear” will normally mean “complex-linear”.

It is often customary to eliminate the parentheses around the argument of a linear transformation, i.e. to write $Tx$ in place of $T(x)$. We will try to avoid doing this, at least at first, in the interest of clarity, to emphasize that a linear transformation is a function.

The first simple observation is:

III.11.3.3. Proposition. Let $V$ and $W$ be vector spaces over $\mathbb{F}$, and $T : V \to W$ a linear transformation. Then $T(0) = 0$.

Proof: By linearity, we have $T(0) = T(0 \cdot 0) = 0 \cdot T(0) = 0$.

This is generalized in III.11.3.6.

III.11.3.4. Examples. (i) An isomorphism of vector spaces (III.11.2.7.) is a linear transformation.
(ii) If $V$ and $W$ are any vector spaces over $\mathbb{F}$, the constant function $0$ from $V$ to $W$ is linear.
(iii) Regard $\mathbb{F}$ as a one-dimensional vector space over $\mathbb{F}$. Let $T : \mathbb{F} \to \mathbb{F}$ be linear, and set $m = T(1)$. Then, for any $x \in \mathbb{F}$, we have $T(x) = T(x \cdot 1) = x \cdot T(1) = x \cdot m = mx$. Thus $T(x) = mx$ for some fixed $m \in \mathbb{F}$. Conversely, by the associative, commutative, and distributive laws in $\mathbb{F}$, any such function is linear.
(iv) Let $\mathcal{P}$ be the real vector space of polynomials with real coefficients. Define $D : \mathcal{P} \to \mathcal{P}$ by $D(f) = f'$. Elementary theorems of calculus (),(,) show that $D$ is a linear transformation.

III.11.3.5. Example (iii) for $\mathbb{F} = \mathbb{R}$ reveals an ambiguity in standard mathematical terminology. It is customary in calculus and precalculus to call a function $f : \mathbb{R} \to \mathbb{R}$ a “linear function” if it is of the form

$$f(x) = mx + b$$

for constants $m$ and $b$. But such a function is a linear function in the linear algebra sense if and only if $b = 0$. To resolve this ambiguity, we will usually call a function of the form $f(x) = mx + b$ an affine function rather than a linear function if $b \neq 0$, reserving the term “linear” for functions which are actually linear in the sense of III.11.3.1.

Null Space and Range

Linear transformations map subspaces to subspaces in the following sense:
III.11.3.6. **Proposition.** Let $V$ and $W$ be vector spaces over $F$, and let $T : V \to W$ be a linear transformation.

(i) Let $X$ be a subspace of $V$. Set 

$$T(X) = \{ T(x) : x \in X \} .$$

Then $T(X)$ is a subspace of $W$.

(ii) Let $Y$ be a subspace of $W$. Set 

$$T^{-1}(Y) = \{ x \in V : T(x) \in Y \} .$$

Then $T^{-1}(Y)$ is a subspace of $V$.

**Proof:**

(i) $T(X)$ is nonempty since $X$ is nonempty. Let $y_1, y_2 \in T(X)$. Then $y_1 = T(x_1)$ and $y_2 = T(x_2)$ for some $x_1, x_2 \in X$. Then $x_1 + x_2 \in X$, and 

$$T(x_1 + x_2) = T(x_1) + T(x_2) = y_1 + y_2 \in T(X)$$

since $T$ is linear. A similar argument shows that $T(X)$ is closed under scalar multiplication.

(ii): Since $T(0_V) = 0_W$ and $0_W \in Y$, $0_V \in T^{-1}(Y)$ and $T^{-1}(Y) \neq \emptyset$. If $x_1, x_2 \in T^{-1}(Y)$, then $T(x_1), T(x_2) \in Y$ and thus 

$$T(x_1 + x_2) = T(x_1) + T(x_2) \in Y$$

so $x_1 + x_2 \in T^{-1}(Y)$. Similarly, $T^{-1}(Y)$ is closed under scalar multiplication.

There are two especially important subspaces associated with any linear transformation:

III.11.3.7. **Definition.** Let $V$ and $W$ be vector spaces over $F$, and let $T : V \to W$ be a linear transformation. The **range** of $T$ is 

$$\mathcal{R}(T) = T(V) = \{ T(x) : x \in V \} \subseteq W .$$

The **null space** of $T$ is 

$$\mathcal{N}(T) = T^{-1}(\{ 0 \}) = \{ x \in V : T(x) = 0 \} \subseteq V .$$

By III.11.3.6., $\mathcal{R}(T)$ and $\mathcal{N}(T)$ are subspaces of $W$ and $V$ respectively.

III.11.3.8. **Proposition.** Let $V$ and $W$ be vector spaces over $F$, and let $T : V \to W$ be a linear transformation. Then $T$ is one-to-one if and only if $\mathcal{N}(T) = \{ 0 \}$.

**Proof:** If $T$ is one-to-one, if $x \in \mathcal{N}(T)$, then $T(x) = 0 = T(0)$, so $x = 0$. Conversely, suppose $\mathcal{N}(T) = \{ 0 \}$. If $x_1, x_2 \in V$ and $T(x_1) = T(x_2)$, then $T(x_1 - x_2) = T(x_1) - T(x_2) = 0$ and $x_1 - x_2 \in \mathcal{N}(T)$, so $x_1 = x_2$ and $T$ is one-to-one.

The next result, a generalization of III.11.3.8., is one of the fundamental facts from linear algebra:
III.11.3.9. Theorem. [First Isomorphism Theorem] Let $V$ and $W$ be vector spaces over $F$, and let $T : V \rightarrow W$ be a linear transformation. Then $T$ induces a well-defined linear isomorphism $\tilde{T}$ from $V/N(T)$ onto $R(T)$ by $\tilde{T}( [x] ) = T(x)$. Proof: We have to show that $\tilde{T}$ is well-defined and one-to-one; then it is automatic that $\tilde{T}$ is linear and $R(\tilde{T}) = R(T)$. So suppose $x_1, x_2 \in V$ and $[x_1] = [x_2]$ in $V/N(T)$. Then $x_1 - x_2 \in N(T)$, so we have

$$T(x_1) = T(x_2 + (x_1 - x_2)) = T(x_2) + T(x_1 - x_2) = T(x_2) + 0 = T(x_2)$$

and $\tilde{T}$ is well defined. The argument is reversible: if $\tilde{T}( [x_1] ) = \tilde{T}( [x_2] )$, then $T(x_1) = T(x_2)$, $T(x_1 - x_2) = 0$, $x_1 - x_2 \in N(T)$, so $[x_1] = [x_2]$. Thus $\tilde{T}$ is one-to-one.

III.11.3.10. Here is an alternate argument. Let $X$ be a subspace of $V$ complementary to $N(T)$ (III.11.2.38.). Then $\tilde{T} := T|_X$ is a linear transformation from $X$ to $W$. It is easily checked that $\tilde{T}$ is one-to-one (since $X \cap N(T) = \{0\}$), and that $R(\tilde{T}) = R(T)$. $T$ is then obtained by composing $\tilde{T}$ with the natural isomorphism from $V/N(T)$ onto $X$ (III.11.2.47.).

Using III.11.2.48., we obtain:

III.11.3.11. Corollary. Let $V$ and $W$ be vector spaces over $F$, and let $T : V \rightarrow W$ be a linear transformation. Then $\dim(\text{R}(T)) = \text{codim}(\text{N}(T))$. In particular,

$$\dim(\text{N}(T)) + \dim(\text{R}(T)) = \dim(V)$$

This corollary is often called the “dimension-counting theorem” for linear transformations. It has an important consequence for finite-dimensional vector spaces (using III.11.3.8):

III.11.3.12. Corollary. Let $V$ be a finite-dimensional vector space over $F$, and let $T : V \rightarrow V$ be a linear transformation. Then $T$ is injective if and only if it is surjective.

This can be false if $V$ is infinite-dimensional ()

III.11.4. More on Matrices and Linear Equations

III.11.5. Affine Subspaces and Affine Functions

General Position

A useful variation of linear independence is general position.

III.11.5.1. Definition. Let $V$ be a vector space over $F$, and $S$ a subset of $V$. Then $S$ is in general position, or the points in $S$ are in general position, if, whenever $\{x_0, x_1, \ldots, x_m\}$ is a set of $m + 1$ points in $S$ with $m \leq \dim(V)$, then $\{x_1 - x_0, \ldots, x_m - x_0\}$ is linearly independent.

This definition is applied most often when $S$ is a finite subset of $V$, but makes sense in general. It is most interesting when the size of $S$ is at least the dimension of $V$.

The proof of the next proposition is nearly trivial, and is left to the reader.
III.11.5.2. **Proposition.** Let \( \mathcal{V} \) be an \( n \)-dimensional vector space over \( F \), and \( S \) a subset of \( \mathcal{V} \) with \( r \) elements (\( r = \infty \) is allowed). Set \( m = \min(r, n + 1) \). Then the following are equivalent:

(i) \( S \) is in general position.

(ii) Whenever \( \{x_0, \ldots, x_{m-1}\} \) is a subset of \( S \) with \( m \) distinct elements, \( \{x_1 - x_0, \ldots, x_{m-1} - x_0\} \) is linearly independent.

(iii) Whenever \( \{x_0, \ldots, x_{m-1}\} \) is a subset of \( S \) with \( m \) distinct elements, the smallest affine subspace of \( \mathcal{V} \) containing \( \{x_0, \ldots, x_{m-1}\} \) has dimension \( m - 1 \).

(iv) Whenever \( \{x_0, \ldots, x_k\} \) is a subset of \( S \) with \( k \) distinct elements (\( k \leq m - 1 \)), the smallest affine subspace of \( \mathcal{V} \) containing \( \{x_0, \ldots, x_k\} \) has dimension \( k \).

III.11.5.3. If \( F \) is a finite field and \( \mathcal{V} \) is an \( n \)-dimensional vector space over \( F \), then there is a limit to the size of a set in general position. But if \( F \) is infinite, there can be infinite sets in general position. For example, if \( \mathcal{V} \) is 1-dimensional, any set of vectors is in general position.

III.11.5.4. **Example.** For a set \( S \) to be in general position, it is not enough that there be an \( x_0 \in S \) such that, whenever \( \{x_1, \ldots, x_{m-1}\} \) is a subset of \( S \) with \( m - 1 \) distinct elements all distinct from \( x_0 \), then \( \{x_1 - x_0, \ldots, x_{m-1} - x_0\} \) is linearly independent: consider the set \( \{(0,0), (2,0), (0,2), (1,1)\} \subset \mathbb{R}^2 \).

III.11.5.5. Any subset of a set in general position is in general position. A set of at most \( n + 1 \) vectors in an \( n \)-dimensional vector space which is in general position is called **affinely independent**. If \( \{x_0, \ldots, x_n\} \) is affinely independent, then any subset of size \( k + 1 \) generates a \( k \)-dimensional affine subspace.

III.11.6. **The Dot Product on \( \mathbb{R}^n \) and Adjoints**

We now discuss a piece of algebraic structure unique to \( \mathbb{R}^n \) (there is a similar structure on \( \mathbb{C}^n \)).

One of the big algebraic things distinguishing \( \mathbb{R}^n, n > 1 \), from \( \mathbb{R} \) is that there is no way to multiply (or divide) vectors. (There is a nonobvious way to define vector multiplication in \( \mathbb{R}^n \) for \( n = 2, 4, \) or \( 8 \), cf. **III.6.2.2.** and **III.6.3.2.**, but for no other \( n > 1 \).) However, there is a simple and very important way to “multiply” two vectors and obtain a scalar, called the dot product or inner product.

III.11.6.1. **Definition.** Let \( \mathbf{x} = (x_1, \ldots, x_n) \) and \( \mathbf{y} = (y_1, \ldots, y_n) \) be vectors in \( \mathbb{R}^n \). The **dot product** of \( \mathbf{x} \) and \( \mathbf{y} \) is

\[
\mathbf{x} \cdot \mathbf{y} = (x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1y_1 + x_2y_2 + \cdots + x_ny_n = \sum_{k=1}^{n} x_ky_k.
\]

Note that \( \mathbf{x} \cdot \mathbf{y} \in \mathbb{R} \), i.e. the dot product defines a function from \( \mathbb{R}^n \times \mathbb{R}^n \) to \( \mathbb{R} \).
The dot product \( \mathbf{x} \cdot \mathbf{y} \) is sometimes written \( \langle \mathbf{x}, \mathbf{y} \rangle \) (other notations are also occasionally used). The dot product is an example of an inner product, discussed in XV.9.1.

The next proposition lists the important algebraic properties of the dot product. The results are either obvious or proved by straightforward calculations, and the proofs are left to the reader.

**Proposition.** Let \( \mathbf{x}, \mathbf{y}, \mathbf{z} \) be vectors in \( \mathbb{R}^n \). Then:

(i) \( \mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x} \).

(ii) \( \mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z} \).

(ii') \( (\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \mathbf{x} \cdot \mathbf{z} + \mathbf{y} \cdot \mathbf{z} \).

(iii) \( (\alpha \mathbf{x}) \cdot \mathbf{y} = \alpha (\mathbf{x} \cdot \mathbf{y}) \) for any \( \alpha \in \mathbb{R} \). In particular, \( 0 \cdot \mathbf{y} = 0 \).

(iii') \( \mathbf{x} \cdot (\alpha \mathbf{y}) = \alpha (\mathbf{x} \cdot \mathbf{y}) \) for any \( \alpha \in \mathbb{R} \). In particular, \( \mathbf{x} \cdot 0 = 0 \).

(iv) \( \mathbf{x} \cdot \mathbf{x} = ||\mathbf{x}||^2 \). In particular, \( \mathbf{x} \cdot \mathbf{x} \geq 0 \), and \( \mathbf{x} \cdot \mathbf{x} = 0 \) implies \( \mathbf{x} = 0 \).

(v) \( \mathbf{x} \cdot \mathbf{y} = \frac{1}{4}[||\mathbf{x} + \mathbf{y}||^2 - ||\mathbf{x} - \mathbf{y}||^2] \).

Property (i) is called *symmetry*; properties (ii)-(ii')-(iii)-(iii') are together called *bilinearity* (note that because of (i), (ii') and (iii') follow from (ii) and (iii) respectively). The last part of (iv) is called *positive definiteness*, and (v) is the *polarization identity*.

**Corollary.** Let \( T \) be a linear isometry of \( \mathbb{R}^n \) into \( \mathbb{R}^m \). If \( \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \), then

\[
(T\mathbf{x}) \cdot (T\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}.
\]

**Proof:** Apply polarization (III.11.6.3.(v)).

There is a nice geometric interpretation of the dot product. If \( \mathbf{x} \) and \( \mathbf{y} \) are nonzero vectors in \( \mathbb{R}^n \), there is a well-defined angle \( \theta \) between them if we take \( 0 \leq \theta \leq \pi \), measured in the two-dimensional subspace spanned by \( \mathbf{x} \) and \( \mathbf{y} \).

**Proposition.** Let \( \mathbf{x} \) and \( \mathbf{y} \) be nonzero vectors in \( \mathbb{R}^n \), and \( \theta \) the angle between them. Then

\[
\mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}|| ||\mathbf{y}|| \cos \theta.
\]

**Proof:** Since there is a linear isometry from \( \mathbb{R}^2 \) onto the two-dimensional subspace of \( \mathbb{R}^n \) spanned by \( \mathbf{x} \) and \( \mathbf{y} \), by III.11.6.4. we may assume \( n = 2 \). And by III.11.6.3.(iii)-(iii') we may assume \( \mathbf{x} \) and \( \mathbf{y} \) are unit vectors.

There is a rotation (isometry) \( T \) of \( \mathbb{R}^2 \) sending \( \mathbf{x} \) to \( (1,0) \). We then have \( T\mathbf{y} = (\cos \theta, \pm \sin \theta) \), so

\[
\mathbf{x} \cdot \mathbf{y} = (T\mathbf{x}) \cdot (T\mathbf{y}) = (1,0) \cdot (\cos \theta, \pm \sin \theta) = \cos \theta.
\]
III.11.6.6. **Corollary. [Cauchy's Inequality or CBS Inequality]** Let \( \mathbf{x} \) and \( \mathbf{y} \) be vectors in \( \mathbb{R}^n \). Then

\[
|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|
\]

**Proof:** If \( \mathbf{x} \) and \( \mathbf{y} \) are nonzero, let \( \theta \) be the angle between them. Then

\[
|\mathbf{x} \cdot \mathbf{y}| = \|\mathbf{x}\| \|\mathbf{y}\| |\cos \theta| \leq \|\mathbf{x}\| \|\mathbf{y}\|
\]

since \( |\cos \theta| \leq 1 \). And if \( \mathbf{x} \) or \( \mathbf{y} \) is \( 0 \), both sides are 0.

III.11.6.7. This inequality is commonly called the Cauchy-Schwarz inequality or even the Schwarz inequality. But it was first stated and proved in this form by Cauchy in 1821, sixty years before Schwarz. For a more direct proof and extensions to general inner products, and a discussion of the history of the inequality and its names, see XV.9.2.

III.11.6.8. **Proposition.** The dot product on \( \mathbb{R}^n \) is jointly continuous: if \( \mathbf{x}_k \to \mathbf{x} \) and \( \mathbf{y}_k \to \mathbf{y} \), then \( \mathbf{x}_k \cdot \mathbf{y}_k \to \mathbf{x} \cdot \mathbf{y} \).

**Proof:** This follows easily from the CBS inequality, or just the fact that the formula is a polynomial function in the entries.

III.11.7. **Determinants**

There is a number (scalar) called the determinant associated with any square matrix, which is of both theoretical and practical significance. In particular, a (square) matrix is invertible if and only if its determinant is nonzero.

Determinants have gotten somewhat of a bad rap. It is widely fashionable in linear algebra (cf. [Axl95]) to try to banish determinants from the subject. Indeed, elementary linear algebra can be efficiently and elegantly developed without reference to determinants (e.g. [Axl97]). Determinants are mostly excluded from numerical linear algebra. There is one good overriding reason for avoiding determinants wherever possible, especially in numerical work: determinants are hard to compute!

However, determinants cannot be entirely avoided, especially in applications to analysis. The reason is that the determinant has a natural and crucial geometric or measure-theoretic interpretation, and hence appears naturally in such analysis formulas as the change-of-variables formula for integrals \( (\cdot) \). (And modern numerical analysis techniques make determinants somewhat more efficient to compute than by the straightforward brute-force method.)

We will therefore review the basics of the theory of determinants from the analysis point of view. Note that the determinant (as we do it here) is strictly a finite-dimensional notion.

**The Determinant of a Matrix**

There are two approaches to defining the determinant of a matrix. From an abstract point of view, the “best” way is to prove the existence and uniqueness theorem \( (\text{III.11.7.27.}) \) from scratch and then derive a formula for the determinant from it. We will take a lower-tech approach, defining the determinant by a (fairly complicated) algebraic formula, and deriving its properties from the formula.
III.11.7.1. **Definition.** Let $A = [a_{ij}]$ be an $n \times n$ matrix over $\mathbb{F}$. The *determinant* of $A$ is the scalar

$$\text{det}(A) = \sum_{\sigma \in S_n} \text{sgn}(\sigma)a_{1\sigma(1)}a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where $S_n$ is the set of permutations (i) of $\{1, \ldots, n\}$ and $\text{sgn}(\sigma) = \pm 1$ is the sign (i) of the permutation $\sigma$.

There are $n!$ terms, each of which is a product of $n$ entries of $A$, exactly one from each row and column, with all possible such products represented.

To see how this works, we write out the determinant explicitly for small $n$:

III.11.7.2. **Examples.**

(i) $n = 1$: if $A = [a]$ is a $1 \times 1$ matrix (scalar), then $\text{det}(A) = a$ since there is only one permutation of $\{1\}$.

(ii) $n = 2$: Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. There are two permutations of $\{1, 2\}$, the identity $\iota$ and the permutation $\sigma$, where $\sigma(1) = 2$ and $\sigma(2) = 1$, and $\text{sgn}(\iota) = 1$, $\text{sgn}(\sigma) = -1$. Thus we have

$$\text{det}(A) = \text{sgn}(\iota)a_{1\iota(1)}a_{2\iota(2)} + \text{sgn}(\sigma)a_{1\sigma(1)}a_{2\sigma(2)} = a_{11}a_{22} - a_{12}a_{21}.$$

This formula is usually remembered as:

$$\text{det}\left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = ad - bc.$$

(iii) $n = 3$: We write this case out explicitly, but the formula more serves as a warning of how rapidly the determinant formula becomes complicated as $n$ increases than as a useful formula. There are 6 permutations of $\{1, 2, 3\}$. Rather than working through all the details, we simply give the end result:

$$\text{det}\left( \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

(iv) $n = 4$: This formula has 24 terms and exceeds the patience of the author to write.

We will obtain a somewhat simpler inductive method of computing determinants in (i); but computing the determinant of even a general $10 \times 10$ matrix is a computational challenge on a computer (there are $10! = 3,628,800$ terms).

III.11.7.3. The notation $|A|$ is also frequently used for $\text{det}(A)$, especially when the matrix $A$ is written out in an array. This notation can be somewhat confusing, since the determinant has nothing to do with absolute value (e.g. the determinant of a matrix with real entries can be negative). Besides, in analysis we often need to work with the absolute value of the determinant, possibly adding to the confusion. We will stick with the notation $\text{det}(A)$.

**Elementary Properties of Determinants**

The first observation is immediate from the definition of determinant:
III.11.7.4. Proposition. For any \( n \), the determinant of the identity matrix \( I_n \) is 1. More generally, the determinant of any diagonal matrix is the product of the diagonal elements.

In fact, it is only slightly harder to show that the determinant of any triangular matrix is the product of the diagonal elements.

Another immediate consequence of the definition is:

III.11.7.5. Proposition. Let \( A \) be an \( n \times n \) matrix over \( \mathbb{F} \). If \( A \) has a zero row or zero column, then \( \det(A) = 0 \).

This result has a generalization ( ).

The next result is also a simple consequence of the definition.

III.11.7.6. Proposition. Let \( A \) be an \( n \times n \) matrix over \( \mathbb{F} \), and \( r \in \mathbb{F} \). Let \( B \) be the matrix obtained from \( A \) by multiplying one row of \( A \) by \( r \) and leaving all other rows unchanged. Then

\[
\det(B) = r \cdot \det(A)
\]

Note that the \( B \) in the proposition is not a scalar multiple of \( A \). In fact, by applying the proposition successively to each row, we obtain that \( \det(rA) = r^n \cdot \det(A) \).

Proof: Comparing the formulas for \( \det(B) \) and \( \det(A) \), each term in the sum for \( \det(B) \) is obtained from the corresponding term for \( \det(A) \) by multiplying one of the factors by \( r \) and leaving the rest unchanged. Thus a factor of \( r \) can be taken out of the sum by the distributive property.

The next result is more esoteric to state, but has a similarly simple proof:

III.11.7.7. Proposition. Let \( B = [b_{ij}] \) and \( C = [c_{ij}] \) be \( n \times n \) matrices which are identical except for some row, i.e. for which there is an \( i_0 \) such that \( b_{ij} = c_{ij} \) for all \( i \neq i_0 \) and all \( j \). Let \( A = [a_{ij}] \) be the matrix with \( a_{ij} = b_{ij} = c_{ij} \) for \( i \neq i_0 \) and all \( j \), and \( a_{i_0j} = b_{i_0j} + c_{i_0j} \) for all \( j \) (i.e. the \( i_0 \)'th row of \( A \) is the sum of the \( i_0 \)'th rows of \( B \) and \( C \) and the other rows of \( A \) are identical to the other rows of \( B \) and \( C \)). Then

\[
\det(A) = \det(B) + \det(C)
\]

Proof: We have

\[
\det(A) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} = \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots (b_{i_0\sigma(i_0)} + c_{i_0\sigma(i_0)}) \cdots a_{n\sigma(n)}
\]

\[
= \sum_{\sigma \in S_n} \text{sgn}(\sigma) b_{1\sigma(1)} b_{2\sigma(2)} \cdots b_{n\sigma(n)} + \sum_{\sigma \in S_n} \text{sgn}(\sigma) c_{1\sigma(1)} c_{2\sigma(2)} \cdots c_{n\sigma(n)} = \det(B) + \det(C)
\]
The previous two results are abbreviated by saying that the determinant is *multilinear* (or, more specifically, *n-linear*) in the rows of the matrix. See () for the general theory of multilinear functions.

Next we obtain a property which is simpler to state but a little trickier to prove:

**III.11.7.9.** **Proposition.** Let $A$ be an $n \times n$ matrix over $F$, and $B$ a matrix obtained by interchanging two rows of $A$. Then

$$
\det(B) = -\det(A).
$$

**Proof:** Suppose the $i_1$’th and $i_2$’th rows of $A$ are interchanged. Let $\tau$ be the permutation of $\{1, \ldots, n\}$ which interchanges $i_1$ and $i_2$ and leaves everything else fixed. Then $\tau$ is a transposition, so $sgn(\tau) = -1$ and, more generally, $sgn(\sigma \tau) = -sgn(\sigma)$ for every $\sigma \in S_n$. By rearranging the order of factors in each term, we have

$$
\det(B) = \sum_{\sigma \in S_n} sgn(\sigma)a_{1\sigma(1)} \cdots a_{n\sigma(n)} = -\sum_{\sigma \in S_n} sgn(\sigma \tau)a_{1\sigma(1)} \cdots a_{n\sigma(n)}.
$$

But as $\sigma$ ranges over $S_n$, so does $\sigma \tau$; so the last sum is exactly $\det(A)$.

This property of the determinant is called the *alternating* property. It implies:

**III.11.7.10.** **Corollary.** Let $A$ be an $n \times n$ matrix. If $A$ has two identical rows, then $\det(A) = 0$.

(This is actually not a corollary of III.11.7.7. if $F$ has characteristic 2. But this result has a direct proof which works also in the characteristic 2 case; see Exercise III.11.11.1.)

Combining this with III.11.7.6., we obtain an extension:

**III.11.7.12.** **Corollary.** Let $A$ be an $n \times n$ matrix over $F$. If one row of $A$ is a scalar multiple of another row, then $\det(A) = 0$.

**Proof:** If the $i_2$’th row of $A$ is $r$ times the $i_1$’th row, then by III.11.7.6., $\det(A) = r \cdot \det(B)$ for a matrix $B$ with two identical rows.

This will be further extended in ().

Now combining this with III.11.7.7., we obtain:
III.11.7.13. Corollary. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$, and let $B$ be the matrix obtained from $A$ by replacing one row by that row plus a scalar multiple of another row. Then $\det(B) = \det(A)$.

Proof: Let $C$ be the matrix obtained from $A$ by replacing the $i_1$'th row by $r$ times the $i_2$'th row for some $r \in R$. If $b_{i_1,j} = a_{i_1,j} + ra_{i_2,j}$ and $b_{i_2,j} = a_{i_2,j}$ for $i \neq i_1$, then by III.11.7.7, we have $\det(B) = \det(A) + \det(C)$. But $\det(C) = 0$ by III.11.7.12.

Another important property of determinants has a proof similar to the proof of III.11.7.9.

III.11.7.14. Proposition. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$. Then $\det(A^t) = \det(A)$.

Proof: We have

$$\det(A^t) = \sum_{\sigma \in S_n} \text{sgn}(\sigma)a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}.$$ 

Since $\text{sgn}(\sigma^{-1}) = \text{sgn}(\sigma)$ for any $\sigma$, by rearranging the order of the factors in each term we get

$$\det(A^t) = \sum_{\sigma \in S_n} \text{sgn}(\sigma^{-1})a_{1\sigma^{-1}(1)}a_{2\sigma^{-1}(2)}\cdots a_{n\sigma^{-1}(n)}.$$ 

But as $\sigma$ ranges over $S_n$, so does $\sigma^{-1}$, hence the last sum is just $\det(A)$.

As a result, the previous results on rows apply also to columns:

III.11.7.15. Corollary. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$.

(i) The determinant is multilinear in the columns of $A$. In particular, if $B$ is obtained from $A$ by multiplying a column of $A$ by a scalar $r$, then $\det(B) = r \cdot \det(A)$.

(ii) If $B$ is obtained from $A$ by interchanging two columns, then $\det(B) = -\det(A)$.

(iii) If $A$ has two identical columns, or if one column of $A$ is a scalar multiple of another column, then $\det(A) = 0$.

(iv) If $B$ is obtained from $A$ by replacing one column of $A$ by that column plus a scalar times another column, then $\det(B) = \det(A)$.

There are many other properties of determinants discussed in abstract algebra books. We mention only a couple which are relevant for analysis. The proofs are quite obvious from the definition.

III.11.7.16. Proposition. (i) Let $A$ be an $n \times n$ matrix with entries in $\mathbb{Z}$. Then $\det(A) \in \mathbb{Z}$.

(ii) Let $A = [a_{ij}]$ be an $n \times n$ matrix with entries in $\mathbb{C}$, and $\bar{A} = [\bar{a}_{ij}]$ be the complex conjugate matrix. Then $\det(\bar{A}) = \det(A)$.

Combining (ii) with III.11.7.14, we obtain:
III.11.7.17. **Corollary.** Let $A$ be an $n \times n$ matrix over $\mathbb{C}$, and $A^*$ its adjoint matrix ($\cdots$). Then $\det(A^*) = \det(A)$.

III.11.7.18. In fact, the theory of determinants works equally well over any commutative ring (some aspects require the ring to be unital): if $R$ is a commutative ring, for any $n$ there is a determinant function from $M_n(R)$ to $R$ which is $n$-linear and alternating, defined by the same formula as in III.11.7.1. If $R$ and $S$ are commutative rings and $\phi : R \to S$ is a ring-homomorphism, then there is an induced coordinatewise ring homomorphism $\phi_n$ from $M_n(R)$ to $M_n(S)$ for any $n$, and the determinant is natural, i.e. $\det(\phi_n(A)) = \phi(\det(A))$ for any $A \in M_n(R)$. See also III.11.7.53.

However, determinants simply do not work for matrices over noncommutative rings; there is no determinant-like function with reasonable properties on $M_n(R)$ for noncommutative $R$.

**Expansion by Minors**

The multilinear and alternating properties of the determinant lead to an inductive procedure for computing the determinant of an $n \times n$ matrix in terms of determinants of smaller submatrices. This procedure, called expansion by minors or expansion by cofactors, is the way most determinants (at least ones of modest size) are computed in practice.

We first define a set of submatrices of an $n \times n$ matrix:

III.11.7.19. **Definition.** Let $A$ be an $n \times n$ matrix over $F$, and $1 \leq i, j \leq n$. The $(i,j)$ minor of $A$ is the $(n-1) \times (n-1)$ matrix $A^m_{ij}$ obtained from $A$ by deleting the $i$th row and $j$th column. The $(i,j)$ cofactor of $A$ is $d_{ij}(A) = (-1)^{i+j} \det(A^m_{ij})$.

The $(-1)^{i+j}$ in the definition of cofactor is conventional, and simplifies later formulas.

III.11.7.20. **Theorem.** Let $A$ be an $n \times n$ matrix over $F$. For any fixed $i_0$, $1 \leq i_0 \leq n$, we have

$$\det(A) = \sum_{j=1}^{n} a_{i_0j} \cdot d_{i_0j}(A) = \sum_{j=1}^{n} (-1)^{i_0+j} a_{i_0j} \cdot \det(A^m_{i_0j}).$$

For any fixed $j_0$, $1 \leq j_0 \leq n$, we have

$$\det(A) = \sum_{i=1}^{n} a_{ij_0} \cdot d_{ij_0}(A) = \sum_{i=1}^{n} (-1)^{i+j_0} a_{ij_0} \cdot \det(A^m_{ij_0}).$$

Thus the computation of an $n \times n$ determinant can be reduced to the computation of $n$ determinants, each $(n-1) \times (n-1)$. Each of these can be calculated by computing $n-1$ determinants, each $(n-2) \times (n-2)$, etc. It does not matter which row or column is used at each stage; as a practical matter, if the matrix has some zero entries it usually saves work to choose the row or column with the most zeroes at each stage.
III.11.7.21. Example. Expanding along the first row \((i_0 = 1)\), we have

\[
\det \begin{bmatrix}
1 & 2 & -1 \\
3 & -2 & 1 \\
1 & 0 & -4
\end{bmatrix} = (-1)^2(1)\det \begin{bmatrix}
-2 & 1 \\
0 & -4
\end{bmatrix} + (-1)^3(2)\det \begin{bmatrix}
3 & 1 \\
1 & -4
\end{bmatrix} + (-1)^4(-1)\det \begin{bmatrix}
3 & -2 \\
1 & 0
\end{bmatrix}
\]

\[
= 8 - 2(-13) - 2 = 32.
\]

Expanding along the second column \((j_0 = 2)\), we have

\[
\det \begin{bmatrix}
1 & 2 & -1 \\
3 & -2 & 1 \\
1 & 0 & -4
\end{bmatrix} = (-1)^1(2)\det \begin{bmatrix}
3 & 1 \\
1 & -4
\end{bmatrix} + (-1)^2(-1)\det \begin{bmatrix}
1 & -1 \\
1 & 1
\end{bmatrix} + (-1)^3(0)\det \begin{bmatrix}
1 & -1 \\
3 & 1
\end{bmatrix}
\]

\[
= -2(-13) - 2(-3) + 0 = 32.
\]

The second way is slightly simpler since one term is zero and only two \(2 \times 2\) determinants must be computed.

Determinants and Elementary Matrices

There is another way to view the multilinear and alternating properties of the determinant. Each describes the behavior of the determinant under an elementary row operation. Each elementary row operation on a matrix \(A\) can be performed by multiplying on the left by an elementary matrix. We can easily compute the determinant of elementary matrices:

III.11.7.22. Proposition. Determinants of elementary matrices are as follows:

(i) If \(\Pi_{ij}\) is the matrix that interchanges \(i\) and \(j\), then \(\det(\Pi_{ij}) = -1\).

(ii) If \(M_i(r)\) is the diagonal matrix with \((i, i)\) entry \(r\) and the rest of the diagonal entries 1, then \(\det(M_i(r)) = r\).

(iii) If \(E_{ij}(r)\) \((i \neq j)\) is the matrix obtained from \(I_n\) by replacing the \((i, j)\) element by \(r\), then \(\det(E_{ij}(r)) = 1\).

This can be proved by applying III.11.7.9. for (i), III.11.7.6. (or just III.11.7.4.) for (ii), and III.11.7.13. for (iii) (or readily proved directly from the definition of determinant).

Comparing the effect of row operations on the determinant with III.11.7.22., we obtain:

III.11.7.23. Corollary. Let \(A\) be an \(n \times n\) matrix over \(F\), and let \(E\) be an elementary \(n \times n\) matrix over \(F\). Then \(\det(EA) = \det(E) \cdot \det(A)\).

Multiplicative Property of Determinants

The two most important properties of determinants for most applications are simple consequences of III.11.7.23. and (i).
III.11.7.24. Theorem. Let $A$ be an $n \times n$ matrix over $F$. Then $A$ is invertible if and only if $\det(A) \neq 0$.

Proof: By (), there are finitely many elementary matrices $E_1, \ldots, E_m$ such that either $E_1 E_2 \cdots E_m A = I_n$ (if $A$ is invertible) or $E_1 E_2 \cdots E_m A$ has a zero row (if $A$ is singular). In the first case we have

$$A = E_m^{-1} E_{m-1}^{-1} \cdots E_1^{-1}$$

so $A$ is a product of elementary matrices. By repeated applications of III.11.7.23., we obtain

$$\det(A) = \det(E_m^{-1}) \cdot \det(E_{m-1}^{-1}) \cdots \det(E_1^{-1}) \neq 0.$$ 

In the second case, we have from III.11.7.23. that $\det(E_1 \cdots E_m A)$ is a nonzero multiple of $\det(A)$. But $E_1 \cdots E_m A$ has a zero row, so $\det(E_1 \cdots E_m A) = 0$ and hence $\det(A) = 0$. ☝

III.11.7.25. Theorem. Let $A$ and $B$ be $n \times n$ matrices over $F$. Then $\det(AB) = \det(A) \cdot \det(B)$.

Proof: If $A$ is not invertible, then $AB$ is not invertible () and $\det(AB) = 0 = \det(A) \cdot \det(B)$. And if $A$ is invertible, then $A$ is a product of elementary matrices, so by repeated applications of III.11.7.23., we obtain that $\det(AB) = \det(A) \cdot \det(B)$. ☃

Proving this result by direct computation from the definition of determinant is a mess even for $n = 2$, and becomes impossibly complicated for larger $n$. The result holds for matrices over general commutative rings, although it needs a different proof there (cf. III.11.7.28.).

III.11.7.26. In the case of $F = \mathbb{R}$, the result follows immediately from the measure-scaling property of the determinant (). Note, however, that in establishing the properties of the determinant it is important to avoid circularity in arguments; our proof of the measure-scaling property () uses the multiplicativity of the determinant, so while the measure-scaling property is a good geometric explanation of multiplicativity, it is not a good way to prove it.

The Uniqueness Theorem

It turns out that the properties we have obtained uniquely characterize the determinant function:

III.11.7.27. Theorem. Let $F$ be a field.

(i) For any $n \in \mathbb{N}$ there is a unique function $\phi : M_n(F) \to F$ which is $n$-linear in the rows, alternating, and such that $\phi(I_n) = 1$. We have $\phi(A) = \det(A)$ for all $A \in M_n(F)$.

(ii) If $\phi : M_n(F) \to F$ is $n$-linear in the rows, alternating, and $\phi(I_n) = 0$, then $\phi$ is identically zero.

Proof: (i): The existence of such a $\phi$, namely $\det$, has already been established. We prove uniqueness by induction on $n$. For $n = 1$, $\phi$ must be a linear function from $F$ to $F$, hence $\phi(a) = ma$ for some constant $m$; since $\phi(1) = 1$, $m = 1$, so $\phi([a]) = a = \det([a])$. Now suppose the statement has been proved for $n$. ***** ☃

403
III.11.7.28. An alternate proof of III.11.7.25. can be based on the Uniqueness Theorem:

PROOF: Fix $B \in M_n(F)$. It is routine to check that the function $\phi : M_n(F) \to F$ given by $\phi(A) = det(AB)$ is $n$-linear and alternating. We have $\phi(I) = det(B)$. If $det(B) = 0$, then $\phi$ is identically zero by III.11.7.27.(ii); thus $det(AB) = 0 = det(A) \cdot det(B)$ for any $A$. And if $det(B) \neq 0$, set $\psi(A) = \frac{1}{det(B)} \phi(A) = \frac{1}{det(B)} det(AB)$.

By III.11.7.27.(i), $\psi(A) = det(A)$ for all $A$, and thus $\phi(A) = det(AB) = det(A) \cdot det(B)$ for any $A$. ✅

Continuity and Differentiability of the Determinant

The Determinant of a Linear Transformation

III.11.7.29. If $V$ is a finite-dimensional vector space over $F$, and $T$ is a linear transformation from $V$ to $V$, then $T$ does not have a canonical matrix representation; to represent $T$ by a matrix, a basis for $V$ must be chosen (). Although the matrix of $T$ depends on the choice of basis, there is not too much flexibility: if $B$ and $B'$ are bases, then the matrices of $T$ with respect to $B$ and $B'$ are similar:

$$[T]_{B'} = P_{B,B'}^{-1} [T]_B P_{B,B'}$$

where $P_{B,B'}$ is the change-of-basis matrix (), and so we have

$$det([T]_{B'}) = det(P_{B,B'})^{-1} det([T]_B) det(P_{B,B'}) = det([T]_B)$$

(using commutativity of multiplication in $F$), i.e. the determinant of $[T]_B$ does not depend on the choice of basis $B$. Thus the linear transformation $T$ has a well-defined associated scalar, the determinant of the matrix representation of $T$ with respect to any basis.

III.11.7.30. Definition. Let $T$ be a linear transformation from a finite-dimensional vector space $V$ over $F$ to itself. Set

$$det(T) = det([T]_B)$$

where $B$ is any basis for $V$ over $F$.

III.11.7.31. Note that if $S,T$ are linear transformations from $V$ to $V$, then

$$det(S \circ T) = det(S)det(T) .$$

III.11.7.32. The determinant of a linear transformation does depend on the choice of base field $F$, however. If $V$ is a vector space over $\mathbb{C}$ and $T : V \to V$ a linear transformation, then $det(T)$ is not the same if $V$ is regarded as a vector space over $\mathbb{R}$ as it is over $\mathbb{C}$ in general. (See Exercise () for the relationship.)

Geometric Interpretation of the Determinant

Although the theory of determinants is purely algebraic and works equally well over any field (in fact, over any commutative ring), there is a geometric interpretation of the determinant which is exclusive to the case where the field $F$ is $\mathbb{R}$ (and, to a limited extent, to the case $F = \mathbb{C}$). This interpretation underlies the importance of the determinant in analysis.
III.11.7.33. **Volume Principle of Determinants:** Let $A$ be an $n \times n$ matrix with real entries. Regard $A$ as giving a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ (with respect to the standard basis). Then $T$ multiplies $n$-dimensional “volume” by a constant factor $|\text{det}(A)|$, i.e. if $E$ is a subset of $\mathbb{R}^n$ with a well-defined $n$-dimensional “volume” $V(E)$ (technically a Jordan region () or, more generally, a Lebesgue measurable subset), then $T(E)$ also has a well-defined $n$-dimensional volume and $V(T(E)) = |\text{det}(A)| \cdot V(E)$.

III.11.7.34. Actually, we should first define what we mean by $n$-dimensional “volume” of a subset of $\mathbb{R}^n$. This is not as easy as it sounds: the general problem is essentially the basic subject matter of measure theory. It would take us too far afield to try to treat this question here, so we will just assume that “volume” is defined for sufficiently well-behaved subsets of $\mathbb{R}^n$, and satisfies

(i) If $R$ is a “rectangle” $[a_1, b_1] \times \cdots \times [a_n, b_n]$, then $V(R) = (b_1 - a_1) \cdots (b_n - a_n)$.

(ii) If $E \subseteq F$, then $V(E) \leq V(F)$.

(iii) If $E_1, \ldots, E_m$ are nonoverlapping regions (“nonoverlapping” means the intersection of any two is contained in the boundary of both), then

$$V(E_1 \cup \cdots \cup E_m) = V(E_1) + \cdots + V(E_m).$$

(iv) Congruent regions (including allowing reflections) have the same volume.

(There is actually a subtle technical problem with (iii), since the boundary of a region can have positive measure, but this will not cause difficulty if we consider only regions bounded by finitely many smooth hypersurfaces. Actually we will only need to work with regions bounded by finitely many hyperplanes.)

III.11.7.35. We first look at the case $n = 2$, which is the only one we will examine in detail. Here 2-dimensional “volume” is “area.” Suppose

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

is a $2 \times 2$ matrix. The $T(1,0) = (a,b)$ and $T(0,1) = (c,d)$. The image of the unit square $[0,1] \times [0,1]$ is the parallelogram with vertices $(0,0)$, $(a,b)$, $(c,d)$, and $(a+c, b+d)$. We claim the area of this parallelogram is $|\text{det}(A)| = |ad - bc|$.

The argument reduces to several special cases. First consider the case where $a > 0, b \geq 0, d > 0, c \geq 0$, and either $c = 0$ or $\frac{d}{c} \leq \frac{b}{a}$ (Figure III.1).

The area of the total rectangle is $(a+c)(b+d)$. The blue triangles fit together into a rectangle of length $a$ and height $b$, so the blue area is $ab$. Similarly, the green area is $cd$. The two red rectangles each have area $bc$. Thus the area of the parallelogram is

$$(a + c)(b + d) - ab - cd - 2bc = ab + ad + bc + cd - ab - cd - 2bc = ad - bc.$$  

The case where $\frac{d}{c} < \frac{b}{a}$ can be handled by interchanging columns of $A$, which multiplies the determinant by $-1$.

III.11.7.36. Now look at the case $c < 0$, but $a, b, d \geq 0$ (Figure III.2).

The area of the large rectangle is $(a-c)(b+d)$. The blue triangles fit together into a rectangle of area $ab$, and the green triangles have total area $-cd$. So the area of the parallelogram is

$$(a - c)(b + d) - ab + cd = ab + ad - bc - cd - ab + cd = ad - bc.$$
III.11.7.37. The other cases can be handled similarly, or by reducing to these cases by interchanging rows or columns and/or multiplying one or both rows and/or columns by \(-1\), none of which change the absolute value of the determinant.

III.11.7.38. If instead of the unit square, we begin with the rectangle with vertices \((0, 0)\), \((r, 0)\), \((0, s)\), and \((r, s)\), a virtually identical argument shows that the image parallelogram has area \(|\det(A)|rs\). The image of any other coordinate rectangle (rectangle with sides parallel to the axes) of length \(rs\) is congruent to this parallelogram, hence has the same area by property (iv). Thus \(T\) multiplies areas of all coordinate rectangles by the constant factor \(|\det(A)|\).

III.11.7.39. For the case \(n = 3\), the image of the unit cube under the transformation \(T\) is the parallelepiped with vertices at the origin, the column vectors of \(A\), and their sums. A more complicated argument along the same lines as the dimension 2 case shows that the volume of this parallelepiped is exactly \(|\det(A)|\). The argument generalizes to the images of rectangular solids with edges parallel to the axes (coordinate rectangles).

III.11.7.40. The case \(n \geq 4\) is difficult to visualize and quite complicated to give a geometric argument for. But some special cases can be easily shown for any \(n\):
(i) If $A$ is a diagonal matrix, the image of any coordinate rectangle $R$ is another coordinate rectangle, whose $n$-dimensional volume is $V(R)$ times the absolute value of the product of the diagonal elements of $A$, which is $|\det(A)|$.

(ii) If $A$ is a permutation matrix ($\cdot$), and $R$ is a coordinate rectangle, then $T(R)$ is a coordinate rectangle congruent to $R$, so $V(T(R)) = V(R)$; and $\det(A) = \pm 1$, so $|\det(A)| = 1$ and the measure-scaling property holds.

(iii) If $A$ is the elementary matrix with $r$ in the $(i,j)$ place ($\cdot$), then the image of the unit cube is a prism whose base is a rhombus of side 1; thus $V(T(R)) = V(R)$, and $\det(A) = 1$ ($\cdot$).

By the multiplicative property of determinants and the fact that any invertible square matrix is a product of elementary matrices, the measure-scaling property holds for coordinate rectangles for general invertible $A$. For singular $A$, see ($\cdot$).

**III.11.7.41.** To handle a general (Jordan) region $E$, cover $E$ by a fine grid so that the inside and outside estimates for $V(E)$ ($\cdot$) from this grid are close together. The image $T(E)$ is then trapped between finite unions of paralleloptopes, both of which have volume close to $|\det(A)| \cdot V(E)$ by property (iii) of volume. The approximations can be made as closely as desired. By property (ii), $V(T(E)) = |\det(A)| \cdot V(E)$. A similar argument ($\cdot$) gives the result for Lebesgue measurable $E$. 407
III.11.7.42. If $A$ is an $n \times n$ matrix over $\mathbb{C}$, then $A$ can be thought of as giving a linear transformation $T$ from $\mathbb{C}^n \cong \mathbb{R}^{2n}$ to itself, and it can be shown that the factor that $T$ multiplies $2n$-dimensional “volume” by is the square of $|\det(A)|$.

The Inverse of a Matrix

There is actually an explicit algebraic formula for the inverse of a (square) matrix, which we develop here.

III.11.7.43. Definition. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$. The classical adjoint of $A$ is the matrix

$$\text{adj}(A) = [d_{ji}(A)] = [d_{ij}(A)]^t$$

where $d_{ij}(A)$ is the $(i,j)$ cofactor of $A$ (III.11.7.19).

The classical adjoint has nothing to do with the (ordinary) adjoint of a matrix over $\mathbb{C}$ as defined in (). Note the transpose in the definition of $\text{adj}(A)$.

The importance of the classical adjoint is:

III.11.7.44. Theorem. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$. Then

$$A \cdot \text{adj}(A) = \text{adj}(A) \cdot A = (\det(A))I_n .$$


As a corollary, we get an algebraic formula for the inverse of a matrix:

III.11.7.45. Corollary. Let $A$ be an $n \times n$ matrix over $\mathbb{F}$. If $A$ is invertible (i.e. $\det(A) \neq 0$), then

$$A^{-1} = \frac{1}{\det(A)} \text{adj}(A) .$$

III.11.7.46. The algebraic formula for the inverse via the classical adjoint is not computationally efficient to use; it is pretty cumbersome even for $n = 4$. It is much more computationally efficient to compute the inverse by row reduction (Gaussian elimination) (). But the formula at least gives a very simple way to write down the inverse of a $2 \times 2$ matrix:

III.11.7.47. Corollary. Let $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ be a $2 \times 2$ matrix over $\mathbb{F}$. Then $A$ is invertible if and only if $ad - bc \neq 0$, and in this case

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} \frac{d}{ad - bc} & \frac{-b}{ad - bc} \\ \frac{-c}{ad - bc} & \frac{a}{ad - bc} \end{bmatrix} .$$

This formula may be easily checked by direct computation.

The algebraic formula for the inverse does have an important theoretical consequence:
III.11.7.48. Corollary. The entries of the inverse of an invertible (square) matrix $A$ over $\mathbb{F}$ are rational functions of the entries in $A$.

If $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, this can be nicely rephrased.

III.11.7.49. Definition. Let $\mathbb{F}$ be a field. Write $GL_n(\mathbb{F})$ for the set of invertible $n \times n$ matrices over $\mathbb{F}$. $GL_n(\mathbb{F})$ is called the general linear group of size $n$ over $\mathbb{F}$.

If $A, B \in GL_n(\mathbb{F})$, then $AB$ and $A^{-1}$ are also in $GL_n(\mathbb{F})$; thus $GL_n(\mathbb{F})$ is a group under matrix multiplication. By III.11.7.24, $GL_n(\mathbb{F}) = \{ A \in M_n(\mathbb{F}) : \det(A) \neq 0 \}$.

III.11.7.50. Theorem. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. Then $GL_n(\mathbb{F})$ is an open set in $M_n(\mathbb{F}) \cong \mathbb{F}^{n^2}$, and $A \mapsto A^{-1}$ is an analytic (hence $C^\infty$) function from $GL_n(\mathbb{F})$ to $GL_n(\mathbb{F}) \subseteq M_n(\mathbb{F})$.

Proof: Since $\det : M_n(\mathbb{F}) \to \mathbb{F}$ is continuous, $GL_n(\mathbb{F})$ is the inverse image of the open set $\mathbb{F} \setminus \{0\}$ under a continuous map, hence open. The other statement follows from III.11.7.48, since rational functions are analytic on their domains.

III.11.7.51. Corollary. Let $A = [a_{ij}]$ be an invertible $n \times n$ matrix over $\mathbb{R}$ or $\mathbb{C}$. Then there is an $\epsilon > 0$ (depending on $A$) such that, whenever $B = [b_{ij}]$ is an $n \times n$ matrix over $\mathbb{R}$ or $\mathbb{C}$ with $|a_{ij} - b_{ij}| < \epsilon$ for all $i$ and $j$, then $B$ is also invertible.

The formula for the inverse of a matrix has other consequences, for example:

III.11.7.52. Corollary. Let $A$ be an $n \times n$ matrix with integer entries. Then $A$ has an inverse with integer entries (i.e. $A$ is invertible in $M_n(\mathbb{Z})$) if and only if $\det(A) = \pm 1$.

Proof: If $A$ has an inverse $A^{-1}$ with integer entries, then $\det(A)$ and $\det(A^{-1}) = \frac{1}{\det(A)}$ are integers, hence $\det(A) = \pm 1$. Conversely, if $\det(A) = \pm 1$, III.11.7.45. (applied with $\mathbb{F} = \mathbb{R}$) shows that $A$ is invertible and that $A^{-1}$ has integer entries.

This result generalizes to unital commutative rings:

III.11.7.53. Theorem. Let $R$ be a commutative unital ring. If $A \in M_n(R)$, then $A$ is invertible in $M_n(R)$ if and only if $\det(A)$ is a unit (invertible element) of $R$.

Here is another consequence. If $R$ is a unital subring of a larger unital ring $R'$, an element of $R$ can in general be invertible in the larger ring $R'$ without being invertible in $R$ (e.g. $R = \mathbb{Z}$, $R' = \mathbb{R}$). But if $\mathbb{F}$ is a field contained in a larger field $\mathbb{F}'$, and we regard $M_n(\mathbb{F})$ as a subring of $M_n(\mathbb{F}')$, this cannot happen:
**III.11.7.54. Corollary.** Let $F \subseteq F'$ be fields. If $A \in M_n(F)$, then $A$ is invertible in $M_n(F')$ if and only if $A$ is invertible in $M_n(F)$, i.e. if $A$ is invertible in $M_n(F')$, then the entries of $A^{-1}$ are in $F$.

This applies for example if $F = \mathbb{R}$ and $F' = \mathbb{C}$, or if $F = \mathbb{Q}$ and $F' = \mathbb{R}$.

Cramer’s Rule

**III.11.8. Eigenvalues and Eigenvectors**

**III.11.9. Tensor Products and Multilinear Algebra**

**III.11.10. Differential Algebra**

**III.11.11. Exercises**

**III.11.11.1.** Let $A$ be an $n \times n$ matrix over $F$. Show directly from the definition that if $A$ has two identical rows, then the terms in the expansion of $\det(A)$ cancel in pairs, and hence $\det(A) = 0$. [If $\tau$ is the transposition interchanging the two rows, for any $\sigma$ consider the terms for $\sigma$ and $\sigma \tau$.]

**III.11.11.2.** Say that a $2m \times 2n$ matrix $A$ over $\mathbb{R}$ is complex if, when $A$ is partitioned into $2 \times 2$ blocks, each subblock is of the form

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

where the $a$ and $b$ are real numbers which vary from block to block (cf. III.6.3.1.).

(a) Show that a sum or product of complex matrices, if defined, is complex. [Use III.11.1.22.]

(b) If a $2n \times 2n$ matrix over $\mathbb{R}$ is complex and invertible (in $M_{2n}(\mathbb{R})$), show that its inverse is also complex.

(c) If $A$ is the complex matrix

$$\begin{bmatrix} a_{11} & -b_{11} & a_{12} & -b_{12} & \cdots & a_{1n} & -b_{1n} \\ b_{11} & a_{11} & b_{12} & a_{12} & \cdots & b_{1n} & a_{1n} \\ a_{21} & -b_{21} & a_{22} & -b_{22} & \cdots & a_{2n} & -b_{2n} \\ b_{21} & a_{21} & b_{22} & a_{22} & \cdots & b_{2n} & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & -b_{m1} & a_{m2} & -b_{m2} & \cdots & a_{mn} & -b_{mn} \\ b_{m1} & a_{m1} & b_{m2} & a_{m2} & \cdots & b_{mn} & a_{mn} \end{bmatrix}$$

identify $A$ with the $m \times n$ matrix

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{11} & c_{12} & \cdots & c_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix}$$

with complex entries, where $c_{jk} = a_{jk} + ib_{jk}$ for $1 \leq j \leq m$, $1 \leq k \leq n$. Show that this correspondence preserves addition, multiplication, and inversion.

(d) Interpret this correspondence as follows. Let $S = \{e_1, \ldots, e_n\}$ be the standard (ordered) basis for $\mathbb{C}^n$ as a vector space over $\mathbb{C}$. Then the set $B = \{e_1, ie_1, \ldots, e_n, ie_n\}$ is an ordered basis for $\mathbb{C}^n$ as a real vector
space. If $T$ is a real-linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^m$, then the matrix for $T$ with respect to the $B$-bases is complex if and only if $T$ is complex-linear, in which case its matrix (over $\mathbb{C}$) with respect to the standard bases is the corresponding matrix in $M_{mn}(\mathbb{C})$.

(e) More generally, show that if $R$ is a ring, and $R'$ is a subring of $M_k(R)$, then the $m \times n$ matrices over $R'$ can be identified with the set of $km \times kn$ matrices over $R$ whose $k \times k$ block entries are in $R'$, and the identification preserves matrix addition, multiplication, and (if $R'$ is closed under inverses in $M_k(R)$) inversion. In particular, if $\mathbb{F}$ is a field and $\mathbb{F}'$ is a finite extension field of degree $k$, then $\mathbb{F}'$ can be embedded as a (unital) subring of $M_k(\mathbb{F})$, so this construction applies.

(f) Interpret this problem as a special case of the tensor product construction ().
Chapter IV

Sequences and Series

Infinite sequences and infinite series can be regarded as the beginning of analysis. The idea that a sequence of numbers can approach a limiting value, or that an infinite set of numbers can be “added”, is the principal concept that underlies the whole development of the subject. The notion of limit of a sequence is a subtle one which takes some getting used to, but which must be mastered before calculus and analysis of any type can be understood.

Sequences and their limits appear in various contexts throughout mathematics. The most basic situation is sequences of real (or complex) numbers, which is where we begin. The subject then moves on to sequences of points in a metric space, and then to sequences in topological spaces; in fact, the entire subject of topology can be regarded as a way of giving a general setting for studying limits of sequences. Another (intertwined) direction of development is the theory of limits of sequences of functions. The study of infinite series of numbers or functions may be regarded as a special case of the theory of sequences, or really the same theory from a different point of view, although much of this theory predated the advent of a careful theory of sequences and developed in parallel.

In this chapter, we will develop the theory of sequences and their limits, first for numbers and then in more general settings, and the theory of infinite series of numbers. Infinite sequences and series of functions will be postponed until the calculus chapter.

IV.1. Sequences

Recall () that a sequence in a set $X$ is a function from $\mathbb{N}$ to $X$. The usual notation uses a subscript for the argument of the function, which is called the index of the sequence. Thus the value of the function (sequence) $a$ at the number $n$ is normally written $a_n$ instead of $a(n)$; $a_n$ is called the $n$'th term of the sequence. Occasionally the index of a sequence takes values in a set of the form $\{n \in \mathbb{Z} : n \geq n_0\}$ instead of in $\mathbb{N}$.

The notation $(x_n)$ is commonly used to denote a sequence; the notation $(x_n)_{n \in \mathbb{N}}$ is a more precise variation. Some authors use the notation $(x_n)$, or something similar. Note the difference between $(x_n)$ and $x_n$: $(x_n)$ refers to the entire sequence, while $x_n$ denotes the $n$'th term of the sequence for a specific single $n$. We sometimes write a formula for the $n$'th term, surrounded by parentheses, to define a sequence; thus, for example, $(\frac{1}{n})$ or, more precisely, $(\frac{1}{n})_{n \in \mathbb{N}}$, denotes the sequence $(x_n)$ with $x_n = \frac{1}{n}$ for each $n$. 

412
IV.1.1. Sequences of Numbers

In this section and the subsequent subsections, we will concentrate on sequences of numbers, real or (occasionally) complex. The structure of the real numbers makes possible some definitions and operations which do not make sense for sequences in general sets.

Terms and Range of Sequences

IV.1.1.1. If \((x_n)\) is a sequence of numbers (or a sequence in a general set \(X\)), a careful distinction must be made between the terms of the sequence and the set \(\{x_n : n \in \mathbb{N}\}\). This set is the range of the function from \(\mathbb{N}\) to \(\mathbb{R}\) (or \(X\)). For example, consider the sequence \((-1)^n = (-1, 1, -1, 1, \ldots)\). There are infinitely many terms in this sequence, as there are in any sequence, but the range of the sequence is the two-point set \(\{1, -1\}\). The terms \(x_2\) and \(x_4\) are different terms of the sequence, even though these terms have the same numerical value. The most extreme example is a constant sequence, where all the terms of the sequence have the same numerical value; they are nonetheless distinct terms.

This distinction causes complications when we try to depict sequences on the number line. Fortunately, sequences which arise naturally in applications tend to have terms with distinct values, i.e. as a function the sequence is injective, and for such sequences the distinction largely disappears, but not entirely: for example, if the order of the terms is rearranged a different sequence is obtained, even though it has the same range.

Bounded and Unbounded Sequences

IV.1.1.2. **Definition.** A sequence \((x_n)\) of real numbers is bounded above if there is an \(M \in \mathbb{R}\) such that \(x_n \leq M\) for all \(n\). The sequence \((x_n)\) is bounded below if there is an \(M \in \mathbb{R}\) such that \(M \leq x_n\) for all \(n\). The sequence \((x_n)\) is bounded if it is both bounded above and bounded below, i.e. if there is an \(M \in \mathbb{R}\) such that \(|x_n| \leq M\) for all \(n\). A sequence which is not bounded is unbounded.

In other words, the sequence \((x_n)\) is bounded [bounded above, bounded below] if and only if the range \(\{x_n : n \in \mathbb{N}\}\) is bounded [bounded above, bounded below] as a subset of \(\mathbb{R}\).

Algebraic Operations on Sequences

IV.1.2. Subsequences and Tails

IV.1.3. Limits of Sequences

IV.1.3.1. Now we turn to the fundamental concept of analysis, the notion of what it means for a sequence to converge to a limit. The idea is that the sequence \((x_n)\) approaches the limit (limiting value) \(a\) if eventually, i.e. from some point on, the terms of the sequence are approximately equal to \(a\) to within any given tolerance. Pictorially, for sequences of numbers, on the number line the number \(x_n\) should become almost indistinguishable from the limiting value \(a\) for all sufficiently large \(n\):

The terms of the sequence, however, do not ever have to be exactly equal to the limit, and “usually” are not. Indeed, the notion of limit is most interesting for sequences which are not eventually constant.

It is important to note that “most” sequences of numbers do not approach a limiting value, i.e. do not converge. In fact, it is quite special for a sequence to have a limit.
IV.1.3.2. A way of making this notion precise is to say that the sequence \((x_n)\) approaches the limit \(a\) if, no matter how tiny an interval \(I\) around \(a\) we specify, all the terms of the sequence eventually lie within the interval \(I\). The smaller the interval \(I\), the longer we might have to wait before the terms of the sequence settle down into \(I\), but eventually they must for any \(I\):

We are thus led to the following formal definition.

IV.1.3.3. Definition. [Limit of a Sequence, Version 1] Let \((x_n)\) be a sequence in \(\mathbb{R}\), and \(a \in \mathbb{R}\). The sequence \((x_n)\) converges to \(a\), or \(a\) is the limit of the sequence \((x_n)\), if for every interval \(I\) in \(\mathbb{R}\) containing \(a\) in its interior, there is an \(N \in \mathbb{N}\) such that \(x_n \in I\) for all \(n \geq N\).

IV.1.3.4. It turns out to usually be more useful in practice to rephrase the definition slightly. It is only necessary to consider open intervals which are centered at \(a\), since any interval containing \(a\) in its interior contains a smaller open interval centered at \(a\). It is customary in analysis to use the Greek letter epsilon (\(\epsilon\)), or the Greek letter delta (\(\delta\)), to denote the length of the interval to the left or right of \(a\); thus any (bounded) interval centered at \(a\) is of the form

\[(a-\epsilon, a+\epsilon)\]

for some positive real number \(\epsilon\), and conversely any \(\epsilon > 0\) defines such an interval:

Thus we can rephrase the definition in an equivalent way:

IV.1.3.5. Definition. [Limit of a Sequence, Version 2] Let \((x_n)\) be a sequence in \(\mathbb{R}\), and \(a \in \mathbb{R}\). The sequence \((x_n)\) converges to \(a\), or \(a\) is the limit of the sequence \((x_n)\), if for every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[a - \epsilon < x_n < a + \epsilon\]

for all \(n \geq N\).

There is one last slight notational variation. Note that \(|x_n - a| < \epsilon\) if and only if \(a - \epsilon < x_n < a + \epsilon\). Thus we can restate the definition in its usual form:

IV.1.3.6. Definition. [Limit of a Sequence, Final Version] Let \((x_n)\) be a sequence in \(\mathbb{R}\), and \(a \in \mathbb{R}\). The sequence \((x_n)\) converges to \(a\), or \(a\) is the limit of the sequence \((x_n)\), if for every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[|x_n - a| < \epsilon\]

for all \(n \geq N\).

The sequence \((x_n)\) diverges if there is a real number \(a\) such that \(a\) is the limit of the sequence \((x_n)\). If there is no such \(a\), the sequence \((x_n)\) diverges.

IV.1.3.7. It is unimportant in this definition whether or not we use strict inequality in the \(|x_n - a| < \epsilon\) and/or the \(n \geq N\), or whether we require that \(N \in \mathbb{N}\); see Exercise IV.1.11.1.
IV.1.3.8. **NOTATION:** There are two standard notations for the limit of a sequence. If \((x_n)\) converges to \(a\), we write

\[ x_n \to a \]

or

\[ \lim_{n \to \infty} x_n = a. \]

The last expression is read “the limit of \(x_n\) as \(n\) goes to infinity is \(a\).”

Note that we write \(\lim_{n \to \infty} x_n = a\), not \(\lim_{n \to \infty} x_n \to a\).

IV.1.3.9. **The definition of limit is often described as a game or bet.** If I claim that \(x_n \to a\) and you challenge me, we play a game. You give me a positive number \(\epsilon\), and I must produce an \(N\) that works in the definition for the \(\epsilon\) you give me. If I can, I win; if I can’t, you win. The smaller the \(\epsilon\) you give me, the harder it will be for me to find an \(N\) that works, and the larger it will have to be in general. But I don’t have to find the \(N\) until you give me your \(\epsilon\); my \(N\) can (and in general will) depend on what your \(\epsilon\) is.

A proof that \(x_n \to a\) really consists of a demonstration that I can always win the game; I must convince you that no matter what \(\epsilon\) you give me, I can always find an \(N\) that works, so that you realize it would be pointless for you to challenge me with any specific \(\epsilon\). On the other hand, a disproof that \(x_n \to a\) normally consists of demonstrating a specific positive \(\epsilon\) which is small enough that there is no \(N\) that works for it.

Note that for a given \(\epsilon\), there is more than one possible \(N\) that works: the \(N\) can always be increased with no penalty. If there is any \(N\) that works, there is always a smallest \(N\), but it may be difficult to find the smallest \(N\) in specific situations, and it is usually not necessary; a larger \(N\) which is easier to find works just as well. (There are instances, e.g. in some numerical work, where it is important to find the smallest, or at least very nearly the smallest, \(N\) that works in a given setting, but such situations are rather unusual in analysis.)

In the same way, if a certain \(N\) works for one \(\epsilon\) for a particular limit, the same \(N\) works for any larger \(\epsilon\), so there is no penalty in first making the \(\epsilon\) smaller before finding an \(N\) that works. It may be hard to believe that it can sometimes be useful to do this, but it is; see, for example, ()

IV.1.3.10. **EXAMPLES.** (i) Let \(x_n = \frac{1}{n}\) for all \(n\). This is one of the most important limits, and will appear constantly. It seems intuitively clear that \(x_n \to 0\). To prove this, we must show that for any \(\epsilon > 0\), there is an \(N\) such that

\[ |x_n - 0| = \left| \frac{1}{n} - 0 \right| = \frac{1}{n} < \epsilon \]

for all \(n \geq N\). So suppose \(\epsilon > 0\) is given. In order to have \(\frac{1}{n} < \epsilon\), it is sufficient (and also necessary) that \(n > \frac{1}{\epsilon}\). Thus we can choose \(N\) to be any natural number greater than \(\frac{1}{\epsilon}\). There is such an \(N\) by the Archimedean property of \(\mathbb{R}\) (). Then if \(n \geq N\), we have

\[ |x_n - 0| = \frac{1}{n} \leq \frac{1}{N} < \epsilon \]

so this \(N\) works for the given \(\epsilon\).

Note that if \(\epsilon\) is very small, then \(\frac{1}{\epsilon}\) is very large, so \(N\) will have to be large; the smaller the \(\epsilon\), the larger the \(N\) will have to be. For example, if \(\epsilon = 10^{-9}\), we will have to take \(N = 1,000,000,001\) (or any larger integer).

Note that none of the terms of this sequence are exactly equal to the limit 0.
(ii) Similarly, suppose \( x_n = 1 + \frac{1}{n} \). We guess that \( x_n \to 1 \). To prove this, we must show that, if \( \epsilon > 0 \), there is an \( N \) such that
\[
|x_n - 1| = \left| \left(1 + \frac{1}{n}\right) - 1\right| = \frac{1}{n} < \epsilon
\]
for all \( n \geq N \). Thus, by an identical argument to the one in (i), we can choose \( N \) to be any integer greater than \( \frac{1}{\epsilon} \).

(iii) Here is an example which is so simple that it is hard. Suppose \((x_n)\) is a constant sequence with \( x_n = a \) for all \( n \). We certainly expect that \( x_n \not\to a \). If \( \epsilon > 0 \), we have
\[
|x_n - a| = |a - a| = 0 < \epsilon
\]
for any \( n \). Thus we can take \( N = 1 \) no matter what \( \epsilon \) is, so \( x_n \to a \). This is practically the only example where the \( N \) does not depend on \( \epsilon \).

(iv) Suppose \( x_n = (-1)^n \) for all \( n \). Since this sequence alternates \( \pm 1 \), we might expect that the sequence has two limits, 1 and \(-1\). However, this is incorrect; in fact, this sequence diverges, i.e. has no limit, since to have a limit \( a \) all the terms of the sequence from some point on must be close to \( a \). (We will later define cluster points for sequences; this sequence has cluster points 1 and \(-1\).)

For example, let us show that \((x_n)\) does not converge to 1. We have \( |x_n - 1| = 0 \) if \( n \) is even, but \( |x_n - 1| = 2 \) for \( n \) odd. If we are given an \( \epsilon > 0 \), we may or may not be able to find an \( N \) that works. If the person challenging us is stupid enough to give us \( \epsilon = 3 \) (or any \( \epsilon > 2 \)), we can find an \( N \) that works: in fact, \( N = 1 \) works for such an \( \epsilon \). But if our challenger is smart enough to give us an \( \epsilon \) which is \( \leq 2 \), we cannot find an \( N \) that works, since no matter what \( N \) we take there will be an odd \( n \) which is greater than \( N \), and for this \( n \) we have
\[
|x_n - 1| = 2 \geq \epsilon .
\]
Thus we have shown that \((x_n)\) does not converge to 1. An almost identical argument shows that \((x_n)\) does not converge to \(-1\).

However, this does not yet show that the sequence \((x_n)\) diverges, i.e. that \( \lim_{n \to \infty} x_n \) does not exist. To do this, we must show that for any \( a \in \mathbb{R} \), the sequence does not converge to \( a \). Fix an \( a \in \mathbb{R} \). It suffices to exhibit an \( \epsilon > 0 \) for which there is no \( N \) that works for this \( \epsilon \) and \( a \). Take \( \epsilon = 1 \) (any smaller positive number also works). Suppose there is an \( N \) that works, i.e.
\[
|x_n - a| < 1
\]
for all \( n \geq N \). If we take \( n \) to be an odd number greater than \( N \), we have that
\[
|x_n - a| = |1 - a| = |1 + a| < 1
\]
and similarly if \( n \geq N \) is even, we have that
\[
|x_n - a| = |1 - a| < 1 .
\]
Thus we necessarily have the inequalities \( |1 + a| < 1 \) and \( |1 - a| < 1 \), so using the triangle inequality we have
\[
2 = |(1 + a) + (1 - a)| \leq |1 + a| + |1 - a| < 1 + 1
\]
which is a contradiction. Thus the sequence cannot converge to any \( a \in \mathbb{R} \), and hence diverges.
IV.1.3.11. In many books and references, the phrase “closer and closer” is used to informally describe the limiting process, i.e. \( x_n \to a \) if \( x_n \) gets “closer and closer” to \( a \) as \( n \to \infty \). But this phrasing is too imprecise to be an adequate definition, and is even not entirely correct. For example, as SAUNDERS MACLANE was fond of pointing out, the sequence \((1 + \frac{1}{n})\) gets closer and closer to 0 as \( n \to \infty \). The missing idea is the notion of eventually becoming \textit{arbitrarily} close. We might add that if, for example, \( x_n = 1 \) for all \( n \), then \( x_n \to 1 \), but \( x_n \) does not get “closer and closer” to 1 as \( n \) increases since one cannot get any closer to 1 than 1 itself. And, anyway, if \( x_n \to a \) the distance between \( x_n \) and \( a \) need not monotonically decrease as \( n \) increases, e.g. if \( x_n = \frac{1}{n^2} \) for \( n \) odd and \( x_n = \frac{1}{n} \) for \( n \) even (with \( a = 0 \)), so \( x_n \) does not really get “closer and closer” to \( a \). Thus it is best not to use this phrase at all even though it has some conceptual legitimacy.

IV.1.3.12. Indeed, the primary intellectual step forward (primarily due to CAUCHY, although reaching full fruition only with WEIERSTRASS) which made a satisfactory definition of limit possible was abandonment of the “dynamic” idea of approach, which previously dominated thought and discourse on the subject and was an apparently insurmountable obstacle to making a careful and rigorous definition, replacing it with a “static” one of simply being approximately equal within an arbitrarily specified tolerance. The dynamic point of view still has considerable intuitive value, however.

**Uniqueness of Limits**

We have used some suggestive language in Definition IV.1.3.6: we have referred to “the limit of a (convergent) sequence.” To justify this language, we must show that a sequence cannot have more than one limit:

IV.1.3.13. **Theorem.** The limit of a convergent sequence is unique: if \((x_n)\) is a sequence in \( \mathbb{R} \), \( a, b \in \mathbb{R} \), and \( x_n \to a \) and \( x_n \to b \), then \( a = b \).

The idea of the proof is simple: if \( I \) and \( J \) are intervals around \( a \) and \( b \) respectively, and both \( x_n \to a \) and \( x_n \to b \) are true, then for all sufficiently large \( n \) we must have \( x_n \in I \), and similarly for all sufficiently large \( n \) we have \( x_n \in J \). But if \( a \neq b \), then if \( I \) and \( J \) are short enough they are disjoint, so we cannot simultaneously have \( x_n \in I \) and \( x_n \in J \) for any \( n \).

The details of the proof give a good example of how the definition of limit is used, and also of some of the common basic techniques used in analysis proofs. A reader unfamiliar with analysis proofs should be sure to examine the proof carefully to understand how it works logically and how it reflects the informal outline of the previous paragraph.

**Proof:** We will prove the statement by contradiction. Suppose \( a \neq b \), i.e. \(|a - b| > 0\). Set \( \epsilon = \frac{1}{2}|a - b| \); then \( \epsilon > 0 \). Since \( x_n \to a \), there is an \( N_1 \) such that \( |x_n - a| < \epsilon \) for all \( n \geq N_1 \). Similarly, there is an \( N_2 \) such that \( |x_n - b| < \epsilon \) for all \( n \geq N_2 \). Set \( n = \max(N_1, N_2) \). Then \( |x_n - a| < \epsilon \) and \( |x_n - b| < \epsilon \) for this \( n \), so using the triangle inequality () we have

\[
|a - b| = |(a - x_n) + (x_n - b)| \leq |a - x_n| + |x_n - b| < \epsilon + \epsilon = 2\epsilon = \frac{2}{3}|a - b|
\]

and dividing both ends by \(|a - b|\), we obtain \( 1 < \frac{2}{3} \), which is a contradiction. ☞
IV.1.3.14. The “trick” of adding and subtracting the same thing within an absolute value and using the triangle inequality is one of the most important techniques in analysis, and will be seen in a great many arguments. (It is said that the difference between a “trick” and a “technique” is that a trick is a procedure which is used once, or a few times, and a technique a procedure which is used multiple times. This is definitely a technique under this distinction.)

IV.1.3.15. There is a subtle point connected with the proof of IV.1.3.13., which can be confusing to beginners. If I am trying to prove that $x_n \to a$, I am not allowed to choose the $\epsilon$; I must show that there is an $N$ that works for any $\epsilon$ someone else may decide to give me. But if I am assuming that $x_n \to a$ and want to use this to prove something else, as in the proof of IV.1.3.13., I am free to choose any $\epsilon$ I please, as long as it is positive, and I am assured (by my assumption) that there is an $N$ that works for this $\epsilon$.

**Boundedness and Convergence**

A convergent sequence is bounded:

IV.1.3.16. **Proposition.** Let $(x_n)$ be a convergent sequence in $\mathbb{R}$. Then $(x_n)$ is bounded: there is an $M \in \mathbb{R}$ such that $|x_n| \leq M$ for all $n$.

**Proof:** Suppose $x_n \to a$. Let $N$ be a natural number that works in the definition of limit for $\epsilon = 1$, i.e. $|x_n - a| < 1$ for all $n \geq N$. Then $a - 1 < x_n < a + 1$ for all $n \geq N$, so $|x_n| \leq |a| + 1$ for all $n \geq N$. Set

$$M = \max(|x_1|, |x_2|, \ldots, |x_{N-1}|, |a| + 1)$$

(this is a finite set of numbers, so the maximum exists). Then $|x_n| \leq M$ for all $n$. 

IV.1.3.17. The converse is false: a bounded sequence does not converge in general. A simple example is $((-1)^n)$. But a bounded sequence always has a convergent subsequence ( ).

A companion result, showing that the limit of a convergent sequence lies within any closed interval containing the sequence, is very important:

IV.1.3.18. **Theorem.** Let $(x_n)$ be a convergent sequence in $\mathbb{R}$, with $x_n \to c$. Then

(i) If $a \in \mathbb{R}$, $a < c$, then there is an $N \in \mathbb{N}$ such that $a < x_n$ for all $n \geq N$.

(ii) If $b \in \mathbb{R}$, $c < b$, then there is an $N \in \mathbb{N}$ such that $x_n < b$ for all $n \geq N$.

(iii) If $a \in \mathbb{R}$, $a \leq x_n$ for all $n$ or more generally if there is an $N$ such that $a \leq x_n$ for all $n \geq N$, then $a \leq c$.

(iv) If $b \in \mathbb{R}$, $x_n \leq b$ for all $n$ or more generally if there is an $N$ such that $x_n \leq b$ for all $n \geq N$, then $c \leq b$.

(v) If $a, b \in \mathbb{R}$, $a \leq x_n \leq b$ for all $n$ or more generally if there is an $N$ such that $a \leq x_n \leq b$ for all $n \geq N$, then $a \leq c \leq b$. 

418
Proof: (i): Set $\epsilon = c - a > 0$. Then there is an $N$ such that

$$a = c - \epsilon < x_n$$

for all $n \geq N$.

(ii): Nearly identical to (i), setting $\epsilon = b - c$.

(iii) and (iv) are the contrapositives of (ii) and (i) respectively, and (v) is a combination of (iii) and (iv).

**Caution:** The forms of the inequalities in this result are very important. In (i) and (ii) strict inequality in the hypothesis is essential. In (iii)-(v), if we have strict inequality in the hypothesis, the result applies, but we do not necessarily get strict inequality in the conclusion. For example, if $x_n = \frac{1}{n}$, then $0 < x_n$ for all $n$, but the limit is not positive.

Here is a simple corollary of (i) and (ii) which is useful:

**Corollary.** Let $(x_n)$ be a sequence of real numbers, and $c \in \mathbb{R}$. If $x_n \to c$ and $c \neq 0$, then there is an $N \in \mathbb{N}$ such that $x_n \neq 0$ for all $n \geq N$.

Proof: Apply IV.1.3.18. (i) with $a = 0$ if $c > 0$, and IV.1.3.18. (ii) with $b = 0$ if $c < 0$.

This corollary also has a simple direct proof which works in $\mathbb{C}$ as well as in $\mathbb{R}$: Set $\epsilon = \frac{|c|}{2}$ and choose $N$ as in the definition of limit for this $\epsilon$.

**Limits and Subsequences**

**Algebraic Operations and Limits**

The algebraic operations (addition, multiplication, division) behave nicely with respect to limits.

**Theorem.** Let $(x_n)$ and $(y_n)$ be sequences in $\mathbb{R}$. Suppose $x_n \to a$ and $y_n \to b$. Then

(i) $x_n + y_n \to a + b$.

(ii) $x_n y_n \to ab$.

(iii) If $y_n \neq 0$ for all $n$ and $b \neq 0$, then $\frac{x_n}{y_n} \to \frac{a}{b}$. 

419
IV.1.3.22. The theorem is often informally stated as “the limit of a sum [product, quotient] is the sum [product, quotient] of the limits,” and written

\[
\lim_{n \to \infty} (x_n + y_n) = \left( \lim_{n \to \infty} x_n \right) + \left( \lim_{n \to \infty} y_n \right)
\]

\[
\lim_{n \to \infty} (x_n y_n) = \left( \lim_{n \to \infty} x_n \right) \left( \lim_{n \to \infty} y_n \right)
\]

\[
\lim_{n \to \infty} \left( \frac{x_n}{y_n} \right) = \frac{\lim_{n \to \infty} x_n}{\lim_{n \to \infty} y_n}
\]

(provided \(\lim_{n \to \infty} y_n \neq 0\)).

However, stating the result this way can be misleading if one is not careful. The results only hold under the assumption that the sequences \((x_n)\) and \((y_n)\) converge, i.e. the limits on the right side exist; existence of the limits on the left in these expressions does not mean the limits on the right exist and that the limits on the left can be calculated in the indicated way. So care must be exercised not to try to apply the converse of the statements of IV.1.3.21.

The following special case of IV.1.3.21. (ii) (with \((y_n)\) a constant sequence) is used very frequently:

IV.1.3.23. Corollary. Let \((x_n)\) be a sequence of real numbers, and \(a \in \mathbb{R}\). If \(x_n \to a\), then for any \(c \in \mathbb{R}\) the sequence \((cx_n)\) converges to \(ca\). In other words, if \((x_n)\) converges, then \((cx_n)\) converges and

\[
\lim_{n \to \infty} (cx_n) = c \cdot \left( \lim_{n \to \infty} x_n \right)
\]

IV.1.3.24. We first give the proof of IV.1.3.21.(i). We will give the reasoning steps in great detail since the technique is a very common one which can be confusing to beginners; note the comments in IV.1.3.15.

Let \(\epsilon > 0\) be given. We must find an \(N\) (i.e. show that one exists) such that

\[
| (x_n + y_n) - (a + b) | < \epsilon
\]

for all \(n \geq N\). We assume that \(x_n \to a\), so for any \(\epsilon_1 > 0\) (not necessarily the given \(\epsilon\)) there is an \(N_1\) such that \(|x_n - a| < \epsilon_1\) for all \(n \geq N_1\). Similarly, since \(y_n \to b\), for any \(\epsilon_2 > 0\) (not necessarily the same as either \(\epsilon\) or \(\epsilon_1\)) there is an \(N_2\) such that \(|y_n - b| < \epsilon_2\) for all \(n \geq N_2\). Thus, if we fix an \(\epsilon_1\) and an \(\epsilon_2\) and take the corresponding \(N_1\) and \(N_2\), and set \(N = \max(N_1, N_2)\), we have

\[
| (x_n + y_n) - (a + b) | = | (x_n - a) + (y_n - b) | \leq |x_n - a| + |y_n - b| < \epsilon_1 + \epsilon_2
\]

for all \(n \geq N\). We are given \(\epsilon\), but we are free to choose \(\epsilon_1\) and \(\epsilon_2\) however we please, and there are automatically corresponding \(N_1\) and \(N_2\), hence an \(N\) that works for the limit of the sum for \(\epsilon_1 + \epsilon_2\). All we need to do is choose \(\epsilon_1\) and \(\epsilon_2\) so that \(\epsilon_1 + \epsilon_2 \leq \epsilon\), and the process gives us an \(N\) that works for the limit of the sum for the given \(\epsilon\). The simplest thing is just to choose

\[
\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}
\]

420
All of this could be regarded as preliminary “scratch work” giving us the idea of the proof. The way we would write the proof formally is:

**Proof:** Let $\epsilon > 0$. Since $x_n \to a$, there is an $N_1$ such that $|x_n - a| < \frac{\epsilon}{2}$ for all $n \geq N_1$. Since $y_n \to b$, there is an $N_2$ such that $|y_n - b| < \frac{\epsilon}{2}$ for all $n \geq N_2$. Set $N = \max(N_1, N_2)$. Then, for all $n \geq N$,

$$|(x_n + y_n) - (a + b)| = |(x_n - a) + (y_n - b)| \leq |x_n - a| + |y_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and so, since $\epsilon > 0$ is arbitrary, $x_n + y_n \to a + b$.

**IV.1.3.25.** Now we turn to the proof of **IV.1.3.21.(ii).** The argument is somewhat similar, but more complicated. We go through the reasoning more quickly.

As before, for any $\epsilon_1$ and $\epsilon_2$ we have an $N$ such that $|x_n - a| < \epsilon_1$ and $|y_n - b| < \epsilon_2$ for all $n \geq N$. Then, by the adding and subtracting technique, we have

$$|x_ny_n - ab| = |(x_ny_n - ay_n) + (ay_n - ab)| \leq |x_ny_n - ay_n| + |ay_n - ab| = |y_n||x_n - a| + |a||y_n - b| < |y_n|\epsilon_1 + |a|\epsilon_2$$

for all $n \geq N$. Thus, if we are given an $\epsilon > 0$, we want to choose $\epsilon_1$ and $\epsilon_2$ so that

$$|y_n|\epsilon_1 + |a|\epsilon_2 \leq \epsilon$$

for all $n$. We might try

$$\epsilon_1 = \frac{\epsilon}{2|y_n|} \quad \text{and} \quad \epsilon_2 = \frac{\epsilon}{2|a|}$$

but there is a serious problem with the first of these since $\epsilon_1$ cannot depend on $n$. (There is a slight problem with the other one too since we could have $a = 0$.)

Fortunately, there is an easy way past the difficulty. Since $(y_n)$ converges, it is bounded (**IV.1.3.16.**); hence there is an $M$ such that $|y_n| \leq M$ for all $n$. We might as well choose $M$ so that $|a| \leq M$ to ease notation; and we can assume $M > 0$ since there is no penalty in increasing $M$ (and the case where we can take $M = 0$ is pretty trivial anyway). Then we can take

$$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2M}$$

for this $M$. Thus the formal proof is:

**Proof:** Let $\epsilon > 0$. Since $(y_n)$ converges, there is an $M$ such that $|y_n| \leq M$ for all $n$ by **IV.1.3.16.**; we may also assume $M > 0$ and $|a| \leq M$ by increasing $M$ if necessary. Since $x_n \to a$, there is an $N_1$ such that $|x_n - a| < \frac{\epsilon}{2M}$ for all $n \geq N_1$, and since $y_n \to b$ there is an $N_2$ such that $|y_n - b| < \frac{\epsilon}{2M}$ for all $n \geq N_2$. Set $N = \max(N_1, N_2)$. If $n \geq N$, then

$$|x_ny_n - ab| = |(x_ny_n - ay_n) + (ay_n - ab)| \leq |x_ny_n - ay_n| + |ay_n - ab| = |y_n||x_n - a| + |a||y_n - b|$$

$$\leq M|x_n - a| + M|y_n - b| < M\frac{\epsilon}{2M} + M\frac{\epsilon}{2M} = \epsilon$$

and so, since $\epsilon > 0$ is arbitrary, $x_ny_n \to ab$. $\diamondsuit$
Finally, we prove (iii). It will actually suffice to prove the following special case:

**Lemma.** Let \((y_n)\) be a convergent sequence of real numbers, \(y_n \to b\), with all \(y_n \neq 0\) and \(b \neq 0\). Then \(\frac{1}{y_n} \to \frac{1}{b}\).

Then (iii) follows by combining the lemma with (ii). (Note also that the hypothesis in (iii) that \(y_n \neq 0\) for all \(n\) is almost automatic: by IV.1.3.20., \(y_n \neq 0\) for all sufficiently large \(n\) if \(b \neq 0\).)

To prove the lemma, note that

\[
\frac{1}{y_n} - \frac{1}{b} = \frac{b - y_n}{by_n}
\]

for all \(n\). Thus we have

\[
\frac{1}{|y_n - b|} = \frac{1}{|by_n|} |y_n - b|
\]

for all \(n\). If we can show that the sequence \(\left(\frac{1}{|y_n|}\right)\) is bounded, we can easily proceed. Note that since \(b \neq 0\), there is an \(N_1\) such that \(|y_n - b| < \frac{|b|}{2}\) for all \(n \geq N_1\), and thus \(|y_n| > \frac{|b|}{2}\) for \(n \geq N_1\). Set

\[
M = \max\left(\frac{1}{|by_1|}, \ldots, \frac{1}{|by_{N_1-1}|}, \frac{2}{|b|^2}\right)
\]

and note that \(\frac{1}{|y_n|} \leq M\) for all \(n\) (and that \(M > 0\)).

Now suppose \(\epsilon > 0\) is given. There is an \(N\) such that

\[
|y_n - b| < \frac{\epsilon}{M}
\]

for all \(n \geq N\). Then, for \(n \geq N\),

\[
\frac{1}{|y_n - b|} = \frac{1}{|by_n|} |y_n - b| \leq M \frac{\epsilon}{M} = \epsilon
\]

**IV.1.3.27.** Note that IV.1.3.21.(iii) does not apply if \(b = 0\), even if \(y_n \neq 0\) for all \(n\). It turns out that if \(b = 0\) and \(a \neq 0\), then \(\lim \limits_{n \to \infty} \left(\frac{x_n}{y_n}\right)\) cannot exist (since if it did, and equaled \(c\), we would have

\[
a = \lim \limits_{n \to \infty} \left(\frac{x_n}{y_n}\right) = \left(\lim \limits_{n \to \infty} y_n\right) \left(\lim \limits_{n \to \infty} \left(\frac{x_n}{y_n}\right)\right) = bc
\]

by (ii), which is impossible). But if \(a = b = 0\), the limit of the quotient sequence can exist, but cannot be computed simply in general. This is a type of indeterminate form. One of the important parts of calculus and analysis is the resolution of indeterminate forms of various types; see (), (), . . .

**Limits in \(\mathbb{C}\)**

Limits in \(\mathbb{C}\) behave essentially identically to those in \(\mathbb{R}\), although the geometric interpretation is a little different (it is essentially like the interpretation in \(\mathbb{R}^n\) described in ()). The definition of convergence in \(\mathbb{C}\) is identical to the final version (IV.1.3.6.) with \(\mathbb{R}\) replaced by \(\mathbb{C}\) throughout and no other changes:
IV.1.3.28. **Definition.** [Limit of a Complex Sequence] Let \((z_n)\) be a sequence in \(\mathbb{C}\), and \(c \in \mathbb{C}\). The sequence \((z_n)\) *converges* to \(c\), or \(c\) is the *limit* of the sequence \((z_n)\), if for every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[ |z_n - c| < \epsilon \]

for all \(n \geq N\).

The sequence \((z_n)\) *converges* if there is a complex number \(c\) such that \(c\) is the limit of the sequence \((z_n)\). If there is no such \(c\), the sequence \((z_n)\) *diverges*.

IV.1.3.29. All the results about real sequences except those depending on the ordering of \(\mathbb{R}\) (e.g. IV.1.3.18. and the results on monotone sequences) hold for sequences in \(\mathbb{C}\), with identical proofs. As an alternate approach, we may reduce the theory of sequences in \(\mathbb{C}\) to the theory for \(\mathbb{R}\) by considering real and imaginary parts.

IV.1.3.30. **Proposition.** Let \((z_n)\) be a sequence in \(\mathbb{C}\). For each \(n\), write \(z_n = x_n + iy_n\) with \(x_n, y_n \in \mathbb{R}\). Then \((x_n)\) and \((y_n)\) are real sequences. If \(c = a + ib \in \mathbb{C}\) with \(a, b \in \mathbb{R}\), then \(z_n \to c\) if and only if \(x_n \to a\) and \(y_n \to b\).

**Proof:** Suppose \(z_n \to c\). Let \(\epsilon > 0\). There is an \(N\) such that \(|z_n - c| < \epsilon\) for all \(n \geq N\). Since the real and imaginary parts of a complex number are not larger than the absolute value,

\[ |x_n - a| \leq |x_n - a + i(y_n - b)| = |z_n - c| < \epsilon \]

for all \(n \geq N\), so \(x_n \to a\). Similarly, \(y_n \to b\).

Conversely, suppose \(x_n \to a\) and \(y_n \to b\). Let \(\epsilon > 0\). Then there is an \(N_1\) such that

\[ |x_n - a| < \frac{\epsilon}{2} \]

for all \(n \geq N_1\), and there is an \(N_2\) such that

\[ |y_n - b| < \frac{\epsilon}{2} \]

for all \(n \geq N_2\). Set \(N = \max(N_1, N_2)\). Then, if \(n \geq N\),

\[ |z_n - c| = |(x_n - a) + i(y_n - b)| \leq |x_n - a| + |y_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \]

and thus \(z_n \to c\).

IV.1.4. **Infinite Limits**

IV.1.5. **Monotone Sequences**

IV.1.6. **Subsequences and Cluster Points of Sequences**

“Most” sequences, even most bounded sequences, do not converge.
The Bolzano-Weierstrass Theorem

IV.1.7. Limit Superior and Limit Inferior

In this subsection, we define two generalizations of limit for sequences of real numbers, limit superior and limit inferior. Every real sequence has a limit superior and a limit inferior, even sequences which do not converge; in fact, these notions are most interesting for nonconvergent sequences (they both just reduce to the usual limit for convergent sequences).

These notions make essential use of the ordering on \(\mathbb{R}\), and thus do not generalize to \(\mathbb{C}\), \(\mathbb{R}^n\), or general metric spaces.

Limit Superior and Limit Inferior for Bounded Sequences

We first consider only bounded sequences for simplicity. We give the formal definitions, then explain them.

IV.1.7.1. Definition. Let \((x_n)\) be a bounded sequence in \(\mathbb{R}\). Define

\[
\limsup_{n \to \infty} x_n = \lim_{n \to \infty} \left[ \sup_{k \geq n} x_k \right] = \inf_n \left[ \sup_{k \geq n} x_k \right]
\]

\[
\liminf_{n \to \infty} x_n = \lim_{n \to \infty} \left[ \inf_{k \geq n} x_k \right] = \sup_n \left[ \inf_{k \geq n} x_k \right]
\]

called the limit superior and limit inferior of \((x_n)\) respectively.

IV.1.7.2. To explain these definitions, first note that the set \(\{x_k : k \geq n\}\) is a bounded nonempty set of numbers for any \(n\), hence has a supremum \(s_n\) and infimum \(t_n\). As \(n\) increases, \(s_n\) decreases and \(t_n\) increases since the set \(\{x_k : k \geq n\}\) decreases. Thus \((s_n)\) is a bounded decreasing (nonincreasing) sequence, hence converges to its infimum; similarly, \((t_n)\) increases and converges to its supremum. These limits are \(\limsup_{n \to \infty} x_n\) and \(\liminf_{n \to \infty} x_n\) respectively.

The \(\limsup_{n \to \infty} x_n\) and \(\liminf_{n \to \infty} x_n\) are often written \(\limsup_{n \to \infty} x_n\) and \(\liminf_{n \to \infty} x_n\) respectively. In all notations the “\(n \to \infty\)” is frequently omitted whenever there is no possibility of confusion.

IV.1.7.3. Examples. (i) Let \(x_n = (-1)^n\). For each \(n\) we have \(\sup_{k \geq n} x_k = 1\) and \(\inf_{k \geq n} x_k = -1\), so \(\limsup_{n \to \infty} x_n = 1\) and \(\liminf_{n \to \infty} x_n = -1\).

(ii) Let \(x_n = (-1)^n \left(1 + \frac{1}{n}\right)\). Then for \(n\) odd we have \(\sup_{k \geq n} x_k = 1 + \frac{1}{n+1}\) and \(\inf_{k \geq n} x_k = -1 - \frac{1}{n+1}\), and for \(n\) even we have \(\sup_{k \geq n} x_k = 1 + \frac{1}{n}\) and \(\inf_{k \geq n} x_k = -1 - \frac{1}{n+1}\), so \(\limsup_{n \to \infty} x_n = 1\) and \(\liminf_{n \to \infty} x_n = -1\).

(iii) Let \(x_n = \sin n\). Then \(-1 < x_n < 1\) for all \(n\). There is a sequence \((n_j)\) in \(\mathbb{N}\) with \(\sin n_j \to 1\) (), so for any \(n\) we have \(\sup_{k \geq n} x_k = 1\) and \(\limsup_{n \to \infty} x_n = 1\). Similarly, there is a sequence \((m_j)\) in \(\mathbb{N}\) with \(\sin m_j \to -1\), so for any \(n\) we have \(\inf_{k \geq n} x_k = -1\) and \(\liminf_{n \to \infty} x_n = -1\).
IV.1.7.4. We obviously have $\lim\inf x_n \leq \lim\sup x_n$ since $t_n \leq s_n$ for each $n$. We will show (IV.1.7.9.) that $\lim\sup x_n$ and $\lim\inf x_n$ are just the largest and smallest cluster points () of the sequence $(x_n)$, which is usually the best way to think of them. To begin with, we make the following observation:

IV.1.7.5. **Proposition.** Let $(x_n)$ be a bounded sequence in $\mathbb{R}$, and let $a \in \mathbb{R}$.

(i) If for every $\epsilon > 0$ there are only finitely many $x_n$ satisfying $x_n > a + \epsilon$, then $\lim\sup x_n \leq a$.

(ii) If for every $\epsilon > 0$ there are infinitely many $x_n$ satisfying $x_n > a - \epsilon$, then $\lim\sup x_n \geq a$.

(iii) If for every $\epsilon > 0$ there are only finitely many $x_n$ satisfying $x_n < a - \epsilon$, then $\lim\inf x_n \geq a$.

(iv) If for every $\epsilon > 0$ there are infinitely many $x_n$ satisfying $x_n < a + \epsilon$, then $\lim\inf x_n \leq a$.

**Proof:**

(i): Let $\epsilon > 0$, and fix $N$ so that $x_k \leq a + \epsilon$ for all $k \geq N$. Then, for all $n \geq N$, $s_n = \sup_{k \geq n} x_k \leq a + \epsilon$, so $\lim\sup x_n = \lim_{n \to \infty} s_n \leq a + \epsilon$. Since this is true for every $\epsilon > 0$, $\lim\sup x_n \leq a$.

(ii) For any $\epsilon > 0$, we have $s_n > a - \epsilon$ for every $n$, so each $s_n$ satisfies $s_n \geq a$ and thus $\inf_n s_n \geq a$.

(iii) and (iv) are essentially identical with inequalities reversed.

A simple rephrasing is often useful:

IV.1.7.6. **Corollary.** Let $(x_n)$ be a bounded sequence. Then

(i) $\lim\sup x_n = \inf \{b \in \mathbb{R} : x_n < b \text{ for all sufficiently large } n\}$.

(ii) $\lim\inf x_n = \sup \{a \in \mathbb{R} : x_n > a \text{ for all sufficiently large } n\}$.

**Proof:** For (i), let $B = \inf \{b \in \mathbb{R} : x_n < b \text{ for all sufficiently large } n\}$ and $L = \lim\sup x_n$. Then we have $L \leq B$ by IV.1.7.5.(i) and $L \geq B$ by IV.1.7.5.(ii). The proof of (ii) is analogous.

The next result gives a useful way of thinking of $\lim\sup$ and $\lim\inf$.

IV.1.7.7. **Theorem.** Let $(x_n)$ be a bounded sequence.

(i) Let $L = \lim\sup x_n$. Then $L$ satisfies

(a) For any $\epsilon > 0$ there are infinitely many $n$ such that $x_n > L - \epsilon$

(b) For any $\epsilon > 0$ there are only finitely many $n$ such that $x_n > L + \epsilon$

and $L$ is the unique real number with these two properties.

(ii) Let $l = \lim\inf x_n$. Then $l$ satisfies

(a) For any $\epsilon > 0$ there are infinitely many $n$ such that $x_n < l + \epsilon$
(b) For any \( \epsilon > 0 \) there are are only finitely many \( n \) such that \( x_n < l - \epsilon \) and \( l \) is the unique real number with these two properties.

**Proof:** (i): By IV.1.7.5. (i)–(ii), any number satisfying (ia)–(ib) must be equal to \( L \). On the other hand, if \( \epsilon > 0 \), then \( s_n = \sup_{k \geq n} x_k \) satisfies \( s_n \geq L \) for all \( n \), so for each \( n \) there must be a \( k \geq n \) with \( x_k > L - \epsilon \), i.e. \( x_k > L - \epsilon \) for infinitely many \( k \). But if \( x_k > L + \epsilon \) for infinitely many \( k \), then \( s_n > L + \epsilon \) for all \( n \), and thus \( L = \inf_n s_n \geq L + \epsilon \), a contradiction.

(ii): Almost identical with inequalities reversed. 🍊

**IV.1.7.8. Corollary.** Let \((x_n)\) be a bounded sequence in \( \mathbb{R} \). Then \( \lim \inf_{n \to \infty} x_n = \lim \sup_{n \to \infty} x_n \) if and only if \( \lim_{n \to \infty} x_n \) exists, in which case

\[
\lim \inf_{n \to \infty} x_n = \lim \sup_{n \to \infty} x_n = \lim_{n \to \infty} x_n.
\]

**Proof:** If \( \lim_{n \to \infty} x_n \) exists and equals \( L \), then \( L \) satisfies all four conditions, so

\[
L = \lim \inf_{n \to \infty} x_n = \lim \sup_{n \to \infty} x_n.
\]

Conversely, if \( \lim \inf_{n \to \infty} x_n = \lim \sup_{n \to \infty} x_n = L \), then \( L \) satisfies (ib) and (iib), hence \( \lim_{n \to \infty} x_n \) exists and equals \( L \). 🍊

Finally, we obtain the characterization of \( \lim \sup \) and \( \lim \inf \) in terms of cluster points:

**IV.1.7.9. Theorem.** Let \((x_n)\) be a bounded sequence. Then \( \lim \sup_{n \to \infty} x_n \) is the largest cluster point of \((x_n)\) and \( \lim \inf_{n \to \infty} x_n \) is the smallest cluster point of \((x_n)\).

**Proof:** Let \( L = \lim \sup_{n \to \infty} x_n \). If \( c \) is a cluster point of \((x_n)\), then for any \( \epsilon > 0 \) there are infinitely many \( x_n \) satisfying \( x_n > c - \epsilon \), so \( c \leq L \) by IV.1.7.5.(ii). On the other hand, for any \( \epsilon > 0 \) there are infinitely many \( x_n \) satisfying \( x_n > L - \epsilon \) by IV.1.7.7.(ia), and all but finitely many of these also satisfy \( x_n < L + \epsilon \) by IV.1.7.7.(ib), so \( L \) is a cluster point of \((x_n)\). The proof for \( \lim \inf \) is analogous. 🍊

**Tail Dependence**

If \((x_n)\) is a bounded sequence, then \( \lim \sup_{n \to \infty} x_n \) and \( \lim \inf_{n \to \infty} x_n \), like \( \lim_{n \to \infty} x_n \) if it exists, depend only on the tails of the sequence. We state two precise versions of this, whose simple proofs are left to the reader ()().

**IV.1.7.10. Proposition.** Let \((x_n)\) be a bounded sequence, and \((y_n)\) any tail sequence of \((x_n)\). Then \( \lim \sup_{n \to \infty} x_n = \lim \sup_{n \to \infty} y_n \) and \( \lim \inf_{n \to \infty} x_n = \lim \inf_{n \to \infty} y_n \).
**IV.1.7.11.** Proposition. Let \((x_n)\) and \((y_n)\) be bounded sequences, with \(x_n = y_n\) for all sufficiently large \(n\). Then \(\limsup_{n \to \infty} x_n = \limsup_{n \to \infty} y_n\) and \(\liminf_{n \to \infty} x_n = \liminf_{n \to \infty} y_n\).

We can improve the last result:

**IV.1.7.12.** Proposition. Let \((x_n)\) and \((y_n)\) be bounded sequences, with \(\lim_{n \to \infty} (x_n - y_n) = 0\). Then \(\limsup_{n \to \infty} x_n = \limsup_{n \to \infty} y_n\) and \(\liminf_{n \to \infty} x_n = \liminf_{n \to \infty} y_n\).

A direct proof can be given, but the result is an immediate corollary of **IV.1.7.17.**

**Algebraic and Order Properties**

We begin with two simple observations, whose proofs are left to the reader. The first generalizes **IV.1.7.13.**:

**IV.1.7.13.** Proposition. Let \((x_n)\) and \((y_n)\) be bounded sequences, with \(x_n \leq y_n\) for all sufficiently large \(n\). Then \(\limsup_{n \to \infty} x_n \leq \limsup_{n \to \infty} y_n\) and \(\liminf_{n \to \infty} x_n \leq \liminf_{n \to \infty} y_n\).

The next observation gives a symmetry between \(\limsup\) and \(\liminf\):

**IV.1.7.14.** Proposition. Let \((x_n)\) be a bounded sequence. Then \(\limsup_{n \to \infty} (-x_n) = -\liminf_{n \to \infty} x_n\) and \(\liminf_{n \to \infty} (-x_n) = -\limsup_{n \to \infty} x_n\).

**IV.1.7.15.** The limit of a sum of two convergent sequences is the sum of the limits (). However, the analogous statement does not hold for \(\limsup\) and \(\liminf\). For example, let \(x_n = (-1)^n\) and \(y_n = (-1)^{n+1}\) for all \(n\). Then \(\limsup_{n \to \infty} x_n = \limsup_{n \to \infty} y_n = 1\) and \(\liminf_{n \to \infty} x_n = \liminf_{n \to \infty} y_n = -1\). But \(x_n + y_n = 0\) for all \(n\), so
\[
\limsup_{n \to \infty} (x_n + y_n) = 0 \neq 2 = \limsup_{n \to \infty} x_n + \limsup_{n \to \infty} y_n
\]
\[
\liminf_{n \to \infty} (x_n + y_n) = 0 \neq -2 = \liminf_{n \to \infty} x_n + \liminf_{n \to \infty} y_n.
\]

However, we do have inequalities:

**IV.1.7.16.** Theorem. Let \((x_n)\) and \((y_n)\) be bounded sequences. Then
\[
\liminf_{n \to \infty} x_n + \liminf_{n \to \infty} y_n \leq \liminf_{n \to \infty} (x_n + y_n) \leq \limsup_{n \to \infty} x_n + \limsup_{n \to \infty} y_n \leq \limsup_{n \to \infty} (x_n + y_n) \leq \limsup_{n \to \infty} x_n + \limsup_{n \to \infty} y_n.
\]

Of course, the \(x_n\) and \(y_n\) may be interchanged in the middle, i.e. the middle expression can be replaced by \(\liminf_{n \to \infty} x_n + \limsup_{n \to \infty} y_n\).
Proof: For the first inequality, let \( l = \liminf_{n \to \infty} x_n \) and \( l' = \liminf_{n \to \infty} y_n \). Let \( \epsilon > 0 \). Then there is an \( N \) such that \( x_n > l - \frac{\epsilon}{2} \) and \( y_n > l' - \frac{\epsilon}{2} \) for all \( n \geq N \). Then \( x_n + y_n > l + l' - \epsilon \) for all \( n \geq N \), so \( \liminf_{n \to \infty} (x_n + y_n) \geq l + l' \) by IV.1.7.5.(iii).

For the second inequality, let \( L = \limsup_{n \to \infty} x_n \) and \( l' = \liminf_{n \to \infty} y_n \). Let \( \epsilon > 0 \). Then there is an \( N \) such that \( x_n < L + \frac{\epsilon}{2} \) for all \( n \geq N \). For infinitely many such \( n \), \( y_n < l' + \frac{\epsilon}{2} \). Thus for infinitely many \( n \), \( x_n + y_n < L + l' + \epsilon \), and thus \( \liminf_{n \to \infty} (x_n + y_n) \leq L + l' \) by IV.1.7.5.(iv).

The proofs of the other two inequalities are analogous.

We do get equalities if one of the sequences converges:

IV.1.7.17. Corollary. Let \( (x_n) \) be a bounded sequence and \( (y_n) \) a convergent sequence. Then

\[
\begin{align*}
\limsup_{n \to \infty} (x_n + y_n) &= \limsup_{n \to \infty} x_n + \lim_{n \to \infty} y_n \\
\liminf_{n \to \infty} (x_n + y_n) &= \liminf_{n \to \infty} x_n + \lim_{n \to \infty} y_n.
\end{align*}
\]

Proof: For the first equality, we have

\[
\limsup_{n \to \infty} x_n + \lim_{n \to \infty} y_n = \limsup_{n \to \infty} x_n + \limsup_{n \to \infty} y_n \leq \limsup_{n \to \infty} (x_n + y_n) \leq \limsup_{n \to \infty} x_n + \limsup_{n \to \infty} y_n = \limsup_{n \to \infty} x_n + \lim_{n \to \infty} y_n.
\]

The second is analogous, using the remark at the end of Theorem IV.1.7.16.

Unbounded Sequences

IV.1.7.18. The same definition of \( \limsup \) and \( \liminf \) makes sense, properly interpreted, for unbounded sequences of real numbers and even extended real numbers; in this case the \( \limsup \) and/or the \( \liminf \) may take the values \( +\infty \), as well as real-number values. It is simplest to just note the following criteria for when they take infinite values:

IV.1.7.19. Proposition. Let \( (x_n) \) be a sequence of real numbers. Then

(i) \( \limsup_{n \to \infty} x_n = +\infty \) if and only \( (x_n) \) is not bounded above.

(ii) \( \liminf_{n \to \infty} x_n = -\infty \) if and only \( (x_n) \) is not bounded below.

Proof: (i): \( \limsup_{n \to \infty} x_n < +\infty \) if and only if \( s_n = \sup_{k \geq n} x_k \) is finite for some \( n \), i.e. a tail of \( (x_n) \) is bounded above. But this implies that the entire sequence is bounded above. Conversely, if \( x_n \leq M \) for all \( n \), then \( s_n \leq M \) for all \( n \) and \( \limsup_{n \to \infty} x_n \leq M \).

The proof of (ii) is analogous.

428
IV.1.7.20. Proposition. Let \((x_n)\) be a sequence of real numbers. Then

(i) \(\lim_{n \to \infty} \sup x_n = -\infty\) if and only if \(\lim_{n \to \infty} x_n = -\infty\) (in the sense of ()).

(ii) \(\lim_{n \to \infty} \inf x_n = +\infty\) if and only if \(\lim_{n \to \infty} x_n = +\infty\).

The simple proof is left to the reader.

IV.1.7.21. All the results about \(\lim \sup\) and \(\lim \inf\) for bounded sequences, suitably interpreted, hold also for unbounded sequences, although some of the proofs need to be modified. Details are left to the reader.

IV.1.8. Cauchy Sequences

IV.1.9. Sequences in \(\mathbb{R}^n\)

IV.1.10. Sequences in Metric Spaces

IV.1.11. Exercises

IV.1.11.1. Show that in the definition of limit (IV.1.3.6., ??) it is unimportant whether strict inequality is used in two places: if \((x_n)\) is a sequence in \(\mathbb{R}\), then the following conditions are equivalent to the condition in IV.1.3.6.

(i) For every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[|x_n - a| \leq \epsilon\]

for all \(n \geq N\). [For one direction, decrease \(\epsilon\) before finding \(N\).]

(ii) For every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[|x_n - a| < \epsilon\]

for all \(n > N\). [For one direction, replace \(N\) by \(N + 1\).]

(iii) For every \(\epsilon > 0\), there is an \(N \in \mathbb{N}\) such that

\[|x_n - a| \leq \epsilon\]

for all \(n > N\).

(iv) Show that the condition \(N \in \mathbb{N}\) can be replaced by \(N \in \mathbb{R}\) without changing the definition of limit, i.e. that the following condition is equivalent to the condition in IV.1.3.6.: For every \(\epsilon > 0\), there is an \(N \in \mathbb{R}\) such that

\[|x_n - a| < \epsilon\]

for all \(n \in \mathbb{N}\), \(n \geq N\). [For one direction, use the Archimedean property of \(\mathbb{R}\) ().]

Do the same for the conditions in (i)–(iii). Show analogous results for limits of sequences in metric spaces.

IV.1.11.2. Let \((x_n)\) and \((y_n)\) be sequences converging to the same limit \(a\). Alternate or shuffle the \(x_n\) and \(y_n\) into a single sequence \((z_n)\) by taking \(z_n = x_m\) if \(n = 2m - 1\) is odd, and \(z_n = y_m\) if \(n = 2m\) is even. Give two proofs that \(z_n \to a\), one directly from the definition of limit and one from ()

429
IV.1.11.3. Prove the following theorem ([?]; cf. []):

**Theorem.**

(i) Let \((\alpha_n)\) be a sequence of positive numbers satisfying \(\alpha_{n+m} \leq \alpha_n \alpha_m\) for all \(n\) and \(m\). Set \(\beta_n = \alpha_n^{1/n}\). Then \(\lim_{n \to \infty} \beta_n\) exists (as a real number) and

\[
\lim_{n \to \infty} \beta_n = \inf_n \beta_n.
\]

(ii) Let \((a_n)\) be a sequence of real numbers satisfying \(a_{n+m} \leq a_n + a_m\) for all \(n\) and \(m\). Set \(b_n = \frac{a_n}{n}\). Then \(\lim_{n \to \infty} b_n\) exists as a real number or \(-\infty\), and

\[
\lim_{n \to \infty} b_n = \inf_n b_n.
\]

(iii) Let \((a_n)\) be a sequence of real numbers satisfying \(a_{n+m} \geq a_n + a_m\) for all \(n\) and \(m\). Set \(b_n = \frac{a_n}{n}\). Then \(\lim_{n \to \infty} b_n\) exists as a real number or \(+\infty\), and

\[
\lim_{n \to \infty} b_n = \sup_n b_n.
\]

(a) Show that (i), (ii), and (iii) are equivalent. So it suffices to prove (iii).

(b) In (iii), replacing \(a_n\) by \(a_n - na_1\) for all \(n\), it suffices to assume \(a_1 = 0\).

(c) For any \(k, m \in \mathbb{N}\), \(a_{km} \geq ka_m\), and hence \(b_{km} \geq b_m\). In particular, \(a_n \geq na_1 = 0\) for all \(n\), and hence \(b_n \geq 0\) for all \(n\).

(d) Fix \(m \in \mathbb{N}\). For \(k \in \mathbb{N}\) and \(1 \leq r \leq m\),

\[a_{km+r} \geq a_{km} + a_r \geq a_{km}\]

and thus

\[
b_{km+r} = \frac{a_{km+r}}{km+r} \geq \frac{a_{km}}{(k+1)m} = \frac{k}{k+1} \cdot b_{km} \geq \frac{k}{k+1} \cdot b_m.
\]

(e) Conclude that

\[
\lim_{n \to \infty} \inf b_n \geq \sup_n b_n.
\]

IV.1.11.4. Let \((a_n)\) and \((b_n)\) be sequences of real numbers, with \((b_n)\) a strictly increasing sequence of nonzero numbers with \(\lim_{n \to \infty} b_n = +\infty\). Suppose

\[
\lim_{n \to \infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}
\]

exists (in either the ordinary or extended sense) and equals \(L\). Show that

\[
\lim_{n \to \infty} \frac{a_n}{b_n}
\]

also exists (in the same sense) and equals \(L\). [Show that for every \(\epsilon > 0\) there is an \(m\) such that

\[(L - \epsilon)(b_n - b_m) < a_n - a_m < (L + \epsilon)(b_n - b_m)\]

for all \(n > m\) by adding a telescoping sum.]

This result can be regarded as a discrete version of l’Hôpital’s Rule Version 2d’ (V.9.2.11.).

430
IV.2. Infinite Series

Infinite series (first of numbers, then of functions) form one of the most important and useful topics in analysis.

An infinite series is an expression of the form

$$\sum_{k=1}^{\infty} a_k$$

(or slight variations of this general form, e.g. we may start the summation at 0 or another integer), where the \(a_k\), called the terms of the series, are numbers or (later) functions.

The summation notation of infinite series suggests that the series represents a sum of numbers (or functions), hence has a numerical value. But an infinite series is, at least at first, just a formal expression and we do not necessarily associate to it a numerical value. We can associate a numerical “sum” to many series, but not all (the ones which have an associated sum are called convergent, and the ones which do not are called divergent). These “sums,” however, do not always behave like finite sums. The following principle must be kept in mind at all times in this subject:

First Fundamental Principle of Infinite Series: An infinite series is NOT a finite sum and cannot be treated like one!

The following chart lists the main differences between infinite series and finite sums:

<table>
<thead>
<tr>
<th>Finite Sum</th>
<th>Infinite Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Always represents a number</td>
<td>May or may not converge</td>
</tr>
<tr>
<td>Purely algebraic expression</td>
<td>Requires limits and analysis to define sum</td>
</tr>
<tr>
<td>Associative and commutative laws hold</td>
<td>Laws of arithmetic do not hold in general</td>
</tr>
</tbody>
</table>

A lot of the theory of infinite series is devoted to establishing that under certain conditions, infinite series can be manipulated and evaluated roughly like finite sums. But operations on infinite series which would be routine or trivial for finite sums must always be carefully justified using the results of the theory. In finding results or formulas, it can be very useful to manipulate infinite series formally as though they were finite sums, but formulas obtained in this way cannot be considered valid unless justified by careful applications of theorems about infinite series; some such formulas turn out to be correct and others don’t.

IV.2.1. Partial Sums and Convergence

IV.2.1.1. Suppose \(\sum_{k=1}^{\infty} a_k\) is an infinite series. We can associate to this series two sequences. One obvious sequence is the sequence \((a_k)\) of terms. The other, much more important, is the sequence \((s_n)\) of partial sums:

IV.2.1.2. Definition. Let \(\sum_{k=1}^{\infty} a_k\) be an infinite series, and \(n \in \mathbb{N}\). The \(n^{th}\) partial sum of the series is the finite sum

$$s_n = \sum_{k=1}^{n} a_k .$$
IV.2.1.3. Conversely, if \((s_n)\) is any sequence of real numbers, there is a unique infinite series \(\sum_{k=1}^{\infty} a_k\) such that \((s_n)\) is the sequence of partial sums: set \(a_1 = s_1\) and \(a_k = s_k - s_{k-1}\) for \(k > 1\). Thus in a sense the theory of infinite series is identical to the theory of sequences, although we usually take a somewhat different point of view. (This correspondence can be regarded as a discrete version of the Fundamental Theorem of Calculus.)

Since the partial sums of an infinite series are finite sums, the usual rules of arithmetic and algebra can be used with them. We can also use facts and results about sequences, applied to the sequence of partial sums, to analyze series. In particular, we can define convergence for an infinite series:

IV.2.1.4. Definition. The infinite series \(\sum_{k=1}^{\infty} a_k\) converges, with sum \(s \in \mathbb{R}\), if the sequence \((s_n)\) of partial sums converges to \(s\). If the sequence \((s_n)\) does not converge, the series is said to diverge.

If the series \(\sum_{k=1}^{\infty} a_k\) converges with sum \(s\), we will often write

\[
\sum_{k=1}^{\infty} a_k = s
\]

and regard the infinite series as representing the number \(s\). Of course, by the uniqueness of limits for convergent sequences (IV.1.3.13.), the sum of a convergent infinite series is uniquely determined.

We reiterate the First Fundamental Principle: use of the word “sum” for a convergent infinite series does not mean that the “sum” is an actual sum in the algebraic sense.

IV.2.1.5. It is sometimes useful, if imprecise, to write an infinite series \(\sum_{k=1}^{\infty} a_k\), especially a convergent series, as

\[
\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots
\]

We will mostly avoid this notation since it is imprecise and can be misleading, although with due care it can be helpful in working with the series. When writing an infinite series this way one must be especially careful not to treat it like a finite sum; even great mathematicians have occasionally been misled into errors by confusing such an expression with a finite sum and applying unjustifiable algebraic operations.

A sequence converges if and only if all “tail” sequences converge (). A similar statement holds for series:

IV.2.1.6. Proposition. Let \(\sum_{k=1}^{\infty} a_k\) be an infinite series. The following are equivalent:
(i) The series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) For some \( n \in \mathbb{N} \), the tail series \( \sum_{k=n+1}^{\infty} a_k \) converges.

(iii) For every \( n \in \mathbb{N} \), the tail series \( \sum_{k=n+1}^{\infty} a_k \) converges.

**Proof:** (i) \( \Rightarrow \) (iii): Suppose the series converges, with sum \( s \), and fix \( n \in \mathbb{N} \). Let \( s_m \) be the \( m \)'th partial sum of the original series, and \( t_m \) the \( m \)'th partial sum of the series \( \sum_{k=n+1}^{\infty} a_k \).

The relationship is that \( t_m = s_{n+m} - s_n \) for any \( m \). We have that

\[
\lim_{m \to \infty} t_m = \left[ \lim_{m \to \infty} s_{n+m} \right] - s_n = s - s_n
\]

so the series \( \sum_{k=n+1}^{\infty} a_k \) converges with sum \( s - s_n \).

(iii) \( \Rightarrow \) (ii) is trivial.

(ii) \( \Rightarrow \) (i): If the tail series converges for some \( n \in \mathbb{N} \), with sum \( t \), we have

\[
\lim_{m \to \infty} s_{n+m} = \left[ \lim_{m \to \infty} t_m \right] + s_n = t + s_n
\]

so the original series converges with sum \( t + s_n \).

**IV.2.1.7.** Note, however, that there is a difference between sequences and series: the limit of a tail of a sequence is the same as the limit of the whole sequence (i.e. the limit of a sequence depends only on the tail terms). But the sum of a tail of a convergent infinite series is not generally the same as the sum of the whole series, i.e. although the convergence of an infinite series depends only on tail behavior, the sum of the series depends on all the terms.

A useful first observation is that for an infinite series to converge, it is necessary for the terms to approach zero (the key observation is that \( s_{n+1} - s_n = a_{n+1} \) for any \( n \)):  

433
IV.2.1.8. **Proposition.** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. If the series converges, then

$$\lim_{k \to \infty} a_k = 0.$$ 

**Proof:** Let $\epsilon > 0$. Since $\lim_{n \to \infty} s_n$ exists and equals, say, $s$, there is an $N$ such that for all $n \geq N$,

$$|s_n - s| < \frac{\epsilon}{2}.$$ 

If $n \geq N$, we also have $|s_{n+1} - s| < \frac{\epsilon}{2}$. So for all $n \geq N$ we have

$$|a_{n+1}| = |s_{n+1} - s_n| = |(s_{n+1} - s) + (s - s_n)| \leq |s_{n+1} - s| + |s - s_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and thus $\lim_{n \to \infty} a_n = 0$. 

IV.2.1.9. It is important to note that this result goes only one way: if the terms of an infinite series approach zero, the series may or may not converge. A simple example is $\sum_{n=1}^{\infty} \frac{1}{n}$: we have $s_n \geq n \cdot \frac{1}{\sqrt{n}} = \sqrt{n}$ for each $n$, so the partial sums are unbounded and cannot converge. We will see many other examples later.

The result IV.2.1.8. is commonly used in its contrapositive form:

IV.2.1.10. **Corollary. [Divergence Criterion]** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. If the sequence $(a_k)$ of terms does not converge to zero, then the infinite series diverges.

We can give an important refinement of IV.2.1.8., using the Cauchy Criterion (). The key observation here is that if $n < m$, then

$$s_m - s_n = \sum_{k=n+1}^{m} a_k.$$ 

IV.2.1.11. **Theorem. [Cauchy Criterion for Series]** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. Then the series converges if and only if, for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$\left| \sum_{k=n+1}^{m} a_k \right| < \epsilon$$

whenever $N \leq n < m$. 

434
In other words, an infinite series converges if and only if the sum of arbitrarily many consecutive terms is small, provided only that the sum starts sufficiently far out in the sequence of terms. This result includes IV.2.1.8., which is the special case where \( m = n + 1 \).

As a variation, we have a version involving sums of “tails” of the series which refines part of IV.2.1.6.: 

**IV.2.1.12. Corollary.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series. If the series converges, then for every \( \epsilon > 0 \) there is an \( N \in \mathbb{N} \) such that
\[
\sum_{k=n+1}^{\infty} a_k
\]
(which converges for all \( n \in \mathbb{N} \) by IV.2.1.6.) satisfies
\[
\left| \sum_{k=n+1}^{\infty} a_k \right| < \epsilon
\]
whenever \( n \geq N \).

**Proof:** Suppose the series converges. Let \( \epsilon > 0 \), and fix \( N \in \mathbb{N} \) such that
\[
\left| \sum_{k=n+1}^{m} a_k \right| < \frac{\epsilon}{2}
\]
whenever \( N \leq n < m \). We have, for \( n \geq N \),
\[
\left| \sum_{k=n+1}^{\infty} a_k \right| = \lim_{m \to \infty} \left| \sum_{k=n+1}^{m} a_k \right| \leq \frac{\epsilon}{2} < \epsilon.
\]

This corollary can also be proved directly with a slight extension of the proof of IV.2.1.6.

**Geometric Series**

One of the most basic examples of an infinite series is the geometric series.

**IV.2.1.13. Definition.** Let \( x \in \mathbb{R} \). The geometric series with ratio \( x \) is the infinite series
\[
\sum_{k=1}^{\infty} x^k.
\]

Eventually, we will want to regard \( x \) as a variable to obtain a function (power series), but for now \( x \) will be just regarded as a fixed number.
**IV.2.1.14. PROPOSITION.** The geometric series with ratio $x$ converges if and only if $|x| < 1$. In this case, we have

$$\sum_{k=1}^{\infty} x^k = \frac{x}{1-x}.$$ 

**Proof:** If $|x| \geq 1$, the series diverges by the Divergence Criterion since the terms do not go to zero. For any $x$, we have, for any $n \in \mathbb{N},$

$$(1 - x)(x + x^2 + \cdots + x^n) = x - x^{n+1}$$

by a simple calculation (or, more precisely, a simple proof by induction), so if $s_n$ is the $n$’th partial sum of the geometric series with ratio $x$, we have

$$(1 - x)s_n = x - x^{n+1}$$

and thus, for $x \neq 1$, we have

$$s_n = \frac{x}{1-x} - \frac{x^{n+1}}{1-x}$$

so, if $|x| < 1$, we have

$$\lim_{n \to \infty} s_n = \frac{x}{1-x} - \lim_{n \to \infty} \frac{x^{n+1}}{1-x} = \frac{x}{1-x}$$

and thus the series converges with sum $\frac{x}{1-x}$. 

**IV.2.1.15. REMARKS.**

(i) This is one of the few infinite series where it is possible to find a simple formula for the exact sum. In many cases, it can be shown easily from one of the convergence tests that the series converges, but the exact sum is not easy (and often even impossible) to find.

(ii) The closely related series

$$\sum_{k=0}^{\infty} x^k$$

is often called the geometric series instead. This series just has one more term at the beginning, equal to 1; thus it also converges if and only if $|x| < 1$, and the sum is

$$1 + \frac{x}{1-x} = \frac{(1-x) + x}{1-x} = \frac{1}{1-x}.$$ 

**Linear Combinations of Convergent Series**

There is some nice algebraic behavior for convergent infinite series: 

436
IV.2.1.16. **Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) and \( \sum_{k=1}^{\infty} b_k \) be convergent infinite series. Then the series \( \sum_{k=1}^{\infty} (a_k + b_k) \) converges, and
\[
\sum_{k=1}^{\infty} (a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k .
\]

**Proof:** Let \( s_n \) and \( t_n \) be the \( n \)’th partial sum of the series \( \sum_{k=1}^{\infty} a_k \) and \( \sum_{k=1}^{\infty} b_k \) respectively. Then
\[
\sum_{k=1}^{n} (a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k = s_n + t_n
\]
for any \( n \) by usual rules of algebra for finite sums. Then, using \( () \), we obtain
\[
\lim_{n \to \infty} \left[\sum_{k=1}^{n} (a_k + b_k)\right] = \lim_{n \to \infty} (s_n + t_n) = \lim_{n \to \infty} s_n + \lim_{n \to \infty} t_n = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k
\]
and thus the series \( \sum_{k=1}^{\infty} (a_k + b_k) \) converges and has the desired sum. \( \triangle \)

IV.2.1.17. Note that this result also goes only one way: \( \sum_{k=1}^{\infty} (a_k + b_k) \) can converge even if the two individual series diverge. A rather trivial example is to take \( a_k = 1 \) and \( b_k = -1 \) for all \( k \).

Similarly, we have a “distributive law” for infinite series:

IV.2.1.18. **Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series and \( c \in \mathbb{R} \). If the series \( \sum_{k=1}^{\infty} a_k \) converges, then the series \( \sum_{k=1}^{\infty} ca_k \) converges, and
\[
\sum_{k=1}^{\infty} ca_k = c \left[ \sum_{k=1}^{\infty} a_k \right] .
\]
Proof: If $s_n$ is the $n$’th partial sum of $\sum_{k=1}^{\infty} a_k$, then for each $n$ we have

$$\sum_{k=1}^{n} c a_k = c \left[ \sum_{k=1}^{n} a_k \right] = c s_n$$

by the distributive law for finite sums. Thus, by (), we have

$$\lim_{n \to \infty} \left[ \sum_{k=1}^{n} c a_k \right] = \lim_{n \to \infty} c s_n = c \left[ \lim_{n \to \infty} s_n \right] = c \sum_{k=1}^{\infty} a_k$$

and the result follows.

IV.2.1.19. Note that there is no corresponding result for term-by-term products; there is no simple relationship between the partial sums of $\sum_{k=1}^{\infty} a_k b_k$ and those of $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$. In fact, if $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ converge, the series $\sum_{k=1}^{\infty} a_k b_k$ may or may not converge (). (However, see Exercise (); see Exercise () for a somewhat related result.)

Grouping and Contraction of Series

There is a limited sense in which the associative law applies in a convergent series.

IV.2.1.20. Definition. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. A grouping or contraction of the series is an infinite series $\sum_{k=1}^{\infty} b_k$, where

$$b_k = \sum_{j=m_{k-1}+1}^{m_k} a_j$$

for some strictly increasing sequence $0 = m_0 < m_1 < m_2 < \cdots$ in $\mathbb{N} \cup \{0\}$.

Informally, the series $\sum_{k=1}^{\infty} b_k$ is obtained by inserting parentheses

$$(a_1 + \cdots + a_{m_1}) + (a_{m_1+1} + \cdots + a_{m_2}) + \cdots + (a_{m_k+1} + \cdots + a_{m_{k+1}}) + \cdots$$

Note that all the terms in the original series appear exactly once, and in the same order; they are simply grouped together.
IV.2.1.21. **Proposition.** Let $\sum_{k=1}^{\infty} a_k$ be a convergent infinite series. Then any grouping or contraction of the series also converges, and has the same sum.

**Proof:** Let $\sum_{k=1}^{\infty} b_k$ be a contraction of the series. Then every partial sum of the contraction is also a partial sum of the original series; hence the sequence of partial sums of $\sum_{k=1}^{\infty} b_k$ is a subsequence of the sequence of partial sums of $\sum_{k=1}^{\infty} a_k$, and thus converges to the same limit. 

IV.2.1.22. This result does not apply to divergent series. A famous example concerns the series $\sum_{k=1}^{\infty} \frac{1}{k}$. We can contract this divergent series to $(1 - 1) + (1 - 1) + (1 - 1) + \cdots$ to obtain a convergent series with sum 0, or to $1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots$ to obtain a convergent series with sum 1.

Before convergence of infinite series was well understood, several great mathematicians argued over what should be the correct sum of this series; some, including Leibniz and Euler, claimed that the true sum should be $\frac{1}{2}$, since among other things this is the average of the two “equally reasonable” groupings above (they had better reasons too; cf. ()). The Cesaro sum $\sum$ of this series is $\frac{1}{2}$.

IV.2.1.23. Applying IV.2.1.21, repeatedly, we may insert parentheses at will in a convergent series (not changing the order of terms) without affecting convergence or changing the sum, so long as any two pairs of inserted parentheses are either nested or disjoint, i.e. only the patterns (]), [)], (][), or []() are allowed, not (][) or ([) (which do not make sense anyway).

If the order of the terms in a convergent series is changed, however, or if some (infinitely many) terms are deleted, the new series may not converge; see (), (), .

IV.2.2. **Nonnegative Series**

IV.2.2.1. A **nonnegative series** is an infinite series in which all the terms are nonnegative. Nonnegative series behave somewhat more nicely than general infinite series, and criteria for convergence are considerably simpler to state. In this subsection and the next two, we examine convergence of nonnegative series and develop the standard tests for convergence.
Of course, as for any infinite series, the terms of a convergent nonnegative infinite series must converge to zero. But a nonnegative series whose terms go to zero does not necessarily converge (IV.2.1.9.). In fact, we will develop a set of results that make precise the following informal principle:

**Fundamental Principle of Nonnegative Series:** A nonnegative series converges if and only if the terms of the series approach zero sufficiently rapidly.

The important fact about the sequence of partial sums of a nonnegative series is that it is nondecreasing. Since a nondecreasing sequence converges if and only if it is bounded (), we obtain:

**Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) be a nonnegative infinite series. Then the series converges if and only if the sequence of partial sums is bounded.

This result can fail for series which are not nonnegative. For example, in the geometric series \( \sum_{k=1}^{\infty} (-1)^k \) the partial sum \( s_n \) is \(-1\) for \( n \) odd and \( 0 \) for \( n \) even, hence the sequence of partial sums is bounded. But the series does not converge by the Divergence Criterion.

Indeed, some authors distinguish between infinite series which are “divergent” (or “definitely divergent”) in the sense that the limit of the partial sums is \(+\infty\) or \(-\infty\), and series which are “indeinitely divergent,” or “oscillatory,” in the sense that the sequence of partial sums has no limit in the extended real line.

**The Comparison Test**

The first test we develop for convergence of nonnegative series, which makes the Fundamental Principle more precise, is the Comparison Test. This test will, however, only have teeth once we develop a reasonable supply of “test series” to compare new series to. (So far the only examples we have are geometric series.)

**Theorem. [Comparison Test]** Let \( \sum_{k=1}^{\infty} a_k \) and \( \sum_{k=1}^{\infty} b_k \) be nonnegative infinite series. Suppose \( a_k \leq b_k \) for all \( k \). Then

(i) If \( \sum_{k=1}^{\infty} b_k \) converges, then \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( \sum_{k=1}^{\infty} a_k \) diverges, then \( \sum_{k=1}^{\infty} b_k \) diverges.
Proof: Statement (ii) is just the contrapositive of statement (i), so we need only prove (i). For $n \in \mathbb{N}$, let $s_n$ and $t_n$ be the $n$'th partial sum of $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ respectively. Then $s_n \leq t_n$ for all $n$. Since the sequence $(t_n)$ converges, it is bounded (); any bound for $(t_n)$ is also a bound for $(s_n)$, hence $(s_n)$ is also bounded. Thus $\sum_{k=1}^{\infty} a_k$ converges by IV.2.2.4.

We can give a slight extension in which finitely many terms can be ignored in each sequence:

**IV.2.2.7.** Corollary. [Extended Comparison Test] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be nonnegative infinite series. Suppose there are $p, q \in \mathbb{N}$ such that $a_{p+k} \leq b_{q+k}$ for all $k$ (this will be satisfied, for example, if there is an $N \in \mathbb{N}$ for which $a_k \leq b_k$ for all $k > N$, by taking $p = q = N$). Then

(i) If $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\sum_{k=1}^{\infty} a_k$ diverges, then $\sum_{k=1}^{\infty} b_k$ diverges.

Proof: Apply the Comparison Test to the tail series $\sum_{k=p+1}^{\infty} a_k$ and $\sum_{k=q+1}^{\infty} b_k$, and apply IV.2.1.6.

**IV.2.2.8.** The substance of these results is that a nonnegative series whose terms go to zero “more rapidly” than those of a convergent nonnegative series also converges, and a nonnegative series whose terms go to zero “more slowly” than those of a divergent nonnegative series also diverges.

The Harmonic Series

**IV.2.2.9.** Perhaps the most important single infinite series is the harmonic series

$$\sum_{k=1}^{\infty} \frac{1}{k}.$$

**IV.2.2.10.** It is not obvious whether the harmonic series converges. Its terms do go to zero, so the Divergence Criterion does not rule out convergence but does not guarantee it either. If we do some numerical experiments calculating some partial sums we obtain

$$s_{10} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10} = 2.9289\cdots$$
\( s_{50} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{50} = 4.499 \ldots \)
\( s_{100} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{100} = 5.187 \ldots \)

and the partial sums might appear to be approaching a limit. If we go on and compute
\( s_{1000} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{1000} = 7.485 \ldots \)
\( s_{10000} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10000} = 9.787 \ldots \)

this conclusion seems less obvious. Are the partial sums bounded?

**IV.2.2.11. Theorem.** The harmonic series diverges, i.e. the partial sums are unbounded.

**Proof:** We will give a quick proof later using the Integral Test, but here we give an alternate elementary proof based on the Comparison Test. Define a new series \( \sum_{k=1}^{\infty} a_k \) by setting \( a_1 = 1 \) and \( a_k = 2^{-r} \) for \( 2^{r-1} + 1 \leq k \leq 2^r, \, r \in \mathbb{N} \). Note that \( a_k \leq \frac{1}{k} \) for all \( k \). If \( s_n \) is the \( n \)'th partial sum of the series \( \sum_{k=1}^{\infty} a_k \), then for \( m \in \mathbb{N} \) and \( n = 2^m \) we have

\[
s_n = s_{2^m} = \sum_{k=1}^{2^m} a_k = 1 + \sum_{r=1}^{m} \sum_{k=2^{r-1}+1}^{2^r} a_k = 1 + \sum_{r=1}^{m} [2^{r-1} \cdot 2^{-r}] = 1 + \sum_{r=1}^{m} \frac{1}{2} = 1 + \frac{m}{2}.
\]

These are unbounded as \( m \to \infty \), so \( \sum_{k=1}^{\infty} a_k \) diverges, and hence the harmonic series also diverges by the Comparison Test.

**IV.2.2.12.** The argument of this proof was discovered by the French mathematician NICOLE ORESME in the fourteenth century, and is one of the most significant pieces of mathematics surviving from medieval Europe. (Oresme also developed the theory of exponents, and was the first to draw graphs of functions.) Despite this argument, convergence of infinite series and its importance were not widely understood by mathematicians until the nineteenth century.

**IV.2.2.13.** The argument indicates that the partial sums of the harmonic series grow only logarithmically in \( n \), and thus go to +\( \infty \) very slowly. This conclusion will follow more carefully from an application of the Integral Test. In fact, it takes \( 15,092,688,622,113,788,323,693,563,264,538,101,449,859,497 \approx 1.5 \times 10^{43} \) terms to make the partial sum exceed 100, for example \([?]\). For more information about the partial sums of the harmonic series, see the website [http://mathworld.wolfram.com/HarmonicNumber.html](http://mathworld.wolfram.com/HarmonicNumber.html).

The divergence of the harmonic series leads to some apparent paradoxes. For example, a stable stack of identical books, or bricks, etc., can be made (in principle) in which the top book overhangs the bottom one by an arbitrary amount: overhang the \((k-1)\)'th book from the \(k\)'th (counting down from the top) by \( \frac{1}{k} \) times the width of the book. For other “paradoxes” involving the harmonic series, see [\(\)].
The \( p \)-Series

The harmonic series is a member of an important class of infinite series to which a similar argument applies, the \( p \)-series:

**IV.2.2.14.** Definition. Let \( p \in \mathbb{R} \). The \( p \)-series with exponent \( p \) is the infinite series

\[
\sum_{k=1}^{\infty} k^{-p} = \sum_{k=1}^{\infty} \frac{1}{k^p}.
\]

The \( p \)-series with exponent \( p \) is a nonnegative series. The only interesting case is when \( p > 0 \), since for \( p \leq 0 \) the terms of the series do not approach zero, so the series automatically diverges.

**IV.2.2.15.** The harmonic series is the \( p \)-series with exponent 1. We know this series diverges (IV.2.2.11.), and by comparison we have that a \( p \)-series for \( p < 1 \) also diverges. We now show that a \( p \)-series with \( p > 1 \) converges. We will give another proof of these facts using the Integral Test (IV.2.3.5.).

**IV.2.2.16.** Theorem. The \( p \)-series with exponent \( p \) converges if and only if \( p > 1 \).

**Proof:** We have already done the cases \( p \leq 1 \), so assume \( p > 1 \). Define a new series \( \sum_{k=1}^{\infty} a_k \) by setting \( a_1 = 1 \) and \( a_k = 2^{-pr} = [2^{-r}]^p \) for \( 2^r \leq k \leq 2^{r+1} - 1 \), \( r \in \mathbb{N} \). Note that \( a_k \geq \frac{1}{2^r} \) for all \( k \). If \( s_n \) is the \( n \)’th partial sum of the series \( \sum_{k=1}^{\infty} a_k \), then for \( m \in \mathbb{N} \) and \( n = 2^{m+1} - 1 \) we have

\[
s_n = s_{2^{m+1} - 1} = \sum_{k=1}^{2^{m+1} - 1} a_k = 1 + \sum_{r=1}^{m} \sum_{k=2^r}^{2^{r+1} - 1} a_k = 1 + \sum_{r=1}^{m} [2^r \cdot 2^{-pr}] = 1 + \sum_{r=1}^{m} [2^{1-p}]^r.
\]

These are partial sums of a geometric series with ratio \( 2^{1-p} < 1 \), so the \( s_n \) converge, and hence the \( p \)-series converges by the Comparison Test.

The Limit Comparison Test

There is a variation of the Comparison Test which is often simpler to use:

**IV.2.2.17.** Theorem. [Limit Comparison Test] Let \( \sum_{k=1}^{\infty} a_k \) and \( \sum_{k=1}^{\infty} b_k \) be infinite series with positive terms. Suppose

\[
L = \lim_{k \to \infty} \frac{a_k}{b_k}
\]

exists as a (nonnegative) extended real number. Then:

443
(i) If $0 \leq L < +\infty$ and $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $0 < L \leq +\infty$ and $\sum_{k=1}^{\infty} b_k$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

**Proof:** For (i), if $0 < L < +\infty$, then we have $a_k < cb_k$ for all sufficiently large $k$. We have that $\sum_{k=1}^{\infty} cb_k$ converges, so $\sum_{k=1}^{\infty} a_k$ converges by the **Extended Comparison Test**. The proof of (ii) is almost identical: if $0 < d < L$, then $\sum_{k=1}^{\infty} db_k$ diverges, so $\sum_{k=1}^{\infty} a_k$ diverges.

A refined version using $\limsup$ and $\liminf$ holds with the same proof:

**IV.2.2.18. Theorem.** [Limit Comparison Test, General Version] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series with positive terms.

(i) If $\limsup_{k \to \infty} \frac{a_k}{b_k} < +\infty$ and $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\liminf_{k \to \infty} \frac{a_k}{b_k} > 0$ and $\sum_{k=1}^{\infty} b_k$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

The next class of examples gives one of the most important applications of the Limit Comparison Test:

**IV.2.2.19. Example.** Let $f(k)$ and $g(k)$ be functions which are finite sums of terms of the form $c_r k^r$ for $r \in \mathbb{R}$. ($f$ and $g$ need not be polynomials, i.e. the powers need not be nonnegative integers.) Suppose $g(k) \neq 0$ for all $k$. Let $c_p k^p$ be the term in $f$ with largest power, and $d_q k^q$ the term in $g$ with largest power ($c_p$ and $d_q$ are nonzero). Then:

(i) If $q - p > 1$, the series $\sum_{k=1}^{\infty} \frac{f(k)}{g(k)}$ converges.

(ii) If $q - p \leq 1$, the series $\sum_{k=1}^{\infty} \frac{f(k)}{g(k)}$ diverges.
For we can do a limit comparison test with the $p$-series $\sum_{k=1}^{\infty} \frac{1}{k^{q-p}}$ since
\[
\lim_{k \to \infty} \frac{k^{q-p} f(k)}{g(k)} = \frac{c_p}{d_q}
\]
which is a nonzero finite number. (If $\frac{c_p}{d_q}$ is negative, first multiply the series by $-1$. The terms of the series will then eventually be strictly positive.)

The Border Between Convergence and Divergence

IV.2.2.20. The set of nonnegative sequences $(a_n)$ converging to zero splits into two disjoint sets, those for which $\sum_{k=1}^{\infty} a_k$ converges and those for which it diverges. Any sequence dominated by a sequence in the first group is in the first group, and any sequence which dominates one in the second is in the second. Is there some identifiable border between the sets? Specifically, we can ask:

(i) Is there a sequence $(a_n)$ such that $a_n \to 0$ most slowly among terms of convergent series, i.e. a nonnegative series $\sum_{k=1}^{\infty} b_k$ converges if and only if $\frac{a_k}{b_k}$ is bounded?

(ii) Is there a sequence $(a_n)$ such that $a_n \to 0$ most rapidly among terms of divergent series, i.e. a nonnegative series $\sum_{k=1}^{\infty} b_k$ diverges if and only if $\frac{a_k}{b_k}$ is bounded?

(We could restrict to nonincreasing sequences to sharpen the questions and eliminate some obvious pathologies; the terms of a nonnegative series can be reordered without affecting convergence – see IV.2.8.4.)

Both questions have negative answers: for any convergent nonnegative series there is another convergent nonnegative series whose terms approach zero more slowly (Exercise IV.2.16.5.), and for every divergent nonnegative series there is a divergent nonnegative series whose terms go to zero more rapidly (Exercise IV.2.16.6.). The border cannot even be countably described: for any countable collection of convergent nonnegative series there is a convergent nonnegative series whose terms go to zero more slowly than all of them, and for any countable collection of divergent nonnegative series there is a divergent nonnegative series whose terms go to zero more rapidly than all of them (Exercise IV.2.16.17.). Thus, although a sequence like $(\frac{1}{k})$ can be regarded as “close to the border between series convergence and divergence” (i.e. “just barely divergent”), there is no precise countable description of an exact border between convergence and divergence, and thus no foolproof countably describable test for convergence of nonnegative series.

Nonetheless, there are many useful tests which together work for most series arising in practice, as described in the next two sections.

IV.2.3. Tests for Convergence of Nonnegative Series

Besides the Comparison Tests of the previous section, which primarily illustrate the Fundamental Principle of Nonnegative Series, there are several standard tests for convergence of such series, along with a large number of specialized tests. In this subsection, we describe the three standard tests, the Integral Test, the Root Test, and the Ratio Test, and give a sampling of some of the more important specialized tests.
The Integral Test

Infinite series bear a close resemblance to improper integrals; in fact, infinite series can be regarded as a discrete analog of improper integrals. At least in the case of nonnegative series, this close relationship can be exploited to give a test for convergence of certain nonnegative series using improper integrals.

IV.2.3.1. Suppose $f$ is a function on the interval $[1, \infty)$ which is (Riemann-)integrable on each finite subinterval (e.g. continuous or at least piecewise-continuous, or monotone). Recall (IV.2.2.2.) that the improper integral of $f$ from 1 to $\infty$ is defined to be

$$\int_1^{\infty} f(t) \, dt = \lim_{b \to \infty} \left[ \int_1^{b} f(t) \, dt \right]$$

if the limit exists (the improper integral is then said to converge). The finite integrals may be regarded as “partial integrals” and some sort of continuous analog of the partial sums of an infinite series.

If $f$ is nonnegative on $[1, \infty)$, the “partial integral” from 1 to $b$ is nondecreasing as $b$ increases, so the improper integral converges if and only if the partial integrals are bounded; and

$$\int_1^{\infty} f(t) \, dt = \sup_{b>1} \left[ \int_1^{b} f(t) \, dt \right]$$

if the supremum is finite. This will be the case if and only if $f(t) \to 0$ “sufficiently rapidly” as $t \to \infty$, in complete analogy with the Fundamental Principle of Nonnegative Series (IV.2.2.2.).

IV.2.3.2. If the function $f$ is nonnegative and nonincreasing, we have that

$$f(k + 1) \cdot 1 \leq \int_k^{k+1} f(t) \, dt \leq f(k) \cdot 1$$

for all $k$ (see Figure (IV.2.3.2.)). It follows that for any $n$, we have

$$\sum_{k=2}^{n+1} f(k) \leq \int_1^{n+1} f(t) \, dt \leq \sum_{k=1}^{n} f(k) .$$

Thus, if the nonnegative infinite series $\sum_{k=1}^{\infty} f(k)$ converges, we obtain that the supremum of the partial integrals from 1 to $n$ is finite, and hence the improper integral $\int_1^{\infty} f(t) \, dt$ also converges. Conversely, we have

$$\sum_{k=1}^{n} f(k) \leq f(1) + \int_1^{n} f(t) \, dt$$

for every $n$, so if the improper integral converges, the partial sums of the series $\sum_{k=1}^{\infty} f(k)$ are bounded and so the series converges. Thus we obtain:
IV.2.3.3. Theorem. [Integral Test] Let \( f \) be a function which is nonnegative and nonincreasing on the interval \([1, \infty)\). Then the infinite series \( \sum_{k=1}^{\infty} f(k) \) converges if and only if the improper integral \( \int_{1}^{\infty} f(t) \, dt \) converges, and we have
\[
\int_{1}^{\infty} f(t) \, dt - f(1) \leq \sum_{k=1}^{\infty} f(k) \leq \int_{1}^{\infty} f(t) \, dt.
\]

IV.2.3.4. Note that this test, when it applies, gives a complete necessary and sufficient condition for convergence of the series \( \sum_{k=1}^{\infty} f(k) \). But although it gives some information about the sum, it does not directly allow a calculation of the exact sum. See ( )

Of course, when using this test, as with any test, one must check that the hypotheses are satisfied.

Application: \( p \)-Series

We use the Integral Test to give an alternate proof of the convergence or divergence of the \( p \)-series (IV.2.14.); cf. IV.2.16.

IV.2.3.5. Theorem. The \( p \)-series with exponent \( p \) converges if and only if \( p > 1 \).

Proof: Suppose \( p > 0, p \neq 1 \). We apply the Integral Test to the function \( f(t) = t^{-p} \), which is continuous, nonnegative, and nonincreasing on \([1, \infty)\). Then we have
\[
\int_{1}^{b} t^{-p} \, dt = \frac{t^{-p+1}}{-p+1} \bigg|_{1}^{b} = \frac{b^{1-p}}{1-p} - \frac{1}{1-p}.
\]
If \( p < 1 \), \( b^{1-p} \to +\infty \) as \( b \to \infty \) and the improper integral diverges, so the series diverges. If \( p > 1 \), \( b^{1-p} \to 0 \) as \( b \to \infty \) and the improper integral converges, so the series converges.

For completeness, we also do the case \( p = 1 \) using the integral test. We have
\[
\int_{1}^{b} t^{-1} \, dt = \log t \bigg|_{1}^{b} = \log b - \log 1 = \log b
\]
which goes to \( +\infty \) as \( b \to \infty \); hence the improper integral diverges and the series diverges too.

The Ratio and Root Tests

The Ratio Test and the Root Test are two of the most widely-used tests for convergence of nonnegative series. They are rather crude, in the sense that the tests fail (give no conclusion) for many series; in fact, they are successful only for series which converge or diverge rather spectacularly. But when they do work they are often quite easy to apply, so are quite useful.

The Root Test and the Ratio Test are sometimes called Cauchy’s Test and d’Alembert’s Test respectively, after their originators (although it seems generous to attribute the Ratio Test to d’ALEMBERT; see [Gra00, p.116-117]).
IV.2.3.6. Theorem. [Basic Root Test] Let \( \sum_{k=1}^{\infty} a_k \) be a nonnegative infinite series. Then

(i) If there is a \( u < 1 \) such that \( \sqrt[k]{a_k} = a_k^{1/k} \leq u \) for all sufficiently large \( k \), then the series converges.

(ii) If \( \sqrt[k]{a_k} = a_k^{1/k} \geq 1 \) for infinitely many \( k \), then the terms do not approach zero and the series diverges.

Proof: We first show (ii). If \( a_k^{1/k} \geq 1 \) for infinitely many \( k \), then \( a_k \leq k \) for infinitely many \( k \), and thus we cannot have \( \lim_{k \to \infty} a_k = 0 \), and the series diverges by the Divergence Criterion.

For (i), if \( a_k^{1/k} < u \) for all \( k \geq N \), then \( a_k \leq u^k \) for all \( k \geq N \). Comparing the series with the geometric series \( \sum_{k=1}^{\infty} u^k \), which converges since \( u < 1 \), we conclude from the Extended Comparison Test that the series converges.

The Root Test is usually phrased in terms of the \( \limsup \), although the divergence part is not quite as strong (but often easier to check):

IV.2.3.7. Corollary. [Root Test] Let \( \sum_{k=1}^{\infty} a_k \) be a nonnegative infinite series, and set

\[ r = \limsup_{k \to \infty} \sqrt[k]{a_k} = \limsup_{k \to \infty} a_k^{1/k}. \]

Then

(i) If \( r < 1 \), then the series converges.

(ii) If \( r > 1 \), then the terms do not approach zero and the series diverges.

(iii) If \( r = 1 \), then the series may or may not converge, and no conclusion may be drawn (the test fails).

In particular, if \( \lim_{k \to \infty} a_k^{1/k} \) exists, the limit is the \( r \) of the theorem and the series converges or diverges according to whether \( r < 1 \) or \( r > 1 \). But the test applies even if this limit does not exist; recall that the \( \limsup \) always exists for any sequence.

Proof: Part (i) is actually equivalent to IV.2.3.6.(i): suppose \( r < 1 \), and fix \( u \) with \( r < u < 1 \). Then there is an \( N \in \mathbb{N} \) with \( a_k^{1/k} < u \) for all \( k \geq N \) (\( ) \), so the hypotheses of IV.2.3.6.(i) are satisfied. Conversely, if there is a \( u \) such that IV.2.3.6.(i) holds, then \( r \leq u \) by (\( ) \).

For (ii), if \( r > 1 \), then \( a_k^{1/k} > 1 \) for infinitely many \( k \), so the hypotheses of IV.2.3.6.(ii) are satisfied. (However, the two are not equivalent: the hypotheses of IV.2.3.6 (ii) do not imply those of IV.2.3.7 (ii).)

To show the test fails in case (iii), it suffices to consider \( p \)-series. For the \( p \)-series, we have

\[ a_k^{1/k} = \frac{1}{[k^{1/p}]^{1/k}} = \frac{1}{[k^{1/k}]^p} \]

which has a limit of 1 for every \( p \) since \( \lim_{k \to \infty} k^{1/k} = 1 \) (\( ) \). But the \( p \)-series can either converge or diverge depending on \( p \). (Actually even the Basic Root Test fails to apply to \( p \)-series for \( p > 0 \).)
IV.2.3.8. Theorem. [Basic Ratio Test] Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Then

(i) If there is a \( u < 1 \) such that \( \frac{a_{k+1}}{a_k} \leq u \) for all sufficiently large \( k \), then the series converges.

(ii) If \( \frac{a_{k+1}}{a_k} \geq 1 \) for all sufficiently large \( k \), then the terms do not approach zero and the series diverges.

Proof: The proof is quite similar to the proof of the Root Test.

(ii): Suppose there is an \( N \in \mathbb{N} \) such that \( \frac{a_{n+1}}{a_n} \geq 1 \), \( a_{n+1} \geq a_n \) for all \( n \geq N \). Then the terms of the series are eventually nondecreasing and so, since they are positive, they cannot converge to zero. Thus the series diverges by the Divergence Criterion.

(i): Suppose there is an \( N \in \mathbb{N} \) such that \( \frac{a_{k+1}}{a_k} < 1 \) for all \( k \geq N \). Then, for all \( k \in \mathbb{N} \), we have

\[
a_{N+k} \leq u a_{N+k-1} \leq u^2 a_{N+k-2} \leq \cdots \leq u^k a_N
\]

and we may apply the Extended Comparison Test to the given series and the series \( \sum_{k=1}^{\infty} a_N u^k \), which converges since \( u < 1 \).

As with the Root Test, the Ratio Test is usually phrased in terms of \( \limsup \) (and \( \liminf \)):

IV.2.3.9. Corollary. [Ratio Test] Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Then

(i) If \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} < 1 \), then the series converges.

(ii) If \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} > 1 \), then the terms do not approach zero and the series diverges.

(iii) If \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} \leq 1 \leq \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} \), then the series may or may not converge, and no conclusion may be drawn (the test fails).

A more special version often applies which is slightly easier to state:

IV.2.3.10. Corollary. [Ratio Test, Limit Version] Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms, and suppose \( r = \lim_{k \to \infty} \frac{a_{k+1}}{a_k} \) exists. Then

(i) If \( r < 1 \), then the series converges.
(ii) If \( r > 1 \), then the terms do not approach zero and the series diverges.

(iii) If \( r = 1 \), then the series may or may not converge, and no conclusion may be drawn (the test fails).

**Proof:** IV.2.3.10. (i)–(ii) are clearly special cases of IV.2.3.9. (i)–(ii). The proof of IV.2.3.9. (i)–(ii) is nearly identical to the proof of IV.2.3.7. from IV.2.3.6. and is left to the reader.

To show the test fails in case (iii), even in the case of IV.2.3.10., it suffices to consider \( p \)-series. For the \( p \)-series, we have

\[
\frac{a_{k+1}}{a_k} = \frac{1}{(k + 1)^p} = \left[ \frac{k}{k + 1} \right]^p
\]

which has a limit of 1 for every \( p \) since \( \lim_{k \to \infty} \frac{k}{k + 1} = 1 \). But the \( p \)-series can either converge or diverge depending on \( p \).

There are even convergent positive series where \( \lim \inf_{n \to \infty} \frac{a_{k+1}}{a_k} = 1 < \lim \sup_{n \to \infty} \frac{a_{k+1}}{a_k} \) (Exercise IV.2.16.2.) and divergent positive series where \( \lim \inf_{n \to \infty} \frac{a_{k+1}}{a_k} < 1 = \lim \sup_{n \to \infty} \frac{a_{k+1}}{a_k} \) (Exercise IV.2.16.3.), and both convergent and divergent positive series where \( \lim \inf_{n \to \infty} \frac{a_{k+1}}{a_k} < 1 < \lim \sup_{n \to \infty} \frac{a_{k+1}}{a_k} \) (Exercise IV.2.16.4.).

To compare the strength of the Ratio Test and Root Test, we use the following fact:

**IV.2.3.11. Proposition.** Let \((x_n)\) be a sequence of positive numbers. Then

\[
\lim \inf_{n \to \infty} \frac{x_{n+1}}{x_n} \leq \lim \inf_{n \to \infty} x_n^{1/n} \leq \lim \sup_{n \to \infty} x_n^{1/n} \leq \lim \sup_{n \to \infty} \frac{x_{n+1}}{x_n}.
\]

**Proof:** We prove only the last inequality; the first is similar (Exercise ()). Let \( r = \lim \sup_{n \to \infty} \frac{x_{n+1}}{x_n} \geq 0 \). If \( r = +\infty \), there is nothing to prove, so assume \( r \in [0, \infty) \). Then, for any \( s > r \), there is an \( N \in \mathbb{N} \) such that \( \frac{x_{n+1}}{x_n} < s \) for all \( n \geq N \). For \( m \in \mathbb{N} \), we have

\[
x_{N+m} < s x_{N+m-1} < s^2 x_{N+m-2} < \cdots < s^m x_N
\]

and so we obtain

\[
x_n < s^n \frac{x_N}{s^N}
\]

for all \( n \geq N \) (note that \( s > 0 \) since \( r \geq 0 \)). Taking \( n \)'th roots, we have

\[
x_n^{1/n} < t^{1/n}
\]

for all \( n \geq N \), where \( t = \frac{x_N}{s^N} \). So

\[
\lim \sup_{n \to \infty} x_n^{1/n} \leq s \cdot \lim \sup_{n \to \infty} t^{1/n} = s
\]

since \( \lim_{n \to \infty} t^{1/n} = 1 \). Since \( s \) is an arbitrary number \( > r \), we have

\[
\lim \sup_{n \to \infty} x_n^{1/n} \leq r.
\]
IV.2.3.12. Any or all of the inequalities in IV.2.3.11. can be strict; see Exercise ().

IV.2.3.13. **Corollary.** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. If the Ratio Test gives a conclusion about the convergence or divergence of the series (i.e. the Ratio Test does not fail), then the Root Test also gives the same conclusion (in particular, it does not fail).

In other words, the Root Test is, at least in principle, a more broadly applicable test than the Ratio Test. It is in fact strictly more broadly applicable (Exercise ()). However, in practice the Ratio Test is frequently simpler to use than the Root Test, so there is utility in having both tests available.

IV.2.4. **More Delicate Tests**

There are several refinements of the Ratio Test which can be used for certain series for which the Ratio Test fails. Most are special cases of the following blockbuster test:

IV.2.4.1. **Theorem. [Kummer’s Test]** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms.

(i) If there are positive real numbers $r$ and $p_1, p_2, \ldots$ such that

$$p_k \frac{a_k}{a_{k+1}} - p_{k+1} \geq r$$

for all sufficiently large $k$, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If there are positive real numbers $p_1, p_2, \ldots$ such that

$$p_k \frac{a_k}{a_{k+1}} - p_{k+1} \leq 0$$

for all sufficiently large $k$, and $\sum_{k=1}^{\infty} \frac{1}{p_k}$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

Note that the Basic Ratio Test is the special case where $p_k = 1$ for all $k$.

**Proof:** For both (i) and (ii), by passing to a tail of the series we may assume the inequalities hold for all $k$.

(i): For each $k$, we have $a_{k+1} \leq \frac{1}{r} [p_k a_k - p_{k+1} a_{k+1}]$, so for each $n > 1$,

$$s_n = \sum_{k=1}^{n} a_k = a_1 + \sum_{k=2}^{n} a_k \leq a_1 + \frac{1}{r} \sum_{k=1}^{n-1} (p_k a_k - p_{k+1} a_{k+1})$$

$$= a_1 + \frac{1}{r} \left[ \sum_{k=1}^{n-1} p_k a_k - \sum_{k=2}^{n} p_k a_k \right] = a_1 + \frac{p_1 a_1}{r} - \frac{p_n a_n}{r} \leq a_1 + \frac{p_1 a_1}{r}$$

451
and so the partial sums are bounded and the series converges.

(ii): We have that

\[
\frac{a_2}{a_2} \geq \frac{p_1 a_1}{p_2}, \quad \frac{a_3}{a_3} \geq \frac{p_2 a_2}{p_3}, \quad \ldots, \quad \frac{a_k}{a_k} \geq \frac{p_l a_{l-1}}{p_k}, \quad \ldots
\]

and so we can apply the Comparison Test to \( \sum_{k=1}^{\infty} a_k \) and the divergent series

\[
a_1 + \sum_{k=2}^{\infty} \frac{p_1 a_1}{p_k}.
\]

IV.2.4.2. Kummer’s Test is, in principle at least, a maximally applicable test: for any infinite series with strictly positive terms, there is a sequence \( \{p_k\} \) of positive numbers for which the test works to give either convergence or divergence (Exercise IV.2.16.1.). However, it can be extremely difficult to find such a sequence for a given series, so the general Kummer Test is of limited applicability in practice.

Special cases of Kummer’s Test with specific sequences \( \{p_n\} \) are more useful. Here is one of the most widely-used ones:

**IV.2.4.3. Theorem. [Raabe’s Test]** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Then

(i) If \( \liminf_{k \to \infty} k \left[ \frac{a_k}{a_{k+1}} - 1 \right] > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( k \left[ \frac{a_k}{a_{k+1}} - 1 \right] \leq 1 \) for all sufficiently large \( k \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

**Proof:** Apply Kummer’s Test with \( p_k = k \) for all \( k \) and, for (i), \( r \) a number with \( 0 < r < \liminf_{k \to \infty} k \left[ \frac{a_k}{a_{k+1}} - 1 \right] - 1 \).

A limit version is frequently applicable:

**IV.2.4.4. Corollary. [Raabe’s Test, Limit Version]** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Suppose

\[
L = \lim_{k \to \infty} k \left[ \frac{a_k}{a_{k+1}} - 1 \right]
\]

exists (as an extended real number). Then
(i) If \( L > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( L < 1 \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

(iii) If \( L = 1 \), then the series \( \sum_{k=1}^{\infty} a_k \) may or may not converge, and no conclusion may be drawn (the test fails).

Raabe’s Test is often written in a slightly different but equivalent form (cf. Exercise IV.2.16.9).

**IV.2.4.5. COROLLARY. [RAABE’S TEST, VERSION 2]** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Suppose

\[
L = \lim_{k \to \infty} k \left[ 1 - \frac{a_{k+1}}{a_k} \right]
\]

exists (as an extended real number). Then

(i) If \( L > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( L < 1 \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

(iii) If \( L = 1 \), then the series \( \sum_{k=1}^{\infty} a_k \) may or may not converge, and no conclusion may be drawn (the test fails).

Yet another formulation of Raabe’s Test is:

**IV.2.4.6. THEOREM. [RAABE’S TEST, VERSION 3]** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Suppose

\[
\frac{a_{k+1}}{a_k} = 1 - \frac{\lambda}{k} + o \left( \frac{1}{k} \right)
\]

for some fixed real number \( \lambda \). Then

(i) If \( \lambda > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.
(ii) If $\lambda < 1$, the series $\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $\lambda = 1$, then the series $\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

**Proof:** Applying IV.2.4.5., we have

$$k \left[ 1 - \frac{a_{k+1}}{a_k} \right] = \lambda + k \cdot o \left( \frac{1}{k} \right) = \lambda + o(1)$$

so we have $L = \lambda$. The converse implication is similar, and left as an exercise.

**IV.2.4.7.** Note that Raabe’s Test (even the limit version IV.2.4.4.) is more delicate than the Ratio Test: if $\limsup_{k \to \infty} \frac{a_{k+1}}{a_k} < 1$, then $L = +\infty$, and if $\liminf_{k \to \infty} \frac{a_{k+1}}{a_k} > 1$, then $L = -\infty$, so Raabe’s Test applies (and gives the same conclusion, of course).

Raabe’s test is delicate enough to handle $p$-series for $p \neq 1$: for the $p$-series we have

$$L = \lim_{k \to \infty} k \left[ \frac{a_k}{a_{k+1}} - 1 \right] = \lim_{k \to \infty} k \left[ \frac{(k+1)^p}{k^p} - 1 \right] = \lim_{k \to \infty} \frac{(k+1)^p - k^p}{k^{p-1}} = p$$

(See Exercise (i).)

However, Raabe’s Test does not work for the harmonic series (although the slight extension to Kummer’s Test for $p_k = k$ does), nor does it work for the log$^p$-series

$$\sum_{k=2}^{\infty} \frac{1}{k^{[\log k]^p}}$$

which converges for $p > 1$ and diverges for $p \leq 1$ (i). For these log$^p$-series, a “second-order Raabe’s Test” does apply:

**IV.2.4.8.** Theorem. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$M = \lim_{k \to \infty} (\log k) \left( k \left[ \frac{a_k}{a_{k+1}} - 1 \right] - 1 \right)$$

exists (as an extended real number). Then

(i) If $M > 1$, the series $\sum_{k=1}^{\infty} a_k$ converges.
(ii) If \( M < 1 \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

(iii) If \( M = 1 \), then the series \( \sum_{k=1}^{\infty} a_k \) may or may not converge, and no conclusion may be drawn (the test fails).

**Proof:** Apply Kummer’s Test with \( p_k = k \log k \). (See Exercise (i).)

There is a variation of this second-order test similar to IV.2.4.6. The proof is left as an exercise.

**IV.2.4.9. Theorem.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Suppose

\[
\frac{a_{k+1}}{a_k} = 1 - \frac{1}{k} - \frac{\mu}{k \log k} + o\left(\frac{1}{k \log k}\right)
\]

for some fixed real number \( \mu \). Then

(i) If \( \mu > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( \mu < 1 \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

(iii) If \( \mu = 1 \), then the series \( \sum_{k=1}^{\infty} a_k \) may or may not converge, and no conclusion may be drawn (the test fails).

**IV.2.4.10.** The second-order test will work for the \( \log^p \)-series with \( p \neq 1 \). It will not work for the \( \log \log^p \)-series

\[
\sum_{k=3}^{\infty} \frac{1}{k \log k [\log \log k]^p}
\]

(which again converges or diverges depending whether \( p > 1 \) or \( p \leq 1 \); cf. (i)), but a “third-order Raabe Test” will (for \( p \neq 1 \)), using

\[
L = \lim_{k \to \infty} \left( \log k \right) \left( k \left[ \frac{a_k}{a_{k+1}} - 1 \right] - 1 \right) - 1
\]

The process can be continued *ad infinitum* to higher orders. Details are left to the interested reader.
Gauss’s Test
As part of his work on hypergeometric series (), Gauss developed an early test which is closely related to Raabe’s test, but both the hypotheses and conclusion are slightly stronger. (Gauss actually only stated this test when the \( O \left( \frac{1}{k^p} \right) \) term is a rational function in \( k \).) Note that this test never fails (when the hypotheses are satisfied).

**IV.2.4.11. Theorem. [Gauss’s Test]** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with strictly positive terms. Suppose

\[
\frac{a_{k+1}}{a_k} = 1 - \frac{\lambda}{k} + O \left( \frac{1}{k^p} \right)
\]

for some fixed real number \( \lambda \) and some fixed \( p > 1 \). Then

(i) If \( \lambda > 1 \), the series \( \sum_{k=1}^{\infty} a_k \) converges.

(ii) If \( \lambda \leq 1 \), the series \( \sum_{k=1}^{\infty} a_k \) diverges.

**Proof:** Since \( O \left( \frac{1}{k^p} \right) = o \left( \frac{1}{k} \right) \), if \( \lambda \neq 1 \) the result follows from IV.2.4.6. And if \( \lambda = 1 \), we have \( O \left( \frac{1}{k^p} \right) = o \left( \frac{1}{k \log k} \right) \) so the result follows from IV.2.4.9. with \( \mu = 0 \).

Cauchy’s Condensation Test and Yermakov’s Test
These two closely related tests reduce the question of convergence of a series with decreasing nonnegative terms to consideration of a “sparse” subseries. Cauchy’s Condensation Test is a generalization of the argument used in IV.2.2.11.

**IV.2.4.12. Theorem. [Cauchy’s Condensation Test]** Let \( (c_k) \) be a strictly increasing sequence of positive integers such that there is a constant \( M \) with

\[
c_{k+2} - c_{k+1} \leq M(c_{k+1} - c_k)
\]

for all \( k \). Then, if \( \sum_{k=1}^{\infty} a_k \) is an infinite series with nonincreasing nonnegative terms, the two series

\[
\sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} (c_{k+1} - c_k)a_{c_k}
\]

either both converge or both diverge.
Proof: If \( c_1 > 1 \), set \( r = \sum_{k=1}^{c_1-1} a_k \); if \( c_1 = 1 \), set \( r = 0 \). Write \( s_n \) and \( t_n \) for the \( n \)’th partial sums of the two series respectively. Then, for \( n \leq c_m \), we have

\[
s_n \leq s_{c_m} \leq r + (a_{c_1} + a_{c_1+1} + \cdots + a_{c_2-1}) + (a_{c_2} + \cdots + a_{c_3-1}) + \cdots + (a_{c_m} + \cdots + a_{c_{m+1}-1})
\]

\[
\leq r + (c_2 - c_1)a_{c_1} + \cdots + (c_{m+1} - c_m)a_{c_m} = r + t_m
\]

and so, if the second series converges, i.e. \((t_m)\) is bounded, the sequence \((s_n)\) is also bounded so the first series converges. On the other hand, if \( n > c_m \), we have

\[
s_n \geq s_{c_m} \geq (a_{c_1+1} + \cdots + a_{c_2}) + (a_{c_2+1} + \cdots + a_{c_3}) + \cdots + (a_{c_{m-1}+1} + \cdots + a_{c_m})
\]

\[
\geq (c_2 - c_1)a_{c_2} + (c_3 - c_2)a_{c_3} + \cdots + (c_m - c_{m-1})a_{c_m}
\]

\[
Ms_n \geq (c_1 - c_2)a_{c_2} + \cdots + (c_{m+1} - c_m)a_{c_m} = t_m - a_{c_1}
\]

and so if the first series converges, i.e. \((s_n)\) is bounded, the sequence \((t_m)\) is also bounded and the second series converges. 

The most commonly used case of this test is to take \( c_k = 2^k \) for all \( k \) (then we can take \( M = 2 \)). We obtain:

**IV.2.4.13.** Corollary. [Cauchy’s Condensation Test, Common Version] If \( \sum_{k=1}^{\infty} a_k \) is an infinite series with nonincreasing nonnegative terms, the two series

\[
\sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} 2^k a_{2^k}
\]

either both converge or both diverge.

This corollary suffices to show that the harmonic series converges, since the terms of the second series then become all 1, so the second series diverges by the Divergence Criterion. (This is essentially the argument of IV.2.2.11.) The corollary also suffices to show convergence and divergence of the series in (\( ) \).

The test of Yermakov (or Ermakoff) is similar in nature, but is based on the Integral Test and gives a direct criterion for convergence or divergence without making a comparison with another series:

**IV.2.4.14.** Theorem. [Yermakov’s Test] Let \( f \) be a nonincreasing strictly positive function on \([1, \infty)\), and \( \phi \) a strictly increasing differentiable function on \([1, \infty)\) with \( \phi(t) > t \) for all \( t \). Then

(i) If \( \limsup_{t \to \infty} \frac{\phi'(t)f(\phi(t))}{f(t)} < 1 \), then \( \sum_{k=1}^{\infty} f(k) \) converges.
(ii) If $\phi'(t)f(\phi(t)) \geq f(t)$ for all sufficiently large $t$ (in particular, if $\liminf_{t \to \infty} \frac{\phi'(t)f(\phi(t))}{f(t)} > 1$), then $\sum_{k=1}^{\infty} f(k)$ diverges.

**Proof:** (i): It suffices to show that $\int_{c}^{\infty} f(t) dt$ converges for some $c$ by the Integral Test. Under the hypothesis, there is an $r < 1$ and an $a$ such that $\phi'(t)f(\phi(t)) < r \cdot f(t)$ for all $t \geq a$. Then, for any $b > a$ (we restrict to $b > \phi(a)$), we have

$$\int_{a}^{b} f(\phi(t)) \phi'(t) dt < r \int_{a}^{b} f(t) dt$$

so, making the substitution $u = \phi(t)$ in the first integral, we obtain

$$\int_{\phi(a)}^{\phi(b)} f(u) du < r \int_{a}^{b} f(t) dt .$$

Writing the first integral as $\int_{\phi(a)}^{\phi(b)} f(t) dt$, we have

$$(1 - r) \int_{\phi(a)}^{\phi(b)} f(t) dt < r \left[ \int_{a}^{b} f(t) dt - \int_{\phi(a)}^{\phi(b)} f(t) dt \right]$$

$$= r \left[ \int_{a}^{\phi(a)} f(t) dt + \int_{\phi(a)}^{b} f(t) dt - \int_{a}^{b} f(t) dt - \int_{b}^{\phi(b)} f(t) dt \right] = r \left[ \int_{a}^{\phi(a)} f(t) dt - \int_{b}^{\phi(b)} f(t) dt \right]$$

and, since $b < \phi(b)$, the last integral is positive, so we have

$$(1 - r) \int_{\phi(a)}^{\phi(b)} f(t) dt < r \int_{a}^{\phi(a)} f(t) dt .$$

The right side is constant, so we have that the left side is bounded as $b \to +\infty$ (and hence $\phi(b) \to +\infty$). Thus $\int_{\phi(a)}^{\infty} f(t) dt$ converges.

(ii): Suppose $\phi'(t)f(\phi(t)) \geq f(t)$ for all $t \geq a$. Then we have, for any $b > a$,

$$\int_{a}^{b} f(\phi(t)) \phi'(t) dt \geq \int_{a}^{b} f(t) dt .$$

Changing variables as before, we obtain

$$\int_{\phi(a)}^{\phi(b)} f(t) dt \geq \int_{a}^{b} f(t) dt$$

$$\int_{\phi(a)}^{b} f(t) dt + \int_{b}^{\phi(b)} f(t) dt \geq \int_{a}^{\phi(a)} f(t) dt + \int_{\phi(a)}^{b} f(t) dt$$

458
\[
\int_b^{\phi(b)} f(t) \, dt \geq \int_a^{\phi(a)} f(t) \, dt.
\]

The right side is a strictly positive constant, and if \( \int_1^{\infty} f(t) \, dt \) converged, the left side would have to converge to 0 as \( b \to \infty \). Thus \( \int_1^{\infty} f(t) \, dt \) diverges, and the series also diverges by the Integral Test.

This test is commonly written in the special case \( \phi(t) = e^t \):

**IV.2.4.15. Corollary.** (Yermakov’s Test, Common Version) Let \( f \) be a nonincreasing strictly positive function on \([1, \infty)\). Then

(i) If \( \limsup_{t \to \infty} \frac{e^t f(e^t)}{f(t)} < 1 \), then \( \sum_{k=1}^{\infty} f(k) \) converges.

(ii) If \( \liminf_{t \to \infty} \frac{e^t f(e^t)}{f(t)} > 1 \), then \( \sum_{k=1}^{\infty} f(k) \) diverges.

**IV.2.4.16.** This version is sufficient to handle all the \( p \)-series, \( \log^{-p} \)-series, \( \log \log^{-p} \)-series, etc., of (Exercise ()). (Of course, these series can also be handled directly by the Integral Test, as well as by Cauchy’s Condensation Test.)

**IV.2.5. Absolute and Conditional Convergence**

Convergence or divergence of an infinite series is a much more subtle matter if it has both positive and negative terms, since there can be complicated cancellation in the partial sums. In fact, there are convergent series whose terms go to zero arbitrarily slowly. But series which converge because their terms go to zero rapidly (absolutely convergent series) are better behaved and do not have the pathologies possible in the general case where convergence depends delicately on such cancellation.

**IV.2.5.1. Definition.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series. The series \( \sum_{k=1}^{\infty} a_k \) converges absolutely if \( \sum_{k=1}^{\infty} |a_k| \) converges. The series \( \sum_{k=1}^{\infty} a_k \) converges conditionally if \( \sum_{k=1}^{\infty} a_k \) converges but \( \sum_{k=1}^{\infty} |a_k| \) diverges.

Of course, convergence and absolute convergence are the same thing for nonnegative series (and, by IV.2.1.18., also for series with only negative terms). But there is a big difference for general series.

The first observation is that absolute convergence implies convergence. Buried in the proof is the Completeness Property of \( \mathbb{R} \), in the form of the Cauchy Criterion; in fact this result is another equivalent characterization of completeness of \( \mathbb{R} \) (cf. ())).

459
IV.2.5.2. Proposition. Every absolutely convergent series converges.

Proof: Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, and suppose $\sum_{k=1}^{\infty} |a_k|$ converges. Let $\epsilon > 0$, and fix $N$ such that, whenever $N \leq n < m$, we have

$$\sum_{k=n+1}^{m} |a_k| < \epsilon$$

(IV.2.1.11.). Then, by the triangle inequality we have for $N \leq n < m$,

$$\left| \sum_{k=n+1}^{m} a_k \right| \leq \sum_{k=n+1}^{m} |a_k| < \epsilon$$

so $\sum_{k=1}^{\infty} a_k$ converges by the Cauchy Criterion (IV.2.1.11.).

For an alternate proof, see IV.2.16.10.

IV.2.5.3. Thus there are only three possibilities for an infinite series $\sum_{k=1}^{\infty} a_k$:

(i) The series converges absolutely, i.e. $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} |a_k|$ both converge.

(ii) The series converges conditionally, i.e. $\sum_{k=1}^{\infty} a_k$ converges but $\sum_{k=1}^{\infty} |a_k|$ diverges.

(iii) The series diverges, i.e. both $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} |a_k|$ diverge.

Of course, conditional convergence is only a possibility when there are infinitely many positive terms and also infinitely many negative terms.

The Fundamental Principle of Nonnegative Series (IV.2.2.2.) asserts that a series converges absolutely if and only if its terms go to zero sufficiently rapidly.

We have not yet seen an example of a conditionally convergent series, but there are (rather degenerate) simple examples where the terms go to zero arbitrarily slowly:

IV.2.5.4. Example. Let $(b_k)$ be a sequence of numbers converging to zero. For each $k$, set $a_{2k-1} = b_k$ and $a_{2k} = -b_k$. Then the partial sums $s_n$ of the series $\sum_{k=1}^{\infty} a_k$ satisfy $s_n = 0$ for $n$ even and $s_n = b_m$ if
\( n = 2m - 1 \) is odd. Thus \( s_n \to 0 \) and the series converges (with sum 0). The series \( \sum_{k=1}^{\infty} a_k \) converges absolutely if and only if \( \sum_{k=1}^{\infty} |b_k| \) converges; hence, for example, if \( b_k = \frac{1}{k} \), the series \( \sum_{k=1}^{\infty} a_k \) converges conditionally.

We will see more interesting examples of conditionally convergent series.

To further illustrate the nature of conditionally convergent series, we note:

**IV.2.5.5. Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) be a conditionally convergent series, and let \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) be the infinite series whose terms are the positive and negative terms of \( \sum_{k=1}^{\infty} a_k \) respectively. Then \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) both diverge (i.e. the partial sums of both series are unbounded).

**Proof:** The series \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) cannot both converge, since then we would have, for any \( n \),

\[
\sum_{k=1}^{n} |a_k| \leq \sum_{k=1}^{\infty} b_k + \sum_{k=1}^{\infty} (-c_k)
\]

and \( \sum_{k=1}^{\infty} a_k \) would converge absolutely. If \( \sum_{k=1}^{\infty} b_k \) diverges but \( \sum_{k=1}^{\infty} c_k \) converges, there is an \( m > 0 \) such that all partial sums of \( \sum_{k=1}^{\infty} c_k \) are \( \geq -m \); for any \( M \) all sufficiently large partial sums of \( \sum_{k=1}^{\infty} b_k \) are \( \geq M \), so for all sufficiently large \( n \) we have

\[
\sum_{k=1}^{n} a_k \geq M - m
\]

and so \( \sum_{k=1}^{\infty} a_k \) cannot converge. The argument in the other case is almost identical.

What the proof actually shows is the following slightly more precise result:

**IV.2.5.6. Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with infinitely many terms of each sign, and let \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) be the infinite series whose terms are the positive and negative terms of \( \sum_{k=1}^{\infty} a_k \) respectively. Then
(i) If $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ both converge, then $\sum_{k=1}^{\infty} a_k$ converges absolutely.

(ii) If one of $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ converges and the other diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

Tests for Absolute Convergence

Any of the tests for convergence of nonnegative series become tests for absolute convergence if absolute values are inserted. We explicitly state the Root and Ratio Tests, and the Comparison Test:

IV.2.5.7. Theorem. [Root Test] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, and set

$$r = \limsup_{k \to \infty} \sqrt[k]{|a_k|} = \limsup_{k \to \infty} |a_k|^{1/k}.$$ 

Then

(i) If $r < 1$, then the series converges absolutely.

(ii) If $r > 1$, then the terms do not approach zero and the series diverges.

(iii) If $r = 1$, then the series may converge absolutely, converge conditionally, or diverge, and no conclusion may be drawn (the test fails).

IV.2.5.8. Theorem. [Ratio Test] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with nonzero terms. Then

(i) If $\limsup_{k \to \infty} \left| \frac{a_{k+1}}{a_k} \right| < 1$, then the series converges absolutely.

(ii) If $\liminf_{k \to \infty} \left| \frac{a_{k+1}}{a_k} \right| > 1$, then the terms do not approach zero and the series diverges.

(iii) If $\liminf_{k \to \infty} \left| \frac{a_{k+1}}{a_k} \right| \leq 1 \leq \limsup_{k \to \infty} \left| \frac{a_{k+1}}{a_k} \right|$, then the series may converge absolutely, converge conditionally, or diverge, and no conclusion may be drawn (the test fails).

The only part of these theorems which may not be obvious from earlier work is the assertion in (iii) of each that the series can converge conditionally. The series of IV.2.5.4. with $b_k = \frac{1}{k}$ is an example in both cases. (In fact, any conditionally convergent series must fall into case (iii) in each theorem.)

462
IV.2.5.9. **Theorem.** [Comparison Test] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series. Suppose $|a_k| \leq |b_k|$ for all $k$. If $\sum_{k=1}^{\infty} b_k$ converges absolutely, then $\sum_{k=1}^{\infty} a_k$ converges absolutely.

IV.2.5.10. **Caution:** We do not get a comparison test for simple convergence of series, i.e. if $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ are infinite series, with $|a_k| \leq |b_k|$ for all $k$, and $\sum_{k=1}^{\infty} b_k$ converges, then we cannot conclude that $\sum_{k=1}^{\infty} a_k$ converges. There are many counterexamples from the Alternating Series Test (IV.2.5.13.), e.g. $a_k = \frac{1}{k}$ and $b_k = \frac{(-1)^k}{k}$ (IV.2.5.14.(i)).

Combining the Comparison Test and the Triangle Inequality with IV.2.1.16., we obtain:

IV.2.5.11. **Proposition.** Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be absolutely convergent infinite series. Then the series $\sum_{k=1}^{\infty} (a_k + b_k)$ converges absolutely, and $\sum_{k=1}^{\infty} (a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k$.

It is more straightforward to see that if $\sum_{k=1}^{\infty} a_k$ is an absolutely convergent infinite series, and $c \in \mathbb{R}$, then $\sum_{k=1}^{\infty} ca_k$ is also absolutely convergent.

**Alternating Series**

There is a simple criterion for the convergence of infinite series in which the terms alternate sign. In such series the cancellation of positive and negative terms is regular enough to make the series converge under quite mild restrictions.

IV.2.5.12. **Definition.** An alternating series is an infinite series in which successive terms are of opposite signs.

The next theorem is the main result of this subsection. It is a special case of Dirichlet’s Test, but the proof is considerably simpler in this case, essentially reducing to the Nested Intervals Property of $\mathbb{R}$. This is also the most commonly used test for series with terms of varying sign.
IV.2.5.13. **Theorem. [Alternating Series Test]** Let $\sum_{k=1}^{\infty} a_k$ be an alternating series. If

(i) The terms decrease in absolute value, i.e. $|a_{k+1}| \leq |a_k|$ for all $k$, and

(ii) $\lim_{k \to \infty} a_k = 0$,

then the series $\sum_{k=1}^{\infty} a_k$ converges. The sum lies between any two consecutive partial sums.

Note that this result says nothing about absolute or conditional convergence; either type can occur under the hypotheses.

**Proof:** By discarding the first term of the series if necessary (cf. IV.2.1.6.), we may assume $a_1 > 0$, and hence $a_k > 0$ for $k$ odd, $a_k < 0$ for $k$ even. (This is not an essential restriction, but simplifies the notation of the proof.)

Let $s_n$ be the $n$'th partial sum of the series. By (i), we have

$s_2 \leq s_4 \leq s_6 \leq \cdots \leq s_5 \leq s_3 \leq s_1$

and thus for any $m$ we have $s_{2m} \leq s_k \leq s_{2m-1}$ for all $k \geq 2m - 1$. (This can be proved carefully by induction on $k$.) Thus, if $k, j \geq 2m - 1$, we have

$|s_k - s_j| \leq s_{2m-1} - s_{2m} = |a_{2m}|$.

For any $\epsilon > 0$, we have $|a_{2m}| < \epsilon$ for all sufficiently large $m$ by (ii), so the sequence $(s_n)$ is a Cauchy sequence and hence converges; the limit $s$ satisfies $s_{2m} \leq s \leq s_{2m-1}$ for any $m$.

IV.2.5.14. **Examples.** (i) The Alternating Series Test provides more interesting examples of conditionally convergent series than IV.2.5.4.; in fact, in the Alternating Series Test the terms can also go to zero arbitrarily slowly. For example, the alternating $p$-series

$$\sum_{k=1}^{\infty} \frac{(-1)^k}{k^p}$$

converges for any $p > 0$; the convergence is absolute for $p > 1$ and conditional for $0 < p \leq 1$.

(ii) In the Alternating Series Test, both hypotheses are necessary. Hypothesis (ii) is clearly necessary by the Divergence Criterion, and is easily remembered. It is more common to forget to check (i), but some version of (i) is necessary: for example, set $a_k = \frac{1}{k}$ for $k$ odd and $a_k = -\frac{1}{k}$ for $k$ even. Then the series $\sum_{k=1}^{\infty} a_k$ is alternating, and the terms approach zero; but the series of positive terms diverges and the series of negative terms converges, and hence $\sum_{k=1}^{\infty} a_k$ diverges by IV.2.5.6.

It does suffice, of course, that conditions (i) and (ii) (and even the fact that the series alternates) just hold from some point on in the series, by IV.2.1.6.; the statement that the sum of the series lies between successive partial sums then only holds beyond the point where the terms satisfy the hypotheses.
IV.2.6. Hypergeometric Series

Hypergeometric series are a generalization of geometric series which arise naturally in applications: many applied problems give rise to power series which are hypergeometric ( ). The theory of hypergeometric series was primarily developed by Gauss.

IV.2.6.1. Definition. A hypergeometric series is an infinite series \( \sum_{k=1}^{\infty} a_k \) in which all \( a_k \) are nonzero, and for which there are polynomials \( p \) and \( q \) such that

\[
\frac{a_{k+1}}{a_k} = \frac{p(k)}{q(k)}
\]

for all \( k \). The summation may also start at 0 or another integer.

IV.2.6.2. Examples. (i) A geometric series is a hypergeometric series where \( p \) and \( q \) are constant: if the ratio is \( x \), take \( p(k) = x \) (\( x \) is a constant!) and \( q(k) = 1 \).

(ii) For fixed \( x \), the series

\[
\sum_{k=0}^\infty \frac{x^k}{k!}
\]

for \( e^x \) is hypergeometric: we have

\[
\frac{a_{k+1}}{a_k} = \frac{k!x^{k+1}((k+1)!x^k)}{x^{k+1}} = \frac{x}{k+1}
\]

so we may take \( p(k) = x \) (constant) and \( q(k) = k + 1 \). The power series for \( \sin x \) and \( \cos x \) (\( x \) fixed) are also hypergeometric.

(iii) The basic hypergeometric series considered by Gauss was

\[
\sum_{k=0}^{\infty} \frac{\alpha(\alpha + 1)\cdots(\alpha + k - 1)\beta(\beta + 1)\cdots(\beta + k - 1)x^k}{k!(\gamma + 1)\cdots(\gamma + k - 1)}
\]

where \( \alpha, \beta, \gamma \) are real numbers which are not integers \( \leq 0 \), \( x \) is a fixed real number, and the \( k = 0 \) term is taken to be 1. We then have

\[
\frac{a_{k+1}}{a_k} = \frac{(\alpha + k)(\beta + k)x}{(1 + k)(\gamma + k)}
\]

so the series is hypergeometric. (The geometric series is the special case where \( \alpha = 1 \) and \( \beta = \gamma \).)

IV.2.6.3. Since \( p(k) \) and \( q(k) \) have constant sign for sufficiently large \( k \), it follows that the terms of a hypergeometric series either are eventually constant in sign or eventually alternate in sign.
IV.2.6.4. The Ratio Test is tailor-made for determining convergence of hypergeometric series, since

\[ \lim_{k \to \infty} \frac{p(k)}{q(k)} \]

always exists (as an extended real number) and is easily computed. (In fact, the Ratio Test was originally specifically developed largely to analyze hypergeometric series.) Thus the series converges absolutely if this limit is less than one in absolute value, and diverges if greater than one. In particular, we have:

IV.2.6.5. PROPOSITION. Let \( \sum_{k=1}^{\infty} a_k \) be a hypergeometric series with

\[ \frac{a_{k+1}}{a_k} = \frac{p(k)}{q(k)} \]

for all \( k \). Let \( c \) and \( d \) be the leading coefficients of \( p \) and \( q \) respectively. Then

(i) If the degree of \( q \) is greater than the degree of \( p \), the series converges absolutely.

(ii) If the degree of \( p \) is greater than the degree of \( q \), the series diverges.

(iii) If \( p \) and \( q \) have the same degree and \( |\frac{c}{d}| < 1 \), the series converges absolutely.

(iv) If \( p \) and \( q \) have the same degree and \( |\frac{c}{d}| > 1 \), the series diverges.

IV.2.6.6. COROLLARY. The basic hypergeometric series of IV.2.6.2.(iii) converges absolutely if \( |x| < 1 \) and diverges if \( |x| > 1 \).

The most interesting situation is when the Ratio Test fails, i.e. when \( p \) and \( q \) have the same degree and \( |\frac{c}{d}| = 1 \). (In the basic hypergeometric series, this is the case \( |x| = 1 \).) GAUSS was led to develop his test (IV.2.4.11.), which always applies in this situation.

IV.2.6.7. Suppose \( p \) and \( q \) have the same degree \( m \) and \( |\frac{c}{d}| = 1 \). Then

\[ \left| \frac{a_{k+1}}{a_k} \right| = \left| \frac{p(k)/c}{q(k)/d} \right| = \left| \frac{k^m + ak^{m-1} + r(k)}{k^m + bk^{m-1} + s(k)} \right| \]

where \( r \) and \( s \) are polynomials of degree \( \leq m - 2 \). Since the quotient in the last term is positive for large \( k \), the absolute values can be removed. We then have

\[ 1 - \left| \frac{a_{k+1}}{a_k} \right| = \frac{(k^m + bk^{m-1} + s(k)) - (k^m + ak^{m-1} + r(k))}{k^m + bk^{m-1} + s(k)} = \frac{(b - a)k^{m-1} + (r(k) - s(k))}{k^m + bk^{m-1} + s(k)} \]

and so

\[ \left| \frac{a_{k+1}}{a_k} \right| = 1 - \frac{(b - a)k^{m-1} + (s(k) - r(k))}{k^m + bk^{m-1} + s(k)} \]

466
\[ 1 - \frac{b - a}{k} + \left[ \frac{b - a}{k} - \frac{(b - a)k^{m-1} + ((s(k) - r(k))}{k^m + bk^{m-1} + s(k)} \right] \]

\[ = 1 - \frac{b - a}{k} + \frac{(b - a)(k^m + bk^{m-1} + s(k))}{k(k^m + bk^{m-1} + s(k))} - \frac{(b - a)k^m + k(s(k) - r(k))}{k(k^m + bk^{m-1} + s(k))} \]

\[ = 1 - \frac{b - a}{k} + \frac{\phi(k)}{k(k^m + bk^{m-1} + s(k))} \]

where \( \phi \) is a polynomial of degree \( \leq m - 1 \). The last term is \( O \left( \frac{1}{k^2} \right) \), so we have

\[ \frac{|a_{k+1}|}{|a_k|} = 1 - \frac{b - a}{k} + O \left( \frac{1}{k^2} \right). \]

Applying Gauss’s Test, we obtain:

**IV.2.6.8. Theorem.**

(i) If \( b - a > 1 \), the series converges absolutely.

(ii) If \( b - a \leq 1 \), the series does not converge absolutely.

This result completely settles the convergence question when the terms of the series eventually have constant sign. \textsc{Gauss} went on to completely analyze the case where the terms alternate in sign. (We are not following \textsc{Gauss}’s approach; see \textsc{Bre07} for a nice exposition of \textsc{Gauss}’s arguments, including a proof of the next result.)

**IV.2.6.9. Theorem.** If the terms of the series alternate in sign, then

(i) If \( b < a \), the terms go to infinity in absolute value so the series diverges.

(ii) If \( b = a \), the absolute values of the terms approach a nonzero finite limit and the series diverges.

(iii) If \( b > a \), the series converges.

The convergence in (iii) is conditional if \( a < b \leq a + 1 \) and absolute if \( b > a + 1 \).

We can then obtain the convergence for the basic hypergeometric series:
IV.2.6.10. Corollary. Let $\sum_{k=1}^{\infty} a_k$ be the hypergeometric series of IV.2.6.2.(iii) with parameters $\alpha, \beta, \gamma$, and $|x| = 1$. Then

(i) If $\alpha + \beta < \gamma$, the series converges absolutely.

(ii) If $\alpha + \beta - \gamma \geq 1$, the series diverges.

(iii) If $x = 1$ and $\alpha + \beta \geq \gamma$, the series diverges.

(iv) If $x = -1$ and $\gamma \leq \alpha + \beta < \gamma + 1$, the series converges conditionally.

Proof: We have

$$\frac{|a_{k+1}|}{|a_k|} = \frac{k^2 + (\alpha + \beta)k + \alpha \beta}{k^2 + (1 + \gamma)k + \gamma}$$

by a simple calculation. The terms of the series are eventually constant in sign if $x > 0$ and eventually alternate if $x < 0$.

IV.2.7. Scaling of Series

In this subsection, we examine what happens to the convergence of a series $\sum_{k=1}^{\infty} a_k$ when the terms $a_k$ are scaled by factors $c_k$. Of course, when the $c_k$ are constant, the scaling does not affect convergence, but when the $c_k$ vary, convergence can be affected very subtly.

The most obvious example where convergence can be affected is where the $c_k$ are unbounded. For bounded scaling, the simplest case where convergence changes is where $\sum_{k=1}^{\infty} a_k$ is a conditionally convergent alternating series and $c_k = (-1)^k$. But even if the $c_k$ are all nonnegative and/or converge to a limit (even 0), convergence or divergence can change by the scaling:

IV.2.7.1. Examples. (i) Let $a_k = \frac{(-1)^k}{\sqrt{k}}$ for all $k$. Set $c_k = \frac{1}{\sqrt{k}}$ for $k$ odd, and $c_k = \frac{1}{k}$ for $k$ even. Then $\sum_{k=1}^{\infty} a_k$ converges by the Alternating Series Test, but the series $\sum_{k=1}^{\infty} c_k a_k$ diverges since the series of positive terms (the even ones) converges, and the series of negative terms (the odd ones) diverges (IV.2.5.6.(ii)).

(ii) Let $a_k = c_k = \frac{(-1)^k}{\sqrt{k}}$ for all $k$. Then $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} c_k$ both converge by the Alternating Series Test, but $\sum_{k=1}^{\infty} c_k a_k$ is the harmonic series, which diverges.

The simplest general result is for absolutely convergent series:
IV.2.7.2. PROPOSITION. Let \( \sum_{k=1}^{\infty} a_k \) be an absolutely convergent series, and \((c_k)\) a bounded sequence of real numbers. Then \( \sum_{k=1}^{\infty} c_k a_k \) also converges absolutely.

This result applies in particular if all the \( a_k \) and \( c_k \) are nonnegative.

PROOF: If \( |c_k| \leq m \) for all \( k \), compare \( \sum_{k=1}^{\infty} |c_k a_k| \) with \( \sum_{k=1}^{\infty} m|a_k| \).

IV.2.7.3. Conversely, if \( \sum_{k=1}^{\infty} a_k \) is an infinite series and \((d_k)\) a sequence of real numbers such that \( \inf d_k > 0 \), and \( \sum_{k=1}^{\infty} d_k a_k \) is absolutely convergent, then \( \sum_{k=1}^{\infty} a_k \) is also absolutely convergent, seen by applying IV.2.7.2. to scaling by \( c_k = \frac{1}{d_k} \).

On the other hand, if \((d_k)\) is a sequence of positive numbers with \( \inf d_k = 0 \), there is an infinite series \( \sum_{k=1}^{\infty} a_k \) with positive terms such that \( \sum_{k=1}^{\infty} a_k \) diverges but \( \sum_{k=1}^{\infty} d_k a_k \) converges absolutely. And if \((c_k)\) is an unbounded sequence of positive numbers, there is an absolutely convergent series \( \sum_{k=1}^{\infty} a_k \) with positive terms such that \( \sum_{k=1}^{\infty} c_k a_k \) diverges. See Exercise IV.2.16.14.

Abel's Partial Summation

To analyze the case where \( \sum_{k=1}^{\infty} a_k \) is not absolutely convergent, it is convenient to use a summation formula due to Abel in 1826. This formula is a discrete version of Integration by Parts, and will be used roughly to interchange the roles of the \( a_k \) and \( c_k \).

IV.2.7.4. PROPOSITION. Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series, with partial sums \( s_n = \sum_{k=1}^{n} a_k \). Set \( s_0 = 0 \). If \((c_k)\) is a sequence of real numbers, then for any \( n \geq 0, m \geq 1 \),

\[
\sum_{k=n+1}^{n+m} c_k a_k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_k - c_{n+1} s_n + c_{n+m+1} s_{n+m}.
\]

PROOF: For any \( k \geq 1 \) we have

\[
c_k a_k = c_k (s_k - s_{k-1}) = (c_k - c_{k+1}) s_k - c_k s_{k-1} + c_{k+1} s_k
\]
and summing from $n + 1$ to $n + m$ the last two terms telescope to $c_{n+m+1}s_{n+m} - c_{n+1}s_n$.

Essentially the same proof shows the following technical strengthening of the formula:

**IV.2.7.5. Proposition.** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, with partial sums $s_n = \sum_{k=1}^{n} a_k$. Set $s_0 = 0$. If $0 \leq r \leq n$, set $s_{n;r} = s_n - s_r = \sum_{k=r+1}^{n} a_k$. If $(c_k)$ is a sequence of real numbers, then for any $n \geq 0$, $m \geq 1$, and any $r$, $0 \leq r \leq n$,

$$\sum_{k=n+1}^{n+m} c_k a_k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_{k;r} - c_{n+1}s_{n;r} + c_{n+m+1}s_{n+m;r}.$$ 

**Proof:** For any $k \geq r + 1$ we have

$$c_k a_k = c_k(s_{k;r} - s_{k-1;r}) = (c_k - c_{k+1}) s_{k;r} - c_{k}s_{k-1;r} + c_{k+1}s_{k;r}$$

and summing from $n + 1$ to $n + m$ the last two terms telescope to $c_{n+m+1}s_{n+m;r} - c_{n+1}s_{n;r}$. 

Proposition IV.2.7.4. is the case $r = 0$. The case $r = n$ is also useful (cf. V.15.3.16.). Note that if $r = n$, one term disappears from the right side since $s_{n;n} = 0$.

The first consequence of Abel’s summation formula is the following general test, which we will call the **Scaling Test**, a nonstandard name:

**IV.2.7.6. Theorem.** [Scaling Test] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with partial sums $s_n$, and $(c_k)$ a sequence of numbers. If

(i) $\sum_{k=1}^{\infty} (c_k - c_{k+1}) s_k$ converges and

(ii) $\lim_{m \to \infty} c_{m+1}s_m$ exists,

Then the series $\sum_{k=1}^{\infty} c_k a_k$ converges.

**Proof:** Abel’s partial summation formula for $n = 0$ and any $m$ gives

$$\sum_{k=1}^{m} c_k a_k = \sum_{k=1}^{m} (c_k - c_{k+1}) s_k + c_{m+1}s_m.$$ 

By hypothesis, the limit of the right side exists as $m \to \infty$. 

470
Abel’s Test
The first special case of the Scaling Test we consider is due to Abel:

**IV.2.7.7. Theorem. [Abel’s Test]** Let $\sum_{k=1}^{\infty} a_k$ be a convergent infinite series, and $(c_k)$ a bounded monotone sequence of numbers. Then $\sum_{k=1}^{\infty} c_k a_k$ converges.

**Proof:** We have that $(c_k)$ converges, and since $(s_m)$ converges by hypothesis, $(c_{m+1}s_m)$ converges. We also have that $\sum_{k=1}^{\infty} (c_k - c_{k+1})$ converges, and hence converges absolutely since the terms are all of the same sign.

Since the $s_k$ are bounded, $\sum_{k=1}^{\infty} (c_k - c_{k+1})s_k$ converges (absolutely) by IV.2.7.2.; thus the hypotheses of the Scaling Test are satisfied. $\blacksquare$

Dirichlet’s Test
Another consequence of the Scaling Test is called Dirichlet’s Test, although it has been suggested that it was probably also known to Abel previously:

**IV.2.7.8. Theorem. [Dirichlet’s Test]** Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with bounded partial sums, and $(c_k)$ a nonincreasing sequence of (nonnegative) numbers converging to zero. Then $\sum_{k=1}^{\infty} c_k a_k$ converges.

The Alternating Series Test is the special case where $a_k = (-1)^k$ or $(-1)^{k+1}$.

**Proof:** As in the previous proof, the series $\sum_{k=1}^{\infty} (c_k - c_{k+1})s_k$ converges absolutely, and we have $\lim_{m \to \infty} c_{m+1}s_m = 0$ by ( ). Thus the hypotheses of the Scaling Test are satisfied. $\blacksquare$

As a nice application, we have:

**IV.2.7.9. Example.** If $x$ is any fixed real number, then the series $\sum_{k=1}^{\infty} \frac{\sin kx}{k}$ converges.
Note that for general \( x \), \( \sin kx \) jumps around through \([-1, 1]\) in an apparently chaotic manner as \( k \) increases, so the terms vary in size and sign in a seemingly unpredictable way. But the essence of the argument is that there is enough regularity to insure that the partial sums of the series

\[
\sum_{k=1}^{\infty} \sin kx
\]

are bounded. We can then take \( c_k = \frac{1}{k} \) and use Dirichlet’s Test.

To show that these partial sums are bounded, we use a clever argument due to Dirichlet involving the sum formula for \( \cos(\alpha + \beta) \). By periodicity, it suffices to prove the result for \( 0 < x < 2\pi \) (the case \( x = 0 \) is trivial). Fix \( x \in (0, 2\pi) \) and set

\[
s_n = \sum_{k=1}^{n} \sin kx.
\]

For any \( k \), we have

\[
\cos \left( \left[k - \frac{1}{2}\right] x \right) - \cos \left( \left[k + \frac{1}{2}\right] x \right)
= \left[ \cos kx \cos \frac{x}{2} + \sin kx \sin \frac{x}{2} \right] - \left[ \cos kx \cos \frac{x}{2} - \sin kx \sin \frac{x}{2} \right]
= 2 \sin kx \sin \frac{x}{2}
\]

so we have, by telescoping,

\[
2s_n \sin \frac{x}{2} = \sum_{k=1}^{n} 2 \sin kx \sin \frac{x}{2}
\]

\[
= \sum_{k=1}^{n} \left[ \cos \left( \left[k - \frac{1}{2}\right] x \right) - \cos \left( \left[k + \frac{1}{2}\right] x \right) \right]
= \cos \frac{x}{2} - \cos \left( \left[n + \frac{1}{2}\right] x \right).
\]

Thus we have, for all \( n \),

\[
|s_n| \leq \frac{\left| \cos \frac{x}{2} - \cos \left( \left[n + \frac{1}{2}\right] x \right) \right|}{2 \left| \sin \frac{x}{2} \right|} \leq \frac{1}{\left| \sin \frac{x}{2} \right|}.
\]

In fact, we get that

\[
\sum_{k=1}^{\infty} \frac{\sin kx}{k^p}
\]

converges for any \( x \in \mathbb{R} \) and \( p > 0 \), and that

\[
\sum_{k=2}^{\infty} \frac{\sin kx}{\log k}
\]

converges for any \( x \in \mathbb{R} \).
IV.2.8. Subseries and Rearrangements of Series

There are two natural procedures for manufacturing a new infinite series out of an old one:

IV.2.8.1. Definition. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series.

(i) An infinite series $\sum_{k=1}^{\infty} b_k$ is a subseries of $\sum_{k=1}^{\infty} a_k$ if there is a strictly increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all $k$.

(ii) An infinite series $\sum_{k=1}^{\infty} b_k$ is a rearrangement of $\sum_{k=1}^{\infty} a_k$ if there is a bijection $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all $k$.

(iii) An infinite series $\sum_{k=1}^{\infty} b_k$ is a subarrangement of $\sum_{k=1}^{\infty} a_k$ if there is a one-to-one function $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all $k$.

A subseries or a rearrangement is a special case of a subarrangement of a series. We have seen natural examples of subseries in IV.2.5.5. A natural example of a rearrangement of a series is to write the terms of a strictly positive convergent series $\sum_{k=1}^{\infty} a_k$ in nonincreasing order (check that this is a rearrangement: in particular, for any $k$ there can be only finitely many $j$ for which $a_j \geq a_k$).

Note that in a subarrangement, each term of the original series occurs at most once (if there are repeated terms in the original series, the term can be repeated in the subarrangement up to as many times as in the original series). In a rearrangement, each term in the original series occurs exactly once in the rearrangement.

The $\phi$ in a subarrangement is uniquely determined. (If there are repeated terms in the original series, more than one subarrangement can give the same new series.) If $\sum_{k=1}^{\infty} b_k$ is a rearrangement of $\sum_{k=1}^{\infty} a_k$ via $\phi$, then $\sum_{k=1}^{\infty} a_k$ is a rearrangement of $\sum_{k=1}^{\infty} b_k$ via $\phi^{-1}$.

Some authors rather colorfully call a rearrangement of a series a derangement of the series.

We have seen in IV.2.5.5. that a subseries of a convergent series need not converge. We will see () that a rearrangement of a convergent series need not converge, and even if it does it need not have the same sum. These difficulties do not occur for convergent nonnegative series (or for absolutely convergent series – see ()).
IV.2.8.2. **Proposition.** Let \( \sum_{k=1}^{\infty} a_k \) be a convergent nonnegative infinite series. Then any subarrangement \( \sum_{k=1}^{\infty} b_k \) of \( \sum_{k=1}^{\infty} a_k \) also converges, and satisfies

\[ \sum_{k=1}^{\infty} b_k \leq \sum_{k=1}^{\infty} a_k . \]

**Proof:** Let \( M = \sum_{k=1}^{\infty} a_k \). Any partial sum \( t_n \) of the nonnegative series \( \sum_{k=1}^{\infty} b_k \) consists of some, but not necessarily all, of the terms in a corresponding larger partial sum \( s_m \) of \( \sum_{k=1}^{\infty} a_k \) (where \( m = \max\{\phi(1), \ldots, \phi(n)\} \)), and hence \( t_n \leq s_m \leq M \). Thus the partial sums of \( \sum_{k=1}^{\infty} b_k \) are bounded by \( M \), so \( \sum_{k=1}^{\infty} b_k \leq M \). \( \Diamond \)

IV.2.8.3. **Corollary.** Let \( \sum_{k=1}^{\infty} a_k \) be an absolutely convergent infinite series. Then any subarrangement \( \sum_{k=1}^{\infty} b_k \) of \( \sum_{k=1}^{\infty} a_k \) also converges absolutely.

**Proof:** \( \sum_{k=1}^{\infty} |b_k| \) is a subarrangement of \( \sum_{k=1}^{\infty} |a_k| \). \( \Diamond \)

IV.2.8.4. **Corollary.** Let \( \sum_{k=1}^{\infty} a_k \) be a convergent nonnegative infinite series. Then any rearrangement \( \sum_{k=1}^{\infty} b_k \) of \( \sum_{k=1}^{\infty} a_k \) also converges, and satisfies

\[ \sum_{k=1}^{\infty} b_k = \sum_{k=1}^{\infty} a_k . \]

This result generalizes to absolutely convergent series:
IV.2.8.5. **Theorem.** Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any rearrangement $\sum_{k=1}^{\infty} b_k$ of $\sum_{k=1}^{\infty} a_k$ also converges absolutely, and satisfies

$$\sum_{k=1}^{\infty} b_k = \sum_{k=1}^{\infty} a_k.$$ 

**Proof:** $\sum_{k=1}^{\infty} b_k$ converges absolutely by IV.2.8.3.

Let $s = \sum_{k=1}^{\infty} a_k$, and write $s_n = \sum_{k=1}^{n} a_k$ and $t_n = \sum_{k=1}^{n} b_k$ for the partial sums of the series. We will show that $t_n \to s$, i.e. for every $\epsilon > 0$ there is an $N$ such that $|t_n - s| < \epsilon$ for all $n > N$.

Let $\epsilon > 0$, and fix $m$ such that

$$|s_m - s| < \frac{\epsilon}{2} \quad \text{and} \quad \sum_{k=m+1}^{\infty} |a_k| < \frac{\epsilon}{2}$$

(the second condition is possible since $\sum_{k=1}^{\infty} a_k$ converges absolutely; in fact, the first condition follows automatically from the second). Now let

$$N = \max\{\phi(k) : 1 \leq k \leq m\}$$

where $\phi : \mathbb{N} \to \mathbb{N}$ is the bijection of the rearrangement. Now suppose $n > N$. We have that

$$t_n - s_m = \sum_{k=1}^{n} b_k - \sum_{k=1}^{m} a_k$$

is a sum consisting of various $a_k$’s and their negatives. Some $a_k$’s occur in both sums and therefore cancel; this happens for all $a_k$ with $k \leq m$ by the choice of $N$. Thus we have

$$t_n - s_m = \sum_{j=1}^{r} (\pm a_{k_j})$$

for some $r$ and some $k_j$’s, with $k_j > m$ for all $j$. So

$$|t_n - s_m| \leq \sum_{j=1}^{r} |a_{k_j}| \leq \sum_{k=m+1}^{\infty} |a_k| < \frac{\epsilon}{2}$$

and we get

$$|t_n - s| \leq |t_n - s_m| + |s_m - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

475
for all \( n > N \).

This is in stark contrast to what happens in a conditionally convergent series. If \( \sum_{k=1}^{\infty} a_k \) is conditionally convergent, then the series \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) of positive and negative terms, respectively, both diverge. Thus, if the series is rearranged to begin with many (or few) positive terms, then many (or few) negative ones, etc., the partial sums of the rearranged series can be made to oscillate wildly or to converge to any limit at all. Precisely:

**IV.2.8.6. Theorem.** Let \( \sum_{k=1}^{\infty} a_k \) be a conditionally convergent series, and let \( (u_n) \) and \( (v_n) \) be sequences of real numbers with \( v_n \leq u_n \) and \( v_n \leq u_{n+1} \) for all \( n \). Then there is a rearrangement \( \sum_{k=1}^{\infty} d_k \) of \( \sum_{k=1}^{\infty} a_k \) and natural numbers \( n_1 < m_1 < n_2 < m_2 < \cdots \) such that

(i) The partial sums \( t_n \) of \( \sum_{k=1}^{\infty} d_k \) increase for \( 1 \leq n \leq n_1 \) and for \( m_j \leq n \leq n_{j+1} \) for all \( j \), and decrease for \( n_j < n \leq m_j \) for all \( j \).

(ii) \( t_{n_j} \geq u_j \) and \( t_{m_j} \leq v_j \) for all \( j \).

(iii) \( \limsup_{n \to \infty} t_n = \limsup_{j \to \infty} u_j \) and \( \liminf_{n \to \infty} t_n = \liminf_{j \to \infty} v_j \).

**Proof:** Set \( d_1 = b_1 \), and \( d_k = b_k \) until the first index \( n_1 \) is reached with \( t_{n_1} = \sum_{k=1}^{n_1} b_k \geq u_1 \) (there is such an \( n_1 \) since \( \sum_{k=1}^{\infty} b_k \) diverges). Then \( u_1 \leq t_{n_1} \leq u_1 + b_{n_1} \). Set \( d_{n_1+1} = c_1 \), and \( d_{n_1+k} = c_k \) until the first index \( m_1 > n_1 \) is reached with \( t_{m_1} \leq v_1 \) (there is such an \( m_1 \) since \( \sum_{k=1}^{\infty} c_k \) diverges). Then \( v_1 + c_{m_1} \leq t_{m_1} \leq v_1 \). Set \( d_{m_1+1} = a_{n_1+1} \) and \( d_{m_1+k} = b_{n_1+k} \) until the first index \( n_2 > m_1 \) is reached with \( t_{n_2} \geq v_2 \). Continue the process inductively to define \( d_k, n_j, \) and \( m_j \) for all \( j \); the construction is possible at each step since all tails of \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) diverge. We have \( u_j \leq t_{n_j} \leq u_j + b_{r_j} \), where \( r_j = n_1 + \sum_{i=1}^{j-1} (n_{i+1} - n_i) \) is the number of \( b_k \) used in the first \( n_j \) steps, and similarly \( v_j + c_{s_j} \leq t_{m_j} \leq v_j \), where \( s_j = \sum_{i=1}^{j-1} (m_i - n_i) \) is the number of \( c_k \) used. Then \( r_j, s_j \) are strictly increasing, so \( b_{r_j} \to 0 \) and \( c_{s_j} \to 0 \) since \( a_k \to 0 \).

By construction, (i) and (ii) are satisfied for the rearrangement \( \sum_{k=1}^{\infty} d_k \) of \( \sum_{k=1}^{\infty} a_k \). We need only check (iii).

It is clear that

\[
\limsup_{n \to \infty} t_n = \limsup_{j \to \infty} t_{n_j}
\]

476
and we have

\[ u := \limsup_{j \to \infty} u_j \leq \limsup_{j \to \infty} t_{n_j} \leq \limsup_{j \to \infty} (u_j + b_{r_j}) = u \]

by (i) since \( b_{r_j} \to 0 \). Thus \( \limsup_{n \to \infty} t_n = u \). The argument for \( \liminf \) is similar. 

**IV.2.8.7.** **COROLLARY.** Let \( \sum_{k=1}^{\infty} a_k \) be a conditionally convergent series.

(i) For any \( t \in \mathbb{R} \) there is a rearrangement \( \sum_{k=1}^{\infty} d_k \) of \( \sum_{k=1}^{\infty} a_k \) which converges with sum \( t \).

(ii) For any \( u \) and \( v \), \( -\infty \leq v \leq u \leq +\infty \), there is a rearrangement \( \sum_{k=1}^{\infty} d_k \) of \( \sum_{k=1}^{\infty} a_k \) whose partial sums \( t_n \) satisfy

\[ \limsup_{n \to \infty} t_n = u \quad \text{and} \quad \liminf_{n \to \infty} t_n = v . \]

**PROOF:** Part (i) is the special case of (ii) with \( u = v = t \). For (ii), if \( u \) and/or \( v \) is in \( \mathbb{R} \) take \( u_j = u \) and/or \( v_j = v \) for all \( j \); if \( u \) and/or \( v \) is \( \pm \infty \), take \( u_j \) and/or \( v_j \) to be \( \pm j \). This result was obtained by RIEMANN in ( ). The argument applies more generally to any series \( \sum_{k=1}^{\infty} a_k \) in which \( a_k \to 0 \) for which both the series of positive terms and of negative terms diverge.

**Revisions of Infinite Series**

**IV.2.8.8.** **DEFINITION.** Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series. A revision of the series is a contraction (IV.2.1.20.) of a rearrangement (IV.2.8.1.) of the series.

**IV.2.8.9.** A revision of the series is thus a new series

\[ (a_{k_1,1} + a_{k_1,2} + \cdots + a_{k_1,n_1}) + (a_{k_2,1} + a_{k_2,2} + \cdots + a_{k_2,n_2}) + \cdots \]

where each \( k \) occurs as exactly one \( k_{i,j} \), i.e. each group consists of a finite number of terms of the series, with each term occurring exactly once (in exactly one group). Then by IV.2.1.21. and IV.2.8.5. we obtain:

**IV.2.8.10.** **PROPOSITION.** Let \( \sum_{k=1}^{\infty} a_k \) be an absolutely convergent infinite series. Then any revision of the series converges, with the same limit.
IV.2.8.11. In this result, each group is allowed to include only finitely many terms of the series. In IV.2.10.16. we will show the result holds also for more drastic revisions where the groups are allowed to be infinite.

IV.2.9. Unordered Summations

IV.2.9.1. Let \( J \) be a set, possibly infinite, possibly even uncountable, and let \( \{a_j : j \in J\} \) be an indexed collection of real (or complex) numbers. We want to make sense out of the expression \( \sum_{j \in J} a_j \). If \( J \) is finite, there is an unambiguous meaning, but if \( J \) is infinite, there is no direct meaning. We have made sense of such a sum if \( J = \mathbb{N} \), but the theory of infinite series depends critically on having a fixed ordering on \( \mathbb{N} \). A general set \( J \) has no natural ordering, and even if a particular \( J \) has an ordering (e.g. if \( J = \mathbb{N} \)) we do not want our definition to depend on the ordering. It is to be expected from the case \( J = \mathbb{N} \) that we can only make sense of the sum under certain conditions, not in general.

Infinite series give a clue as to how to proceed: we will take a limit of sums over finite subsets of \( J \).

IV.2.9.2. Definition. Let \( J \) be a set, and let \( \{a_j : j \in J\} \) be an indexed collection of real numbers. We say the unordered sum \( \sum_{j \in J} a_j \) converges with sum \( s \) if for every \( \epsilon > 0 \) there is a finite subset \( F \) of \( J \) such that, for every finite subset \( E \) of \( J \) containing \( F \),

\[
| s - \sum_{j \in E} a_j | < \epsilon .
\]

When we write \( \sum_{j \in J} a_j = s \), we mean that the unordered sum converges to \( s \).

IV.2.9.3. Note that if \( J = \mathbb{N} \), this definition is not the same as the usual definition of convergence of the sum as an infinite series, since for an infinite series to converge only the partial sums, which are sums of initial segments, must converge, not, for example, finite sums made up mostly of even-numbered terms. If \( (a_k) \) is a sequence of real numbers, and the unordered sum \( \sum_{j \in J} a_j \) converges with sum \( s \), then the infinite series

\[
\sum_{k \in \mathbb{N}} a_k
\]

also converges, with sum \( s \). But we will see that the converse is false in general; in fact, it follows easily from IV.2.8.5. and IV.2.8.7. that the unordered sum converges if and only if the infinite series converges absolutely (cf. IV.2.9.9.).

IV.2.9.4. Proposition. Let \( (a_k) \) be a sequence of real numbers.

(i) If the unordered sum \( \sum_{k \in \mathbb{N}} a_k \) converges, then the infinite series \( \sum_{k=1}^{\infty} a_k \) converges, and the sums are the same.
(ii) If all the \( a_k \) are nonnegative, then the unordered sum \( \sum_{k=1}^{\infty} a_k \) converges if and only if the infinite series \( \sum_{k=1}^{\infty} a_n \) converges.

**Proof:** (i) Let \( s_n \) be the \( n \)'th partial sum of the infinite series. If the unordered sum converges to \( s \), and \( \epsilon > 0 \), choose \( F \) so that
\[
|s - \sum_{k \in E} a_k| < \epsilon
\]
for all finite \( E \) containing \( F \). If \( n \) is large enough that \( F \subseteq \{1, \ldots, n\} \), then \( |s - s_n| < \epsilon \). Thus the infinite series converges with sum \( s \).

(ii) Suppose all \( a_k \) are nonnegative, and that the infinite series converges to \( s \). If \( \epsilon > 0 \), choose \( n \) so that \( |s - s_n| < \epsilon \). If \( E \) is any finite subset of \( \mathbb{N} \) containing \( F = \{1, \ldots, n\} \), and \( m \) is the largest number in \( E \), then
\[
s_n \leq \sum_{k \in E} a_k \leq s_m \leq s
\]
and so
\[
|s - \sum_{k \in E} a_k| < \epsilon
\]
and \( \sum_{k \in \mathbb{N}} a_k \) converges with sum \( s \).

The first observation is an analog of the Divergence Criterion:

**IV.2.9.5. Proposition.** Let \( J \) be a set, and let \( \{a_j : j \in J\} \) be an indexed collection of real numbers. If \( \sum_{j \in J} a_j \) converges, then for any \( \epsilon > 0 \) there are only finitely many \( j \) for which \( |a_j| > \epsilon \).

**Proof:** Let \( s \) be the sum of the series, and \( \epsilon > 0 \). Then there is a finite subset \( F \) of \( J \) such that
\[
|s - \sum_{j \in E} a_j| < \frac{\epsilon}{2}
\]
for all finite subsets \( E \) of \( J \) containing \( F \). Let \( j_0 \in J \setminus F \), and \( E = F \cup \{j_0\} \). Then
\[
|a_{j_0}| = \left| \sum_{j \in E} a_j - \sum_{j \in F} a_j \right| \leq \left| \sum_{j \in E} a_j - s \right| + \left| s - \sum_{j \in F} a_j \right| < \epsilon.
\]
This is true for all \( j_0 \notin F \).

This has an interesting consequence. The index set \( J \) could very well be uncountable, but we can effectively reduce to the case where \( J \) is countable by the following result:
**IV.2.9.6.** Proposition. Let \( J \) be a set, and let \( \{a_j : j \in J\} \) be an indexed collection of real numbers. If \( \sum_{j \in J} a_j \) converges, then the set \( K = \{j \in J : a_j \neq 0\} \) is countable and \( \sum_{j \in K} a_j \) converges to \( \sum_{j \in J} a_j \).

Proof: By IV.2.9.5., the set \( K_n = \{j \in J : |a_j| > \frac{1}{n}\} \) is finite for each \( n \), and \( K = \cup_n K_n \). If \( F \) is any finite subset of \( J \), then \( \sum_{j \in F} a_j = \sum_{j \in F \cap K} a_j \).

As with infinite series, we can extend IV.2.9.5. to obtain a Cauchy criterion:

**IV.2.9.7.** Proposition. Let \( J \) be a set, and let \( \{a_j : j \in J\} \) be an indexed collection of real numbers. Then \( \sum_{j \in J} a_j \) converges if and only if, for any \( \epsilon > 0 \), there is a finite subset \( F \) of \( J \) such that

\[
\left| \sum_{j \in D} a_j \right| < \epsilon
\]

for every finite subset \( D \) of \( J \) disjoint from \( F \).

Proof: Suppose the unordered sum converges to \( s \), and let \( \epsilon > 0 \). Then there is a finite subset \( F \) of \( J \) such that

\[
\left| s - \sum_{j \in E} a_j \right| < \frac{\epsilon}{2}
\]

for all finite subsets \( E \) of \( J \) containing \( F \). Let \( D \) be a finite subset of \( J \) disjoint from \( F \). Then

\[
\left| \sum_{j \in D} a_j \right| = \left| \sum_{j \in D \cup F} a_j - \sum_{j \in F} a_j \right| \leq \left| \sum_{j \in D \cup F} a_j - s \right| + \left| s - \sum_{j \in F} a_j \right| < \epsilon.
\]

Conversely, suppose the other condition holds. Choose an increasing sequence \( (F_n) \) of finite subsets of \( J \) such that

\[
\left| \sum_{j \in D} a_j \right| < \frac{1}{n}
\]

for all finite \( D \) disjoint from \( F_n \). Then, for any finite \( E \) containing \( F_n \), we have

\[
\left| \sum_{j \in E} a_j - \sum_{j \in F_n} a_j \right| = \left| \sum_{j \in E \setminus F_n} a_j \right| \leq \frac{1}{n}.
\]

Set

\[
\alpha_n = \inf \left\{ \sum_{j \in E} a_j : F_n \subseteq E \text{ finite} \right\} \geq \sum_{j \in F_n} a_j - \frac{1}{n}.
\]
\[
\beta_n = \sup \left\{ \sum_{j \in E} a_j : F_n \subseteq \mathbb{R} \text{ finite} \right\} \leq \sum_{j \in F_n} a_j + \frac{1}{n}
\]
and set \( I_n = [\alpha_n, \beta_n] \). We have \( I_{n+1} \subseteq I_n \) since \( F_n \subseteq F_{n+1} \). Then \( \cap_n I_n = \{ s \} \) for some \( s \in \mathbb{R} \), and if \( E \) is any finite set containing \( F_n \) we have

\[
| s - \sum_{j \in E} a_j | \leq \frac{2}{n}
\]

(both numbers are in \( I_n \)), which can be made arbitrarily small. Thus \( \sum_{j \in J} a_j \) converges to \( s \). ☑

**IV.2.9.8.** We can extend the notion of unordered sum to arbitrary nonnegative sums. Let \( J \) be a set, and let \( \{ a_j : j \in J \} \) be an indexed collection of real numbers. If all the \( a_j \) are nonnegative, we can always make sense of the unordered sum as

\[
\sum_{j \in J} a_j = \sup \sum_{j \in F} a_j
\]

where the supremum is over all finite subsets \( F \) of \( J \). If the supremum is \( s < +\infty \), this definition agrees with the previous one: given \( \epsilon > 0 \), choose a finite \( F \) with

\[
| s - \sum_{j \in F} a_j | < \epsilon
\]

and this \( F \) works in the definition of convergence. If the supremum is \( +\infty \), we will not say the unordered sum converges, but we will write \( \sum_{j \in J} a_j = +\infty \).

We can now state the main result of this subsection:

**IV.2.9.9.** *Theorem.* Let \( J \) be a set, and let \( \{ a_j : j \in J \} \) be an indexed collection of real numbers. Then \( \sum_{j \in J} a_j \) converges if and only if \( \sum_{j \in J} |a_j| \) converges (i.e. \( \sum_{j \in J} |a_j| < +\infty \)).

**Proof:** If \( \sum_{j \in J} |a_j| \) converges, then for any \( \epsilon > 0 \) there is a finite set \( F \) such that

\[
\sum_{j \in D} |a_j| < \epsilon
\]

for every finite subset \( D \) of \( J \) disjoint from \( F \) by IV.2.9.7.. But then

\[
\left| \sum_{j \in D} a_j \right| \leq \sum_{j \in D} |a_j| < \epsilon
\]

for every finite subset \( D \) of \( J \) disjoint from \( F \), so \( \sum_{j \in J} a_j \) converges by IV.2.9.7..
Conversely, suppose $\sum_{j \in J} a_j$ converges. Let $F$ be a finite subset of $J$ such that

$$\left| \sum_{j \in D} a_j \right| < 1$$

for every finite subset $D$ of $J$ disjoint from $F$. Suppose $\sum_{j \in J} |a_j| = +\infty$; then there are finite subsets $D$ of $J$ disjoint from $F$ with $\sum_{j \in D} |a_j|$ arbitrarily large. In particular, there is such a $D$ with $\sum_{j \in D} |a_j| > 2$. For this $D$, let $D_+ = \{ j \in D : a_j > 0 \}$, $D_- = \{ j \in D : a_j < 0 \}$. Then either $\left| \sum_{j \in D_+} a_j \right| = \sum_{j \in D_+} |a_j| > 1$ or $\left| \sum_{j \in D_-} a_j \right| = \sum_{j \in D_-} |a_j| > 1$, a contradiction. Thus $\sum_{j \in J} |a_j| < +\infty$.

Thus there is no difference between convergence and “absolute convergence” for unordered sums.

**IV.2.9.10. Corollary.** Let $J$ be a countable index set, and $\{a_j : j \in J\}$ an indexed set of numbers. Then the following are equivalent:

(i) The unordered sum $\sum_{j \in J} a_j$ converges.

(ii) The infinite series $\sum_{k=1}^{\infty} a_{\phi(k)}$ converges absolutely for some bijection $\phi : \mathbb{N} \to J$.

(iii) The infinite series $\sum_{k=1}^{\infty} a_{\phi(k)}$ converges absolutely for every bijection $\phi : \mathbb{N} \to J$.

If the series converge, the sums are all equal.

Thus a convergent unordered sum can be converted to an ordinary infinite series by ordering the terms in an arbitrary way; the resulting infinite series will converge absolutely and have the same sum independently of how the terms are ordered.

**IV.2.9.11. Corollary.** Let $(a_k)$ be a sequence of real numbers. Then the unordered series $\sum_{k \in \mathbb{N}} a_k$ converges if and only if the infinite series $\sum_{k=1}^{\infty} a_k$ converges absolutely. If the series converge, the sums are equal.
IV.2.10. Double Summations

Suppose we have a double array

\[
\begin{array}{ccc}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33} \\
  \vdots & \vdots & \vdots \\
\end{array}
\]

of numbers, and we want to “add them all up.” There are many potentially reasonable ways to proceed, and we might hope (or even expect from commutativity of addition) that all these ways give the same answer. Experience with rearranging conditionally convergent series might make us suspicious that this works, and these suspicions are well founded: the various methods do not always give the same answer. But if the sums are all “absolutely convergent” in an appropriate sense, the methods all give the same sum.

IV.2.10.1. A first, somewhat naive, approach is to simply write all the elements down in some order to make an ordinary infinite series, and take the sum of this series (if it exists). More precisely, take the infinite series

\[
\sum_{k=1}^{\infty} a_{\phi(k)}
\]

where \( \phi \) is a bijection from \( \mathbb{N} \) to \( \mathbb{N} \times \mathbb{N} \). Of course, there are many ways to do this, and the series obtained are rearrangements of one another (and any rearrangement of one of the series is another such series). If one such series converges absolutely, so does any other and all have the same sum (IV.2.8.5.), so the procedure seems to be satisfactory; but if the series do not converge absolutely, the answers will differ drastically (IV.2.8.7.), and not all will even converge.

IV.2.10.2. As a variation, we could consider the unordered sum (IV.2.9.2.)

\[
\sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk} .
\]

This unordered sum will converge if and only if the sums in IV.2.10.1. converge absolutely (IV.2.9.11.), with the same sum. This unordered sum can be regarded as the double sum of the array.

IV.2.10.3. A somewhat different approach is to consider sums of the form

\[
\sum_{j,k=1}^{n} a_{jk} = \sum_{j=1}^{n} \left[ \sum_{k=1}^{n} a_{jk} \right]
\]

and let \( n \to \infty \). These are finite sums, so each sum can be computed in any order; in particular, we have

\[
\sum_{k=1}^{n} \sum_{j=1}^{n} a_{jk} = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} := \sum_{j=1}^{n} \left[ \sum_{k=1}^{n} a_{jk} \right]
\]

(such double sums are usually written without the brackets). These sums are often called square sums since the terms in a square in the upper left-hand corner of the array are added. As a variation, we could take
rectangular sums

\[
\sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}
\]

and take a limit as \( m, n \to \infty \).

The rectangular sum procedure is less likely to work (i.e. to have a limit), but if it does the square sum procedure will work and give the same sum. If the unordered sum procedure converges (i.e. the procedure of IV.2.10.1. gives absolutely convergent series), these rectangular and square sum procedures will also work and give the same sum, since they are limits of partial sums of contractions of certain of the series from IV.2.10.1..

**IV.2.10.4.** Another variation is to consider

\[
\sum_{n=2}^{\infty} \sum_{j+k=n} a_{jk}
\]

This is called a triangular sum, since the terms on each northeast-southwest diagonal are first added (a finite sum) and then the diagonal sums are added. This procedure is particularly natural when products of power series are considered (IV.2.11.4.). If the unordered sum converges, this procedure will also converge and give the same sum, for the same reason as for square and rectangular sums.

**IV.2.10.5.** As a quite different procedure, we can sum along each row, and then add the row sums, or sum along each column and then add the column sums. Specifically, for each fixed \( j \) we can set

\[
r_j = \sum_{k=1}^{\infty} a_{jk}
\]

if the series converges; \( r_j \) is the sum of the terms in the \( j \)'th row. We then take

\[
\sum_{j=1}^{\infty} r_j
\]

which is usually written

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}
\]

(when an expression like this is written, it is implicitly assumed that the inside sum converges for each \( j \)).

We can similarly first add each column: for each fixed \( k \) set

\[
c_k = \sum_{j=1}^{\infty} a_{jk}
\]

if the series converges; \( c_k \) is the sum of the terms in the \( k \)'th column. We then take

\[
\sum_{k=1}^{\infty} c_k = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk}
\]
It is not immediately clear whether these two iterated summation procedures are well defined (i.e. all the
infinite series will converge), or give the same answer if they are; and the relation between the convergence
of these procedures and the previous ones is also not obvious. The main results of this section will establish
the relationship in good cases.

Note that it is not quite correct to call such iterated summations infinite series; they are a different (but
closely related) type of infinite sum which involves many infinite series.

IV.2.10.6. Example. Here is the standard example illustrating some of what can go wrong. Set

\[
 a_{jk} = \begin{cases} 
 1 & \text{if } j = k \\
 -1 & \text{if } j = k + 1 \\
 0 & \text{otherwise}
\end{cases}
\]

i.e. the terms on the main diagonal are all 1, the terms on the first subdiagonal are −1, and all other terms
are 0. The row and column sums each have only finitely many nonzero terms (no more than 2), hence
converge absolutely. But we have \( r_1 = 1, r_j = 0 \) for \( j > 1 \), and \( c_k = 0 \) for all \( k \), so

\[
 1 = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \neq \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk} = 0.
\]

The square sum procedure gives a sum of 1; but the rectangular and triangular procedures do not converge
since the rectangular or triangular sums can be either 0 or 1. The unordered sum diverges in this case; the
procedure of IV.2.10.1. always gives a divergent series since the terms do not approach 0.

To examine the convergence of such iterated summations, we first consider the case where all the terms
are nonnegative. In this case we can make sense of any iterated sum by taking the convention that an infinite
series with nonnegative terms and at least one term +1 has sum +∞. The next result is a special case of
Tonelli’s Theorem for integrals on product spaces (\()\).

IV.2.10.7. Theorem. [Tonelli’s Theorem for Double Summations] Let \( \{a_{jk} : j, k \in \mathbb{N}\} \) be an
array with all \( a_{jk} \geq 0 \). If one of the three sums

\[
 \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}, \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk}, \sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk}
\]

converges (i.e. is finite), then all three converge and have the same sum.

Proof: First note that for any \( m, n \in \mathbb{N} \),

\[
 \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \leq \sup_{(j,k) \in \mathbb{N} \times \mathbb{N}} \sum_{F \subseteq \mathbb{N} \times \mathbb{N}} a_{jk}
\]

since \( \{1, \ldots, m\} \times \{1, \ldots, n\} \) is a finite subset of \( \mathbb{N} \times \mathbb{N} \). Conversely, for any finite \( F \subseteq \mathbb{N} \times \mathbb{N} \), we have

\[
 \sum_{(j,k) \in F} a_{jk} \leq \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}
\]
for sufficiently large \( m, n \) (large enough that \( F \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\} \)); thus

\[
\sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk} = \sup_{m,n \in \mathbb{N}} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}.
\]

Also, for any \( m, n \),

\[
\begin{align*}
\sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} &\leq \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} \leq \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk},
\end{align*}
\]

so we get

\[
\sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk} = \sup_{m,n \in \mathbb{N}} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \leq \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}.
\]

For the reverse inequality, suppose

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} = \sup_{m,n \in \mathbb{N}} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} = s \leq +\infty
\]

and fix \( r < s \). Then there is an \( m \) such that

\[
r < \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk}.
\]

But by repeated applications of IV.2.1.16., we have

\[
\sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} = \sum_{k=1}^{\infty} \sum_{j=1}^{m} a_{jk}
\]

so

\[
\sum_{k=1}^{\infty} \sum_{j=1}^{m} a_{jk} = \sup_{n \in \mathbb{N}} \sum_{k=1}^{n} \sum_{j=1}^{m} a_{jk} > r
\]

and thus there is an \( n \) such that

\[
\sum_{k=1}^{n} \sum_{j=1}^{m} a_{jk} > r
\]

and we conclude that

\[
\sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk} = \sup_{m,n \in \mathbb{N}} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \geq s.
\]

Thus

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} = \sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk}.
\]

A nearly identical argument shows that

\[
\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk} = \sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk}.
\]
We now turn to the general absolutely convergent case. This result is a special case of Fubini’s Theorem for integrals on product spaces (\(\cdot\)).

**IV.2.10.8. Theorem. [Fubini’s Theorem for Double Summations]** Let \(\{a_{jk} : j, k \in \mathbb{N}\}\) be an array of numbers. If one (hence each by IV.2.10.7.) of the three sums

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |a_{jk}|, \quad \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} |a_{jk}|, \quad \sum_{(j,k)\in \mathbb{N}\times \mathbb{N}} |a_{jk}|
\]

converges (i.e. is finite), then each of the three summations

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}, \quad \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk}, \quad \sum_{(j,k)\in \mathbb{N}\times \mathbb{N}} a_{jk}
\]

converges absolutely, and the sums are equal. In particular, the iterated summations are equal.

**Proof:** The argument is similar to the proof of IV.2.8.5. First note that the three sums in the conclusion (and all the infinite series in the iterated summations) converge absolutely, hence converge, so we need only show the three sums are equal. We will show

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} = s := \sum_{(j,k)\in \mathbb{N}\times \mathbb{N}} a_{jk}
\]

and the proof that the other iterated sum equals \(s\) is almost identical.

Let \(\epsilon > 0\). Then, since \(\sum_{(j,k)\in \mathbb{N}\times \mathbb{N}} |a_{jk}| < \infty\), there is a finite set \(F \subseteq \mathbb{N} \times \mathbb{N}\) such that

\[
\sum_{(j,k)\in \mathbb{N}\times \mathbb{N}\setminus F} |a_{jk}| < \frac{\epsilon}{3}
\]

and \(M, N \in \mathbb{N}\) such that \(F \subseteq \{1, \ldots, M\} \times \{1, \ldots, N\}\). Thus if \(m \geq M, n \geq N\),

\[
\left| s - \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \right| < \frac{\epsilon}{3}.
\]

We have

\[
\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} = \lim_{m \to \infty} \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk}.
\]

Fix \(m \geq M\) such that

\[
\left| \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} - \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \right| < \frac{\epsilon}{3}.
\]

487
Since $m$ is fixed, we have

$$\sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} = \lim_{n \to \infty} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}$$

by (), so there is an $n \geq N$ such that

$$\left| \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} - \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} \right| < \frac{\epsilon}{3}.$$  

We then have

$$\left| s - \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \right| \leq \left| s - \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \right| + \left| \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} - \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} \right| + \left| \sum_{j=1}^{m} \sum_{k=1}^{\infty} a_{jk} - \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \right| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

and since $\epsilon > 0$ is arbitrary, the iterated sum equals $s$.

**IV.2.10.9.** We have retained the commonly-used names “Tonelli’s Theorem” and “Fubini’s Theorem” for IV.2.10.7. and IV.2.10.8. since they are special cases of the general results commonly called “Tonelli’s Theorem” () and “Fubini’s Theorem” (), even though it is likely that the double sum results predated FUBINI and TONELLI (at least IV.2.10.7. was undoubtedly “known” to EULER, although it was probably only really proved later). I have been unable to discover who first explicitly stated and proved the double sum versions (very likely sometime in the nineteenth century). See XIV.3.4.34. for additional historical comments.

**IV.2.10.10.** Actually, we must be careful in saying what it means for an iterated summation

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}$$

to converge absolutely. The strong (and proper) interpretation is that the iterated summation

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |a_{jk}|$$

should converge. Fubini’s Theorem IV.2.10.8. says that if an iterated sum

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}$$

converges absolutely in this sense, then the iterated sum

$$\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk}$$

488
in the opposite order (which also converges absolutely by IV.2.10.7.) has the same sum, i.e. the order of summation can be reversed.

We could instead consider a weaker sense in which an iterated summation
\[ \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} \]
converges absolutely: we could just require

(i) For each \( j \), the infinite series \( \sum_{k=1}^{\infty} a_{jk} \) converges absolutely with sum \( b_j \).

(ii) The infinite series \( \sum_{j=1}^{\infty} b_j \) converges absolutely, i.e.
\[ \sum_{j=1}^{\infty} \left| \sum_{k=1}^{\infty} a_{jk} \right| \]
converges.

But example IV.2.10.6. shows that even if both iterated summations
\[ \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk}, \quad \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk} \]
converge absolutely in this weak sense, the sums may not be equal. In fact, if the iterated summation in one order converges absolutely in the weak sense, the iterated sum in the other order need not, and need not even make sense: consider, for example, the case where for all \( j \) \( a_{j1} = 1 \), \( a_{j2} = -1 \), and \( a_{jk} = 0 \) for all \( k > 2 \).

Combining IV.2.10.8. with the previous observations, we obtain:

**IV.2.10.11. Corollary.** Let \( \{a_{jk} : j, k \in \mathbb{N}\} \) be an array of numbers. If one (hence each) of the three sums
\[ \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |a_{jk}|, \quad \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} |a_{jk}|, \quad \sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} |a_{jk}| \]
converges (i.e. is finite), then each of the following summations converges absolutely, and the sums are equal:

(i) \[ \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{jk} . \]

(ii) \[ \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{jk} . \]

(iii) \[ \sum_{(j,k) \in \mathbb{N} \times \mathbb{N}} a_{jk} . \]
(iv) The rectangular sum \( \lim_{m,n \to \infty} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk} \).

(v) The square sum \( \lim_{n \to \infty} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} \).

(iv) The triangular sum \( \sum_{n=2}^{\infty} \sum_{j+k=n} a_{jk} \).

(vii) Any sum of the form \( \sum_{k=1}^{\infty} a_{\phi(k)} \) where \( \phi \) is a bijection from \( \mathbb{N} \) to \( \mathbb{N} \times \mathbb{N} \).

IV.2.10.12. These results are stated for double summations, but by repeated applications of the double summation result the same conclusions hold for triple or \( n \)-fold summations for any \( n \).

**Drastic Revisions of an Infinite Series**

**IV.2.10.13.** A drastic revision of an infinite series

\[ \sum_{k=1}^{\infty} a_k \]

is a sum of the form

\[ (a_{k_1,1} + a_{k_1,2} + \cdots) + (a_{k_2,1} + a_{k_2,2} + \cdots) + \cdots \]

which resembles a revision (IV.2.8.8.) of the series except that the groups are allowed to be infinite. More precisely:

**IV.2.10.14.** **Definition.** A drastic revision of the infinite series \( \sum_{k=1}^{\infty} a_k \) is an iterated summation of the form

\[ \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{\phi_j(k)} \]

where \( \phi_1, \phi_2, \ldots \) is a sequence of one-to-one functions from \( \mathbb{N} \) to \( \mathbb{N} \times \mathbb{N} \) with disjoint ranges whose union is all of \( \mathbb{N} \).

**IV.2.10.15.** We can also allow some of the groups to be finite; such a sum is technically a drastic revision of a series obtained by inserting an infinite number of 0’s into the original series, which makes no essential change to the series. Thus an ordinary revision is a drastic revision.

The main result about drastic revisions is:
IV.2.10.16. Theorem. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any drastic revision of the series converges absolutely, and has the same sum as the original series.

Proof: This is an almost immediate corollary of Fubini’s Theorem. Set $a_{jk} = a_{\phi_j(k)}$. Since the original series converges absolutely, we have that the unordered sum

$$\sum_{(j,k)\in\mathbb{N} \times \mathbb{N}} a_{jk} = \sum_{k\in\mathbb{N}} a_k$$

converges absolutely to the sum of the original series by IV.2.9.11. and IV.2.9.4.(i), and hence the iterated sum does too by IV.2.10.8.

IV.2.10.17. Fubini’s Theorem is usually thought of as a result justifying interchanging the order of summation or integration in an iterated sum or integral. But IV.2.10.16. is an interesting and useful application not involving interchange of order; the relevant part of Fubini’s Theorem is that the iterated sum is equal to the unordered sum (double sum).

IV.2.10.18. We could more generally consider drastic revisions of drastic revisions of a series, etc. If the series converges absolutely, all these will converge absolutely to the same sum. The process can be taken to any finite level, or to an infinite or even transfinite level. Details are left to the interested reader.

Drastic Revisions of Unordered Sums

IV.2.10.19. Suppose $\sum_{j\in J} a_j$ is an unordered summation, and we have a decomposition $J = \cup_{k\in K} J_k$ of $J$ as a disjoint union of subsets. For each $k$ we may form the unordered sum

$$\sum_{j\in J_k} a_j$$

and, if these sums all converge, we can form the unordered sum

$$\sum_{k\in K} \left( \sum_{j\in J_k} a_j \right).$$

Such a sum is called a drastic revision of the unordered sum $\sum_{j\in J} a_j$.

IV.2.10.20. Theorem. Let $\sum_{j\in J} a_j$ be a convergent unordered sum, and $J = \cup_{k\in K} J_k$ a decomposition of $J$. Then

(i) The unordered sum $\sum_{j\in J_k} a_j$ converges for each $k$.  

491
(ii) The unordered sum (drastic revision)
\[
\sum_{k \in K} \left( \sum_{j \in J_k} a_j \right)
\]
converges, and
\[
\sum_{k \in K} \left( \sum_{j \in J_k} a_j \right) = \sum_{j \in J} a_j .
\]

**Proof:** We have \(a_j \neq 0\) for only countably many \(J\) (IV.2.9.6.), and there are only countably many \(k\) for which \(J_k\) contains a nonzero term. Thus there is no loss of generality in assuming \(J\) and \(K\) are countable.

Fix bijections from \(\mathbb{N}\) to \(K\) and to \(J_k\) for each \(k\) (the **Countable AC** is needed to do this; zero terms may have to be added to the unordered sum if some \(J_k\) are finite). The unordered sums thus become ordinary infinite series using these orderings, and the infinite series corresponding to IV.1 is a drastic revision of the original sum as an infinite series. The result then follows from IV.2.10.16. and IV.2.9.11.

**IV.2.10.21.** The converse of this theorem is true if all the terms of the unordered sum are nonnegative. But the converse is not true in general: if a drastic revision of an unordered sum converges, the original unordered sum need not converge, and if two drastic revisions of an unordered sum both converge, the sums may not be equal. Example IV.2.10.6. gives a counterexample if the series are regarded as unordered sums. The situation is essentially the same as the problem described in IV.2.10.10.

**IV.2.10.22.** Just as with ordinary series, we can take such drastic revisions to multiple levels. If the original unordered sum converges, so will all multiple-level drastic revisions, with the same sum.

**IV.2.11. Products of Infinite Series**

**IV.2.11.1.** Suppose we have two convergent infinite series
\[
\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k
\]
with sums \(s\) and \(t\) respectively (it is notationally convenient to start the sums at 0 instead of 1). What should we mean by the product of the two series
\[
\left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} b_k \right)
\]
and what should its sum be? However we define the product its “sum” should be \(st\).

**IV.2.11.2.** A natural approach is to consider the double array \((c_{jk})\), where \(c_{jk} = a_j b_k\); this array would be obtained by formally multiplying out the series
\[
(a_0 + a_1 + a_2 + \cdots)(b_0 + b_1 + b_2 + \cdots)
\]
as though they were finite sums. Thus any of the sums considered in the last section would be a potential interpretation of the product of the series. Several of these automatically converge:
**IV.2.11.3. Proposition.** Using the above notation, the following sums converge with sum $st$:

(i) The iterated sum $\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} a_j b_k$.

(ii) The iterated sum $\sum_{k=0}^{\infty} \sum_{j=0}^{\infty} a_j b_k$.

(iii) The rectangular sum $\lim_{m,n \to \infty} \sum_{j=0}^{m} \sum_{k=0}^{n} a_j b_k$.

(iv) The square sum $\lim_{n \to \infty} \sum_{j=0}^{n} \sum_{k=0}^{n} a_j b_k$.

**Proof:** (i) and (ii) each follow directly from two applications of **IV.2.11.3.**. For (iii) and (iv), set

$$s_m = \sum_{j=0}^{m} a_j \quad \text{and} \quad t_n = \sum_{k=0}^{n} b_k$$

and note that

$$\sum_{j=0}^{m} \sum_{k=0}^{n} a_j b_k = s_m t_n$$

and that

$$\lim_{m,n \to \infty} s_m t_n = \lim_{n \to \infty} s_n t_n = st.$$ 

**IV.2.11.4.** However, the triangular sum is more problematic. And it is the triangular sum which is the most natural definition of the product of the two series. For one thing, it is the only systematically obtained sum of the types previously considered which is actually a single infinite series. More importantly, it is especially natural when power series are considered (cf. **III.5.3.2., V.16.3.11.**): if

$$\sum_{k=0}^{\infty} a_k x^k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k x^k$$

are power series, if we collect together like powers of $x$ in the array $c_{jk} = a_j b_k x^{j+k}$ we get the triangular sum

$$\sum_{n=0}^{\infty} \sum_{j+k=n} a_j b_k x^n = \sum_{n=0}^{\infty} \left[ \left( \sum_{k=0}^{n} a_k b_{n-k} \right) x^n \right].$$

We thus make the definition:
IV.2.11.5. Definition. Let
\[ \sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k \]
be infinite series. Then the product, or Cauchy product, of the series is the triangular sum
\[ \left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} b_k \right) = \sum_{n=0}^{\infty} \sum_{j+k=n} a_j b_k = \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k} \].

IV.2.11.6. Example. The product of two convergent infinite series can diverge. Set
\[ a_k = b_k = \frac{(-1)^k}{\sqrt{k}} \]
for all \( k \geq 1 \). Then
\[ \sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} b_k \]
converge by the Alternating Series Test. But the product series
\[ \sum_{n=2}^{\infty} (-1)^n \left( \sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}} \right) \]
diverges since the terms do not approach 0:
\[ \frac{1}{\sqrt{k(n-k)}} \geq \frac{2}{n} \quad \text{for} \quad 1 \leq k \leq n - 1. \]

IV.2.11.7. If the series
\[ \sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k \]
both converge absolutely to \( s \) and \( t \) respectively, then the iterated sum
\[ \sum_{j=0}^{\infty} \sum_{k=0}^{j} |a_j b_k| \]
converges, and thus by IV.2.10.11. all the associated sums, and in particular the triangular sum, converge absolutely to the same sum, which must be \( st \). Thus we obtain:

IV.2.11.8. Theorem. Let
\[ \sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k \]
be absolutely convergent infinite series with sums \( s \) and \( t \) respectively. Then the product series
\[ \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k} \]
converges absolutely, with sum $st$.

While the absolutely convergent case is usually the most important one, the convergence result can be extended to the case where only one of the series converges absolutely. Note, however, that absolute convergence of the product series is not guaranteed in this case. This theorem is due to MERTENS.

**IV.2.11.9.** Theorem. Let
\[ \sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k \]
be convergent infinite series with sums $s$ and $t$ respectively. Suppose at least one of the series converges absolutely. Then the product series
\[ \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k} \]
converges, with sum $st$.

Actually, convergence of the product series is the only issue: if the product series converges, its sum is automatically the right thing. This result is due to Abel.

**IV.2.11.10.** Theorem. Let
\[ \sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k \]
be convergent infinite series with sums $s$ and $t$ respectively. If the product series
\[ \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k} \]
converges, its sum is $st$.

**IV.2.11.11.** We have considered products of two series. But we may iterate to obtain products of any finite number of series, and the extension of IV.2.11.8. applies: if each of the series converges absolutely, so does the product series. We should note the following technicality, which is proved by a straightforward calculation left to the reader:

**IV.2.11.12.** Proposition. The product of infinite series is associative and commutative: if
\[ \sum_{k=0}^{\infty} a_k, \quad \sum_{k=0}^{\infty} b_k, \quad \sum_{k=0}^{\infty} c_k \]
are infinite series, then
\[
\left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} b_k \right) \left( \sum_{k=0}^{\infty} c_k \right) = \left( \sum_{k=0}^{\infty} a_k \right) \left[ \left( \sum_{k=0}^{\infty} b_k \right) \left( \sum_{k=0}^{\infty} c_k \right) \right]
\]
\[
\left( \sum_{k=0}^{\infty} a_k \right) \left( \sum_{k=0}^{\infty} b_k \right) = \left( \sum_{k=0}^{\infty} b_k \right) \left( \sum_{k=0}^{\infty} a_k \right).
\]

IV.2.12. Infinite Products

IV.2.12.1. An infinite product is an expression of the form
\[
\prod_{k=1}^{\infty} u_k
\]
where the \( u_k \) are numbers (or later functions). The \( u_k \) are called the factors of the product.

Infinite products have a long history, and there are some famous examples such as the following expressions involving \( \pi \) due respectively to Viète (1646) and Wallis (1653):
\[
\frac{2}{\pi} = \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}}} \cdots
\]
\[
\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdots \cdot \frac{2m}{2m+1} \cdot \frac{2m+2}{2m+1} \cdots
\]

In this section, we will make sense out of such expressions and discuss convergence.

IV.2.12.2. In analogy with the definition for infinite series, we could define partial products
\[
p_n = \prod_{k=1}^{n} u_k
\]
and say that the infinite product converges, with product \( p \), if the sequence \((p_n)\) converges to \( p \). But there are two serious difficulties with this approach in general. First, if one of the \( u_k \) is zero, then all the partial products from that point on will be zero no matter what the rest of the factors are. And secondly, even if all the \( u_k \) are nonzero, so all the \( p_k \) are nonzero, we can easily have \( p = 0 \). In either of these cases it is not reasonable to say that the infinite product converges. Thus we make the following definition:

IV.2.12.3. Definition. Let \( \prod_{k=1}^{\infty} u_k \) be an infinite product. Then the product converges, with product \( p \), if \( p \neq 0 \) and the sequence \((p_n)\) of partial products converges to \( p \). If there is no such \( p \), the infinite product diverges.
The factors of a convergent infinite product are necessarily nonzero. We can extend the definition by allowing a finite number of zero factors and determining convergence by the tail products after the last zero factor (this is useful when the factors are functions, to allow the product to be defined when one factor is zero).

The next fundamental observation is analogous to the Divergence Criterion for series:

**IV.2.12.4.** Prop. Let \( \prod_{k=1}^{\infty} u_k \) be a convergent infinite product. Then \( \lim_{k \to \infty} u_k = 1 \).

**Proof:** We have \( \lim_{k \to \infty} p_k = p \neq 0 \) and \( \lim_{k \to \infty} p_{k+1} = p \), and hence by ()

\[
\lim_{k \to \infty} u_k = \lim_{k \to \infty} \frac{p_{k+1}}{p_k} = \frac{\lim_{k \to \infty} p_{k+1}}{\lim_{k \to \infty} p_k} = \frac{p}{p} = 1.
\]

Note that \( \lim_{k \to \infty} u_k = 1 \) is a necessary condition for convergence of the infinite product, but not a sufficient one, just as with infinite series.

**IV.2.12.5.** Two important consequences of **IV.2.12.4.** are:

(i) Since the factors of a convergent infinite product are eventually close to 1, they are eventually positive, i.e. a convergent infinite product can have only finitely many negative factors. For convenience, we will usually only consider infinite products with positive factors; ones with finitely many negative (or zero) factors can be reduced to this case by considering a tail product (in obvious analogy with **IV.2.1.6.**, an infinite product converges if and only if some (every) tail product converges).

(ii) In light of **IV.2.12.4.**, it is customary to write the factor \( u_k \) as \((1 + a_k)\) and to call the \( a_k \) the terms of the infinite product. **IV.2.12.4.** then says that the terms of a convergent infinite product tend to zero.

We will eventually see that the infinite product \( \prod_{k=1}^{\infty} (1 + a_k) \) converges roughly (but not exactly) if and only if \( \sum_{k=1}^{\infty} a_k \) converges.

As a strengthening of (i), for convenience we often only consider infinite products with \(|a_k| < 1\) for all \( k \); this will be true in some tail of any convergent infinite product.

The next result in a sense completely reduces the convergence question for infinite products to a question about infinite series:
IV.2.12.6. Proposition. Let \( \prod_{k=1}^{\infty} u_k \) be an infinite product with \( u_k > 0 \) for all \( k \). Then the infinite product converges if and only if the infinite series \( \sum_{k=1}^{\infty} \log u_k \) converges.

Proof: If we have \( p_n \to p > 0 \), we have

\[
\log p_n = \sum_{k=1}^{n} \log u_k \to \log p
\]

by continuity of the logarithm function, so \( \sum_{k=1}^{\infty} \log u_k \) converges. Conversely, if \( s_n = \sum_{k=1}^{n} \log u_k \) converges to \( s \), we have \( p_n = e^{s_n} \to e^s \) by continuity of the exponential function, so the infinite product converges.

IV.2.12.7. Thus, if we want to analyze the convergence of the infinite product \( \prod_{k=1}^{\infty} (1 + a_k) \) with all \( |a_k| < 1 \), we can just consider convergence of the infinite series \( \sum_{k=1}^{\infty} \log(1 + a_k) \). But it would be preferable to have a criterion in terms of the closely related but simpler series \( \sum_{k=1}^{\infty} a_k \). We can do this using the following lemma:

IV.2.12.8. Lemma. There is a \( \delta > 0 \) such that

\[
x - x^2 < \log(1 + x) < x - \frac{1}{3} x^2
\]

whenever \( |x| < \delta \).

Proof: It suffices to show that

\[
-1 < \lim_{x \to 0} \frac{\log(1 + x) - x}{x^2} < -\frac{1}{3}.
\]

But this limit is \(-\frac{1}{2}\) by l’Hôpital’s Rule.

IV.2.12.9. Proposition. Let \((a_n)\) be a sequence of numbers with \( |a_n| < 1 \) for all \( n \). If \( \sum_{k=1}^{\infty} a_k^2 \) converges absolutely, then \( \sum_{k=1}^{\infty} a_k \) converges if and only if \( \sum_{k=1}^{\infty} \log(1 + a_k) \) converges.
Proof: We have \( a_k \to 0 \). For each \( k \), write \( \log(1 + a_k) = a_k + \theta_k a_k^2 \). Then \( (\theta_k) \) is a bounded sequence since \(-1 < \theta_k < -\frac{1}{4}\) for sufficiently large \( k \) by IV.2.12.8. (actually \( \theta_k \to -\frac{1}{2} \)). Since \( \sum_{k=1}^{\infty} a_k^2 \) converges absolutely, we have that \( \sum_{k=1}^{\infty} \theta_k a_k^2 \) converges (absolutely) by IV.2.7.2. Hence \( \sum_{k=1}^{\infty} \log(1 + a_k) \) and \( \sum_{k=1}^{\infty} a_k \) either both converge or both diverge by IV.2.16.

**IV.2.12.10.** Corollary. Let \( \prod_{k=1}^{\infty} (1 + a_k) \) be an infinite product. If \( \sum_{k=1}^{\infty} a_k^2 \) converges absolutely, then the infinite product converges if and only if \( \sum_{k=1}^{\infty} a_k \) converges.

**IV.2.12.11.** If the \( a_k \) are real numbers, then \( \sum_{k=1}^{\infty} a_k^2 \) is a nonnegative series, so convergence and absolute convergence are the same thing. But these results also hold for complex series and products, where absolute convergence of \( \sum_{k=1}^{\infty} a_k^2 \) does not necessarily follow from convergence.

**IV.2.12.12.** The convergence of \( \sum_{k=1}^{\infty} a_k^2 \) is not necessary for the convergence of the infinite product, however; see the example in Exercise IV.2.16.11. See also Exercise IV.2.16.12. for an example of a divergent infinite product where \( \sum_{k=1}^{\infty} a_k \) converges.

However, if all terms have the same sign, convergence of \( \sum_{k=1}^{\infty} a_k \) is both necessary and sufficient for convergence of the infinite product:

**IV.2.12.13.** Theorem. Let \( \prod_{k=1}^{\infty} (1 + a_k) \) be an infinite product. If all \( a_k \) have the same sign, then the infinite product converges if and only if \( \sum_{k=1}^{\infty} a_k \) converges.

Proof: One direction follows from IV.2.12.10: if \( \sum_{k=1}^{\infty} a_k \) converges, it converges absolutely, and hence \( \sum_{k=1}^{\infty} a_k^2 \) converges (absolutely) by the Comparison Test. The converse does not follow from IV.2.12.10, however.
To prove the converse, suppose first that all $a_k > 0$. If $s_n = \sum_{k=1}^{n} a_k$, then we have

$$p_n = (1 + a_1)(1 + a_2) \cdots (1 + a_n) = 1 + a_1 + a_2 + \cdots + a_n + \text{(higher-order terms)} \geq s_n$$

since the higher-order terms are all nonnegative. If the infinite product converges, the $p_n$ are bounded, so the $s_n$ are bounded and $\sum_{k=1}^{\infty} a_k$ converges.

Now assume the $a_k$ are all negative. By passing to a tail, we may assume $|a_k| < 1$ for all $k$. Assume the infinite product converges with product $p$; then $p_n \geq p$ for all $n$. We have

$$1 - x \leq \frac{1}{1 + x}$$

for any $x > -1$, since $(1 + x)(1 - x) = 1 - x^2$. Thus we have

$$1 - a_k \leq \frac{1}{1 + a_k}$$

for all $k$, and thus

$$\prod_{k=1}^{n} (1 - a_k) \leq \frac{1}{p_n} \leq \frac{1}{p}.$$ 

Thus the infinite product $\prod_{k=1}^{\infty} (1 - a_k)$ converges, and hence $\sum_{k=1}^{\infty} (-a_k)$ converges by the previous argument.

Absolutely Convergent Infinite Products

**IV.2.12.14.** Definition. An infinite product $\prod_{k=1}^{\infty} (1 + a_k)$ converges absolutely if $\prod_{k=1}^{\infty} (1 + |a_k|)$ converges.

Just as in the sum case (IV.2.5.2.), we have:

**IV.2.12.15.** Proposition. An absolutely convergent infinite product converges.

The proof is left to the reader (Exercise IV.2.16.13.).

The next result is an immediate corollary of IV.2.12.14., and along with the previous examples shows that there are many convergent infinite products which are not absolutely convergent.

**IV.2.12.16.** Corollary. An infinite product $\prod_{k=1}^{\infty} (1 + a_k)$ converges absolutely if and only if $\sum_{k=1}^{\infty} a_k$ converges absolutely.
Rearrangement

Just as with infinite series, we may rearrange the order of factors in an absolutely convergent infinite product without affecting convergence or changing the product; however, rearrangement of order in a non-absolutely convergent infinite product can affect convergence and/or change the product.

IV.2.13. Infinite Series of Complex Numbers

Infinite series of complex numbers behave almost identically to series of real numbers. Almost all results about real infinite series carry over to the complex case, with identical proofs. We state the essentials here.

IV.2.13.1. Definition. The infinite series \( \sum_{k=1}^{\infty} a_k \) of complex numbers converges, with sum \( s \in \mathbb{R} \), if the sequence \((s_n)\) of partial sums converges to \( s \) in \( \mathbb{C} \).

The infinite series \( \sum_{k=1}^{\infty} a_k \) of complex numbers converges absolutely if \( \sum_{k=1}^{\infty} |a_k| \) converges. The series converges conditionally if it converges but does not converge absolutely.

The next observation allows many aspects of the theory of complex infinite series to be reduced to the real case.

IV.2.13.2. Proposition. Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series of complex numbers. For each \( k \), write \( a_k = b_k + ic_k \) with \( b_k, c_k \in \mathbb{R} \). Then \( \sum_{k=1}^{\infty} a_k \) converges if and only if both \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) converge, and

\[
\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} b_k + i \sum_{k=1}^{\infty} c_k .
\]

The series \( \sum_{k=1}^{\infty} a_k \) converges absolutely if and only if both \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) converge absolutely.

Proof: Let \( r_n = \sum_{k=1}^{n} a_k, s_n = \sum_{k=1}^{n} b_k, t_n = \sum_{k=1}^{n} c_k \) be the partial sums. Then \( r_n = s_n + it_n \) for each \( n \), so the first result follows from (\). The second result follows immediately from the Comparison Test and IV.2.1.16. since, for every \( k \),

\[
|b_k|, |c_k| \leq |a_k| \leq |b_k| + |c_k| .
\]

IV.2.13.3. Proposition. Every absolutely convergent series of complex numbers converges.

The proof is identical to the proof of IV.2.5.2. Alternately, the result follows immediately from IV.2.13.2.
Since $\sum_{k=1}^{\infty} |a_k|$ is an infinite series with nonnegative real terms, all the tests for convergence of nonnegative real series are tests for absolute convergence of complex series as well.

The theorems about termwise addition and scalar multiplication (IV.2.1.16., IV.2.1.18.), including multiplication by complex scalars, hold also for complex series. The theorems about rearrangements and subarrangements (IV.2.8.3., IV.2.8.5.), on unordered and iterated summations, and the theorem on products of series (IV.2.11.8.), hold for absolutely convergent complex series. All the proofs are the same as in the real case, and/or the results follow immediately from the real case using IV.2.13.2. (Riemann’s Theorem on rearranging conditionally convergent series (IV.2.8.6.) is more subtle, however: see Exercise ().)
(a) Show that \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} = 2 \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} = 1 \).

(b) Show that \( \sum a_k \) converges.

(c) Construct a similar convergent positive series \( \sum b_k \) with \( \limsup_{k \to \infty} \frac{b_{k+1}}{b_k} = +\infty \) and \( \liminf_{k \to \infty} \frac{b_{k+1}}{b_k} = 1 \).

**IV.2.16.3.** Choose an increasing sequence \( (k_n) \) such that \( \frac{1}{k_n} > 4^n \), and for \( k_n < k \leq k_{n+1} \) let \( a_k = \frac{2 - n}{k} \).

(a) Show that \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} = 1 \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} = \frac{1}{2} \).

(b) Show that \( \sum a_k \) diverges.

(c) Construct a similar divergent positive series \( \sum b_k \) with \( \limsup_{k \to \infty} \frac{b_{k+1}}{b_k} = 1 \) and \( \liminf_{k \to \infty} \frac{b_{k+1}}{b_k} = 0 \).

**IV.2.16.4.** Let \( p > 0 \) be fixed, and set

\[
 a_k = \begin{cases} 
  \frac{1}{k^p} & \text{if } \ k \ \text{odd} \\
  \frac{2}{k^p} & \text{if } \ k \ \text{even}
\end{cases}
\]

(a) Show that \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} = 2 \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} = \frac{1}{2} \).

(b) Show that \( \sum a_k \) converges if \( p > 1 \) and diverges if \( p \leq 1 \).

**IV.2.16.5.** [?] Let \( \sum_{k=1}^{\infty} a_k \) be a convergent infinite series with strictly positive terms, and for each \( n \) let \( t_n = \sum_{k=n}^{\infty} a_k \) be the \( n \)‘th tail sum. Set \( b_k = \sqrt{t_k} - \sqrt{t_{k+1}} \) for each \( k \). Show that \( \sum_{k=1}^{\infty} b_k \) also converges, and that \( \lim_{k \to \infty} \frac{b_k}{a_k} = +\infty \) (note that \( a_k = t_k - t_{k+1} \)), i.e. \( (b_k) \) converges to zero “more slowly” than \( (a_k) \).

**IV.2.16.6.** Let \( \sum_{k=1}^{\infty} a_k \) be a divergent infinite series with strictly positive terms, and for each \( n \) let \( s_n = \sum_{k=1}^{n} a_k \) be the \( n \)‘th partial sum. Set \( b_k = \frac{\sqrt{s_k}}{a_k} \) for each \( k \). Show that \( \sum_{k=1}^{\infty} b_k \) also diverges (its \( n \)‘th partial sum is at least \( \sqrt{s_n} \)), and that \( \lim_{k \to \infty} \frac{b_k}{a_k} = 0 \), i.e. \( (b_k) \) converges to zero “more rapidly” than \( (a_k) \).
IV.2.16.7. Let \( \sum_{k=1}^{\infty} a_k \) be an infinite series with nonincreasing nonnegative terms. Use the Cauchy Condensation Test to show that \( \sum_{k=1}^{\infty} a_k \) converges if and only if \( \sum_{k=1}^{\infty} (2k+1)a_{k^2} \) converges. Then show by the Limit Comparison Test that this series converges if and only if \( \sum_{k=1}^{\infty} ka_{k^2} \) converges.

IV.2.16.8. Let \((x_k)\) be a sequence of positive numbers. Show that
\[
\liminf_{k \to \infty} [k(x_k - 1)] = \liminf_{k \to \infty} \left[k \left(1 - \frac{1}{x_k}\right)\right] \quad \text{and} \quad \limsup_{k \to \infty} [k(x_k - 1)] = \limsup_{k \to \infty} \left[k \left(1 - \frac{1}{x_k}\right)\right]
\]
as follows:
(a) Suppose \( L = \liminf_{k \to \infty} [k(x_k - 1)] \neq -\infty \), and let \( r \in \mathbb{R}, \ r < L \). Then \( k(x_k - 1) > r \) for all sufficiently large \( k \). Show that
\[
\liminf_{k \to \infty} \left[k \left(1 - \frac{1}{x_k}\right)\right] \geq \liminf_{k \to \infty} \frac{kr}{k+r} = r.
\]
(b) Suppose \( L' = \liminf_{k \to \infty} \left[k \left(1 - \frac{1}{x_k}\right)\right] \neq -\infty \), and let \( r \in \mathbb{R}, \ r < L' \). Then \( k \left(1 - \frac{1}{x_k}\right) > r \) for all sufficiently large \( k \). Show that
\[
\liminf_{k \to \infty} [k(x_k - 1)] \geq \liminf_{k \to \infty} \frac{kr}{k-r} = r.
\]
(c) Prove the \( \liminf \) equality in the case where one side is \( -\infty \).
(d) Replace \( x_k \) by \( \frac{1}{x_k} \) in the \( \liminf \) equality to obtain the \( \limsup \) equality.

IV.2.16.9. (a) Use Exercise IV.2.16.8. to prove the version IV.2.4.5. of Raabe’s Test from IV.2.4.4., and the converse implication.
(b) Similarly prove the equivalence of IV.2.4.8. and IV.2.4.9..

IV.2.16.10. This problem gives an alternate proof that an absolutely convergent series of real numbers converges (IV.2.5.2.).

Let \( \sum_{k=1}^{\infty} a_k \) be an absolutely convergent infinite series. For each \( k \) set
\[
b_k = \frac{1}{2}(|a_k| + a_k), \quad c_k = \frac{1}{2}(|a_k| - a_k).
\]
Note that \( a_k = b_k - c_k \) and \( |a_k| = b_k + c_k \).

(a) Show that \( \sum_{k=1}^{\infty} b_k \) and \( \sum_{k=1}^{\infty} c_k \) are nonnegative series which converge by the comparison test.
(b) Conclude from IV.2.16. and IV.2.18. that \( \sum_{k=1}^{\infty} a_k \) also converges.

(c) Where was completeness of \( \mathbb{R} \) used in this proof?
This proof is arguably simpler than the one in IV.2.5.2., but the one in IV.2.5.2. is more widely applicable since it works verbatim for infinite series in \( \mathbb{C} \) and in Banach spaces ().

**IV.2.16.11.** For each \( k \) set \( a_{2k-1} = \frac{1}{\sqrt{k}} \) and \( a_{2k} = \frac{1}{k} - \frac{1}{\sqrt{k}} \).

(a) Show that \( \sum_{k=1}^{\infty} a_k \) and \( \sum_{k=1}^{\infty} a_k^2 \) both diverge.

(b) Show that \( \prod_{k=1}^{\infty} (1 + a_k) \) converges.

**IV.2.16.12.** For each \( k \) set \( a_k = \frac{(-1)^k}{\sqrt{k}} \).

(a) Show that \( \sum_{k=1}^{\infty} a_k \) converges and \( \sum_{k=1}^{\infty} a_k^2 \) diverges.

(b) Show that \( \prod_{k=1}^{\infty} (1 + a_k) \) diverges.

**IV.2.16.13.** (a) Prove the following Cauchy Criterion for infinite products: An infinite product \( \prod_{k=1}^{\infty} u_k \) converges if and only if, for every \( \epsilon > 0 \) there is an \( N \) such that
\[
\left| \left[ \prod_{k=n+1}^{m} u_k \right] - 1 \right| < \epsilon
\]
whenever \( N \leq n < m \).

(b) If \( a_1, \ldots, a_n \) are any numbers, prove that
\[
\left| \left[ \prod_{k=1}^{n} (1 + a_k) \right] - 1 \right| \leq \left[ \prod_{k=1}^{n} (1 + |a_k|) \right] - 1 .
\]

(c) Prove that an absolutely convergent infinite product converges.

**IV.2.16.14.** (a) Let \( (c_k) \) be an unbounded sequence of positive numbers. Choose a strictly increasing sequence \( (k_j) \) such that \( c_{k_j} \geq 2^j \). Set \( a_{k_j} = 2^{-j} \), and \( a_k = 2^{-k} \) for \( k \) not one of the \( k_j \). Show that \( \sum_{k=1}^{\infty} a_k \) converges and \( \sum_{k=1}^{\infty} c_k a_k \) diverges.
(b) Let \((d_k)\) be a sequence of positive numbers with \(\inf_k d_k = 0\). Choose a strictly increasing sequence \((k_j)\) such that \(d_{k_j} \leq 2^{-j}\). Set \(a_{k_j} = 1\); for \(k\) not one of the \(k_j\) set \(a_k = 2^{-k}\) if \(d_k \leq 1\) and \(a_k = \frac{1}{2^n d_k}\) if \(d_k > 1\).

Show that \(\sum_{k=1}^{\infty} a_k\) diverges but \(\sum_{k=1}^{\infty} d_k a_k\) converges.

IV.2.16.15. (a) Let \(\sum_{k=1}^{\infty} a_k\) be a divergent nonnegative infinite series, with partial sums \(s_n > 0\). Show that \(\sum_{k=1}^{\infty} \frac{a_k}{s_k}\) also diverges. [Show that if \(n < m\), then
\[
\sum_{k=n+1}^{m} \frac{a_k}{s_k} \geq 1 - \frac{s_n}{s_m}
\]
and use that \(s_n \to +\infty\).]

(b) Show that \(\sum_{k=1}^{\infty} \frac{a_k}{s_k}\) converges if \(p > 1\). (The case \(p = 1/2\) is Exercise IV.2.16.6.)

This result was proved independently by Abel and Dini.

IV.2.16.16. (a) Let \(\sum_{k=1}^{\infty} a_k\) be a convergent nonnegative infinite series, with tail sums \(t_n > 0\). Show that \(\sum_{k=1}^{\infty} \frac{a_k}{t_k}\) diverges.

(b) Show that \(\sum_{k=1}^{\infty} \frac{a_k}{t_k^{1/p}}\) converges if \(p < 1\).

This result was proved by Dini.

IV.2.16.17. Let \(\sum_{k=1}^{\infty} a_k^{(1)}, \sum_{k=1}^{\infty} a_k^{(2)}, \ldots\) be a sequence of convergent positive infinite series.

(a) Inductively find for each \(m\) a convergent positive series \(\sum_{k=1}^{\infty} b_k^{(m)}\) such that \(\lim_{k \to \infty} \frac{a_k^{(j)}}{b_k^{(m)}} = 0\) for \(1 \leq j \leq m\) and \(\lim_{k \to \infty} \frac{b_k^{(j)}}{b_k^{(m)}} = 0\) for \(j < m\). [Apply the argument of Exercise IV.2.16.5. to the series \(\sum_{k=1}^{\infty} (b_k^{(m-1)} + a_k^{(m)})\).]

(b) Construct a convergent positive infinite series \(\sum_{k=1}^{\infty} b_k\) such that \(\lim_{k \to \infty} \frac{a_k^{(m)}}{b_k} = 0\) for all \(m\), i.e. such that the \(b_k\) go to zero “more slowly” than \(a_k^{(m)}\) for every \(m\). [Let \(b_k = b_k^{(m)}\) for \(n_{m-1} + 1 \leq k \leq n_m\), for a suitably chosen increasing sequence \((n_m)\).]
(c) If each \( \sum_{k=1}^{\infty} a_k^{(m)} \) diverges, similarly construct a divergent positive infinite series \( \sum_{k=1}^{\infty} b_k \) such that \( \lim_{k \to \infty} \frac{b_k}{a_k^{(m)}} = 0 \) for all \( m \), i.e. such that the \( b_k \) go to zero “more rapidly” than \( a_k^{(m)} \) for every \( m \).

**IV.2.16.18. (Euler)** Let \( p_k \) be the \( k \)th prime number (in increasing order). The goal of this problem is to prove that \( \sum \frac{1}{p_k} \) diverges. (This provides another proof that there are infinitely many primes.)

(a) For each \( n \), let \( m_n \) be the integer part of \( \log_2 n \), i.e. \( 2^{m_n} \) is the largest power of 2 which is \( \leq n \). Show that, if \( r \leq n \) and \( r = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s} \), then \( e_j \leq m_n \) for all \( j \).

(b) Let \( n \in \mathbb{N} \), and let \( r \) be the largest index for which \( p_r \leq n \). Show that

\[
\left( \sum_{j=0}^{m_n} \frac{1}{p_1^j} \right) \left( \sum_{j=0}^{m_n} \frac{1}{p_2^j} \right) \cdots \left( \sum_{j=0}^{m_n} \frac{1}{p_r^j} \right) \geq \sum_{k=1}^{n} \frac{1}{p_k} .
\]

[Multiply out the left side.]

(c) Show that the left side of the inequality in (b) is dominated by

\[
\prod_{k=1}^{r} \frac{p_k}{p_k - 1} .
\]

[The factors are partial sums of geometric series.]

(d) Conclude that there are infinitely many \( p_k \), and that the infinite product

\[
\prod_{k=1}^{\infty} \frac{p_k}{p_k - 1}
\]

diverges, and hence that

\[
\sum_{k=1}^{\infty} \frac{1}{p_k}
\]

do also diverges. [Use IV.2.12.16.]

**IV.2.16.19.**

(a) Show that a revision of a revision of an infinite series \( \sum a_k \) is a revision of \( \sum a_k \).

(b) Say that two infinite series are **associates** if they are both revisions of the same infinite series. Find necessary and sufficient conditions that two given series are associates.

(c) Say that two absolutely convergent infinite series are **strong associates** if they are both revisions of the same absolutely convergent infinite series. Find necessary and sufficient conditions that two given absolutely convergent series are strong associates.

(d) Is the property of being associates [resp. strong associates] an equivalence relation on infinite series [resp. absolutely convergent infinite series]?

(e) Answer the same questions if “revisions” is replaced by “drastic revisions.”

507
IV.3. Convergence of Sequences of Functions

If \((x_n)\) is a sequence of numbers, the meaning of an expression like \(x_n \to x\) is unambiguous: there is only one notion of convergence for sequences of numbers (except in rare instances where an exotic topology is being considered, which need not concern us here). Similarly, there is one standard and universal notion of convergence for sequences of points in Euclidean space.

The situation with functions is dramatically different. If \((f_n)\) is a sequence of functions, there is no unambiguous meaning for an expression like \(f_n \to f\). There are many possible meanings (some dependent on context); some of the most common ones are pointwise convergence, uniform convergence, u.c. convergence, a.e. convergence, convergence in measure, mean convergence, weak convergence, \ldots. The reader should form the habit of clearly specifying, or at least clearly understanding, what kind of convergence is being considered in any situation, to avoid writing “meaningless” expressions like \(f_n \to f\). The conceptual understanding that sequences of functions can converge in some senses but not others is one of the most important and basic mindsets an analyst must develop.

Some of the notions of convergence of sequences of functions are stronger (more restrictive) than others. Properties of functions (e.g. continuity) are sometimes automatically preserved under certain types of convergence, but not others; the stronger the notion of convergence, the more properties are preserved, and the larger the number of desirable consequences of the convergence. There are a number of useful results in analysis and topology which allow one to deduce a stronger type of convergence from a weaker one under certain circumstances.

In this section, we will discuss a few of the most important and widely-used notions of convergence of sequences of functions; other notions will appear in certain contexts in other sections. We will emphasize the setting of functions from \(\mathbb{R}\) to \(\mathbb{R}\), but point out how the notions can be extended to more general situations.

IV.3.1. Pointwise and Uniform Convergence

IV.3.1.1. More than once CAUCHY asserted that a (pointwise) limit of a sequence (or sum of an infinite series) of continuous functions is continuous, even after ABEL had observed in 1826 that the statement in CAUCHY’s 1821 text *Cours d’analyse* was false; CAUCHY still claimed in 1833 that the sum of a pointwise convergent Fourier series is continuous, and even in 1853 when he in effect added uniform convergence to the statement in the *Cours d’analyse* (and to its proof), he pointedly declined to retract his previous assertions, representing the change as a “clarification” rather than a “correction.”

A. ROBINSON [Rob66], I. LAKATOS [Lak78], and D. LAUGWITZ [Lau87–Lau89] have tried to claim that CAUCHY could not have just made a mistake about such a thing and that he was thinking either about transfinite “sequences” of functions or the nonstandard real line (or both). This is no more than a combination of wishful thinking and hero worship. CAUCHY was simply wrong. He was a great mathematician (and a strong proponent of increased rigor in mathematics), but he was somewhat prone to errors in his writing (e.g. his “proof” of the Mean Value Theorem; cf. [Bre07]), and his understanding of the foundations of analysis, while far ahead of his predecessors, was still fairly murky compared to today; the mistake in this case was a lack of understanding of the distinction between pointwise and uniform convergence, which was not definitively clarified (by WEIERSTRASS) until considerably later. CAUCHY’S concept of limit was still based on infinitesimals, but to CAUCHY an infinitesimal was a dynamic concept, a variable real number converging to zero. See [Gra00], [Jah03], [Koe91], and [Bre07] for a discussion.
IV.3.2. U.C. Convergence

Uniform convergence is very nice and has important consequences, but it is in a sense too nice – in many natural situations we do not have uniform convergence. But it turns out that a slightly weaker form of convergence is almost as good, and much more common: uniform convergence on compact sets, or u.c. convergence.

IV.3.2.1. Definition. Let \( X \) be a topological space, and \((f_n)\) a sequence of functions from \( X \) to \( \mathbb{R} \) (or a metric space \( Y \)). If \( f : X \to \mathbb{R} \) (or \( Y \)), then \( f_n \to f \) uniformly on compact sets (u.c.) on \( X \) if \( f_n|_K \to f|_K \) uniformly for every compact subset \( K \) of \( X \).

Note that since singletons are compact, u.c. convergence implies pointwise convergence (but usually turns out to be considerably stronger). Uniform convergence obviously implies u.c. convergence. If \( X \) itself is compact, u.c. convergence on \( X \) is the same as uniform convergence.

IV.3.2.2. Perhaps the most common and important situation is when \( X = \mathbb{R}^n \) or a (usually open) subset. Since a subset of \( \mathbb{R}^n \) is compact if and only if it is closed and bounded, \( f_n \to f \) u.c. on \( X \) if and only if \( f_n|_K \to f|_K \) uniformly for every closed bounded subset \( K \) of \( X \) (note that “closed” means closed in \( \mathbb{R}^n \), not just relatively closed in \( X \)).

The name “u.c. convergence” is not as standard as it should be; it is a very convenient and appropriate abbreviation which we will use frequently (and not just because the author got his Ph.D. at Berkeley!) In Complex Analysis, u.c. convergence of sequences of holomorphic functions on an open set in \( \mathbb{C} \) is usually called normal convergence (X.4.1.1.), but this name seems to be rarely used in a broader context.

IV.3.2.3. By considering partial sums, we can also say what is meant for an infinite series of functions to converge u.c.

Here are some examples showing that u.c. convergence is common and considerably weaker than uniform convergence.

IV.3.2.4. Examples. (i) Let \( f_n : \mathbb{R} \to \mathbb{R} \) be given by \( f_n(x) = x/n \). Then \( f_n \to 0 \) pointwise, but not uniformly. But restricted to any bounded set, \( f_n \to 0 \) uniformly, i.e. \( f_n \to 0 \) u.c.

(ii) On \((0, +\infty)\) (or any subset, e.g. \((0, 1)\)), let \( f_n(x) = \frac{1}{nx} \). Again the sequence converges pointwise to 0 but not uniformly. But any compact subset of \((0, +\infty)\) is contained in an interval \([\epsilon, M]\) for some \( \epsilon, M, 0 < \epsilon < M < +\infty \), and on any such interval the sequence converges uniformly to 0. Thus \( f_n \to 0 \) u.c.

(iii) Consider the infinite series \( \sum_{k=0}^{\infty} x^k \) where the terms are regarded as real-valued functions on \((-1, 1)\) (which is the set of \( x \) for which the series converges). The infinite series converges pointwise to \( f(x) = \frac{1}{1-x} \). More precisely,

\[
\begin{align*}
s_n &= \sum_{k=0}^{n} x^k = \frac{1}{1-x} - \frac{x^n}{1-x} \\
\end{align*}
\]

so, if \( 0 < r < 1 \), for \( x \in [-r, r] \), we have

\[
\left| \frac{1}{1-x} - s_n \right| = \left| \frac{x^n}{1-x} \right| \leq \frac{r^n}{1-r}
\]
which can be made uniformly small by taking \( n \) large enough (depending on \( r \)). Since every closed (in \( \mathbb{R} \)) subset of \((-1, 1)\) is contained in \([-r, r]\) for some \( r < 1 \), the infinite series converges u.c. on \((-1, 1)\).

More generally, any power series converges u.c. on its interval (or open disk) of convergence (\( ) \).

(iii) Let \( (f_n) \) be the sequence of \( ) \) of functions from \([0, 1]\) to \( \mathbb{R} \). Then \( f_n \to 0 \) pointwise but not u.c. (i.e. not uniformly since \([0, 1]\) is compact).

**IV.3.2.5.** **Proposition.** Let \( X \) be a locally compact topological space, and \( f_n, f : X \to \mathbb{R} \) (or a metric space). Then \( f_n \to f \) u.c. if and only if \( f_n \to f \) uniformly on some neighborhood of each point of \( X \).

**Proof:** One direction is obvious, since every point of a locally compact space has a compact neighborhood. Conversely, suppose \( f_n \to f \) on some neighborhood of each point of \( X \). If \( K \) is a compact subset of \( X \), for each \( x \in K \) there is an open neighborhood \( U_x \) of \( x \) on which \( f_n \to f \) uniformly. Finitely many \( U_{x_1}, \ldots, U_{x_m} \) of the \( U_x \) cover \( K \), and \( f_n \to f \) uniformly on \( U_{x_1} \cup \cdots \cup U_{x_m} \) and hence on \( K \).

**IV.3.2.6.** A uniform limit of continuous functions is continuous \( ) \). The same is not true for u.c. convergence in complete generality (XI.16.4.1 (b)); the problem is that a general \( X \) may not have enough compact subsets, illustrated by the fact that there are nondiscrete topological spaces in which every compact subset is finite, and on such a space u.c. convergence is just pointwise convergence, which does not necessarily preserve continuity. But under a mild restriction, u.c. convergence does preserve continuity:

**IV.3.2.7.** **Theorem.** Let \( X \) be a topological space, and \( (f_n) \) a sequence of functions from \( X \) to \( \mathbb{R} \) (or a metric space) converging u.c. to a function \( f \). If each \( f_n \) is continuous on compact sets, then \( f \) is continuous on compact sets.

**Proof:** This is an immediate corollary of \( ) \).

**IV.3.2.8.** **Corollary.** Let \( X \) be a compactly generated (e.g. first countable) topological space (XI.11.11.1.), and \( (f_n) \) a sequence of functions from \( X \) to \( \mathbb{R} \) (or a metric space) converging u.c. to a function \( f \). If each \( f_n \) is continuous, then \( f \) is continuous.

**IV.3.2.9.** The corollary applies if \( X \) is any subset of \( \mathbb{R}^n \), or more generally any subset of a metrizable space. The conclusion is obvious from \( ) \) if \( X \) is locally compact, since continuity is a local property.

**IV.3.3.** **Mean and Mean-Square Convergence**

**IV.3.4.** **Exercises**

**IV.3.4.1.** [Dug78, XII.7.5] Let \( X \) and \( Y \) be topological spaces, and \( f, f_n (n \in \mathbb{N}) \) functions from \( X \) to \( Y \). Say \( f_n \to f \) **continuously** if, whenever \( (x_n) \) is a sequence in \( X \) and \( x_n \to x \), then \( f_n(x_n) \to f(x) \) in \( Y \). Show that if \( X \) is first countable and \( Y \) is a metric space, and \( f \) and the \( f_n \) are continuous, then \( f_n \to f \) continuously if and only if \( f_n \to f \) u.c.
Chapter V

Calculus
V.1. Limits of Functions

The limit of a function at a point is a variation of the notion of limit of a sequence, and is one of the most fundamental ideas in calculus and analysis. In this section, we consider only limits of functions from $\mathbb{R}$ to $\mathbb{R}$; in later sections, the situation will be generalized in a number of ways.

V.1.1. Ordinary Limits

V.1.1.1. Suppose $f$ is a real-valued function defined at least on an interval around a real number $a$, except possibly at $a$ itself (i.e. defined at least on a deleted neighborhood of $a$). If we don’t know the value $f(a)$ or even whether this value exists, but can calculate $f(x)$ for all $x$ near $a$ but different from $a$, we want to determine what $f(a)$ “ought to be” (it will be impossible to determine whether this prediction is correct without knowledge about the value precisely at $a$). Geometrically, we consider the graph $y = f(x)$ and cover up the vertical line $x = a$, and try to determine where the point of the graph on this line should be from what can be seen:

When I teach limits in Calculus, I take a piece of string to class and cover up the line $x = a$ on various graphs, and ask the students where the point, if any, under the string is and where it should be. Note that it will be impossible to determine from what can be seen whether there is in fact a point of the graph on this line, or, if so, where it is, but it can in many (but not all) cases be determined where it ought to be. Informally, we will say the limit of the function as $x$ approaches $a$ will be the $y$-coordinate of the point on
this vertical line (i.e. with $x$-coordinate $a$) which “should” be on the graph, if it can in fact be determined, and otherwise that the limit does not exist.

**V.1.1.2.** More formally, as in the case of limits of sequences, we will say that the limit of $f(x)$ as $x$ approaches $a$ is $L$, written

$$\lim_{x \to a} f(x) = L$$

or $f(x) \to L$ as $x \to a$, if $f(x)$ can be made arbitrarily close to $L$ by making $x$ close enough to $a$, but never equal to $a$. This can be phrased similarly to the definition of limit of a sequence, with a slight variation:

**V.1.1.3.** Definition. Let $V$ be a deleted neighborhood of $a \in \mathbb{R}$, $F : V \to \mathbb{R}$ a function, and $L \in \mathbb{R}$. Then the limit of $f(x)$ as $x$ approaches $a$ is $L$ if, for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$|f(x) - L| < \epsilon$$

for all $x \in V$ for which $0 < |x - a| < \delta$.

Note that the inequality “$0 < |x - a| < \delta$” is just a slick way of writing “$a - \delta < x < a + \delta$ and $x \neq a$.”

**V.1.1.4.** Geometrically, $\lim_{x \to a} f(x) = L$ if, for any $\epsilon > 0$, no matter how small, if we draw the horizontal lines $y = L + \epsilon$ and $y = L - \epsilon$ above and below $y = L$, then there are vertical lines $x = a - \delta$ and $x = a + \delta$ to the left and right of $a$, such that the part of the graph $y = f(x)$ between these two vertical lines lies between the two horizontal lines, except possibly when $x = a$ (Figure V.2).

**V.1.1.5.** As with sequences (IV.1.3.9.), the definition of limit of a function can be described as a game or bet. If I claim that $\lim_{x \to a} f(x) = L$ and you challenge me, you give me an $\epsilon > 0$ and I must produce a $\delta > 0$ that works in the definition for your $\epsilon$. If I can, I win; if I can’t, you win. The smaller your $\epsilon$ is, the harder it is for me to find a $\delta$, and the smaller it will have to be in general. But I don’t have to choose my $\delta$ until you give me your $\epsilon$; my $\delta$ can (and in general will) depend on your $\epsilon$. If one $\delta$ works for a given $\epsilon$, then any smaller $\delta$ also works for the same $\epsilon$, so I don’t have to find the largest possible $\delta$ for your $\epsilon$; a smaller one which is easier to find will work just as well. Similarly, a $\delta$ which works for one $\epsilon$ will also work for any larger $\epsilon$. A limit proof will be a demonstration that I can always win the bet no matter what $\epsilon$ you give me.

**V.1.1.6.** To apply the definition of limit directly, the limit $L$ must be “guessed.” Thus, to just show that a limit exists, and especially to show that a limit does not exist, one must work somewhat indirectly in a manner similar to the one in ( ).

**V.1.1.7.** Note that in the definition of limit, the $x$ is allowed to be either to the right of $a$ ($x > a$) or to the left of $a$ ($x < a$). Thus the limit is often called the “two-sided limit” for emphasis. (But “limit” will always for us mean “two-sided limit” unless otherwise specified.) The “limit” as defined in V.1.1.3. is also sometimes called the “ordinary limit” to distinguish it from the extended notions of limit discussed in ( ).

**V.1.1.8.** Of course, $\lim_{x \to a} f(x)$ depends not only on the function $f$ but also on the number $a$: for a fixed $f$, the limit can potentially be defined for any $a$, although in practice the limit is particularly important only for certain $a$’s depending on $f$. 513
Note that there is a huge conceptual difference for a function $f$ between the value of $f$ at $a$ and the limit of $f$ as $x$ approaches $a$. Either can exist without the other, and if they both exist they may or may not be numerically equal. It turns out that for “most” functions encountered naturally, and for “most” $a$, the two both exist and are numerically equal (the notion of continuity discussed in ()); this fact is ultimately very pleasant and significant, but can be confusing for beginners trying to distinguish between the concepts.

The following principle is basic in the study of limits:

**Proposition.** Let $f$ and $g$ be functions defined in at least a deleted neighborhood of $a$. If $f(x) = g(x)$ for all $x$ in a deleted neighborhood of $a$, and $\lim_{x \to a} f(x)$ exists, then $\lim_{x \to a} g(x)$ also exists and equals $\lim_{x \to a} f(x)$.

Note that we do not require that $f(a) = g(a)$, or even that either or both of these numbers are defined.

**Examples.** (i) Let $f(x) = x + 1$, $g(x) = \frac{x^2 - 1}{x - 1}$,

$$h(x) = \begin{cases} x + 1 & \text{if } x \neq 1 \\ 3 & \text{if } x = 1 \end{cases}.$$
We have that $f(1) = 2$, $g(1)$ does not exist, and $h(1) = 3$. The graph of $f$ is a straight line with slope 1 and $y$-intercept 1; the graph of $g$ is the same line with the point $(1, 2)$ removed, and the graph of $h$ is the line with $(1, 2)$ removed and $(1, 3)$ added. If the vertical line $x = 1$ is covered, the graphs appear identical.

We guess that $\lim_{x \to 1} f(x) = 2$. To prove this, let $\epsilon > 0$. We must find a $\delta > 0$ such that

$$|f(x) - 2| = |x + 1 - 2| = |x - 1| < \epsilon$$

whenever $0 < |x - 1| < \delta$. We can clearly take $\delta = \epsilon$, proving the result. This simple formula makes sense since the graph is a straight line with slope 1 (if the line had slope $m$ we would take $\delta = \frac{\epsilon}{m}$; if the function were nonlinear the formula would generally be more complicated).

Since $g(x) = h(x) = f(x)$ for $x \neq 1$, we also have

$$\lim_{x \to 1} g(x) = \lim_{x \to 1} h(x) = 2$$

by V.1.1.10.; these can be proved directly with exactly the same argument as for $f$. Note that $g$ has a limit but no value at 1, while $h$ has both a value and a limit at 1 but they are not numerically equal.

(ii) Set

$$f(x) = \frac{|x|}{x} = \begin{cases} 
1 & \text{if } x > 0 \\
-1 & \text{if } x < 0 
\end{cases}$$

$$g(x) = \begin{cases} 
1 & \text{if } x \geq 0 \\
-1 & \text{if } x < 0 
\end{cases}$$

Then neither $\lim_{x \to 0} f(x)$ nor $\lim_{x \to 0} g(x)$ exists, by an argument almost identical to the one in (i). (These functions have one-sided limits at 0.) Although $f(0)$ does not exist, we have that $g(0) = 1$. Thus $g$ has a value but no limit at 0.

(iii) Let $f(x) = \sin \frac{1}{x}$, $g(x) = x \sin \frac{1}{x}$ for $x \neq 0$. Neither $f$ nor $g$ has a value at 0. Also, $\lim_{x \to 0} f(x)$ does not exist; there are not even one-sided limits. See Figure (). But $\lim_{x \to 0} g(x)$ does exist: it equals 0. For let $\epsilon > 0$. Set $\delta = \epsilon$. If $0 < |x - 0| = |x| < \delta$, we have

$$|g(x) - 0| = \left| x \sin \frac{1}{x} \right| \leq |x| < \delta = \epsilon$$

since $|\sin \frac{1}{x}| \leq 1$ for all $x \neq 0$. See Figure ().

(iv) Let $f(x) = \sqrt{x}$. If $a > 0$, $\lim_{x \to a} f(x)$ exists and equals $\sqrt{a}$. But $\lim_{x \to 0} f(x)$ cannot exist since $f$ is not defined in any entire deleted neighborhood of 0. We do have a one-sided limit of 0 from the right in this case ().
V.1.2. The Sequential Criterion
V.1.3. Limit Theorems and Indeterminate Forms
V.1.4. One-Sided Limits
V.1.5. Extended Limits
V.1.6. Limit Superior and Limit Inferior
V.1.7. Exercises

V.1.7.1. ([dlVP46, 87, Ex.6], attributed to CAUCHY) Let $\phi$ be a real-valued function on an interval $(a, +\infty)$. Suppose $\lim_{x \to +\infty} [\phi(x + 1) - \phi(x)]$ exists (in the usual or extended sense) and equals $L$.

(a) If $\phi$ is bounded on $[b, b + 1]$ for all sufficiently large $b$ (e.g., if $\phi$ is continuous), show that $\lim_{x \to +\infty} \frac{\phi(x)}{x}$ also exists and equals $L$.

(b) Show that $\lim_{x \to +\infty} \frac{\phi(x)}{x}$ need not exist if no boundedness assumption is made. [Consider a Hamel basis for $\mathbb{R}$ over $\mathbb{Q}$ which contains 1 and is not bounded above, and let $\phi(x)$ be the coefficient of 1 in the expansion of $x$ with respect to this basis.] But if the limit does exist, it must equal $L$.

(c) Suppose $\psi$ is a function from $(a, +\infty)$ to $(0, +\infty)$ for which $\lim_{x \to +\infty} \frac{\psi(x + 1)}{\psi(x)}$ exists and equals $L$. Find similar conditions on $\psi$ insuring that $\lim_{x \to +\infty} [\psi(x)]^{1/x}$ exists (and then necessarily equals $L$). [Consider $\phi = \log \psi$.] Compare with IV.2.3.11.
V.2. Continuity on $\mathbb{R}$

“[Mathematicians before the early 19th century] assumed that any sensible formula, such as a polynomial, automatically defined a continuous function. But they had never proved this. In fact, they couldn’t prove it, because they hadn’t defined ‘continuous.’ The whole area was awash with vague intuitions, most of which were wrong.”

Ian Stewart

V.2.1. Continuity at a Point

V.2.2. Continuity on a Set

In discussions and uses of continuity, it is much more common to use the notion of continuity on a set rather than continuity at a point. The transition between the two is simple but involves a subtlety which must be understood. It should be noted that conventions vary in references as to the meaning of the phrase “continuity on a set $A$,” and the convention in any reference should be noted carefully. We have picked the convention which seems the most common and useful.

The definition is a slight variation of ():

V.2.2.1. Definition. Let $f$ be a real-valued function defined on a subset of $\mathbb{R}$, and let $A$ be a subset of the domain of $f$. Then $f$ is continuous on $A$ if, for every $a \in A$ and $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ whenever $x \in A$ and $|x - a| < \delta$.

V.2.2.2. If $f$ is continuous (in the sense of ()) at every point of $A$, then $f$ is continuous on $A$. The converse is false in general, since in V.2.2.1. only values of the function at points of $A$ are considered; if a point $a$ of $A$ is a limit point of $A^c$ the behavior of $f$ on $A^c$ can force $f$ to be discontinuous at $a$, but will not affect whether $f$ is continuous on $A$.

V.2.2.3. Examples. (i) Let

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}.$$  

Then $f$ is discontinuous at 0, but it is continuous on $[0, \infty)$.

(ii) More generally, if $f = \chi_A$ is the indicator function of a set $A \subseteq \mathbb{R}$, then $f$ is continuous on $A$ since it is a constant function on $A$. Similarly, $f$ is continuous on $A^c$. But $f$ can be discontinuous at many numbers: e.g. if $A = \mathbb{Q}$, $f$ is discontinuous everywhere (i.e. at every $a \in \mathbb{R}$).

(iii) If $A = \{a\}$, then any real-valued function defined at $a$ is continuous on $\{a\}$ since the condition in the definition is essentially vacuous. Thus “continuity on $\{a\}$” is quite different from “continuity at $a$.” The similarity of the terminology is somewhat unfortunate, but such near-ambiguity is unavoidable.

V.2.2.4. Continuity on a set $A$ is most frequently used when $A$ is either an open set or a closed or half-open interval. In these cases, the ambiguities largely disappear:

1[?, p. 151].

517
**V.2.2.5.** Proposition. Let \( f \) be a real-valued function on a subset of \( \mathbb{R} \), and \( U \) an open set in \( \mathbb{R} \) contained in the domain of \( f \). Then \( f \) is continuous on \( U \) if and only if \( f \) is continuous (in the sense of (i)) at \( a \) for every \( a \in U \).

**V.2.2.6.** Proposition. Let \( f \) be a real-valued function on a subset of \( \mathbb{R} \), and \( I \) an interval in \( \mathbb{R} \) contained in the domain of \( f \). Then \( f \) is continuous on \( I \) if and only if \( f \) is continuous (in the sense of (i)) at \( a \) for every \( a \) in the interior of \( I \), and right continuous () at the left endpoint of \( I \) and left continuous () at the right endpoint of \( I \) when there are such endpoints in \( I \).

In other words, \( f \) is continuous on \([a, b]\) if and only if it is continuous at every point of \((a, b)\), right continuous at \( a \), and left continuous at \( b \); it is continuous on \([a, b]\) if and only if it is continuous at every point of \((a, b)\) and right continuous at \( a \), etc.

**V.2.2.7.** We say a function \( f \) defined on a subset of \( \mathbb{R} \) is **continuous** if it is continuous on its domain.

### Three Characterizations of Continuity

There is a sequential criterion for continuity on a set:

**V.2.2.8.** Theorem. [Sequential Criterion] Let \( f \) be a real-valued function on a subset of \( \mathbb{R} \), and \( A \) a subset of \( \mathbb{R} \) contained in the domain of \( f \). Then \( f \) is continuous on \( A \) if and only if, whenever \( (x_n) \) is a sequence in \( A \), \( x_n \to x \), and \( x_n \to x \), we have \( f(x_n) \to f(x) \).

Note that in this criterion, the terms \( x_n \) of the sequence and the limit \( x \) must all be in \( A \).

**Proof:** Suppose \( f \) is continuous on \( A \), and let \( (x_n) \) be a sequence in \( A \) converging to \( x \in A \). Let \( \epsilon > 0 \). Since \( f \) is continuous at \( x \), there is a \( \delta > 0 \) such that \( |f(y) - f(x)| < \epsilon \) for all \( y \in A \) with \( |y - x| < \delta \). There is an \( N \) such that \( |x_n - x| < \delta \) for all \( n \geq N \). Then \( |f(x_n) - f(x)| < \epsilon \) for all \( n \geq N \). Thus \( f(x_n) \to f(x) \).

Conversely, suppose the sequential criterion is satisfied. If \( f \) is not continuous on \( A \), then there is an \( x \in A \) and \( \epsilon > 0 \) for which there is no corresponding \( \delta \) with \( |f(y) - f(x)| < \epsilon \) whenever \( y \in A \) and \( |y - x| < \delta \); thus, for this \( x \) and \( \epsilon \), for each \( n \in \mathbb{N} \) there is an \( x_n \in A \) with \( |x_n - x| < \frac{1}{n} \) but \( |f(x_n) - f(x)| \geq \epsilon \). So \( x_n \to x \) but \( f(x_n) \not\to f(x) \), a contradiction.

Note that the Countable AC is used in this proof. (In fact, this theorem cannot be proved in ZF.)

There is a third standard characterization of continuity in terms of preimages:

**V.2.2.9.** Theorem. Let \( A \subseteq \mathbb{R} \), and \( f : A \to \mathbb{R} \) a function. Then the following are equivalent:

(i) \( f \) is continuous on \( A \).

(ii) The preimage of every open set in \( \mathbb{R} \) is (relatively) open in \( A \).

(iii) The preimage of every open interval in \( \mathbb{R} \) is (relatively) open in \( A \).

(iv) The preimage of every closed set in \( \mathbb{R} \) is (relatively) closed in \( A \).
Proof: (ii) and (iv) are transparently equivalent since preimages respect complements (). (ii) \( \Rightarrow \) (iii) is trivial.

(i) \( \Rightarrow \) (ii): Suppose \( f \) is continuous, and let \( U \) be an open set in \( \mathbb{R} \), and \( x \in f^{-1}(U) \). There is an \( \epsilon > 0 \) such that \( \{ y \in \mathbb{R} : |y - f(x)| < \epsilon \} \subseteq U \) since \( U \) is open, and there is a \( \delta > 0 \) such that \( |f(z) - f(x)| < \epsilon \) whenever \( z \in A \) and \( |z-x| < \delta \) since \( f \) is continuous at \( x \). Thus if \( z \in A \), \( |z-x| < \delta \), then \( f(z) \in U \), i.e. \( z \in f^{-1}(U) \).

So \( f^{-1}(U) \) is open.

(iii) \( \Rightarrow \) (i): Let \( x \in A \) and \( \epsilon > 0 \). Let \( U = \{ y \in \mathbb{R} : |y - x| < \epsilon \} \). Then \( U \) is an open interval, so \( f^{-1}(U) \) is relatively open in \( A \) by assumption. Since \( x \in f^{-1}(U) \), there is a \( \delta > 0 \) such that \( A \cap (x-\delta, x+\delta) \subseteq f^{-1}(U) \).

If \( z \in A \), \( |z-x| < \delta \), then \( z \in f^{-1}(U) \), i.e. \( f(z) \in U \), \( |f(z) - f(x)| < \epsilon \). So \( f \) is continuous at \( x \).

V.2.2.10. Thus we have three equivalent characterizations of continuous functions:

(i) The \( \epsilon - \delta \) characterization (definition).

(ii) The sequential criterion.

(iii) The preimage characterization.

Each is useful in different contexts: some applications are much easier using one or the other characterization.

There are more general settings for continuity: metric spaces () and, even more generally, topological spaces (). Characterization (i) is special to metric spaces. The sequential criterion is valid more generally in topological spaces for which sequences determine the topology (first-countable topological spaces). Only the third is appropriate in general topological spaces, and is taken as the definition of continuity in general spaces. The three characterizations are equivalent in metric spaces (). The definition V.2.2.1. precisely says that a function from \( \mathbb{R} \) (or a subset) to \( \mathbb{R} \) is continuous on a subset \( A \) of its domain if and only if the restriction of \( f \) to \( A \) is continuous in this generalized sense when \( A \) is given the relative or subspace topology.

V.2.3. The Intermediate Value Theorem

One of the most important properties of a continuous function on an interval is that its range is also an interval, i.e. that it takes all intermediate values between any two of its values:

V.2.3.1. **Theorem.** [Intermediate Value Theorem (IVT)] Let \( I \) be an interval in \( \mathbb{R} \), \( f : I \to \mathbb{R} \) a continuous function, and \( a, b \in I \). If \( d \) is any real number between \( f(a) \) and \( f(b) \), then there is at least one \( c \) between \( a \) and \( b \) with \( f(c) = d \).

V.2.3.2. Geometrically, the graph of \( f \) gives an “unbroken” curve between \( (a, f(a)) \) and \( (b, f(b)) \), and thus any horizontal line of intermediate height must intersect it (Figure () is the case \( f(a) < f(b) \); the other case is similar).

This picture, although informal, gives a clue how to prove the theorem from the Completeness Axiom (): take the supremum of the \( x \)-coordinates of all points on the graph below (or on) the line; this should be the \( x \)-coordinate of the “last” point of intersection (we could equally well take the infimum for points above the line, giving the “first” point of intersection; of course, there may be many other points of intersection too). In fact, the IVT is equivalent to the Completeness Axiom (Exercise ()).
Proof: We may assume without loss of generality that $a < b$. Suppose $f(a) < f(b)$ (the case $f(a) > f(b)$ is similar, and if $f(a) = f(b)$ there is nothing to prove). Fix $d$ with $f(a) < d < f(b)$, and set

$$S = \{ x \in [a,b] : f(x) \leq d \}.$$  

Then $S$ is a nonempty subset of $[a,b]$ since $a \in S$. Set $c = \sup S$. We claim $f(c) = d$. If $f(c) \neq d$, then there is a $\delta > 0$ such that $|f(x) - f(c)| < |f(c) - d|$ for $x \in [a,b]$, $|x - c| < \delta$. If $f(c) > d$, then $f(x) > d$ for $c - \delta < x < c$ and $c - \delta$ is an upper bound for $S$, a contradiction. And if $f(c) < d$, then $c \neq b$ and if $x \in [a,b]$, $c < x < c + \delta$, then $f(x) < d$, contradicting that $c$ is an upper bound for $S$. 

**V.2.3.3.** Corollary. Let $f$ be a continuous real-valued function on an interval $I$. Then $f(I)$ is an interval (or single point if $f$ is constant on $I$).

**V.2.3.4.** This result can be done in the more general context of connected subsets of topological spaces: a continuous image of a connected set is connected (XI.13.2.1). The connected subsets of $\mathbb{R}$ are precisely the intervals and singletons (degenerate intervals); cf. XI.13.1.7.

**V.2.3.5.** The IVT has many applications. For example, we get a simple proof of III.1.12.2: if $f(x) = x^n$, then $f(0) = 0$ and $f(1) = 1$, and $f$ is continuous, so $f$ takes all values between 0 and 1 on $[0,1]$, i.e. every number between 0 and 1 has an $n$’th root between 0 and 1 (which is unique since $f$ is strictly increasing). For numbers larger than 1, consider reciprocals.

We also get the following important algebraic fact:

**V.2.3.6.** Theorem. Every polynomial of odd order with real coefficients has at least one real root.

Proof: There is no loss of generality to consider monic polynomials. So let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$$

be a monic polynomial with $n$ odd and $a_k \in \mathbb{R}$. We have

$$\frac{f(x)}{x^n} = 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}$$

for $x \neq 0$, so if

$$|x| > \max (1, n|a_0|, \ldots, n|a_{n-1}|)$$

we have $\frac{f(x)}{x^n} > 0$. Thus, if $x$ is sufficiently large positive, $f(x) > 0$. And, since $n$ is odd, $x^n < 0$ if $x < 0$, so $f(x) < 0$ if $x$ is sufficiently large negative. Thus $f$ takes both positive and negative values, so also takes the value 0 by the IVT since it is continuous.
V.2.3.7. This can also be proved from the Fundamental Theorem of Algebra (X.1.4.2.), since for polynomials with real coefficients the nonreal complex roots come in conjugate pairs, so there are an even number of nonreal roots. This argument gives the slightly stronger result that a polynomial of odd degree with real coefficients always has an odd number of real roots (counted with multiplicity); this can be proved directly from V.2.3.6. by dividing out a linear factor for each root. If an even number of factors is divided out, the resulting polynomial still has odd degree.

V.2.3.8. A similar argument can be efficiently used to locate real roots of a polynomial or any other continuous function: find a decreasing sequence of subintervals on which the function changes sign, say by successive bisection; the point of intersection of the intervals is a root.

V.2.4. The Max-Min Theorem
A second crucial result about continuous functions concerns attaining maximum and minimum values on closed bounded sets. The proof uses the Bolzano-Weierstrass Theorem (); in fact this result is also equivalent to the Completeness Axiom (Exercise ()).

V.2.4.1. Theorem. [Max-Min Theorem or Extreme Value Theorem] Let \( f \) be a continuous real-valued function on a closed bounded subset \( A \) of \( \mathbb{R} \). Then \( f \) is bounded on \( A \), and attains its maximum and minimum on \( A \): there are \( c,d \in A \) such that \( f(d) \leq f(x) \leq f(c) \) for all \( x \in A \).

Proof: We first show that \( f \) is bounded above. If not, then there is a sequence \( (x_n) \) in \( A \) with \( f(x_n) \geq n \) for all \( n \). By Bolzano-Weierstrass, there is a convergent subsequence \( (x_{k_n}) \) since \( A \) is bounded, say \( x_{k_n} \to a \). We have \( a \in A \) since \( A \) is closed. But then \( f(x_{k_n}) \to f(a) \) by the Sequential Criterion, a contradiction since \( (f(x_{k_n})) \) is unbounded (IV.1.3.16.).

Let \( M = \sup_{x \in A} f(x) \). Then, for each \( n \in \mathbb{N} \), there is an \( x_n \in A \) with \( f(x_n) > M - \frac{1}{n} \). Thus \( f(x_n) \to M \). By B-W, there is a convergent subsequence \( (x_{k_n}) \). Set \( c = \lim_{n \to \infty} x_{k_n} \). Then \( c \in A \), and \( f(x_{k_n}) \to f(c) \) by the Sequential Criterion. But \( f(x_{k_n}) \to M \) by (); thus \( f(c) = M \).

The argument that \( f \) is bounded below and attains its minimum is almost identical. \( \Box \)

Note that this proof uses the Countable AC. (But the result does not require any form of choice; cf. XI.11.1.8.) Also note that, as in the IVT, the \( c \) and \( d \) are not necessarily unique.

V.2.4.2. What is really going on here is compactness: a subset of \( \mathbb{R} \) is compact if and only if it is closed and bounded (XI.11.1.5.), and a continuous image of a compact set is compact (XI.11.1.7.).

V.2.4.3. Examples. The result may fail if the set \( A \) is not closed or not bounded, or if \( f \) is not continuous on \( A \).

(i) The function \( f(x) = x \) is not bounded on \( [0, \infty) \). It is bounded on \( A = [0,1) \) but does not attain a maximum on \( A \).

(ii) The function \( g(x) = \frac{1}{x} \) is continuous on the bounded set \( (0,1] \) but is not bounded above. If we set

\[
g(x) = \begin{cases} \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}
\]
then $g$ is not bounded above on the closed bounded interval $[0, 1]$ (of course, it is not continuous on $[0, 1]$).

(iii) Set

$$f(x) = \begin{cases} x & \text{if } x \neq 1 \\ 0 & \text{if } x = 1 \end{cases}$$

Then $f$ is bounded above on $[0, 1]$, but does not attain a maximum on $[0, 1]$ (again, it is not continuous on $[0, 1]$).

**V.2.5. Continuity and Oscillation**

It is often fruitful to think of continuity of functions in terms of oscillations.

**V.2.5.1. Definition.** Let $f$ be a real-valued function defined on an interval $I$.

(i) If $\emptyset \neq A \subseteq I$, define the oscillation of $f$ on $A$ to be

$$\omega(f, A) = \sup\{|f(x) - f(y)| : x, y \in A\} \in [0, +\infty].$$

(ii) If $x \in I$, define the oscillation of $f$ at $x$ to be

$$\omega(f, x) = \inf\{\omega(f, U) : U \text{ a neighborhood of } x\} \in [0, +\infty].$$

The next proposition is obvious.

**V.2.5.2. Proposition.** (i) If $\emptyset \neq A \subseteq B \subseteq I$, then $\omega(f, A) \leq \omega(f, B)$.

(ii) If $x \in I$, then $\omega(f, x) = \inf\{\omega(f, U) : U \text{ an open neighborhood of } x\}$.

(iii) $f$ is continuous at $x \in I$ if and only if $\omega(f, x) = 0$.

*Proof of (iii):* If $\omega(f, x) = 0$, then for every $\epsilon > 0$ there is a neighborhood $U$ of $x$ such that $\omega(f, U) < \epsilon$. There is a $\delta > 0$ such that $(x - \delta, x + \delta) \subseteq U$. Then $\omega(f, (x - \delta, x + \delta)) < \epsilon$ by (i), so if $y \in (x - \delta, x + \delta)$, we have $|f(x) - f(y)| < \epsilon$.

Conversely, suppose $f$ is continuous at $x$, and $\epsilon > 0$. There is a $\delta > 0$ such that if $y \in (x - \delta, x + \delta)$, $|f(x) - f(y)| < \frac{\epsilon}{2}$. Then, if $y, z \in (x - \delta, x + \delta)$,

$$|f(y) - f(z)| \leq |f(y) - f(x)| + |f(x) - f(z)| < \epsilon$$

so $\omega(f, (x - \delta, x + \delta)) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $\omega(f, x) = 0$.

The next observation is important and useful:
V.2.5.3. **Proposition.** Let \( f \) be a real-valued function on an open interval \( I \), and \( \epsilon > 0 \). Then
\[
\{ x \in I : \omega(f, x) < \epsilon \}
\]
is an open set.

**Proof:** Call the set \( A \). If \( x \in A \), then there is an open neighborhood \( U \) of \( x \) with \( \omega(f, U) < \epsilon \). If \( y \in U \), then \( U \) is also an open neighborhood of \( y \), so \( \omega(f, y) \leq \omega(f, U) < \epsilon \). Thus \( y \in A \), \( U \subseteq A \).

V.2.5.4. **Corollary.** Let \( f \) be a real-valued function on an interval \( I \). Then the set of points where \( f \) is continuous is a \( G_\delta \), i.e. the set of discontinuities of \( f \) is an \( F_\sigma \).

**Proof:** We may assume that \( I \) is open (why?) If \( A \) is the set of points where \( f \) is continuous, then
\[
A = \{ x \in I : \omega(f, x) = 0 \} = \bigcap_{n=1}^{\infty} \left\{ x \in I : \omega(f, x) < \frac{1}{n} \right\}.
\]
V.3. Differentiation and Derivatives

The derivative of a function is of great importance, both theoretical and practical. It is the principal topic in any calculus course. The process of finding and using the derivative of a function is called differentiation.

V.3.1. The Derivative

The derivative of a function can be described in many ways: the instantaneous rate of change of the function, the slope of the tangent line to the graph of the function, and the best linear approximation to the function are three of the most common.

The Tangent Line to the Graph of a Function

V.3.1.1. There is an intuitive notion of “tangent line” to a curve at a point on the curve. Informally, the tangent line to the curve at the point $P$ is the line which just “touches” the curve at $P$ but does not cross it (indeed, the word “tangent” comes from the Latin tangere, which means “to touch”). This definition is adequate for some simple curves: for example, the tangent line to a circle at a point $P$ is the line perpendicular to the diameter line of the circle through $P$ (this is easily seen to be the only line through $P$ which does not “cross” the curve). Indeed, this turns out to be the “generic” situation.

V.3.1.2. But simple examples show that this intuitive definition is inadequate in general. For example, the $x$-axis is clearly the only reasonable candidate for the tangent line to the curve $y = x^3$ at the origin, but it “crosses” the curve there:

![Maple plot of graph of $f(x) = x^3$](maple_plot.png)

Figure V.3: Maple plot of graph of $f(x) = x^3$
V.3.1.3. For more complicated curves, such as the graph of \( f(x) = x^2 \sin \frac{1}{x^2} \) (extended by setting \( f(0) = 0 \)), this definition is even more inadequate (Figure V.4).

![Maple plot of graph of \( f(x) = x^2 \sin \frac{1}{x^2} \)](image)

Figure V.4: Maple plot of graph of \( f(x) = x^2 \sin \frac{1}{x^2} \)

It is clear that the “tangent line” at the origin, if it exists, must again be the \( x \)-axis, but this line crosses the curve infinitely often in any interval around 0.

V.3.1.4. In addition, if there is a line through \( P \) which does not “cross” the curve, there may be more than one. For example, consider the curve \( y = |x| \). Any line through the origin with slope strictly between \(-1\) and \(1\) just “touches” the curve and does not cross it. Which, if any, of these lines should be considered the “tangent line”?

V.3.1.5. Finally, suppose the curve itself is a straight line. What could the “tangent line” be but the line itself? But how does this line just “touch” the original line?
The Difference Quotient

**V.3.1.6.** Mathematicians wrestled for a long time with the difficulties in making a proper definition of tangent line, and finally came up with a satisfactory dynamic definition, now recognized as a limit.

If we want to define the tangent line at \( P \), we take an indirect approach. Suppose \( Q \) is another point on the curve. Then there is a unique line through \( P \) and \( Q \), called the *secant line* to the curve through \( P \) and \( Q \). (Note that, as with tangent lines, the term “secant line” has nothing to do with trigonometry; the word “secant” comes from the Latin *secare*, “to cut”.) If \( P \) is held fixed and \( Q \) “approaches” \( P \), the secant line should “approach” the tangent line at \( P \).

**V.3.1.7.** To make this precise, suppose the curve is the graph of a function \( f \), and that \( P = (a; f(a)) \). If \( Q = (b; f(b)) \), the line through \( P \) and \( Q \) has slope
\[
\frac{f(b) - f(a)}{b - a}
\]
as long as \( P \neq Q \), i.e. \( b \neq a \).

We want to regard \( a \) as fixed and let \( b \) vary, so we usually use a variable name like \( x \) or \( t \) instead of \( b \).

**V.3.1.8.** Definition. Let \( f \) be a function defined on an open interval \( I \) around \( a \in \mathbb{R} \). Then the *difference quotient* of \( f \) at \( a \) with variable \( x \) is
\[
\frac{f(x) - f(a)}{x - a}
\]
which is defined for \( x \in I, x \neq a \).

**Definition of the Derivative**

We can then make the main definition:

**V.3.1.9.** Definition. Let \( f \) be defined in an open interval around \( a \in \mathbb{R} \). Then \( f \) is *differentiable at \( a \)* if
\[
\lim_{x \to a} \frac{f(x) - f(a)}{x - a}
\]
exists. If it exists, it is called the *derivative* of \( f \) at \( a \), denoted \( f'(a) \).

Note that this is an *ordinary two-sided limit*, i.e. \( x \) can be either larger or smaller than \( a \), and the limit must be finite, not \( \pm \infty \).

Then we can make a careful definition of tangent line:

**V.3.1.10.** Definition. Let \( f \) be a function defined on an open interval \( I \), and \( a \in I \). If \( f \) is differentiable at \( a \), then the *tangent line* to the graph \( y = f(x) \) at \( a \) is the line through \( (a, f(a)) \) with slope \( f'(a) \).

**V.3.1.11.** If \( f \) is not differentiable at \( a \), we will generally say that the graph has no tangent line at \( (a, f(a)) \). (This may not always be accurate: the graph of a function can have a vertical tangent line at some points. We will discuss this more carefully in ( ).)
Differentiability vs. Continuity

We will see in the next few sections that “most” functions which are naturally encountered are differentiable wherever they are defined. However, there is one obvious situation which prevents differentiability: a discontinuity. Since a differentiable function is “approximately linear” near \( a \), and linear functions are continuous, it is not surprising that differentiable functions are continuous:

**V.3.1.12. Proposition.** Let \( f \) be a function defined on an open interval \( I \), and \( a \in I \). If \( f \) is differentiable at \( a \), then \( f \) is continuous at \( a \).

**Proof:** If \( f \) is differentiable at \( a \), then

\[
\lim_{x \to a} \frac{f(x) - f(a)}{x - a}
\]

exists. Since \( \lim_{x \to a} (x - a) = 0 \), we must also have \( \lim_{x \to a} (f(x) - f(a)) = 0 \) by (). But this is just the definition of continuity of \( f \) at \( a \). \( \blacksquare \)

**V.3.1.13.** However, continuity of \( f \) at \( a \) is not sufficient to give differentiability at \( a \), as the following example shows. Thus differentiability at \( a \) is a strictly stronger than condition than continuity.

**V.3.1.14. Example.** Let \( f(x) = |x| \). Then \( f \) is continuous at \( 0 \) (). But \( f \) is not differentiable at \( 0 \): the difference quotient

\[
\frac{f(x) - f(0)}{x - 0} = \frac{|x|}{x}
\]

is 1 if \( x > 0 \) and \(-1\) if \( x < 0 \). Thus the limit as \( x \to 0 \) does not exist ().

Here is a dramatic example that shows that differentiability at \( a \) does not imply continuity anywhere except at \( a \):

**V.3.1.15. Example.** Let

\[
f(x) = \begin{cases} 
x^2 & \text{if } x \in \mathbb{Q} \\
0 & \text{if } x \notin \mathbb{Q}
\end{cases}
\]

Then \( f \) is discontinuous everywhere except at \( 0 \) (). But \( f \) is differentiable at \( 0 \) since

\[
\left| \frac{f(x) - f(0)}{x - 0} \right| = \left| \frac{f(x)}{x} \right| \leq |x|
\]

for \( x \neq 0 \), and hence \( f'(0) \) exists and equals 0 by the Squeeze Theorem ().
**Geometric Interpretation of the Tangent Line**

**V.3.1.16.** A somewhat informal but conceptually excellent way to interpret the tangent line to a curve is by magnification. If a curve has a tangent line at the point \( P \), then when smaller and smaller pieces of the curve around \( P \) are magnified more and more, they become increasingly almost indistinguishable from the tangent line at \( P \) (see e.g. https://www.ima.umn.edu/~arnold/calculus/tangent/tangent-g.html, http://archives.math.utk.edu/visual.calculus/2/tangents.5/, or https://www.geogebra.org/material/simple/id/335793).

Conversely, if the curve becomes more and more indistinguishable from a straight line in small neighborhoods of \( P \), then this line is the tangent line at \( P \), and if the curve is the graph of a function and the tangent line is not vertical, its slope is the derivative of the function at that point.

Although this approach to the tangent line is not rigorous without additional work, it is much more precise and universally applicable than the idea of the tangent line just “touching” the curve at the point of tangency.

**V.3.1.17.** The magnification interpretation gives a satisfactory geometric explanation of why a function such as \( f(x) = |x| \) is not differentiable. If an arbitrarily small segment of this graph around 0 is magnified arbitrarily much, it still looks exactly like the original curve and does not ever resemble a straight line segment.

**V.3.1.18.** Note, however, that even if a curve has a tangent line at a point \( P \), and a small segment of the curve near \( P \) when magnified looks approximately like the tangent line, if an even smaller piece of the magnified segment not including the point \( P \) is further magnified, it does not need to bear any resemblance to the tangent line:

The magnification principle only applies to segments of the curve around \( P \). In fact, the curve need not even have a tangent line at any other point than \( P \), i.e. if it is the graph of a function, the function need not be differentiable except at \( P \); it can even fail to be continuous except at \( P \)! For examples, see V.3.1.15. and V.3.9.6.

**Rate of Change Interpretation**

**V.3.1.19.** The difference quotient and derivative can be interpreted as rates of change. If \( f \) is a function defined on an interval containing \( a \) and \( b \), then the average rate of change of \( f \) from \( a \) to \( b \) can be reasonably defined as the total change in \( f \) divided by the total change in \( x \). The total change in \( f \) from \( a \) to \( b \) is \( f(b) - f(a) \), and the total change in \( x \) from \( a \) to \( b \) is \( b - a \). Thus the average rate of change in \( f \) from \( a \) to \( b \) is

\[
\frac{f(b) - f(a)}{b - a}
\]

provided \( b \neq a \). (Here we could have \( a < b \) or \( a > b \).)

**V.3.1.20.** If we hold \( a \) fixed and let \( b \) vary, i.e. replace \( b \) by a variable \( x \), then as \( x \to a \) the average rate of change of \( f \) from \( a \) to \( x \) should approach the “instantaneous rate of change of \( f \) at \( a \)” Thus

\[
f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}
\]

can be interpreted as the instantaneous rate of change of \( f \) at \( a \).
V.3.1.21. The most natural instance of this interpretation is if \( f(t) \) denotes the position of an object at time \( t \). Then
\[
\frac{f(b) - f(a)}{b - a}
\]
naturally denotes the average velocity of the object between \( t = a \) and \( t = b \), and
\[
f'(a) = \lim_{t \to a} \frac{f(t) - f(a)}{t - a}
\]
has a natural interpretation as the instantaneous velocity of the object at \( t = a \).

The Derivative as a Function

Other Notations for the Derivative

V.3.1.22. The formula for the derivative \( f'(a) \) in V.3.1.9. is frequently rewritten by writing \( h = x - a \), so that \( x = a + h \); then
\[
f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}.
\]

V.3.1.23. Another variation is that the \( h = x - a \) is written \( \Delta x \), thought of as the “change in \( x \).” Note that \( \Delta x \) is regarded as a single symbol and that the \( \Delta \) and \( x \) have no independent meaning related to this symbol. Thus
\[
f'(a) = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}.
\]
This notation is more commonly used when the \( a \) is replaced by \( x \), which is treated like a constant in the calculation of the limit, and then as a variable:
\[
f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.
\]
If another symbol such as \( t \) is used for the independent variable (e.g. when it is interpreted as time), the corresponding replacement is normally made in the change symbol:
\[
f'(t) = \lim_{\Delta t \to 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}.
\]

V.3.1.24. Carrying this a step further, we often work with the equation \( y = f(x) \) and define
\[
\Delta y = \Delta f = f(x + \Delta x) - f(x)
\]
where \( x \) and \( x + \Delta x \) are in the domain of \( f \). With this notation,
\[
f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}.
\]
This notation nicely matches up with the “rate of change” interpretation: \( \frac{\Delta y}{\Delta x} \) is the average rate of change of \( y \) with respect to \( x \) between \( x \) and \( x + \Delta x \).
V.3.1.25. This leads to the notation originally used by Leibniz in his development of calculus, not surprisingly called Leibniz notation. We use the symbol $\frac{dy}{dx}$ or $\frac{df}{dx}$ to denote the derivative:

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$  

Note that $\frac{dy}{dx}$ (pronounced “d-y-d-x”) is regarded as a single symbol and is not a fraction, i.e. the $dy$ and $dx$ have no independent meaning (although meanings are sometimes attached to them in such a way as to make the quotient equal to $\frac{dy}{dx}$). The $x$ and $y$ can be replaced by any other variable names.

V.3.1.26. This notation arose because Leibniz thought of $dx$ as an “infinitesimal change” in $x$ and $dy$ the corresponding “infinitesimal change” in $y = f(x)$; the quotient $\frac{dy}{dx}$ should then (up to an “infinitesimal error”) be the exact slope of the tangent line. (Actually functional notation did not exist at that time, and one thought only about independent and dependent variables related by some formula.) The exact meaning of the term “infinitesimal change” was unclear, and was the principal logical weak point of calculus until the nineteenth century when infinitesimals were effectively eliminated from the rigorous calculus developed then. They have reappeared in a rigorous way in “nonstandard analysis” (1). Infinitesimals have never disappeared as a useful intuitive device in calculus and its applications, however.

V.3.1.27. In some ways, the Leibniz notation is inferior to the prime notation, mainly because in Leibniz notation it is not explicit where the derivative is being evaluated. However, for many purposes Leibniz notation is efficient and suggestive, so it continues to be commonly used. For example, if $y$ is a function of $t$, then $\frac{dy}{dt}$ has a natural interpretation as the (instantaneous) rate of change of $y$ with respect to $t$.

V.3.1.28. As a variation of Leibniz notation, we define an operator $\frac{d}{dx}$, which operates on (differentiable) functions and yields the derivative, i.e.

$$\frac{d}{dx} \left( f(x) \right) = f'(x).$$

This notation is often convenient for expressing differentiation formulas; cf. V.3.2.1. Again, $\frac{d}{dx}$ is regarded as a single symbol.

V.3.2. Elementary Rules of Differentiation

V.3.2.1. Theorem. [Power Rule, Version 1] Let $n \in \mathbb{N}$, and $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^n$. Then $f$ is differentiable at any $a \in \mathbb{R}$, and

$$f'(a) = na^{n-1}.$$  

As the Power Rule is often written,

$$\frac{d}{dx} (x^n) = nx^{n-1}.$$  

Proof: The following formula for any $x$ and $a$ in $\mathbb{R}$ can be quickly verified by multiplying out the right side and telescoping:

$$x^n - a^n = (x - a) \sum_{k=0}^{n-1} x^{n-1-k}a^k.$$
(This formula should technically be proved by induction.) Thus we have

\[ f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a} \frac{x^n - a^n}{x - a} = \lim_{x \to a} \sum_{k=0}^{n-1} x^{n-1-k} a^k. \]

The last expression is a polynomial \( p(x) \), hence by (1) the last limit exists and equals \( p(a) = na^{n-1} \).

**V.3.2.2.** This result will be extended to negative integer \( n \) in (1), to fractional \( n \) in (1), and to arbitrary real \( n \) in (1). The same derivative formula holds in all cases.

The Power Rule can be proved for positive integer \( n \) more easily by induction using the Product Rule (cf. Exercise V.3.10.7.). It can also be proved using the Binomial Theorem (Exercise V.3.10.1.).

**The Sum and Constant Multiple Rules**

**V.3.2.3.** Proposition. [Sum Rule] Let \( f \) and \( g \) be functions from an open interval \( I \) to \( \mathbb{R} \), and \( h = f + g : I \to \mathbb{R} \). If \( f \) and \( g \) are differentiable at \( a \in I \), then \( h \) is differentiable at \( a \) and

\[ h'(a) = f'(a) + g'(a). \]

**Proof:** By (1), we have

\[
\lim_{x \to a} \left[ \frac{f(x) + g(x)}{x - a} - \frac{f(a) + g(a)}{x - a} \right] = \lim_{x \to a} \left[ \frac{f(x) - f(a) + g(x) - g(a)}{x - a} \right]
\]

\[
= \lim_{x \to a} \left[ \frac{f(x) - f(a)}{x - a} + \frac{g(x) - g(a)}{x - a} \right] = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} + \lim_{x \to a} \frac{g(x) - g(a)}{x - a}.
\]

**V.3.2.4.** Proposition. [Constant Multiple Rule] Let \( f \) be a function from an open interval \( I \) to \( \mathbb{R} \), \( c \in \mathbb{R} \), and \( h = cf : I \to \mathbb{R} \). If \( f \) is differentiable at \( a \in I \), then \( h \) is differentiable at \( a \) and

\[ h'(a) = cf'(a). \]

**Proof:** By (1), we have

\[
\lim_{x \to a} \frac{cf(x) - cf(a)}{x - a} = \lim_{x \to a} \left[ c \cdot \frac{f(x) - f(a)}{x - a} \right] = c \cdot \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.
\]
V.3.2.5. Note that the Sum Rule and the Constant Multiple Rule are not merely formulas. They are (easy) theorems, with hypotheses and a conclusion with a theoretical part (that \( h \) is differentiable at \( a \)) and a formula part (the derivative of \( h \) at \( a \)).

V.3.3. The Product and Quotient Rules

V.3.3.1. It is not true that a derivative of a product of two functions is the product of the derivatives; almost any product where one is able to compute derivatives directly (e.g. a product of two power functions) is a counterexample. In fact, if it were true that the derivative of a product is the product of the derivatives, it would force the derivative of any differentiable function to be zero: if \( f \) is differentiable, we would have

\[
\lim_{x \to 0} \frac{f(x)}{x} = \lim_{x \to 0} \left( \frac{1}{x} \right) = 0.
\]

Actually, the derivative of a composition of two functions is the product of the derivatives; see V.3.4.3. for a precise statement.

V.3.3.2. Instead, the derivative of a product of two differentiable functions is given by a more complicated formula with two terms, in each of which one of the factors is differentiated and the other left unchanged.

V.3.3.3. Theorem. [Product Rule] Let \( f \) and \( g \) be functions from an open interval \( I \) to \( \mathbb{R} \), and \( h = fg : I \to \mathbb{R} \). If \( f \) and \( g \) are differentiable at \( a \in I \), then \( h \) is differentiable at \( a \) and

\[
h'(a) = f'(a)g(a) + f(a)g'(a).
\]

Proof: Write \( u = f(a) \) and \( v = g(a) \), and \( y = h(a) = uv \). If \( \Delta x \neq 0 \), write \( \Delta u = f(a + \Delta x) - f(a) \), \( \Delta v = g(a + \Delta x) - g(a) \), \( \Delta y = h(a + \Delta x) - h(a) = (u + \Delta u)(v + \Delta v) - uv \). Then we have

\[
\frac{\Delta y}{\Delta x} = \frac{(u + \Delta u)(v + \Delta v) - uv}{\Delta x} = \frac{v \cdot \Delta u + u \cdot \Delta v + \Delta u \cdot \Delta v}{\Delta x} + \frac{v \Delta u}{\Delta x} + u \Delta v \Delta x.
\]

By assumption, \( \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \) exists and equals \( f'(a) \), and \( \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} \) exists and equals \( g'(a) \). We also have \( \lim_{\Delta x \to 0} \Delta u = 0 \) since \( f \) is differentiable at \( a \), hence continuous at \( a \) (VIII.3.1.13.). Thus, by (1) and (2), since \( u \) and \( v \) are constants we have that \( h'(a) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \) exists and equals

\[
v \left[ \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \right] + u \left[ \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} \right] + \left[ \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \right] \left[ \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} \right] = f'(a)g(a) + f(a)g'(a) + 0 \cdot g'(a).
\]

532
V.3.3.4. The above proof is a natural one, but requires introduction of some new notation. We can write essentially the same proof more compactly in straightforward notation, but when the proof is done in this way it involves the “trick” of adding and subtracting the same quantity, so does not seem as natural:

**Proof:** We have, for \( x \neq a \),

\[
\frac{h(x) - h(a)}{x - a} = \frac{f(x)g(x) - f(a)g(a)}{x - a} = \frac{[f(x)g(x) - f(a)g(x)] + [f(a)g(x) - f(a)g(a)]}{x - a}
\]

\[
= g(a) \frac{f(x) - f(a)}{x - a} + f(x) \frac{g(x) - g(a)}{x - a}.
\]

We have that \( \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \) exists and equals \( f'(a) \), and that \( \lim_{x \to a} \frac{g(x) - g(a)}{x - a} \) exists and equals \( g'(a) \). Also, \( \lim_{x \to a} f(x) \) exists and equals \( f(a) \) since \( f \) is differentiable at \( a \), hence continuous at \( a \). Thus by (1) and (2) we have that \( h'(a) = \lim_{x \to a} \frac{h(x) - h(a)}{x - a} \) exists and equals

\[
g(a) \left[ \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right] + \left[ \lim_{x \to a} f(x) \right] \left[ \lim_{x \to a} \frac{g(x) - g(a)}{x - a} \right] = g(a)f'(a) + f(a)g'(a).
\]

\[\Box\]

V.3.3.5. **Corollary.** Let \( f \) and \( g \) be differentiable functions on an interval \( I \). Then the product \( fg \) is also differentiable on \( I \).

V.3.3.6. Now we turn to the Quotient Rule. Just as with products, the derivative of a quotient is not the quotient of the derivatives (try taking as denominator the constant function 1!) Again, the Quotient Rule is a theorem with hypotheses and a theoretical part to the conclusion as well as a formula for the derivative of the quotient. The one additional technical complication concerns the domain of the quotient function.

The quotient of the derivatives of numerator and denominator of a quotient of two functions does appear in l’Hôpital’s Rule (1), but there it does not represent the derivative of the quotient.

V.3.3.7. **Theorem.** **[Quotient Rule]** Let \( f \) and \( g \) be functions from an open interval \( I \) to \( \mathbb{R} \). If \( f \) and \( g \) are differentiable at \( a \in I \) and \( g(a) \neq 0 \), then \( h = \frac{f}{g} \) is defined in an interval around \( a \) and differentiable at \( a \), and

\[
h'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{(g(a))^2}.
\]

**Proof:** Since \( g \) is differentiable at \( a \) and hence continuous at \( a \), and \( g(a) \neq 0 \), there is an \( \epsilon > 0 \) such that, whenever \( |x - a| < \epsilon \), \( x \in I \) and \( g(x) \neq 0 \). Thus \( h \) is defined at least on the interval \((a - \epsilon, a + \epsilon)\).

Write \( u = f(a) \) and \( v = g(a) \), and \( y = h(a) = \frac{u}{v} \). If \( 0 < |\Delta x| < \epsilon \), write \( \Delta u = f(a + \Delta x) - f(a) \), \( \Delta v = g(a + \Delta x) - g(a) \) (note that \( v + \Delta v \neq 0 \)), \( \Delta y = h(a + \Delta x) - h(a) = \frac{u + \Delta u}{v + \Delta v} - \frac{u}{v} \). Then, using a common denominator, we have

\[
\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x} \left[ \frac{u + \Delta u}{v + \Delta v} - \frac{u}{v} \right]
\]

533
By assumption, $\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x}$ exists and equals $f'(a)$, and $\lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x}$ exists and equals $g'(a)$. We also have

$$h'(a) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \cdot \frac{1}{v(v + \Delta v)} = \frac{1}{v} \lim_{\Delta x \to 0} \left( v \frac{\Delta u}{\Delta x} - u \frac{\Delta v}{\Delta x} \right).$$

**V.3.3.8.** Corollary. Let $f$ and $g$ be differentiable functions on an interval $I$, with $g(x) \neq 0$ for all $x \in I$. Then $\frac{f}{g}$ is differentiable on $I$.

**V.3.3.9.** Corollary. Every rational function $\left( \frac{p(x)}{q(x)} \right)$ is differentiable on its domain.

**V.3.3.10.** The proof of V.3.3.7. can be rewritten in straightforward notation as in V.3.3.4.; cf. Exercise V.3.10.2. We write this version only in a simplified special case, which can be combined with the Product Rule to give an alternate proof of the Quotient Rule (cf. V.3.10.3):

**V.3.3.11.** Corollary. Let $g$ be a function from an open interval $I$ to $\mathbb{R}$ which is differentiable at $a \in I$, with $g(a) \neq 0$. Then $h = \frac{1}{g}$ is defined in an open interval around $a$ and differentiable at $a$, and

$$h'(a) = -\frac{g'(a)}{|g(a)|^2}.$$  

**Proof:** This is the special case of V.3.3.7. with $f(x) = 1$. We give an alternative proof. Let $\epsilon > 0$ be as in the proof of V.3.3.7.; then $h$ is defined at least on the interval $(a - \epsilon, a + \epsilon)$.

If $0 < |x - a| < \epsilon$, we have

$$\frac{h(x) - h(a)}{x - a} = \frac{1}{g(x)} - \frac{1}{g(a)} = \frac{1}{x - a} \left( \frac{g(a) - g(x)}{g(x)g(a)} \right) = -\frac{1}{g(x)g(a)} \cdot \frac{g(x) - g(a)}{x - a}.\]$$

Since $\lim_{x \to a} \frac{g(x) - g(a)}{x - a}$ exists and equals $g'(a)$, and $\lim_{x \to a} g(x) = g(a) \neq 0$ since $g$ is continuous at $a$, we have that $h'(a) = \lim_{x \to a} \frac{h(x) - h(a)}{x - a}$ exists and equals

$$-\left[ \lim_{x \to a} \frac{1}{g(x)g(a)} \right] \left[ \lim_{x \to a} \frac{g(x) - g(a)}{x - a} \right] = -\frac{1}{|g(a)|^2} g'(a).$$
V.3.3.12. Corollary. [Power Rule, Version 2] Let $n$ be a negative integer. Then the function $f(x) = x^n$ is differentiable on $\mathbb{R} \setminus \{0\}$, with derivative

$$f'(x) = nx^{n-1}.$$  

Proof: Let $m = -n \in \mathbb{N}$, and let $g(x) = x^m$. Then $f(x) = \frac{1}{g(x)}$, so $f$ is differentiable wherever $g$ is nonzero and

$$f'(x) = -\frac{g'(x)}{(g(x))^2} = -\frac{mx^{m-1}}{|x^m|^2} = -m \cdot \frac{x^{m-1}}{x^{2m}} = -mx^{m-1-2m} = nx^{n-1}.$$  

V.3.3.13. Caution: The derivative of a power function is a multiple of a power function with exponent one smaller than the exponent in the original. But if the original exponent is negative, the exponent in the derivative is more negative. Thus, for example,

$$\frac{d}{dx}(x^5) = 5x^4$$

but

$$\frac{d}{dx}(x^{-5}) = -5x^{-6}.$$  

V.3.4. The Chain Rule

The Chain Rule is a result about differentiation of composite functions. As with the Product Rule and Quotient Rule, the Chain Rule is a theorem, not simply a formula; it has a theoretical part (that a composition of differentiable functions is differentiable) and a computational part (a formula for the derivative of the composite function).

V.3.4.1. The motivation and general idea of the Chain Rule is that if $y$ is indirectly a function of $x$ via an intermediate variable $u$, i.e. $u = f(x)$ and $y = g(u)$ for some functions $f$ and $g$, and $f$ and $g$ are differentiable, we want to describe the rate of change $\frac{dy}{dx}$ in terms of $\frac{du}{dx} = g'(u)$ and $\frac{du}{dx} = f'(x)$. Roughly, if $x$ is changed by a small amount $\Delta x$, then $u$ is changed by an approximately proportional amount with constant of proportionality $f'(x)$, i.e. $\Delta u \approx f'(x) \cdot \Delta x$. Similarly, $\Delta y \approx g'(u) \cdot \Delta u$. So

$$\Delta y \approx g'(u) \cdot \Delta u \approx g'(u) f'(x) \cdot \Delta x.$$  

Thus we should have that $y = g(f(x))$ is a differentiable function of $x$ and

$$\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}.$$  

This formula is easily remembered by pretending that $\frac{du}{dx}$ and $\frac{du}{dx}$ are fractions (which they are not!) and “dividing out” the $du$’s. This is a prime example of the utility of the Leibniz notation for derivatives, although this notation is somewhat imprecise as to exactly where the relevant functions are evaluated.
If $f$ and $g$ are linear functions, this all works out properly and quite simply. If $u = f(x) = mx+b$ and $y = g(u) = nu + c$, then $f'(x) = \frac{du}{dx}$ and $g'(u) = \frac{dy}{du}$ are the constant functions $m$ and $n$ respectively. If $h = g \circ f$, then $y = h(x) = g(f(x))$ is also a linear function with formula

\[ y = h(x) = n f(x) + c = n(mx + b) + c = nm x + (nb + c) \]

and we have

\[ \frac{dy}{dx} = h'(x) = nm = \frac{dy}{du} \frac{du}{dx} . \]

Thus the slope of a composition of two linear functions is the product of the slopes of the component functions.

The general Chain Rule asserts that if $f$ and $g$ are “approximately linear” (i.e. differentiable), then $g \circ f$ is also approximately linear, with slope equal to the product of the slopes of $f$ and $g$. But one must be careful about where the derivatives are evaluated since they are not constant. Here is the precise statement.

**Theorem.** [Chain Rule] Let $f$ be defined on a neighborhood $U$ of $a \in \mathbb{R}$ and differentiable at $a$, with $f(a) = b \in \mathbb{R}$, and let $g$ be defined on a neighborhood $V$ of $b$ and differentiable at $b$, with $f(U) \subseteq V$. Set $h = g \circ f$ on $U$. Then $h$ is differentiable at $a$, and

\[ h'(a) = (g \circ f)'(a) = g'(b) f'(a) = g'(f(a)) f'(a) . \]

A straightforward and essentially correct proof can be given which contains a subtle fallacy; see Exercise V.3.10.5. A fully correct proof is not much harder if done right, but requires a slightly modified approach which is not so obvious. We first define a new function we will use in the proof:

**Definition.** Let $g$ and $b$ be as in V.3.4.3. Define a function $G_b$ on $V$ by

\[ G_b(u) = \begin{cases} g(u) - g(b) & \text{if } u \neq b \\ \frac{g(u) - g(b)}{u - b} & \text{if } u = b \end{cases} . \]

**Lemma.** $G_b$ is continuous at $b$.

**Proof:** We need to check that $G_b(b) = \lim_{u \to b} G_b(u)$. But this is true by definition of $g'(b)$. \(\Box\)

536
V.3.4.6. Lemma. Using the notation of V.3.4.3., if \( x \in \mathcal{U}, x \neq a \), we have

\[
\frac{h(x) - h(a)}{x - a} = G_b(f(x)) \cdot \frac{f(x) - f(a)}{x - a}.
\]

Proof: If \( f(x) \neq b = f(a) \), we have

\[
G_b(f(x)) \cdot \frac{f(x) - f(a)}{x - a} = \frac{g(f(x)) - g(b)}{f(x) - b} \cdot \frac{f(x) - f(a)}{x - a}
\]

and, since \( b = f(a) \), \( h(x) = g(f(x)) \), and \( h(a) = g(b) \), the formula holds. If \( f(x) = b \), then both sides of the equation are zero, so the formula again holds.

We now prove V.3.4.3.

Proof: Since \( f \) is continuous at \( a \) by () and \( G_b \) is continuous at \( b = f(a) \), \( G_b \circ f \) is continuous at \( a \) by () and

\[
\lim_{x \to a} G_b(f(x)) = G_b(f(a)) = G_b(b) = g'(b).
\]

Also, we have

\[
\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a)
\]

So by (), we have

\[
\lim_{x \to a} \frac{h(x) - h(a)}{x - a} = \left[ \lim_{x \to a} G_b(f(x)) \right] \left[ \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right] = g'(b)f'(a).
\]

Here are some of the most important standard applications of the Chain Rule:

V.3.4.7. Corollary. [Chain Rule for Powers] Let \( f \) be differentiable at \( a \), \( f(a) > 0 \), and \( r \in \mathbb{R} \). Then \( h(x) = [f(x)]^r \) is differentiable at \( a \), and \( h'(a) = r[f(a)]^{r-1}f'(a) \). Alternatively,

\[
\frac{d}{dx} ([f(x)]^r) = r[f(x)]^{r-1}f'(x).
\]

Proof: Apply V.3.4.3. with \( g(u) = u^r \).

V.3.4.8. Corollary. [Chain Rule for Exponentials] Let \( f \) be differentiable at \( a \). Then \( h(x) = e^{f(x)} \) is differentiable at \( a \), and \( h'(a) = e^{f(a)}f'(a) \). Alternatively,

\[
\frac{d}{dx} (e^{f(x)}) = f'(x)e^{f(x)}.
\]
V.3.4.9. Corollary. [Chain Rule for Logarithms] Let \( f \) be differentiable at \( a \), with \( f(a) \neq 0 \). Then \( h(x) = \log |f(x)| \) is differentiable at \( a \), and \( h'(a) = \frac{1}{f(a)} f'(a) \). Alternatively,

\[
\frac{d}{dx} (\log |f(x)|) = \frac{f'(x)}{f(x)}.
\]

The proofs of V.3.4.8. and V.3.4.9. are similar to the proof of V.3.4.7., and are left to the reader.

V.3.5. Derivatives of Inverse Functions

V.3.6. Critical Points and Local Extrema

One of the most important elementary applications of the derivative is the location and analysis of local extrema.

V.3.6.1. Definition. Let \( f \) be a function defined on an interval \( I \), and \( a \) an interior point of \( I \). The function \( f \) has a local maximum, or relative maximum, at \( a \) if there is a \( \delta > 0 \) such that \( f(x) \leq f(a) \) for all \( x \) with \( |x - a| < \delta \). The function \( f \) has a local minimum, or relative minimum, at \( a \) if there is a \( \delta > 0 \) such that \( f(x) \geq f(a) \) for all \( x \) with \( |x - a| < \delta \). A local extremum, or relative extremum, is a local maximum or local minimum.

V.3.6.2. A continuous function \( f \) on a closed bounded interval attains a maximum and minimum on the interval \( I \). If the maximum is attained at an interior point \( a \) of the interval, then \( f \) has a local maximum at \( a \). The converse is false: if \( f \) has a local maximum at \( a \), its value there may not be the maximum value of \( f \) on the entire interval (it is only the maximum on some sufficiently small subinterval around \( a \)). Indeed, \( f \) may have two or more local maxima in the interval, where it may or may not attain the maximum on the whole interval, and the maximum on the interval can occur at an endpoint and not at a local maximum; in fact, \( f \) need not have any local maxima. Corresponding statements hold for local minima.

The crucial relation between local extrema and the derivative is:

V.3.6.3. Theorem. Let \( f \) be a function defined on an interval \( I \), and \( a \) an interior point of \( I \). If \( f \) has a local extremum at \( a \), and \( f \) is differentiable at \( a \), then \( f'(a) = 0 \).

Proof: We prove the statement for a local maximum; the proof for a local minimum is identical with a few inequalities reversed. Fix \( \delta > 0 \) such that \( f(x) \leq f(a) \) for \( |x - a| < \delta \). If \( a < x < a + \delta \), then we have

\[
\frac{f(x) - f(a)}{x - a} \leq 0
\]

and thus by ()

\[
\lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \leq 0
\]
(if the limit exists). Similarly, If \( a - \delta < x < a \), then we have
\[
\frac{f(x) - f(a)}{x - a} \geq 0
\]
since numerator and denominator are both negative (nonpositive), and thus
\[
\lim_{x \to a^-} \frac{f(x) - f(a)}{x - a} \geq 0
\]
(assuming the limit exists). So, if \( f \) is differentiable at \( a \), both limits exist and are equal, hence 0; thus
\[
f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a^-} \frac{f(x) - f(a)}{x - a} = 0 .
\]

V.3.6.4. A continuous function need not be differentiable at a local extremum: consider \( f(x) = |x| \) at 0 (cf. V.4.2.7.). And if \( f \) is differentiable at \( a \) and \( f'(a) = 0 \), then \( f \) need not have a local extremum at \( a \): consider \( f(x) = x^3 \) at 0. However, the places where \( f'(a) = 0 \) (or does not exist) are the candidates for where \( f \) can have local extrema. Thus we make the following definition:

V.3.6.5. DEFINITION. Let \( f \) be a function defined on an interval \( I \), and \( a \) an interior point of \( I \). Then \( a \) is a critical number for \( f \) if either \( f \) is not differentiable at \( a \), or \( f \) is differentiable at \( a \) and \( f'(a) = 0 \).

V.3.6.6. Thus every local extremum is at a critical number, but not conversely in general. The term critical point is often used as a synonym for critical number, but “critical point” is also sometimes used to refer to the corresponding point on the graph, so “critical number” is less ambiguous. Technically, the terms local maximum, etc., also refer to points on the graph; thus we say “\( f \) has a local extremum at \( a \)” rather than “\( a \) is a local extremum of \( f \).”

Increasing and Decreasing Functions at a Point

There is a sharper form of the contrapositive to V.3.6.3. We first make a definition:

V.3.6.7. DEFINITION. Let \( f \) be a function defined on an interval \( I \), and \( a \) an interior point of \( I \). Then \( f \) is increasing at \( a \) if there is a \( \delta > 0 \) such that \( f(x) < f(a) \) for \( a - \delta < x < a \) and \( f(x) > f(a) \) for \( a < x < a + \delta \). Similarly, \( f \) is decreasing at \( a \) if there is a \( \delta > 0 \) such that \( f(x) > f(a) \) for \( a - \delta < x < a \) and \( f(x) < f(a) \) for \( a < x < a + \delta \).

Note that this is strictly a pointwise definition: we only compare \( f(x) \) with \( f(a) \) for \( x \) near \( a \), we do not compare \( f(x_1) \) and \( f(x_2) \) for \( x_1, x_2 \) near \( a \), which is a much more problematic matter (cf. VIII.6.6.2.).
V.3.6.8. COROLLARY. Let \( f \) be a function defined on an interval \( I \), and \( a \) an interior point of \( I \). Suppose \( f \) is differentiable at \( a \). Then

(i) If \( f'(a) > 0 \), then \( f \) is increasing at \( a \).

(ii) If \( f'(a) < 0 \), then \( f \) is decreasing at \( a \).

Proof: The proof of each part is essentially a piece of the proof of V.3.6.3. For (i), suppose \( f \) is not increasing at \( a \). Then for any \( \varepsilon > 0 \) there is an \( x \) with \( 0 < |x - a| < \delta \) and

\[
\frac{f(x) - f(a)}{x - a} \leq 0.
\]

Thus

\[
f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \leq 0.
\]

The proof of (ii) is the same with inequalities reversed.

V.3.6.9. Thus, if \( f \) is differentiable on an interval \( I \) and \( f'(x) > 0 \) for all \( x \in I \), then \( f \) is increasing at every point of \( I \). It stands to reason that \( f \) should actually be strictly increasing on \( I \), i.e. that \( f(x_1) < f(x_2) \) if \( x_1, x_2 \in I, \ x_1 < x_2 \). This turns out to be true, but is subtle to prove (V.8.2.1., V.8.6.12.). A symmetric result holds if \( f' < 0 \) on \( I \).

Endpoint Extrema

We can analyze endpoint behavior in the same way.

V.3.6.10. DEFINITION. (i) Let \( f \) be a function defined on an interval with left endpoint \( a \). Then \( f \) has an endpoint maximum at \( a \) if there is a \( \delta > 0 \) such that \( f(x) \leq f(a) \) for all \( x \) with \( a < x < a + \delta \). Similarly, \( f \) has an endpoint minimum at \( a \) if there is a \( \delta > 0 \) such that \( f(x) \geq f(a) \) for \( a < x < a + \delta \).

(ii) Let \( f \) be a function defined on an interval with right endpoint \( b \). Then \( f \) has an endpoint maximum at \( b \) if there is a \( \delta > 0 \) such that \( f(x) \leq f(b) \) for all \( x \) with \( b - \delta < x < b \). Similarly, \( f \) has an endpoint minimum at \( b \) if there is a \( \delta > 0 \) such that \( f(x) \geq f(b) \) for \( b - \delta < x < b \).

An endpoint extremum is an endpoint maximum or endpoint minimum.

V.3.6.11. COROLLARY. Let \( f \) be a function on an interval \( I \), and \([a, b]\) a closed bounded interval in the interior of \( I \). Then

(i) If \( f \) has an endpoint maximum at \( a \) and \( f \) is differentiable at \( a \), then \( f'(a) \leq 0 \).

(ii) If \( f \) has an endpoint minimum at \( a \) and \( f \) is differentiable at \( a \), then \( f'(a) \geq 0 \).

(iii) If \( f \) has an endpoint maximum at \( b \) and \( f \) is differentiable at \( b \), then \( f'(b) \geq 0 \).

(iv) If \( f \) has an endpoint minimum at \( b \) and \( f \) is differentiable at \( b \), then \( f'(b) \leq 0 \).

The proof is again pieces of the proof of V.3.6.3.
V.3.7. Higher Derivatives

V.3.7.1. If $f$ is a real-valued function defined on a subset $A$ of $\mathbb{R}$, then the derivative $f'$ is another function defined on a subset (proper subset in general) of $A$. Thus $f'$ itself has a derivative on a subset of its domain, called the second derivative of $f$. The process can be repeated to define the third derivative and higher derivatives of all orders.

V.3.7.2. Definition. Let $A \subseteq \mathbb{R}$, $a \in A$, and $f : A \to \mathbb{R}$ a function. If $f'$ is differentiable at $a$, then $f$ is second-order differentiable, or twice differentiable, at $a$ and

$$f''(a) = (f')(a) = \lim_{x \to a} \frac{f'(x) - f'(a)}{x - a}$$

is the second derivative of $f$ at $a$, also denoted $f^{(2)}(a)$. (Also write $f^{(0)} = f$ and $f^{(1)} = f'$; $f'$ is called the first derivative of $f$.)

Inductively, if $n \in \mathbb{N}$ and $f^{(n-1)}$ is differentiable at $a$, then $f$ is differentiable to order $n$ at $a$, or $n$-times differentiable at $a$, and $f^{(n)}(a) = (f^{(n-1)})'(a)$ is the $n$'th derivative of $f$ at $a$.

If $I$ is an interval contained in $A$, then $f$ is differentiable to order $n$ on $I$, or $n$-times differentiable on $I$, if $f^{(n)}(a)$ is differentiable at all $a \in I$, i.e. $f^{(n-1)}$ is differentiable on $I$. If $f$ is $n$'th-order differentiable on $I$ and $f^{(n)}$ is continuous on $I$, then $f$ is $C^n$ on $I$ (continuously $n$'th order differentiable on $I$).

V.3.7.3. Note that for $f^{(n)}(a)$ to be defined, it is necessary for $f^{(n-1)}$ to be defined in an entire interval around $a$ and continuous at $a$ (VIII.3.1.13.). Thus if $f$ is $n$'th order differentiable on $I$, then it is $C^{n-1}$ on $I$ (the converse is false in general). Note that “$f$ is $C^n$ on $I$” just means “$f$ is continuous on $I$.”

V.3.7.4. Notation: We can use additional primes to denote higher derivatives: $f''$ for the third derivative $f^{(3)}$, $f'''$ for $f^{(4)}$, etc. Three seems to be the conventional dividing line: the second derivative is usually denoted $f''$, the third derivative commonly denoted either $f'''$ or $f^{(3)}$, and only rarely are primes used for $f^{(n)}$, $n > 3$.

Leibniz notation is also commonly used for higher derivatives: we write

$$\frac{d^2 f}{dx^2} = \frac{d^2 f}{dx^2}(x) = \frac{d^2 f}{dx^2}(f(x)) = \frac{d}{dx} \left( \frac{d}{dx} (f(x)) \right) = \left( \frac{d}{dx} \right)^2 (f(x)) = f''(x)$$

and more generally

$$\frac{d^n f}{dx^n} = \frac{d^n f}{dx^n}(x) = f^{(n)}(x)$$

(note the position of the exponents: we write $d^2 f$, not $\frac{d^2 f}{dx^2}$ or $\frac{d^2 f}{dx^2}$; think of $\frac{d^n f}{dx^n} = \left( \frac{d}{dx} \right)^n$ as an operator).

V.3.7.5. Second derivatives are especially important in physics: for example, acceleration is the second derivative of position. Higher derivatives beyond the second are not so commonly used in applications.

Geometric Interpretation

V.3.7.6. All the higher derivatives of a function $f$ have geometric interpretations in the shape of the graph of $f$, but the interpretations of $f^{(n)}$ are increasingly subtle as $n$ increases and are only easily described for $n = 2$ (V.8.3.).
The Formula of Faà di Bruno

V.3.7.7. Under the hypotheses of the Chain Rule V.3.4.3., if \( f \) and \( g \) are differentiable to order \( n \) at \( a \), it follows inductively from the Product Rule and Chain Rule that the composite function \( g \circ f \) is also differentiable to \( n \)-th order at \( a \). If \( n = 2 \), we have

\[
(g \circ f)''(a) = g'(f(a))f''(a) + g''(f(a))[f'(a)]^2
\]

and if \( n = 3 \),

\[
(g \circ f)'''(a) = g'(f(a))f'''(a) + 3g''(f(a))f'(a)f''(a) + g'''(f(a))[f'(a)]^3.
\]

What is the general formula for \( (g \circ f)^{(n)} \)? It is a nontrivial exercise to even find a general expression, much less prove it. The most commonly written version of the formula is:

\[
h^{(n)}(a) = (g \circ f)^{(n)}(a) = \sum \frac{n!}{k_1! \cdots k_n!} g^{(k)}(b) \left( \frac{f'(a)}{k_1!} \right)^{k_1} \cdots \left( \frac{f^{(n)}(a)}{k_n!} \right)^{k_n}
\]

where the sum is over all sets of nonnegative integers \( k_1, \ldots, k_n \) with \( k_1 + 2k_2 + \cdots nk_n = n \), and \( k = k_1 + k_2 + \cdots + k_n \).

The formula is named after Francesco Faà di Bruno, a respected Italian mathematician of the mid-nineteenth century; however, the attribution is not historically accurate since he was not the first to prove or publish it. The formula is essentially combinatorial, and there are many forms and variations, arising in various contexts throughout mathematics. In principle it can be proved by induction, but this is very difficult in practice, and the standard proofs generally employ power series, Bell polynomials, or umbral calculus instead. The variations, proofs, and the complicated history of the formula are discussed in detail in [Joh02], which contains an extensive list of references.

V.3.8. Continuous but Nondifferentiable Functions

It came as quite a shock to many mathematicians when in ( \( \text{Weierstrass} \) gave an example of a function which is continuous everywhere but differentiable nowhere. There have since been many other explicit examples given (sometimes essentially the same examples have been rediscovered and republished), and it is now known (XI.12.2.5.) that “most” continuous functions are not differentiable anywhere, quite contrary to our standard naive mental picture of a “typical” continuous function.

Van der Waerden’s Example

V.3.8.1. Example. Here is one of the simplest examples of a continuous function which is not differentiable anywhere, due to B. L. Van der Waerden [vdW30] (cf. [? , ]). (The first example of this sort was given by T. Takagi in 1903 [?], and other similar examples were constructed by several other mathematicians.)
For each $k \in \mathbb{N} \cup \{0\}$, define $f_k : \mathbb{R} \to \mathbb{R}$ by letting $f_k(x)$ be the distance from $x$ to the closest decadent rational number of the form $\frac{m}{10^k}$ for $m \in \mathbb{Z}$. See Figure V.5.

It is clear that $f_k$ is continuous, periodic with period $\frac{1}{10^k}$, and that

$$0 \leq f_k(x) \leq \frac{1}{2 \cdot 10^k}$$

for all $x$. Then set

$$f(x) = \sum_{k=0}^{\infty} f_k(x).$$

Since the series is uniformly convergent, $f$ is continuous on all of $\mathbb{R}$. See Fig. V.6.

We claim that $f$ is not differentiable anywhere. Fix $x \in \mathbb{R}$, and write

$$x = r.d_1 d_2 d_3 \cdots$$

in a decimal expansion. (If $x < 0$, take the decimal expansion of $x + s$ for some $s \in \mathbb{N}$, $s > |x|$ instead. The $d_n$ are independent of $s$. There is the usual ambiguity if $x$ is a decadent rational; in this case, either decimal expansion will do.) We will show that $f$ is not differentiable at $x$. 

543
For each \( n \in \mathbb{N} \) set \( x_n = x - \frac{1}{10^n} \) if \( d_n \) is 4 or 9, and \( x_n = x + \frac{1}{10^n} \) otherwise. Then \( x_n \to x \), so if \( f \) is differentiable at \( x \) we must have
\[
\lim_{n \to \infty} \frac{f(x_n) - f(x)}{x_n - x} = f'(x) \in \mathbb{R}
\]
and thus the limit must exist (in the ordinary sense).

However, if \( k < n \), we have that the closest number \( y \) of the form \( \frac{m}{10^n} \) to \( x_n \) is also the closest one to \( x \), and \( x \) and \( x_n \) are on the same side of \( y \), so we have
\[
|f_k(x_n) - f_k(x)| = |(x_n - y) - (x - y)| = |x_n - x|.
\]
On the other hand, if \( k \geq n \), then \( \frac{1}{10^n} = |x_n - x| \) is an integer multiple of \( \frac{1}{10^n} \) and \( f_k(x_n) - f_k(x) = 0 \) by periodicity of \( f_k \). We thus have
\[
f(x_n) - f(x) = \sum_{k=0}^{n-1} |f_k(x_n) - f_k(x)| = p_n(x_n - x)
\]
where \( p_n \) is an integer of the same parity (odd or even) as \( n \). Thus
\[
\frac{f(x_n) - f(x)}{x_n - x} = p_n
\]
cannot have a limit in \( \mathbb{R} \) (i.e. be eventually constant) as \( n \to \infty \).

V.3.8.2. Unlike Weierstrass’s example, for this example there are uncountably many points where some or all of the Dini derivatives are finite (cf. Exercise V.3.10.13.), although such points form a set of measure zero by the Denjoy-Young-Saks Theorem.

V.3.9. The Bush-Wunderlich-Swift Example

We now describe a particularly simple example of a continuous nondifferentiable function. The example we give is from [Swi61], a special case of an example from [Bus52] and [Wun52].

V.3.9.1. Let \( x \) be a real number. Write the ternary (base 3) decimal expansion of the fractional part of \( x \) as
\[.t_1t_2t_3\cdots\]
where each \( t_k \) is 0, 1, or 2 (if \( x \) is a triadic rational number, there will be two such expansions, one ending in a string of 0’s and one in a string of 2’s). Define \( f(x) \) to be the number in the unit interval with binary expansion
\[.d_1d_2d_3\cdots\]
\((d_k = 0 \text{ or } 1)\) defined by the following inductive rule: \( d_1 = 1 \) if and only if \( t_1 = 1 \), and \( d_{n+1} = d_n \) if and only if \( t_{n+1} = t_n \).
V.3.9.2. **Proposition.** The function $f$ is well defined.

**Proof:** The definition of $f$ is via a definite procedure which gives a well-defined binary decimal for every ternary decimal. Thus we need only show that if $x$ is a ternary rational, the two ternary decimal representations of $x$ give binary decimals representing the same number. This is completely elementary, but a little ugly. Note that if a triadic expansion ends in a constant sequence, the corresponding dyadic expansion also ends in a constant sequence. Denote by $t_1t_2\cdots$ the ternary expansion of $x$ ending in 2's, and $t'_1t'_2\cdots$ the ternary expansion ending in 0's, and $d_1d_2\cdots$ and $d'_1d'_2\cdots$ the corresponding dyadic expansions. There are six cases to check. The first is

$$t_1t_2\cdots = \cdots 00222\cdots$$

$$t'_1t'_2\cdots = \cdots 0100\cdots$$

where the omitted initial segments are equal, and the displayed segments start with $t_n$ and $t'_n$. Then either $d_n = d'_n = 0$, in which case

$$d_1d_2\cdots = \cdots 00111\cdots$$

$$d'_1d'_2\cdots = \cdots 0100\cdots$$

with the omitted initial segments equal, or $d_n = d'_n = 1$, in which case

$$d_1d_2\cdots = \cdots 1100\cdots$$

$$d'_1d'_2\cdots = \cdots 1011\cdots$$

and in either case the dyadic rational numbers represented are equal. The other five cases where $t_1t_2\cdots$ is $\cdots 10222\cdots$, $\cdots 0222\cdots$, $\cdots 01222\cdots$, $\cdots 1222\cdots$, or $\cdots 21222\cdots$ can be checked similarly. The cases $0222\cdots$ and $1222\cdots$ must be checked separately. 

V.3.9.3. **Proposition.** The function $f$ is continuous.

**Proof:** Recall () that if $x_n, x \in \mathbb{R}$, then $x_n \to x$ if and only if, for each $k$ there is an $N$ such that, for all $n \geq N$, the triadic expansion of $x_n$ (at least one expansion if $x_n$ is triadic rational) agrees with the triadic expansion of $x$ (at least one expansion if $x$ is triadic rational) at least up to the $k$‘th decimal place, and similarly for dyadic expansions. Since first $k$ decimal places of the dyadic expansion of $f(x)$ are completely determined by the first $k$ decimal places of the triadic expansion of $x$, it follows that if $x_n \to x$, then $f(x_n) \to f(x)$. 

V.3.9.4. **Proposition.** Let $x \in \mathbb{R}$. Then either $D^+f(x) = +\infty$ or $D_+f(x) = -\infty$, and either $D^-f(x) = +\infty$ or $D_-f(x) = -\infty$. In particular, $f$ is not differentiable anywhere.

**Proof:** Let $t_1t_2\cdots$ be the ternary expansion for the fractional part of $x$ not ending in a string of 2’s, and let $d_1d_2\cdots$ be the corresponding expansion of $f(x)$. There are infinitely many $n$ such that $t_{n+1} = t_n$ and $t_{n+1} \neq 2$. Fix such an $n$. Let $t'_{n+1} = 2$ if $t_{n+1} = t_n < 2$ and $t'_{n+1} = t_n$ if $t_{n+1} < t_n$. Then successively choose $t'_k$ for $k > n + 1$ so that the number $y$ with integer part the same as $x$ and fractional part $\cdots t_1t_2t_{n+1}t'_{n+1}t'_{n+2}\cdots$
has image

\[ f(y) = d_1 \cdots d_n d'_{n+1} d_{n+2} \cdots \]
i.e. \( d'_{n+1} \neq d_{n+1} \) but \( d'_k = d_k \) for all \( k \neq n + 1 \). We then have

\[ 0 < y - x \leq 3^{-n} \]
\[ |f(y) - f(x)| = 2^{-(n+1)} \]

so

\[ \left| \frac{f(y) - f(x)}{y - x} \right| \geq \frac{1}{2} \left( \frac{3}{2} \right)^n . \]

Since there are infinitely many such \( n \), the result for \( D^+ f(x) \) and \( D^- f(x) \) follows. The proof for \( D^- f(x) \) and \( D^- f(x) \) is similar.

V.3.9.5. The function takes values in \([0, 1]\), and takes the value 0 at the integers; thus, if \( x \) is an integer, we must have \( D^- f(x) \leq 0 \) and \( D_+ f(x) \geq 0 \), so \( D^+ f(x) = +\infty \) and \( D^- f(x) = -\infty \). Similarly, \( f(1/2) = 1 \), so if \( x \) is \( 1/2 \) or an integer translate, then \( D_+ f(x) \geq 0 \) and \( D^+ f(x) \leq 0 \), so \( D^- f(x) = +\infty \) and \( D_+ f(x) = -\infty \).

Continuous Functions Differentiable at Only One Point

V.3.9.6. Example. Let \( f \) be a continuous function from \( \mathbb{R} \) to \( \mathbb{R} \) which is not differentiable anywhere, e.g. one of the above examples. Set \( g(x) = x^2 f(x) \). Then \( g \) is a continuous function from \( \mathbb{R} \) to \( \mathbb{R} \). We claim that \( g \) is differentiable at 0, and only at 0.

To see that \( g \) is differentiable at 0, note that \( f \) is bounded on the interval \([-1, 1]\) since it is continuous there; say \( |f(x)| \leq M \) for \( |x| \leq 1 \). Then, for \( 0 < |x| \leq 1 \), we have

\[ \left| \frac{g(x) - g(0)}{x - 0} \right| = \left| \frac{x^2 f(x)}{x} \right| \leq M|x| \]

and thus \( g'(0) \) exists and equals 0 by the Squeeze Theorem.

But \( g \) is not differentiable anywhere except at 0, since if it were \( f(x) = \frac{g(x)}{x^2} \) would also be differentiable there by the Quotient Rule.
Figure V.6: Maple plot of graph of $f$
Figure V.7: Maple plot of $f(x) = \sum_{k=1}^{\infty} 2^{-k} \cos(4^k x)$
V.3.10. Exercises

V.3.10.1. Write out the proof of the power rule by writing, for \( f(x) = x^n \) (\( n \in \mathbb{N} \)) and \( a \in \mathbb{R} \),

\[
f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h} = \lim_{h \to 0} \frac{(a + h)^n - a^n}{h}
\]

and expanding \((a + h)^n\) using the Binomial Theorem (V.17.7.3).

V.3.10.2. Write the proof of the full Quotient Rule along the lines of the proofs of V.3.3.4 and V.3.3.11.

V.3.10.3. Here is an alternate derivation of the quotient formula and “proof” of the Quotient Rule which has appeared in some textbooks.

Let \( f \) and \( g \) be defined in a neighborhood of \( a \) and differentiable at \( a \), with \( g(a) \neq 0 \).

(i) Set \( h = \frac{f}{g} \). Then, since \( g \) is continuous at \( a \) and \( g(a) \neq 0 \), \( h \) is defined in a neighborhood of \( a \), and \( f = gh \) on this neighborhood.

(ii) By the Product Rule, \( f'(a) = g'(a)h(a) + g(a)h'(a) \).

(iii) Solve this equation for \( h'(a) \) to obtain the Quotient Rule formula.

(iv) What is wrong with this argument as a proof of the Quotient Rule? What does this argument actually prove?

V.3.10.4. (a) Show by direct calculation that if \( a \neq 0 \), then \( f(x) = \frac{1}{x} \) is differentiable at \( a \) and \( f'(a) = \frac{-1}{a^2} \).

(b) Use (a) and the Chain Rule to give an alternate proof of V.3.3.11.

(c) Conclude that the Quotient Rule is a consequence of (a), the Product Rule, and the Chain Rule.

V.3.10.5. What is wrong with the following “proof” of the Chain Rule? [Hint: the problem is very similar to the problem in ().]

Let \( f \) be defined on a neighborhood \( U \) of \( a \) and differentiable at \( a \), with \( f(a) = b \), and let \( g \) be defined on a neighborhood \( V \) of \( b \) and differentiable at \( b \), with \( f(U) \subseteq V \). Set \( h = g \circ f \) on \( U \).

(i) For \( x \in U \), \( x \neq a \), write

\[
\frac{h(x) - h(a)}{x - a} = \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \cdot \frac{f(x) - f(a)}{x - a}
\]

(ii) For \( x \in U \), write \( u = f(x) \). Since \( \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \) exists and \( \lim_{x \to a} (x - a) = 0 \), we have \( \lim_{x \to a} (u - b) = 0 \).

(iii) Combining (i) and (ii), we have

\[
\lim_{x \to a} \frac{h(x) - h(a)}{x - a} = \left[ \lim_{u \to b} \frac{g(u) - g(b)}{u - b} \right] \left[ \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right] = g'(b)f'(a).
\]
V.3.10.6. What is wrong with the following “proof” of the Inverse Function Rule? What does this argument actually prove?

Let \( f \) be continuous and strictly monotone in a neighborhood of \( a \), and differentiable at \( a \). If \( f(a) = b \), then \( g = f^{-1} \) is defined, continuous, and strictly monotone in a neighborhood of \( b \) by (i).

(i) Set \( h = g \circ f = f^{-1} \circ f \). Then \( h(x) = x \) in a neighborhood of \( a \). Thus \( h \) is differentiable at \( a \), and \( h'(a) = 1 \).

(ii) Using the Chain Rule, differentiate \( h \) at \( a \) to obtain

\[
1 = h'(a) = g'(f(a))f'(a) = (f^{-1})'(b)f'(a) .
\]

(iii) If \( f'(a) \neq 0 \), divide both ends of the equation by \( f'(a) \) to obtain

\[
(f^{-1})'(b) = \frac{1}{f'(a)} .
\]

V.3.10.7. Prove the formula \( \frac{d}{dx} ([f(x)]^n) = n[f(x)]^{n-1}f'(x) \) for \( n \in \mathbb{N} \) by induction using the Product Rule instead of the Chain Rule.

V.3.10.8. (S. Chang\(^2\)) Let \( f(x) = \sum_{k=1}^{100} \frac{\sin(2\pi k^2 x)}{4\pi^2 k^3} + x^2 \). Note that \( f \) is \( C^\infty \) (in fact, analytic) on \( \mathbb{R} \).

Use a computer algebra system to plot \( f \), \( f' \), and \( f'' \). Could you have predicted the behavior of \( f'' \) from the appearance of the graph of \( f \), or even the graph of \( f' \)?

V.3.10.9. Formally define the “derivative” of a polynomial \( p(x) = \sum_{k=0}^{n} a_k x^k \) to be

\[
p#(x) = \sum_{k=1}^{n} k a_k x^{k-1} .
\]

Give a purely algebraic proof that \( p \rightarrow p# \) is linear and satisfies the “Product Rule”

\[
(pq)# = p#q + pq#
\]

for polynomials \( p \) and \( q \). (This argument is valid over any field or even any ring.)

V.3.10.10. Let \( f(x) = 2x^2 + x^2 \sin \frac{1}{x^2} \) for \( x \neq 0 \), \( f(0) = 0 \).

(a) Show that \( f \) is differentiable everywhere.

(b) Show that \( f \) has a strict local minimum (actually absolute minimum) at 0.

(c) Show that the first derivative test fails at 0.

V.3.10.11. There is a unique differentiable function \( f \) with \( f'(x) = \frac{1}{x} x + x^2 \sin \frac{1}{x^2} \) for \( x \neq 0 \) and \( f(0) = f'(0) = 0 \) (V.8.5.9.).

(a) Show that by the first derivative test, \( f \) has a local minimum at 0 (which is in fact a strict absolute minimum).

(b) In what sense is the graph of \( f \) concave upward around 0?

Let $f$ and $g$ be differentiable on an open interval $I$, and suppose that $f'g = fg'$ on $I$.

(a) Can we conclude from the quotient rule applied to $\frac{f}{g}$ that $\frac{f}{g}$ is constant on $I$, i.e. that $f$ is a constant multiple of $g$ on $I$?

(b) Consider the case where $I = \mathbb{R}$, $g(x) = x^2$ for all $x$, $f(x) = x^2$ if $x \leq 0$, and $f(x) = 2x^2$ for $x > 0$.

(c) Suppose we assume $f$ and $g$ are $C^\infty$. [Consider Cauchy’s examples V.17.2.3., especially the last two and their reflections.]

Let $f$ be van der Waerden’s example (V.3.8.1.).

(a) Let $a$ be a real number whose decimal part is $d_1d_2d_3\cdots$, where $d_n$ is in the range $0-4$ for $n$ odd and in the range $5-9$ for $n$ even. Show that $D_-f(a) = D_+f(a) = 0$ and $D^-f(a) = D^+f(a) = 1$.

(b) Show that if $m, n \in \mathbb{Z}$, $m < n$, there are uncountably many points $a$ in $[0, 1]$ with $D_-f(a) = D_+f(a) = m$ and $D^-f(a) = D^+f(a) = n$. [If $n > 0$, take the first $n$ decimal places in the range $0-4$, the next $n-m$ in the range $5-9$, etc.]

(c) Extend the argument of $b$ to cover the cases where $m$ and/or $n$ is $\pm \infty$. (Both can be $+\infty$ or both $-\infty$, or we can have $m = -\infty$ and $n = +\infty$.)

(d) For how many points $a$ is $D_-f(a) \neq D_+f(a)$ and/or $D^-f(a) \neq D^+f(a)$? What are the possibilities?
V.3.11. Calculus of Finite Differences

“The analysis of the infinite ... is nothing but a special case of the method of differences, ... wherein the differences are infinitely small ...”

*L. Euler*

In this section, we discuss a discrete version of derivatives. This procedure is important historically as a motivation of much of differential and integral calculus (cf. e.g. [Eul00]), and is still very useful in numerical analysis.

V.3.11.1. We start with a functional relation $y = f(x)$, where $f$ is a function defined, say, on all of $\mathbb{R}$ (simply to make the discussion simpler by not having to worry about domain questions). Fix an $x_0 \in \mathbb{R}$, and an “increment” $h > 0$ ($h$ is often written $\Delta x$). Form the sequence

$$x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_n = x_0 + nh, \ldots$$

which is an arithmetic sequence whose successive differences are each equal to $h = \Delta x$. Set

$$\Delta x_n = x_{n+1} - x_n$$

for each $n$; of course, $\Delta x_n = h$ for all $n$ since the $x_n$ were chosen evenly spaced.

V.3.11.2. Now form the sequence

$$y_0, y_1, y_2, \ldots$$

where $y_n = f(x_n)$ for each $n$. This sequence is rarely an arithmetic sequence (it will be arithmetic if $f$ is linear or affine); its properties are closely connected to the properties of the function $f$. Set

$$\Delta y_n = y_{n+1} - y_n = f(x + (n + 1)h) - f(x + nh)$$

for each $n$. The $\Delta y_n$ are called the sequence of first differences of $y$ (or $f$) corresponding to the increment $h = \Delta x$. This sequence is rarely constant (it will be constant if $f$ is linear or affine); its properties are closely connected to the properties of the function $f$. If $f$ is a quadratic polynomial, the first differences will form an arithmetic progression.

V.3.11.3. Go on to form the second differences

$$\Delta^2 y_n = \Delta y_{n+1} - \Delta y_n$$

for each $n$. The second differences are all zero if and only if the first differences are constant, e.g. if $f$ is affine. If $f$ is a quadratic polynomial, the second differences will be constant.

V.3.11.4. Successively form higher differences

$$\Delta^{k+1} x_n = \Delta^k y_{n+1} - \Delta^k y_n$$

for each $k$ and $n$. The sequence of $(k+1)$’st differences of $(y_n)$ is the sequence of first differences of $(\Delta^k y_n)$. If $f$ is a polynomial of degree $\leq k$, then the sequence of $k$’th differences will be constant, so the sequence of $(k+1)$’st differences will be $0$.

---

3[Eul00, p. 64].
V.3.11.5. The successive differences can be conveniently displayed in a table

\[
\begin{array}{cccccc}
  y_0 & y_1 & y_2 & y_3 & \cdots \\
\Delta y_0 & \Delta y_1 & \Delta y_2 & \Delta y_3 & \cdots \\
\Delta^2 y_0 & \Delta^2 y_1 & \Delta^2 y_2 & \cdots \\
\Delta^3 y_0 & \Delta^3 y_1 & \Delta^3 y_2 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
\]

where the first row is computed from the function \( f \) and each other entry is the difference of the two entries just above it.

V.3.11.6. The sequence of \( k \)'th differences is a discrete version of the \( k \)'th derivative \( f^{(k)} \) (although the differences are defined for any \( f \) whether or not it is differentiable or even continuous).

V.3.11.7. Examples. (i) suppose \( f(x) = mx \). Then the table becomes

\[
\begin{array}{cccccc}
  mx_0 & mx_1 & mx_2 & mx_3 & \cdots \\
  m\Delta x & m\Delta x & m\Delta x & m\Delta x & \cdots \\
  0 & 0 & 0 & \cdots \\
  0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{array}
\]

since \( mx_{n+1} - mx_n = m\Delta x \) for all \( n \).

(ii) Let \( f(x) = x^2 \). Then the table becomes

\[
\begin{array}{cccccc}
  x_0^2 & x_1^2 & x_2^2 & x_3^2 & \cdots \\
  (x_0 + x_1)\Delta x & (x_1 + x_2)\Delta x & (x_2 + x_3)\Delta x & (x_3 + x_4)\Delta x & \cdots \\
  2(\Delta x)^2 & 2(\Delta x)^2 & 2(\Delta x)^2 & \cdots \\
  0 & 0 & 0 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{array}
\]

since \( x_{n+1}^2 - x_n^2 = (x_n + x_{n+1})(x_{n+1} - x_n) = (x_n + x_{n+1})\Delta x \) for all \( n \). Since \( x_{n+1} = x_n + \Delta x \), we have
\[ \Delta y_n = 2x_n \Delta x + (\Delta x)^2, \] so the table becomes

\[
\begin{array}{cccccc}
& x_0^2 & x_1^2 & x_2^2 & x_3^2 & \cdots \\
2x_0 \Delta x + (\Delta x)^2 & 2x_1 \Delta x + (\Delta x)^2 & 2x_2 \Delta x + (\Delta x)^2 & 2x_3 \Delta x + (\Delta x)^2 & \cdots \\
\end{array}
\]

\[
\begin{array}{cccccc}
& 2(\Delta x)^2 & 2(\Delta x)^2 & 2(\Delta x)^2 & \cdots \\
0 & 0 & 0 & \cdots & \cdots \\
\end{array}
\]

(iii) Let \( f(x) = x^k, k \in \mathbb{N}. \) The binomial theorem shows that

\[ \Delta y_n = (x_n + \Delta x)^k - x_n^k = kx_n^{k-1} \Delta x + \binom{k}{2} x_n^{k-2} (\Delta x)^2 + \cdots + k x_n (\Delta x)^{k-1} + (\Delta x)^k \]

\[ = \left( x_n^{k-1} + x_n^{k-2} x_{n+1} + x_n^{k-3} x_{n+1}^2 + \cdots + x_n x_{n+1}^{k-2} + x_{n+1}^{k-1} \right) \Delta x. \]

The succeeding rows can be similarly calculated. The general formula is complicated, but the \((k+1)\)st row is constant, \( \Delta^k y_n = k!(\Delta x)^k \) for all \( n \), and the succeeding rows are 0.

(iv) Let \( f(x) = \frac{1}{x}. \) (This is not defined for all \( x \in \mathbb{R} \), but we only need that it is defined for \( x \geq x_0 \) to make the above construction of the differences work, so we just need to take \( x_0 > 0 \).) Then

\[ \Delta y_n = \frac{1}{x_n + \Delta x} - \frac{1}{x_n} = -\frac{1}{x_n(x_n + \Delta x)} \Delta x = -\frac{1}{x_n x_{n+1}} \Delta x \]

\[ \Delta^2 y_n = -\frac{1}{x_{n+1} x_{n+2}} \Delta x + \frac{1}{x_n x_{n+1}} \Delta x = \frac{x_{n+2} - x_n}{x_n x_{n+1} x_{n+2}} \Delta x = \frac{2 \Delta x}{x_n x_{n+1} x_{n+2}} (\Delta x)^2 \]

and higher differences can be calculated similarly. In this case the rows never become identically zero.

(v) Let \( f(x) = \sqrt{x} = x^{1/2}. \) Then

\[ \Delta y_n = \sqrt{x_n + \Delta x} - \sqrt{x_n} = \left( \sqrt{x_n + \Delta x} - \sqrt{x_n} \right) \sqrt{x_n + \Delta x} + \sqrt{x_n} = \frac{1}{\sqrt{x_n + \Delta x} + \sqrt{x_n}} \Delta x = \frac{1}{\sqrt{x_n} + \sqrt{x_{n+1}}} \Delta x. \]

Again, no row of higher differences becomes zero.

(vi) Let \( f(x) = a^x \) for fixed \( a > 0 \) (the comment from (iv) applies to this function). Then the sequence \( (y_0, y_1, \ldots) \) is a geometric progression with successive ratios \( b = a^h \). We have

\[ \Delta y_n = a^{x_0 + (n+1)h} - a^{x_0 + nh} = a^{x_0 + nh}(a^h - 1) = cy_n \]

where \( c = a^h - 1 \). Thus each succeeding row is also a geometric progression.
V.3.11.8. The $x_0$ is fixed in these calculations. But at the end we want to allow the $x_0$ to be any number (in the domain), and thus often write $x$ in place of $x_0$ and $y$ in place of $y_0$, and hence $\Delta^m y$ in place of $\Delta^m y_0$. Thus the above formulas become:

(i) If $f(x) = mx$, then $\Delta y = m\Delta x$ and $\Delta^k y = 0$ for $k > 1$.

(ii) If $f(x) = x^2$, then $\Delta y = 2x\Delta x + (\Delta x)^2$, $\Delta^2 y = 2(\Delta x)^2$, and $\Delta^m y = 0$ for $m > 2$.

(iii) If $f(x) = x^k$, then

\[
\Delta y = kx^{k-1}\Delta x + \left(\frac{k}{2}\right)x^{k-2}(\Delta x)^2 + \cdots + kx(\Delta x)^{k-1} + (\Delta x)^k
\]

and $\Delta^k y = k!(\Delta x)^k$.

(iv) If $f(x) = \frac{1}{x}$, then $\Delta y = -\frac{1}{x(x+\Delta x)}\Delta x$.

(v) If $f(x) = \sqrt{x}$, then $\Delta y = \frac{1}{\sqrt{x+\sqrt{x+\Delta x}}}\Delta x$.

(vi) If $f(x) = a^x$, then $\Delta y = ca^x = cy$, where $c = a^{\Delta x} - 1$.

V.3.11.9. The formulas from these examples are analogous to differentiation formulas, and suggest that the derivative of a power function should be a power function whose exponent is reduced by one, and that the derivative of an exponential function should be another exponential function. Here is a nonrigorous but suggestive closer link to differentiation. If we regard $h = \Delta x$ as “infinitesimally small,” so that the $x_n$ are essentially equal to $x = x_0$ (at least for $n$ small), in the case $f(x) = x^k$ we have

\[
\Delta y_n = (x_n^{k-1} + x_n^{k-2}x_n+1 + x_n^{k-3}x_n^{2} + \cdots + x_n^{k-2}x_n+1 + x_n+1)\Delta x \approx kx^{k-1}\Delta x
\]

for small $n$, and if $f(x) = a^x$ we have

\[
\Delta y_n = a^{x+(n+1)h} - a^{x+nh} = a^{x+nh}(a^h - 1) \approx cy = ca^x
\]

where $c = a^h - 1$.

V.3.11.10. More rigorously, we have

\[
f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}
\]

(to be completely rigorous we should also allow $h < 0$) so in the case where $f(x) = x^k$ we get $f'(x) = kx^{k-1}$, and in the case $f(x) = a^x$ we get

\[
f'(x) = a^x \left[ \lim_{h \to 0} \frac{a^h - 1}{h} \right].
\]

Of course, these are in fact the correct formulas. Note also that in these cases we have

\[
f^{(m)}(x) = \lim_{\Delta x \to 0} \frac{\Delta^m y}{(\Delta x)^m}
\]

for all $m$. This is one of the origins of the notation $\frac{d^m y}{dx^m}$ for $f^{(m)}(x)$. 
V.3.11.11. One can also consider the inverse problem, which is the discrete antiderivative problem: given the sequence of first differences \((\Delta y_n)\), find the sequence \((y_n)\) it comes from. The \(y_n\) are not uniquely determined: an arbitrary constant can be added to all the \(y_n\). This amounts to making an arbitrary choice of \(y_0\). Once \(y_0\) has been chosen, there is a unique solution:

\[
y_n = y_0 + \sum_{m=0}^{n-1} \Delta y_m .
\]
V.4. Extensions of the Derivative

There are a number of ways to extend the derivative concept which prove to be useful. We discuss several of these generalizations in this section.

V.4.1. One-Sided Derivatives

The first generalization we consider is one-sided derivatives. The definition is quite straightforward:

V.4.1.1. Definition. (i) Let $f$ be a real-valued function on the interval $[x_0, b)$. If

$$\lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}$$

exists (as a real number), then $f$ is right differentiable at $x_0$, and this number is the right derivative of $f$ at $x_0$, denoted $D_R f(x_0)$.

(ii) Let $f$ be a real-valued function on the interval $(a, x_0]$. If

$$\lim_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}$$

exists (as a real number), then $f$ is left differentiable at $x_0$, and this number is the left derivative of $f$ at $x_0$, denoted $D_L f(x_0)$.

Various notations are used for left and right derivatives, such as $D_l f$ and $f'$ for $D_L f$ and $D_r f$ and $f'_+$ for $D_R f$.

The next proposition is obvious from (i).

V.4.1.2. Proposition. Let $f$ be a real-valued function on an interval $(a, b)$, and $x_0 \in (a, b)$. Then $f$ is differentiable at $x_0$ if and only if it is both left and right differentiable at $x_0$ and $D_L f(x_0) = D_R f(x_0)$. We then have

$$f'(x_0) = D_L f(x_0) = D_R f(x_0).$$

We can have left and right derivatives which are not equal, or one or both may fail to exist:

V.4.1.3. Examples. (i) Let $f(x) = |x|$. Then $D_L f(0) = -1$ and $D_R f(x) = 1$. Thus $f$ is both left and right differentiable at 0, but not differentiable at 0.

(ii) Let $f(x) = \sqrt{x}$ on $[0, \infty)$. Then $D_R f(0)$ does not exist (we sometimes say $D_R f(0) = +\infty$), and $D_L f(0)$ is undefined since $f$ is undefined to the left of 0. So $f$ is neither left or right differentiable at 0.

(iii) Let $f(x) = \sqrt{x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Then $D_R f(0)$ does not exist and $D_L f(0) = 0$. So $f$ is left differentiable but not right differentiable at 0.

(iv) Let $f(x) = x \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then neither $D_R f(0)$ nor $D_L f(0)$ exists. So $f$ is neither left or right differentiable at 0.
All these examples are continuous. We have the following slight generalization of (\ref{eq:cont}), whose proof is nearly identical and is left to the reader. Recall the definition of left and right continuity (\ref{eq:lrcont}).

V.4.1.4. **PROPOSITION.** (i) Let $f$ be a real-valued function on an interval $[x_0, b)$. If $f$ is right differentiable at $x_0$, then $f$ is right continuous at $x_0$.

(ii) Let $f$ be a real-valued function on an interval $(a, x_0]$. If $f$ is left differentiable at $x_0$, then $f$ is left continuous at $x_0$.

**Endpoint Differentiability and Differentiability on an Interval**

If $x_0$ is an endpoint of the domain of a function, it does not make sense to say that $f$ is differentiable at $x_0$. But one-sided differentiability does make sense, and is called *endpoint differentiability* in this context. So we may make the following definition in analogy with (\ref{eq:cont}):

V.4.1.5. **DEFINITION.** Let $f$ be a real-valued function on an interval $[a, b]$. Then $f$ is differentiable on $[a, b)$ if it is differentiable at all points of $(a, b)$, right differentiable at $a$, and left differentiable at $b$. Differentiability on an interval of the form $[a, b)$ or $(a, b]$ is defined analogously.

V.4.2. **Dini Derivatives**

A further generalization is very useful. Recall the definition of $\limsup$ and $\liminf$ from (\ref{eq:limsupinf}).

V.4.2.1. **DEFINITION.** (i) Let $f$ be a real-valued function on the interval $[x_0, b)$. Set

\[
D^+ f(x_0) = \limsup_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} = \limsup_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}
\]

\[
D_+ f(x_0) = \liminf_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} = \liminf_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}
\]

(as extended real numbers). $D^+ f(x_0)$ and $D_+ f(x_0)$ are the *upper and lower right Dini derivatives* of $f$ at $x_0$.

(i) Let $f$ be a real-valued function on the interval $(a, x_0]$. Set

\[
D^- f(x_0) = \limsup_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \limsup_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}
\]

\[
D_- f(x_0) = \liminf_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} = \liminf_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0}
\]

(as extended real numbers). $D^- f(x_0)$ and $D_- f(x_0)$ are the *upper and lower left Dini derivatives* of $f$ at $x_0$.

V.4.2.2. **Note that the Dini derivatives are always defined, provided only that the function is defined on an interval to the appropriate side of $x_0$. We clearly always have $D_+ f(x_0) \leq D^+ f(x_0)$ and $D_- f(x_0) \leq D^- f(x_0)$.**

We then have an obvious characterization of left and right differentiability (cf. (\ref{eq:limsupinf})): 558
V.4.2.3. **Proposition.** (i) Let $f$ be a real-valued function on an interval $[x_0, b)$. Then $f$ is right differentiable at $x_0$ if and only if $D^+ f(x_0)$ and $D_+ f(x_0)$ are equal and finite. We then have

$$D_R f(x_0) = D^+ f(x_0) = D_+ f(x_0).$$

(ii) Let $f$ be a real-valued function on an interval $(a, x_0]$. Then $f$ is left differentiable at $x_0$ if and only if $D^- f(x_0)$ and $D_- f(x_0)$ are equal and finite. We then have

$$D_L f(x_0) = D^- f(x_0) = D_- f(x_0).$$

(iii) Let $f$ be a real-valued function on an interval $(a, b)$, and $x_0 \in (a, b)$. Then $f$ is differentiable at $x_0$ if and only if all four Dini derivatives at $x_0$ are equal and finite. We then have

$$f'(x_0) = D^+ f(x_0) = D_+ f(x_0) = D^- f(x_0) = D_- f(x_0).$$

V.4.2.4. **Examples.** (i) Let $f(x) = |x|$. Then $D^- f(0) = D_- f(0) = -1$ and $D^+ f(0) = D_+ f(x) = 1$.

(ii) Let $f(x) = \sqrt{x}$ on $[0, \infty)$. Then $D^+ f(0) = D_+ f(0) = +\infty$, and $D^- f(0)$ and $D_- f(0)$ are undefined since $f$ is undefined to the left of 0.

(iii) Let $f(x) = \sqrt{|x|}$. Then $D^+ f(0) = D_+ f(0) = +\infty$ and $D^- f(0) = D_- f(0) = -\infty$.

(iv) Let $f(x) = x \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then $D^+ f(0) = D_- f(0) = 1$ and $D_+ f(0) = D_- f(0) = -1$.

(v) Let $f(x) = x \sin \frac{1}{x}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Then $D^+ f(0) = 1$, $D_+ f(0) = -1$, and $D^- f(0) = D_- f(0) = 0$.

(vi) Let $f(x) = \sqrt{x} \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then $D^+ f(0) = D^- f(0) = +\infty$ and $D_+ f(0) = D_- f(0) = -\infty$.

(vii) Let $f(x) = x^{1/3}$. Then $D^+ f(0) = D_+ f(0) = D^- f(0) = D_- f(0) = +\infty$.

We have a version of (i) also for Dini derivatives:

V.4.2.5. **Proposition.** (i) Let $f$ be a real-valued function on an interval $[x_0, b)$. If $f$ is not right continuous at $x_0$, then at least one right Dini derivative at $x_0$ is infinite.

(ii) Let $f$ be a real-valued function on an interval $(a, x_0]$. If $f$ is not left continuous at $x_0$, then at least one left Dini derivative at $x_0$ is infinite.

**Proof:** The proofs of both (i) and (ii) consist of four nearly identical cases. We do only one of the cases for (i), and leave the rest of the cases to the reader.

In (i), if $f$ is not right continuous at $x_0$, then either $f(x_0) \neq \liminf_{x \to x_0^+} f(x)$ or $f(x_0) \neq \limsup_{x \to x_0^+} f(x)$ (or both). Suppose $f(x_0) < \liminf_{x \to x_0^+} f(x)$. Then

$$\liminf_{x \to x_0^+} [f(x) - f(x_0)] > 0$$

559
and, since $\lim_{x \to x_0^+} \frac{1}{x - x_0} = +\infty$, by (i) we have

$$D_+ f(x_0) = \liminf_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0} = +\infty.$$  

V.4.2.6. Note that the converse is not true by Examples V.4.2.4.(iii),(v),(vi). In fact, there are (many!) continuous functions $f$ on $\mathbb{R}$ for which $D^+ f(x) = D^- f(x) = +\infty$ and $D_+ f(x) = D_- f(x) = -\infty$ for every $x \in \mathbb{R}$.

Local Extrema

Here is a general version of V.3.6.3. The proof is essentially the same.

V.4.2.7. **Proposition.** Let $f$ be a function on an interval $(a,b)$, and $x_0 \in (a,b)$.

(i) If $f$ has a local maximum at $x_0$, then $D^- f(x_0) \geq 0$ and $D^+ f(x_0) \leq 0$.

(ii) If $f$ has a local minimum at $x_0$, then $D^- f(x_0) \leq 0$ and $D^+ f(x_0) \geq 0$.

This result also applies when $x_0$ is an endpoint of an interval in the domain of $f$, except that the Dini derivatives on the outside of the interval are not defined (or not relevant):

V.4.2.8. **Proposition.** Let $f$ be a real-valued function on $[a,b]$. Then

(i) If $f$ has an endpoint maximum at $a$, then $D^+ f(a) \leq 0$.

(ii) If $f$ has an endpoint minimum at $a$, then $D_- f(a) \geq 0$.

(iii) If $f$ has an endpoint maximum at $b$, then $D_- f(a) \geq 0$.

(iv) If $f$ has an endpoint minimum at $b$, then $D^+ f(a) \leq 0$.

V.4.2.9. Converses of these statements are, of course, false; the counterexamples in the differentiable case, such as $f(x) = x^3$ at 0, are also counterexamples here. But there is a version of the First Derivative Test in this generality (i).

V.4.2.10. At a single point, there may be no relation between the left and right Dini derivatives. However, at “most” points both upper Dini derivatives are greater than both lower derivatives:
V.4.2.11. Theorem. Let \( f \) be a real-valued function on an open interval \( I \). Set
\[
A = \{ x \in I : D^- f(x) < D_+ f(x) \}
\]
\[
B = \{ x \in I : D^+ f(x) < D_- f(x) \}.
\]
Then \( A \) and \( B \) are countable.

Proof: We prove that \( A \) is countable; the proof for \( B \) is essentially identical. Let \( x \in A \). Choose \( r_x \in \mathbb{Q} \) with \( D^- f(x) < r_x < D_+ f(x) \), and then fix \( s_x, t_x \in \mathbb{Q} \cap I \) with \( s_x < x < t_x \) and
\[
\frac{f(y) - f(x)}{y - x} < r_x \text{ for all } y, s_x < y < x
\]
\[
\frac{f(y) - f(x)}{y - x} > r_x \text{ for all } y, x < y < t_x.
\]
Then, for any \( y \neq x \) with \( s_x < y < t_x \), we have
\[
f(y) - f(x) > r_x (y - x).
\]
Define \( \phi : A \to \mathbb{Q}^3 \) by \( \phi(x) = (r_x, s_x, t_x) \). We show that \( \phi \) is one-to-one; then, since \( \mathbb{Q}^3 \) is countable, it will follow that \( A \) is countable. Suppose \( \phi(x) = \phi(z) \) for \( x, z \in A, x \neq z \). Since \( s_x = s_z < z < t_z = t_x \), we have
\[
f(z) - f(x) > r_x (z - x).
\]
Similarly, since \( s_z < x < t_z \), we have
\[
f(x) - f(z) > r_z (x - z).
\]
But \( r_x = r_z \), a contradiction. \( \diamondsuit \)

We get the following immediate consequence, which is not so obvious geometrically:

V.4.2.12. Corollary. Let \( f \) be a function on an interval \( I \). Then the set of \( x \in I \) for which \( D^- f(x) \) and \( D^+ f(x) \) are both defined (even allowing infinite values) but unequal is countable. In particular, if \( f \) is left and right differentiable everywhere in \( I \), then \( f \) is differentiable at all but countably many points of \( I \).

A far deeper result is the remarkable theorem proved successively for larger classes of functions by A. Denjoy, Grace Chisholm Young, and S. Saks:

V.4.2.13. Theorem. [Denjoy-Young-Saks] Let \( f \) be an arbitrary real-valued function on an interval \([a, b]\). Then for almost all \( x \in [a, b] \), one of the following four conditions holds:

1. \( f \) is differentiable at \( x \).
2. \( D^- f(x) = D_+ f(x) \in \mathbb{R}, D_+ f(x) = -\infty \), and \( D^- f(x) = +\infty \).
3. \( D^+ f(x) = D_- f(x) \in \mathbb{R}, D_- f(x) = -\infty \), and \( D^+ f(x) = +\infty \).
(4) \( D_- f(x) = D_+ f(x) = -\infty \) and \( D^- f(x) = D^+ f(x) = +\infty \).

In particular, for almost all \( x \in [a, b] \), either \( f \) is differentiable at \( x \) or else \( \max(D^- f(x), D^+ f(x)) = +\infty \) and \( \min(D_- f(x), D_+ f(x)) = -\infty \).

If \( f \) is measurable, then for almost all \( x \), either condition (1) or condition (4) holds (i.e. conditions (2) and (3) hold only on sets of measure zero).

The proof uses some sophisticated machinery, and will be given in \( () \). This result is closely related to a result about approximate tangents to rectifiable sets in geometric measure theory [Mor09, 3.12].

V.4.3. Symmetric First and Second Derivatives

V.4.4. Approximate and Weak Derivatives

V.4.5. Exercises

V.4.5.1. Define a function \( f : [0, 1) \rightarrow [0, 1) \) as follows. If \( x \in [0, 1) \), write

\[
x = .d_1d_2d_3\ldots
\]

in its decimal expansion not ending in an infinite string of 9’s. Set

\[
f(x) = .0d_10d_20d_3\ldots.
\]

(a) Show that \( f \) is a strictly increasing function.

(b) Show that \( f \) is right continuous everywhere.

(c) Show that \( f \) is left continuous except at the decadent rationals \( () \).

(d) Show that \( D_R f(x) \) exists and equals 0 everywhere.
V.5. Exponential and Trigonometric Functions

Anyone who has studied mathematics through calculus knows that exponential and trigonometric functions are of great importance, both theoretical and practical. However, the theoretical subtleties involved in properly defining these functions and establishing their properties are often underappreciated. Since the functions are so important and useful, they are usually introduced in the mathematics curriculum before the necessary tools (from calculus) for a proper treatment are available; and by the time the right tools are discussed, the functions are so familiar that they are often taken for granted. (These comments also apply to the historical development of exponential and trigonometric functions.)

A careful and rigorous treatment of exponential and/or trigonometric functions can be done in several ways. The two most common methods are: (1) defining one of the inverse functions (log or arcsin) as the antiderivative of a familiar function, or (2) using infinite series. In the case of exponential functions, there is a third approach which proceeds directly from algebraic rules of exponents. This third approach is the one we will take for exponential and logarithm functions in this section; see Exercise () for the first approach via integrals, and () for the infinite series approach. Our treatment of trigonometric functions will follow approach (1) (see () for the infinite series approach here).

V.5.1. Rational Exponents

V.5.1.1. Recall that if \( x \in \mathbb{R} \) and \( n \in \mathbb{N} \), then \( x^n = x \cdot x \cdots x \) (\( n \) times). If \( x \neq 0 \), extend this to \( n \in \mathbb{Z} \) by \( x^0 = 1 \) and \( x^{-n} = \frac{1}{(x^{-1})^n} = (x^n)^{-1} \). The usual rules of exponents hold: \( x^{m+n} = x^m x^n \), \( x^{m-n} = \frac{x^m}{x^n} \), \( (x^m)^n = x^{mn} \), \( (xy)^n = x^ny^n \) for all \( x, y \neq 0 \) and all \( m, n \in \mathbb{Z} \).

If \( n \in \mathbb{N} \) and \( x > 1 \), then \( x^n > x^m \). So if \( n \in \mathbb{N} \) and \( 0 < x < y \), then \( x^n < y^n \). If \( x > 1 \), then \( f(x) = x^n \) is strictly increasing on \((0, \infty)\). Similarly, if \( n \in \mathbb{Z} \), \( n < 0 \), then \( f(x) = x^n \) is strictly decreasing on \((0, \infty)\). For any \( n \in \mathbb{Z} \), \( f(x) = x^n \) is continuous and even differentiable on \((0, \infty)\), and its range is unbounded above and below in \((0, \infty)\) if \( n \neq 0 \).

Thus, by the IVT, if \( n \in \mathbb{Z} \), \( n \neq 0 \), then the range of \( f(x) = x^n \) is all of \((0, \infty)\), and by () \( f \) has a continuous inverse function \( f^{-1} \) which also maps \((0, \infty)\) onto \((0, \infty)\), and \( f^{-1} \) is strictly increasing if \( n > 0 \), strictly decreasing if \( n < 0 \). If \( x > 0 \), write \( x^{1/n} \) for \( f^{-1}(x) \). If \( n > 0 \), the definition of \( x^{1/n} \) can also be extended to \( x = 0 \) by setting \( 0^{1/n} = 0 \), and if \( n \) is odd, to \( x < 0 \) by \( x^{1/n} = -(x^{-1})^{1/n} \); the extended function is continuous on its domain.

V.5.1.2. Proposition. If \( p, q \in \mathbb{Z} \), \( q \neq 0 \), for \( x > 0 \) define \( x^{p/q} = (x^p)^{1/q} \). We then have:

(i) For \( p, q \in \mathbb{Z} \), \( q \neq 0 \), and \( x > 0 \), \( [(x^p)^{1/q}]^q = x^p = [(x^{1/q})^q]^p = [(x^{1/q})^p]^q \), so \( (x^p)^{1/q} = (x^{1/q})^p \) since \( f(x) = x^q \) is one-to-one.

(ii) Similarly, if \( p, q, r, s \in \mathbb{Z} \), \( q, s \neq 0 \), and \( x > 0 \), then \( [(x^{p/q})^{r/s}]^{q/s} = x^{p/r} = [(x^{r/s})^{p/q}]^{q/s} \), so \( (x^{p/q})^{r/s} = (x^{r/s})^{p/q} \). In particular, \( x^{p/s} = x^{p/q} \), and hence \( x^t \) is well defined for any \( t \in \mathbb{Q} \).

(iii) If \( r, s \in \mathbb{Q} \), \( x, y \in (0, \infty) \), we have \( (x^r)^s = x^{rs} \), \( x^{1+s} = x^{1}x^{s} \), \( x^{-s} = \frac{x}{x^s} \), \( (xy)^r = x^ry^r \).

(iv) If \( r \in \mathbb{Q} \), \( r \neq 0 \), the function \( f(x) = x^r \) is a composition of two strictly monotone continuous functions from \((0, \infty)\) onto \((0, \infty)\), hence strictly monotone and continuous, strictly increasing if \( r > 0 \) and strictly decreasing if \( r < 0 \).

(v) \( 1^r = 1 \) for all \( r \in \mathbb{Q} \), so if \( x > 1 \), then \( x^r > 1 \) for \( r > 0 \) and \( 0 < x^r < 1 \) for \( r < 0 \).
V.5.1.3. **Proposition.** If \( r \in \mathbb{Q} \), then \( f(x) = x^r \) is differentiable on \((0, \infty)\), and \( f'(x) = rx^{r-1} \) on \((0, \infty)\).

**Proof:** If \( n \in \mathbb{N} \), then \( \frac{d}{dx} (x^n) = nx^{n-1} \) \((v)\). The formula also obviously holds if \( n = 0 \). If \( n \in \mathbb{N} \), then using the Quotient Rule \( x^{-n} \) is differentiable on \((0, \infty)\) and

\[
\frac{d}{dx} (x^{-n}) = \frac{d}{dx} \left( \frac{1}{x^n} \right) = \frac{0 - nx^{-n-1}}{x^{2n}} = -nx^{-n-1}
\]

so the formula holds for all \( r \in \mathbb{Z} \). If \( q \in \mathbb{Z} \), \( q \neq 0 \), then by the Inverse Function Theorem \((v)\) \( x^{1/q} \) is differentiable on \((0, \infty)\) and

\[
\frac{d}{dx} \left( x^{1/q} \right) = \frac{1}{q(x^{1/q})^{q-1}} = \frac{1}{q x^{1-1/q}} = \frac{1}{q} x^{(1/q)-1}
\]

If \( r = \frac{p}{q} \), then by the Chain Rule \( x^r \) is differentiable on \((0, \infty)\) and

\[
\frac{d}{dx} \left( x^{p/q} \right) = \frac{d}{dx} \left( (x^p)^{1/q} \right) = \frac{1}{q} (x^p)^{(1/q)-1} \cdot px^{p-1} = \frac{p}{q} x^{(p/q)-1} = \frac{p}{q} x^{(p/q)-1}.
\]

V.5.2. **Irrational Exponents and Exponential Functions**

Now fix \( b > 1 \) and look at the function \( f(x) = b^x \) with domain \( \mathbb{Q} \).

V.5.2.1. **Proposition.** \( b^x \) is strictly increasing.

**Proof:** If \( x, y \in \mathbb{Q} \) with \( x < y \), then \( b^y = b^{x + (y-x)} = b^x b^{y-x} \), and \( b^{y-x} > 1 \) by \((v)\).

V.5.2.2. **Proposition.** \( \{b^n : n \in \mathbb{N}\} \) is unbounded.

**Proof:** If \( S = \{b^n : n \in \mathbb{N}\} \) is bounded, let \( a = \sup(S) \). Since \( b > 1 \), \( \frac{a}{b} < a \), so there is an \( n \in \mathbb{N} \) with \( b^n > \frac{a}{b} \). Then \( b^{n+1} > a \), a contradiction.

V.5.2.3. **Corollary.** \( \lim_{n \to \infty} b^{1/n} = 1 \).

**Proof:** The sequence \( (b^{1/n}) \) is decreasing and bounded below by 1, hence \( \lim_{n \to \infty} b^{1/n} \) exists and equals \( c := \inf_{n \in \mathbb{N}} b^{1/n} \). Then \( c \geq 1 \) and \( c^n \leq b \) for all \( n \in \mathbb{N} \), so \( c = 1 \) by V.5.2.2..
V.5.2.4. Proposition. If \( a \in \mathbb{Q} \), then the function \( f(x) = b^x \) is uniformly continuous on \((-\infty, a) \cap \mathbb{Q}\).

Proof: If \( \epsilon > 0 \), then there is an \( n \in \mathbb{N} \) with \( b^{1/n} - 1 < b^{-a} \epsilon \). Set \( \delta(\epsilon) = 1/n \). Then, if \( x, y \in \mathbb{Q} \), \( x < y < a \), \( y - x < \delta(\epsilon) \), we have

\[
0 < b^y - b^x = b^x (b^{y-x} - 1) < b^x (b^{y-x} - 1) < \epsilon.
\]

V.5.2.5. Corollary. The function \( f(x) = b^x \ (x \in \mathbb{Q}) \) extends to a continuous strictly increasing function on all of \( \mathbb{R} \), also denoted \( b^x \), with range \((0, \infty)\).

Proof: The extension is guaranteed by ??, and is continuous and increasing. To see that it is strictly increasing, let \( x, y \in \mathbb{R} \), \( x < y \), and let \( r, s \in \mathbb{Q} \), \( x < r < s < y \). Then \( b^x \leq b^r < b^s \leq b^y \). The range is an interval (??) which must be all of \((0, \infty)\) by V.5.2.2.

V.5.2.6. Definition. The function \( f(x) = b^x \) thus has a continuous strictly increasing inverse function \( f^{-1} \) with domain \((0, \infty)\) and range \( \mathbb{R} \). The number \( f^{-1}(x) \) is called the logarithm of \( x \) with base \( b \), denoted \( \log_b(x) \).

V.5.2.7. Proposition. The following laws of exponents and logarithms hold: if \( x, y \in \mathbb{R} \) and \( b, c > 1 \), then

\[
\begin{align*}
    b^{x+y} &= b^x b^y, \\
b^{x-y} &= \frac{b^x}{b^y}, \\
b^{xy} &= (b^x)^y, \\
b^{(bc)x} &= b^x c^x; \\
    \text{if } s, t > 0, r \in \mathbb{R}, \text{ then } \log_b(st) &= \log_b(s) + \log_b(t), \\
    \log_b(st) &= \log_b(s) - \log_b(t), \\
    \log_b(s^r) &= r \log_b(s). 
\end{align*}
\]

If \( c > 1 \), then \( \log_b(s) = \log_c(s) \log_b(c) \). In particular, \( \log_b 1 = 0 \) and \( \log_b(c) = \frac{1}{\log_b(b)} \).

Proof: The laws of exponents have already been established for \( x, y \in \mathbb{Q} \). They follow for general \( x, y \) by continuity: for example, if \( (x_n) \) and \( (y_n) \) are sequences of rational numbers with \( x_n \to x \) and \( y_n \to y \), then \( b^{x_n} \to b^x \), \( b^{y_n} \to b^y \), \( b^{x_n + y_n} \to b^{x+y} \), and \( b^{x_n y_n} \to b^{x+y} \) by continuity of \( b^x \) and standard limit theorems. The laws of logarithms follow from the laws of exponents by setting \( x = \log_b(s) \) and \( y = \log_b(t) \), so \( s = b^x \) and \( t = b^y \): for example, \( st = b^x b^y = b^{x+y} \), so \( \log_b(st) = x + y \). For the last formula, if \( z = \log_c(s) \), then \( s = c^z = (b^{\log_b c})^z = b^{z \log_b c} \), so \( \log_b(s) = z \log_b c \).

If \( 0 < b < 1 \) and \( x \in \mathbb{R} \), we can define \( b^x = (b^{-1})^{-x} \). Then \( b^x \) is strictly decreasing, as is the inverse function \( \log_b(x) \). The same rules of exponents and logarithms hold. We can also define \( 1^x = 1 \) for all \( x \).

An important property of the exponential function \( b^x \ (b > 0, b \neq 1) \) is that it is strictly convex (??):

V.5.2.8. Lemma. If \( b > 1, k \in \mathbb{R}, \) and \( x > 0 \), then \( b^{k+2x} - b^k > 2(b^{k+x} - b^k) \).

Proof:

\[
b^{k+2x} - b^k = b^k (b^{2x} - 1) = b^k (b^x + 1)(b^x - 1) = (b^x + 1)(b^{k+x} - b^k).
\]

565
V.5.2.9. Proposition. If $b > 1$ and $x, y, z \in \mathbb{R}$ with $z < y < x$, then

$$
\frac{b^y - b^z}{y - z} < \frac{b^x - b^z}{x - z} < \frac{b^x - b^y}{x - y}.
$$

(It follows that $f(x) = b^x$ is a strictly convex function.)

Proof: Lemma V.5.2.8. means that on any interval $[x_1, x_2]$ with midpoint $c$, the point $(c, b^c)$ lies below the line between $(x_1, b^{x_1})$ and $(x_2, b^{x_2})$. Thus the first inequality follows from repeated applications of V.5.2.8. if $y - z$ is a dyadic rational multiple of $x - z$ (see Exercise 1). By continuity the first inequality holds with $a$ for all $x, y, z$. To show that the strict inequality holds in general, choose $w$ with $y < w < x$ and $y - z$ a dyadic rational multiple of $w - z$. Then

$$
\frac{b^y - b^z}{y - z} < \frac{b^w - b^z}{w - z} < \frac{b^x - b^z}{x - z}.
$$

For the second inequality, suppose that $\frac{b^x - b^y}{x - y} < \frac{b^x - b^z}{x - z}$. Then

$$
b^x - b^z = (y - z)\frac{b^y - b^z}{y - z} + (x - y)\frac{b^x - b^y}{x - y} < (y - z)\frac{b^x - b^z}{x - z} + (x - y)\frac{b^x - b^z}{x - z} = b^x - b^z
$$

which is a contradiction.

V.5.2.10. Corollary. Let $b > 1$, and $0 < z < y < x$. Then

$$
\frac{\log_b x - \log_b y}{x - y} < \frac{\log_b x - \log_b z}{x - z} < \frac{\log_b y - \log_b z}{y - z}.
$$

Note that V.5.2.9. and V.5.2.10. also hold for $0 < b < 1$. For V.5.2.9.,

$$
\frac{(b^{-1})^{-y} - (b^{-1})^{-x}}{-y + x} < \frac{(b^{-1})^{-z} - (b^{-1})^{-x}}{-z + x} < \frac{(b^{-1})^{-z} - (b^{-1})^{-y}}{-z + y}.
$$

V.5.3. The Number $e$

V.5.3.1. Lemma. If $x > 0$ and $k > 1$, then $(1 + x)^k > 1 + kx$.

Proof: Fix $k > 1$ rational, and let $f(x) = (1 + x)^k - 1 - kx$ for $x \geq 0$. Then $f$ is differentiable for $x \geq 0$, $f(0) = 0$, and $f'(x) = k(1 + x)^{k-1} - k = k[(1 + x)^{k-1} - 1] > 0$ for $x > 0$. The formula thus holds for $k$ rational, and thus for $k$ irrational with $\geq$ by continuity. To see that the strict inequality holds for general $k > 1$, choose rational $r$ with $1 < r < k$. Then

$$
(1 + x)^k = [(1 + x)^r]^k/r > [1 + rx]^{k/r} \geq 1 + \frac{k}{r}rx.
$$

566
V.5.3.2. **Corollary.** The function \( g(x) = (1 + x)^{1/x} \) is decreasing for \( x > 0 \).

**Proof:** Let \( x > 0 \) and \( k > 1 \). Then \( (1 + kx)^{1/kx} < ([1 + x]^k)^{1/kx} = (1 + x)^{1/x} \).

Thus, to show that \( \lim_{x \to 0^+} (1 + x)^{1/x} \) exists, it suffices to show that \( (1 + x)^{1/x} \) is bounded for \( x > 0 \).

V.5.3.3. **Theorem.** Let \( h(x) = (1 + x)^{1+1/x} \) for \( x > 0 \). Then

(i) \( h \) is increasing.

(ii) \( g(x) < h(z) \) for all \( x, z > 0 \).

(iii) \( \lim_{x \to 0^+} g(x) \) and \( \lim_{x \to 0^+} h(x) \) exist and are equal.

V.5.3.4. **Lemma.** Let \( b > 1 \), \( x > 0 \), and \( 0 < y < 1 \). Then

\[
\log_b([1 + x]^{1/x}) = \frac{\log_b(1 + x)}{x} < -\frac{\log_b(1 - y)}{y} = \log_b\left(\frac{1}{1 - y}\right)^{1/y}.
\]

**Proof:** Apply V.5.2.10. to \( 1 - y, 1, 1 + x \).

We now prove Theorem V.5.3.3.

**Proof:** For (i) and (ii), fix \( b > 1 \) and for \( z > 0 \) set \( y = \frac{z}{1+z} \). Then \( 0 < y < 1 \) and \( z = \frac{1}{1-y} - 1 \), and \( z \) increases from 0 to \( +\infty \) as \( y \) increases from \( 0 \) to \( 1 \). By V.5.2.10., \( -\frac{\log_b(1 - y)}{y} = \log_b\left(\frac{1}{1 - y}\right)^{1/y} \) increases in \( y \) for \( 0 < y < 1 \) [if \( 0 < y_1 < y_2 < 1 \), apply V.5.2.10. to \( 1 - y_2, 1 - y_1, 1 \), so \( \log_b([1 + z]^{1+z}) \) increases in \( z \) for \( z > 0 \), proving (i), and V.5.3.4. with \( y \) converted to \( z \) gives (ii).

Thus \( g(x) \) is increasing and bounded above as \( x \to 0^+ \), and \( h(x) \) is decreasing and bounded below, so \( \lim_{x \to 0^+} g(x) = \sup_{x>0} g(x) \) and \( \lim_{x \to 0^+} h(x) = \inf_{x>0} h(x) \) exist; call the limits \( d \) and \( e \) respectively. Then

\[
2 = (1 + 1)^1 < d \leq e,
\]

and \( \lim_{x \to 0^+} \frac{h(x)}{g(x)} = \frac{e}{d} \). But \( \frac{h(x)}{g(x)} = 1 + x \), so \( \lim_{x \to 0^+} \frac{h(x)}{g(x)} = 1 = \frac{e}{d}, \ d = e. \)

V.5.3.5. **Definition.** The number \( e \) is the limit \( \lim_{x \to 0^+} (1 + x)^{1/x} \).

The number \( e \) is one of the most important constants in mathematics, comparable in importance to \( \pi \).

V.5.3.6. **Corollary.** \( e > (1 + 1)^1 = 2 > 1 \), so \( f(x) = e^x \) is strictly increasing and continuous, and its inverse function \( \ln x = \log_e x \) is continuous.

The function \( \ln x \) is called the natural logarithm function, and mathematicians usually write it \( \log x \).
V.5.3.7. We can of course get better estimates of the numerical value of \(e\) using V.5.3.3. for small \(x\): for example, if \(x = .01\), we get \(2.704 < 1.01^{100} < e < 1.01^{101} < 2.732\). The average of \(1.01^{100}\) and \(1.01^{101}\), \(2.7183\), is an even better estimate (the geometric mean, \(2.7183\), is better yet). The exact numerical value of \(e\) is \(2.718281828\) (the decimal does not continue to repeat; \(e\) is an irrational number, in fact transcendental – see III.9.3.1.). There are more efficient ways to compute \(e\) using infinite series; see ??.

V.5.4. Derivatives of Exponential and Logarithm Functions

V.5.4.1. Theorem. The function \(f(x) = e^x\) is differentiable, and \(f'(x) = e^x\).

V.5.4.2. Lemma. If \(f(x) = e^x\), then \(f\) is differentiable at 0 and \(f'(0) = \lim_{h \to 0} \frac{e^h - 1}{h} = 1\).

Proof: Let \(x = e^h - 1\). Then \(h = \ln(1 + x)\) and \(x \to 0\) as \(h \to 0\) (i.e. \(\lim_{h \to 0} e^h - 1 = 0\) and \(\lim_{x \to 0} \ln(1 + x) = 0\) by continuity of \(e^x\) and \(\ln x\), and \(x > 0\) if and only if \(h > 0\). Thus

\[
\lim_{h \to 0^+} \frac{e^h - 1}{h} = \lim_{x \to 0^+} \frac{x}{\ln(1 + x)} = \lim_{x \to 0^+} \frac{1}{x \ln(1 + x)} = \lim_{x \to 0^+} \frac{1}{(1 + x)^{1/2}} = \frac{1}{\ln e} = 1
\]

since \(\lim_{x \to 0^+} (1 + x)^{1/2} = e\) and \(\ln x\) is continuous at \(e\). We also have

\[
\lim_{h \to 0^-} \frac{e^h - 1}{h} = \lim_{h \to 0^+} \frac{e^{-h} - 1}{-h} = \lim_{h \to 0^+} \frac{e^{-h}(1 - e^h)}{-h} = \left[ \lim_{h \to 0^+} e^{-h} \right] \left[ \lim_{h \to 0^+} \frac{e^h - 1}{h} \right] = 1
\]

We now prove Theorem V.5.4.1.

Proof: Fix \(x\). Then

\[
f'(x) = \lim_{h \to 0} \frac{e^{x+h} - e^x}{h} = \lim_{h \to 0} \frac{e^x e^h - e^x}{h} = \lim_{h \to 0} \frac{e^x (e^h - 1)}{h} = e^x \lim_{h \to 0} \frac{e^h - 1}{h} = e^x
\]

V.5.4.3. Corollary. Let \(b > 0\), and \(f(x) = b^x\). Then \(f\) is differentiable everywhere and \(f'(x) = (\ln b) b^x\) for all \(x\).

Proof: Since \(b^x = e^{x \ln b}\), the result follows from V.5.4.1. and the Chain Rule.

V.5.4.4. Corollary. If \(b > 0\), \(b \neq 1\), then \(f(x) = \log_b(x)\) is differentiable on \((0, \infty)\), and \(\frac{d}{dx} (\log_b(x)) = \frac{1}{x \ln b}\). In particular, \(\frac{d}{dx} (\ln x) = \frac{1}{x}\).

Proof: This follows immediately from V.5.4.3. and the Inverse Function Theorem:

\[
\frac{d}{dx} (\log_b x) = \frac{1}{(\ln b) b^{\log_b x}} = \frac{1}{x \ln b}.
\]
V.5.4.5. Corollary. For all \( r \in \mathbb{R} \), the function \( f(x) = x^r \) is differentiable (hence continuous) on \((0, \infty)\), and the formula \( \frac{d}{dx} (x^r) = r x^{r-1} \) holds on \((0, \infty)\).

Proof: Since \( x^r = e^{r \ln x} \) for \( x > 0 \), \( f \) is differentiable by V.5.4.1, V.5.4.4, and the Chain Rule, and we have

\[
\frac{d}{dx} (x^r) = \frac{d}{dx} (e^{r \ln x}) = e^{r \ln x} \frac{r}{x} = r x^{r-1}.
\]

\( \Box \)

V.5.5. Exercises

V.5.5.1. (a) Write out the details of the proof of V.5.1.2(ii).
(b) Prove the rules of exponents in V.5.1.2(iii), using arguments similar to those in (i). (Note that the first rule has already been proved in (ii).)

V.5.5.2. (a) Prove the assertion in the first sentence in the proof of V.5.2.9.
(b) Prove the rules of exponents in V.5.2.9, by induction, using the statement \( P(n) \): “For every \( u, v, w \in \mathbb{R} \) with \( w < v < u \) and \( \frac{w-u}{v-u} = \frac{k}{2^n} \) for some \( k \in \mathbb{N} \), \( k < 2^n \), \( \frac{u-v}{v-w} < \frac{k}{u-v} \cdot \frac{v-u}{w-u} \).

To prove \( P(n+1) \) for a set \( u, v, w \), apply \( P(n) \) to either the first or second half of the interval \([w, u]\) depending where \( v \) is.

V.5.5.3. Here is an alternate approach to the definition of \( e \) (cf. [?, III.4].) Let \( x_n = (1 + \frac{1}{n})^n \).
(a) Using the Binomial Theorem, show that

\[
x_n = 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right).
\]

(b) Using (a), show that \( x_n < x_{n+1} \) for any \( n \in \mathbb{N} \).
(c) Let \( y_n = \sum_{k=0}^n \frac{1}{k!} \). Show that \( x_n < y_n \) for all \( n \).
(d) Since \( k! > 2^{k-1} \) for all \( k > 2 \), show that \( y_n < 2 + \sum_{k=2}^n \frac{1}{2^{k-1}} < 3 \) for \( n > 2 \). Thus the sequence \( (x_n) \) is a bounded increasing sequence, hence convergent. Call the limit \( e \). Show that \( 2 < e < 3 \).

V.5.5.4. Use the notation of V.5.5.3.
(a) If \( 1 < k < n \), set

\[
z_{k,n} = 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right).
\]

Show that \( z_{k,n} < x_n \), so \( z_{k,n} < e \).
(b) Fix \( k \). Show \( \lim_{n \to \infty} z_{k,n} = y_k \). Conclude that \( y_k \leq e \) for all \( k \). (In fact, since \( y_k < y_{k+1} \leq e \), we have \( y_k < e \) for all \( k \).)

(c) Since \( x_k < y_k < e \) for all \( k \), and \( \lim_{k \to \infty} x_k = e \), we have \( \lim_{k \to \infty} y_k = e \).

569
V.5.5.5. Use the notation of V.5.5.3–V.5.5.4. The sequence \((y_n)\) converges to \(e\) fairly rapidly (unlike \((x_n)\), which converges rather slowly):

(a) For each \(n\) and each \(m > 1\), show that

\[
y_{n+m} - y_n = \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{m} \frac{1}{(n+2)\cdots(n+k)}\right)
\]

and thus, since \(\lim_{m \to \infty} y_{n+m} = e\), \(e - y_n \leq \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{\infty} \frac{1}{(n+2)\cdots(n+k)}\right)\).

(b) By comparing terms and using \(\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}\) for \(|r| < 1\),

\[
e - y_n \leq \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{\infty} \frac{1}{(n+2)\cdots(n+k)}\right)
\]

Thus, for example, \(e - y_{10} < \frac{12}{101} \approx 2.17 \times 10^{-8}\), so \(y_{10}\) agrees with \(e\) to 7 decimal places. \(e - y_{100} < \frac{102}{100.107^2} \approx 10^{-156}\), so \(y_{100}\) (which can be quickly computed on a programmable hand calculator) agrees with \(e\) to at least 155 decimal places. (In contrast, \(x_{100} = 2.704\cdots\) only agrees with \(e\) to one decimal place.)

(c) Since \(\frac{n(n+2)}{(n+1)^2} = \frac{n^2 + 2n}{n^2 + 2n + 1} < 1\), we have \(y_n \leq e < y_n + \frac{1}{n+1}\). We can write \(e = y_n + \frac{\theta_n}{n+1}\) for some \(\theta_n\), \(0 < \theta_n < 1\).

(d) Using (b) and \(y_{n+1} < e\), show that

\[
\frac{n}{n+1} < \theta_n < \frac{n^2 + 2n}{n^2 + 2n + 1}
\]

and thus \(\theta_n \to 1\) as \(n \to \infty\).

V.5.5.6. Continue to use the notation of V.5.5.3–V.5.5.5. Show that \(e\) is irrational as follows. Suppose \(e = \frac{m}{n}\). Then

\[
m = 1 + \frac{1}{2!} + \cdots + \frac{1}{n!} + \frac{\theta_n}{n!} n
\]

Multiply both sides by \(n!\) to conclude that \(\frac{\theta_n}{n!}\) is an integer. This is a contradiction since \(0 < \theta_n < 1\).

V.5.5.7. Let \(f : \mathbb{R} \to \mathbb{R}\) be a continuous function with \(f(x + y) = f(x) + f(y)\) for all \(x, y \in \mathbb{R}\).

(a) Let \(a = f(1)\). Show that \(f(n) = an\) for all \(n \in \mathbb{Z}\).

(b) Show that \(f(r) = ar\) for all \(r \in \mathbb{Q}\).

(c) Use continuity to show that \(f(x) = ax\) for all \(x \in \mathbb{R}\). Thus \(f\) is linear.

(d) The continuity hypothesis cannot be removed. (However, the continuity hypothesis can be relaxed to Lebesgue measurability; see ??.) \(\mathbb{R}\) is a vector space over \(\mathbb{Q}\) of dimension \(2^{\aleph_0}\). Let \(B\) be a basis for \(\mathbb{R}\) over \(\mathbb{Q}\). Then every function from \(B\) to \(\mathbb{R}\) extends to a \(\mathbb{Q}\)-linear (hence additive) function from \(\mathbb{R}\) to \(\mathbb{R}\). There are \(2^{2^{\aleph_0}}\) functions from \(B\) to \(\mathbb{R}\). But there are only \(2^{\aleph_0}\) \(\mathbb{R}\)-linear functions from \(\mathbb{R}\) to \(\mathbb{R}\). Using the Axiom of Choice, construct a discontinuous example.

570
V.5.5.8. Let \( g \) be a continuous function from \( \mathbb{R} \) to \( \mathbb{R} \) which is not identically 0 and with the property that \( g(x+y) = g(x)g(y) \) for all \( x, y \in \mathbb{R} \).

(a) Show that \( g \) takes only positive values. \([g(x) = g(x/2)^2]\)

(b) Show using Exercise V.5.5.7. that \( g(x) = b^x \) for all \( x \), where \( b = g(1) > 0 \). So \( g \) is an exponential function. \([\text{Consider } f(x) = \log g(x).]\)

(c) The continuity hypothesis cannot be removed here either: construct a discontinuous example.

(d) Prove similarly that a continuous function \( h : (0, \infty) \to \mathbb{R} \) which is not identically 0 and which has the property that \( h(xy) = h(x) + h(y) \) for all \( x, y > 0 \) must be a logarithm function.

V.5.5.9. Redevop the entire theory of exponential and logarithm functions by first defining \( \log x = \int_1^x \frac{1}{t} \, dt \) for \( x > 0 \) and then letting \( f(x) = e^x \) be the inverse function of \( \log x \).

V.5.5.10. Here is an alternate approach to proving differentiability of exponential functions. Let \( b > 1 \). Then \( g_b(x) = b^x \) is continuous on \( \mathbb{R} \) (??).

(a) If \( k \in \mathbb{R} \), \( \int_{-k}^{-1} b^t \, dt < b^{-k} \). Thus, if \( n \in \mathbb{N} \), \( \int_{-n}^{0} b^t \, dt < \sum_{k=0}^{n-1} b^{-k} = \frac{1-b^{-n}}{1-b^{-1}} \).

(b) Conclude that the improper integral \( \int_{-\infty}^{\infty} b^t \, dt \) converges, and hence \( \int_{-\infty}^{\infty} b^t \, dt \) converges for any \( x \). For \( x \in \mathbb{R} \), set \( f_b(x) = \int_{-\infty}^{x} b^t \, dt \). Show that \( f_b \) is differentiable and \( f'_b(x) = g_b(x) = b^x \) for all \( x \).

(c) By changing variables in the improper integral, show that \( f_b(x) = b^x \cdot f_b(0) \) for any \( x \). Thus \( g_b \) is infinitely differentiable, and \( g'_b = k_b g_b \), where \( \frac{1}{k_b} = \int_{-\infty}^{0} b^t \, dt \).

(d) If \( a, b > 1 \), apply the product rule to \( g_{ab} = g_a g_b \) to conclude that \( k_{ab} = k_a + k_b \). Prove that \( \ell(b) = k_b \) is continuous. Thus \( \ell \) is a logarithm function.

V.6. Trigonometric Functions

There is a dark secret behind the usual (precalculus) development of the theory of trigonometric functions: the entire theory is based on logical quicksand. All of trigonometry is based on the assumption that a circle has a finite “circumference” (“arc length”), which is never really defined; the number \( \pi \) is “defined” to be the ratio of the “circumference” to the diameter, and empirically calculated to be a little more than three. This assumption, while eminently believable and experimentally verifiable, is logically untenable, especially in the absence of a real definition of arc length. The notion of arc length can only be rigorously defined using limits (\( \)). See I.5.1.3.ff. for a fuller discussion.

In this section, we will give a rigorous and efficient treatment of trigonometric functions, the properties and identities they satisfy, and their calculus. We will actually define one of the inverse trigonometric functions first, an approach which may seem oddly backwards to readers familiar with the traditional approach, but which has the advantage of being logically impeccable.

V.6.1. Arc Length of a Circle and the Arcsine Function

We will consider the unit circle \( \{(x, y) : x^2 + y^2 = 1\} \) in \( \mathbb{R}^2 \).

V.6.1.1. PROPOSITION. The arc length of the circle is finite.

To prove this, consider the arc from \((0,1)\) to \((1,0)\), which is a quarter circle. By symmetry, it is enough to prove that this arc has finite length. We give two arguments.
V.6.1.2. First consider the arc from \((0, 1)\) to \((b, \sqrt{1-b^2})\) for \(0 < b < 1\). This arc is the graph of the continuously differentiable function \(y = \sqrt{1 - x^2}\) on \([0, b]\), so by ?? the arc length is finite and given by

\[
L(0, b) = \int_0^b \sqrt{1 + \frac{x^2}{1-x^2}} \, dx = \int_0^b \frac{1}{\sqrt{1-x^2}} \, dx
\]

Thus, for example, the arc from \((0, 1)\) to \((\frac{1}{2}, \frac{\sqrt{3}}{2})\) has finite length. Since this arc is congruent to the arc from \((\frac{1}{2}, \frac{\sqrt{3}}{2})\) to \((1, 0)\) (via interchanging \(x\) and \(y\)), this arc also has finite length, and the quarter circle is the union, hence also has finite length which we will call \(L(0, 1)\).

V.6.1.3. We give an alternate argument which gives a (rather crude) estimate of the length. As \(b \to 1\), \(L(0, b) \to L(0, 1)\) (??), and \(L(0, b)\) is increasing as \(b\) increases, so we only need show that \(\{L(0, b) : 0 \leq b < 1\}\) is bounded. But if \(0 < x < 1\), \(x^2 < x\), so \(\frac{1}{\sqrt{1-x^2}} < \frac{1}{\sqrt{1-x}}\), and

\[
L(0, b) = \int_0^b \frac{1}{\sqrt{1-x^2}} \, dx < \int_0^b \frac{1}{\sqrt{1-x}} \, dx = -2\sqrt{1-x}\bigg|_0^b = 2 - 2\sqrt{1-b} < 2
\]

for all \(b < 1\); thus \(L(0, 1)\) is finite and less than 2.

V.6.1.4. To get a lower bound for \(L(0, 1)\), since it is the length of an arc from \((0, 1)\) to \((1, 0)\), it must be greater than the straight-line distance between the points, which is \(\sqrt{2}\). A better lower bound, for example, is the sum of the distance from \((0, 1)\) to \((\frac{1}{2}, \frac{\sqrt{3}}{2})\) and the distance from \((\frac{1}{2}, \frac{\sqrt{3}}{2})\) to \((1, 0)\), which is \(\sqrt{\frac{3}{1}} + \sqrt{\frac{2}{1}} = 1.52688 \cdots\).

V.6.1.5. Definition. The number \(\pi\) is the arc length of half the unit circle.

By symmetry, \(\pi = 2L(0, 1)\). By the above estimates, we have \(2(\sqrt{\frac{3}{1}} + \sqrt{\frac{2}{1}}) = 3.05 \cdots < \pi < 4\). Of course, it is well known that the exact value of \(\pi\) is \(3.1415926535 \cdots\); we will discuss efficient ways to calculate \(\pi\) later (??).

V.6.1.6. Definition. If \(-1 \leq x \leq 1\), the \textit{arcsine} of \(x\) is the number

\[
\arcsin x = \int_0^x \frac{1}{\sqrt{1-t^2}} \, dt
\]

(If \(x = \pm 1\), this is an improper integral; we can define \(\arcsin 1 = L(0, 1) = \frac{\pi}{2}\) and \(\arcsin(-1) = -L(0, 1) = -\frac{\pi}{2}\).)

Thus \(\arcsin x\) is the antiderivative of \(\frac{1}{\sqrt{1-x^2}}\), i.e. by the Fundamental Theorem of Calculus, \(\frac{d}{dx} \left(\arcsin x\right) = \frac{1}{\sqrt{1-x^2}}\) for \(-1 < x < 1\). Thus \(f(x) = \arcsin x\) is continuous on \([-1, 1]\) and differentiable on \((-1, 1)\), and \(f'(x) = \frac{1}{\sqrt{1-x^2}} > 0\) on \((-1, 1)\), so \(\arcsin x\) is strictly increasing on \([-1, 1]\). We have \(\arcsin 0 = 0\) and \(\arcsin(-x) = -\arcsin x\) for \(-1 \leq x \leq 1\).

V.6.2. The Sine and Cosine Functions

Since \(f(x) = \arcsin x\) is continuous and strictly increasing on \([-1, 1]\), with range \([-\frac{\pi}{2}, \frac{\pi}{2}]\), it has a continuous inverse function \(g(x)\) on \([-\frac{\pi}{2}, \frac{\pi}{2}]\).
V.6.2.1. **Definition.** If \( x \in \left[ -\frac{\pi}{2}, \frac{\pi}{2} \right] \), the *sine* of \( x \) is \( \sin x = g(x) \), and the *cosine* of \( x \) is \( \cos x = \sqrt{1 - g(x)^2} = \sqrt{1 - (\sin x)^2} \).

We usually write \( \sin^2 x \) instead of \( (\sin x)^2 \), i.e. \( \cos^2 x = (\cos x)^2 \). Thus \( \sin^2 x + \cos^2 x = 1 \) for all \( x \in \left[ -\frac{\pi}{2}, \frac{\pi}{2} \right] \) (of course, \( \cos^2 x = (\cos x)^2 \)).

V.6.2.2. Since \( \arcsin(x) = \arcsin(x) \), \( \sin x \) has the same property:

\[
\sin(-x) = -\sin x \\
\cos(-x) = \cos x
\]

We have \( \sin 0 = 0, \cos 0 = 1, \sin(\frac{\pi}{2}) = 1, \cos(\frac{\pi}{2}) = 0, \sin(-\frac{\pi}{2}) = -1, \cos(-\frac{\pi}{2}) = 0. \)

V.6.2.3. From the Inverse Function Theorem, \( g(x) = \sin x \) is differentiable on \( \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \) and

\[
g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\sqrt{1 - \sin^2 x}} = \sqrt{1 - \sin^2 x} = \cos x
\]

and from the Chain Rule \( h(x) = \cos x \) is differentiable on \( \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \) and

\[
h'(x) = \frac{d}{dx} \left( \sqrt{1 - \sin^2 x} \right) = \frac{1}{2} \left( \sqrt{1 - \sin^2 x} \right)^{-1/2} (-2 \sin x \cos x)
\]

\[
= -\frac{\sin x \cos x}{\sqrt{1 - \sin^2 x}} = -\frac{\sin x \cos x}{\cos x} = -\sin x
\]

and thus we obtain:

V.6.2.4. **Theorem.** The functions \( \sin x \) and \( \cos x \) are infinitely differentiable on \( \left( -\frac{\pi}{2}, \frac{\pi}{2} \right) \), and

\[
\frac{d}{dx} (\sin x) = \cos x, \quad \frac{d}{dx} (\cos x) = -\sin x
\]

for \( -\frac{\pi}{2} < x < \frac{\pi}{2} \).

V.6.2.5. By the properties of inverse functions, if \( 0 \leq x \leq \frac{\pi}{2} \), \( \sin x = b \) means

\[
x = \arcsin b = L(0, b) = \int_0^b \frac{1}{\sqrt{1 - t^2}} dt
\]

so the point \( P \) on the circle (in the first quadrant) for which the arc from \( (0,1) \) to \( P \) has length \( x \) has first coordinate \( \sin x \). The second coordinate is then \( \sqrt{1 - \sin^2 x} = \cos x \), i.e. \( P = (\sin x, \cos x) \). It is more customary to measure arcs of the circle counterclockwise from \( (1,0) \), and then the coordinates of the point of arc length \( x \) are \( (\cos x, \sin x) \) (just apply the transformation interchanging \( x \) and \( y \)).
We extend the definition of \( \sin x \) and \( \cos x \) to all real numbers by taking \((\cos x, \sin x)\) to be the coordinates of the point of arc length \(x\) traveling counterclockwise around the circle from \((1,0)\) (clockwise a distance \(|x|\) if \(x < 0\)). Since the circumference of the circle is \(2\pi\), we obtain that \(\sin\) and \(\cos\) are periodic with period \(2\pi\): for all \(x\),
\[
\sin (x + 2\pi) = \sin x, \quad \cos (x + 2\pi) = \cos x
\]
It can in fact be easily checked that, for all \(x\),
\[
\sin(-x) = -\sin x, \quad \cos(-x) = \cos x
\]
\[
\sin \left(x + \frac{\pi}{2}\right) = \cos x, \quad \cos \left(x + \frac{\pi}{2}\right) = -\sin x
\]
since when the point \((a, b)\) is rotated a quarter circle counterclockwise, its coordinates become \((-b, a)\). Then replacing \(x\) by \(x - \frac{\pi}{2}\),
\[
\sin \left(x - \frac{\pi}{2}\right) = -\cos x, \quad \cos \left(x - \frac{\pi}{2}\right) = \sin x
\]
Repeating the process twice, we obtain
\[
\sin (x \pm \pi) = -\sin x, \quad \cos (x \pm \pi) = -\cos x
\]
for all \(x\). Of course, \(|\sin x| \leq 1\) and \(|\cos x| \leq 1\) for all \(x\).

**Theorem.** The functions \(\sin x\) and \(\cos x\) are infinitely differentiable on all of \(\mathbb{R}\), and
\[
\frac{d}{dx} (\sin x) = \cos x, \quad \frac{d}{dx} (\cos x) = -\sin x
\]
for all \(x \in \mathbb{R}\).

**Proof:** By V.6.2.6., the result holds on \((-\frac{\pi}{2}, \frac{\pi}{2})\). Then, since \(\sin x = \cos \left(x - \frac{\pi}{2}\right)\), \(\sin x\) is differentiable on \((0, \pi)\), and by the chain rule
\[
\frac{d}{dx} (\sin x) = \frac{d}{dx} \left(\cos \left(x - \frac{\pi}{2}\right)\right) = -\sin \left(x - \frac{\pi}{2}\right) = \cos x
\]
on \((0, \pi)\), and similarly \(\cos x\) is differentiable on \((0, \pi)\) with derivative \(-\sin x\). From \(\sin x = -\cos \left(x + \frac{\pi}{2}\right)\) we similarly get that \(\sin x\) is differentiable on \((-\pi, 0)\) with derivative \(\cos x\), and similarly for \(\cos x\). Thus the result holds on the interval \((-\pi, \pi)\). Since \(\sin (x \pm \pi) = -\sin x\) and \(\cos (x \pm \pi) = -\cos x\) for all \(x\), the result holds on \(((n-1)\pi, (n+1)\pi)\) for any \(n \in \mathbb{Z}\), and hence for the whole real line.

### V.6.3. Other Trigonometric Functions

**Definition.** If \(x \in \mathbb{R}\), define the *tangent*, *cotangent*, *secant*, and *cosecant* of \(x\), respectively denoted \(\tan x\), \(\cot x\), \(\sec x\), \(\csc x\), by
\[
\tan x = \frac{\sin x}{\cos x}, \quad \cot x = \frac{\cos x}{\sin x} = \frac{1}{\tan x}
\]
\[
\sec x = \frac{1}{\cos x}, \quad \csc x = \frac{1}{\sin x}
\]
V.6.3.2. These functions are defined wherever the denominators are nonzero. Thus the tangent and secant are defined except at $\pm \frac{\pi}{2} \pm n\pi$, and the cotangent and cosecant are defined except at (integer) multiples of $\pi$. We have

$$\lim_{x \to \frac{\pi}{2}^-} \tan x = +\infty, \quad \lim_{x \to \frac{\pi}{2}^+} \tan x = -\infty$$

$$\lim_{x \to \frac{\pi}{2}^-} \sec x = +\infty, \quad \lim_{x \to \frac{\pi}{2}^+} \sec x = +\infty$$

$$\lim_{x \to \pi^-} \cot x = -\infty, \quad \lim_{x \to 0^+} \cot x = +\infty$$

$$\lim_{x \to \pi^-} \csc x = +\infty, \quad \lim_{x \to 0^+} \csc x = +\infty$$

Since $|\sin x|$ and $|\cos x|$ are always $\leq 1$, we have $|\sec x| \geq 1$ and $|\csc x| \geq 1$ whenever defined.

The functions sec and csc are periodic with period $2\pi$. But since $\sin(x+\pi) = -\sin x$ and $\cos(x+\pi) = -\cos x$, the functions tan and cot are actually periodic with period $\pi$.

If $0 \leq x < \frac{\pi}{2}$, the number tan $x$ is the slope of the line through the origin and the point $(\cos x, \sin x)$ in the first quadrant whose arc length along the unit circle from $(1, 0)$ is $x$.

V.6.3.3. These functions, where defined, satisfy the identities

$$1 + \tan^2 x = \frac{\cos^2 x}{\cos^2 x} + \frac{\sin^2 x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

and similarly $1 + \cot^2 x = \csc^2 x$ (where, as usual, $\tan^2 x = (\tan x)^2$, etc.)

V.6.3.4. These functions are infinitely differentiable on their domains, and their derivatives can be calculated using the quotient rule and power rule:

$$\frac{d}{dx} (\tan x) = \frac{(\cos x)(\cos x) - (\sin x)(-\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

$$\frac{d}{dx} (\cot x) = \frac{(\sin x)(-\sin x) - (\cos x)(\cos x)}{\sin^2 x} = -\frac{1}{\sin^2 x} = -\csc^2 x$$

$$\frac{d}{dx} (\sec x) = \frac{d}{dx} ((\cos x)^{-1}) = (-1)(\cos x)^{-2}(-\sin x) = \frac{\sin x}{\cos^2 x} = \sec x \tan x$$

$$\frac{d}{dx} (\csc x) = \frac{d}{dx} ((\sin x)^{-1}) = (-1)(\sin x)^{-2}(\cos x) = -\frac{\cos x}{\sin^2 x} = -\csc x \cot x$$

V.6.4. The Inverse Trigonometric Functions

We have already defined an inverse function, arcsin $x$, for $\sin x$. Actually, if we start with the sine function and try to define an inverse, since $\sin x$ is periodic and not one-to-one, we must restrict it to an interval on which it is one-to-one; it is conventional to take the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Similarly, for $\cos x$ we conventionally take the interval $[0, \pi]$, which is mapped one-one onto $[-1, 1]$.

V.6.4.1. Definition. The inverse function to $\cos x$ with domain $[-1, 1]$ and range $[0, \pi]$ is called the arccosine function, denoted $\arccos x$. 575
V.6.4.2. Since the function \( g(x) = \cos x \) is continuous on \([0, \pi]\) and differentiable on \((0, \pi)\) with nonzero derivative, the function \( f(x) = \arccos x \) is continuous on \([-1, 1]\) and differentiable on \((-1, 1)\) with derivative given by the Inverse Function Theorem:

\[
\frac{d}{dx} \arccos x = f'(x) = \frac{1}{g'(f(x))} = -\frac{1}{\sin(\arccos x)}
\]

\[
= -\frac{1}{\sqrt{1 - \cos^2(\arccos x)}} = -\frac{1}{\sqrt{1 - x^2}}
\]

Actually, from \( \sin x = \cos \left( \frac{\pi}{2} - x \right) \) it follows that \( \arccos x = \frac{\pi}{2} - \arcsin x \).

V.6.4.3. Now consider \( \tan x \). The interval \((-\frac{\pi}{2}, \frac{\pi}{2})\) is one period of the tangent, and \( \tan x \) maps this interval one-one onto \((-\infty, \infty)\), so we conventionally take the inverse function for the tangent function to have this range:

V.6.4.4. Definition. The inverse function to \( \tan x \) with domain \((-\infty, \infty)\) and range \((-\frac{\pi}{2}, \frac{\pi}{2})\) is called the \textit{arctangent} function, denoted \( \arctan x \).

Since the function \( g(x) = \tan x \) is differentiable on \((-\frac{\pi}{2}, \frac{\pi}{2})\) with nonzero derivative, \( f(x) = \arctan x \) is differentiable on \((-\infty, \infty)\) with derivative

\[
\frac{d}{dx} \arctan x = f'(x) = \frac{1}{g'(f(x))} = \frac{1}{\sec^2(\arctan x)} = \frac{1}{1 + \tan^2(\arctan x)} = \frac{1}{1 + x^2}
\]

V.6.4.5. We can similarly define an \textit{arccotangent} function with domain \((-\infty, \infty)\), range \((0, \pi)\), and derivative \(-\frac{1}{1+x^2}\). Actually, \( \arccot x = \frac{\pi}{2} - \arctan x \) since \( \cot x = \tan \left( \frac{\pi}{2} - x \right) \).

We could also define arcsecant and arccosecant functions, but these are rarely used.

It is very common to write \( \sin^{-1} x \) for \( \arcsin x \), etc. This notation is consistent with the usual notation for inverse functions, but incompatible with the convention of writing \( \sin^2 x \) for \( (\sin x)^2 \), etc., and thus can be confusing. (Note that \( \sin^{-1} x \neq (\sin x)^{-1} \))

V.6.5. The Area of a Circle

V.6.5.1. We can use trigonometry to calculate the area of a circle. Actually, we should first define what we mean by “area” of a subset of the plane. This is not as easy as it sounds: the general problem is essentially the basic subject matter of measure theory. It would take us too far afield to try to treat this question here (cf. ()), so we will just assume that area is defined for sufficiently well-behaved subsets of the plane, and satisfies

(i) If \( f \) is a nonnegative continuous function on a closed bounded interval \([a, b]\), then the region \( R = \{(x, y) : a \leq x \leq b, 0 \leq y \leq f(x)\} \) has area \( A(R) = \int_a^b f(x) \, dx \).

(ii) If \( R \subseteq S \), then \( A(R) \leq A(S) \).

(iii) If \( R_1, \ldots, R_n \) are nonoverlapping regions (“nonoverlapping” means the intersection of any two is contained in the boundary of both), then

\[
A(R_1 \cup \cdots \cup R_n) = A(R_1) + \cdots + A(R_n).
\]
(iv) Congruent regions have the same area.

(There is actually a subtle technical problem with (iii), since the boundary of a region can have positive measure, but this will not cause difficulty if we consider only regions bounded by finitely many smooth curves.)

A circle (technically a disk) or any wedge of a circle has a well-defined area under these assumptions. If \( D \) is the unit disk, then \( A(D) = 2 \int_1^1 \sqrt{1-x^2} \, dx < 4 \). We want to calculate the exact value of \( A(D) \).

**V.6.5.2. Theorem.** \( A(D) = \pi \).

The simplest way to calculate this area is using vector calculus, but we will proceed in an alternate way. We have that \( A(D) = \pi p \) for some \( p > 0 \); we want to show that \( p = 1 \).

**V.6.5.3.** Define a function \( \phi : [0, 2\pi] \to \mathbb{R} \) by setting \( \phi(x) \) equal to the area of a wedge of a circle with arc length \( x \). Any two such wedges are congruent, so \( \phi(x) \) is well defined, and strictly increasing; we have \( \phi(0) = 0 \) and \( \phi(2\pi) = A(D) = \pi p \). For each \( n \in \mathbb{N} \), we have \( \phi \left( \frac{2\pi}{n} \right) = \frac{\pi p}{n} \) by (iii), i.e. \( \phi(x) = \frac{p}{2} x \) if \( x = \frac{2\pi}{n} \) for some \( n \). (Actually this formula holds for all \( x \in [0, 2\pi] \), although we will not need this fact for the Proof: it holds for \( x = \frac{2\pi m}{n} \) for \( m \leq n \) by (iii), and hence for all \( x \) since such numbers are dense in \( [0, 2\pi] \) and \( \phi \) is strictly increasing.)

![Figure V.8](image-url)

If \( 0 < x < \frac{\pi}{2} \), draw the picture in Figure 1, with \( OA = 1, OX = 1 \), and \( \angle AOX = x \), so \( X = (\cos x, \sin x) \).

Then \( BX = \sin x \) and \( AZ = \tan x \). Let \( T_1 \) be the triangle \( \triangle AOX \) and \( T_2 \) the triangle \( \triangle AOZ \). Then \( A(T_1) \leq \phi(x) \leq A(T_2) \), so if \( x_n = \frac{2\pi}{n} \) for \( n > 4 \) we have

\[
\frac{1}{2} \sin x_n \leq \frac{p}{2} x_n \leq \frac{1}{2} \tan x_n
\]

\[
\frac{\sin x_n}{x_n} \leq p \leq \frac{\tan x_n}{x_n}
\]

577
Since \( \lim_{n \to \infty} \frac{\sin x_n}{x_n} = \lim_{n \to \infty} \frac{\tan x_n}{x_n} = 1 \), we have \( p = 1 \). This completes the proof of the theorem.

**V.6.5.4.** It follows from the argument of the proof that \( \phi(x) = \frac{1}{2}x \) for all \( x \), giving an alternate geometric interpretation of the angle \( x \) as area rather than arc length. See Exercise () for another derivation of this formula.

**V.6.6. The Hyperbolic Functions**

The trigonometric functions are sometimes called the *circular functions*, for obvious reasons. We can define analogous functions related to a hyperbola which are formally similar to the circular functions; it turns out that the connections are much more than formal, although these deep connections are beyond the scope of this chapter (cf. X.1.1.8.).

There are actually two apparently totally different ways the hyperbolic functions can be defined, one using the hyperbola in the way the circle was used for the trigonometric functions, and the other using exponential functions. It is remarkable that the same functions are obtained by the two methods. We will describe both methods in this section.

**V.6.6.1.** We begin with the unit hyperbola \( y^2 - x^2 = 1 \). (It is more customary to use the hyperbola \( x^2 - y^2 = 1 \) with base point \((1,0)\), but our approach facilitates computations.) Since the upper half of the hyperbola has equation \( y = \sqrt{1 + x^2} \), in analogy with the formula defining \( \arcsin x \) we define the *hyperbolic arcsine* \( \arcsinh x \) by

\[
\arcsinh x = \int_0^x \frac{1}{\sqrt{1 + t^2}} \, dt
\]

Unfortunately, there is no interpretation of this as arc length in the hyperbolic case. (See Exercise () for a connection with area similar to the one for the trigonometric functions.)

Then \( \arcsinh x \) is defined for all \( x \) and is differentiable on \((-\infty, \infty)\) with derivative \( \frac{1}{\sqrt{1 + x^2}} \). We obviously have \( \arcsinh(-x) = -\arcsinh x \) for all \( x \).

**V.6.6.2.** Proposition. The range of \( \arcsinh x \) is \((-\infty, \infty)\).

Proof: Since \( \lim_{x \to +\infty} \frac{x}{\sqrt{1 + x^2}} = 1 \), there is an \( n \) such that \( \frac{1}{\sqrt{1 + x^2}} > \frac{1}{2x} \) for \( x > n \). Thus, if \( b > n \), we have

\[
\int_n^b \frac{1}{\sqrt{1 + x^2}} \, dx > \int_n^b \frac{1}{2x} \, dx = (\ln b - \ln n)
\]

which goes to infinity as \( b \to +\infty \). So \( \arcsinh x \) becomes arbitrarily large as \( x \to +\infty \). Since \( \arcsinh(-x) = -\arcsinh x \) and the range is an interval, it must be the whole real line.

**V.6.6.3.** So \( f(x) = \arcsinh x \) is strictly increasing and has an inverse function which we call the *hyperbolic sine*, denoted \( \sinh x \). By V.6.6.2., the domain of \( \sinh x \) is all of \( \mathbb{R} \). We then define the *hyperbolic cosine* \( \cosh x \) to be \( \sqrt{1 + \sinh^2 x} \), so \( \cosh^2 x - \sinh^2 x = 1 \) for all \( x \). We have \( \sinh(-x) = -\sinh x \) and \( \cosh(-x) = \cosh x \) for all \( x \), \( \sinh 0 = 0 \), and \( \cosh 0 = 1 \).
V.6.6.4. The function $g(x) = \sinh x$ is differentiable on $\mathbb{R}$, with derivative given by
\[
\frac{d}{dx} (\sinh x) = g'(x) = \frac{1}{\cosh^2 x} = \frac{1}{\sqrt{1 + \sinh^2 x}} = \cosh x
\]
and by the Chain Rule,
\[
\frac{d}{dx} (\cosh x) = \frac{d}{dx} \left( \sqrt{1 + \sinh^2 x} \right) = \frac{1}{2} (1 + \sinh^2 x)^{-1/2} (2 \sinh x \cosh x)
= \frac{\sinh x \cosh x}{\sqrt{1 + \sinh^2 x}} = \sinh x
\]
and thus $\sinh x$ and $\cosh x$ are infinitely differentiable.

The functions $\sinh x$ and $\cosh x$ thus behave formally in a similar way to $\sin x$ and $\cos x$, with occasional sign differences. The analogy only goes so far, however: $\sinh x$ and $\cosh x$ are not periodic and unbounded.

In fact, $\cosh x \geq 1$ for all $x$.

V.6.6.5. We can define analogs of the other trigonometric functions, called the hyperbolic tangent, cotangent, secant, and cosecant:
\[
\tanh x = \frac{\sinh x}{\cosh x}, \quad \coth x = \frac{\cosh x}{\sinh x} = \frac{1}{\tanh x}
\]
\[
\text{sech} x = \frac{1}{\cosh x}, \quad \text{csch} x = \frac{1}{\sinh x}
\]

It is useful to have a one-syllable pronunciation of the names of the hyperbolic functions, since the names “hyperbolic sine,” etc., are cumbersome. There is an obvious pronunciation for “cosh”, and also for the seldom-used “coth” and “sech”. The names “sinh” and “tanh” are normally pronounced “cinch” and “tanch” respectively. Your guess is as good as mine about how to pronounce “csch” (or “csc” for that matter!)

V.6.6.6. The domain of $\tanh x$ and $\text{sech} x$ is all of $\mathbb{R}$; the domain of $\coth x$ and $\text{csch} x$ is $\mathbb{R} \setminus \{0\}$. We have $\tanh 0 = 0$, $\tanh(-x) = -\tanh x$ and $|\tanh x| < 1$ for all $x$, and $\lim_{x \to \pm \infty} \tanh x = 1$, $\lim_{x \to -\infty} \tanh x = -1$. Similarly, $\text{sech} 0 = 1$, $\text{sech}(-x) = \text{sech} x$ and $|\text{sech} x| \leq 1$ for all $x$, and $\lim_{x \to \pm \infty} \text{sech} x = \lim_{x \to -\infty} \text{sech} x = 0$. They satisfy identities similar to the ones for trigonometric functions; for example:
\[
1 - \tanh^2 x = 1 - \frac{\sinh^2 x}{\cosh^2 x} = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x} = \text{sech}^2 x
\]

V.6.6.7. These hyperbolic functions are all infinitely differentiable on their domains, with derivatives given by standard rules:
\[
\frac{d}{dx} (\tanh x) = \frac{(\cosh x)(\cosh x) - (\sinh x)(\sinh x)}{\cosh^2 x} = \frac{1}{\cosh^2 x} = \text{sech}^2 x
\]
\[
\frac{d}{dx} (\coth x) = \frac{(\sinh x)(\sinh x) - (\cosh x)(\cosh x)}{\sinh^2 x} = -\frac{1}{\sinh^2 x} = -\text{csch}^2 x
\]
\[
\frac{d}{dx} (\text{sech} x) = \frac{d}{dx} \left((\cosh x)^{-1}\right) = -\frac{1}{\text{cosh}^2 x} = -\text{sech} x \tanh x
\]
\[
\frac{d}{dx} (\text{csch} x) = \frac{d}{dx} \left((\sinh x)^{-1}\right) = -\frac{1}{\text{cosh}^2 x} = -\text{csch} x \coth x
\]
V.6.6.8. Other inverse hyperbolic functions can be defined. For example, the one most often used, the inverse hyperbolic tangent \( f(x) = \text{arctanh} \, x \) is the inverse function of \( g(x) = \tanh x \), with domain \((-1, 1)\), range \((-\infty, \infty)\), and derivative

\[
\frac{d}{dx} \text{arctanh} \, x = f'(x) = \frac{1}{g'(f(x))} = \frac{1}{\text{sech}^2(\text{arctanh} \, x)} = \frac{1}{1 - \text{tanh}^2(\text{arctanh} \, x)} = \frac{1}{1 - x^2}
\]

Since \( \frac{1}{1-x} = \frac{1}{2} \frac{1}{1+x} + \frac{1}{2} \frac{1}{1-x} \) by partial fractions, the derivative of \( \text{arctanh} \, x \) is the same as the derivative of

\[
h(x) = \frac{1}{2} \ln(1+x) - \frac{1}{2} \ln(1-x) = \frac{1}{2} \ln \left( \frac{1+x}{1-x} \right)
\]

Since \( h(0) = \text{arctanh} \, 0 = 0 \), we have \( h(x) = \text{arctanh} \, x \) for all \( x \), i.e.

\[
\text{arctanh} \, x = \frac{1}{2} \ln(1+x) - \frac{1}{2} \ln(1-x) = \frac{1}{2} \ln \left( \frac{1+x}{1-x} \right), \quad (-1 < x < 1)
\]

There is therefore a close connection between inverse hyperbolic functions and logarithm functions, and hence the hyperbolic functions are actually closely related to exponential functions. Since the connection is not so easy to motivate, we will simply define some new functions and show that they are the same as the hyperbolic functions.

V.6.6.9. Definition. Define \( s(x) = \frac{e^x - e^{-x}}{2} \) and \( c(x) = \frac{e^x + e^{-x}}{2} \).

V.6.6.10. Theorem. We have \( s(x) = \sinh x \) and \( c(x) = \cosh x \) for all \( x \).

To prove this theorem, note that \( s(x) \) and \( c(x) \) are differentiable on all of \( \mathbb{R} \), and that \( s'(x) = c(x) \), \( c'(x) = s(x) \). Since \( c(x) > 0 \) for all \( x \), it follows that \( s(x) \) is strictly increasing on \( \mathbb{R} \). We also have that \( s(0) = 0, s(-x) = -s(x) \) for all \( x \), and \( \lim_{x \to +\infty} s(x) = +\infty \), so it follows that the range of \( s(x) \) is all of \( \mathbb{R} \). We have, for all \( x \),

\[
1 + (s(x))^2 = 1 + \frac{(e^x - e^{-x})^2}{4} = \frac{4 + e^{2x} - 2 + e^{-2x}}{4} = \frac{e^{2x} + 2 + e^{-2x}}{4} = \frac{(e^x + e^{-x})^2}{4} = (c(x))^2
\]

and thus \( c(x) = \sqrt{1 + (s(x))^2} \) for all \( x \).

Thus \( s(x) \) has an inverse function \( h(x) \), which is defined and differentiable on all of \( \mathbb{R} \), with derivative given by the Inverse Function Theorem:

\[
h'(x) = \frac{1}{s'(h(x))} = \frac{1}{c(h(x))} = \frac{1}{\sqrt{1 + (s(h(x)))^2}} = \frac{1}{\sqrt{1 + x^2}}
\]

so the derivative of \( h(x) \) is the same as the derivative of \( \text{arcsinh} \, x \). Since \( h(0) = 0 = \text{arcsinh} \, 0 \), we have \( h(x) = \text{arcsinh} \, x \) for all \( x \), and thus the inverse functions are also equal: \( s(x) = \sinh x \). The derivatives of these functions are then also equal: \( c(x) = \cosh x \). This completes the proof of the theorem.
V.6.6.11. From this, we get formulas for the other hyperbolic functions:

\[
\begin{align*}
\tanh x &= \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \\
\coth x &= \frac{\cosh x}{\sinh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}} \\
\text{sech} x &= \frac{1}{\cosh x} = \frac{2}{e^x + e^{-x}}, \\
\text{csch} x &= \frac{1}{\sinh x} = \frac{2}{e^x - e^{-x}}
\end{align*}
\]

See Exercise () for a direct algebraic derivation of the formula for \( \tanh x \).

V.6.6.12. Actually, the trigonometric functions also have similar formulas in terms of exponential functions, but the formulas involve complex numbers:

\[
\begin{align*}
\cos x &= \frac{e^{ix} + e^{-ix}}{2}, \\
\sin x &= \frac{e^{ix} - e^{-ix}}{2i}
\end{align*}
\]

(we will not attempt to explain these formulas here.) The rich and beautiful subject of complex analysis (calculus for functions on the complex numbers) reveals many deep and fascinating connections between trigonometric, hyperbolic, and exponential functions; see X.1.1.8.

V.6.7. Exercises

V.6.7.1. If \( 0 \leq x \leq \frac{\pi}{2} \), show that the function \( \phi(x) \) of the proof of V.6.5.2. is given by

\[
\phi(x) = \frac{1}{2} \cos x \sin x + \int_{\cos x}^{1} \frac{\sqrt{1-t^2}}{1-t^2} dt
\]

(see Figure ()). Using (), show that \( \phi'(x) \) is the constant function \( \frac{1}{2} \), so \( \phi(x) = \frac{1}{2} x \).

V.6.7.2. If \( x \geq 0 \), let \( \psi(x) \) be the area of the wedge bounded by the hyperbola \( x^2 - y^2 = 1 \), the line through the origin and \( (\cosh x, \sinh x) \), and the \( x \)-axis (Figure ()). Show that \( \psi(x) \) is given by

\[
\psi(x) = \frac{1}{2} \cosh x \sinh x - \int_{1}^{\cosh x} \sqrt{t^2 - 1} dt
\]

As in Exercise (), show that \( \psi'(x) \) is the constant function \( \frac{1}{2} \), so \( \psi(x) = \frac{1}{2} x \).

V.6.7.3. Derive the exponential formula for \( \tanh x \) from the formula \( \arctanh x = \frac{1}{2} \ln \left( \frac{1+x}{1-x} \right) \) by solving the equation

\[
2x = \ln \left( \frac{1 + y}{1 - y} \right)
\]

for \( y \).
V.6.7.4. [Bri98] Prove the sum formulas for sine and cosine

\[
\sin(\alpha + \beta) = \sin \alpha \cos \beta + \cos \alpha \sin \beta \\
\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta
\]

for \( \alpha, \beta \in \mathbb{R} \), as follows.

(a) Fix \( \beta \in \mathbb{R} \), and replace \( \alpha \) by a variable \( x \), and define

\[
f(x) = [\sin(x + \beta) - \sin x \cos \beta - \cos x \sin \beta]^2 + [\cos(x + \beta) - \cos x \cos \beta + \sin x \sin \beta]^2 .
\]

Then \( f \) is differentiable on \( \mathbb{R} \).

(b) Compute \( f' \) using the Chain Rule, and show that \( f' \) is identically 0. So \( f \) is constant.

(c) Since \( f(-\beta) = 0 \), \( f \) is identically 0. Conclude that the sum formulas hold.

(d) Prove the following identities from the above sum formulas, for every \( \alpha, \beta \in \mathbb{R} \) which make sense in the formulas:

\[
\sin(\alpha - \beta) = \sin \alpha \cos \beta - \cos \alpha \sin \beta \\
\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta \\
\sin \alpha \cos \beta = \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)] \\
\cos \alpha \cos \beta = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)] \\
\sin \alpha \sin \beta = \frac{1}{2} [\cos(\alpha - \beta) - \cos(\alpha + \beta)] \\
\sin 2\alpha = 2 \sin \alpha \cos \alpha \\
\cos 2\alpha = \cos^2 \alpha - \sin^2 \alpha = 2 \cos^2 \alpha - 1 = 1 - 2 \sin^2 \alpha \\
\cos^2 \frac{\alpha}{2} = \frac{1 + \cos \alpha}{2} \\
\sin^2 \frac{\alpha}{2} = \frac{1 - \cos \alpha}{2} \\
\tan(\alpha + \beta) = \frac{\tan \alpha + \tan \beta}{1 - \tan \alpha \tan \beta}
\]
V.7. Big-O and Little-o Notation

In this section, we introduce some notation which is commonly used to describe the asymptotic behavior of functions or sequences, the big-O and little-o notation. Although some care must be exercised in using this notation, it is very helpful in simplifying formulas and arguments containing constants whose exact value is unimportant or functions which become negligible as a limit is approached, in addition to giving a clean and precise description of the essential quantitative behavior of limits of functions. Not only does use of this system often simplify arguments in form, but it also sometimes helps conceptually in that the essential ideas of the argument are clarified and highlighted; in addition, arguments written this way are usually closer in spirit to the way the founding fathers of calculus phrased their work (although our arguments have a much higher level of rigor than those of the early days of the subject.) In fact, a version of the little-o notation was used by Newton (it is unclear whether he invented it), who wrote:

“When I am investigating a truth or the solution to a Probleme I use all sorts of approximations and neglect to write down the letter o, but when I am demonstrating a Proposition I always write down the letter o & proceed exactly by the rules of Geometry.”

Isaac Newton

DEFINITION. Let \( a \) be an extended real number, and \( f \) and \( g \) functions defined in a deleted neighborhood of \( a \) (this does not preclude the possibility that they are also defined at \( a \) if \( a \in \mathbb{R} \)).

(i) If \( a \in \mathbb{R} \), then \( g(x) = O(f(x)) \) as \( x \to a \) if there exist constants \( M \geq 0 \) and \( \delta > 0 \) such that \( |g(x)| \leq M|f(x)| \) whenever \( 0 < |x - a| < \delta \).

(ii) If \( a \in \mathbb{R} \), then \( g(x) = o(f(x)) \) as \( x \to a \) if for every \( \epsilon > 0 \) there is a \( \delta > 0 \) such that \( |g(x)| \leq \epsilon |f(x)| \) whenever \( 0 < |x - a| < \delta \).

If \( a = \pm \infty \), the \( \delta \) and \( 0 < |x - a| < \delta \) must be replaced in the usual way by \( B \) and \( x \neq B \) for \( a = +\infty \), and by \( A \) and \( x < A \) if \( a = -\infty \).

An analogous definition can be made for one-sided approach if \( f \) and \( g \) are defined in a deleted one-sided neighborhood of \( a \), and for sequences. Thus we may write “\( g(x) = O(f(x)) \) as \( x \to a^+ \)” or “\( y_n = o(x_n) \) as \( n \to \infty \)”.

In the definition of \( O \), we can clearly assume that \( M > 0 \) since there is no harm in increasing \( M \).

The \( O \) and \( o \) stand for “order.” The expression \( g(x) = O(f(x)) \) as \( x \to a \) roughly means “\( |g(x)| \) is of the same order as (or not essentially larger than) \( |f(x)| \) for \( x \) near \( a \)”;

\( g(x) = o(f(x)) \) roughly means “\( |g(x)| \) is of smaller order than (or negligibly small compared to) \( |f(x)| \) for \( x \) near \( a \)”.

EXAMPLES.

(i) \( g(x) = O(f(x)) \) as \( x \to a \) roughly means that \( \frac{g(x)}{f(x)} \) is bounded in a neighborhood of \( a \).” In particular, \( g(x) = O(1) \) as \( x \to a \) exactly means that \( g \) is bounded in a neighborhood of \( a \). If \( \lim_{x \to a} \frac{g(x)}{f(x)} \) exists, then \( g(x) = O(f(x)) \) as \( x \to a \). But note that \( g(x) = O(f(x)) \) as \( x \to a \) does not imply that \( \lim_{x \to a} \frac{g(x)}{f(x)} \) exists; for example, \( \sin \frac{1}{x} = O(1) \) as \( x \to 0 \).
Proposition. Here are some rules that are valid:

(ii) \( g(x) = o(f(x)) \) as \( x \to a \) roughly means that \( \lim_{x \to a} \frac{g(x)}{f(x)} = 0 \); if \( \lim_{x \to a} \frac{g(x)}{f(x)} = 0 \), then \( g(x) = o(f(x)) \) as \( x \to a \), but not quite conversely: \( x \sin \frac{1}{x} = o(\sin \frac{1}{x}) \) as \( x \to 0 \). But \( g(x) = o(1) \) as \( x \to a \) exactly means that \( \lim_{x \to a} g(x) = 0 \).

(iii) \( x^m = O(x^n) \) as \( x \to 0 \) if and only if \( m \geq n; x^m = o(x^n) \) as \( x \to 0 \) if and only if \( m > n \). But \( x^m = O(x^n) \) as \( x \to +\infty \) if and only if \( m \leq n \), and \( x^m = o(x^n) \) as \( x \to +\infty \) if and only if \( m < n \).

Proposition. Let \( a \) be an extended real number, and \( f, g, \) and \( h \) real-valued functions defined in a neighborhood of \( a \).

(i) If \( g(x) = o(f(x)) \) as \( x \to a \), then \( g(x) = O(f(x)) \) as \( x \to a \).

(ii) If \( g(x) = O(f(x)) \) and \( h(x) = O(g(x)) \) as \( x \to a \), then \( h(x) = O(f(x)) \) as \( x \to a \).

(iii) If \( g(x) = o(f(x)) \) and \( h(x) = O(g(x)) \) as \( x \to a \), then \( h(x) = o(f(x)) \) as \( x \to a \).

(iv) If \( g(x) = O(f(x)) \) and \( h(x) = o(g(x)) \) as \( x \to a \), then \( h(x) = o(f(x)) \) as \( x \to a \).

In particular, if \( g(x) = O(f(x)) \) and \( f(x) = O(1) \), then \( g(x) = O(1) \); if \( g(x) = o(f(x)) \) and \( f(x) = O(1) \), or \( g(x) = O(f(x)) \) and \( f(x) = o(1) \), then \( g(x) = o(1) \).

Proof: (i) is obvious from the definitions.

(ii): By hypothesis, there is an \( M > 0 \) and \( \delta_1 > 0 \) such that \( |g(x)| \leq M|f(x)| \) for \( 0 < |x - a| < \delta_1 \), and an \( N > 0 \) and \( \delta_2 > 0 \) such that \( |h(x)| \leq N|g(x)| \) for \( 0 < |x - a| < \delta_2 \). Set \( \delta = \min(\delta_1, \delta_2) \). Then, if \( 0 < |x - a| < \delta \), we have

\[
|h(x)| \leq N|g(x)| \leq MN|f(x)|.
\]

(iii): Let \( \epsilon > 0 \). By hypothesis, there is an \( M > 0 \) and \( \delta_1 > 0 \) such that \( |h(x)| \leq M|g(x)| \) for \( 0 < |x - a| < \delta_1 \). There is then a \( \delta_2 > 0 \) such that \( |g(x)| \leq \frac{\epsilon}{M} |f(x)| \) for \( 0 < |x - a| < \delta_2 \). Set \( \delta = \min(\delta_1, \delta_2) \). Then, for \( 0 < |x - a| < \delta \),

\[
|h(x)| \leq M|g(x)| \leq M \cdot \frac{\epsilon}{M} |f(x)| = \epsilon |f(x)|.
\]

(iv) is almost identical to (iii), and is left to the reader.

We often just write “\( O(f(x)) \)” or “\( o(f(x)) \)” (along with “\( x \to a \)” if it is not understood from the context) to denote some function \( g(x) \) which satisfies \( g(x) = O(f(x)) \) or \( g(x) = o(f(x)) \) (\( x \to a \)).

Note that the expressions \( O(f(x)) \) and \( o(f(x)) \) do not necessarily represent specific functions of \( x \) (or, more precisely, different uses of \( O(f(x)) \) or \( o(f(x)) \) may represent different specific functions of \( x \)). Consequently, these expressions do not necessarily follow the usual algebraic rules. For example, from

\[
g(x) + o(f(x)) = h(x) + o(f(x))
\]

it cannot be concluded that \( g(x) = h(x) \); it only follows that \( g(x) = h(x) + o(f(x)) \) (i.e. \( g(x) - h(x) = o(f(x)) \)).

Here are some rules that are valid:

Proposition.

(i) If \( g_1(x) = O(f(x)) \) and \( g_2(x) = O(f(x)) \) as \( x \to a \), then \( g_1(x) + g_2(x) = O(f(x)) \) as \( x \to a \).
Proof: of the last two results is that $O$ and in particular is nonzero, so there is a

Proof: \[ \lim_{x \to a} g(x) \] and \[ k(x) \] and \[ h(x) \] and \[ O(g(x)) = O(f(x)g(x)), \] \[ o(g(x)) = o(f(x)g(x)) \]

One of the simplest and most common applications of the last two results is that $K \cdot O(f(x)) = O(f(x))$ and $K \cdot o(f(x)) = o(f(x))$ for any constant $K$.

PROOF: We prove only (iv); the other proofs are very similar (but slightly easier). Let $\epsilon > 0$. By hypothesis, there is an $M > 0$ and $\delta_2 > 0$ such that $|g_2(x)| \leq M |f_2(x)|$ for $0 < |x - a| < \delta_2$. Then there exists a $\delta_1 > 0$ such that $|g_1(x)| \leq \frac{\epsilon}{M} |f_1(x)|$ for $0 < |x - a| < \delta_1$. Set $\delta = \min(\delta_1, \delta_2)$. Then, for $0 < |x - a| < \delta$,

$$|g_1(x)g_2(x)| \leq \frac{\epsilon}{M} |f_1(x)| \cdot M |f_2(x)| = \epsilon |f_1(x)f_2(x)|.$$  

$\blacksquare$

The next result is fairly special, but is useful in dealing with quotients. It is awkward to even state this result without using the $o$-notation (try it!)

PROPOSITION. Let $a$ be an extended real number, and $g$, $h$, and $k$ real-valued functions defined on a deleted neighborhood of $a$, with $h(x) \neq 0$ for all $x$ in a deleted neighborhood of $a$. Suppose $g(x) = O(1), \frac{1}{h(x)} = O(1)$, and $k(x) = o(1)$ as $x \to a$ (i.e. $g$ is bounded and $h$ is bounded away from 0 in a deleted neighborhood of $a$, and $\lim_{x \to a} k(x) = 0$). Then

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} - \frac{g(x)}{h(x)} \cdot \frac{k(x)}{h(x)^2} + O((k(x))^2) \text{ as } x \to a.$$  

PROOF: The proof uses the algebraic formula

$$\frac{1}{1 + t} = 1 - t + \frac{t^2}{1 + t}$$

valid for $t \neq -1$, which is easily proved by multiplying both sides by $1 + t$. (This formula is a special case of the formula of $\cdot$.)

Under the hypotheses, there is an $M > 0$ and $\delta_1 > 0$ such that $|h(x)| \geq M$ for $0 < |x - a| < \delta_1$, and there is a $\delta_2 > 0$ such that $|k(x)| < M/2$ for $0 < |x - a| < \delta_2$. There is also an $N \geq 0$ and $\delta_3 > 0$ such that $|g(x)| \leq N$ for $0 < |x - a| < \delta_3$. Set $\delta = \min(\delta_1, \delta_2, \delta_3)$. Then if $0 < |x - a| < \delta$, $|h(x) + k(x)| > M/2$ and in particular is nonzero, so $\frac{g(x)}{h(x) + k(x)}$ is defined in a deleted neighborhood of 0. Also, if $0 < |x - a| < \delta$,

$$\frac{k(x)}{h(x)} < 1/2, \text{ so } \frac{k(x)}{h(x)} \neq -1.$$  

Thus, for $0 < |x - a| < \delta$, we have

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} \cdot \frac{1}{1 + \frac{k(x)}{h(x)}}.$$  

585
The last term is \( O([k(x)]^2) \) as \( x \to a \) since

\[
\left| \frac{g(x)}{h(x)} - \frac{g(x)}{[h(x)]^2} k(x) + \frac{g(x)}{[h(x)]^2(h(x) + k(x))} [k(x)]^2 \right| \leq \frac{N}{M^2 \left( \frac{M}{2} \right)} = \frac{2N}{M^3}
\]

for \( 0 < |x - a| < \delta \).  

**Corollary.** Let \( a \) be an extended real number, and \( f, g, h, \) and \( k \) real-valued functions defined on a deleted neighborhood of \( a \), with \( h(x) \neq 0 \) for all \( x \) in a deleted neighborhood of \( a \). Suppose \( g(x) = O(1) \), 

\[
\frac{1}{h(x)} = O(1),
\]

and \( f(x) = o(1) \) as \( x \to a \) (i.e. \( g \) is bounded and \( h \) is bounded away from 0 in a deleted neighborhood of \( a \), and \( \lim_{x \to a} f(x) = 0 \)).

(i) If \( k(x) = O(f(x)) \) as \( x \to a \), then

\[
\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} + O(f(x)) \text{ as } x \to a.
\]

(ii) If \( k(x) = o(f(x)) \) as \( x \to a \), then

\[
\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} + o(f(x)) \text{ as } x \to a.
\]

This result is sometimes abbreviated by the following curious algebraic formula: “Under the hypotheses on \( f, g, \) and \( h \),

\[
\frac{g(x)}{h(x) + o(f(x))} = \frac{g(x)}{h(x)} + o(f(x))
\]

(and similarly for \( O(f(x)) \).) The explanation of why this expression does not violate the rules of algebra is that the \( o(f(x)) \)'s on the two sides of the equation stand for (usually) different functions of \( x \).

**Proof:** First observe that \( \frac{g(x)}{[h(x)]^2} \) is bounded in a deleted neighborhood of \( a \). Using the notation of the proof of **V.7**, it is bounded by \( \frac{N}{M^2} \) for \( 0 < |x - a| < \delta \). Thus the second term is \( O(k(x)) \), which is \( O(f(x)) \) if \( k(x) = O(f(x)) \), and \( o(f(x)) \) if \( k(x) = o(f(x)) \). Also, the third term is \( O([k(x)]^2) = o(k(x)) = o(f(x)) \) in either part since \( k(x) = o(1) \).

The \( o \) is probably used more than the \( O \), since \( o \) denotes a function which becomes negligible and thus is very useful in limit or differentiability arguments (although the \( O \), if used carefully, gives a more precise description of asymptotic behavior). The next result, an almost immediate consequence of the definition of differentiability, is probably the most important example of use of the little-\( o \) notation:
Proposition. Let $f$ be a real-valued function defined in a neighborhood of $a \in \mathbb{R}$. Then $f$ is differentiable at $a$ if and only if there is an $m \in \mathbb{R}$ such that
\[ f(x) = f(a) + m(x-a) + o(x-a) \text{ as } x \to a \]
(or, more precisely, $f(x) - f(a) - m(x-a) = o(x-a)$ as $x \to a$). Such an $m$, if it exists, is unique and equals $f'(a)$.

As an example of use of the little-$o$ notation in simplifying arguments, here is a proof of the Product Rule (\ref{product-rule}). Note that this proof shows how the product rule follows naturally from the algebraic rule for multiplying multinomials. In addition, it avoids having to use the trick/technique of adding and subtracting something from an expression which may not be obvious, and also avoids explicitly using the result that a differentiable function is continuous.

Proof: Let $f$ and $g$ be real-valued functions defined in a neighborhood of $a \in \mathbb{R}$, and differentiable at $a$. Then
\[ f(x) = f(a) + f'(a)(x-a) + o(x-a) \]
and
\[ g(x) = g(a) + g'(a)(x-a) + o(x-a) \]
in a neighborhood of $a$. So, in a neighborhood of $a$,
\[ f(x)g(x) = [f(a) + f'(a)(x-a) + o(x-a)][g(a) + g'(a)(x-a) + o(x-a)] \]
\[ = f(a)g(a) + f'(a)g(a)(x-a) + f(a)g'(a)(x-a) + f'(a)g'(a)(x-a)^2 + g'(a)(x-a)g(x-a) + o(x-a) \cdot o(x-a) \]
\[ = f(a)g(a) + [f'(a)g(a) + f(a)g'(a)](x-a) \]
\[ + [g(a)g(x-a) + f'(a)g'(a)(x-a)^2 + g'(a)(x-a)g(x-a) + f(a)g(x-a) + f'(a)(x-a)g(x-a) + o(x-a) \cdot o(x-a)] \]
The expressions in the last brackets are all $o(x-a)$ since $f(a)$, $g(a)$, $f'(a)g(a)$, $f'(a)(x-a)$, $g'(a)(x-a)$, and $o(x-a)$ are all $O(1)$ as $x \to a$, and $(x-a)^2 = o(x-a)$ as $x \to a$. Thus
\[ f(x)g(x) = f(a)g(a) + [f'(a)g(a) + f(a)g'(a)](x-a) + o(x-a) \]
in a neighborhood of $a$, and so by \ref{v.7}, $fg$ is differentiable at $a$ and
\[ (fg)'(a) = f'(a)g(a) + f(a)g'(a) \cdot \]

We now give a proof of the Quotient Rule using the $o$-notation. This proof is not as clean or satisfying as the above proof of the Product Rule, but gives a better idea of where the Quotient Rule comes from algebraically than the usual proof, and also avoids the “adding and subtracting” trick/technique. It does, however, use in one place the fact that a differentiable function is continuous.

Proof: Let $f$ and $g$ be defined in a neighborhood of $a \in \mathbb{R}$ and differentiable at $a$, with $g(a) \neq 0$. Then, since $f$ and $g$ are continuous at $a$, there is a $\delta_1 > 0$ such that $|f(x)| < |f(a)| + 1$ for $|x-a| < \delta_1$, and there is $\delta_2 > 0$ such that $|g(x)| > \frac{|g(a)|}{2}$ for $|x-a| < \delta_2$. Thus $g$ is nonzero in a neighborhood of $a$, so $\frac{f}{g}$ is defined in a neighborhood of $a$, and $f$ and $\frac{1}{g}$ are bounded in a neighborhood of $a$. By \ref{v.7}, we can write
\[ f(x) = f(a) + f'(a)(x-a) + o(x-a) \]
and
\[ g(x) = g(a) + g'(a)(x-a) + o(x-a) \]
in a neighborhood of $a$, so
\[ \frac{f(x)}{g(x)} = \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a) + g'(a)(x-a) + o(x-a)} \]
First applying V.7(ii), and then applying V.7., we obtain
\[
\frac{f(x)}{g(x)} = \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a) + g'(a)(x-a) + o(x-a)}
\]
\[
= \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a) + g'(a)(x-a)} + o(x-a)
\]
\[
= \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a)}
\]
\[
- \frac{f(a) + f'(a)(x-a) + o(x-a)}{[g(a)]^2} g'(a)(x-a) + O([x-a]^2) + o(x-a)
\]
\[
= \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a)}
\]
\[
- \frac{f(a) + f'(a)(x-a) + o(x-a)}{[g(a)]^2} g'(a)(x-a) + o(x-a)
\].

Expanding and rearranging terms, we obtain:
\[
\frac{f(x)}{g(x)} = \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a) + g'(a)(x-a)} + \frac{f(a)g'(a)}{[g(a)]^2} (x-a)
\]
\[
- \frac{f'(a)g'(a)}{[g(a)]^2} (x-a)^2 - \frac{o(x-a)}{[g(a)]^2} (x-a) + o(x-a)
\]
\[
= \frac{f(a)}{g(a)} + \left[ \frac{f'(a)}{g(a)} - \frac{f(a)g'(a)}{[g(a)]^2} \right] (x-a)
\]
\[
+ \left[ \frac{o(x-a)}{g(a)} - \frac{f'(a)g'(a)}{[g(a)]^2} (x-a)^2 - \frac{o(x-a)}{[g(a)]^2} (x-a) + o(x-a) \right]
\].

All the terms in the last brackets are o(x-a), so in a neighborhood of a we have
\[
\frac{f(x)}{g(x)} = \frac{f(a)}{g(a)} + \left[ \frac{f'(a)}{g(a)} - \frac{f(a)g'(a)}{[g(a)]^2} \right] (x-a) + o(x-a)
\].

Thus by V.7., \( \frac{f}{g} \) is differentiable at a and
\[
\left( \frac{f}{g} \right)'(a) = \frac{f'(a)}{g(a)} - \frac{f(a)g'(a)}{[g(a)]^2} = \frac{g(a)f'(a) - f(a)g'(a)}{[g(a)]^2}.
\]

V.7.1. Exercises

1. Give careful definitions of (1) “g(x) = O(f(x)) as x → +∞” (2) “g(x) = O(f(x)) as x → a+” (3) “y_n = O(x_n) as n → ∞”, and similarly for o, and give versions of all the results of this section for these notions.
V.8. The Mean Value Theorem

“The Mean Value Theorem is the midwife of calculus – not very important or glamorous by itself, but often helping to deliver other theorems that are of major significance.”

E. Purcell and D. Varberg

The Mean Value Theorem (MVT) is one of the major theoretical results of calculus. It is the primary tool allowing information about a differentiable function to be obtained from properties of its derivative.

V.8.1. Rolle’s Theorem and the Mean Value Theorem

V.8.1.1. Theorem. [Mean Value Theorem] Let \( a, b \in \mathbb{R}, a < b \). Let \( f \) be a function which is continuous on \([a, b] \) and differentiable on \((a, b)\). Then there is at least one \( c \in (a, b) \) for which

\[
f(b) - f(a) = f'(c)(b - a).
\]

Alternatively, there is at least one \( c \in (a, b) \) with

\[
f'(c) = \frac{f(b) - f(a)}{b - a}.
\]

A geometric interpretation is that there is at least one point on the graph between \( a \) and \( b \) at which the tangent line is parallel to the secant line between \((a, f(a))\) and \((b, f(b))\) (Figure V.9).

V.8.1.2. The number \( f'(c) \) can be interpreted as the instantaneous rate of change of \( f \) at \( c \), while the number \( \frac{f(b) - f(a)}{b - a} \) is the average rate of change of \( f \) between \( a \) and \( b \). Thus, for example, if a driver’s average speed is 60 MPH over a period of time, there must be at least one instant of time during the period at which his instantaneous speed is exactly 60 MPH.

V.8.1.3. Three important points to note about the conclusion of the MVT:

(i) The \( c \) depends on the numbers \( a \) and \( b \), as well as on the function \( f \).

(ii) For a given \( f, a, \) and \( b \), there may be more than one \( c \), even infinitely many, satisfying the conclusion. (In Figure V.9 there are five, denoted \( c_1 - c_5 \).) Indeed, in the extreme case where \( f \) is linear, every \( c \in (a, b) \) works.

(iii) In practice it may be difficult if not impossible to explicitly find a \( c \) that works for a given \( a \) and \( b \), even for relatively well-behaved \( f \). The MVT is a prime example of a pure existence theorem, in which something is shown to exist without any clue as to how to find a specific example. As we will see, however, just the knowledge that such a \( c \) exists will be of great importance in obtaining conclusions about the function \( f \).
Rolle’s Theorem

We will give the standard modern proof of the MVT, due to O. Bonnet in 1868. Cauchy previously gave a more complicated and flawed proof (cf. [Bre07]). The essential step in Bonnet’s proof is to prove a special case, which is important enough in its own right to have a standard name:

V.8.1.4. **Theorem. [Rolle’s Theorem]** Let \( a, b \in \mathbb{R}, a < b \), and let \( f \) be continuous on \([a, b]\) and differentiable on \((a, b)\). If \( f(a) = f(b) \), then there is at least one \( c \in (a, b) \) with \( f'(c) = 0 \).

It is ironic that a theorem of calculus should be named after Rolle, since according to [Bal60, p. 318], Rolle, a contemporary of Newton and Leibniz, believed that calculus was “nothing but a collection of ingenious fallacies.” (It has been suggested that he later changed his mind.) Many references claim that Rolle first stated and proved this theorem; however, Rolle’s statement and proof only referred to polynomials and did not use any explicit calculus. It is unclear who first stated the result in its modern form. See the chapter by J. Barrow-Green in [RS09] for the full story. See also Exercise V.8.6.11.

**Proof:** Since \( f \) is continuous on a closed bounded interval, it attains a maximum \( M \) and a minimum \( m \) on \([a, b]\) (Max-Min Theorem). Then, since \( m \leq f(a) \leq M \), at least one of the following is true:

(i) \( M > f(a) \).

(ii) \( m < f(a) \).
(iii) \( f \) is constant on \([a, b] \).

In case (i), there is a \( c \in (a, b) \) with \( f(c) = M \). Then \( f'(c) = 0 \) by V.3.6.3.

In case (ii), there is a \( c \in (a, b) \) with \( f(c) = m \). Again by V.3.6.3. we have \( f'(c) = 0 \).

And in case (iii), \( f'(c) = 0 \) for all \( c \in (a, b) \).

Before going on to the MVT, we record some simple consequences of Rolle’s theorem which are basically just restatements or special cases of the theorem.

**V.8.1.5. Corollary.** Let \( a, b \in \mathbb{R}, a < b \), and let \( f \) be continuous on \([a, b]\) and differentiable on \((a, b)\).

If \( f'(x) \neq 0 \) for all \( x \in (a, b) \), then \( f(a) \neq f(b) \).

The next corollary is useful in locating the zeros of a function such as a polynomial:

**V.8.1.6. Corollary.** Let \( f \) be differentiable on an interval \( I \). Then between any two zeros of \( f \) in \( I \) there is at least one zero of \( f' \).

### The Mean Value Theorem

We now turn to the MVT. This theorem was first stated by Lagrange in (). We will actually prove the following more general theorem, due to Cauchy (with a flawed proof) in 1821; the usual MVT is the special case \( g(x) = x \).

**V.8.1.7. Theorem.** [Cauchy Mean Value Theorem] Let \( a, b \in \mathbb{R}, a < b \). Let \( f \) and \( g \) be functions which are continuous on \([a, b]\) and differentiable on \((a, b)\). Then there is at least one \( c \in (a, b) \) for which

\[
g'(c)[f(b) - f(a)] = f'(c)[g(b) - g(a)] .
\]

**Proof:** For \( x \in [a, b] \), set

\[
h(x) = f(x)[g(b) - g(a)] - g(x)[f(b) - f(a)] .
\]

Then \( h \) is continuous on \([a, b]\) and differentiable on \((a, b)\), and \( h(a) = h(b) \). By V.8.1.4. there is a \( c \in (a, b) \) such that

\[
0 = h'(c) = f'(c)[g(b) - g(a)] - g'(c)[f(b) - f(a)] .
\]

The following special case is commonly stated as Cauchy’s Mean Value Theorem:
**V.8.1.8.** Corollary. Let \( a, b \in \mathbb{R}, a < b \). Let \( f \) and \( g \) be functions which are continuous on \([a, b]\) and differentiable on \((a, b)\), such that \( g'(x) \neq 0 \) for all \( x \in (a, b) \). Then \( g(a) \neq g(b) \), and there is at least one \( c \in (a, b) \) for which

\[
\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}.
\]

Proof: The first conclusion follows from V.8.1.5., and then the second conclusion is V.8.1.7.

**V.8.1.9.** This corollary has a geometric interpretation. If \( \phi(t) = (g(t), f(t)) \), then \( \phi \) is the parametrization of a smooth curve in \( \mathbb{R}^2 \). The slope of the line through the endpoints of the curve is

\[
\frac{f(b) - f(a)}{g(b) - g(a)}
\]

and the slope of the tangent line at the point \( \phi(t) \) is \( \frac{f'(t)}{g'(t)} \). Cauchy’s MVT says that there is at least one point on the curve where the tangent line is parallel to the line through the two endpoints. (The requirement that \( g' \neq 0 \) on \((a, b)\) insures that this curve has no vertical tangents, and it turns out to actually be the graph of a differentiable function (cf. V.8.2.14.). The geometric result is true more generally as long as the endpoints are distinct and \( f' \) and \( g' \) do not simultaneously vanish (Exercise V.8.6.1.).)

**V.8.2. Applications of the Mean Value Theorem**

There are many important theoretical consequences of the MVT. We begin with a crucial one:

**V.8.2.1.** Corollary. Let \( a, b \in \mathbb{R}, a < b \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( f'(x) \geq 0 \) for all \( x \in (a, b) \), then \( f \) is nondecreasing on \([a, b]\), and in particular \( f(a) \leq f(b) \). If \( f' \) is not identically 0 on \((a, b)\), then \( f(a) < f(b) \).

Proof: If \( a \leq x_1 < x_2 \leq b \), applying the MVT to the interval \([x_1, x_2]\) we have

\[
f(x_2) - f(x_1) = f'(c)(x_2 - x_1) \geq 0
\]

for some \( c \in (x_1, x_2) \). Since \( f \) is nondecreasing on \([a, b]\), if \( f(a) = f(b) \) we must have \( f \) constant on \([a, b]\), and hence \( f' \) is identically 0.

Similarly, we have:

**V.8.2.2.** Corollary. Let \( a, b \in \mathbb{R}, a < b \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( f'(x) \leq 0 \) for all \( x \in (a, b) \), then \( f \) is nonincreasing on \([a, b]\), and in particular \( f(a) \geq f(b) \). If \( f' \) is not identically 0 on \((a, b)\), then \( f(a) > f(b) \).

Putting the last two together, we get a result of basic significance:
V.8.2.3. **Corollary.** Let \(a, b \in \mathbb{R}, a < b\). Let \(f\) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \(f'(x) = 0\) for all \(x \in (a, b)\), then \(f\) is constant on \([a, b]\).

V.8.2.4. **Corollary.** Let \(a, b \in \mathbb{R}, a < b\). Let \(f\) and \(g\) be functions which are continuous on \([a, b]\) and differentiable on \((a, b)\). If \(f'(x) = g'(x)\) for all \(x \in (a, b)\), then \(f\) and \(g\) differ by a constant.

**Proof:** Apply V.8.2.3. to \(f - g\).

V.8.2.5. **Corollary.** Let \(a, b \in \mathbb{R}, a < b\). Let \(f\) and \(g\) be functions which are continuous on \([a, b]\) and differentiable on \((a, b)\). If \(f(a) = g(a)\) and \(f'(x) \geq g'(x)\) for all \(x \in (a, b)\), then \(f(x) \geq g(x)\) for all \(x \in [a, b]\), with strict inequality unless \(f' = g'\) on \((a, x)\).

**Proof:** Apply V.8.2.1. to \(f - g\).

The First Derivative Test for Local Extrema

One of the standard tests for local extrema is a direct consequence of the MVT:

V.8.2.6. **Theorem.** [First Derivative Test] Let \(f\) be continuous on an interval \(I\), and \(a\) an interior point of \(I\). Suppose

(i) There is an \(r \in I, r < a\), with \(f\) differentiable on \((r, a)\) and \(f'(x) \geq 0\) for all \(x \in (r, a)\).

(ii) There is an \(s \in I, s > a\), with \(f\) differentiable on \((a, s)\) and \(f'(x) \leq 0\) for all \(x \in (a, s)\).

Then \(f\) has a local maximum at \(a\). It is a strict local maximum unless \(f'\) is identically zero in an interval to the left or right of \(a\).

Similarly, suppose

(iii) There is an \(r \in I, r < a\), with \(f\) differentiable on \((r, a)\) and \(f'(x) \leq 0\) for all \(x \in (r, a)\).

(iv) There is an \(s \in I, s > a\), with \(f\) differentiable on \((a, s)\) and \(f'(x) \geq 0\) for all \(x \in (a, s)\).

Then \(f\) has a local minimum at \(a\). It is a strict local minimum unless \(f'\) is identically zero in an interval to the left or right of \(a\).

Note that differentiability of \(f\) at \(a\) is not assumed (if \(f'(a)\) exists, it must be 0 as a result of this theorem and V.3.6.3.).

**Proof:** In the first case, apply V.8.2.1. to \([r, a]\) and V.8.2.2. to \([a, s]\), and vice versa in the second case.

A one-sided version of this test also applies at endpoints of \(I\); the statement is left to the reader.

Differential Approximation with Error Estimates

We can obtain a variation of differential approximation with error estimates:
**V.8.2.7.** Corollary. Let \( a, b, M \in \mathbb{R} \), \( a < b \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( f'(x) < M \) for all \( x \in (a, b) \), then
\[
f(x_2) - f(x_1) < M(x_2 - x_1)
\]
for all \( x_1, x_2 \in [a, b], x_1 < x_2 \). The result also holds if \(< \) is replaced by \( \leq \) throughout.

**Proof:** Let \( x_1, x_2 \in [a, b], x_1 < x_2 \). Then, applying the MVT to \([x_1, x_2]\), for some \( c \in (x_1, x_2) \) we have
\[
f(x_2) - f(x_1) = f'(c)(x_2 - x_1) < M(x_2 - x_1).
\]

Similarly, we have

**V.8.2.8.** Corollary. Let \( a, b, m \in \mathbb{R} \), \( a < b \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( m < f'(x) \) for all \( x \in (a, b) \), then
\[
m(x_2 - x_1) < f(x_2) - f(x_1)
\]
for all \( x_1, x_2 \in [a, b], x_1 < x_2 \). The result also holds if \(< \) is replaced by \( \leq \) throughout.

Combining, we get the two-sided version:

**V.8.2.9.** Corollary. Let \( a, b, m, M \in \mathbb{R} \), \( a < b \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( m < f'(x) < M \) for all \( x \in (a, b) \), then
\[
m(x_2 - x_1) < f(x_2) - f(x_1) < M(x_2 - x_1)
\]
for all \( x_1, x_2 \in [a, b], x_1 < x_2 \). The result also holds if \(< \) is replaced by \( \leq \) throughout.

Taylor's theorem gives a higher-order version of these inequalities ().

**V.8.2.10.** V.8.2.7. is often called the **Mean Value Inequality**. It is an immediate consequence of the special case V.8.2.1. and linearity of the derivative. There is a direct proof of V.8.2.1. (in the special case where \( f \) is differentiable on \([a, b]\)) by the Method of Successive Bisections () (Exercise V.8.6.11.). This proof is arguably simpler and more elementary than the proof of the MVT since it only uses the Nested Intervals Theorem and not the Extreme Value Theorem, and several authors, e.g. [?, p. 152], [vRS82, p. 93], [Boa95, p. 8], [Kör04, p. 59], argue that the Mean Value Inequality is sufficient for most applications in analysis and that the MVT itself is overrated. I personally do not regard the proof of the MVT as very difficult given the Extreme Value Theorem, which one would want to include anyway in any treatment of real analysis.

**Lipschitz Continuity**

A similar consequence which will be important later is:
**V.8.2.11. Corollary.** Let \( a, b, M \in \mathbb{R}, a < b, M \geq 0 \). Let \( f \) be a function which is continuous on \([a, b]\) and differentiable on \((a, b)\). If \( |f'(x)| \leq M \) for all \( x \in (a, b) \), then 

\[
|f(x_2) - f(x_1)| \leq M|x_2 - x_1|
\]

for all \( x_1, x_2 \in [a, b] \) (i.e. if \( f' \) is bounded on \((a, b)\), then \( f \) is a Lipschitz function \()\) on \([a, b]\).

**Proof:** Let \( x_1, x_2 \in [a, b], x_1 < x_2 \). Then, applying the MVT to \([x_1, x_2]\), for some \( c \in (x_1, x_2) \) we have

\[
|f(x_2) - f(x_1)| = |f'(c)||x_2 - x_1| \leq M|x_2 - x_1| .
\]


There is also a generalized version whose statement is left to the reader (it is very similar to V.8.4.9.).

**The Intermediate Value Property for Derivatives**

Although derivatives can be quite discontinuous (cf. V.12.), they rather remarkably satisfy the conclusion of the Intermediate Value Theorem.

**V.8.2.12. Definition.** A function \( g \) defined on an interval \( I \) has the Intermediate Value Property, or Darboux Property, if, whenever \( a, b \in I, a < b \), and \( r \) is a number strictly between \( g(a) \) and \( g(b) \), there is a \( c \in (a, b) \) with \( g(c) = r \).

A function which has the Intermediate Value Property is sometimes called Darboux continuous.

Every continuous function has the Intermediate Value Property by V.2.3.1. So do derivatives:

**V.8.2.13. Theorem.** If \( f \) is a differentiable function on an interval \( I \), then \( f' \) has the Intermediate Value Property on \( I \).

**Proof:** Let \( a, b \in I, a < b \), and \( r \) a number strictly between \( f'(a) \) and \( f'(b) \). Set \( g(x) = f(x) - rx \); then \( g \) is differentiable on \( I \) and \( g'(x) = f'(x) - r \) for \( x \in I \).

If \( f'(a) > r > f'(b) \), then \( g'(a) > 0 \) and \( g'(b) < 0 \). Since \( g \) is continuous on \([a, b]\), it has a maximum on this interval at some \( c \in [a, b] \) (V.2.4.1.). The maximum cannot occur at \( a \) or \( b \) since \( g'(a) > 0 \) and \( g'(b) < 0 \) (V.3.6.11.); thus \( c \in (a, b) \), and \( g'(c) = 0 \) by V.3.6.3.. Then \( f'(c) = r \).

The case \( f'(a) < r < f'(b) \) is nearly identical, considering the minimum of \( g \) on \([a, b]\).

**V.8.2.14. Corollary.** If \( f \) is differentiable on an interval \( I \), and \( f'(x) \neq 0 \) for all \( x \in I \), then \( f' \) does not change sign on \( I \), i.e. either \( f'(x) > 0 \) for all \( x \in I \) or \( f'(x) < 0 \) for all \( x \in I \). In particular, \( f \) is either strictly increasing on \( I \) or strictly decreasing on \( I \) (V.8.2.1., V.8.2.2.).

Here is another related approximate continuity property of derivatives:
**V.8.2.15.** **Theorem.** Let \( f \) be differentiable on an open interval \((a, b)\), and \( c \in (a, b) \). Then

\[
\liminf_{x \to c^+} f'(x) \leq f'(c) \leq \limsup_{x \to c^+} f'(x)
\]

\[
\liminf_{x \to c^-} f'(x) \leq f'(c) \leq \limsup_{x \to c^-} f'(x)
\]

In particular, if \( \lim_{x \to c^+} f'(x) \) exists, it equals \( f'(c) \), and similarly for \( \lim_{x \to c^-} f'(x) \).

**Proof:** We prove only the first inequality; the others are virtually identical. If \( f'(c) < \liminf_{x \to c^+} f'(x) \) there is an \( \epsilon > 0 \) and a \( d, c < d \leq b \), such that \( f'(x) \geq f'(c) + \epsilon \) for all \( x \in (c, d) \). Then, by the MVT, for every \( x, y \in (c, d) \), \( x \neq y \), we have

\[
\frac{f(y) - f(x)}{y - x} \geq f'(c) + \epsilon.
\]

Fix \( y \in (c, d) \); we then get

\[
\frac{f(y) - f(c)}{y - c} = \lim_{x \to c^+} \frac{f(y) - f(x)}{y - x} \geq f'(c) + \epsilon
\]

since \( f \) is continuous at \( c \). But then we have

\[
f'(c) = \lim_{y \to c^+} \frac{f(y) - f(c)}{y - c} \geq f'(c) + \epsilon
\]

which is a contradiction.

In fact, every Darboux continuous function has this approximate continuity property (Exercise V.8.6.4.).

See V.9.4.11. for an extension of the last statement. See also V.12.2.5.

**V.8.3.** **Geometric Interpretation of the Second Derivative**

Using the MVT, we can now give careful justification of the interpretation of the second derivative of a function in terms of concavity.

**V.8.3.1.** **Definition.** Let \( f \) be a continuous function on an interval \( I \). Then \( f \) is **concave upward** on \( I \) if, whenever \( a, b, c \in I \) with \( a < c < b \), we have

\[
f(c) < f(a) + \frac{f(b) - f(a)}{b - a}(c - a).
\]

The function \( f \) is **concave downward** on \( I \) if, whenever \( a, b, c \in I \) with \( a < c < b \), we have

\[
f(c) > f(a) + \frac{f(b) - f(a)}{b - a}(c - a).
\]

Geometrically, \( f \) is concave upward on \( I \) if on every closed bounded subinterval of \( I \), the graph of \( f \) on the interior lies strictly below the secant line through the endpoints, and similarly the graph lies strictly above the secant line if \( f \) is concave downwards. “Concave upward” is synonymous with “strictly convex” (cf. XIV.15.1.1.).

596
V.8.3.2. THEOREM. Let \( f \) be continuous on \([a, b]\) and differentiable on \((a, b)\). Then \( f \) is concave upward on \([a, b]\) if and only if \( f' \) is strictly increasing on \((a, b)\), and \( f \) is concave downward on \([a, b]\) if and only if \( f' \) is strictly decreasing on \((a, b)\).

**Proof:** We prove the first statement. Suppose \( f' \) is strictly increasing on \([a, b]\). Let \( L(x) \) be the linear function

\[
L(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)
\]

whose graph is the secant line through \((a, f(a))\) and \((b, f(b))\). Suppose there is an \( x_0 \in (a, b) \) with \( f(x_0) \geq L(x_0) \). Then by the MVT there are \( c_1 \in (a, x_0) \) and \( c_2 \in (x_0, b) \) with

\[
f'(c_1) = \frac{f(x_0) - f(a)}{x_0 - a} \geq \frac{L(x_0) - L(a)}{x_0 - a} = \frac{f(b) - f(a)}{b - a} \geq \frac{f(b) - f(x_0)}{b - x_0} = f'(c_2)
\]

which contradicts that \( f' \) is strictly increasing on \((a, b)\) since \( a < c_1 < c_2 < b \). The converse is a special case of XIV.15.1.16. The proof of the second statement is identical with inequalities reversed.

V.8.3.3. COROLLARY. Let \( f \) be continuous on \([a, b]\) and twice differentiable on \((a, b)\). If \( f''(x) \geq 0 \) for all \( x \in (a, b) \), and is not identically 0 on any subinterval, then \( f \) is concave upward on \([a, b]\). If \( f''(x) \leq 0 \) for all \( x \in (a, b) \), and is not identically 0 on any subinterval, then \( f \) is concave downward on \([a, b]\). In particular, if \( f'' > 0 \) [resp. \( f'' < 0 \)] on an interval \( I \), then \( f \) is concave upward [resp. concave downward] on \( I \).

**Proof:** If \( f'' \geq 0 \) on \((a, b)\) and not 0 on any subinterval, then \( f' \) is strictly increasing on \((a, b)\) (V.8.2.1.), and similarly if \( f'' \leq 0 \).

V.8.3.4. A function which is concave upward need not be differentiable everywhere, for example \( f(x) = |x| + x^2 \) (but it must be differentiable except on a countable set, cf. XIV.15.1.16.). A linear function is twice differentiable, but not concave up or concave down on any interval. And even a nonlinear twice differentiable function need not be either concave upward or concave downward on any interval: consider an antiderivative for the example of \((x)\).

There is a pointwise analog:

V.8.3.5. THEOREM. Let \( I \) be an interval, \( f \) be a function from \( I \) to \( \mathbb{R} \), and \( a \) in the interior of \( I \). Suppose \( f \) is twice differentiable at \( a \) (which implies that \( f' \) exists in a neighborhood of \( a \) and is continuous at \( a \)). Let \( L \) be the linear function whose graph is the tangent line to the graph of \( f \) at \( a \), i.e. \( L(x) = f(a) + f'(a)(x-a) \).

(i) If \( f''(a) > 0 \), then there is an \( \epsilon > 0 \) such that \( f(x) > L(x) \) for all \( x \), \( 0 < |x-a| < \epsilon \).
(ii) If \( f''(a) < 0 \), then there is an \( \epsilon > 0 \) such that \( f(x) < L(x) \) for all \( x \), \( 0 < |x-a| < \epsilon \).

Geometrically, if \( f''(a) > 0 \), the graph of \( f \) near \( a \) lies strictly above the tangent line at \( a \), and strictly below if \( f''(a) < 0 \).
Proof: (i): We have that $f'$ is increasing at $a$ (V.3.6.8.), i.e. there is an $\epsilon > 0$ such that $f'(x) < f'(a)$ if $x \in (a-\epsilon,a)$ and $f'(x) > f'(a)$ if $x \in (a,a+\epsilon)$. If $x \in (a-\epsilon,a)$, then by the MVT there is a $c \in (x,a)$ with

$$\frac{f(x) - f(a)}{x-a} = f'(c) < f'(a) = \frac{L(x) - L(a)}{x-a}$$

and it follows that $f(x) - f(a) > L(x) - L(a) = L(x) - f(a)$ since $x - a < 0$. The argument if $x \in (a,a+\epsilon)$ is essentially identical. The proof of (ii) is the same with inequalities reversed.

A special case is particularly important:

V.8.3.6. Corollary. [Second Derivative Test] Let $I$ be an interval, $f$ be a function from $I$ to $\mathbb{R}$, and $a$ in the interior of $I$. Suppose $f$ is twice differentiable at $a$ (which implies that $f'$ exists in a neighborhood of $a$ and is continuous at $a$).

(i) If $f'(a) = 0$ and $f''(a) > 0$, then $f$ has a strict local minimum at $a$.
(ii) If $f'(a) = 0$ and $f''(a) < 0$, then $f$ has a strict local maximum at $a$.

Proof: The function $L(x)$ is the constant function $f(a)$.

V.8.3.7. This result only goes one way. If $f$ has a relative extremum at $a$, it need not be twice differentiable (or even once differentiable!) at $a$, and even if it is we can have $f''(a) = 0$ (consider $f(x) = x^4$ at 0). See V.10.6.13. for an extension of the second derivative test.

Points of Inflection

A point of inflection of a function is a place where the function changes concavity:

V.8.3.8. Definition. Let $f$ be continuous on an interval $I$, and $c$ an interior point of $I$. Then $f$ has a point of inflection at $c$ if there are $a,b \in I$ with $a < c < b$ such that $f$ is concave upward on $[a,c]$ and concave downward on $[c,b]$, or vice versa.

V.8.3.9. If $f$ is differentiable on $I$, and has a point of inflection at $c \in I$, then by V.8.3.2. and V.8.2.15. $f'$ must have a strict local extremum at $c$. In particular, if $f$ is twice differentiable at $c$, then $f''(c) = 0$. However, the converses are not true: if $f(x) = x^4$, then $f''(0) = 0$, but $f$ does not have a point of inflection at 0, and in fact is concave up on all of $\mathbb{R}$. And if $g$ is an antiderivative for the function $f$ of V.3.10.11., then $g$ is twice differentiable everywhere with $g''(0) = 0$, and $g'$ has a strict local minimum at 0, but $g$ does not have a point of inflection at 0: in fact, $g$ is neither concave up or concave down in any interval with endpoint 0.

V.8.4. Refinements of the Mean Value Theorem

We can give a refinement of V.8.2.1.. We do not obtain this as a consequence of the MVT, but give a direct proof in the spirit of the MVT.
V.8.4.1. **Theorem.** Let $f$ be a continuous function on $[a, b]$, and $A$ a countable subset of $[a, b]$. Suppose $f$ is differentiable on $[a, b] \setminus A$.

(i) If $f'(x) > 0$ for all $x \in [a, b] \setminus A$, then $f$ is strictly increasing on $[a, b]$.

(ii) If $f'(x) \geq 0$ for all $x \in [a, b] \setminus A$, then $f$ is nondecreasing on $[a, b]$.

(iii) If $f'(x) = 0$ for all $x \in [a, b] \setminus A$, then $f$ is constant on $[a, b]$.

Similar statements to (i) and (ii) hold if $f$ is strictly increasing on $[a, b]$, and in fact this may fail (cf. V.8.4.2). Thus, to obtain a more general version of V.8.4.2. in this setting, we would have to add a clumsy set of additional hypotheses, although there are situations where a generalization is possible (e.g. XIV.17.1.17.). We can state part (i) in this setting:

**Proof:** (i): It suffices to show that $f$ is nondecreasing on $[a, b]$, since if $f$ is nondecreasing but not strictly increasing it would be constant on some subinterval, hence have $f' = 0$ on some interval. In fact, it suffices to show that $f(a) \leq f(b)$, since the interval $[a, b]$ can be replaced by any closed subinterval.

So suppose $f(a) > f(b)$. Since $f(A)$ is countable, we can find a number $d \notin f(A)$ with $f(a) > d > f(b)$. Let $B = \{x \in [a, b] : f(x) = d\}$. $B$ is nonempty by the Intermediate Value Theorem. Since $f$ is continuous, $B$ is closed, and thus has a largest element $c$. We have $a < c < b$ since $f(a) > d > f(b)$, and $c \notin A$ since $f(c) \notin f(A)$. So $f$ is differentiable at $c$ and $f'(c) > 0$. Since $f(b) < d$ and $f$ never takes the value $d$ on $(c, b)$, we must have $f(x) < d$ for all $x \in (c, b)$ by the Intermediate Value Theorem. Hence

$$
\lim_{x \to c^+} \frac{f(x) - f(c)}{x - c} \leq 0
$$

contradicting the assumption that $f'(c) > 0$.

(ii): Let $\epsilon > 0$. Apply (i) to $g(x) = f(x) + \epsilon x$ to conclude that $f(a) \leq f(b) + \epsilon(b - a)$. Since $\epsilon$ is arbitrary, $f(a) \leq f(b)$.

(iii): Apply (ii) to $f$ and $-f$.

The differentiability of $f$ on $[a, b] \setminus A$ can be relaxed to the assumption that the upper right Dini derivative of $f$ is positive (nonnegative, zero) on $[a, b] \setminus A$ (the proof of (iii) is a bit more subtle in this case, however; see (i)). Hence a more general statement can be proved with the same argument:

V.8.4.2. **Theorem.** Let $f$ be a continuous function on an interval $[a, b]$, and let $A$ be a countable subset of $[a, b]$. Then

(i) If $D^+ f(x) > 0$ for all $x \in [a, b] \setminus A$, then $f$ is strictly increasing on $[a, b]$.

(ii) If $D^+ f(x) \geq 0$ for all $x \in [a, b] \setminus A$, then $f$ is nondecreasing on $[a, b]$.

(iii) If $D^+ f(x) = 0$ for all $x \in [a, b] \setminus A$, then $f$ is constant on $[a, b]$.

Similar statements to (i) and (ii) hold if $D^+ f(x) < 0$ or $D^+ f(x) \leq 0$ for all $x \in [a, b] \setminus A$.

V.8.4.3. Note that the only property of $A$ needed for the proof of (i) is that $f(A)$ does not contain any interval. However, there is a subtlety in the proof of (ii) (and hence (iii)) if only this assumption is made: if $f(A)$ does not contain any interval, if $A$ is uncountable it is not obvious that $g(A)$ contains no interval if $g(x) = f(x) + \epsilon x$, and in fact this may fail (cf. [vRS82, 21.G]). Thus, to obtain a more general version of V.8.4.2. in this setting, we would have to add a clumsy set of additional hypotheses, although there are situations where a generalization is possible (e.g. XIV.17.1.17.). We can state part (i) in this setting:
V.8.4.4. **Theorem.** Let $f$ be a continuous function on an interval $[a, b]$, and let $A$ be a subset of $[a, b]$ such that $f(A)$ contains no interval. Then

(i) If $D^+ f(x) > 0$ for all $x \in [a, b] \setminus A$, then $f$ is strictly increasing on $[a, b]$.

(ii) If $D^+ f(x) < 0$ for all $x \in [a, b] \setminus A$, then $f$ is strictly decreasing on $[a, b]$.

Perhaps the most general version of V.8.4.2 (ii) which can be stated reasonably simply is:

V.8.4.5. **Theorem.** Let $f$ be a continuous function on an interval $[a, b]$. If $D^+ f(x) \geq 0$ for all $x$ not in a set $A$ of measure $0$, and $B = \{ x \in [a, b] : D^+ f(x) = -\infty \}$ is at most countable, then $f$ is nondecreasing on $[a, b]$.

(Under the hypotheses, the sets $B$, and even $A$, would turn out to be empty, but this need not be assumed. Actually, the result is interesting and useful even in the case $B = \emptyset$.)

**Proof:** Let $g$ be a nondecreasing function on $[a, b]$ with $D_+ g(x) = +\infty$ for all $x \in A$. Fix $\epsilon > 0$ and let $h_{\epsilon} = f + \epsilon g$. Then $D^+ h_{\epsilon}(x) \geq D^+ f(x) \geq 0$ for $x \notin A \cup B$ since $g$ is nondecreasing. If $x \in A \setminus B$, then $D^+ f(x) > -\infty$, so $D^+ h_{\epsilon}(x) = +\infty$ by ($\cdot$). Thus $D^+ h_{\epsilon}(x) \geq 0$ for all $x \notin B$, so $h_{\epsilon}$ is nondecreasing by V.8.4.2 (ii). Since $f = \lim_{\epsilon \to 0} h_{\epsilon}$, $f$ is also nondecreasing.

V.8.4.6. Note that the hypothesis that $B$ be countable cannot be replaced by an assumption that $B$ is a null set: the negative of the Cantor function ($\cdot$) is a counterexample.

Using V.8.4.1., we can obtain generalizations of V.8.2.7. and V.8.2.8.. As explained in V.8.4.3., to obtain a clean statement we must assume the conditions hold except on a countable subset.

V.8.4.7. **Corollary.** Let $a, b, M \in \mathbb{R}$, $a < b$. Let $f$ be a function which is continuous on $[a, b]$. Let $A$ be a countable subset of $[a, b]$. If $D^+ f(x) < M$ for all $x \in [a, b] \setminus A$, then

$$f(x_2) - f(x_1) < M(x_2 - x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

**Proof:** Apply V.8.4.2. to $h(x) = Mx - f(x)$.

V.8.4.8. **Corollary.** Let $a, b, m \in \mathbb{R}$, $a < b$. Let $f$ be a function which is continuous on $[a, b]$. Let $A$ be a countable subset of $[a, b]$. If $m < D^+ f(x)$ for all $x \in [a, b] \setminus A$, then

$$m(x_2 - x_1) < f(x_2) - f(x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

**Proof:** Apply V.8.4.2. to $g(x) = f(x) - mx$.
**V.8.4.9.** Corollary. Let \(a, b, m, M \in \mathbb{R}, a < b\). Let \(f\) be a function which is continuous on \([a, b]\). Let \(A\) be a countable subset of \([a, b]\). If \(m < D^+ f(x)\) and \(D_+ f(x) < M\) for all \(x \in [a, b] \setminus A\), then
\[
m(x_2 - x_1) < f(x_2) - f(x_1) < M(x_2 - x_1)
\]
for all \(x_1, x_2 \in [a, b], x_1 < x_2\). The result also holds if \(<\) is replaced by \(\leq\) throughout.

The following special case of V.8.4.2.(iii) is useful:

**V.8.4.10.** Corollary. Let \(f\) and \(g\) be continuous functions on an interval \(I\). If there is a countable subset \(A\) of \(I\) such that \(D_R f(x)\) and \(D_R g(x)\) exist and are equal for all \(x \in I \setminus A\), then \(f\) and \(g\) differ by a constant on \(I\).

**Proof:** Apply V.8.4.2.(iii) to \(f - g\) (on any closed bounded subinterval of \(I\)).

---

**V.8.5. Antiderivatives and Primitives**

Antiderivatives are usually associated with integration, but we can use the results of the previous subsection to obtain antiderivatives for certain well-behaved functions without discussing integration, at least explicitly.

**V.8.5.1.** Definition. Let \(f\) be a function on an interval \(I\). An antiderivative for \(f\) on \(I\) is a function \(F\) on \(I\) which is differentiable on \(I\) and with \(F' = f\) on \(I\).

A primitive for \(f\) on \(I\) is a continuous function \(F\) on \(I\) such that, for some countable subset \(A\) of \(I\), \(F'(x)\) exists and equals \(f(x)\) for all \(x \in I \setminus A\).

**V.8.5.2.** Since a differentiable function on an interval is continuous, every antiderivative for \(f\) on \(I\) is a primitive for \(f\) on \(I\) (with \(A = \emptyset\)). The terms “antiderivative” and “primitive” are often used as synonyms, but we make a slight technical distinction between them which will be useful.

**V.8.5.3.** Antiderivatives and primitives are not unique: a constant can be added to any antiderivative (primitive) to obtain another antiderivative (primitive). By V.8.2.5., this is the only change possible for antiderivatives: any two antiderivatives for a function \(f\) differ by a constant. This is also true for primitives by V.8.4.10.

**Step Functions**

Recall () that a step function on an interval \([a, b]\) corresponding to a partition
\[
\mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\}
\]
is a function which is constant on \((x_{k-1}, x_k)\) for \(1 \leq k \leq n\).
V.8.5.4. Proposition. Let \( f \) be a step function on \([a, b]\). Then \( f \) has a primitive on \([a, b]\).

Proof: This is quite simple and straightforward, but notationally a bit complicated. The function \( f \) will be “piecewise linear” or “polygonal”. Suppose \( f \) takes the value \( a_k \) on \((x_{k-1}, x_k)\). Define \( f \) as follows:

For \( x_0 \leq x \leq x_1 \), let \( F(x) = a_1(x - x_0) \).

For \( x_1 < x \leq x_2 \), let \( F(x) = a_1(x_1 - x_0) + a_2(x - x_1) \).

For \( x_2 < x \leq x_3 \), let \( F(x) = a_1(x_1 - x_0) + a_2(x_2 - x_1) + a_3(x - x_2) \).

\[ \cdots \]

For \( x_{k-1} < x \leq x_k \), let \( F(x) = \sum_{j=1}^{k-1} a_j(x_j - x_{j-1}) + a_k(x - x_{k-1}) \).

\[ \cdots \]

It is obvious that \( F \) is continuous on \([a, b]\), and that \( F'(x) = f(x) \) for \( x \in [a, b] \setminus \{x_0, \ldots, x_n\} \). ☐

The next theorem is the most important result of this subsection. It will be used for other purposes than the ones of this subsection, such as justifying term-by-term differentiation and integration of infinite series of functions. In many, but not all, applications, the set \( A \) in the statement of the theorem will be the empty set.

V.8.5.5. Theorem. Let \((F_n)\) and \((f_n)\) be sequences of functions on an interval \( I \), and \( A \) a countable subset of \( I \). Suppose

(a) Each \( F_n \) is continuous on \( I \).

(b) For every \( x \in I \setminus A \), each \( F_n \) is differentiable at \( x \) and \( F'_n(x) = f_n(x) \).

(c) The sequence \((f_n)\) converges u.c. on \( I \) to a function \( f \).

(d) There is an \( x_0 \in I \) such that the sequence \((F_n(x_0))\) converges.

Then:

(i) The sequence \((F_n)\) converges u.c. on \( I \) to a (continuous) function \( F \).

(ii) For each \( x \in I \setminus A \), \( F \) is differentiable at \( x \) and \( F'(x) = f(x) \).

Proof: (i): Let \( J \) be a closed bounded subinterval of \( I \). We will show that \((F_n)\) is a uniform Cauchy sequence on \( J \), hence uniformly convergent to a (continuous) limit function \( F \) by \( (\cdot) \). By expanding \( J \), we may assume \( x_0 \in J \).

Let \( \epsilon > 0 \). Choose \( N \) such that for all \( n, m \geq N \),

\[
\sup_{x \in J} |f_n(x) - f_m(x)| < \frac{\epsilon}{2\ell(J)}
\]

602
Now fix for all case where \( V.x.y \), and thus whenever \( 0 < \epsilon < \frac{\epsilon}{2} \). Then, if \( n, m \geq N \) and \( x \in J \), by V.8.4.7. we have, since \((F_n - F_m)' = f_n - f_m \) on \( J \setminus A \),

\[
|F_n(x_0) - F_m(x_0)| < \frac{\epsilon}{2}.
\]

Then, if \( n, m \geq N \) and \( x \in J \), by V.8.4.7. we have, since \((F_n - F_m)' = f_n - f_m \) on \( J \setminus A \),

\[
|F_n(x) - F_m(x)| - |F_n(x_0) - F_m(x_0)| < \frac{\epsilon}{2f(J)} |x - x_0| \leq \frac{\epsilon}{2}
\]

and so

\[
|F_n(x) - F_m(x)| = |[F_n(x) - F_n(x_0)] - [F_m(x) - F_m(x_0)] + [F_n(x_0) - F_m(x_0)]| \\
\leq |[F_n(x) - F_n(x_0)] - [F_m(x) - F_m(x_0)]| + |F_n(x_0) - F_m(x_0)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\]

(ii): Let \( J \) be as above, and \( \epsilon > 0 \). There is an \( N \) such that

\[
\sup_{x \in J} |f_n(x) - f(x)| < \frac{\epsilon}{4}
\]

for all \( n \geq N \). Fix \( n \geq N \). For all \( m \geq N \), we have

\[
\sup_{x \in J} |f_n(x) - f_m(x)| < \frac{\epsilon}{2}
\]

and hence by V.8.4.7. we have

\[
|[F_m(y) - F_m(x)] - [F_n(y) - F_n(x)]| = |[F_m(y) - F_n(y)] - [F_m(x) - F_n(x)]| < \frac{\epsilon}{2} |y - x|
\]

for all \( x, y \in J \). Since \( F_m \to F \) uniformly on \( J \), letting \( m \to \infty \) we have

\[
|[F(y) - F(x)] - [F_n(y) - F_n(x)]| \leq \frac{\epsilon}{2} |y - x|
\]

for all \( x, y \in J \).

Now fix \( x \in J \setminus A \). Then there is a \( \delta > 0 \) such that

\[
\left| \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) \right| < \frac{\epsilon}{4}
\]

whenever \( 0 < |y - x| < \delta \). Then, for \( 0 < |y - x| < \delta \), we have

\[
\left| \frac{F(y) - F(x)}{y - x} - f(x) \right| = \left| \frac{F(y) - F(x)}{y - x} - \frac{F_n(y) - F_n(x)}{y - x} + \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) + f_n(x) - f(x) \right| \\
\leq \left| \frac{F(y) - F(x)}{y - x} - \frac{F_n(y) - F_n(x)}{y - x} \right| + \left| \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) \right| + |f_n(x) - f(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon
\]

and thus \( F'(x) = f(x) \).

\[\Box\]

V.8.5.6. A simpler proof using the Fundamental Theorem of Calculus can be given in the crucial special case where \( A = \emptyset \) and the \( f_n \) are continuous (XIV.2.9.15.).

603
V.8.5.7. **Corollary.** Let \((f_n)\) be a sequence of functions on an interval \(I\), converging u.c. on \(I\) to a function \(f\).

(i) If each \(f_n\) has a primitive on \(I\), then \(f\) has a primitive on \(I\).

(ii) If each \(f_n\) has an antiderivative on \(I\), then \(f\) has an antiderivative on \(I\).

**Proof:** Fix \(x_0 \in I\). By adjusting by a constant, \(f_n\) has a primitive (antiderivative in case (ii)) \(F_n\) on \(I\) with \(F_n(x_0) = 0\). Let \(A_n\) be the set of \(x \in I\) for which \(F'_n(x) = f_n(x)\) does not hold; then \(A_n\) is countable, and \(A_n = \emptyset\) in case (ii). Set \(A = \bigcup_{n=1}^{\infty} A_n\). Then \((f_n), (F_n), A\) satisfy the hypotheses of V.8.5.5.  

V.8.5.8. **Proposition.** Let \(f\) be a continuous function on an interval \(I\). Then there is a sequence \((f_n)\) of step functions on \(I\) converging u.c. to \(f\) on \(I\). If \(A\) is any dense subset of \(I\), the \(f_n\) may be chosen to have discontinuities only in \(A\).

**Proof:** This is an almost immediate consequence of (i). Let \([a_n, b_n]\) be an increasing sequence of closed bounded intervals with union \(I\). Then \(f\) is uniformly continuous on \([a_n, b_n]\). For each \(n\) fix \(\delta_n > 0\) such that \(x, y \in [a_n, b_n], |x - y| < \delta_n\) implies \(|f(x) - f(y)| < \frac{1}{n}\). Then let \(\mathcal{P}_n\) be a partition of \([a_n, b_n]\) with interior points in \(A\) with mesh \(< \delta_n\), and \(t_k\) a point in the \(k\)th subinterval of the partition. Let \(f_n\) be the step function with value \(f(t_k)\) on the \(k\)th subinterval of the partition \(\mathcal{P}_n\) (and, say, left continuous at the points of the partition), and extend \(f_n\) to \(I\) by making it constant to the left and right of \([a_n, b_n]\) if necessary. Then \(f_n \to f\) u.c. on \(I\) (Exercise V.8.6.6.).

V.8.5.9. **Theorem.** Let \(f\) be a continuous function on an interval \(I\). Then \(f\) has an antiderivative on \(I\).

**Proof:** Fix \(x_0 \in I\), and let \(A\) and \(B\) be disjoint countable dense subsets of \(I\) (e.g. \(A = \mathbb{Q} \cap I, B = [\mathbb{Q} + \sqrt{2}] \cap I\)). Then there is a sequence \((f_n)\) of step functions on \(I\) with discontinuities only in \(A\), converging u.c. to \(f\) on \(I\). It follows from V.8.5.5. that \(f\) has a primitive \(F_1\) on \(I\) satisfying \(F'_1(x) = f(x)\) for all \(x \in I \setminus A\) and \(F_1(x_0) = 0\). Similarly, there is a primitive \(F_2\) of \(f\) on \(I\) such that \(F'_2(x) = f(x)\) for all \(x \in I \setminus B\) and \(F_2(x_0) = 0\). By V.8.4.10. (applied using the countable set \(A \cup B\)), \(F_1\) and \(F_2\) differ by a constant. But \(F_1(x_0) = F_2(x_0)\), so \(F_1 = F_2\). Thus \(F'_1(x)\) exists and equals \(f(x)\) for all \(x \in (I \setminus A) \cup (I \setminus B) = I\).

V.8.5.10. We will give another proof of this theorem in XIV.2.9.4. (Deep down, it is really the same proof.)

Another example where a similar argument applies is monotone functions:

V.8.5.11. **Proposition.** Let \(f\) be a monotone function on an interval \(I\). Then there is a sequence \((f_n)\) of step functions on \(I\) converging u.c. to \(f\) on \(I\).
V.8.5.12. Corollary. Let \( f \) be a monotone function on an interval \( I \). Then \( f \) has a primitive on \( I \).

V.8.5.13. It can be shown with a little additional work that if \( f \) is monotone on \( I \) and \( F \) is a primitive of \( f \) on \( I \), then the (countable) set \( A \) of \( x \) for which \( F'(x) = f(x) \) does not hold is exactly the set of discontinuities of \( f \). See V.8.6.7. for a more general result.

V.8.6. Exercises

V.8.6.1. Let \( C \) be a smooth curve from \( p \) to \( q \) in \( \mathbb{R}^2 \) (i.e. with a parametrization \( (g(t), f(t))_{a \leq t \leq b} \) with \( f \) and \( g \) differentiable on \([a, b]\) and \( f' \) and \( g' \) not simultaneously 0). If \( p \neq q \) and \( r \) is a point on \( C \) of maximum distance from the line \( L \) through \( p \) and \( q \) (why does such a point exist?), show that the tangent line to \( C \) at \( r \) is parallel to \( L \).

V.8.6.2. (a) Prove V.8.4.2.(i) and (ii) in the case where the upper right Dini derivative \( D_+ f \) is positive or nonnegative on \([a, b] \setminus A \). Show why (iii) does not immediately follow from (ii) in this case.

(b) Suppose \( D_+ f(x) = 0 \) for all \( x \in [a, b] \setminus A \). Apply (ii) to conclude that \( f \) is nondecreasing on \([a, b] \). Thus the lower right Dini derivative \( D_- f \) satisfies \( D_- f(x) \geq 0 \) for all \( x \in [a, b] \).

(c) In the situation of (b), conclude that \( D_+ f(x) = 0 \) for all \( x \in [a, b] \setminus A \); thus (ii) can be applied to \(-f\) to conclude that \(-f\) is increasing on \([a, b] \). Thus \( f \) is constant on \([a, b]\).

V.8.6.3. Use the Intermediate Value Property for derivatives () to give an alternate proof of ().

V.8.6.4. Let \( g \) be a Darboux continuous function on an open interval \((a, b)\), and \( c \in (a, b) \). Show that

\[
\liminf_{x \to c^+} g(x) \leq g(c) \leq \limsup_{x \to c^+} g(x)
\]

\[
\liminf_{x \to c^-} g(x) \leq g(c) \leq \limsup_{x \to c^-} g(x)
\]

V.8.6.5. Let \( f \) be a continuous function on an open interval \( I \). Set

\[
M^+ = \sup\{D^+ f(x) : x \in I\}
\]

\[
m^+ = \inf\{D^+ f(x) : x \in I\}
\]

\(m^+\) and/or \(M^+\) may be infinite), and let

\[
S = \left\{ \frac{f(y) - f(x)}{y - x} : x, y \in I, x \neq y \right\}
\]

(a) Show that \((m^+, M^+) \subseteq S \subseteq [m^+, M^+]\). To show \((m^+, M^+) \subseteq S\), fix \( \lambda \in (m^+, M^+) \) and let \( g(x) = f(x) - \lambda x \). Then \( g \) has both positive and negative right Dini derivatives on \( I \) and hence cannot be one-to-one by () . For \( S \subseteq [m^+, M^+]\), let \( \lambda \leq m^+ \) and \( \mu \geq M^+ \), and apply V.8.4.2.(ii) to \( f(x) - \lambda x \) and \( \mu x - f(x)\).

(b) Show that \( m^+ \in S \) if and only if \( f \) is linear with slope \( m^+ \) on a subinterval of \( I \), and similarly for \( M^+ \).
(c) Repeat (a) with

\[ M_+ = \sup \{ D_+ f(x) : x \in I \} \]
\[ m_+ = \inf \{ D_+ f(x) : x \in I \} \]

and with \( M^- \), \( m^- \) and \( M_- \), \( m_- \) defined similarly using \( D^- f \) and \( D_- f \). Conclude that

\[ m^+ = m_+ = m^- = m_- \quad \text{and} \quad M^+ = M_+ = M^- = M_- . \]

(d) Use (a) to give an alternate proof of the Darboux property for derivatives. [Show that if \( f \) is differentiable on an open interval \( I \), then \( f'(J) \) is an interval for every open subinterval \( J \) of \( I \).]

\[ \text{V.8.6.6.} \quad \text{Show that if } f \text{ is as in V.8.5.8., and } (f_n) \text{ is the sequence defined in the proof, then } f_n \to f \text{ u.c. on } I. \]

\[ \text{V.8.6.7.} \quad \text{In [?], IX, Book IV, Chapter II, §1.3], a function on an interval } I \text{ is called regulated ("réglée") on } I \text{ if it is a u.c. limit of a sequence of step functions on } I. \text{ Note that every regulated function has a primitive by V.8.5.5. and V.8.5.4.. By V.8.5.8. and V.8.5.11., every continuous function and every monotone function is regulated.}

(a) Show that a function on an interval \( I \) which has left and right limits at every interior point of \( I \) has only countably many discontinuities on \( I \). [Mimic the proof of V.4.2.11..]

(b) Show that a function \( f \) on an interval \( I \) is regulated if and only if it has left and right limits at every interior point of \( I \) and one-sided limits at any endpoints of \( I \) which are in \( I \).

(c) Show that if \( f \) is regulated on \( I \) and \( F \) is a primitive for \( f \) on \( I \), then the (countable) set \( A \) of all \( x \in I \) for which \( F'(x) = f(x) \) does not hold is precisely the set of discontinuities of \( f \).

(d) Show that a regulated function on an interval \( I \) is bounded on every closed bounded subinterval of \( I \). [Step functions are bounded.]

(e) Any (finite) linear combination of regulated functions on an interval \( I \) is regulated. Thus every function of bounded variation () on every closed bounded subinterval of \( I \) is regulated on \( I \). Not every regulated function (indeed, not every continuous function) has bounded variation, even on closed bounded subintervals of \( I \).

See XIV.3.11.4. for more on regulated functions.

\[ \text{V.8.6.8.} \quad \text{(a) Show by direct calculation that every polynomial has an antiderivative on } \mathbb{R}. \]

(b) Use V.8.5.5. and XV.8.1.1. to give an alternate proof of V.8.5.9..

\[ \text{V.8.6.9.} \quad \text{[?], Exercise 7.6.21] (a) Prove the following generalization of Cauchy’s Mean Value Theorem:}

\[ \text{THEOREM. Let } f, g, h \text{ be continuous on } [a, b] \text{ and differentiable on } (a, b). \text{ Then there is a } c \in (a, b) \text{ such that}

\[ 0 = \text{Det} \begin{bmatrix} f(a) & g(a) & h(a) \\ f(b) & g(b) & h(b) \\ f'(c) & g'(c) & h'(c) \end{bmatrix} = f'(c)[g(a)h(b) - g(b)h(a)] - g'(c)[f(a)h(b) - f(b)h(a)] + h'(c)[f(a)g(b) - f(b)g(a)]. \]

606
Cauchy’s MVT is the special case where \( h \) is the constant function 1. [Consider the function

\[
\phi(x) = \det \begin{bmatrix} f(a) & g(a) & h(a) \\ f(b) & g(b) & h(b) \\ f(x) & g(x) & h(x) \end{bmatrix}
\]

and apply Rolle’s Theorem. Recall that a (square) matrix with two identical rows has determinant 0.]

(b) Extend the result to the case \( n > 3 \):

**Theorem.** Let \( n \geq 4 \), and let \( f_1, \ldots, f_n \) be continuous on \([a, b]\) and differentiable on \((a, b)\). Let \( \{d_{ij} : 3 \leq i \leq n-1, 1 \leq j \leq n\} \) be arbitrary constants. Then there is a \( c \in (a, b) \) such that

\[
0 = \det \begin{bmatrix} f_1(a) & f_2(a) & \cdots & f_n(a) \\ f_1(b) & f_2(b) & \cdots & f_n(b) \\ d_{31} & d_{32} & \cdots & d_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n-1,1} & d_{n-1,2} & \cdots & d_{n-1,n} \\ f'_1(c) & f'_2(c) & \cdots & f'_n(c) \end{bmatrix}.
\]

(c) Here is another version:

**Theorem.** Let \( n \geq 3 \), and let \( \mathcal{P} = \{a = x_1, x_2, \ldots, x_{n-1} = b\} \) be a partition of \([a, b]\) into \( n-2 \) subintervals. Let \( f_1, \ldots, f_n \) be continuous on \([a, b]\) and differentiable on \([a, b] \setminus \mathcal{P}\). Then, for each \( k \leq n-2 \), there is a \( c \in (x_k, x_{k+1}) \) such that

\[
0 = \det \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_{n-1}) & f_2(x_{n-1}) & \cdots & f_n(x_{n-1}) \\ f'_1(c) & f'_2(c) & \cdots & f'_n(c) \end{bmatrix}.
\]

**V.8.6.10. The Mean Value Inequality.** This problem gives a direct proof of V.8.2.7. (in the case where \( f \) is differentiable on all of \([a, b]\)) by the Method of Successive Bisections ().

(a) Show that to prove V.8.2.7. it suffices to prove V.8.2.1.

(b) Show that to prove V.8.2.1. it suffices to prove the following. [For each \( \epsilon > 0 \) set \( f_\epsilon(x) = f(x) + \epsilon x \). To show that \( f(a) \leq f(b) \), it suffices to show \( f_\epsilon(a) \leq f_\epsilon(b) \) for all \( \epsilon > 0 \).]

**Proposition.** Let \( f \) be differentiable on \([a, b]\). If \( f'(x) > 0 \) for all \( x \in [a, b] \), then \( f(a) \leq f(b) \).

(c) To prove this proposition, set \( I_1 = [a, b] \), and \( c \) the midpoint of \( I_1 \). If \( f(a) > f(b) \), then either \( f(a) > f(c) \) or \( f(c) > f(b) \). Let \( I_2 = [a, c] \) if the first holds and \( I_2 = [c, b] \) otherwise. Repeat the process to obtain a nested sequence \((I_n)\) of intervals, where \( I_n = [a_n, b_n] \) with \( f(a_n) > f(b_n) \) for all \( n \).

(d) Let \( \cap_{n=1}^\infty I_n = \{x\} \). Show that \( f'(x) \leq 0 \), contradicting that \( f'(x) > 0 \).
V.8.6.11. The aim of this problem is to use a purely algebraic argument, along with the Intermediate Value Theorem for polynomials, to prove Rolle’s Theorem for polynomials: between any two real roots of a polynomial $p$ with real coefficients is a root of $p'$.

(a) Let $p(x)$ be a nonzero polynomial with real coefficients, and $a$ and $b$ consecutive real roots of $p$. Write

$$p(x) = (x-a)^r(x-b)^s q(x)$$

where $r, s \in \mathbb{N}$ and $q$ is a polynomial with real coefficients with no roots in $[a,b]$, and which therefore does not change sign on $[a,b]$.

(b) Use the product rule to show that

$$p'(x) = (x-a)^{r-1}(x-b)^{s-1} [r(x-b)q(x) + s(x-a)q(x) + (x-a)(x-b)q'(x)].$$

(c) Show that the polynomial in brackets in (b) takes values of opposite sign at $a$ and $b$.

(d) Apply the Intermediate Value Theorem for polynomials to conclude that $p'$ has a root in $(a,b)$.

No actual differential calculus is needed for this problem, since the derivative of a polynomial can be defined and calculated formally, and the product rule for polynomials proved algebraically from the formal definition of derivative; cf. Exercise V.3.10.9. (Even the Intermediate Value Theorem for polynomials has a purely algebraic proof in real-closed fields; see, for example, [BPR06].) In fact, ROLLE did something like this, and it is unclear whether he even realized he was computing derivatives: he generated the sequence of derivatives of the given polynomial $p$ in a somewhat different (algebraic) way and called them the “cascade of polynomials” in using them to locate the (real) roots of $p$. See the chapter by J. BARROW-GREEN in [RS09].

V.8.6.12. Let $f$ be differentiable on an interval $I$, and suppose $f'(x) > 0$ for all $x \in I$. Use V.3.6.8. and the Max-Min Theorem () to prove that $f$ is strictly increasing on $I$ without using the MVT. [If $x_1, x_2 \in I, x_1 < x_2$, apply the Max-Min Theorem to $[x_1, x_2].$]

V.8.6.13. Let $f$ be continuous on an interval $I$ and $n$-times differentiable on the interior of $I$. Suppose $f$ has $m$ distinct zeroes in $I$. Show by repeated application of Rolle’s Theorem that for $1 \leq k \leq \min(n, m)$, $f^{(k)}$ has at least $m - k$ distinct zeroes in the interior of $I$.

V.8.6.14. (a) Find an elementary closed formula for

$$\int \sqrt{\tan x} \, dx$$

(a computer algebra system is very helpful).

(b) (LIOUVILLE, 1835) Show that

$$\int \sqrt{\sin x} \, dx$$

has no elementary closed formula.
V.9. L’Hôpital’s Rule

L’Hôpital’s Rule is a very useful technique for evaluating limits of indeterminate forms involving quotients. There are a number of variations of the rule, which all fall under the general name “l’Hôpital’s Rule.”

The rule is named for the French nobleman MARQUIS GUILLAUME FRANÇOIS ANTOINE DE L’HÔPITAL (sometimes spelled l’Hospital; his own spelling was apparently Lhospital) and was first published in 1696 in the calculus text _L'Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes_ (the first actual calculus text!) written by L’HÔPITAL; this name for the rule apparently first appeared in print in [Gou59] in 1902. The rule was actually discovered by JOHANN (or JEAN) BERNOULLI, who had taught calculus to L’HÔPITAL, in 1694; BERNOULLI had a business arrangement with L’HÔPITAL giving L’HÔPITAL, who was himself a respectable mathematician, the right to publish BERNOULLI’s results. (See [?] for the full story.) L’HÔPITAL’s text, which fairly closely followed lecture notes by BERNOULLI, was very influential, and editions remained in print for almost 100 years (a copy is available online at http://www.archive.org/details/infinimentpetits1716lhos00uoft); before this text became available, calculus was known and understood by only a handful of top European mathematicians. L’HÔPITAL, besides being a gifted expositor (he also wrote a text on conic sections which was the standard for that subject throughout the eighteenth century), seems to have been a considerably more pleasant and genial person than BERNOULLI: BERNOULLI, although he was a great mathematician and had a reputation as an inspiring teacher (EULER was his prize pupil), was hypercompetitive and was continually embroiled in bitter disputes with other mathematicians, the most vicious of which were with his own brother JAKOB (who was himself described as “self-willed, obstinate, aggressive, vindictive, beset by feelings of inferiority and yet firmly convinced of his own abilities.”) [Hof71] [Do not confuse the various BERNOULLIS; there were several outstanding mathematicians in this family, extending over at least three generations, but JOHANN and JAKOB (or JACQUES), and JOHANN’s son DANIEL, were the best.]

V.9.1. L’Hôpital’s Rule, Version 1

The first version we give of l’Hôpital’s Rule is an easy consequence of the definition of derivative, and is often not regarded as a form of l’Hôpital’s Rule at all.

V.9.1.1. THEOREM. [L’Hôpital’s Rule, Version 1] Let \( f \) and \( g \) be functions on an interval \( I \), and \( a \) an interior point of \( I \). Suppose

1. \( \lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0 \).
2. \( f \) and \( g \) are differentiable at \( a \) and \( g'(a) \neq 0 \).

Then \( \lim_{x \to a} \frac{f(x)}{g(x)} \) exists and equals \( \frac{f'(a)}{g'(a)} \).

PROOF: First note that since \( f \) and \( g \) are differentiable at \( a \), they are continuous at \( a \), and thus \( f(a) = \)
\[ \lim_{x \to a} f(x) = 0 \text{ and } g(a) = \lim_{x \to a} g(x) = 0. \]

We have
\[
\frac{f'(a)}{g'(a)} = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \cdot \lim_{x \to a} \frac{g(x) - g(a)}{x - a}.
\]

Note that \( g(x) = g(x) - g(a) \neq 0 \) in a deleted neighborhood of \( a \) since
\[
g'(a) = \lim_{x \to a} \frac{g(x) - g(a)}{x - a} \neq 0.
\]

So, by (1),
\[
\frac{f'(a)}{g'(a)} = \lim_{x \to a} \frac{f(x) - f(a)}{g(x) - g(a)} = \lim_{x \to a} \frac{f(x) - f(a)}{g(x) - g(a)}.
\]

Version 1 of the rule is adequate for many of the elementary applications.

**Example.** To evaluate
\[ \lim_{x \to 0} \frac{e^x - 1}{\sin x} \]

let \( f(x) = e^x - 1, g(x) = \sin x, a = 0 \). We have \( \lim_{x \to 0} f(x) = \lim_{x \to 0} g(x) = 0, f'(0) = g'(0) = 1 \), so the hypotheses of Version I are satisfied and
\[ \lim_{x \to 0} \frac{e^x - 1}{\sin x} = \frac{f'(0)}{g'(0)} = \frac{1}{1} = 1. \]

(Note that \( g(x) \neq 0 \) on \((-\pi, \pi) \setminus \{0\}\).)

**L’Hôpital’s Rule, Version 2**

There is another version (the commonly stated version) of l’Hôpital’s Rule which is more complicated to prove, but more widely applicable. There are several slight variations of this version, which will collectively be called Version 2 of l’Hôpital’s Rule.

**Theorem.** [L’Hôpital’s Rule, Version 2] Let \( f \) and \( g \) be differentiable functions on an interval \((a, b)\), with \( g' \) nonzero on \((a, b)\). Suppose

(1) \( \lim_{x \to a^+} f(x) = \lim_{x \to a^+} g(x) = 0 \).

(2) \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \) exists in either the usual () or extended () sense.
Then \( \lim_{x \to a^+} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \).

Note that it is not quite correct to say that if \( f \) and \( g \) are differentiable on \((a, b)\) with \( g' \) nonzero, and \( \lim_{x \to a^+} f(x) = \lim_{x \to a^+} g(x) = 0 \), then
\[
\lim_{x \to a^+} \frac{f(x)}{g(x)} = \lim_{x \to a^+} \frac{f'(x)}{g'(x)};
\]
it can happen that \( \lim_{x \to a^+} \frac{f(x)}{g(x)} \) exists but \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \) does not exist (Exercise V.9.4.4), even if \( f \) and \( g \) are \( C^\infty \) and \( g(x) \) and \( g'(x) \) are nonzero for \( x > a \) (Exercise V.9.4.7.).

**Proof:** If necessary, extend \( f \) and \( g \) to be continuous on \([a, b]\) by setting \( f(a) = g(a) = 0 \). If \( x \in (a, b) \), since \( g \) is continuous on \([a, x]\) and differentiable on \((a, x)\), by the Mean Value Theorem () we have
\[
g(x) - g(a) = g'(d)(x - a)
\]
for some \( d \in (a, x) \); since \( g'(d) \neq 0 \), we have \( g(x) = g(x) - g(a) \neq 0 \). Thus \( \frac{f}{g} \) is defined on \((a, b)\).

Set \( L = \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \in \mathbb{R} \cup \{\pm \infty\} \). By the Sequential Criterion (), it suffices to show that if \((x_n)\) is a sequence in \((a, b)\) converging to \( a \), we have \( \lim_{n \to \infty} \frac{f(x_n)}{g(x_n)} = L \). By the Cauchy Mean Value Theorem we have, for each \( n \),
\[
\frac{f(x_n)}{g(x_n)} = \frac{f(x_n) - f(a)}{g(x_n) - g(a)} = \frac{f'(c_n)}{g'(c_n)}
\]
for some \( c_n \in (a, x_n) \). Since \( a < c_n < x_n \) and \( x_n \to a \), we have \( c_n \to a \) by the Squeeze Theorem () and hence
\[
\frac{f(x_n)}{g(x_n)} = \frac{f'(c_n)}{g'(c_n)} \to L
\]
by the Sequential Criterion applied to \( \lim_{x \to a^+} \frac{f(x)}{g'(x)} = L \).

**V.9.2.4.** Actually, the hypothesis that \( g' \) is nonzero on \((a, b)\) is redundant: it is implicit in the existence of the limit \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \) that \( \frac{f}{g} \) is defined, i.e. \( g' \neq 0 \), in an open interval \((a, r)\) for some \( r > a \). Thus the hypothesis can be eliminated at the possible price of reducing the \( b \) to a smaller number still strictly larger than \( a \). We have left the requirement in the statement of the theorem as a reminder that this hypothesis must be checked when the theorem is applied (cf. Exercise V.9.4.5..)

Note that the hypothesis that \( g' \neq 0 \) on \((a, b)\) implies that \( g' \) does not change sign on \((a, b)\) (V.8.2.14.).

An essentially identical argument shows:
V.9.2.5. Theorem. [L’Hôpital’s Rule, Version 2b] Let \( f \) and \( g \) be differentiable functions on an interval \((a, b)\), with \( g' \) nonzero on \((a, b)\). Suppose

1. \( \lim_{x \to b^-} f(x) = \lim_{x \to b^-} g(x) = 0 \).

2. \( \lim_{x \to b^-} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to b^-} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to b^-} \frac{f'(x)}{g'(x)} \).

Combining the two previous versions together, we get:

V.9.2.6. Theorem. [L’Hôpital’s Rule, Version 2c] Let \( f \) and \( g \) be differentiable functions on a deleted neighborhood \( U \) of \( a \in \mathbb{R} \), with \( g' \) nonzero on \( U \). Suppose

1. \( \lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0 \).

2. \( \lim_{x \to a} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to a} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to a} \frac{f'(x)}{g'(x)} \).

There is also a version for limits at \( \infty \):

V.9.2.7. Theorem. [L’Hôpital’s Rule, Version 2d] Let \( f \) and \( g \) be differentiable functions on an interval \((a, +\infty)\), with \( g' \) nonzero on \((a, +\infty)\). Suppose

1. \( \lim_{x \to +\infty} f(x) = \lim_{x \to +\infty} g(x) = 0 \).

2. \( \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to +\infty} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} \).

A similar statement holds for limits as \( x \to -\infty \) for functions differentiable on an interval \((-\infty, a)\).

Proof: We may assume \( a > 0 \). Set \( L = \lim_{x \to +\infty} \frac{f(x)}{g(x)} \). For \( t \in (0, \frac{1}{a}) \), set \( F(t) = f(1/t) \) and \( G(t) = g(1/t) \). Then \( F \) and \( G \) are differentiable on \((0, \frac{1}{a})\), \( G'(t) = -\frac{1}{t^2}g'(1/t) \neq 0 \), and \( G(t) \neq 0 \). We have

\[
\lim_{t \to 0^+} \frac{F'(t)}{G'(t)} = \lim_{t \to 0^+} \frac{-\frac{1}{t^2}f'(1/t)}{-\frac{1}{t^2}g'(1/t)} = \lim_{t \to 0^+} \frac{f'(1/t)}{g'(1/t)} = \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} = L
\]
and so, by V.9.2.3.,
\[
\lim_{x \to +\infty} \frac{f(x)}{g(x)} = \lim_{t \to 0^+} \frac{F(t)}{G(t)} = L .
\]

L'Hôpital's Rule also applies to indeterminate forms of type \( \frac{\infty}{\infty} \). The statements are essentially the same, but the proof is a little more complicated. In fact, it is unnecessary to assume the limit of the numerator is \( \frac{\infty}{\infty} \) (if \( \lim f(x) \neq 0 \) and \( \lim g(x) = \pm \infty \), it follows automatically that \( \lim f(x) = \pm \infty \); this can also be proved directly from \( \lim f(x)g'(x) \neq 0 \) without applying l'Hôpital's rule, cf. V.9.4.8.).

V.9.2.8. **Theorem.** [L'Hôpital's Rule, Version 2a] Let \( f \) and \( g \) be differentiable functions on an interval \((a, b)\), with \( g' \) nonzero on \((a, b)\). Suppose

1. \( \lim_{x \to a^+} g(x) = \pm \infty \).
2. \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to a^+} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to a^+} \frac{f'(x)}{g'(x)} \).

V.9.2.9. **Theorem.** [L'Hôpital's Rule, Version 2b'] Let \( f \) and \( g \) be differentiable functions on an interval \((a, b)\), with \( g' \) nonzero on \((a, b)\). Suppose

1. \( \lim_{x \to b^-} g(x) = \pm \infty \).
2. \( \lim_{x \to b^-} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to b^-} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to b^-} \frac{f'(x)}{g'(x)} \).

V.9.2.10. **Theorem.** [L'Hôpital's Rule, Version 2c'] Let \( f \) and \( g \) be differentiable functions on a deleted neighborhood \( U \) of \( a \in \mathbb{R} \), with \( g' \) nonzero on \( U \). Suppose

1. \( \lim_{x \to a} g(x) = \pm \infty \).
2. \( \lim_{x \to a} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to a} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to a} \frac{f'(x)}{g'(x)} \).
V.9.2.11. Theorem. [L'Hôpital's Rule, Version 2d'] Let \( f \) and \( g \) be differentiable functions on an interval \((a, +\infty)\), with \( g' \) nonzero on \((a, +\infty)\). Suppose

1. \( \lim_{x \to +\infty} g(x) = \pm \infty \).
2. \( \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} \) exists in either the usual or extended sense.

Then \( \lim_{x \to +\infty} \frac{f(x)}{g(x)} \) exists (in the same sense) and equals \( \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} \).

A similar statement holds for limits as \( x \to -\infty \) for functions differentiable on an interval \((-\infty, a)\).

Proof: We give the proof of Version 2d' for variety; proofs of the other versions are similar.

We first show that we may assume \( g \neq 0 \) on \((a, \infty)\). We could give an argument using the Mean Value Theorem (cf. Exercise()), but there is a simpler argument: since \( \lim_{x \to +\infty} g(x) = \pm \infty \), there is automatically an interval \((r, +\infty)\) on which \( |g| > 1 \), say. Thus by increasing \( a \) if necessary we may assume that \( g \neq 0 \) on \((a, \infty)\).

Let \( L = \lim_{x \to +\infty} \frac{f'(x)}{g'(x)} \). As before, it suffices to show that if \((x_n)\) is a sequence in \((a, +\infty)\) with \( x_n \to +\infty \), then \( \lim_{n \to \infty} \frac{f(x_n)}{g(x_n)} = L \). We will assume \( L \in \mathbb{R} \); the case \( L = \pm \infty \) is an easy variation. By Cauchy’s Mean Value Theorem, for each \( m \) and \( n \) for which \( x_m \neq x_n \) there is a \( d_{m,n} \) between \( x_m \) and \( x_n \) such that

\[
\frac{f(x_n) - f(x_m)}{g(x_n) - g(x_m)} = \frac{f'(d_{m,n})}{g'(d_{m,n})}.
\]

We then have

\[
\frac{f(x_n)}{g(x_n)} - L = \left[ \frac{f'(d_{m,n})}{g'(d_{m,n})} - L \right] \cdot \frac{g'(d_{m,n})}{g'(d_{m,n})} \frac{g(x_n)}{g'(d_{m,n})} + \frac{f'(d_{m,n})}{g'(d_{m,n})} \frac{g(x_m) - g(x_n)}{g'(d_{m,n})}.
\]

Fix \( \epsilon > 0 \) and let \( R \) such that

\[
\left| \frac{f'(x)}{g'(x)} - L \right| < \frac{\epsilon}{3}
\]

for \( x > R \), and fix \( m_0 \) such that \( x_n > R \) for all \( n \geq m_0 \). Thus, for \( n > m_0 \),

\[
\left| \frac{f'(d_{m_0,n})}{g'(d_{m_0,n})} - L \right| < \frac{\epsilon}{3}.
\]

Set \( M = |L| + \frac{\epsilon}{3} \). Since \( \lim_{x \to +\infty} |g(x)| = +\infty \), choose \( N > m_0 \) such that \( \left| \frac{f(x_{m_0})}{g(x_{m_0})} \right| < \frac{\epsilon}{3} \) and \( \left| \frac{g(x_{m_0})}{g(x_n)} \right| < \frac{\epsilon}{3M} \) for all \( n \geq N \). Then, for \( n \geq N \),

\[
\left| \frac{f(x_n)}{g(x_n)} - L \right| \leq \left| \frac{f'(d_{m_0,n})}{g'(d_{m_0,n})} - L \right| + \left| \frac{f'(d_{m_0,n})}{g'(d_{m_0,n})} \right| \left| \frac{g(x_{m_0})}{g'(d_{m_0,n})} \right| + \left| \frac{f(x_{m_0})}{g(x_{m_0})} \right| < \epsilon.
\]

\( \Box \)
V.9.2.12. In fact, if \( g' \) is never zero and 
\[
\lim_{x \to x_0} \frac{f'(x)}{g'(x)}
\]
exists, then \( \lim_{x \to x_0} \frac{f(x)}{g(x)} \) always exists, but does not necessarily equal \( \lim_{x \to x_0} \frac{f'(x)}{g'(x)} \) unless \( \lim_{x \to x_0} f(x) = \lim_{x \to x_0} g(x) = 0 \) or \( \lim_{x \to x_0} g(x) = \pm \infty \) (Exercise V.9.4.9.; statements apply to any kind of one-sided limits, but not two-sided).

V.9.2.13. Version 2 of l’Hôpital’s Rule (in its various forms) is commonly used even when Version 1 applies. There are also situations where Version 2 can be used although Version 1 does not apply. (The reverse can also happen: there are situations where Version 1 applies but Version 2 does not; cf. Exercise V.9.4.4.)

V.9.3. Applications

V.9.3.1. Examples. (i) As in V.9.1.2., we can apply Version 2c to obtain
\[
\lim_{x \to 0} \frac{e^x - 1}{\sin x} = \lim_{x \to 0} \frac{e^x}{\cos x} = 1 .
\]

(ii) An example of a limit where Version 1 does not apply but Version 2 does is:
\[
\lim_{x \to 0^+} \frac{\sin x}{\sqrt{x}}.
\]

Version 2a says that if
\[
\lim_{x \to 0^+} \frac{\cos x}{\frac{1}{2}x^{-1/2}} = \lim_{x \to 0^+} 2\sqrt{x} \cos x
\]
exists, the first limit also exists and has the same value. But the second limit is 0 by (); thus
\[
\lim_{x \to 0^+} \frac{\sin x}{\sqrt{x}} = 0 .
\]

This limit could have also been evaluated by
\[
\lim_{x \to 0^+} \frac{\sin x}{\sqrt{x}} = \lim_{x \to 0^+} \left[ \sqrt{x} \cdot \frac{\sin x}{x} \right] = 0 \cdot 1 = 0 .
\]

(iii) Besides applications of Versions 2d and 2a’–2d’, which have no Version 1 analogues, the most common situation where Version 2 is necessary (i.e. Version 1 does not apply) is illustrated by the following example. Let \( f(x) = 1 - \cos x, \ g(x) = x^2 \). Then
\[
\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{1 - \cos x}{x^2}
\]
is an indeterminate form of type \( \frac{0}{0} \) to which we might try to apply l’Hôpital’s Rule. If
\[
\lim_{x \to 0} \frac{f'(x)}{g'(x)} = \lim_{x \to 0} \frac{\sin x}{2x}
\]
exists, then Version 2c implies that the first limit exists and has the same value. The second limit is also an indeterminate form of type $0/0$, so we may apply Version 2c again to convert it to

$$\lim_{x \to 0} \frac{f''(x)}{g''(x)} = \lim_{x \to 0} \frac{\cos x}{2}$$

i.e. if this limit exists, then the second limit and thus also the first limit exists and has the same value. But the last limit is not an indeterminate form, and equals $1/2$ by direct evaluation. We usually just abbreviate the whole argument as

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \lim_{x \to 0} \frac{\sin x}{2x} = \lim_{x \to 0} \frac{\cos x}{2} = \frac{1}{2}$$

where the first two steps are justified by “l’Hôpital’s Rule” (specifically, Version 2c); logically, we technically do the second application first and use the result to justify the first application. We could also justify the first step by Version 2c and calculate the second limit directly by Version 1.

(iv) In applying l’Hôpital’s Rule, it must be checked that the hypotheses apply. This pitfall is particularly important to watch for when the Rule is used repeatedly as in (iii). For example, one might argue

$$\lim_{x \to 0} \frac{x^2}{x^2 + \sin x} = \lim_{x \to 0} \frac{2x}{2x + \cos x} = \lim_{x \to 0} \frac{2}{2 - \sin x} = \frac{2}{2} = 1.$$

However, the second limit is not an indeterminate form and the second application of l’Hôpital’s Rule is not justified. The second limit can be evaluated directly to be 0, and thus the first limit is also 0. See Exercise V.9.4.7. for a more subtle example where the hypotheses of l’Hôpital’s Rule are not satisfied.

Sometimes some cleverness is required to successfully use l’Hôpital’s Rule:

**V.9.3.2.** **Example.** Consider

$$\lim_{x \to 0^+} \frac{e^{-1/x}}{x}.$$

The hypotheses of Version 2a are satisfied with $f(x) = e^{-1/x}$ and $g(x) = x$. Thus the limit can be converted to

$$\lim_{x \to 0^+} \frac{e^{-1/x}}{x} = \lim_{x \to 0^+} \frac{e^{-1/x}}{x^2}.$$

But this limit is worse than the one we started with, and repeating l’Hôpital’s Rule will make things successively even worse.

Instead, we can rewrite the limit as

$$\lim_{x \to 0^+} \frac{1/x}{e^{1/x}}$$

and note that Version 2a’ applies to give

$$\lim_{x \to 0^+} \frac{1/x}{e^{1/x}} = \lim_{x \to 0^+} \frac{-1/x^2}{e^{-1/x} x} = \lim_{x \to 0^+} \frac{e^{-1/x}}{x} = 0.$$
We could also make the substitution $u = 1/x$ and use Version 2d' to give

$$
\lim_{x \to 0^+} \frac{1}{e^{1/x}} = \lim_{u \to +\infty} \frac{u}{e^u} = \lim_{u \to +\infty} \frac{1}{e^u} = 0.
$$

Indeterminate forms of other types can often be converted into a form to which l’Hôpital’s rule can be applied, by algebraic manipulation or taking logarithms or exponentials:

V.9.3.3. Examples. (i) $\lim_{x \to +\infty} xe^{-x}$ is an indeterminate form of type $\infty \cdot 0$. It can be converted to $\lim_{x \to +\infty} x$ and evaluated by Version 2d’. Another similar example is in Exercise (d).

(ii) $\lim_{x \to 0^+} \left[ \frac{1}{\sin x} - \frac{1}{x} \right]$ is an indeterminate form of type $\infty - \infty$. It can be converted into a fraction and evaluated by

$$
\lim_{x \to 0^+} \frac{1 - \cos x}{\sin x + x \cos x} = \lim_{x \to 0^+} \frac{\sin x}{2 \cos x - x \sin x} = 0
$$

by two applications of l’Hôpital’s Rule. (For the first application, it must be verified that $h(x) = \sin x + x \cos x$ is nonzero in a deleted neighborhood of 0; this follows from the fact that $h'(0) \neq 0$.)

(iii) $\lim_{x \to 0^+} x^x$ is an indeterminate form of type $0^0$. By taking logarithms we can reduce the problem to computing $\lim_{x \to 0^+} x \log x$, which is 0 (Exercise (d)). Thus

$$
\lim_{x \to 0^+} x^x = \lim_{x \to 0^+} e^{x \log x} = e^0 = 1
$$

by continuity of the exponential function at 0.

(iv) $\lim_{x \to +\infty} \left(1 + \frac{1}{x}\right)^x$ is of the form $1^\infty$. Taking logarithms, we reduce to evaluating $\lim_{x \to +\infty} x \log \left(1 + \frac{1}{x}\right)$. This is of type $\infty \cdot 0$, so we can convert it to

$$
\lim_{x \to +\infty} \frac{\log(1 + 1/x)}{1/x} = \lim_{x \to +\infty} \frac{1 + \frac{1}{x}(-1/x^2)}{1 - 1/x^2} = \lim_{x \to +\infty} \frac{1}{1 + 1/x} = 1
$$

so

$$
\lim_{x \to +\infty} \left(1 + \frac{1}{x}\right)^x = e^1 = e
$$

again by continuity of the exponential function. (This example is somewhat artificial, since direct evaluation of this limit was used in defining $e$ and the exponential function.)
V.9.4. Exercises

V.9.4.1. Express and justify Version 1 of l’Hôpital’s Rule using little-o notation (V.7.) by writing $f(x) = f'(a)(x-a) + o(x-a)$ and $g(x) = g'(a)(x-a) + o(x-a)$.

V.9.4.2. Show that Version 1 and Version 2c of l’Hôpital’s Rule are equivalent in the (common) case where $f$ and $g$ are differentiable in a neighborhood of $a$, $f'$ and $g'$ are continuous at $a$, $f(a) = g(a) = 0$, and $g'(a) \neq 0$.

V.9.4.3. Let $f(x) = x^2$ for $x \in \mathbb{Q}$ and $f(x) = 0$ for $x \not\in \mathbb{Q}$, and let $g(x) = x$ for all $x$. Then Version 1 of l’Hôpital’s Rule applies for $a = 0$. But $f$ is not differentiable, or even continuous, except at 0, so no variant of Version 2 applies.

V.9.4.4. Let $f(x) = x^2 \sin \frac{1}{x}$ for $x \neq 0$, and $f(0) = 0$; and let $g(x) = \sin x$. Then $f$ and $g$ are differentiable everywhere, and $f'(0) = 0$, $g'(0) = 1$.

(a) Version 1 of l’Hôpital’s Rule applies to show that
\[
\lim_{x \to 0} \frac{f(x)}{g(x)} = \lim_{x \to 0} \frac{x^2 \sin \frac{1}{x}}{\sin x} = 0.
\]

(b) $\lim_{x \to 0} f'(x) = \lim_{x \to 0} \left[2x \sin \frac{1}{x} - \frac{2}{x^2 \sin \frac{1}{x}} \right]$ and $\lim_{x \to 0} \frac{f'(x)}{g(x)}$ do not exist. In particular, Version 2 of l’Hôpital’s rule does not apply.

(c) The existence of $\lim_{x \to a} \frac{f(x)}{g(x)}$ does not imply the existence of $\lim_{x \to a} \frac{f'(x)}{g'(x)}$ even for everywhere differentiable $f$ and $g$ with $g' \neq 0$ in a neighborhood of $a$. Thus Version 2 is not more general than Version 1 even for functions which are everywhere differentiable.

(d) There are simpler examples for limits at $\infty$: for example, let $f(x) = x + \sin x$, $g(x) = x$. Then $\lim_{x \to \infty} f(x) = \lim_{x \to \infty} g(x) = +\infty$ and $\lim_{x \to \infty} \frac{f(x)}{g(x)}$ exists and equals 1, but $\lim_{x \to \infty} \frac{f'(x)}{g'(x)}$ does not exist.

V.9.4.5. ([?]; cf. [?]) What is wrong with the following argument?

(a) Let $f(x) = x + \sin x \cos x$. Show that $f'(x) = 2 \cos^2 x$.

(b) Let $g(x) = e^{\sin^2 x} f(x)$. Show that $\lim_{x \to \infty} f(x) = \lim_{x \to \infty} g(x) = +\infty$.

(c) We have
\[
\lim_{x \to \infty} \frac{f'(x)}{g'(x)} = \lim_{x \to \infty} \frac{2 \cos^2 x}{e^{\sin x} [f(x) \cos x + 2 \cos^2 x]} = \lim_{x \to \infty} \frac{2 \cos x}{e^{\sin x} [f(x) + 2 \cos x]} = 0
\]
since the numerator is bounded and the denominator approaches $+\infty$. Thus, by l’Hôpital’s Rule, Version 2d',
\[
\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} e^{-\sin x} = 0.
\]

But $e^{-1} \leq e^{-\sin x} \leq e$ for all $x$, and $\lim_{x \to \infty} e^{-\sin x}$ does not exist.
V.9.4.6.  
(a) Using l'Hôpital's Rule, prove that, for every $a > 0$ and $r \in \mathbb{R}$,
\[
\lim_{x \to +\infty} e^{ax} x^r = +\infty.
\]
[Prove this by induction for $r \in \mathbb{N}$, and use the Squeeze Theorem for other $r$.]
(b) Using l'Hôpital's Rule, prove that for all $r > 0$,
\[
\lim_{x \to +\infty} \frac{\log x}{x^r} = 0.
\]
(c) Use (b) to give a proof of (a) not requiring induction.
(d) Using l'Hôpital's Rule, show that for all $r > 0$,
\[
\lim_{x \to 0+} x^r \log x = \lim_{x \to 0+} \frac{\log x}{x^{-r}} = 0.
\]
For a proof of (a) not using l'Hôpital's Rule, note that, for every $x$,
\[
e^{ax} = \sum_{k=0}^{\infty} \frac{a^k x^k}{k!}
\]
and all terms are positive if $a > 0$ and $x > 0$, and hence for any $n$ and $m$ we have
\[
e^{ax} > \frac{a^{n+m} x^{n+m}}{(n+m)!}
\]
\[
e^{ax} x^n > \frac{a^{n+m}}{(n+m)!} x^m.
\]

V.9.4.7.  
[?] Let $f(x) = x \sin \left(\frac{1}{x^2}\right) e^{-1/x^2}$ and $g(x) = e^{-1/x^2}$ for $x \neq 0$ and $f(0) = g(0) = 0$.
(a) Show that $f$ and $g$ are $C^\infty$ on $\mathbb{R}$, and that $g(x)$ and $g'(x)$ are nonzero for $x \neq 0$.
(b) Show that $\lim_{x \to 0} \frac{f(x)}{g(x)} = 0$.
(c) Show that $\frac{f'(x)}{g'(x)}$ is unbounded in every neighborhood of 0, and that $\lim_{x \to 0} \frac{f'(x)}{g'(x)}$ does not exist.

V.9.4.8.  
Suppose $f$ and $g$ satisfy the hypotheses Version 2d' of l'Hôpital's Rule. Recall (V.8.2.14.) that $g'$ does not change sign on $(a, +\infty)$.
(a) Suppose $g' > 0$ on $(a, \infty)$. If $\lim_{x \to +\infty} g(x) = +\infty$ (it cannot be $-\infty$!), and $\lim_{x \to +\infty} \frac{f'(x)}{g'(x)}$ exists and equals $L$, $0 < L \leq +\infty$, then there is an $M > 0$ such that $f'(x) > Mg'(x)$ for sufficiently large $x$. Show that this implies $\lim_{x \to +\infty} f(x) = +\infty$ [$f - Mg$ is eventually increasing]. Modify the argument for the case $g' < 0$ on $(a, \infty)$ and/or $-\infty \leq L < 0$.
(b) Modify the argument to apply to Version 2a', and similarly for Version 2b' and 2c'.

619
V.9.4.9.  (a) Suppose $f$ and $g$ are differentiable on $(a, +\infty)$ and $g'$ is never zero, and that $\lim_{x \to +\infty} \frac{f'(x)}{g'(x)}$ exists and is finite. If $\lim_{x \to +\infty} g(x) = \pm \infty$ is false, show that $\lim_{x \to +\infty} g(x)$ and $\lim_{x \to +\infty} f(x)$ exist and are finite, and $\lim_{x \to +\infty} \frac{f(x)}{g(x)}$ exists (possibly in the extended sense). [Note that $g$ must be strictly monotone. For some $M$, $M|g| \pm f$ are eventually increasing.] If $\lim_{x \to +\infty} g(x) \neq 0$, the condition that $\lim_{x \to +\infty} \frac{f'(x)}{g'(x)}$ exists can be relaxed to just a requirement that $\frac{f(x)}{g(x)}$ be bounded.

(b) Show that $\lim_{x \to +\infty} \frac{f(x)}{g(x)}$ is not necessarily the same as $\lim_{x \to +\infty} \frac{f'(x)}{g'(x)}$ (which need not even exist if $\lim_{x \to +\infty} g(x) \neq 0$), by considering
\[ \lim_{x \to +\infty} \frac{1}{\frac{1}{x} - \frac{1}{2}} \quad \text{or} \quad \lim_{x \to +\infty} \frac{1 - \frac{1}{x^2}}{1 - \frac{1}{x}} \quad \text{or} \quad \lim_{x \to +\infty} \frac{1 - \sin \frac{x}{x}}{1 - \frac{1}{x}} \]
or by noting that constants can be added to $f$ and/or $g$ without affecting the hypotheses. In particular, the assumption $\lim_{x \to +\infty} g(x) = \pm \infty$ cannot be eliminated from Version 2d'.

(c) Suppose $f$ and $g$ are differentiable on $(a, +\infty)$ and $g'$ is never zero, and that $\lim_{x \to +\infty} \frac{f'(x)}{g'(x)} = \pm \infty$. If $\lim_{x \to +\infty} g(x) = \pm \infty$ is false, show that $\lim_{x \to +\infty} g(x)$ and $\lim_{x \to +\infty} f(x)$ exist, $\lim_{x \to +\infty} g(x)$ is finite, and $\lim_{x \to +\infty} \frac{f(x)}{g(x)}$ exists (possibly, but not necessarily, in the extended sense).

(d) Obtain versions for $\lim_{x \to -\infty}$ and, for an interval $(a, b)$, for $\lim_{x \to a^+}$ and $\lim_{x \to b^-}$. (There is no good two-sided version.)

V.9.4.10.  Show that hypothesis (1) in Version 2c' can be relaxed to

$$ (1') \lim_{x \to a} |g(x)| = +\infty . $$

(The same change can be made in Versions 2a', 2b', and 2d', but this is not an actual change in these cases by V.8.2.14.)

V.9.4.11.  Let $I$ be an open interval, $c \in I$, and let $f$ be continuous on $I$ and differentiable on $I \setminus \{c\}$. If $\lim_{x \to c} f'(x)$ exists and equals $L \in \mathbb{R}$, show that $f$ is necessarily differentiable at $c$ and that $f'(c) = L$. [Apply l'Hôpital's Rule Version 2c to $\lim_{x \to c} \frac{f(x) - f(c)}{x - c}$.] Give a similar statement for one-sided limits.

Compare with V.8.2.15. and V.12.2.5.

V.9.4.12.  Let $x_0 \in \mathbb{R}$, and $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ and $\sum_{k=0}^{\infty} b_k (x - x_0)^k$ be power series () centered at $x_0$ with positive radius of convergence, defining analytic functions $f$ and $g$ in an open interval around $x_0$. Let $a_m$ be the first nonzero coefficient in the first series, i.e. $a_m \neq 0$ and $a_k = 0$ for $k < m$. Similarly, let $b_n$ be the first nonzero coefficient in the second series.
(a) Use V.15.3.26. and l’Hôpital’s Rule to show that if \( m \geq n \), then
\[
\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{a_n}{b_n}
\]
and in particular, the limit always exists (in the usual sense), and is zero if \( m > n \) and nonzero if \( m = n \).

(b) Obtain the same result from basic limit and power series theorems by first dividing \( f \) and \( g \) by a suitable power of \( x - x_0 \), and show additionally that
\[
\lim_{x \to x_0} \frac{f(x)}{g(x)}
\]
does not exist (in the usual sense) if \( m < n \).

(c) Also conclude the result from (), although this argument is less elementary (but shows more).

V.9.4.13. [?, 11.10] Find
\[
\lim_{x \to 0} \frac{\sin(\tan(x)) - \tan(\sin(x))}{\arcsin(\arctan(x)) - \arctan(\arcsin(x))}
\]
[To apply l’Hôpital’s Rule, or Exercise V.9.4.12., try expanding numerator and denominator in Maclaurin series; a computer algebra system can be very helpful.] Does graphing the function near 0 with a computer algebra system help in guessing the answer?

V.9.4.14. [?, 4.4] Let
\[
h(x) = \frac{x}{\sqrt{x^2 + 1}}
\]
and note that \( h(x) > 0 \) for \( x > 0 \). Applying (), (), and l’Hôpital’s Rule Version 2d’ to \( \lim_{x \to +\infty} h(x) \), the only conclusion that can be drawn is that either the limit is 1 or it does not exist. Use another method to determine which is the case.

V.9.4.15. Let \( f \) be a function defined on an open interval around \( a \).

(a) Suppose \( f''(a) \) exists. Show that
\[
\lim_{h \to 0} \frac{f(a + h) - 2f(a) + f(a - h)}{h^2}
\]
exists and equals \( f''(a) \).

(b) Show that the limit in (a) can exist even if \( f''(a) \) does not exist, and even if \( f \) is not continuous at \( a \).
[Consider an odd function with \( a = 0 \).]
V.10. Taylor’s Theorem

Taylor’s Theorem is one of the most important and useful results in calculus. It can be thought of as a higher-order version of differential approximation or of the Mean Value Theorem, and is crucial in representing functions by power series, which will be discussed later. There are many other applications of both theoretical and practical importance.

The statement and a “proof” of Taylor’s Theorem were first published by Brook Taylor in 1715, although the result was previously known to several mathematicians, and apparently first obtained by Gregory about 1670. Taylor’s “proof” fell far short of modern standards of precision and rigor. For an account of the history of the result, see[].

V.10.1. Taylor Polynomials

V.10.1.1. Suppose $f$ is a function which is differentiable to some high order near a number $x_0$. We would like to approximate $f$ near $x_0$ as closely as possible by a polynomial of fixed degree. It makes intuitive sense that the approximating polynomial should duplicate the values of $f$ and its derivatives at $x_0$ as closely as possible, since these values should determine the location and shape of the graph of $f$ near $x_0$. There is a limit to how far this can be done, since if $p$ is a polynomial of degree $n$, then $p^{(k)}$ is identically 0 for $k > n$; so we can only hope to make the derivatives of a degree $n$ polynomial up through the $n$th derivative agree with those of $f$. In differential approximation, which is the case $n = 1$, we have seen that there is a unique polynomial $p_1$ of degree $\leq 1$ (i.e. linear or affine function) with $p_1(x_0) = f(x_0)$ and $p_1'(x_0) = f'(x_0)$, namely $p_1(x) = f(x_0) + f'(x_0)(x - x_0)$, and $p_1$ is indeed the polynomial of degree $\leq 1$ most closely approximating $f$ near $x_0$, in the sense that

$$\lim_{x \to x_0} \frac{f(x) - p_1(x)}{x - x_0} = 0$$

and $p_1$ is the only linear function with this property.

The higher-order case gets its justification from the following result:

V.10.1.2. THEOREM. Let $g$ be a real-valued function on an interval $I$, and $x_0$ an interior point of $I$, and $n \geq 1$. Suppose $g$ is $n$-times differentiable at $x_0$ (this implies that for all $k < n$, $g^{(k)}$ is defined in a neighborhood of $x_0$ and is continuous at $x_0$, and, for $1 \leq r \leq n - 1$, $g^{(n-r-1)}$ is $r$-times differentiable in a neighborhood of $x_0$). Then

$$\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n}$$

exists if and only if $g^{(k)}(x_0) = 0$ for $0 \leq k < n$. If the limit exists, it equals $\frac{g^{(n)}(x_0)}{n!}$.

V.10.1.3. COROLLARY. Let $g$ be a real-valued function on an interval $I$, and $x_0$ an interior point of $I$, and $n \geq 1$. Suppose $g$ is $n$-times differentiable at $x_0$ (this implies that for all $k < n$, $g^{(k)}$ is defined in a neighborhood of $x_0$ and is continuous at $x_0$, and, for $1 \leq r \leq n - 1$, $g^{(n-r-1)}$ is $r$-times differentiable in a neighborhood of $x_0$). Then

$$\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n} = 0$$

if and only if $g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0$. 

622
Proof: By (complete) induction on \( n \). First suppose \( n = 1 \). By assumption, \( g \) is differentiable at \( x_0 \) and hence continuous at \( x_0 \). If

\[
\lim_{x \to x_0} \frac{g(x)}{x - x_0}
\]

exists, then \( \lim_{x \to x_0} g(x) \) must be 0 since \( \lim_{x \to x_0} (x - x_0) = 0 \). Thus \( g(x_0) = 0 \). And if \( g(x_0) = 0 \), then

\[
\lim_{x \to x_0} \frac{g(x)}{x - x_0} = \lim_{x \to x_0} \frac{g(x) - g(x_0)}{x - x_0}
\]

exists and equals \( g'(x_0) \) by definition.

Now let \( n > 1 \). Assume the result for all \( k \), \( 1 \leq k < n \), and suppose \( g \) is \( n \)-times differentiable at \( x_0 \). Suppose

\[
\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n}
\]

exists and equals \( L \). Then, if \( 0 \leq k < n \), we have

\[
\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n} = \left[ \lim_{x \to x_0} \frac{g(x)}{(x - x_0)^{n-k}} \right] \left[ \lim_{x \to x_0} (x - x_0)^{n-k} \right] = L \cdot 0 = 0
\]

and thus by the inductive hypothesis \( g^{(k)}(x_0) = 0 \) (if \( k = 0 \), we obtain \( g(x_0) = 0 \) by continuity of \( g \) at \( x_0 \)). Conversely, if \( g(x_0) = g'(x_0) = \cdots = g^{(n-1)}(x_0) = 0 \), we apply l'Hôpital’s rule Version 2c (V.9.2.6.): if

\[
\lim_{x \to x_0} \frac{g(x)}{n(x - x_0)^{n-1}}
\]

exists and equals \( L \), then

\[
\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n}
\]

also exists and equals \( L \). But the inductive hypothesis applies to \( g' \) since it is \( (n-1) \)-times differentiable at \( x_0 \) and \( g'(x_0) = (g')'(x_0) = \cdots = (g')^{(n-2)}(x_0) = 0 \), so the first limit exists and

\[
L = \frac{1}{n} \frac{(g')^{(n-1)}(x_0)}{(n-1)!} = \frac{g^{(n)}(x_0)}{n!}.
\]

\( \blacksquare \)

V.10.1.4. In particular, if \( p \) is a polynomial of degree \( \leq n \) and \( f \) is \( n \)-times differentiable at \( x_0 \), then

\[
\lim_{x \to x_0} \frac{f(x) - p(x)}{(x - x_0)^n} = 0
\]

if and only if \( p^{(k)}(x_0) = f^{(k)}(x_0) \) for \( 0 \leq k \leq n \). Such a polynomial (if it exists) is unique, and is the “best \( n \)’th order polynomial approximation” to \( f \) near \( x_0 \).

So the problem is to construct a polynomial of degree \( \leq n \) with specified derivatives at \( x_0 \). This is easily done:
V.10.1.5. Proposition. Let \( p(x) = a_0 + a_1 x + \cdots + a_n x^n \) be a polynomial of degree \( n \). Then, for \( 0 \leq k \leq n, f^{(k)}(0) = k! a_k \).

The proof is a simple calculation. If derivatives at \( x_0 \) instead of 0 are desired, a simple trick does the shift: if
\[
q(x) = p(x - x_0) = a_0 + a_1 (x - x_0) + \cdots + a_n (x - x_0)^n,
\]
then \( q \) is also a polynomial of degree \( \leq n \) and \( q^{(k)}(x_0) = k! a_k \).

V.10.1.6. Thus, if \( f \) is \( n \)-times differentiable at \( x_0 \) and we want a polynomial \( p_n \) of degree \( \leq n \) whose derivatives at \( x_0 \) up to the \( n \)’th are the same as those of \( f \), we may (and must) take
\[
p_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!} (x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n
\]
This is the \( n \)’th Taylor Polynomial of \( f \) at \( x_0 \) (it will have degree strictly less than \( n \) if \( f^{(n)}(x_0) = 0 \)).

Although our notation does not reflect it, note that \( p_n \) depends not only on \( n \) (and, of course, \( f \)), but also on the choice of \( x_0 \), i.e. the function \( f \) will generally have different \( n \)’th order Taylor polynomials at different points (just as the best linear or affine approximation to \( f \) generally varies from point to point.)

We summarize our work in the next theorem (see V.10.1.2. for the last statement):

V.10.1.7. Theorem. [Taylor’s Theorem] Let \( f \) be defined in a neighborhood of \( x_0 \), and \( n \in \mathbb{N} \). Suppose \( f^{(n)}(x_0) \) exists (this implies that for all \( k < n \), \( f^{(k)} \) is defined in a neighborhood of \( x_0 \) and is continuous at \( x_0 \), and, for \( 1 \leq r \leq n - 1 \), \( f^{(n-r)} \) is \( r \)-times differentiable in a neighborhood of \( x_0 \)). Then the \( n \)’th Taylor polynomial \( p_n \) of \( f \) at \( x_0 \) satisfies
\[
\lim_{x \to x_0} \frac{f(x) - p_n(x)}{(x - x_0)^n} = 0
\]
and \( p_n \) is the unique polynomial of degree \( \leq n \) with this property. If \( f \) is \( (n + 1) \)-times differentiable at \( x_0 \), then
\[
\lim_{x \to x_0} \frac{f(x) - p_n(x)}{(x - x_0)^{n+1}} = \frac{f^{(n+1)}(x_0)}{(n + 1)!}.
\]

Note that if \( n = 1 \), we obtain the differential approximation polynomial \( p_1 \). Thus the Taylor polynomials can be regarded as giving “higher-order differential approximation.”

V.10.1.8. The Taylor polynomials thus give good approximations to \( f \) sufficiently near \( x_0 \); rephrasing the limit characterization, for any \( \epsilon > 0 \) and \( n \in \mathbb{N} \) there is a \( \delta > 0 \) such that
\[
|f(x) - p_n(x)| \leq \epsilon |x - x_0|^n \quad \text{whenever} \quad |x - x_0| < \delta
\]
and thus, for \( x \) close enough to \( x_0 \), the difference between \( f(x) \) and \( p_n(x) \) is small even compared to the very small number \( |x - x_0|^n \).

In general, the \( \delta \) depends on \( x_0 \) and \( n \), as well as on \( f \) and \( \epsilon \). As a result, for a fixed \( x \) near \( x_0 \) but different from \( x_0 \), we should not necessarily expect \( p_n(x) \) to be a good approximation to \( f(x) \) for any particular \( n \), or
even that \( p_n(x) \) becomes a better and better approximation to \( f(x) \) as \( n \) increases, although fortunately this often happens. In fact, to paraphrase [Act90], the Taylor polynomials give a very, very good approximation close enough to \( x_0 \), then get increasingly worse as we move away, and finally (often) explode into complete unusability. These points will be discussed more carefully later.

### V.10.2. Vanishing Order

#### V.10.2.1. Definition. Let \( f \) be a function defined in a deleted neighborhood of \( x_0 \), and let \( \alpha \in [0, \infty) \). Then \( f \) **vanishes to order** \( \alpha \) at \( x_0 \) if

\[
|f(x)| = O(|x - x_0|^{\alpha}) \quad \text{as} \quad x \to x_0,
\]

i.e. there are positive \( K \) and \( \delta \) such that

\[
|f(x)| \leq K|x - x_0|^{\alpha} \quad \text{for} \quad 0 < |x - x_0| < \delta.
\]

We say \( f \) **vanishes to order** \( \alpha^+ \) at \( x_0 \) if

\[
f(x) = o(|x - x_0|^{\alpha}) \quad \text{as} \quad x \to x_0.
\]

If \( f \) vanishes to order \( \alpha^+ \) at \( x_0 \), then it vanishes to order \( \alpha \). If it vanishes to order \( \beta \) for some \( \beta > \alpha \) at \( x_0 \), then it vanishes to order \( \alpha^+ \). A function vanishes to order \( 0^+ \) at \( x_0 \) if and only if it can be extended to be continuous at \( x_0 \) by setting \( f(x_0) = 0 \).

#### V.10.2.2. Examples. (i) The function \( f(x) = |x - x_0|^{\alpha} \) vanishes at \( x_0 \) to order \( \alpha \) but not to order \( \alpha^+ \).

(ii) The function \( f(x) = \frac{|x-x_0|^{\alpha}}{\log|x-x_0|} \) vanishes at \( x_0 \) to order \( \alpha^+ \), but not to order \( \beta \) for any \( \beta > \alpha \).

Taylor’s Theorem can be rephrased (the last statement becomes slightly weaker):

#### V.10.2.3. Theorem. [Taylor’s Theorem] Let \( f \) be \( n \)-times differentiable at \( x_0 \), and let \( p_n \) be the \( n \)’th order Taylor polynomial of \( f \) at \( x_0 \). Then \( f - p_n \) vanishes to order \( n^+ \) at \( x_0 \), and \( p_n \) is the unique polynomial of degree \( \leq n \) with this property. If \( f \) is \((n+1)\)-times differentiable at \( x_0 \), then \( f - p_n \) vanishes to order \( n+1 \) at \( x_0 \).

### V.10.3. Remainder Formulas

#### V.10.3.1. It is convenient to think of the situation in terms of the **remainder function** \( r_n = f - p_n \). Thus

\[
r_n(x) = f(x) - p_n(x) = f(x) - f(x_0) - f'(x_0)(x-x_0) - \cdots - \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n
\]

(like \( p_n \), the function \( r_n \) depends on \( x_0 \) as well as on \( f \) and \( n \)). The approximation in Taylor’s Theorem can be summarized by

\[
r_n(x) = o(|x - x_0|^{\alpha}) \quad \text{as} \quad x \to x_0.
\]

However, for practical applications of Taylor’s approximation, as well as for many theoretical applications, it is important to have specific estimates for the size of \( r_n(x) \). For example, one might expect that for fixed \( x \) near \( x_0 \), \( r_n(x) \to 0 \) as \( n \to \infty \), i.e. that the Taylor polynomials for \( f \) around \( x_0 \) converge to \( f \) pointwise near
for some times differentiable on an interval formula due to Lagrange Remainder Formula can be regarded as a higher-order generalization of the Mean Value Theorem.

There are many formulas known for \( r_n(x) \). The first and still most widely used was found by Joseph Lagrange about 100 years after Taylor’s work (published 1813). Lagrange’s formula and another useful formula due to Cauchy can be derived by the same procedure using the Mean Value Theorem.

Good formulas for the remainder function can only be obtained when \( f \) is \((n+1)\)-times differentiable in a neighborhood of \( x_0 \). Let us assume this, and fix \( x \) in such a neighborhood. For specificity we will assume that \( x > x_0 \), but an essentially identical argument yields the same formulas in the case \( x < x_0 \).

**V.10.3.2.** The key trick is to keep \( x \) fixed in the formula for \( r_n \), and replace \( x_0 \) (which we also regard as fixed) by a variable \( t \). Set

\[
\phi(t) = f(x) - f(t) - f'(t)(x-t) - \cdots - \frac{f^{(n)}(t)}{n!} (x-t)^n - r_n(x) \left( \frac{x-t}{x-x_0} \right)^{n+1}
\]

for \( x_0 \leq t \leq x \). Then \( \phi \) is differentiable on \([x_0,x]\), \( \phi(x_0) = \phi(x) = 0 \), and

\[
\phi'(t) = -f'(t) - [f''(t)(x-t) - f''(t)] - \cdots - \left[ \frac{f^{(n+1)}(t)}{n!} (x-t)^n - \frac{f^{(n)}(t)}{(n-1)!} (x-t)^{n-1} \right] + \frac{r_n(x)}{(x-x_0)^{n+1}} (n+1)(x-t)^{n+1}
\]

\[
= -\frac{f^{(n+1)}(t)}{n!} (x-t)^n + \frac{r_n(x)}{(x-x_0)^{n+1}} (n+1)(x-t)^n.
\]

By Rolle’s Theorem, there is a \( c \in (x_0,x) \) with \( \phi'(c) = 0 \). We then have

\[
r_n(x) = \frac{(x-x_0)^{n+1}}{(n+1)(x-c)^n} \frac{f^{(n+1)}(c)}{n!} (x-c)^n = \frac{f^{(n+1)}(c)}{(n+1)!} (x-x_0)^{n+1}.
\]

So we obtain:

**V.10.3.3.** Theorem. [Taylor’s Theorem With Remainder, Lagrange’s Form] Let \( f \) be \((n+1)\)-times differentiable on an interval \( I \), \( x_0 \in I \), \( p_n \) the \( n \)th Taylor polynomial for \( f \) at \( x_0 \), and \( x \in I \). Then, for some \( c \) between \( x_0 \) and \( x \),

\[
r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x-x_0)^{n+1}
\]

or, equivalently,

\[
f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!} (x-x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!} (x-x_0)^{n+1}.
\]

**V.10.3.4.** The case \( n = 0 \) is just the usual Mean Value Theorem. Thus Taylor’s Theorem with the Lagrange Remainder Formula can be regarded as a higher-order generalization of the Mean Value Theorem.
V.10.3.5. We can obtain another formula for the remainder by a more straightforward variation of the argument: set

$$\psi(t) = f(x) - f(t) - f'(t)(x-t) - \cdots - \frac{f^{(n)}(t)}{n!}(x-t)^n$$

for $x_0 \leq t \leq x$. Then $\psi$ is differentiable on $[x_0, x]$, $\psi(x_0) = r_n(x)$, $\psi(x) = 0$, and

$$\psi'(t) = -f'(t) - [f''(t)(x-t) - f'(t)] - \cdots - \left[ \frac{f^{(n+1)}(t)}{n!} (x-t)^n - \frac{f^{(n)}(t)}{(n-1)!} (x-t)^{n-1} \right]$$

$$= -\frac{f^{(n+1)}(t)}{n!} (x-t)^n.$$

We apply the Mean Value Theorem to $\psi$ to obtain a formula for $r_n(x)$:

V.10.3.6. **Theorem.** Let $f$ be $(n+1)$-times differentiable on an interval $I$, $x_0 \in I$, $p_n$ the $n$’th Taylor polynomial for $f$ at $x_0$, and $x \in I$. Then, for some $c$ between $x_0$ and $x$,

$$r_n(x) = f(x) - p_n(x) = \phi(x_0) - \phi(x) = \phi'(c)(x_0 - x) = \frac{f^{(n+1)}(c)}{n!} (x-c)^n (x-x_0).$$

This formula for the remainder is usually written in a different form. Since $c$ is between $x_0$ and $x$, $c = x_0 + \theta(x-x_0)$ for some $\theta$, $0 < \theta < 1$. Then $x-c = x-x_0 - \theta(x-x_0) = (1-\theta)(x-x_0)$, so

V.10.3.7. **Corollary.** [Taylor’s Theorem With Remainder, Cauchy’s Form] Let $f$ be $(n+1)$-times differentiable on an interval $I$, $x_0 \in I$, $p_n$ the $n$’th Taylor polynomial for $f$ at $x_0$, and $x \in I$. Then, for some $\theta$ between 0 and 1,

$$r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(x_0 + \theta(x-x_0))}{n!} (1-\theta)^n (x-x_0)^{n+1}.$$

V.10.3.8. Note that the $c$ is not necessarily unique, and depends on $n$, as well as on $f$, $x$, and $x_0$, and on which formula (Cauchy or Lagrange) is used. There is usually no reasonable way to explicitly find a $c$ that works in a given example. However, as we will see, mere existence of such a $c$ often gives good error estimates and crucial information about the degree of approximation given by the Taylor polynomials.

V.10.3.9. See V.11.6.7. for a generalization of V.10.3.3. to the case where $f$ is approximated at more than one point.
V.10.4. The Integral Remainder Formula

V.10.4.1. There is another formula for \( r_n \) given by an integral. Let us fix \( f, x_0 \) and \( x \). Let \( J \) be the closed interval from \( x_0 \) to \( x \) (\( x \) may be greater than or less than \( x_0 \)). If \( f \) is differentiable on \( J \) and \( f' \) is Riemann integrable on \( J \), then

\[
f(x) = f(x_0) + \int_{x_0}^{x} f'(t) \, dt .
\]

If \( f'' \) is defined and Riemann integrable on \( J \), do the integral by parts \(^1\) the “wrong way” by letting

\[
F(t) = f'(t), \quad G(t) = -(x - t)
\]

to obtain

\[
f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^{x} f''(t)(x-t) \, dt.
\]

If \( f'''(t) \) exists and is Riemann integrable on \( J \), do this integral again by parts using

\[
F(t) = f''(t), \quad G(t) = -\frac{1}{2}(x-t)^2
\]

and continue in the same fashion \( n \) steps if \( f \) is \((n+1)\)-times differentiable on \( J \) and \( f^{(n+1)} \) is Riemann integrable on \( J \). The formula for \( f(x) \) thus obtained consists of the \( n \)’th Taylor polynomial \( p_n \) for \( f \) at \( x_0 \) plus an integral:

\[
f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{n!} \int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n \, dt
\]

so we obtain:

V.10.4.2. **Theorem.** [Taylor’s Theorem With Remainder, Integral Form] Let \( f \) be \((n+1)\)-times differentiable on an interval \( I \) and \( x, x_0 \in I \). Suppose \( f^{(n+1)} \) is Riemann integrable on the interval between \( x_0 \) and \( x \). If \( p_n \) is the \( n \)’th Taylor polynomial of \( f \) at \( x_0 \), and \( r_n = f - p_n \), then

\[
r_n(x) = \frac{1}{n!} \int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n \, dt
\]

V.10.4.3. Using more advanced integration, the hypotheses can be slightly relaxed to only require that \( f^{(n)} \) be absolutely continuous on the interval between \( x_0 \) and \( x \) \(^2\), or that \( f^{(n+1)} \) be defined everywhere between \( x_0 \) and \( x \) \(^2\).
V.10.5. Applications

V.10.6. Exercises

V.10.6.1. In V.10.1.3., the hypothesis that \( g^{(n)}(x_0) \) exists is necessary if \( n > 1 \):

(a) Let \( h_n(x) = \begin{cases} x^{n+1} & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases} \). Then \( \lim_{x \to 0} \frac{h_n(x)}{x^n} = 0 \), but \( h_n \) is not even continuous except at 0, so \( h_n', \ldots, h_n^{(n)} \) do not exist at 0 (or anywhere else).

(b) The hypothesis that \( g^{(n)}(x_0) \) exists is automatic if \( n = 1 \): if \( g(x_0) = 0 \) and \( \lim_{x \to 0} \frac{g(x)}{x-x_0} = 0 \), then \( g'(x_0) \) exists and equals 0.

V.10.6.2. Give an alternate direct proof of Theorem V.10.1.2. not using l'Hôpital's Rule.

V.10.6.3. (a) Mimic the proof of V.10.1.2. to prove the following generalization of V.10.1.3.:

Theorem. Let \( g \) be \( n \)-times differentiable in an open interval around \( x_0 \), with \( g^{(n)} \) Hölder continuous () at \( x_0 \) with exponent \( \alpha \) (\( 0 < \alpha < 1 \)) and constant 0, i.e.

\[
\lim_{x \to x_0} \frac{g^{(n)}(x) - g^{(n)}(x_0)}{(x-x_0)^\alpha} = 0 .
\]

Then

\[
\lim_{x \to x_0} \frac{g(x)}{(x-x_0)^{n+\alpha}} = 0
\]

if and only if \( g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0 \).

(b) Derive V.10.1.3. as a corollary of the case \( \alpha = 1 \) of (a).

V.10.6.4. Write out the details of the arguments for the Lagrange and Cauchy remainder formulas in the case \( x < x_0 \).

V.10.6.5. The proof of V.10.3.3. is simple, but requires a nonobvious guess of a function to apply Rolle's theorem to. Phrase the proof in an alternate way not requiring a guess:

(a) Using the hypotheses and the notation of the theorem, show that there is an \( M \) such that

\[
f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + M(x-x_0)^{n+1} .
\]

(b) For \( x_0 \leq t \leq x \) set

\[
\phi(t) = f(x) - f(t) - f'(t)(x-t) - \frac{f''(t)}{2}(x-t)^2 - \cdots - \frac{f^{(n)}(t)}{n!}(x-t)^n - M(x-t)^{n+1} .
\]

(c) Apply Rolle's Theorem to \( \phi \) on \([x_0, x]\) (check that the hypotheses apply) and deduce Lagrange's form of the remainder. (The \( M \) will have to be calculated in terms of \( r_n(x) \).)

Let $f$ be $(n+1)$-times differentiable on an interval $I$, $x_0 \in I$, $p_n$ the $n$'th Taylor polynomial for $f$ at $x_0$, and $x \in I$. Let $g$ be a differentiable function on $I$ such that $g'(t) \neq 0$ for all $t$ in the open interval between $x_0$ and $x$. Then there is a $c$ between $x_0$ and $x$ for which

$$r_n(x) = -\frac{g(x_0) - g(x)}{g'(c)} \frac{f^{(n+1)}(c)}{n!} (x-c)^n$$

[Set

$$\omega(t) = \frac{\psi(t)}{g(t)}$$

where $\psi$ is as in V.10.3.5., and apply the Cauchy Mean Value Theorem.]

(b) Obtain the Lagrange form of the remainder by taking $g(t) = (x-t)^{n+1}$.

Theorem. [Taylor's Theorem With Remainder, Schlömilch's Form] Let $f$ be $(n+1)$-times differentiable on an interval $I$, $x_0 \in I$, $p_n$ the $n$'th Taylor polynomial for $f$ at $x_0$, and $x \in I$. Fix $m \in \mathbb{N}$, $1 \leq m \leq n+1$. Then, for some $\theta$ between 0 and 1,

$$r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(x_0 + \theta(x-x_0))}{n!m} (1-\theta)^{n+1-m}(x-x_0)^{n+1}.$$

The case $m = 1$ is Cauchy’s form, and $m = n + 1$ is Lagrange’s form. The $\theta$ depends on $f$, $x_0$, $x$, $n$, and $m$.

(a) Prove by complete induction that $\lim_{x \to 0} \frac{f(x)}{x^n} = 0$ for all $n \in \mathbb{N} \cup \{0\}$. [Hint: Use l'Hôpital’s Rule, but not in the most straightforward way.]

(b) Prove by induction that $f^{(n)}(0)$ exists and equals 0 for all $n$. Thus $f$ is $C^\infty$ on all of $\mathbb{R}$. [You will need to use part (a) and the definition of the derivative.]

(c) Find the $n$'th Taylor polynomial $p_n$ for $f$ around 0, and show that $p_n(x)$ does not converge to $f(x)$ for any fixed $x \neq 0$.

Let $f(x) = e^x$, and let $p_n$ be the $n$'th Taylor polynomial for $f$ around 0. Since $f^{(k)}(x) = e^x$ for all $k$, $p_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$.

(a) Show that for any fixed $x \in \mathbb{R}$, $\lim_{n \to \infty} \frac{x^n}{n!} = 0$. [Consider the ratios of successive terms, and use ()]

(b) What is wrong with the following argument? Fix $x \in \mathbb{R}$. Then by Lagrange’s remainder formula there is a $c$ between 0 and $x$ such that

$$f(x) - p_n(x) = \frac{e^c x^{n+1}}{(n+1)!} x^{n+1}.$$  

Since $\lim_{n \to \infty} e^c x^{n+1} = 0$, $p_n(x) \to f(x)$ as $n \to \infty$.

(c) Modify the argument of (b) to give a valid proof that $p_n(x) \to f(x)$ for any $x \in \mathbb{R}$.
[Extended Second Derivative Test] Let $f$ be defined in a neighborhood of $x_0$, and $n \in \mathbb{N}$, $n \geq 2$. Suppose $f^{(n)}(x_0)$ exists, and that $f'(x_0) = \cdots = f^{(n-1)}(x_0) = 0$ and $f^{(n)}(x_0) \neq 0$. Using Exercise V.10.6.12 (a), show that

(a) If $n$ is even, then $f$ has a strict local extremum at $x_0$, a local minimum if $f^{(n)}(x_0) > 0$ and a local maximum if $f^{(n)}(x_0) < 0$.

(b) If $n$ is odd, then $f$ is increasing or decreasing at $x_0$ in the sense of V.3.6.7., increasing if $f^{(n)}(x_0) > 0$ and decreasing if $f^{(n)}(x_0) < 0$, and $f$ has a point of inflection at $x_0$.  

V.10.6.10. Give an alternate proof of the integral form of the remainder for Taylor’s Theorem not requiring integration by parts, by computing the iterated integral

$$
\int_{x_0}^{x} \int_{x_0}^{x} \cdots \int_{x_0}^{x} f^{(n+1)}(t) \, dt \, dx_1 \cdots dx_n
$$

where $x_1, \ldots, x_n$ are variables ranging over the interval between $x_0$ and $x$, by successive applications of the Fundamental Theorem of Calculus. Note that the first integral is not necessarily a Riemann integral unless $f^{(n+1)}$ is well-behaved between $x_0$ and $x$, but can be interpreted as a Lebesgue integral or an H-K integral in more general cases. The rest of the integrals are always Riemann integrals.

V.10.6.11. Let $f$ be $C^n$ ($0 \leq n < \infty$) on an open interval $I$, and $x_0 \in I$. Suppose there is a sequence $(x_k)$ in $I \setminus \{x_0\}$ with $x_k \to x_0$ and $f(x_k) = 0$ for all $k$. Show that

$$
f(x_0) = f'(x_0) = \cdots = f^{(n)}(x_0) = 0
$$

and hence

$$
\lim_{x \to x_0} \frac{f(x)}{(x-x_0)^n} = 0.
$$

[Apply Rolle’s Theorem to the intervals between $x_k$ and $x_{k+1}$ for each $k$, and iterate.] Show by example that

$$
\lim_{x \to x_0} \frac{f(x)}{(x-x_0)^{n+1}} = 0
$$

do not necessarily hold.

V.10.6.12. Let $f$ be defined in a neighborhood of $x_0$, and $n \in \mathbb{N}$. Suppose $f^{(n)}(x_0)$ exists (this implies that $f^{(n-1)}$ is defined in a neighborhood $U$ of $x_0$ and is continuous at $x_0$).

(a) Show that there is a unique function $g_n$ defined in $U$ and continuous at $x_0$ such that

$$
f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!} (x-x_0)^{n-1} + (x-x_0)^ng_n(x)
$$

for all $x \in U$. Show that $g_n$ is $(n-1)$-times differentiable in $U \setminus \{x_0\}$.

(b) If $n > 2$, show that $g_n$ is $(n-2)$-times differentiable at $x_0$. [Hint: show that $g_1$ is given by the formula

$$
g_1(x) = \int_0^1 f'(tx + (1-t)x_0) \, dt.
$$

(c) By considering $f(x) = |x|^{n+1}$ for $n$ even, and a similar example for $n$ odd, with $x_0 = 0$, show that $g_n$ need not be more than $(n-2)$-times differentiable at $x_0$, even if $f$ is $C^n$ on $U$ and $C^\infty$ on $U \setminus \{x_0\}$. 

V.10.6.13. [Extended Second Derivative Test] Let $f$ be defined in a neighborhood of $x_0$, and $n \in \mathbb{N}$, $n \geq 2$. Suppose $f^{(n)}(x_0)$ exists, and that $f'(x_0) = \cdots = f^{(n-1)}(x_0) = 0$ and $f^{(n)}(x_0) \neq 0$. Using Exercise V.10.6.12 (a), show that

(a) If $n$ is even, then $f$ has a strict local extremum at $x_0$, a local minimum if $f^{(n)}(x_0) > 0$ and a local maximum if $f^{(n)}(x_0) < 0$.

(b) If $n$ is odd, then $f$ is increasing or decreasing at $x_0$ in the sense of V.3.6.7., increasing if $f^{(n)}(x_0) > 0$ and decreasing if $f^{(n)}(x_0) < 0$, and $f$ has a point of inflection at $x_0$.  

631
V.11. Interpolation and Extrapolation

In this section, we discuss some topics normally treated only in numerical analysis texts but which are very helpful and relevant to understanding Taylor’s Theorem (our exposition is largely based on the nice treatments in [? and http://www.maths.lancs.ac.uk/~jameson/interpol.pdf]). No actual calculus is needed for the basic definitions and formulas, but the analysis is closely related in spirit to differential approximation and Taylor’s theorem (indeed, Taylor obtained his formula as a limiting case of the Gregory-Newton interpolation formula), and results from calculus are needed for error estimates.

The basic problem is as follows. Suppose we have a function \( f \) defined on an interval \( I \), which we presume to be suitably nicely behaved, e.g. differentiable to some high order. We know values of \( f \) at some points \( x_0, \ldots, x_n \) of \( I \) (typically forming a partition of \( I \), but not necessarily), and we want to estimate \( f(x) \) for a general \( x \in I \). This type of approximation is called \textit{interpolation}, especially if \( x \) is in the smallest interval containing \( x_0, \ldots, x_n \); if \( x \) is not in this interval, the approximation is \textit{extrapolation}. (Extrapolation is generally much more difficult and problematic to do with any accuracy.)

One promising way to proceed is to approximate \( f \) by the unique polynomial \( p \) of degree \( n \) whose values at \( x_0, \ldots, x_n \) are the same as \( f \), called the \textit{interpolating polynomial} for \( f \) with respect to \( x_0, \ldots, x_n \).

V.11.1. The Interpolating Polynomial

We first consider the problem of the existence and uniqueness of the interpolating polynomial and how to find it.

V.11.1.1. The most direct approach is a simple exercise in linear algebra. Given distinct \( x_0, \ldots, x_n \) in \( \mathbb{R} \) and real numbers \( b_0, \ldots, b_n \) (not necessarily distinct), we want a polynomial

\[
p(x) = a_0 + a_1 x + \cdots + a_n x^n
\]

with real coefficients, of degree \( \leq n \) (it could have degree less than \( n \) since \( a_n \) could be zero), such that \( p(x_k) = b_k \) for \( 0 \leq k \leq n \). Such a polynomial, if it exists, must be unique: if \( q(x) \) is another such polynomial, then \( p(x) - q(x) \) is a polynomial of degree \( \leq n \) with at least \( n + 1 \) roots (at \( x_0, \ldots, x_n \)), hence must be identically zero.

To show the polynomial exists, we must solve the system

\[
\begin{bmatrix}
1 & x_0 & x_0^2 & \cdots & x_0^n \\
1 & x_1 & x_1^2 & \cdots & x_1^n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^n
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
\vdots \\
a_n
\end{bmatrix}
= 
\begin{bmatrix}
b_0 \\
b_1 \\
\vdots \\
b_n
\end{bmatrix}
\]

for \( \{a_0, \ldots, a_n\} \). The matrix on the left is a \textit{Vandermonde matrix} which is invertible (the determinant is \( \prod_{j<k} (x_k - x_j) \); cf. Exercise V.11.8.1); thus the system has a (unique) solution which can be found by Gaussian elimination. (This works over any \textit{field}.)

V.11.1.2. This procedure for finding \( p \) works fine in principle, but is regarded as a poor method in numerical analysis, since it is computationally inefficient and the system is often ill-conditioned, especially if the \( x_k \) are close together, leading to large roundoff errors.
V.11.1.3. There is another form of the interpolating polynomial which is more computationally efficient. This form is usually attributed to Lagrange and is often called the Lagrange form of the interpolation polynomial although it was originally due to E. Waring in 1779.

Fix distinct real numbers \(x_0, \ldots, x_n\). The polynomial

\[ p_k(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \]

of degree \(n\) has the property that \(p_k(x_j) = \delta_{kj}\) for \(0 \leq j \leq n\). Thus, if \(b_0, \ldots, b_n \in \mathbb{R}\), the polynomial \(p\) of degree \(n\) with \(p(x_k) = b_k\) for all \(k\) is

\[ p(x) = \sum_{k=0}^{n} b_k p_k(x) \, . \]

V.11.1.4. There is an even better form of the interpolating polynomial, the Newton form (V.11.3.2).

V.11.2. Linear Interpolation and Extrapolation

V.11.2.1. If we only know the value of \(f\) at one point \(x_0\) of \(I\), the best estimate we can make of \(f(x)\) for \(x \in I\) is

\[ f(x) = f(x_0) \]

i.e. \(f\) is constant on \(I\) (this is generally a pretty crude estimate, but the best we can do with only one data point, and reasonably accurate for \(x\) near \(x_0\) if \(f\) is continuous). We will use the notation \(f[x_0]\) to denote the number \(f(x_0)\); thus we have the 0’th order approximation

\[ f(x) = f[x] \approx f[x_0] = f(x_0) \, . \]

Denote by \(p_0\) the constant function with value \(f(x_0)\) (we should really write \(p_{x_0}\), but we will think of \(x_0\) as being fixed and adding additional points \(x_1, x_2, \ldots\) successively).

V.11.2.2. If we know the values of \(f\) at two points \(x_0\) and \(x_1\) of \(I\), we can do a little better: we can assume \(f\) is linear, with slope

\[ f[x_0, x_1] := \frac{f(x_1) - f(x_0)}{x_1 - x_0} \]

(i.e. \(f[x_0, x_1]\) is the difference quotient (V.3.1.8), usually called a divided difference in numerical analysis.) We then have the approximation

\[ f(x) \approx f(x_0) + f[x_0, x_1](x - x_0) = f[x_0] + (x - x_0)f[x_0, x_1] \]

for \(x \in I\) and, of course, the exact formula

\[ f(x) = f[x_0] + (x - x_0)f[x_0, x] \]

\((x \neq x_0)\) which is useless for approximation.

V.11.2.3. The linear function

\[ p_{0,1}(x) = f[x_0] + (x - x_0)f[x_0, x_1] \]

is the unique linear function with the same values as \(f\) at \(x_0\) and \(x_1\).
V.11.2.4. Linear interpolation and extrapolation can be written in the equivalent form

\[ f(\lambda x_0 + (1 - \lambda)x_1) \approx \lambda f(x_0) + (1 - \lambda)f(x_1) \]

(interpolation if \(0 < \lambda < 1\), extrapolation if \(\lambda < 0\) or \(\lambda > 1\)).

V.11.3. Higher-Order Interpolation and Extrapolation

V.11.3.1. Now suppose we know the values of \(f\) at \(x_0, x_1, x_2\) (assumed distinct). We want to find the quadratic polynomial \(p_{0,1,2}\) whose values are the same as \(f\) at \(x_0, x_1, x_2\). This can be done as in section V.11.1, but we take a different approach. The polynomial will be of the form

\[ p_{0,1,2}(x) = f[x_0] + (x-x_0)f[x_0, x_1] + K(x-x_0)(x-x_1) \]

for some constant \(K\). A little algebra gives

\[ K = \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)} + \frac{f(x_1)}{(x_1-x_0)(x_1-x_2)} + \frac{f(x_2)}{(x_2-x_0)(x_2-x_1)} = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2-x_0}. \]

We thus define

\[ f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2-x_0} \]

so we have, for \(x \in I\),

\[ f(x) \approx p_{0,1,2}(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2] \]

and the exact formula

\[ f(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x] \]

(the last expression is not defined if \(x = x_0\) or \(x = x_1\), but the last term can be taken to be 0 in these cases, which makes the overall formula valid).

V.11.3.2. Continuing inductively in the same manner, if we have \(n+1\) distinct points \(x_0, \ldots, x_n \in I\), we have

\[ p_{0,1,\ldots,n}(x) = f[x_0]+(x-x_0)f[x_0, x_1]+(x-x_0)(x-x_1)f[x_0, x_1, x_2]+\cdots+(x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_n] \]

where

\[ f[x_0, \ldots, x_n] = \frac{f(x_0)}{(x_0-x_1)\cdots(x_0-x_n)} + \cdots + \frac{f(x_k)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} + \cdots + \frac{f(x_n)}{(x_n-x_0)\cdots(x_n-x_{n-1})} \]

are called higher-order divided differences. There are many other formulas for these divided differences; see below for some of them.

Thus, for \(x \in I\), we have the \(n\)’th order approximation

\[ f(x) \approx f[x_0]+(x-x_0)f[x_0, x_1]+(x-x_0)(x-x_1)f[x_0, x_1, x_2]+\cdots+(x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_n] \]

and the exact formula \((x \neq x_0, \ldots, x_{n-1})\)

\[ f(x) = f[x_0]+(x-x_0)f[x_0, x_1]+(x-x_0)(x-x_1)f[x_0, x_1, x_2]+\cdots+(x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_n, x]. \]
V.11.3.3. Divided differences and the form of the interpolating polynomial in V.11.3.2. are attributed to Newton and called the Newton form, although they were first obtained by James Gregory around 1670, at least in the case of evenly-spaced data points.

V.11.3.4. The Newton form of the interpolating polynomial is regarded as the most generally useful and efficient form for numerical analysis purposes. Not only is it computationally efficient, but it also has the nice property that the coefficients do not have to all be recomputed if an additional data point is added.

Properties of Higher-Order Divided Differences

V.11.3.5. In the divided difference \( f[x_0, \ldots, x_n] \), the numbers \( x_0, \ldots, x_n \) are typically taken in increasing order. However, this is not necessary; in fact, it is obvious from the definition V.11.3.2. that the order in which the \( x_k \) are taken is irrelevant: the numerical value is invariant under any permutation of the \( x_k \).

This can be seen in another way. The number \( f[x_0, \ldots, x_n] \) is the coefficient of \( x^n \) in the interpolating polynomial corresponding to \( \{x_0, \ldots, x_n\} \). Since this polynomial obviously depends only on the set \( \{x_0, \ldots, x_n\} \) and not on the order of the numbers, so does \( f[x_0, \ldots, x_n] \).

V.11.3.6. There is also a recursive formula

\[
f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}
\]

which justifies the name “higher-order divided difference” (Exercise V.11.8.2.). There are many variations obtained by combining this formula with permutations of the \( x_k \) (V.11.3.5.).

V.11.4. The Mean Value Theorem for Divided Differences

We now bring some calculus into the picture.

V.11.4.1. The \( n \)'th order divided difference \( f[x_0, \ldots, x_n] \) can be regarded as a function \( f[^n] \) defined on the subset \( I_d^{n+1} \) of \( I^{n+1} \) consisting of points with all coordinates distinct. The set \( I_d^{n+1} \) is a dense open set in \( I^{n+1} \). If \( f \) is continuous on \( I \), this \( n \)'th order divided difference is clearly continuous on \( I_d^{n+1} \). We will eventually extend it to all of \( I^{n+1} \) if \( f \) is sufficiently differentiable on \( I \) (V.11.5.7.).

The most important calculus result about divided differences is the following version of the MVT:

V.11.4.2. Theorem. Let \( f \) be continuous on \([a, b]\) and \( n \)-times differentiable everywhere on \((a, b)\). Let \( x_0, \ldots, x_n \) be \( n + 1 \) distinct points in \([a, b]\). Then there is at least one \( c \in (a, b) \) with

\[
f[x_0, \ldots, x_n] = \frac{f^{(n)}(c)}{n!}.
\]

Proof: The function

\[
g(x) = f(x) - p_{0,1,\ldots,n}(x)
\]

635
has at least \( n+1 \) distinct zeroes in \([a, b]\) (at \( x_0, \ldots, x_n \)). Applying Rolle’s Theorem repeatedly to the intervals between \( x_{k-1} \) and \( x_k \), \( g'(x) \) has at least \( n \) zeroes in \((a, b)\), \( g''(x) \) has at least \( n-1 \) zeroes, \( \ldots \), \( g^{(n)}(x) \) has at least one zero \( c \). But \( p_{0,\ldots,n} \) is a polynomial of degree \( \leq n \) whose degree \( n \) term is

\[
 f[x_0, \ldots, x_n]x^n
\]

so its \( n \)’th derivative is the constant function \( n!f[x_0, \ldots, x_n] \), and thus

\[
 0 = g^{(n)}(c) = f^{(n)}(c) - p_{0,\ldots,n}^{(n)}(c) = f^{(n)}(c) - n!f[x_0, \ldots, x_n].
\]

**V.11.4.3.** Note that the case \( n = 1 \) of this theorem is the usual MVT, so this theorem is a generalization of the MVT. But the MVT, or at least the special case of Rolle’s Theorem, is used in the proof.

**V.11.4.4.** We can use this result to calculate the error in using \( p_{0,\ldots,n}(x) \) to approximate \( f(x) \) for \( x \in I \). Set

\[
 E_n(x) = f(x) - p_{0,\ldots,n}(x)
\]

for \( x \in I \). If \( x \) is not one of \( x_0, \ldots, x_n \), then we have from **V.11.3.2.** that

\[
 E_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)f[x_0, \ldots, x_n, x]
\]

and combining this with **V.11.4.2.**, we obtain:

**V.11.4.5.** Corollary. Let \( f \) be \((n + 1)\)-times differentiable on \( I \), and \( x_0, \ldots, x_n \) distinct points of \( I \). Then, for any \( x \in I \), there is a \( c \) in the interior of the smallest subinterval of \( I \) containing \( x_0, \ldots, x_n, x \) such that

\[
 E_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)\frac{f^{(n+1)}(c)}{(n + 1)!}.
\]

**Proof:** If \( x \) is not one of the \( x_k \), the result follows from **V.11.4.2.**, and if \( x \) is one of the \( x_k \) both sides are zero no matter what \( c \) is chosen.

**V.11.5. Repeated Points**

**V.11.5.1.** In defining \( f[x_0, x_1] \), the \( x_0 \) and \( x_1 \) must be distinct. But if (and only if) \( f \) is differentiable at \( x_0 \), we have, for fixed \( x_0 \),

\[
 \lim_{x_1 \to x_0} f[x_0, x_1] = \lim_{x_1 \to x_0} \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(x_0)
\]

so we may define

\[
 f[x_0, x_0] = f'(x_0)
\]

if \( f \) is differentiable at \( x_0 \). This definition is especially suitable if \( f \) is strictly differentiable at \( x_0 \) (\( f' \) is continuous at \((x_0, x_0)\)). In particular, if \( f \) is \( C^1 \) on \( I \), then \( f[\cdot, \cdot] \) is continuous on \( I^2 \).
V.11.5.2. We want to extend this to higher-order divided differences. We first consider the case where all the coordinates are the same. Fix $n$, and for $x_0 \in I$ write $x_0 = (x_0, \ldots, x_0) \in I^{n+1}$. If $x = (y_0, \ldots, y_n) \in I^{n+1}$, write $f[x] = f[y_0, \ldots, y_n]$.

V.11.5.3. Proposition. Let $f$ be $n$-times differentiable on $I$, and $x_0 \in I$. If $f^{(n)}$ is continuous at $x_0$, then

$$\lim_{x \to x_0} f[x] = \frac{f^{(n)}(x_0)}{n!}.$$

Proof: Let $\epsilon > 0$. Then there is a $\delta > 0$ such that, for all $y \in I$, $|y - x_0| < \delta$, we have

$$\left| \frac{f^{(n)}(y)}{n!} - \frac{f^{(n)}(x_0)}{n!} \right| < \epsilon.$$

Let $J = I \cap (x_0 - \delta, x_0 + \delta)$. If $x = (y_0, \ldots, y_n) \in J^{n+1}$, then by V.11.4.2. there is a $c_x \in J$ with

$$f[x] = \frac{f^{(n)}(c_x)}{n!}$$

so

$$\left| f[x] - \frac{f^{(n)}(x_0)}{n!} \right| < \epsilon.$$

V.11.5.4. If $f$ is $C^n$ on $I$, we can thus extend the $n$th order divided difference to points with all coordinates the same by setting

$$f[x_0, \ldots, x_0] = \frac{f^{(n)}(x_0)}{n!}$$

for $x_0 \in I$. The extended function is continuous on the subset of $I^{n+1}$ consisting of points whose coordinates are either all the same or all different.

V.11.5.5. We can extend higher-order divided differences to points where some, but not all, coordinates coincide. Rather than trying to state things in maximum generality (cf. V.11.5.13.), we will just assume $f$ is $C^n$ on $I$. The most efficient approach is to prove an integral formula for the higher-order divided differences which is of general interest:

V.11.5.6. Theorem. Let $f$ be $C^n$ on an interval $I$, and $x_0, \ldots, x_n$ distinct points of $I$. Then

$$f[x_0, \ldots, x_n] = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1})) \, dt_n \cdots dt_2 dt_1.$$

637
Proof: By induction on $n$. The case $n = 1$ is essentially the Fundamental Theorem of Calculus in slight disguise: to evaluate the integral
$$\int_0^1 f'(x_0 + t(x_1 - x_0)) \, dt$$
for $f \in C^1$, make the substitution
$$u = x_0 + t(x_1 - x_0)$$
$$du = (x_1 - x_0) \, dt$$
to obtain
$$\int_0^1 f'(x_0 + t(x_1 - x_0)) \, dt = \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} f'(u) \, du = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$ 
Thus $f[x_0, x_1]$ is the average value of $f'$ over the interval $[x_0, x_1]$ (or $[x_1, x_0]$ if $x_1 < x_0$).

Now suppose the formula holds for $n - 1$. Compute the first integral
$$\int_0^{t_{n-1}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1})) \, dt_n$$
making the substitution
$$u = x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1})$$
$$du = (x_n - x_{n-1}) \, dt_n$$
so the first integral evaluates to
$$\frac{1}{x_n - x_{n-1}} [f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1}))]_0^{t_{n-1}}$$
$$= \frac{1}{x_n - x_{n-1}} [f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}) + t_n(x_n - x_{n-1})) - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad = \frac{1}{x_n - x_{n-1}} [f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2})) - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad = \frac{1}{x_n - x_{n-1}} [f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2})) - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad - f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]
\quad = \frac{1}{x_n - x_{n-1}} [f[x_0, x_1] - f[x_0, x_{n-2}, x_n] - f[x_0, x_{n-2}, x_{n-1}]) = f[x_0, \ldots, x_n]
\quad = \frac{1}{x_n - x_{n-1}} (f[x_0, \ldots, x_{n-2}, x_n] - f[x_0, \ldots, x_{n-2}, x_{n-1}]) = f[x_0, \ldots, x_n]
\quad = \frac{1}{x_n - x_{n-1}} (f[x_0, \ldots, x_{n-2}, x_n] - f[x_0, \ldots, x_{n-2}, x_{n-1}]) = f[x_0, \ldots, x_n]"
V.11.5.7. The formula in V.11.5.6. makes perfect sense even if \( x_0, \ldots, x_n \) are not distinct, and clearly gives a continuous function of \((x_0, \ldots, x_n)\) on \( I^{n+1} \). We denote this extended function also by \( f[x_0, \ldots, x_n] \). The extended function agrees with the extension given in V.11.5.4. in the case all the coordinates are equal. Thus we obtain

\[ f[x_0, \ldots, x_n] \]

V.11.5.8. Corollary. If \( f \) is \( C^n \) on an interval \( I \), then \( f[\cdot, \ldots, \cdot] \) \((n+1\) arguments), extended as above to points with repeated coordinates, is continuous on \( I^{n+1} \).

V.11.5.9. If \( f \) is \( C^n \) on \([a, b]\), then the conclusion of Theorem V.11.4.2. holds also (with \( c \in [a, b] \)) if not all the \( x_k \) are distinct, as can be seen by a straightforward limiting argument using V.11.5.8. and the Bolzano-Weierstrass Theorem \((\cdot)\). In particular, if \( f \) is \( C^n \), then for any \( x_0 \) and any \([a, b]\) containing \( x_0 \) there is a \( c \in [a, b] \) for which \( f[x_0, \ldots, x_0] = \frac{f^{(n)}(c)}{m!} \). Taking the limit as \( a, b \to x_0 \), we obtain an alternate proof of V.11.5.3.\

To extend \( f[x_0, \ldots, x_n] \) to points with some repetitions in somewhat greater generality, we use the following result, which is proved inductively using V.11.5.8. and the recursion formula for divided differences with distinct entries:

V.11.5.10. Theorem. Let \( f \) be a function on an interval \( I \), \( J_0, \ldots, J_r \) disjoint subintervals of \( I \), and \( m_0, \ldots, m_r \in \mathbb{N} \). Suppose \( f \) is \( C^{m_k-1} \) on \( J_k \) for each \( k \). Then

\[ f[x_0(0), \ldots, x_R(0, m_0-1), x_1(1,0), \ldots, x_1(1,m_1-1), \ldots, x_R(r, m_r-1)] \]

extends to a continuous function on \( J_0^{m_0} \times \cdots \times J_r^{m_r} \).

V.11.5.11. We can obtain an explicit formula for \( f[x_0, \ldots, x_n] \) when the points are not all distinct. Since the function values are independent of the order of the coordinates on a dense set, by continuity this holds everywhere. Thus to evaluate the function we may put the entries in increasing order, with the repeated coordinates grouped together. We write \( x_0, \ldots, x_r \) for the distinct values, and suppose \( x_k \) occurs \( m_k \) times, so \( m_0 + \cdots + m_r = n + 1 \).

V.11.5.12. Theorem. Suppose \( f \) is \( C^n \) on \( I \). Then \( f[x_0, \ldots, x_r] \) is \( C^{n-r} \) on \( I_d^{r+1} \), and

\[ f[x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r] = \frac{1}{(m_0 - 1)!(m_1 - 1)! \cdots (m_r - 1)!} \frac{\partial^{n-r}}{\partial x_0^{m_0-1} \partial x_1^{m_1-1} \cdots \partial x_r^{m_r-1}} (f[x_0, \ldots, x_r]) \]

where \( x_k \) occurs \( m_k \) times on the left and once on the right.

V.11.5.13. Actually, for this formula to hold and define \( f[x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r] \) continuously at \( (x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r) \), it suffices that \( f \) be \( C^{m_k-1} \) in an interval around \( x_k \) for each \( k \) (although the \( C^{n-r} \) statement does not hold in general). See [\( ? \), p. 72-73] for an outline of the proof. Note that the formula agrees with the previous cases where \( r = n \) (all \( m_k = 1 \)) or \( r = 0 \) (\( m_0 = n + 1 \)).

639
V.11.6. Osculating Polynomials

V.11.6.1. We want to extend the notion of interpolating polynomial to the case where the points \( x_0, \ldots, x_n \) are not all distinct. The first thing we must decide is what this should mean: if \( x_0 \) is repeated \( m \) times, how should the approximating polynomial \( p \) agree with \( f \) at \( x_0 \)?

The approximation should be of “order \( m \)” in the sense that \( f - p \) should have a zero of “order \( m \)” at \( x_0 \). The broadest sense in which this can be interpreted is to demand that

\[
f(x) - p(x) = o(|x - x_0|^{m-1})
\]
as \( x \to x_0 \) in \( I \). By V.10.1.3., if \( f \) is \( (m - 1) \)-times differentiable at \( x_0 \), this will happen if and only if

\[
p(x_0) = f(x_0), \quad p'(x_0) = f'(x_0), \ldots, p^{(m-1)}(x_0) = f^{(m-1)}(x_0).
\]

So, if \( f \) is differentiable to high enough order, we need to find a polynomial with specified values and derivatives up to a certain level at each of the points. This is possible because of the following extension of V.11.1.1.:

V.11.6.2. Theorem. Let \( n \in \mathbb{N}, x_0, \ldots, x_r \) distinct real numbers, \( m_0, \ldots, m_r \in \mathbb{N} \) with \( m_0 + \cdots + m_r = n + 1 \) (so necessarily \( r \leq n \)), and

\[
b_0^{(0)}, b_0^{(1)}, \ldots, b_0^{(m_0-1)}, b_1^{(0)}, \ldots, b_r^{(m_r-1)}
\]
real numbers (not necessarily distinct). Then there is a unique polynomial \( p \) with real coefficients, of degree \( \leq n \), for which \( p^{(j)}(x_k) = b_k^{(j)} \) for \( 0 \leq k \leq r, 0 \leq j \leq m_k - 1 \).

Proof: To find

\[
p(x) = a_0 + a_1 x + \cdots + a_n x^n
\]
we need to solve the system

\[
\begin{bmatrix}
1 & x_0 & x_0^2 & \cdots & x_0^n \\
0 & 1 & 2x_0 & \cdots & nx_0^{n-1} \\
0 & 0 & 2 & \cdots & n(n-1)x_0^{n-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_1 & x_1^2 & \cdots & x_1^n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & n(n-1) \cdots (n-m_r+2)x_r^{n-m_r+1}
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
\vdots \\
a_n
\end{bmatrix}
= \begin{bmatrix}
b_0^{(0)} \\
b_0^{(1)} \\
\vdots \\
b_r^{(m_r-1)}
\end{bmatrix}
\]

where there are \( m_k \) rows for each \( x_k \). It can be shown (Exercise V.11.8.1.) that the matrix on the left is invertible if the \( x_k \) are distinct, so the system has a unique solution. 

V.11.6.3. Corollary. Let \( f \) be a function on an interval \( I \). Let \( n \in \mathbb{N}, x_0, \ldots, x_r \) distinct numbers in \( I \), and \( m_0, \ldots, m_r \in \mathbb{N} \) with \( m_0 + \cdots + m_r = n + 1 \) (so necessarily \( r \leq n \)). Suppose \( f \) is \( (m_k - 1) \)-times differentiable at \( x_k \) for each \( k \). Then there is a unique polynomial \( p \) of degree \( \leq n \), with real coefficients, such that \( p^{(j)}(x_k) = f^{(j)}(x_k) \) for \( 0 \leq k \leq r, 0 \leq j \leq m_k - 1 \).
V.11.6.4. Definition. The polynomial $p$ is called the osculating polynomial of $f$ with respect to $(x_0, \ldots, x_r; m_0, \ldots, m_r)$.

V.11.6.5. The osculating polynomial generalizes both the interpolating polynomial (the case $r = n$ and all $m_k = 1$) and the Taylor polynomial (the case $r = 0$, $m_0 = n + 1$). If $f$ is sufficiently differentiable, the osculating polynomial is a limiting case of interpolation polynomials for distinct points approaching the repeated points (this is essentially how TAYLOR found the Taylor polynomials).

V.11.6.6. In fact, there is a formula for the osculating polynomial using the extended definitions of higher-order divided differences which is just an extension of the Newton form of the interpolation polynomial:

$$p(x) = f[x_0] + (x - x_0)f[x_0, x_0] + \cdots + (x - x_0)^{m_0 - 1}f[x_0, \ldots, x_0] + (x - x_0)^{m_0}f[x_0, \ldots, x_0, x_1]$$

$$+ (x - x_0)^{m_0}(x - x_1)f[x_0, \ldots, x_0, x_1, x_1] + \cdots + (x - x_0)^{m_0}(x - x_1)^m(x - x_r)^{m - 1}f[x_0, \ldots, x_0, x_1, \ldots, x_r]$$

where $x_{k+1}$ starts appearing when $x_k$ has been repeated $m_k$ times. There is also an exact formula

$$f(x) = f[x_0] + (x - x_0)f[x_0, x_0] + \cdots + (x - x_0)^{m_0 - 1}f[x_0, \ldots, x_0] + (x - x_0)^{m_0}f[x_0, \ldots, x_0, x_1]$$

$$+ (x - x_0)^{m_0}(x - x_1)f[x_0, \ldots, x_0, x_1, x_1] + \cdots + (x - x_0)^{m_0}(x - x_1)^m(x - x_r)^{m - 1}f[x_0, \ldots, x_0, x_1, \ldots, x_r, x]$$

where $x_k$ is repeated $m_k$ times, which is valid for $x \neq x_0, \ldots, x_r$, and even also for $x = x_k$ if $f$ is $m_k$-times differentiable at $x_k$.

We thus get a formula for the error from approximation by the osculating polynomial: if $E(x) = f(x) - p(x)$, we have

$$E(x) = (x - x_0)^{m_0}(x - x_1)^m \cdots (x - x_r)^{m_r}f[x_0, \ldots, x_0, x_1, \ldots, x_r, x]$$

and, combining this with V.11.5.9., we obtain:

V.11.6.7. Corollary. Let $f$ be $(n+1)$-times differentiable on $I$, $x_0, \ldots, x_r$ distinct points of $I$, and $m_0, \ldots, m_r \in \mathbb{N}$ with $m_0 + \cdots + m_r = n + 1$ (so necessarily $r \leq n$). Let $p$ be the osculating polynomial of $f$ with respect to $(x_0, \ldots, x_r; m_0, \ldots, m_r)$. Then, for any $x \in I$, there is a $c$ in the smallest subinterval of $I$ containing $x_0, \ldots, x_r, x$ such that

$$f(x) - p(x) = (x - x_0)^{m_0}(x - x_1)^m \cdots (x - x_r)^{m_r}f^{(n+1)}(c) \frac{f^{(n+1)}(c)}{(n + 1)!}.$$  

Note that both V.11.4.5. and the Lagrange form of the error in Taylor’s Theorem (V.10.3.3.) are special cases.

V.11.6.8. An osculating polynomial where all $m_k$ are 2 ($n$ is then necessarily odd) is called a Hermite polynomial, and approximation by such a polynomial is called Hermite interpolation. (Approximation with general $m_k \geq 2$ is also sometimes called Hermite interpolation.)
V.11.6.9. In practical approximation problems, it can happen, although it is somewhat unusual, that one has data allowing approximation of an unknown function by an osculating polynomial, especially a Hermite polynomial. Such an approximation, when possible, will generally (but not always) be more accurate than approximation by the interpolation polynomial for the same data points.

V.11.7. How Good are the Approximations?

It is beyond the scope of this book to give a comprehensive discussion of the accuracy (not to mention the efficiency) of the various ways to approximate functions by polynomials. But we will briefly survey some of the aspects of accuracy in this section. Almost any book on numerical analysis contains a much more thorough treatment. The article [http://people.maths.ox.ac.uk/trefethen/mythspaper.pdf](http://people.maths.ox.ac.uk/trefethen/mythspaper.pdf) has some important observations about polynomial approximation. See also [?].

Approximation With Evenly-Spaced Nodes

V.11.7.1. If $I = [a,b]$, the most obvious choice of $x_0, \ldots, x_n$ is as a partition of $I$ into $n$ subintervals of equal length $h = \frac{b-a}{n}$, i.e. $x_k = x_0 + kh$. In this case, the denominators in the recursive formula (V.11.3.6.) for the divided differences are all equal to $h$, so we may instead consider just the ordinary differences

$$\Delta f(x_j) = f(x_{j+1}) - f(x_j)$$

$$\Delta^{k+1} f(x_j) = \Delta^k f(x_{j+1}) - \Delta^k f(x_j)$$

(these are usually called forward differences), giving the formula (easily proved by induction)

$$f[x_0, \ldots, x_k] = \frac{1}{k! h^k} \Delta^k f(x_0).$$

We also use the notation

$$\pi_k(t) = t(t-1) \cdots (t-k+1)$$

for $k \geq 1$ (we may set $\pi_0(t) = 1$; $\pi_k$ is called the $k$’th factorial polynomial). The Newton formula for the interpolation polynomial then becomes, for $0 \leq t \leq n$,

$$p_{0,\ldots,n}(x_0 + th) = f(x_0) + \sum_{k=1}^n \frac{\pi_k(t)}{k!} \Delta^k f(x_0) = \sum_{k=0}^n \frac{\pi_k(t)}{k!} \Delta^k f(x_0)$$

if we set $\Delta^0 f(x_0) = f(x_0)$. This is the Gregory-Newton interpolation formula. We may further rewrite this using the generalized binomial coefficient

$$\binom{t}{k} = \frac{\pi_k(t)}{k!}$$

(cf. V.17.7.1.), so

$$p_{0,\ldots,n}(x_0 + th) = \sum_{k=0}^n \binom{t}{k} \Delta^k f(x_0).$$

We may get similar formulas for backward differences or centered differences; see books on numerical analysis for details.
If \( f \) is continuous on \([a, b]\) and \((n + 1)\)-times differentiable on \((a, b)\), we get a corresponding formula for the error term for \(0 \leq t \leq n\):

\[
E_n(x_0 + th) = \pi_n(t) h^{n+1} \frac{f^{(n+1)}(c)}{(n+1)!}
\]

for some \(c \in (a, b)\).

We also get a formula for the Lagrange form of the interpolation formula using evenly-spaced nodes. Using the same notation as above, we have

\[
p_{0, \ldots, n}(x_0 + th) = \frac{\pi_n(t)}{n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \frac{f(x_k)}{t - k}.
\]

The Runge Phenomenon

Suppose \( f \) is a function on \([a, b]\). One might expect that as \( n \to \infty \), the interpolation polynomials \( p_{0, \ldots, n} \) for \( f \) corresponding to a subdivision of \([a, b]\) into \( n \) subintervals of equal length should converge to \( f \) pointwise or even uniformly on \([a, b]\) if \( f \) is sufficiently smooth. But this fails dramatically in general, even if \( f \) is analytic. This phenomenon was found by C. Runge in 1901, although related work had previously been done by C. Meray.

Runge’s classic example is the innocent-looking function

\[
g(x) = \frac{1}{x^2 + 1}
\]

on the interval \([-5, 5]\), or equivalently by rescaling

\[
f(x) = \frac{1}{25x^2 + 1}
\]

on \([-1, 1]\). These functions are real-analytic on all of \(\mathbb{R}\), and complex-analytic on \(\mathbb{C}\) except for two poles. Figure V.10 (from http://www.math.utk.edu/~ccollins/M471/runge.jpg) shows some approximants of small order, which suggest there is an approximation problem near the ends of the interval.

In fact, the approximants for \( g \) converge pointwise for

\[
|x| < x_c \approx 3.63
\]

and u.c. on \((-x_c, x_c)\), but diverge pointwise for \(x_c < |x| < 5\). The demonstration http://demonstrations.wolfram.com/RungePhenomenon shows just how wildly the approximations diverge near the endpoints.

It should not be too surprising that there are problems near the endpoints, since interpolation near the endpoints is almost extrapolation, for which one would expect problems. In fact, the polynomial \( \pi_n \) has large values near the ends of the interval \([0, n]\), so the error in V.11.7.2. might tend to blow up (and often does) toward the ends of the interval as \( n \) increases.
V.11.7.8. The real “explanation” of why approximation fails in this example is that the poles of the function in \( \mathbb{C} \) are too close to the interval of approximation (cf. Exercise V.11.8.10.). A detailed discussion can be found in [?], (), ()..


Chebyshev Nodes

V.11.7.10. It would appear that one gets a better approximation using nodes which are not evenly spaced, but are more concentrated near the endpoints. This is true in reasonable cases. We will work on the interval \([-1,1]\) for simplicity; other intervals can be handled via an affine transformation.

V.11.7.11. If \( f \) is smooth, the error in approximating by \( p_{0,\ldots,n} \) for \( \{x_0,\ldots,x_n\} \) is given by

\[
(x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(c)}{(n+1)!}
\]

for some \( c \in (-1,1) \). To try to minimize this, we can choose the \( x_k \) so that the polynomial

\[
q(x) = (x - x_0)(x - x_1) \cdots (x - x_n)
\]

is uniformly as small as possible on \([-1,1]\). CHEBYSHEV showed that the minimum occurs when

\[
x_k = \cos \left( \frac{2k + 1}{2(n+1)} \pi \right).
\]
These are called the Chebyshev nodes (note that they do not form a partition of \([-1, 1]\); the endpoints are not nodes). If the \(x_k\) are the Chebyshev nodes, the maximum value of \(q(x)\) is \(2^{-n}\), so the error in the approximation can be bounded:

\[
|f(x) - p_0,\ldots,n(x)| \leq \frac{1}{2^n(n+1)!} \sup_{-1 < c < 1} |f^{(n+1)}(c)|
\]

for \(x \in [-1, 1]\).

**V.11.7.12.** In fact, the Chebyshev approximations converge uniformly to \(f\) if \(f\) is Hölder continuous, and in particular if \(f\) is \(C^1\) or even Lipschitz (the minimal requirement to insure convergence is somewhat weaker than Hölder continuity; cf. ())). They converge rapidly if \(f\) is differentiable to high order. There are, however, examples of continuous \(f\) where the Chebyshev approximations do not converge even pointwise. In fact, for any scheme of choosing interpolation points an example of a continuous \(f\) can be concocted for which the scheme does not converge (Faber’s theorem).

**Extrapolation**

**V.11.7.13.** Because the polynomial

\[
q(x) = (x - x_0)(x - x_1) \cdots (x - x_n)
\]

grows rapidly outside the smallest interval containing \(x_0, \ldots, x_n\), the error estimate from V.11.4.5. or V.11.6.6. quickly blows up outside this interval. In fact, extrapolation using an interpolating or osculating polynomial is not generally too accurate outside the interval of nodes.

**V.11.7.14.** Using a Taylor polynomial for \(f\) at \(x_0\) to approximate \(f(x)\) for \(x \neq x_0\) is really always extrapolation, so it is not surprising that the Taylor polynomials do not always give good approximations to \(f(x)\) for a fixed \(x \neq x_0\). It is actually more surprising (from the interpolation-extrapolation point of view) that they really do give good approximations in so many important cases.

**Other Types of Approximation**

**V.11.7.15.** The Chebyshev polynomials are not usually the best uniform approximants to a smooth function \(f\). It can be shown that for any \(f\) which is continuous on \([a, b]\), there is a unique polynomial \(p\) of degree \(\leq n\) which approximates \(f\) uniformly most closely on \([-1, 1]\) (minimax approximation), i.e. for which \(\|f - p\|_{\infty}\) is minimum [existence of a minimizing polynomial is a simple compactness argument, but uniqueness is not so obvious since \(\| \cdot \|_{\infty}\) is not strictly convex], and \(p\) is an interpolation polynomial for \(f\) for some choice of \(x_0, \ldots, x_n\) (Chebyshev’s equioscillation theorem, cf. Exercise V.11.8.9.). But there is no reasonable procedure for finding the best \(x_0, \ldots, x_n\) in general, even if \(f\) is smooth, although there is an iterative procedure (the Remez algorithm) which converges rapidly in good cases. In reasonable cases, the Chebyshev polynomials come close to realizing the minimum. See [May06] for a nice discussion of Chebyshev’s equioscillation theorem with illustrative applets.

**V.11.7.16.** Another way of uniformly approximating a continuous \(f\) is by Bernstein polynomials (XV.8.1.2.). These are not interpolation polynomials and converge uniformly to \(f\) relatively slowly even if \(f\) is smooth, but they are guaranteed to converge uniformly to \(f\) for any continuous \(f\).
V.11.7.17. One can also try for approximation of \( f \) in other norms besides the uniform norm. One alternative which can be efficiently computed is using the 2-norm (least-square approximation). See (x).

V.11.7.18. One can also use other functions besides polynomials to approximate, e.g. rational functions or trigonometric functions. There is an extensive theory of such approximations, discussed in numerical analysis texts.

V.11.8. Exercises

V.11.8.1. (a) Verify the formula for the determinant of the Vandermonde matrix (V.11.1.1). [Regard \( x_0, \ldots, x_n \) successively as variables. The determinant is a polynomial in these variables which vanishes when \( x_k = x_j \) for \( j \neq k \), so \( x_k - x_j \) is a factor.]

(b) Similarly compute the determinant of the matrix in the proof of V.11.6.2. and show it is nonzero. [Experiments with small-size examples using a computer algebra system can be helpful in guessing the correct formula.]

(c) Give an alternate proof that the system of V.11.6.2. (and the special case V.11.1.1.) always has a unique solution \([\text{a solution without computing determinants, as follows. First show that the homogeneous system (all } b_k^{(j)} = 0) \text{ has a unique solution [a solution } q \text{ is a polynomial of degree } \leq n \text{ with at least } n + 1 \text{ roots, counted with multiplicity]. This implies that the matrix is invertible, so the system is uniquely solvable for any } b_k^{(j)} \text{'s.}}\]

V.11.8.2. (a) Verify the formula for \( K \) in V.11.3.1. and the formula in V.11.3.2. [Use that \( f(x_0) + (x - x_0)f(x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1})f(x_0, \ldots, x_n) \)

(b) Verify the recursive formula for \( f[x_0, \ldots, x_n] \) in V.11.3.6. [Show that \( f[x_0, \ldots, x_n, x] = \sum_{k=0}^{n} f[x_k, x] \prod_{j \neq k} (x_k - x_j) \).]

V.11.8.3. Suppose \( f \) is \( C^r \) on an interval \( I \) \((r \geq 1)\), and \( x_0 \in I \). Set 

\[
g(x) = \begin{cases} 
 f[x_0, x] & \text{if } x \neq x_0 \\
 f'(x_0) & \text{if } x = x_0 
\end{cases}
\]

Show that \( g \) is \( C^{r-1} \) on \( I \).

V.11.8.4. (a) Suppose \( f \) is a function on an interval \( I \), and \( x_0, \ldots, x_n \) are distinct points of \( I \). Show that, for \( x \in I \), \( x \) not equal to any \( x_k \),

\[
f[x_0, \ldots, x_n, x] = \sum_{k=0}^{n} f[x_k, x] \prod_{j \neq k} (x_k - x_j) .
\]

(b) Suppose \( f \) is \( C^r \) on \( I \) \((r \geq 1)\). Set 

\[
g(x) = f[x_0, \ldots, x_n, x] = \sum_{k=0}^{n} f[x_k, x] \prod_{j \neq k} (x_k - x_j) 
\]

where \( f[x_k, x] \) is extended to \( x = x_k \) as in V.11.8.3.. Show that \( g \) is \( C^{r-1} \) on \( I \).
V.11.8.5. Write out the details of the proof of V.11.5.10. and V.11.5.13..

V.11.8.6. Show that the formula in V.11.6.6. gives the osculating polynomial.

V.11.8.7. Derive the formula for $p_0, \ldots, p_n$ in V.11.7.1. from V.11.3.2., and the Lagrange form in V.11.7.3. from V.11.1.3..

V.11.8.8. For $n \geq 0$, define

$$T_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$$

where $x_0, \ldots, x_n$ are the Chebyshev nodes (V.11.7.11.). $T_{n+1}$ is called the $(n+1)$’st monic Chebyshev polynomial of the first kind. (References differ on whether to multiply by a conventional factor of $2^n$, which is why we use the term “monic Chebyshev polynomial.” We will not discuss Chebyshev polynomials of other kinds.)

(a) Show that

$$T_n(x) = 2^{-n+1} \cos(n \cos^{-1} x)$$

for $n \geq 1$ and $-1 \leq x \leq 1$. (This formula can also be used to define $T_n$ for $n = 0$.)

(b) Show that $\max_{|x| \leq 1} |T_n(x)| = 2^{-n+1}$, and that the maximum occurs precisely at the $n+1$ points $\cos \frac{k}{n}$, $0 \leq k \leq n$.

(c) Show that $T_n$ ($n \geq 1$) has the smallest uniform norm of all monic polynomials of degree $\leq n$, with real coefficients, with $n$ roots in $[-1,1]$ counted with multiplicity.

(d) Prove the recurrence formula $T_{n+1}(x) = xT_n(x) - \frac{1}{2}T_{n-1}(x)$ for $n \geq 1$.

(e) Show that $T_n$ is a solution to the differential equation

$$(1-x^2)y'' = xy' - n^2 y.$$ 

The Chebyshev polynomials also arise in least-squares approximation.

V.11.8.9. Let $f$ be a continuous function on a closed bounded interval $[a,b]$, and $n \in \mathbb{N}$.

(a) Prove using the Heine-Borel theorem that there is a polynomial $p$ of degree $\leq n$ for which $\|f-p\|_\infty$ is minimum among all polynomials of degree $\leq n$.

(b) If $p$ is a polynomial of degree $\leq n$ and there are $n+2$ distinct points $y_0, \ldots, y_{n+1}$ in $[a,b]$ (in increasing order) such that

$$f(y_k) - p(y_k) = \pm \|f-p\|_\infty$$

with alternating signs, then $\|f-p\|_\infty$ is minimum among all polynomials of degree $\leq n$. [If $q$ is a polynomial of degree $\leq n$ with $\|f-q\|_\infty < \|f-p\|_\infty$, then $q(y_k)$ is alternately greater than and less than $p(y_k)$. Thus $p-q$ is a nonzero polynomial of degree $\leq n$ with at least $n+1$ roots by the Intermediate Value Theorem.]
(c) Chebyshev’s Equioscillation Theorem states the converse to (b): a polynomial $p$ for which $\|f - p\|_\infty$ is minimum among all polynomials of degree $\leq n$ has the equioscillation property of (b).

(d) Show that the result of (c) implies that the polynomial of degree $\leq n$ of minimum uniform distance from $f$ is unique. [If $p$ and $q$ minimize distance from $f$, then so does $\phi = \frac{1}{2}(p + q)$, and thus $\phi$ has the equioscillation property.]

(e) Show by the Intermediate Value Theorem that the closest approximant $p$ is an interpolation polynomial for $f$ with nodes $x_0, \ldots, x_n$ with $y_k < x_k < y_{k+1}$ for all $k$.

V.11.8.10. Let $f(x) = \frac{1}{x^2 + 1}$. Sketch the graphs of the Taylor polynomials of $f$ around 0. On what interval do they converge to $f$? What is the radius of convergence of the Taylor series of $f$ around 0? What about the Taylor series around a general $x_0 \in \mathbb{R}$? Justify this geometrically by considering $f$ as an analytic function with poles on $\mathbb{C}$. This suggests some connection between complex poles and the Runge phenomenon; one would expect interpolating polynomials to approximate somewhat better over an interval than Taylor polynomials, but perhaps not to an unlimited extent.
V.12. How Discontinuous Can a Derivative Be?

If \( f \) is a differentiable function (on an interval, or all of \( \mathbb{R} \)), what can be said about the properties of the function \( f' \)?

V.12.1. Continuity and Measurability of Derivatives

V.12.1.1. Example. If \( f \) is differentiable, the function \( f' \) need not itself be differentiable everywhere. For perhaps the simplest example, let
\[
f(x) = \begin{cases} 
  x^2 & \text{if } x \geq 0 \\
  -x^2 & \text{if } x < 0 
\end{cases}.
\]

Then we have
\[
f'(x) = \begin{cases} 
  2x & \text{if } x \geq 0 \\
  -2x & \text{if } x < 0 
\end{cases}.
\]

(It must be shown directly from the definition of derivative that \( f'(0) \) exists and equals 0.) Thus \( f \) is differentiable everywhere, but \( f' \) is not differentiable at 0.

V.12.1.2. In this example, \( f' \) is continuous. In fact, every continuous function occurs as the derivative of a differentiable function: if \( g \) is continuous on an interval \( I \), and \( x_0 \in I \), for \( x \in I \) define
\[
f(x) = \int_{x_0}^{x} g(t) \, dt.
\]

Then, by the Fundamental Theorem of Calculus, \( f'(x) = g(x) \).

However, a derivative need not even be continuous:

V.12.1.3. Example. Let \( f(x) = x^2 \sin \frac{1}{x} \) for \( x \neq 0 \), and \( f(0) = 0 \). Then, for \( x \neq 0 \),
\[
f'(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}.
\]

It can be checked from the definition (Exercise) that \( f'(0) \) exists and equals 0. Thus \( f \) is differentiable everywhere, but \( f' \) is not continuous at 0 since \( \lim_{x \to 0} f'(x) \) does not exist.

In this example, although \( f' \) is not continuous, it is bounded on bounded intervals, and measurable, hence Lebesgue integrable on any bounded interval, and \( f \) can be recovered from \( f' \) by
\[
f(x) = \int_0^x f'(t) \, dt
\]
where the integral is a Lebesgue integral. (In fact, since \( f' \) is bounded and continuous except at one point, it is Riemann integrable on any bounded interval, so the integral can even be taken to be a Riemann integral in this case.)

But \( f' \) need not be even locally Lebesgue integrable:
**V.12.1.4.** Example. Let $f(x) = x^2 \sin \frac{1}{x}$ for $x \neq 0$, and $f(0) = 0$. Then, for $x \neq 0$, 

$$f'(x) = 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}.$$ 

It can again be checked from the definition (Exercise ()) that $f'(0)$ exists and equals 0, so $f$ is differentiable, but $f'$ is unbounded and not Lebesgue integrable in any neighborhood of 0.

**V.12.1.5.** However, if $f$ is any differentiable function, then $f'$ is always locally H-K integrable to $f()$. In fact, the function $f'$ of V.12.1.4. is one of the standard examples of a function which is H-K integrable but not Lebesgue integrable.

**V.12.1.6.** The above examples of derivatives are continuous except at one point. A differentiable function can be easily constructed whose derivative has a dense set of discontinuities (Exercise V.12.3.2.). Much more extreme examples can be constructed; see Exercise V.12.3.3. for an example of a differentiable function whose derivative is discontinuous almost everywhere. One of the most extreme examples is Example () of Katznelson and Stromberg ([KS74]; cf. [vRS82, 13.2]):

**V.12.1.7.** Example. There is a differentiable function $f$ with the following properties:

(i) $|f'(x)| \leq 1$ for all $x$.

(ii) Each of the following sets is dense in $\mathbb{R}$:

- $A = \{x : f'(x) > 0\}$
- $B = \{x : f'(x) = 0\}$
- $C = \{x : f'(x) < 0\}$

(iii) $f$ is not monotone on any interval.

In fact, “many” differentiable functions satisfy (ii) and (iii) (XI.12.5.7., XI.12.5.8.). It follows from (ii) that the discontinuities of $f'$ are dense; in fact, $f'$ must be discontinuous wherever it is nonzero.

By (), $f'$ is not Riemann integrable on any interval, hence $f'$ is not continuous a.e. on any interval (). In fact, the sets $A$ and $C$ both have positive Lebesgue measure.

But the next result, combined with (), shows that the derivative of a differentiable function cannot be “too discontinuous.”

**V.12.1.8.** Proposition. If $f$ is differentiable, then $f'$ is of Baire Class 1 (), and in particular must be Lebesgue measurable, and continuous except on a meager set (which may, however, have positive Lebesgue measure (V.12.1.7., V.12.3.3.)).

**Proof:** Define

$$g_n(x) = \frac{f\left(x + \frac{1}{n}\right) - f(x)}{1/n} = n \left[f\left(x + \frac{1}{n}\right) - f(x)\right].$$

(The definition may need slight modification if working on a closed interval; see Exercise V.12.3.1.). Then each $g_n$ is continuous, and $g_n \to f'$ pointwise. \(\Box\)
V.12.2. Other Properties of Derivatives

V.12.2.1. Derivatives have another property characteristic of continuous functions: they have the Intermediate Value Property, or Darboux Property, i.e. they satisfy the conclusion of the Intermediate Value Theorem (V.8.2.13.).

V.12.2.2. So which functions are derivatives, i.e. if $g$ is a function on an interval $I$, when is there a differentiable function $f$ on $I$ with $f' = g$? This is a difficult question, and no reasonable necessary and sufficient conditions are known. It is sufficient that $g$ be continuous, but not necessary. And it is necessary that $g$ be a Baire Class 1 function with the Intermediate Value Property, but this is not sufficient either:

V.12.2.3. Example. Let $g_t(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$ for $x \neq 0$, and $g_t(0) = t$ (cf. V.12.1.3.). Then $g_t$ is Baire Class 1 for any $t$, and has the Intermediate Value Property if $-1 \leq t \leq 1$. But $g_t$ is a derivative if and only if $t = 0$. (The only differentiable function $f$ with $f(0) = 0$ for which $f'$ could be $g_t$ is $f(x) = x^2 \sin \frac{1}{x}$; the derivative of this $f$ is $g_0$.)

V.12.2.4. A derivative must also be H-K integrable over every closed bounded subinterval of $I$, which is a quite mild restriction, although not automatic for a Baire Class 1 function with the Intermediate Value Property (e.g. the function $g(x) = \frac{1}{x} \sin^2 \frac{1}{x}$, $g(0) = 0$.)

V.12.2.5. One can give one necessary and sufficient condition: we must have that

$$\lim_{h \to 0^+} \frac{1}{h} \int_{x_0}^{x_0+h} g(t) dt \quad \text{and} \quad \lim_{h \to 0^+} \frac{1}{h} \int_{x_0-h}^{x_0} g(t) dt$$

must both exist and be equal, for every $x_0$, and the value of $g$ at $x_0$ must be equal to this common value, i.e. for every $x_0$, the value of $g$ at $x_0$ must be approximately equal to the average value of $g$ over short intervals to the left and right of $x_0$. However, this condition is not a useful one to check in most cases. Compare with V.8.2.15. and V.9.4.11..

V.12.2.6. Example. Let $g(t) = 1$ for $t > 0$ except for thin “spikes” down to 0 around 1/n, and $g(-t) = -g(t)$ for $t > 0$. (See Fig. 651.) Then $g$ is bounded, continuous except at 0 (hence Riemann-integrable over every interval), but $\frac{1}{h} \int_0^h g(t) dt$ is approximately 1 and $\frac{1}{h} \int_{-h}^0 g(t) dt$ is approximately $-1$ if the spikes are thin enough. So $g$ is not a derivative no matter how $g(0)$ is defined, i.e. $g$ is not equal a.e. to the derivative of an everywhere differentiable function. If $-1 \leq g(0) \leq 1$, then $g$ has the Intermediate Value Property.

See the article [] for more information on which functions are derivatives.

V.12.3. Exercises

V.12.3.1. Let $f$ be a differentiable function on a closed bounded interval $[a, b]$. Find a sequence of continuous functions on $[a, b]$ converging pointwise to $f'$. [Modify the functions $g_n$ in the proof of V.12.1.8. on $[b - \frac{1}{n}, b]$.]
V.12.3.2.  Let $\phi(x) = x^2 \sin \left( \frac{x}{2} \right)$ for $x \neq 0$ and $\phi(0) = 0$. Then $\phi$ is differentiable everywhere (\).

(a) Show that $|\phi(x)| \leq 1$ and $|\phi'(x)| \leq 3$ for $x \in [-1, 1]$.

(b) Let $(q_k)$ be an enumeration of the rationals in $[0, 1]$, and for each $k$ set $f_k(x) = 2^{-k} \phi(x - q_k)$. Then $f_k$ is differentiable everywhere, $|f_k(x)| \leq \frac{1}{2^k}$ and $|f'(x)| \leq \frac{1}{2^k}$ for $x \in [0, 1]$.

(c) Show that $f_k(x)$ is differentiable everywhere.

(d) Show that $f_k(x)$ is differentiable everywhere.

V.12.3.3.  [vRS82, 14.3–14.5] If $I = [a, b]$ is a closed bounded interval, let $g_I$ be the function with graph as in Figure (\), with $L_I = \frac{b-a}{4}$.

(a) Let $f_I$ be an antiderivative of $g_I$ with $f_I(a) = 0$. Show that $f(b) = 0$ and $0 \leq f(x) \leq L_I^2$ for all $x \in [a, b]$.

(b) Let $K$ be a Cantor set of positive measure in $[0, 1]$, and $(I_n)$ an enumeration of the closures of the open intervals in $[0, 1] \setminus K$. Let $f$ be the function on $[0, 1]$ which is 0 on $K$ and $f_{I_n}$ on $I_n$, and similarly $g$ the function which is 0 on $K$ and $g_{I_n}$ on $I_n$. Show that $f$ is differentiable on $[0, 1]$ and $f' = g$. [It is only necessary to show that $f'(a) = 0$ for $a \in K$. If $f(x) \neq 0$, then $x \in I_n$ for some $n$, so

$$\left| \frac{f(x) - f(a)}{x - a} \right| = \frac{f_{I_n}(x)}{|x - a|} \leq \frac{L_{I_n}^2}{|x - a|} \leq |x - a|$$

since $|x - a| \geq L_{I_n}$.]

(c) Show that for $a \in K$, $\limsup_{x \to a} g(x) = 1$, so $g$ is discontinuous at each point of $K$, i.e. $f'$ is discontinuous on a set of positive measure.

(d) Show that for $a \in K$, $\liminf_{x \to a} g(x) = 0$, so $g$ is lower semicontinuous.

(e) Let $K_n$ be a Cantor set in $[0, 1]$ of measure $1 - \frac{1}{n}$, and $h_n$ the $g$ constructed as above for $K_n$. Set $h = \sum_{n=1}^{\infty} 2^{-n} h_n$. Then $h$ has an antiderivative $\phi$ on $[0, 1]$ (\).

(f) Since each $h_n$ is lower semicontinuous, $h$ is discontinuous wherever one of the $h_n$ is discontinuous (\). Thus $h$ is discontinuous at all points of $\cup K_n$, i.e. $\phi'$ is discontinuous a.e.

V.12.3.4.  [vRS82, 14.J] Let $g$ be a Darboux continuous function on an interval $I$ which is of Baire Class 1 (e.g. $g = f'$ for a differentiable function $f$ on $I$). Show that the graph of $g$ is connected.
When is a function $f$ from $\mathbb{R}$ (or $[a,b]$) to $\mathbb{R}$ the derivative of an everywhere differentiable function? Necessary conditions:

(1) $f$ is Baire class 1, i.e. a pointwise limit of a sequence of continuous functions [if $f = F'$, set $f_n(t) = n[F(t + 1/n) - F(t)]$; then $f_n$ is continuous and $f_n \to f$.] This implies:

   (a) $f$ is Lebesgue measurable.

   (b) $f$ is continuous except on a set of first category [\cite{?}].

(2) $f$ is Henstock-Kurzweil integrable on every bounded interval [\cite{?}].

(3) $f$ has the Intermediate Value Property.

These properties are not sufficient:

**Examples.**

(i) Let $f(t) = \sin(1/t)$ for $t \neq 0$. For any $\alpha$, $-1 \leq \alpha \leq 1$, define $f(0) = \alpha$. Then $f$ has the above properties, but is a derivative if and only if $\alpha = 0$.

(ii) Let $f(t) = 1$ for $t > 0$ except for thin “spikes” down to 0 around $1/n$, and $f(-t) = -f(t)$ for $t > 0$, $f(0) = 0$. Then $f$ has the above properties, is bounded, continuous except at 0, but not equal a.e. to the derivative of an everywhere differentiable function.
V.13. The Gamma Function

The Gamma function, while not as well known as trigonometric, exponential, and logarithm functions, is of comparable importance in both theoretical and applied mathematics, and notably in probability and statistics. In this section, we will only develop the most basic aspects of the theory of this function.

The origin of the Gamma function was an attempt to find an analytic expression for the factorial function and extend its definition from \( \mathbb{N} \) to \( \mathbb{R} \). Euler found both infinite product and improper integral formulas for \( n! \), both of which also make sense for nonintegral \( n \):

\[
n! = \prod_{k=1}^{\infty} \left( 1 + \frac{1}{k} \right)^n = \int_0^1 (-\log s)^n \, ds .
\]

V.13.1. Definition of the Gamma Function

Legendre (1811) used the integral expression of Euler, modified by the substitution \( t = -\log s \), to give the first formal definition of the Gamma function:

V.13.1.5. Definition. If \( x > 0 \), define

\[
\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt .
\]

Note that Legendre’s definition, for unknown reasons, has a shift of 1 in the argument, so that \( \Gamma(n) = (n-1)! \) for \( n \in \mathbb{N} \). It has remained the convention in mathematics to this day to include this shift, although it has dubious practical merit.

V.13.1.6. Proposition. The improper Riemann integral \( \int_0^\infty t^{x-1} e^{-t} \, dt \) converges for \( x > 0 \) (and only for \( x > 0 \)).

Proof: The integrand is continuous and positive on \((0, \infty)\), hence the Riemann integral over any closed bounded subinterval is defined. The integral \( \int_0^\infty \) converges if and only if \( I_1 = \int_0^1 \) and \( I_2 = \int_1^\infty \) both converge. For \( I_1 \), note that \( t^{x-1} e^{-t} < t^{x-1} \) for \( 0 < t \leq 1 \), and \( \int_0^1 t^{x-1} \, dt \) converges for \( x > 0 \). (For the parenthetical assertion, \( t^{x-1} e^{-t} \geq e^{-1} t^{x-1} \) for \( 0 < t \leq 1 \), and \( \int_0^1 t^{x-1} \, dt \) diverges for \( x \leq 0 \).)

For \( I_2 \), note that, for any \( x \in \mathbb{R} \),

\[
\lim_{t \to +\infty} \frac{t^{x-1}}{e^{t/2}} = 0
\]

by (), and hence \( t^{x-1} e^{-t} < e^{-t/2} \) for all sufficiently large \( t \). Since \( \int_1^\infty e^{-t/2} \, dt \) converges, so does \( I_2 \) for any \( x \in \mathbb{R} \).

The next result gives a crucial factorial-like recurrence property of the Gamma function.
V.13.7. PROPOSITION. If \( x > 0 \), then \( \Gamma(x + 1) = x\Gamma(x) \).

PROOF: Let \( 0 < a < b < \infty \). Compute the integral \( \int_{a}^{b} t^{x} e^{-t} dt \) by parts:

\[
\int_{a}^{b} t^{x} e^{-t} dt = \left[ -t^{x} e^{-t} \right]_{a}^{b} - \int_{a}^{b} x t^{x-1} (-e^{-t}) dt
\]

\[
= -a^{x} e^{-a} + b^{x} e^{-b} + x \int_{a}^{b} t^{x-1} e^{-t} dt .
\]

As \( a \to 0 \) and \( b \to +\infty \), the original integral approaches \( \Gamma(x + 1) \) and the final expression approaches \( x\Gamma(x) \).

V.13.8. COROLLARY. If \( n \in \mathbb{N} \cup \{0\} \), then \( \Gamma(n + 1) = n! \). [Recall that \( 0! = 1 \) by definition.]

PROOF: This is an easy proof by induction using the recurrence formula and the direct computation

\[
\Gamma(1) = \int_{0}^{\infty} e^{-t} dt = \lim_{b \to +\infty} \int_{0}^{b} e^{-t} dt = \lim_{b \to +\infty} [-e^{-t}]_{0}^{b} = \lim_{b \to +\infty} [-e^{-b} + 1] = 1 = 0! .
\]

So the Gamma function extends the (shifted) factorial function to \((0, \infty)\). This extension has the desired analytic property:

V.13.9. THEOREM. The Gamma function is real analytic (and in particular \( C^{\infty} \)) on \((0, \infty)\), and

\[
\Gamma^{(n)}(x) = \int_{0}^{\infty} t^{x-1} e^{-t} (\log t)^{n} dt .
\]

This is a special case of the general result concerning differentiation under the integral sign. It is routine to show that the integral giving \( \Gamma^{(n)}(x) \) converges for \( x > 0 \), and this follows anyway from the general theory.

V.13.10. NOT MANY VALUES OF \( \Gamma(x) \) FOR NONINTEGER \( x \) CAN BE COMPUTED EXACTLY (OF COURSE, MODERN COMPUTERS CAN EASILY CALCULATE VALUES TO GREAT NUMERICAL ACCURACY). ONE IMPORTANT VALUE WHICH CAN BE COMPUTED IS

\[
\Gamma \left( \frac{1}{2} \right) = \sqrt{\pi} .
\]
V.13.2. Log-Convexity and a Characterization of the Gamma Function

V.13.2.11. The Gamma function is far from the only function on \((0, \infty)\) extending the (shifted) factorial: an extension can be defined arbitrarily for nonintegers. \(\Gamma\) is not even the only real-analytic function extending the factorial: if \(f\) is any real-analytic function vanishing at the positive integers, e.g. \(f(x) = \sin \pi x\), then \(\Gamma + f\) is an extension.

Satisfying the functional equation \(g(x + 1) = xg(x)\) for all \(x > 0\) is only slightly more difficult to achieve. In fact, if a function \(g\) is defined arbitrarily on \((0, 1)\), set \(g(1) = 1\) and extend \(g\) by repeatedly applying the functional equation. It is not even hard to find another real-analytic extension satisfying the functional equation, e.g. \(g(x) = \Gamma(x)(1 + \sin 2\pi x)\) \((1 + \sin 2\pi x\) can be replaced by any periodic real-analytic function \(f\) with period 1 and \(f(1) = 1\).

V.13.2.12. However, \(\Gamma\) seems to be the “simplest” and/or “most natural” extension, due to the multitude of formulas for it and its appearance in many contexts. Many leading mathematicians over almost a 200-year period tried to make this precise with a natural characterization, with varied success. The best characterization (to date) was obtained in 1922 by H. Bohr and J. Mollerup using the notion of log-convexity.

V.13.2.13. It is immediate from V.13.1.9. that \(\Gamma''(x) > 0\) for \(x \in (0, \infty)\), and thus that \(\Gamma\) is a convex function on \((0, \infty)\). But more is true.

V.13.2.14. **Definition.** If \(I\) is an interval, a function \(g : I \to (0, \infty)\) is log-convex on \(I\) if \(\log g\) is convex on \(I\).

V.13.2.15. Alternatively, a positive function \(g\) on an interval \(I\) is log-convex on \(I\) if, for all \(a, b \in I\), \(0 \leq \lambda \leq 1\),

\[
g(\lambda a + (1 - \lambda)b) \leq g(a)^\lambda g(b)^{1 - \lambda}.
\]

The right-hand side is a sort of weighted geometric mean of \(g(a)\) and \(g(b)\). If \(g\) is continuous, then by (\() \) \(g\) is log-convex if and only if, for every \(a, b \in I\),

\[
g \left( \frac{a + b}{2} \right) \leq \sqrt{g(a)g(b)}.
\]

V.13.2.16. **Proposition.** A log-convex function is convex.

**Proof:** Let \(g\) be log-convex on \(I\), and let \(a, b \in I\), and \(0 \leq \lambda \leq 1\). Since the log function is concave, we have

\[
\log(\lambda g(a) + (1 - \lambda)g(b)) \geq \lambda \log g(a) + (1 - \lambda) \log g(b).
\]

On the other hand, since \(g\) is log-convex, we have

\[
\log g(\lambda a + (1 - \lambda)b) \leq \lambda \log g(a) + (1 - \lambda) \log g(b).
\]

Combining these, we get

\[
\log g(\lambda a + (1 - \lambda)b) \leq \log(\lambda g(a) + (1 - \lambda)g(b))
\]

and it follows that \(g(\lambda a + (1 - \lambda)b) \leq \lambda g(a) + (1 - \lambda)g(b)\) since the log function is increasing. \(\diamond\)
V.13.2.17. The converse is false. \( g(x) = x^2 \) is convex on \( (0, \infty) \), but not log-convex on any subinterval. In fact, an increasing log-convex function on an interval \( (a, \infty) \) must grow at least exponentially as \( x \to +\infty \).

V.13.2.18. Since a sum of convex functions is convex, it is obvious that a product of log-convex functions is log-convex. It is also clear that a positive scalar multiple of a log-convex function is log-convex. It is less obvious, but true, that a sum of log-convex functions is log-convex (XIV.15.5.7.).

V.13.2.19. Lemma. Let \( a_1, a_2, b_1, b_2, c_1, c_2 \) be positive real numbers. If \( a_1c_1 - b_1^2 \geq 0 \) and \( a_2c_2 - b_2^2 \geq 0 \), then

\[
(a_1 + a_2)(c_1 + c_2) - (b_1 + b_2)^2 \geq 0.
\]

Proof: If \( a, b, c \) are positive real numbers, then the polynomial \( f(x) = ax^2 + 2bx + c \) takes only nonnegative real values (for \( x \) real) if and only if \( ac - b^2 \geq 0 \). By assumption, the polynomials \( f_1(x) = a_1x^2 + 2b_1x + c_1 \) and \( f_2(x) = a_2x^2 + 2b_2x + c_2 \) take only nonnegative real values, and hence

\[
f_1(x) + f_2(x) = (a_1 + a_2)x^2 + 2(b_1 + b_2)x + (c_1 + c_2)
\]

also takes only nonnegative real values.

V.13.2.20. Proposition. A sum of log-convex functions on an interval \( I \) is log-convex on \( I \).

Proof: Since log-convex functions are continuous, by XIV.15.5.2. it suffices to show that if \( f \) and \( g \) are log-convex on \( I \) and \( x, y \in I \), then

\[
\left[ f \left( \frac{x+y}{2} \right) + g \left( \frac{x+y}{2} \right) \right]^2 \leq (f(x) + g(x))(f(y) + g(y))
\]

Apply XIV.15.5.6. with \( a_1 = f(x), a_2 = g(x), b_1 = f \left( \frac{x+y}{2} \right), b_2 = g \left( \frac{x+y}{2} \right), c_1 = f(y), c_2 = g(y) \).

V.13.2.21. The conclusion of XIV.15.5.7. extends to arbitrary finite sums. Since a positive pointwise limit of log-convex functions is obviously log-convex, it also extends to a positive pointwise limit of finite sums of log-convex functions, e.g. a pointwise-convergent infinite series. By approximating the integrals by Riemann sums, we also obtain:

V.13.2.22. Proposition. Let \( f \) be a function on \( I \times [a,b] \) for some interval \( I \) and \( a, b \in \mathbb{R}, a < b \). If \( f(x, t) \) is log-convex on \( I \) for each fixed \( t \), and \( f(x, t) \) is continuous on \( [a, b] \) for each fixed \( x \), then

\[
g(x) = \int_a^b f(x, t) \, dt
\]

is log-convex on \( I \).

Proof: Let \( n \in \mathbb{N} \) and \( h = \frac{b-a}{n} \). Then \( g_n(x) = h(f(x, a) + f(x, a+h) + \cdots + f(x, a+(n-1)h) \) is log-convex, and \( g_n \to g \) pointwise. Since \( f \) is strictly positive, it is obvious that \( g \) is also strictly positive.

This result extends to improper integrals:
V.13.2.23. **Corollary.** Let \( f \) be a function on \( I \times (a, b) \) for some interval \( I \) and \( a, b \in [-\infty, \infty], \ a < b. \) If \( f(x, t) \) is log-convex on \( I \) for each fixed \( t, f(x, t) \) is continuous on \( (a, b) \) for each fixed \( x, \) and the improper Riemann integral \( \int_a^b f(x, t) \, dt \) converges for all \( x \in I, \) then

\[ g(x) = \int_a^b f(x, t) \, dt \]

is log-convex on \( I. \)

**Proof:** Let \( (a_n) \) be a decreasing sequence with \( a_n \to a, \) \( (b_n) \) an increasing sequence with \( b_n \to b, \) \( a_n < b_n \) for all \( n. \) Set

\[ g_n(x) = \int_{a_n}^{b_n} f(x, t) \, dt. \]

Then \( g_n \) is log-convex on \( I \) for all \( n \) by XIV.15.5.9., and \( g_n \to g \) pointwise on \( I. \) It is also clear that \( g \) is strictly positive.

The next theorem is the main result of this section. Our proof is adapted from the Planet Math website http://planetmath.org/?op=getobj&from=objects&id=3808.

V.13.2.24. **Theorem.** The Gamma function is log-convex on \( (0, \infty). \)

**Proof:** We have that \( \Gamma(x) \in (0, \infty) \) for \( x > 0. \) Let \( a, b > 0 \) and \( 0 \leq \lambda \leq 1. \)

\[ \log \Gamma(\lambda a + (1 - \lambda)b) = \log \int_0^\infty e^{-t} t^{\lambda a + (1 - \lambda)b - 1} \, dt = \log \int_0^\infty (e^{-t} t^{a - 1})^\lambda (e^{-t} t^{b - 1})^{1 - \lambda} \, dt. \]

Apply Hölder’s inequality \( ( ) \) with \( p = \frac{1}{\lambda} \) and \( q = \frac{1}{1 - \lambda} \) to obtain

\[ \log \int_0^\infty (e^{-t} t^{a - 1})^\lambda (e^{-t} t^{b - 1})^{1 - \lambda} \, dt \leq \log \left( \left( \int_0^\infty e^{-t} t^{a - 1} \, dt \right)^\lambda \left( \int_0^\infty e^{-t} t^{b - 1} \, dt \right)^{1 - \lambda} \right) \]

\[ = \lambda \log \Gamma(a) + (1 - \lambda) \log \Gamma(b). \]

Theorem. [Bohr-Mollerup] The Gamma function is the unique function \( g \) on \( (0, \infty) \) which is log-convex and satisfies \( g(1) = 1 \) and \( g(x + 1) = xg(x) \) for all \( x > 0. \)

**Proof:** Since \( \Gamma \) satisfies the conditions, we need only prove uniqueness. In fact, it suffices to show that the conditions uniquely determine a function on \( (0, 1), \) since the rest of the values are then uniquely determined by the recurrence formula.

So let \( g \) be a function satisfying the conditions, \( 0 < x < 1, \) and \( n \in \mathbb{N}. \) Using \( n + x = (1 - x)n + x(n + 1), \) we have

\[ g(n + x) \leq g(n)^{1-x} g(n + 1)^x = g(n)^{1-x} g(n)^x n^x = (n - 1)! n^x. \]
Similarly, \( n + 1 = x(n + x) + (1 - x)(n + 1 + x) \), so \( n! \leq g(n + x)(n + x)^{1-x} \).

Combining these two we get

\[
n!(n + x)^{x-1} \leq g(n + x) \leq (n - 1)!n^x
\]

and by using the recurrence relation to obtain \( g(n + x) = x(x + 1)(x + 2) \cdots (x + n - 1) \),

\[
a_n := \frac{n!(n + x)^{x-1}}{x(x + 1) \cdots (x + n - 1)} \leq g(x) \leq \frac{(n - 1)!n^x}{x(x + 1) \cdots (x + n - 1)} =: b_n .
\]

These inequalities hold for every positive integer \( n \). We have

\[
\lim_{n \to \infty} \frac{b_n}{a_n} = \lim_{n \to \infty} \left( \frac{n}{n + x} \right)^{x-1} = 1
\]

so

\[
1 \leq \frac{g(x)}{a_n} \leq \frac{b_n}{a_n} \to 1
\]

and thus \( g(x) = \lim_{n \to \infty} a_n \), so \( g \) is uniquely determined.

As a corollary of the proof, we get a formula for the unique solution, which must be \( \Gamma \). This formula is due to Gauss, and is equivalent to Euler’s infinite product formula from the introduction to the section.

\textbf{V.13.2.26. Corollary.} For \( 0 < x \leq 1 \)

\[
\Gamma(x) = \lim_{n \to \infty} \frac{n!n^x}{x(x + 1) \cdots (x + n)} .
\]

\textbf{V.13.3. Miscellaneous Additional Results}

We give only a few additional sample results about the Gamma function. More are pursued in the Exercises.

\textbf{V.13.3.27.} The integral definition of \( \Gamma(x) \) makes sense if \( x \) is a complex number with positive real part, and defines a complex analytic function on the right half of the complex plane. This function can be analytically continued to the entire complex plane, with simple poles at the nonpositive integers. In fact, Gauss’s formula (V.13.2.26) holds for any complex number \( x \) which is not a nonpositive integer. It turns out that this complex \( \Gamma \) has no zeros, so the reciprocal is an entire function.

\textbf{V.13.3.28.} Unlike the trigonometric, exponential, and logarithm functions, the Gamma function does not satisfy any \textit{algebraic differential equation} (roughly, a differential equation whose coefficient functions are algebraic functions). This theorem was proved by Hölder in 1887. In particular, there is no formula for the derivative or antiderivative of the Gamma function in terms of standard functions, including the Gamma function itself.
There is an almost bewildering array of other formulas, identities, and applications related to the Gamma function. There are entire books (e.g. [Art64]) written on the subject. See [Dav59] for a good survey of the history and importance of this function. In Davis’s words, “[E]ach generation has found something of interest about the gamma function. Perhaps the next generation will also.”
V.13.4. Stirling’s Formula

In this section, we establish an important and useful approximation to the factorial and gamma functions, attributed to James Stirling (1730), but primarily due to A. de Moivre.

V.13.4.1. Theorem. [Stirling’s Formula] For any \( n \in \mathbb{N} \), we have

\[
\sqrt{2\pi n} \, n^n e^{-n} < n! < \sqrt{2\pi n} \, n^n e^{-n} e^{\frac{1}{12n}}.
\]

In particular,

\[
\lim_{n \to \infty} \frac{n!}{\sqrt{2\pi n} \, n^n e^{-n}} = 1.
\]

More generally,

\[
\lim_{x \to +\infty} \frac{\Gamma(x)}{\sqrt{\frac{2\pi}{x} x^x e^{-x}}} = 1.
\]

Informally, we have the approximation

\[
n! \approx \sqrt{2\pi n} \, n^n e^{-n} = \sqrt{2\pi n} \left( \frac{n}{e} \right)^n
\]

\[
\Gamma(x) \approx \sqrt{\frac{2\pi}{x} x^x e^{-x}} = \sqrt{\frac{2\pi}{x} \left( \frac{x}{e} \right)^x}
\]

for \( n \in \mathbb{N}, x > 0 \). The theorem gives precise error estimates. Better error estimates can be obtained by making more careful estimates along the way in the proof; see Exercise ()

V.13.4.2. Stirling’s formula can be written as

\[
n \log n - n + \frac{1}{2} \log(2\pi n) < \log(n!) < n \log n - n + \frac{1}{2} \log(2\pi n) + \frac{1}{12n}.
\]

A crude version of this formula can be obtained by approximating

\[
\log(n!) = \sum_{k=1}^{n} \log k = \sum_{k=2}^{n} \log k
\]

by integrals. We have

\[
\int_{1}^{n} \log x \, dx < \log(n!)
\]

\[
< \int_{2}^{n+1} \log x \, dx = (n + 1) \log(n + 1) - (n + 1) - (2 \log 2 - 2) = n \log(n + 1) - n + \log(n + 1) + (1 - 2 \log 2).
\]

This gives

\[
e \left( \frac{n}{e} \right)^n = n^n e^{-n+1} < n! < \frac{e}{4} (n+1)^{n+1} e^{-n} = \frac{e(n+1)}{4} \left( \frac{n+1}{e} \right)^n.
\]
Better approximations can be made using the Euler-Maclaurin formula to compute the difference between 
\( \log(n!) \) and \( \int_1^n \log x \, dx \), and the theorem can be proved this way. See ( ).

We will proceed differently. The proof we give consists of a sequence of calculations, some of which yield other formulas of interest.

V.13.4.3. Recall that if \( |x| < 1 \), we have

\[ \log(1 + x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k. \]

Thus, we have, for \( |x| < 1 \), combining terms and using some reindexing,

\[
\log \left( \frac{1 + x}{1 - x} \right) = \log(1 + x) - \log(1 - x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k - \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (-x)^k
= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k + \sum_{k=1}^{\infty} -\frac{1}{k} x^k = \sum_{k=0}^{\infty} \frac{2}{2k + 1} x^{2k+1} = \frac{2x}{2k + 1} x^{2k+1}.
\]

This is in itself a useful formula.

V.13.4.4. Now set \( x = \frac{1}{2n+1} \) for \( n \in \mathbb{N} \). Then \( \frac{1+x}{1-x} = \frac{n+1}{n} = 1 + \frac{1}{n} \), so we obtain

\[
\left( n + \frac{1}{2} \right) \log \left( 1 + \frac{1}{n} \right) = \sum_{k=0}^{\infty} \frac{1}{2k + 1} \left( \frac{n+1}{n} \right)^{2k}.
\]

Therefore

\[
\left( n + \frac{1}{2} \right) \log \left( 1 + \frac{1}{n} \right) < 1 + \frac{1}{3} \sum_{k=1}^{\infty} \frac{1}{(2n+1)^{2k}}
\]

and the last sum is a geometric series with sum

\[
\frac{1}{1 - \left( \frac{1}{2n+1} \right)^2} = \frac{1}{(2n+1)^2 - 1} = \frac{1}{4n(n+1)}
\]

so we obtain the estimates

\[
1 < \left( n + \frac{1}{2} \right) \log \left( 1 + \frac{1}{n} \right) < 1 + \frac{1}{12n(n+1)}
\]

\[
e < \left( 1 + \frac{1}{n} \right)^{n+\frac{1}{2}} < e^{\frac{1}{12n(n+1)}}
\]

\[
1 < \frac{1}{e} \left( 1 + \frac{1}{n} \right)^{n+\frac{1}{2}} < e^{\frac{1}{12n(n+1)}}
\]

for \( n \in \mathbb{N} \).
V.13.4.5. Now set \( a_n = n!e^{-n}n^{-\frac{1}{2}} \), \( y_n = e^{-\frac{1}{12n}} \).

We have
\[
1 < \frac{a_n}{a_{n+1}} = \frac{1}{e} \left(1 + \frac{1}{n}\right)^{n+\frac{1}{2}} < e^{\frac{1}{12(n+1)}} = \frac{y_{n+1}}{y_n}.
\]

Thus the sequence \((a_n)\) is nonnegative and decreasing, hence converges to a limit \( a \). The sequence \((a_ny_n)\) is increasing, and since \( y_n \to 1 \), \( a_ny_n \to a \), and \( a_ny_n < a < a_n \) for all \( n \). Thus
\[
a < a_n < ae^{\frac{1}{12n}}
\]
for all \( n \). Since
\[
n! = a_n\sqrt{n}n^ne^{-n}
\]
we obtain that
\[
a\sqrt{n}n^ne^{-n} < n! < a\sqrt{n}n^ne^{-n}e^{\frac{1}{12n}}
\]
for \( n \in \mathbb{N} \). (This version is due to DE MOIVRE.)

V.13.4.6. It remains to show that \( a = \sqrt{2\pi} \) (this was STIRLING’s contribution to the formula).

V.13.4.7. To obtain the formula for the Gamma function, note that
\[
\lim_{n \to \infty} \frac{n\Gamma(n)}{\sqrt{2\pi n}n^ne^{-n}} = \lim_{n \to \infty} \frac{n!}{\sqrt{2\pi n}n^ne^{-n}} = 1
\]
\[
\lim_{n \to \infty} \frac{\Gamma(n)}{\sqrt{\frac{2\pi}{n}}n^ne^{-n}} = 1.
\]

To transition to the limit as \( x \to +\infty \) through positive real numbers, use XIV.15.7.10. and log convexity of the Gamma function:

The proof using the Euler-Maclaurin formula works directly for the Gamma function for \( x > 0 \) (and even for complex numbers away from the negative real axis). See, for example, [Str81, p. 468].

V.13.4.8. There is a more precise series giving an approximation to the factorial and the gamma function:
\[
n! \approx \sqrt{2\pi n}n^ne^{-n}\left[1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} - \frac{571}{2488320n^4} + \ldots\right]
\]
\[
\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}x^xe^{-x}\left[1 + \frac{1}{12x} + \frac{1}{288x^2} - \frac{139}{51840x^3} - \frac{571}{2488320x^4} + \ldots\right]
\]

(the coefficients have a closed formula which is complicated; see e.g. http://mathworld.wolfram.com/StirlingsSeries.html.)

This is an asymptotic expansion () and not a convergent series: for fixed \( n \) or \( x \) adding more terms eventually gives a worse approximation. The relative error after taking finitely many terms is always smaller than the next term.

663
V.13.4.9. By approximating the terms in the asymptotic expansion, we obtain a better asymptotic formula for $n!$:

$$n! \approx \sqrt{\pi(n/e)^n} \left(2n + \frac{1}{3}\right)^{n}.$$ 

This approximation is due to R. W. Gosper.

V.13.4.10. Another numerically efficient approximation, due to G. Nemes [?], is

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}} \left[ \frac{1}{e} \left( x + \frac{1}{12x - \frac{1}{10x}} \right) \right]^x.$$ 

V.13.5. Exercises
V.14. The Volume of the Unit Ball in $\mathbb{R}^n$

Throughout this section, we will use the Euclidean norm
\[
\| (x_1, \ldots, x_n) \| = \sqrt{x_1^2 + \cdots + x_n^2}.
\]

Recall that the closed unit ball and unit sphere in $\mathbb{R}^n$ are
\[
B^n = \{ x \in \mathbb{R}^n : \| x \| \leq 1 \}
\]
\[
S^{n-1} = \{ x \in \mathbb{R}^n : \| x \| = 1 \}.
\]

Our objective in this section is to compute the Lebesgue measure $\beta_n = \lambda^n(B^n)$ of the unit ball in $\mathbb{R}^n$. These numbers will be needed for a number of purposes, such as the development of Hausdorff measure, and are intimately tied up with the measure theory of hypersurfaces and other subsets of Euclidean space (such as the $(n-1)$-dimensional “hypersurface area” of $S^{n-1}$).

V.14.0.1. We can compute $\beta_n$ for small values of $n$ by elementary or calculus methods:

We have $\beta_1 = \lambda([-1,1]) = 2$.

We have that $\beta_2$ is the area of the unit disk in $\mathbb{R}^2$. This can be computed as in (1) or (1) to obtain $\beta_2 = \pi$.

Using spherical coordinates (1), we may obtain $\beta_3 = \frac{4}{3} \pi$.

Actually, $\beta_n$ can be calculated for any $n$ just using calculus, either directly (Exercise V.14.4.1.) or using “hyperspherical” coordinates (Exercise V.14.4.2.). However, the formula obtained in this way becomes increasingly complicated for large $n$. (Try to guess the right formula!)

Instead, we will use some machinery from measure theory to rather painlessly compute a simple formula for $\beta_n$. This argument is adapted from [Fol99, 2.7].

V.14.1. Radial Borel Sets and Functions

Fix an $n$. We will examine radially symmetric sets in $\mathbb{R}^n$.

V.14.1.2. Definition. A subset $S$ of $\mathbb{R}^n$ is radially symmetric if $x \in S$, $\| y \| = \| x \|$ imply $y \in S$.

$B^n$ and $S^{n-1}$ are radially symmetric sets in $\mathbb{R}^n$. A set is radially symmetric if and only if it is a union of spheres of various radii around the origin.

V.14.1.3. If $A \subseteq [0, \infty)$, the symmetrization of $A$ in $\mathbb{R}^n$ is
\[
s_n(A) = \{ x \in \mathbb{R}^n : \| x \| \in A \}.
\]

It is clear that $s_n(A)$ is radially symmetric. Conversely, identify $[0, \infty)$ with the ray
\[
R_+^{(n)} = \{ (t, 0, \ldots, 0) : t \in [0, \infty) \} \subseteq \mathbb{R}^n.
\]

If $S \subseteq \mathbb{R}^n$ is radially symmetric, set $A = S \cap R_+^{(n)}$; then $S = s_n(A)$. 665
V.14.1.4. If $A$ is a subset of $[0, \infty)$, then $s_n(A) = f^{-1}(A)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is the norm function $f(x) = \|x\|$. Since $f$ is continuous, if $B \subseteq [0, \infty)$ is a Borel set, then $s_n(B)$ is a Borel set in $\mathbb{R}^n$. Conversely, if $S$ is a radially symmetric Borel set in $\mathbb{R}^n$, then $B = S \cap R_+^{(n)}$ is a Borel set in $[0, \infty)$ with $S = s_n(B)$.

V.14.1.5. Definition. A radial Borel set is a radially symmetric Borel subset of $\mathbb{R}^n$. Let $R^{(n)}$ be the set of all radial Borel sets in $\mathbb{R}^n$.

The next proposition is obvious, since the collection of radially symmetric subsets of $\mathbb{R}^n$ is closed under unions, intersections, and complements.

V.14.1.6. Proposition. $R^{(n)}$ is a sub-$\sigma$-algebra of $B^{(n)}$.

V.14.1.7. Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is radially symmetric if $\|x\| = \|y\|$ implies $f(x) = f(y)$. If $g : [0, \infty) \to \mathbb{R}$, the symmetrization $s_n(g)$ is defined by $[s_n(g)](x) = g(\|x\|)$. Then $s_n(g)$ is radially symmetric, and every radially symmetric function is of this form.

V.14.1.8. Proposition. (i) A function $f : \mathbb{R}^n \to \mathbb{R}$ is $R^{(n)}$-measurable if and only if it is Borel measurable and radially symmetric.

(ii) If $g : [0, \infty) \to \mathbb{R}$, then $s_n(g)$ is $R^{(n)}$-measurable if and only if $g$ is Borel measurable.

Proof: If $f$ is $R^{(n)}$-measurable, then $f$ is Borel measurable since $R^{(n)} \subseteq B^{(n)}$, and $\{x : f(x) = \alpha\} \in R^{(n)}$ for any $\alpha \in \mathbb{R}$, so $f$ is radially symmetric. The converse and (ii) are obvious from the fact that $R^{(n)}$ consists precisely of the radially symmetric Borel sets, the symmetrizations of Borel sets in $[0, \infty)$.

The main result of this subsection is:

V.14.1.9. Theorem. Let $g$ be a nonnegative Borel measurable function on $[0, \infty)$. Then

$$\int_{\mathbb{R}^n} s_n(g) d\lambda^{(n)} = n \beta_n \int_0^\infty r^{n-1} g(r) d\lambda(r).$$

Proof: We first prove the result for “annuli.” Let $S = s_n((a, b])$, where $0 \leq a < b < \infty$. Then $S$ is a set-theoretic difference of two closed balls around the origin, so by the scaling invariance of $\lambda^{(n)}$ we have

$$\int_{\mathbb{R}^n} \chi_S d\lambda^{(n)} = \lambda^{(n)}(S) = \beta_n(b^n - a^n) = n \beta_n \int_0^\infty r^{n-1} \chi_{(a, b]}(r) d\lambda(r).$$

Now let $S$ be the collection of all subsets of $\mathbb{R}^n$ of the form $s_n((a, b])$ as above. Then $S$ is a semialgebra, and the $\sigma$-algebra generated by $S$ is $R^{(n)}$. Let $\mu$ be the measure on $(\mathbb{R}^n, R^{(n)})$ given by

$$\mu(s_n(B)) = n \beta_n \int_B r^{n-1} d\lambda(r).$$
Since we have shown that $\mu|_S = \lambda^{(n)}|_S$, we have $\mu = \lambda^{(n)}|_{\mathbb{R}^n}$ by (1). Thus the formula in the theorem holds for radially symmetric characteristic functions, and hence for radially symmetric simple functions by linearity.

Finally, let $(f_k)$ be an increasing sequence of nonnegative simple functions on $[0, \infty)$ converging pointwise to $g$. Then $(s_n(f_k))$ is an increasing sequence of radially symmetric simple functions increasing pointwise to $s_n(g)$, so by two applications of the Monotone Convergence Theorem we get

$$\int_{\mathbb{R}^n} s_n(g) \, d\lambda^{(n)} = \lim_{k \to \infty} \int_{\mathbb{R}^n} s_n(f_k) \, d\lambda^{(n)} = \lim_{k \to \infty} n\beta_n \int_0^\infty r^{n-1} f_k(r) \, d\lambda(r) = n\beta_n \int_0^\infty r^{n-1} g(r) \, d\lambda(r) .$$

The proof of this theorem is somewhat simpler and more elementary if $g$ is continuous, since the second step can be bypassed by using an improper Riemann integral on the right side. This is actually the only case we will use in the calculation of $\beta_n$.

**V.14.2. Calculation of $\beta_n$**

**V.14.2.10.** The trick we will use to calculate $\beta_n$ is to consider the function $f_n : \mathbb{R}^n \to \mathbb{R}$ given for $x = (x_1, \ldots, x_n)$ by

$$f_n(x) = e^{-\|x\|^2} = e^{-x_1^2} \cdots e^{-x_n^2} = e^{-x_1^2} e^{-x_2^2} \cdots e^{-x_n^2} .$$

This function is nonnegative, continuous, and radially symmetric, and is also a product of functions of each coordinate separately; moreover, the integrals can be computed in two separate ways.

**V.14.2.11.** On the one hand, from V.14.1.9. we have

$$I_n := \int_{\mathbb{R}^n} f_n \, d\lambda^{(n)} = n\beta_n \int_0^\infty r^{n-1} e^{-r^2} \, dr$$

(the last integral may be regarded as an improper Riemann integral).

**V.14.2.12.** On the other hand, since $f_n(x_1, \ldots, x_n) = f_1(x_1)f_1(x_2) \cdots f_1(x_n)$, by Tonelli’s theorem (the vector calculus version () suffices) we have

$$I_n = \int_{\mathbb{R}^n} f_n \, d\lambda^{(n)} = \int_{\mathbb{R}} f_1(x_1) \left[ \int_{\mathbb{R}} f_1(x_2) \left[ \cdots \left[ \int_{\mathbb{R}} f_1(x_n) \, d\lambda(x_n) \right] \cdots \right] \, d\lambda(x_2) \right] \, \lambda(x_1)$$

$$= \left[ \int_{\mathbb{R}} f_1(x) \, d\lambda(x) \right]^n = (I_1)^n .$$

**V.14.2.13.** The integral which can be computed by elementary means is $I_2$. Since $\beta_2 = \pi$ (V.14.0.1.), we have

$$I_2 = 2\pi \int_0^\infty r e^{-r^2} \, dr = 2\pi \lim_{b \to \infty} \left[ -\frac{1}{2} e^{-r^2} \right]_0^b = \pi$$

and so $I_1 = \sqrt{\pi}$. 667
So we have
\[(I_1)^n = \pi^{n/2} = I_n = n\beta_n \int_0^\infty r^{n-1}e^{-r^2} \, dr.\]

Making the substitution \(u = r^2\), we have
\[\pi^{n/2} = \frac{n\beta_n}{2} \int_0^\infty u^{\frac{n-2}{2}}e^{-u} \, du = \frac{n\beta_n}{2} \Gamma\left(\frac{n}{2}\right)\]
where \(\Gamma\) is the Gamma function. Solving for \(\beta_n\) and using \(\Gamma\left(\frac{n}{2} + 1\right) = \frac{\pi}{2} \Gamma\left(\frac{n}{2}\right)\), we get the final result:

V.14.2.15. **Theorem.**

\[\beta_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)}.\]

So if \(n = 2m\) is even, we have \(\beta_n = \frac{\pi^m}{m!}\).

V.14.2.16. Using \(\beta_1 = 2\), we can obtain \(\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2} = \frac{1}{2} \Gamma\left(\frac{1}{2}\right)\), so \(\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\). Thus, if \(n = 2m + 1\) is odd, we have
\[\Gamma\left(\frac{n}{2} + 1\right) = \frac{1}{2} \cdot \frac{3}{2} \cdot \frac{5}{2} \cdots \frac{n}{2} \sqrt{\pi} = \frac{1 \cdot 3 \cdot 5 \cdots n}{2^{m+1}} \sqrt{\pi}\]
so we obtain
\[\beta_n = \frac{2^{m+1} \pi^m}{1 \cdot 3 \cdot 5 \cdots n}.\]

V.14.2.17. Thus \(\beta_4 = \frac{\pi^2}{2\pi}, \beta_5 = \frac{8\pi^2}{15}\pi\), etc. Note that each time \(n\) increases by 2, one more factor of \(\pi\) appears. The power of \(\pi\) that occurs in the final formulas is always an integer.

V.14.2.18. We can also define \(\beta_s\) for any \(s \in [0, \infty)\) by the formula
\[\beta_s = \frac{\pi^{s/2}}{\Gamma\left(\frac{s}{2} + 1\right)}.\]

This is the theoretical \(s\)-dimensional “measure of the closed unit ball in \(\mathbb{R}^s\).” This definition will be used in the construction of Hausdorff measure.

V.14.3. Surface Area of a Sphere
V.14.3.19. The arguments of this section can be taken much farther. Fix \(n\). We can define “spherical coordinates” \((\rho, \Theta)\) in \(\mathbb{R}^n \setminus \{0\}\) by setting
\[\rho(x) = \|x\|, \quad \Theta(x) = \frac{x}{\|x\|}.\]

Then \(\rho(x) \in (0, \infty)\) and \(\Theta(x) \in S^{n-1}\), and \(\Phi = (\rho, \Theta)\) is a homeomorphism from \(\mathbb{R}^n \setminus \{0\}\) to \((0, \infty) \times S^{n-1}\); the inverse function \(\Phi^{-1}\) is given by \(\Phi^{-1}(r, x) = rx\). Since these functions are homeomorphisms, they send Borel sets to Borel sets.
\textbf{V.14.3.20.} We can also transfer Lebesgue measure $\lambda^{(n)}$ to the product space $(0, \infty) \times S^{n-1}$ via $\Phi$; call the image measure $\mu$. Thus, if $E$ is a Borel subset of $(0, \infty) \times S^{n-1}$, we have $\mu(E) = \lambda^{(n)}(\Phi^{-1}(E))$. Theorem V.14.1.9. can be rephrased to say that if $A$ is a Borel subset of $(0, \infty)$, then

$$
\mu(A \times S^{n-1}) = n\beta_n \int_A r^{n-1} \, d\lambda(r).
$$

Define a measure $\nu$ on $(0, \infty)$ by $\nu(A) = \int_A r^{n-1} \, d\lambda(r)$. We will show that there is a finite Borel measure $\sigma$ on $S^{n-1}$, with $\sigma(S^{n-1}) = n\beta_n$, such that $\mu = \nu \times \sigma$.

\textbf{V.14.3.21.} If $B$ is a Borel subset of $S^{n-1}$ (in either the abstract sense regarding $S^{n-1}$ as a topological space, or a subset of $S^{n-1}$ which is a Borel set in $\mathbb{R}^n$; since $S^{n-1}$ is a closed, hence Borel, subset of $\mathbb{R}^{n-1}$, the two notions are equivalent), then $\Phi^{-1}((0, 1] \times B)$ is a Borel subset of $\mathbb{R}^n$. Define

$$
\sigma(B) = n\lambda^{(n)}((0, 1] \times B).
$$

It is almost immediate that $\sigma$ is a Borel measure on $S^{n-1}$. The main result is:

\textbf{V.14.3.22.} \textbf{THEOREM.} The measures $\mu$ and $\nu \times \sigma$ coincide on the Borel sets in $(0, \infty) \times S^{n-1}$.

\textbf{PROOF:} In a manner almost identical to the proof of V.14.1.9., for a fixed $B$ we have $\mu(A \times B) = \nu(A)\sigma(B)$ for every Borel subset $A$ of $(0, \infty)$ (first show this for $A = (a, b]$ by homogeneity, and then bootstrap to all Borel sets). Thus, if $A$ is the $\sigma$-algebra of Borel sets in $(0, \infty)$ and $\mathcal{B}$ the Borel sets in $S^{n-1}$, we have $\mu = \nu \times \sigma$ on $A \times \mathcal{B}$. So by () $\mu$ and $\nu \times \sigma$ agree on $A \otimes \mathcal{B}$, which is the $\sigma$-algebra of Borel sets in $(0, \infty) \times S^{n-1}$. \&

\textbf{V.14.3.23.} The measure $\sigma$ is called the \textit{surface measure} of $S^{n-1}$. In the cases $n = 2$ and $n = 3$ this surface measure represents arc length and surface area respectively. This measure can be defined in several other ways: it is the restriction of $(n-1)$-dimensional Hausdorff measure () on $\mathbb{R}^n$ to $S^{n-1}$, it is the Riemannian measure on $S^{n-1}$ as a smooth submanifold of $\mathbb{R}^n ()$, and it can be defined via a parametrization ()

\textbf{V.14.3.24.} The total surface measure $\sigma(S^{n-1})$ is just $n\beta_n$. Thus $\sigma(S^1) = 2\pi$, the circumference of the unit circle, and $\sigma(S^2) = 4\pi$, the surface area of the unit sphere in $\mathbb{R}^3$. By scaling, the surface measure of the sphere of radius $r$ in $\mathbb{R}^n$ is $r^{n-1}\sigma(S^{n-1})$. This is the derivative of the function $V(r) = \beta_n r^n$ giving the Lebesgue measure of a ball in $\mathbb{R}^n$ of radius $r$. An interpretation of this is: if $S_h$ is the spherical shell of inside radius 1 and outside radius $1 + h$ in $\mathbb{R}^n$, we have:

$$
\sigma(S^{n-1}) = \lim_{h \to 0} \frac{\lambda^{(n)}(S_h)}{h}.
$$

A similar formula holds for any Borel subset of $S^{n-1}$. 

669
V.14.4.  Exercises

V.14.4.1.  Let $n \geq 2$. Use iterated integration to compute $\beta_n$ as follows.

(a) Write $\beta_n = \int_{x_1^2 + \cdots + x_n^2 \leq 1} 1 \, d(x_1, \ldots, x_n) = \int_{-1}^{1} \left[ \int_{x_1^2 + \cdots + x_n^2 \leq 1 - x_1^2} 1 \, d(x_2, \ldots, x_n) \right] \, dx_1$.

(b) The quantity in brackets is the $(n-1)$-dimensional measure of a ball of radius $\sqrt{1 - x_1^2}$ in $\mathbb{R}^{n-1}$. By scaling invariance, this integral equals $\beta_{n-1}(1 - x_1^2)^{\frac{n-1}{2}}$.

(c) Obtain a recurrence formula

$$\beta_n = \beta_{n-1} \int_{-1}^{1} (1 - x^2)^{\frac{n-1}{2}} \, dx = 2\beta_{n-1} \int_{0}^{1} (1 - x^2)^{\frac{n-1}{2}} \, dx.$$

(d) Make the change of variables $x = \cos \theta$ to convert the recurrence formula to

$$\beta_n = \beta_{n-1} \int_{0}^{\pi} \sin^n \theta \, d\theta = 2\beta_{n-1} \int_{0}^{\pi/2} \sin^n \theta \, d\theta.$$

(e) Evaluate the integral using (c) and obtain the formula of V.14.2.15. using $\beta_1 = 2$.

V.14.4.2.  Let $n \geq 3$. Compute the $n$-dimensional Lebesgue measure of the ball $B_e^{(n)}$ of radius $r$ around the origin in $\mathbb{R}^n$ as follows.

(a) Introduce “hyperspherical” coordinates $(\rho, \theta_1, \ldots, \theta_{n-1})$ in $\mathbb{R}^n$, where

$$x_1 = \rho \cos \theta_1$$

$$x_2 = \rho \sin \theta_1 \cos \theta_2$$

$$x_{n-1} = \rho \sin \theta_1 \cdots \sin \theta_{n-2} \cos \theta_{n-1}$$

$$x_n = \rho \sin \theta_1 \cdots \sin \theta_{n-2} \sin \theta_{n-1}$$

(for $n = 3$, the usual spherical coordinates $(\rho, \phi, \theta)$ have $\theta_1 = \phi$, $\theta_2 = \theta$, and the coordinates come out in the order $(z, x, y)$).

(b) Show that the $n$-dimensional volume $\lambda^{(n)}(B_e^{(n)})$ is given by

$$\int_{0}^{r} \int_{0}^{\pi} \cdots \int_{0}^{\pi} \rho^{n-1} \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 \cdots \sin \theta_{n-2} \, d\theta_1 \, d\theta_2 \cdots d\theta_{n-1} \, d\rho$$

$$= \left( \int_{0}^{r} \rho^{n-1} \, d\rho \right) \left( \int_{0}^{\pi} d\theta_{n-1} \right) \prod_{k=1}^{n-2} \left( \int_{0}^{\pi} \sin^k \theta_{n-k-1} \, d\theta_{n-k-1} \right)$$

$$= \frac{2\pi}{n} r^n \prod_{k=1}^{n-2} \left( \int_{0}^{\pi} \sin^k \theta \, d\theta \right).$$

(This formula also holds with a simplified argument if $n = 2$ if the last product is taken to be the “empty product” 1.)
(c) Conclude that if \( n \geq 2 \), then
\[
\beta_n = \frac{2\pi}{n} \prod_{k=1}^{n-2} \left( \int_0^\pi \sin^k \theta \, d\theta \right)
\]
(with the last product equal to 1 if \( n = 2 \)). Show that this formula coincides with the formula of V.14.2.15. (cf. ()).

**V.14.4.3.**

(a) Find for which \( n \in \mathbb{N} \) we have \( \beta_n \) maximum.
(b) Find for which \( s \in [0, \infty) \) we have \( \beta_s \) maximum.
(c) Show that \( \lim_{s \to \infty} \beta_s = 0 \).

**V.14.4.4.** [BB92] Show that
\[
\left( 10^{-5} \sum_{-\infty}^{\infty} e^{-\frac{n^2}{10^{10}}} \right)^2
\]
agrees with \( \pi \) to at least 42 billion decimal places, but the numbers are not equal. Which is larger? [Show that
\[
\pi = \left( 10^{-5} \int_{-\infty}^{\infty} e^{-\frac{x^2}{10^{10}}} \, dx \right)^2
\]
(cf. V.14.2.13.).]
V.15. Infinite Series of Functions

V.15.1. Infinite Series of Functions

Term-by-Term Differentiation

V.15.1.1. It is important to understand that even a uniformly convergent infinite series of differentiable functions cannot be differentiated term-by-term in general. For a simple example, the infinite series

\[ \sum_{k=1}^{\infty} 2^{-k} \sin 2^k x \]

converges uniformly on \( \mathbb{R} \) to a (continuous) function \( f \). But the series of term-by-term derivatives

\[ \sum_{k=1}^{\infty} \cos 2^k x \]

do not even converge pointwise at 0 or most other points. (In fact, \( f \) is not even differentiable at most places.) See I.5.1.13. for a heuristic explanation.

We do have the following series version of V.8.5.5.:

V.15.1.2. Theorem. Let \( (F_n) \) and \( (f_n) \) be sequences of functions on an interval \( I \), and \( A \) a countable subset of \( I \). Suppose

(a) Each \( F_n \) is continuous on \( I \).

(b) For every \( x \in I \setminus A \), each \( F_n \) is differentiable at \( x \) and \( F'_n(x) = f_n(x) \).

(c) The series \( \sum_{n=1}^{\infty} f_n \) converges u.c. on \( I \) to a function \( f \).

(d) There is an \( x_0 \in I \) such that the series \( \sum_{n=1}^{\infty} F_n(x_0) \) converges.

Then:

(i) The series \( \sum_{n=1}^{\infty} F_n \) converges u.c. on \( I \) to a (continuous) function \( F \).

(ii) For each \( x \in I \setminus A \), \( F \) is differentiable at \( x \) and \( F'(x) = f(x) \).

As usual, the proof is just an application of V.8.5.5. to the sequence of partial sums.

V.15.1.3. The important special case where \( A = \emptyset \) and each \( f_n \) is continuous follows from XIV.2.9.15.
V.15.2. Exercises

V.15.2.1. (a) Prove the sophomore’s dream:

\[ \int_0^1 x^{-x} \, dx = \int_0^1 \frac{1}{x^x} \, dx = \sum_{k=1}^{\infty} \frac{1}{k^k} \quad \text{and} \quad \int_0^1 x^x \, dx = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^k} . \]

[Write \( x^x = e^{x \log x} \), expand in an infinite series using the Maclaurin series for \( e^x \), and integrate term-by-term.]

The name “sophomore’s dream” was coined in [BBG04].

(b) Another sophomore’s dream:

\[ \int_0^\infty \frac{x}{x^3 + 1} \, dx = \int_0^\infty \frac{1}{x^3 + 1} \, dx . \]

V.15.2.2. Does the iterated integral

\[ \int_0^8 \int_0^{\sqrt{1-z}} 4y \, dy \, dx \]

exist? If so, what is its value? Interpret the integral geometrically.
V.15.3. Power Series

Power series are probably the most important kinds of infinite series of functions (although trigonometric series are also very important). A great many of the most important functions can be represented as power series, and power series have some special properties which make them relatively easy to use and work with.

V.15.3.1. Definition. Let $x_0$ be a real number. A (real) power series centered at $x_0$ is an expression of the form

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k$$

where the $a_k$ are real numbers. Complex power series can be defined similarly: if $z_0$ is a complex number, a complex power series centered at $z_0$ is an expression of the form

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

where the $a_k$ are complex numbers.

While we will (except in Chapter X) generally only consider real power series, the theory of power series is most naturally done for complex power series as part of complex analysis, and there are places in the real theory where the complex point of view is very helpful. Any real power series can, of course, be regarded as a complex power series by replacing $x$ by $z$.

V.15.3.2. A power series is regarded as an infinite series of functions. It is conventional when working with power series to regard $(x - x_0)^0 = 1$ no matter how $x - x_0$ is interpreted; thus the $k = 0$ term of the power series is just the constant function $a_0$. The same convention holds for complex power series.

V.15.3.3. A power series is, at first, only a formal expression, but it can be regarded as defining a function $f$ of $x$, whose domain is the set of all $x$ for which the series converges. Because of our convention, $x_0$ is always in the domain of the defined function $f$, with $f(x_0) = a_0$. It can happen that the domain of $f$ consists only of $\{x_0\}$ (such power series are not very interesting as functions).

The most important basic property of power series is V.15.3.5., which says that the domain of the defined function is an interval, or degenerate interval, centered at $x_0$.

V.15.3.4. Definition. Let $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ be a power series, and define

$$L = \limsup_{k \to \infty} |a_k|^{1/k} = \limsup_{k \to \infty} \sqrt[k]{|a_k|}$$

and let $R = \frac{1}{L}$ if $0 < L < +\infty$, $R = 0$ if $L = +\infty$, and $R = +\infty$ if $L = 0$. $R$ is called the radius of convergence of the power series.

An identical definition can be made for complex power series. Note that the radius of convergence of a power series depends only on the coefficients $a_k$, not on the center $x_0$, of the power series. The slower the coefficients grow in size, or the faster they approach zero, the larger the radius of convergence.

The term “radius of convergence” is justified by the next result:
V.15.3.5. **Theorem.** Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence $R$. Then

(i) If $x \in \mathbb{R}$ and $|x-x_0| < R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges absolutely.

(ii) If $x \in \mathbb{R}$ and $|x-x_0| > R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ diverges.

(iii) If $r \in \mathbb{R}$ and $0 < r < R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges uniformly on the interval $[x_0-r, x_0+r]$.

(Of course, (i) and (iii) are vacuous if $R = 0$ and (ii) is vacuous if $R = +\infty$.)

**Proof:** In cases (i) and (ii), apply the Root Test to the series $\sum_{k=0}^{\infty} |a_k(x-x_0)^k|$. We have

\[
\limsup_{k \to \infty} |a_k(x-x_0)^k|^{1/k} = |x-x_0| \limsup_{k \to \infty} |a_k|^{1/k} = \frac{|x-x_0|}{R}
\]

(it is $+\infty$ if $x \neq x_0$ and $R = 0$, and 0 if $R = +\infty$ or $x = x_0$). Thus the series converges absolutely if $|x-x_0| < R$ or $x = x_0$ and diverges since the terms do not go to zero if $|x-x_0| > R$.

For (iii), use the Weierstrass $M$-test. Set $M_k = |a_k|r^k$. Then $|a_k(x-x_0)^k| \leq M_k$ for all $k$ and for all $x$, $|x-x_0| \leq r$. We have

\[
\limsup_{k \to \infty} M_k^{1/k} = r \limsup_{k \to \infty} |a_k|^{1/k} = \frac{r}{R} < 1
\]

if $0 < R < \infty$, and $\limsup_{k \to \infty} M_k^{1/k} = 0$ if $R = +\infty$. Thus $\sum_{k=0}^{\infty} M_k$ converges by the Root Test. $\Box$

The statement and proof of the theorem work identically for complex power series.

V.15.3.6. **Corollary.** Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence $R > 0$. Then the series converges u.c. on the open interval $(x_0 - R, x_0 + R)$, and the sum is continuous on this interval.

**Proof:** If $[a, b] \subseteq (x_0 - R, x_0 + R)$, then $[a, b] \subseteq [x_0 - r, x_0 + r]$ for some $r < R$. Continuity of the sum follows from (i). $\Box$

The sum is actually far more than just continuous on this open interval.

V.15.3.7. Thus the set on which a power series centered at $x_0$ converges is an interval centered at $x_0$, possibly of length zero, i.e. the degenerate interval $\{x_0\}$, or of infinite length, i.e. all of $\mathbb{R}$, possibly including one or both endpoints, called the *interval of convergence* of the power series. (If $0 < R < \infty$, the theorem has nothing to say about behavior at the points where $x-x_0 = \pm R$, the endpoints of the interval of convergence; endpoint behavior will be discussed below.) Similarly, the set on which a complex power series centered at $z_0$ converges is a disk in $\mathbb{C}$ centered at $z_0$, possibly degenerate or all of $\mathbb{C}$, and possibly including some or all points on the boundary circle of the disk.

675
V.15.3.8. The radius of convergence of a power series \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) can always be computed as

\[
R = \frac{1}{\limsup |a_k|^{1/k}}
\]

(with the usual conventions if the lim sup is 0 or \(+\infty\)), but it is not always easy to evaluate this expression. Sometimes the radius of convergence can be computed more simply by the following:

**V.15.3.9. Proposition.** Let \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) be a power series, with all \( a_k \) nonzero. If

\[
\lim_{k \to \infty} \left| \frac{a_k}{a_{k+1}} \right|
\]

exists (in the usual or extended sense), it equals the radius of convergence of the power series.

**Proof:** Let \( S \) be this limit. Apply the Ratio Test (\()\) to show that \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) converges absolutely if \( |x - x_0| < S \) and diverges if \( |x - x_0| > S \). Thus \( S \) must be exactly the radius of convergence. 🙌

Note that in the limit of V.15.3.9., the \( a_k \) and \( a_{k+1} \) are reversed from the way they appear in the Ratio Test (\()\).

Another convenient characterization of the radius of convergence is the following:

**V.15.3.10. Proposition.** Let \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) be a power series. The power series has radius of convergence \( R > 0 \) if and only if, for all \( r, 0 < r < R \), there is a constant \( M \) such that \( |a_k| \leq \frac{M}{r} \) for all \( k \), and there is no such \( M \) if \( r > R \).

**Proof:** Suppose the series has radius of convergence \( R > 0 \), and \( 0 < r < R \). Then

\[
\limsup |a_k|^{1/k} < \frac{1}{r}
\]

so there is a \( k_0 \) such that \( |a_k| < \frac{1}{r} \) for all \( k \geq k_0 \). There is an \( M \geq 1 \) such that \( |a_k| \leq \frac{M}{r} \) for \( 0 \leq k < k_0 \). This \( M \) works for all \( k \).

Conversely, if the condition holds for all \( r \) with \( 0 < r < R \) and no \( r > R \), we have, for each such \( r \) with \( 0 < r < R \),

\[
\limsup |a_k|^{1/k} \leq \limsup \frac{M^{1/k}}{r} = \frac{1}{r}
\]

for some constant \( M \), so the radius of convergence \( R' \) is \( \geq R \). If \( R' > R \), choose \( r \) so that \( R < r < R' \); then the condition holds for \( r \) by the first half of the proof, a contradiction. 🙌
V.15.3.11. Examples. (i) Let $c \in \mathbb{R}$, $c \neq 0$. The power series $\sum_{k=0}^{\infty} c^k (x - x_0)^k$ has radius of convergence $\frac{1}{|c|}$, and interval of convergence $\left( x_0 - \frac{1}{|c|}, x_0 + \frac{1}{|c|} \right)$.

(ii) The power series $\sum_{k=0}^{\infty} \frac{1}{k!} (x - x_0)^k$ has radius of convergence $+\infty$ and interval of convergence $\mathbb{R}$.

(iii) The power series $\sum_{k=0}^{\infty} k! (x - x_0)^k$ has radius of convergence 0 and (degenerate) interval of convergence $\{x_0\}$.

V.15.3.12. The convergence behavior of an infinite series does not depend on $x_0$, and if $x_0$ is changed, leaving the $a_k$ the same, the function defined is simply changed by a translation. In particular, if $f$ is the function defined by $\sum_{k=0}^{\infty} a_k x^k$ on its interval of convergence $I$, then the function defined by $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ is simply the function $g(x) = f(x - x_0)$ on $x_0 + I$. Thus in analyzing behavior of functions defined by power series, we generally may (and often will) take $x_0 = 0$ to simplify notation.

Endpoint Behavior

Almost any kind of behavior can occur at endpoints of the interval of convergence, as the following examples show.

V.15.3.13. Examples. (i) The power series $\sum_{k=0}^{\infty} x^k$ has radius of convergence 1. The series diverges if $x = \pm 1$ since the terms do not go to zero, so the interval of convergence is $(-1, 1)$.

(ii) The power series $\sum_{k=1}^{\infty} \frac{1}{k} x^k$ (a power series with constant term 0) has radius of convergence 1. The series diverges if $x = 1$ (harmonic series), but converges conditionally if $x = -1$, so the interval of convergence is $[-1, 1)$. The power series $\sum_{k=1}^{\infty} \frac{(-1)^k}{k} x^k$ has interval of convergence $(-1, 1]$.

(iii) The power series $\sum_{k=1}^{\infty} \frac{1}{k^2} x^k$ has radius of convergence 1. The series converges absolutely if $x = \pm 1$, so the interval of convergence is $[-1, 1]$.

(iv) The series $\sum_{k=1}^{\infty} \frac{(-1)^k}{k} x^{2k}$ (this is a power series with odd coefficients 0) has radius of convergence 1. The series converges conditionally if $x = \pm 1$, so the interval of convergence is $[-1, 1]$. 

677
V.15.3.14. If a power series converges absolutely at one endpoint, it converges absolutely at the other endpoint too, since the absolute series \( \sum_{k=0}^{\infty} |a_k(x-x_0)^k| \) is the same at both endpoints. This is the only restriction on endpoint behavior for real infinite series.

V.15.3.15. Convergence behavior of a complex series on the boundary circle of the disk of convergence is much more subtle. The series can diverge everywhere on this circle (e.g. \( \sum_{k=0}^{\infty} (z-z_0)^k \)) or it can converge absolutely at all points of the circle (e.g. \( \sum_{k=1}^{\infty} \frac{1}{k^2} (z-z_0)^k \)), or it can converge conditionally on some subset \( S \) of the boundary circle. But \( S \) cannot be an arbitrary subset of the circle (it must be an \( F_\alpha \delta \) ), and not even all of these occur; see [Kör83]. The study of boundary behavior of analytic functions is a big topic in complex analysis, with important applications.

We will primarily be interested in using power series on open intervals. But we note one important result about the behavior of a power series near an endpoint or boundary point where it converges:

V.15.3.16. Theorem. [Abel’s Theorem] Let \( \sum_{k=0}^{\infty} a_k(x-x_0)^k \) be a power series. If \([a,b]\) is a closed bounded interval contained in the interval of convergence of the series, then the series converges uniformly on \([a,b]\). (In other words, the series converges u.c. on its entire interval of convergence.)

If \([a,b]\) is contained in the interior of the interval of convergence, the result follows immediately from V.15.3.6., so Abel’s Theorem only gives additional information in the case where \([a,b]\) contains one or both endpoints of the interval of convergence, and this is the case it is designed for.

V.15.3.17. Corollary. Let \( \sum_{k=0}^{\infty} a_k(x-x_0)^k \) be a power series, with interval of convergence \( I \). Then the function on \( I \) defined by the series is continuous on \( I \), including at any endpoints of \( I \).

Proof: Combine V.15.3.16. with ()..

V.15.3.18. Abel’s Theorem also has a complex version, which we do not state precisely. But even the real version is enough to almost immediately show that if \( \sum_{k=0}^{\infty} a_k(z-z_0)^k \) is a complex power series, and it converges at some \( z_1 \) on the boundary circle of the disk of convergence, then the value at \( z_1 \) is the radial limit of the value at \( z \) as \( z \to z_1 \) along the ray from \( z_0 \) to \( z_1 \). See Exercise ()

We now give the proof of Abel’s Theorem.

Proof: By applying V.15.3.6. to expand the interval if necessary, we may assume \( a \leq x_0 \leq b \). We prove that the series converges uniformly on \([x_0,b]\); the proof that it converges uniformly on \([a,x_0]\) is essentially identical.
Let $\epsilon > 0$. Since $\sum_{k=1}^{\infty} a_k (b - x_0)^k$ converges, by (1) there is an $N$ such that

$$\left| \sum_{k=n+1}^{n+m} a_k (b - x_0)^k \right| < \epsilon$$

whenever $n \geq N$ and $m \in \mathbb{N}$. If $x \in [x_0, b]$, set

$$c_k = \left( \frac{x - x_0}{b - x_0} \right)^k.$$ 

Then by Abel’s Partial Summation Formula (IV.2.7.5.) with $r = n$ we have, for any $n \geq N$ and $m \in \mathbb{N}$,

$$\sum_{k=n+1}^{n+m} a_k (x - x_0)^k = \sum_{k=n+1}^{n+m} c_k a_k (b - x_0)^k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_{k,n} + c_{n+m+1} s_{n+m,n}$$

where $s_{k,n} = \sum_{j=n+1}^{k} a_j (b - x_0)^j$. Since $|s_{k,n}| < \epsilon$ for all $k \geq n$, and $0 \leq c_{k+1} \leq c_k \leq 1$ for all $k$, we have

$$\left| \sum_{k=n+1}^{n+m} a_k (x - x_0)^k \right| \leq \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) |s_{k,n}| + c_{n+m+1} |s_{n+m,n}|$$

$$\leq \epsilon \left[ \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) + c_{n+m+1} \right] = (c_{n+1} - c_{n+m+1}) \epsilon \leq \epsilon$$

and so the series converges uniformly on $[x_0, b]$.

Term-by-Term Integration

Because of the u.c. convergence of a power series on its interval of convergence $I$, we can apply (1) to integrate a power series term-by-term on any closed bounded interval contained in $I$:

**V.15.3.19. Corollary.** Let $\sum_{k=0}^{\infty} c_k (x - x_0)^k$ be a power series with interval of convergence $I$, and $[a, b]$ a closed bounded subinterval of $I$. Then the function $f$ defined by the power series is integrable on $[a, b]$, and

$$\int_{a}^{b} f(x) \, dx = \int_{a}^{b} \left[ \sum_{k=0}^{\infty} c_k (x - x_0)^k \right] \, dx = \sum_{k=0}^{\infty} \left[ \int_{a}^{b} c_k (x - x_0)^k \, dx \right]$$

$$= \sum_{k=0}^{\infty} c_k \left[ (b - x_0)^{k+1} - (a - x_0)^{k+1} \right] = \sum_{k=0}^{\infty} \frac{c_k}{k+1} (b - x_0)^{k+1} - \sum_{k=0}^{\infty} \frac{c_k}{k+1} (a - x_0)^{k+1}$$. 

679
In particular, the series on the last line converge. This result is a consequence of V.15.3.6. if \([a, b]\) is contained in the open interval of convergence, but requires Abel’s Theorem if \(a\) or \(b\) is an endpoint of the interval of convergence. If \([a, b]\) is contained in the open interval of convergence, the series in the last expression will automatically converge absolutely by comparison to the power series evaluated at \(a\) and \(b\) (which converge absolutely by V.15.3.5.(i)). If \(a\) or \(b\) is an endpoint of the interval of convergence, the original series evaluated at \(a\) or \(b\) may not converge absolutely, but the next result shows that the integrated series still converge in this case.

V.15.3.20. **Proposition.** Let \(\sum_{k=0}^{\infty} a_k(x-x_0)^k\) be a power series. If this power series converges for \(x = x_1\), then the integrated series \(\sum_{k=0}^{\infty} \frac{a_k}{k+1} (x-x_0)^{k+1}\) also converges for \(x = x_1\).

**Proof:** Apply Abel’s Test (IV.2.7.7.) with \(c_k = \frac{x_1-x_0}{k+1}\).

Term-by-Term Differentiation

Term-by-term differentiation of a power series is a dicier matter. One of the beautiful and crucial properties of power series is that they can in fact be differentiated term-by-term, in contrast to more general series of functions.

The key fact making this possible is the next result. Note that if \(\sum_{k=0}^{\infty} a_k(x-x_0)^k\) is a power series, the series \(\sum_{k=1}^{\infty} k a_k(x-x_0)^{k-1}\) of term-by-term derivatives is also a power series.

V.15.3.21. **Proposition.** Let \(\sum_{k=0}^{\infty} a_k(x-x_0)^k\) be a power series, with radius of convergence \(R\). Then the radius of convergence of the series \(\sum_{k=1}^{\infty} k a_k(x-x_0)^{k-1}\) of term-by-term derivatives is also exactly \(R\).

**Proof:** First note that the series \(\sum_{k=1}^{\infty} k a_k(x-x_0)^{k-1}\) and \(\sum_{k=1}^{\infty} k a_k(x-x_0)^k\) converge for exactly the same \(x\)’s: both converge for \(x = x_0\), and for \(x \neq x_0\) each series is a constant scalar multiple of the other. Thus it suffices to show that the radius of convergence \(R’\) of the series \(\sum_{k=1}^{\infty} k a_k(x-x_0)^{k}\) is equal to \(R\).

Let \(L = \limsup_{k \to \infty} |a_k|^{1/k}\). Then \(R = \frac{1}{L}\) (with the usual conventions if \(L\) is 0 or \(\infty\). Similarly, \(R’ = \frac{1}{L’}\), where \(L’ = \limsup_{k \to \infty} |ka_k|^{1/k}\). So it suffices to show that \(L’ = L\). But \(\lim_{k \to \infty} k^{1/k} = 1\) (1), so by (1)

\[
L’ = \limsup_{k \to \infty} |ka_k|^{1/k} = \left(\lim_{k \to \infty} k^{1/k}\right) \left(\limsup_{k \to \infty} |a_k|^{1/k}\right) = 1 \cdot L = L.
\]
This result and proof are also valid for complex power series.

**V.15.3.22.** Note that this result does not say that the *interval* of convergence of the derived series is the same as the interval of convergence of the original series; in fact, endpoint behavior can be different. In examples V.15.3.13 (ii)-(iv), the interval of convergence of the derived series is strictly smaller than that of the original series; it cannot be larger (cf. V.15.3.20.).

We now apply () to get the main result:

**V.15.3.23.** **Theorem.** Let \( \sum_{k=0}^{\infty} a_k (x-x_0)^k \) be a power series, with radius of convergence \( R > 0 \). Then the function \( f \) defined by the power series is differentiable on \( (x_0 - R, x_0 + R) \), and

\[
  f'(x) = \sum_{k=1}^{\infty} k a_k (x-x_0)^{k-1}
\]

for all \( x \in (x_0 - R, x_0 + R) \).

**V.15.3.24.** But why stop there? The power series representing \( f' \) can also be differentiated term-by-term, and the theorem applies to show that \( f' \) is also differentiable on \( (x_0 - R, x_0 + R) \), i.e. \( f \) has a second derivative on this interval given by taking term-by-term derivatives twice. This argument can be iterated arbitrarily many times. Thus we obtain:

**V.15.3.25.** **Corollary.** Let \( \sum_{k=0}^{\infty} a_k (x-x_0)^k \) be a power series, with radius of convergence \( R > 0 \). Then the function \( f \) defined by the power series is \( C^\infty \) on \( (x_0 - R, x_0 + R) \), and

\[
  f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1) \cdots (k-n+1) a_k (x-x_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} a_k (x-x_0)^{k-n}
\]

for every \( n \) and every \( x \in (x_0 - R, x_0 + R) \).

Note that V.15.3.23. and hence V.15.3.26. apply also on any open subinterval of the interval of convergence. Thus we can rephrase:

**V.15.3.26.** **Corollary.** Let \( \sum_{k=0}^{\infty} a_k (x-x_0)^k \) be a power series which converges to a function \( f \) on an open interval \( I \). Then \( f \) is \( C^\infty \) on \( I \), and

\[
  f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1) \cdots (k-n+1) a_k (x-x_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} a_k (x-x_0)^{k-n} = \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} a_{k+n} (x-x_0)^{k+n}
\]

for every \( n \) and every \( x \in I \) (in particular, this series converges for all \( n \) and all \( x \in I \)).
Uniqueness of Power Series Representations

V.15.3.27. V.15.3.26. has another important consequence: if \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) is a power series with positive radius of convergence, then the coefficients \( a_k \) can be recovered from the function \( f \) defined by the power series. We have that \( f \) is infinitely differentiable in an interval around \( x_0 \), and \( f^{(n)}(x_0) \) is given by the \( n \)’th derived power series. In particular, \( f^{(n)}(x_0) \) is the constant term of the \( n \)’th derived power series, which is \( n! a_n \). Thus we obtain:

V.15.3.28. Corollary. Let \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) be a power series with positive radius of convergence, and \( f \) the function defined by the series on its open interval of convergence. Then

\[
a_n = \frac{f^{(n)}(x_0)}{n!}
\]

for each \( n \). In particular, the given power series is the only possible power series centered at \( x_0 \) that can represent \( f \) on an open interval containing \( x_0 \). Thus a function can be represented near \( x_0 \) by at most one power series centered at \( x_0 \).

V.15.3.29. Definition. Let \( f \) be \( C^\infty \) on an interval around \( x_0 \in \mathbb{R} \). The Taylor series for \( f \) around \( x_0 \) is the infinite series

\[
\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.
\]

Note that the partial sums of the Taylor series for \( f \) around \( x_0 \) are exactly the Taylor polynomials (V.10.1.6.) of \( f \) around \( x_0 \). The corollary can be rephrased:

V.15.3.30. Corollary. Let \( f \) be \( C^\infty \) on an interval around \( x_0 \in \mathbb{R} \). Then the only power series centered at \( x_0 \) which can converge to \( f \) in any open neighborhood of \( x_0 \) is the Taylor series for \( f \) around \( x_0 \).

V.15.3.31. Note two important things about the corollary:

(i) The result only applies to power series centered at \( x_0 \). It does not imply that \( f \) cannot be represented near \( x_0 \) by a different power series centered at some \( x_1 \neq x_0 \). In fact, if \( f \) can be represented in some interval by a power series centered at \( x_0 \), there will be such a power series centered at any \( x_1 \) sufficiently close to \( x_0 \).

(ii) The result says nothing about whether a function defined near \( x_0 \) can in fact be represented by a power series centered at \( x_0 \), only that if it can, the series is unique and is the Taylor series for \( f \) around \( x_0 \). An obvious necessary condition for \( f \) to be representable is that it be \( C^\infty \) in an interval around \( x_0 \), but this is far from sufficient. The question of which functions can be represented by power series is a substantial one which will be discussed in the next section.
V.16. Analytic Functions

Functions which can be locally represented by their Taylor series, called analytic functions, are particularly well-behaved. It turns out that most standard functions, and many of the functions naturally encountered in mathematics, are analytic.

V.16.1. Definitions and Basic Properties

V.16.1.1. Definition. (i) Let \( U \) be an open set in \( \mathbb{R} \). A function \( f : U \rightarrow \mathbb{R} \) is real analytic on \( U \) if for every \( x_0 \in U \), \( f \) is represented in some neighborhood of \( x_0 \) by a (real) power series centered at \( x_0 \).

(ii) Let \( V \) be an open set in \( \mathbb{C} \). A function \( f : V \rightarrow \mathbb{C} \) is complex analytic on \( V \) if for every \( z_0 \in V \), \( f \) is represented in some neighborhood of \( z_0 \) by a complex power series centered at \( z_0 \).

In mathematics it is most common to use the term analytic to mean “complex analytic.” However, in this book, except in Chapter X, it will normally mean “real analytic” or occasionally “real or complex analytic.”

V.16.1.2. Some remarks on the definition:

(i) The phrase “represented in some neighborhood of \( x_0 \) by a (real) power series centered at \( x_0 \)” means that there is some \( \epsilon > 0 \) and some (real) power series centered at \( x_0 \) which converges to \( f(x) \) for all \( x \in (x_0 - \epsilon, x_0 + \epsilon) \). The meaning in (ii) is analogous.

(ii) If \( f \) is analytic (in either the real or complex sense), then \( f \) is \( C^1 \) and the (unique) power series centered at \( x_0 \) representing \( f \) in a neighborhood of \( x_0 \) is the Taylor series of \( f \) around \( x_0 \) (V.15.3.29.).

(iii) Since the partial sums of the Taylor series for \( f \) around \( x_0 \) are the Taylor polynomials (V.10.1.6.) of \( f \) around \( x_0 \), \( f \) is analytic on \( U \) if and only if, for each \( x_0 \in U \), the Taylor polynomials for \( f \) around \( x_0 \) converge to \( f \) in a neighborhood of \( x_0 \), i.e. the remainder terms \( r_n \) converge to zero on a neighborhood of \( x_0 \) (cf. V.10.1.8.).

(iv) It is important to note that if \( f \) is analytic on \( U \) and \( x_0 \in U \), then the Taylor series of \( f \) around \( x_0 \) does not converge to \( f \) on all of \( U \) in general; it only converges to \( f \) in some open interval around \( x_0 \). Thus the property of being analytic on \( U \) is strictly a local property. If \( f : U \rightarrow \mathbb{R} \) is a function, and for every \( x \in U \) there is a neighborhood of \( x \) in \( U \) on which \( f \) is analytic, then \( f \) is analytic on \( U \). In particular, if \( f \) is analytic on every open subinterval of \( U \), then \( f \) is analytic on \( U \); thus for convenience we often just work with open intervals when dealing with analytic functions.

A similar comment applies in the complex analytic case. See ().

V.16.1.3. One of the important theorems of complex analysis is a strong form of the converse to V.16.1.2.(ii): if \( V \) is an open subset of \( \mathbb{C} \) and \( f : V \rightarrow \mathbb{C} \) is differentiable in the complex sense at every point of \( V \), then \( f \) is complex analytic on \( V \). There is nothing resembling an analog of this result for real analytic functions, and this is one of the most fundamental differences in nature between real analysis and complex analysis. More specifically:

If \( f \) is differentiable on \( I \), then \( f' \) need not even be continuous on \( I \) (), much less differentiable. In particular, a differentiable real function is “usually” not even \( C^1 \), much less \( C^\infty \).

Even most \( C^\infty \) functions are not real analytic; see section () for details.

V.16.1.4. Notation: The set of real analytic functions on an open set \( U \) is denoted \( C^\infty(U) \).
V.16.1.5. Examples. (i) Let \( f \) be a polynomial of degree \( n \). Then, for any \( x_0 \in \mathbb{R} \), by Taylor’s Theorem the \( n \)th order Taylor polynomial for \( f \) around \( x_0 \) is \( f \) itself, i.e. the Taylor series for \( f \) around \( x_0 \) (which has only finitely many nonzero terms) represents \( f \) on all of \( \mathbb{R} \). Thus \( f \) is analytic on \( \mathbb{R} \).

(ii) Let \( f(x) = \sin x \). Then \( f \) is \( C^\infty \) on \( \mathbb{R} \), and \(|f^{(k)}(x)| \leq 1\) for all \( k \) and \( x \). Fix \( x_0 \in \mathbb{R} \), and let \( p_n \) be the \( n \)th order Taylor polynomial for \( f \) around \( x_0 \), and \( r_n = f - p_n \). Then, for any \( x \in \mathbb{R} \), we have

\[
 r_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}
\]

for some \( c \) between \( x_0 \) and \( x \), and thus

\[
 |r_n(x)| \leq \frac{|x - x_0|^{n+1}}{(n+1)!} \to 0
\]

as \( n \to \infty \), so the Taylor series for \( f \) around \( x_0 \) converges to \( f(x) \) for all \( x \) and \( f \) is analytic on \( \mathbb{R} \). An identical argument shows that \( g(x) = \cos x \) is analytic on \( \mathbb{R} \).

(iii) Let \( f(x) = e^x \). Then \( f \) is \( C^\infty \) on \( \mathbb{R} \), and \( f^{(k)}(x) = e^x \) for all \( k \) and \( x \). Fix \( x_0 \in \mathbb{R} \) and \( s > 0 \), and let \( m = e^{x_0 + s} \). Then \( |f^{(k)}(x)| < m \) if \( |x - x_0| < s \). Let \( p_n \) be the \( n \)th order Taylor polynomial for \( f \) around \( x_0 \), and \( r_n = f - p_n \). Then, for any \( x \) with \( |x - x_0| < s \), we have

\[
 r_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}
\]

for some \( c \) between \( x_0 \) and \( x \), and thus

\[
 |r_n(x)| \leq \frac{m|x - x_0|^{n+1}}{(n+1)!} \to 0
\]

as \( n \to \infty \), so the Taylor series for \( f \) around \( x_0 \) converges to \( f(x) \) for all \( x \), \( |x - x_0| < s \). Since \( s \) is arbitrary, the Taylor series converges to \( f(x) \) for all \( x \in \mathbb{R} \). Thus \( f \) is analytic on \( \mathbb{R} \).

Analyticity of Functions Defined by Power Series

The definition of an analytic function is hard to check directly: to show that \( f \) is analytic on \( U \), one must show for every \( x_0 \in U \) that the Taylor series for \( f \) around \( x_0 \) converges to \( f \) in an interval around \( x_0 \), an (uncountably) infinite number of conditions since there is a different Taylor series at each point. The following theorem is one of the most important facts about analytic functions, and saves an enormous amount of work in checking that specific functions are analytic since only one \( x_0 \) must be considered.

V.16.1.6. Theorem. Let \( \sum_{k=0}^{\infty} a_k(x - x_0)^k \) be a power series with radius of convergence \( R > 0 \), and let \( f \) be the function defined on \( I = (x_0 - R, x_0 + R) \) defined by the power series. Then \( f \) is analytic on \( I \).

In fact, the following more precise result holds:
\textbf{V.16.1.7. Theorem.} Let \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) be a power series with radius of convergence \( R > 0 \), and let \( f \) be the function defined on \( I = (x_0 - R, x_0 + R) \) defined by the power series. Let \( x_1 \in I \), and set \( S = R - |x_1 - x_0| \). Then the Taylor series for \( f \) around \( x_1 \) has radius of convergence at least \( S \) and converges to \( f \) on \( (x_1 - S, x_1 + S) \).

See Figure ( ). Note that \( f \) is \( C^\infty \) on \( I \), so the Taylor series for \( f \) around \( x_1 \) is defined.

To prove these theorems, we first need a lemma which will also be used later.

\textbf{V.16.1.8. Lemma.} Let \( \sum_{k=0}^{\infty} a_k (x - x_0)^k \) be a power series with radius of convergence \( R > 0 \), and let \( f \) be the function defined on \( I = (x_0 - R, x_0 + R) \) defined by the power series. Fix \( r, 0 < r < R \), and \( s, 0 < s < R - r \). Then there is a constant \( m \) such that
\[
|f^{(k)}(x)| \leq \frac{m \cdot k!}{s^k}
\]
for all \( k \geq 0 \) and all \( x \) with \( |x - x_0| \leq r \).

\textbf{Proof:} First note that the power series \( \sum_{j=0}^{\infty} x^j \) converges to \( f(x) = \frac{1}{1-x} \) on \((-1, 1)\), and thus the \( k \)'th derived series
\[
\sum_{j=k}^{\infty} j(j-1)\cdots(j-k)x^{j-k}
\]
converges on \((-1, 1)\) to \( f^{(k)}(x) = \frac{k!}{(1-x)^{k+1}} \) for any \( k \).

Let \( r \) and \( s \) be as in the statement. Since \( r + s < R \), by () there is an \( M \) such that \( |a_k| \leq \frac{M}{(r+s)^k} \) for all \( k \). If \( |x - x_0| \leq r \), then we have, for any \( k \),
\[
f^{(k)}(x) = \sum_{j=k}^{\infty} j(j-1)\cdots(j-k)a_j(x-x_0)^{j-k}
\]
\[
|f^{(k)}(x)| \leq \sum_{j=k}^{\infty} j(j-1)\cdots(j-k) \frac{M}{(r+s)^j} \frac{1}{r^j} = \frac{M}{(r+s)^k} \sum_{j=k}^{\infty} j(j-1)\cdots(j-k) \left( \frac{r}{r+s} \right)^{j-k}
\]
\[
= \frac{M}{(r+s)^k} \frac{k!}{(1-\frac{r}{r+s})^{k+1}} = \frac{M}{(r+s)^k} \frac{k!(r+s)^{k+1}}{s^{k+1}} = \frac{m \cdot k!}{s^k}
\]
where \( m = \frac{M(r+s)}{s} \).

We now give the proof of Theorem V.16.1.6. and most of Theorem V.16.1.7.
Proof: Let $\sum_{k=0}^{\infty} b_k(x-x_1)^k$ be the Taylor series for $f$ around $x_1$. We will show that for every $r$, $|x_1-x_0| < r < R$, and $s$, $0 < s < R-r$, this Taylor series has radius of convergence at least $s$ and converges to $f(x)$ whenever $|x-x_1| < \min(s, r-|x_1-x_0|)$. This will show that this Taylor series has radius of convergence at least $S$ and that it converges to $f$ in a neighborhood of $x_1$; it follows that $f$ is analytic on $I$.

Fix such an $r$ and $s$. By Lemma V.16.1.8, there is an $m$ such that $|f^{(k)}(x)| \leq \frac{m}{s^k}$ for all $k \geq 0$ whenever $|x-x_0| \leq r$. In particular, we have

$$|b_k| = \frac{|f^{(k)}(x_1)|}{k!} \leq \frac{m}{s^k}$$

for all $k$, so the radius of convergence of $\sum_{k=0}^{\infty} b_k(x-x_1)^k$ is

$$\limsup_{k \to \infty} \frac{1}{|b_k|^{1/k}} \geq \frac{1}{\limsup_{k \to \infty} \left(\frac{m}{s^k}\right)^{1/k}} = s.$$

Since $s$ can be made arbitrarily close to $S$ by choosing $r$ close enough to $|x_1-x_0|$, the radius of convergence is at least $S$.

Let $g$ be the function on $J = (x_1 - S, x_1 + S)$ defined by $\sum_{k=0}^{\infty} b_k(x-x_1)^k$, and let

$$g_n(x) = \sum_{k=0}^{n} b_k(x-x_1)^k$$

for $n \in \mathbb{N}$. Then $g_n \to g$ on $J$. Set $\phi_n = f - g_n$. Then $\phi_n$ is $C^\infty$ on $I$, $\phi_n^{(k)}(x_1) = 0$ for $0 \leq k \leq n$, and $\phi_n^{(k)} = f^{(k)}$ on $I$ for $k > n$ since $g_n$ is a polynomial of degree $\leq n$.

Set $t = \min(s, r-|x_1-x_0|)$, and let $x$ satisfy $|x-x_1| < t$. By the Lagrange Remainder Formula (), we have

$$\phi_n(x) = \frac{\phi_n^{(n+1)}(c_n)}{(n+1)!}(x-x_1)^{n+1} = \frac{f^{(n+1)}(c_n)}{(n+1)!}(x-x_1)^{n+1}$$

for some $c_n$ between $x_1$ and $x$; since $|c_n-x_0| \leq r$ we have

$$|\phi_n(x)| \leq \frac{m}{s^{n+1}}|x-x_1|^{n+1} = m \left(\frac{|x-x_1|}{s}\right)^{n+1}.$$ 

Since $|x-x_1| < s$, $\phi_n(x) \to 0$, $g_n(x) \to f(x)$, and thus $f = g$ on $(x_1-t, x_1+t)$.

\[ \Diamond \]

V.16.1.9. We have now proved V.16.1.6., and all that remains to finish the proof of V.16.1.7. is to show that $f = g$ on all of $J$. We have that $f$ and $g$ are both analytic on $J$ by V.16.1.6., and $f = g$ on a neighborhood of $x_1$. The conclusion then follows by an application of V.16.2.2..

V.16.1.10. An analogous theorem holds for complex analytic functions: if $f$ is defined by a complex power series around $z_0$ on an open disk $D$ of radius $R > 0$, $z_1 \in D$, and $S = R - |z_1 - z_0|$, then the Taylor series for $f$ around $z_1$ has radius of convergence at least $S$ and converges to $f$ on the open disk of radius $S$ around $z_1$. This can be proved by an identical proof to the one above, but there is a simpler proof using other theorems of complex analysis.
Growth of Derivatives of Analytic Functions

We can obtain a closely related characterization of analytic functions: they are exactly the $C^\infty$ functions whose derivatives “do not grow too rapidly” locally. Precisely:

**V.16.1.11. Theorem.** Let $f$ be a real-valued function on an open set $U$ in $\mathbb{R}$. Then $f$ is analytic on $U$ if and only if $f$ is $C^\infty$ on $U$ and, for each $x_0 \in U$, there is an open interval $J$ containing $x_0$ and natural numbers $m$ and $n$ such that

$$|f^{(k)}(x)| \leq m \cdot k! \cdot n^k$$

for all $k \geq 0$ and all $x \in J$.

**Proof:** Suppose $f$ is analytic on $U$ and $x_0 \in U$. Then the Taylor series for $f$ around $x_0$ has positive radius of convergence $R$, and there is an $r$, $0 < r < R$, such that this Taylor series represents $f$ on $J = (x_0 - r, x_0 + r)$. Fix $s$ with $0 < S < R - r$. Then by Lemma V.16.1.8., there is an $m$ such that, for any $k \geq 0$ and $x \in J$, $|f^{(k)}(x)| \leq \frac{m \cdot k!}{s^k}$. We may assume $m \in \mathbb{N}$, and we may take $n$ any natural number $\geq \frac{1}{s}$.

Conversely, suppose $f$ is $C^\infty$ on $U$ and the growth condition is satisfied. Fix $x_0 \in U$, and let $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ be the Taylor series of $f$ around $x_0$. Let $J$, $m$, and $n$ be as in the statement, and set $s = \frac{1}{n}$. Then there is an $r$, $0 < r < s$, such that $(x_0 - r, x_0 + r) \subseteq J$. For each $k$, let $p_k$ be the $k$'th Taylor polynomial for $f$ around $x_0$, and $r_k = f - p_k$ the $k$'th remainder. If $|x - x_0| < r$, then we have

$$r_k(x) = \frac{f^{(k+1)}(c_k)}{(k+1)!} (x - x_0)^{k+1}$$

for some $c_k$ between $x_0$ and $x$; since $c_k \in (x_0 - r, x_0 + r)$ we have

$$|r_k(x)| \leq m \cdot n^{k+1} r^{k+1} = m \left( \frac{r}{s} \right)^{k+1} \to 0$$

as $k \to \infty$; thus the Taylor series for $f$ around $x_0$ converges to $f$ on $(x_0 - r, x_0 + r)$ and $f$ is analytic on $U$. \(\blacksquare\)

**V.16.2. Rigidity of Analytic Functions**

One of the remarkable properties of an analytic function $f$ on an open interval $I$ is that the values of $f$ on a small part of $I$ (e.g. a small subinterval of $I$) completely determine the values of $f$ on all of $I$. In particular, if an analytic function $f$ on an interval $I$ is identically 0 on a small subinterval of $I$, then $f$ is identically 0 on all of $I$, and in fact the zeroes of a nonzero analytic function are (relatively) isolated in $I$.

**V.16.2.1. Theorem.** Let $f$ and $g$ be analytic on an open interval $I$, and $(x_n)$ a sequence of distinct points in $I$ converging to $x_0 \in I$. If $f(x_n) = g(x_n)$ for all $n$, then $f(x) = g(x)$ for all $x \in I$. 

687
**V.16.2.2.** Corollary. Let $f$ and $g$ be analytic on an open interval $I$, and $J$ a (nondegenerate) subinterval of $I$. If $f(x) = g(x)$ for all $x \in J$, then $f(x) = g(x)$ for all $x \in I$.

**V.16.2.3.** Corollary. Let $f$ be analytic and not identically 0 on an open interval $I$. Then the zeros of $f$ in $I$ are isolated in $I$: if $x_0 \in I$ and $f(x_0) = 0$, then there is a deleted neighborhood $U$ of $x_0$ in $I$ such that $f(x) \neq 0$ for all $x \in U$.

**V.16.2.4.** Caution: Two analytic functions on an interval $I$ can agree at a sequence of points in $I$ without being identical, so long as the sequence has no cluster points in $I$. For example, $f(x) = \sin x$ takes the same values as the constant function 0 at $n\pi$ for all $n$. The sequence where $f$ and $g$ agree can even have a limit which is not in $I$: $f(x) = \sin \left( \frac{x}{2} \right)$ is analytic on $(0, +\infty)$ and agrees with the constant function 0 at a sequence of points $\left( \frac{1}{n\pi} \right)$ converging to 0.

To prove Theorem V.16.2.1., we need a lemma which is of independent interest:

**V.16.2.5.** Lemma. Let $f$ be $C^\infty$ on an open interval $I$, and $(x_n)$ a sequence of distinct points in $I$ converging to $x_0 \in I$. If $f(x_n) = 0$ for all $n$, then $f^{(k)}(x_0) = 0$ for all $k$.

**Proof:** We have $f(x_0) = \lim_{n \to \infty} f(x_n) = 0$ by continuity, and

$$f'(x_0) = \lim_{n \to \infty} \frac{f(x_n) - f(x_0)}{x_n - x_0} = 0.$$  

Suppose $f(x_0) = f'(x_0) = \cdots = f^{(k-1)}(x_0) = 0$ and $f^{(k)}(x_0) \neq 0$, and set $a = \frac{f^{(k)}(x_0)}{k!}$. Then by Taylor’s Theorem () we have

$$f(x) = a(x - x_0)^k + o((x - x_0)^k)$$

for all $x \in I$. Thus there is a $\delta > 0$ such that

$$|f(x) - a(x - x_0)^k| < \frac{|a|}{2}|x - x_0|^k$$

whenever $0 < |x - x_0| < \delta$. But there is an $n$ with $0 < |x_n - x_0| < \delta$, a contradiction. Thus there is no such $k$.  

See V.10.6.11. for an alternate argument. We now give the proof of Theorem V.16.2.1.

**Proof:** First suppose $g$ is identically 0, i.e. $f(x_n) = 0$ for all $n$. Then by V.16.2.5. the Taylor series for $f$ around $x_0$ is identically 0, i.e. $f$ is identically 0 in a neighborhood of $x_0$ in $I$. Let $E$ be the set of all $x \in I$ such that $f$ is identically 0 in a neighborhood of $x$ in $I$. Then $E$ is nonempty since $x_0 \in E$, and it is obviously open in $I$. If $c \in I$ is in the closure of $E$, another application of V.16.2.5. shows that $f^{(k)}(c) = 0$ for all $k$, and hence $f$ is identically 0 in a neighborhood of $c$ in $I$, i.e. $c \in E$. Thus $E$ is relatively closed in $I$. Since $I$ is connected, $E = I$.

Now consider the case of general $g$. We have that $f - g$ is analytic on $I$ and vanishes at each $x_n$, so $f - g$ is identically 0 on $I$.  

688
V.16.2.6. There are (nonanalytic) $C^\infty$ functions on $I$ satisfying the hypotheses of Lemma V.16.2.5., which are not identically zero (V.17.2.2.). Thus $C^\infty$ functions do not generally have the same rigidity properties as analytic functions. In fact, the difference is dramatic; see V.17.2.9.

There are many important consequences of V.16.2.1., for example:

V.16.2.7. Corollary. Let $f$ be an analytic function on an open interval $I$, and $x_0 \in I$. Let $J$ be the interval of convergence of the Taylor series for $f$ around $x_0$. Then the Taylor series for $f$ around $x_0$ represents $f$ on all of $I \cap J$, and $f$ extends uniquely to an analytic function on $I \cup J^o$.

Proof: Let $g$ be the function defined on $J$ by the Taylor series for $f$ around $x_0$. Then $g$ is analytic on $J^o$. The functions $f$ and $g$ are analytic on $I \cap J^o$, and agree in a neighborhood of $x_0$ by V.16.1.7., so they agree on all of $I \cap J^o$ by V.16.2.2.. If $x$ is an endpoint of $J$ contained in $I$, then $x$ is in the closure of $I \cap J^0$ and both $f$ and $g$ are continuous at $x$ (V.15.3.17.), so $f(x) = g(x)$.

For the last statement, define $h(x) = f(x)$ if $x \in I$ and $h(x) = g(x)$ if $x \in J^o$. Then $h$ is well defined by the first part of the proof, and is obviously analytic on $I \cup J^o$. Since $I \cup J^o$ is an interval, and any extension of $f$ to $I \cup J^o$ agrees with $f$ on $I$, uniqueness follows from V.16.2.2.

V.16.2.8. Example. Note that the interval $I$ cannot be replaced by an open set $U$ in this result: let $U = \mathbb{R} \setminus \{0\}$, and let $f(x) = 0$ for $x < 0$ and $f(x) = 1$ for $x > 0$. Then $f$ is analytic on $U$, but V.16.2.7. fails for any $x_0 \in U$.

V.16.2.9. The exact complex analog of V.16.2.1. holds, with essentially identical proof: if $f$ and $g$ are complex analytic on a connected open set $V$, $(z_n)$ is a sequence in $V$ converging to $z_0 \in V$, and $f(z_n) = g(z_n)$ for all $n$, then $f(z) = g(z)$ for all $z \in V$. V.16.2.7. also extends to the complex case, with “interval” replaced by “disk.” (There are topological obstructions to extending this result to connected open sets in $\mathbb{C}$, even simply connected open sets, but it does extend to convex open sets.)

This is needed to prove the following important fact:

V.16.2.10. Theorem. Let $f$ be a real analytic function on an open set $U$ in $\mathbb{R}$. Then there is an open set $V$ in $\mathbb{C}$ with $V \cap \mathbb{R} = U$, and a complex analytic function $g$ on $V$, such that $g|U = f$.

Proof: For each $x_0 \in U$, the Taylor series for $f$ around $x_0$, regarded as a complex power series, defines a complex analytic function $g_{x_0}$ on an open disk in $\mathbb{C}$ centered at $x_0$ whose intersection with the real axis is contained in $U$, and on which $g_{x_0}$ agrees with $f$. It is only necessary to show that these functions agree on the overlaps of their domains; then $g$ can be defined on the union of these disks. But any two of these functions agree on the part of the overlap of their domain disks contained in $U$ (which is an open interval in $\mathbb{R}$), and the overlap is connected, so they agree on the entire overlap.
V.16.3. Combinations of Analytic Functions

In this section, we show that all usual ways of manufacturing new functions out of old preserve analyticity. Specifically:

(i) Linear combinations of analytic functions are analytic.

(ii) Derivatives and antiderivatives of analytic functions are analytic (V.15.3.20., V.15.3.23.).

(iii) Products of analytic functions are analytic.

(iv) Quotients of analytic functions are analytic if the denominator is nonzero.

(v) Compositions of analytic functions are analytic.

(vi) Inverse functions of analytic functions with nonzero derivatives are analytic.

These results, among other things, provide a large supply of standard analytic functions. The direct proofs range from very easy (i) to easy consequences of previous results ((ii)–(iii)) to moderately involved ((iv)–(v)) to quite difficult (vi). The “easy” way to prove (iii)–(vi) is to use V.16.2.10. and theorems from complex analysis, but we will describe direct proofs not using complex analysis.

Linear Combinations of Analytic Functions

V.16.3.1. Proposition. Let \( f \) and \( g \) be analytic on an open set \( U \), and \( \alpha \) a scalar. Then \( f + g \) and \( \alpha f \) are analytic on \( U \).

Proof: The conclusions are immediate consequences of IV.2.1.16. and IV.2.1.18. respectively (plus the observation that linear combinations of power series centered at the same place are also power series). ☐

Unordered Power Series

V.16.3.2. Definition. An unordered power series is an unordered sum (IV.2.9.2.)

\[
\sum_{j \in J} a_j X^{n_j}
\]

where \( J \) is an index set, \( a_j \) is a scalar, \( X \) is an indeterminate symbol, and \( n_j \in \mathbb{N} \cup \{0\} \) for each \( j \). (Thus we may have unordered real power series or unordered complex power series).

The differences between an unordered power series and an ordinary formal power series are:

(i) The order of the terms is not specified.

(ii) There may be many terms, even infinitely many, with the same power of \( X \).
An unordered power series \( \sum_{j \in J} a_j x^{n_j} \) with coefficients in \( \mathbb{F} (= \mathbb{R} \) or \( \mathbb{C} \)) defines a function by setting
\[
f(x) = \sum_{j \in J} a_j x^{n_j}
\]
whose domain is the set of \( x \) for which the unordered sum converges (i.e. converges absolutely by IV.2.9.9.). The domain may be empty (we adhere to the convention that \( x^0 = 1 \) even if \( x = 0 \)). We may more generally fix an \( x_0 \) and consider the function
\[
f(x) = \sum_{j \in J} a_j (x - x_0)^{n_j}.
\]

While the index set \( J \) can in principle be uncountable, if it is not countable, or at least if \( \{ j : a_j \neq 0 \} \) is uncountable (terms with \( a_j = 0 \) can be effectively ignored in the sum), the unordered sum cannot converge for any \( x \neq 0 \) by IV.2.9.6.. Thus in examining such series expansions of functions there is no essential loss of generality in considering only countable index sets. So from now on we will restrict to the case where \( J \) is countable.

We can formally rearrange an unordered power series into an ordinary power series. To do this, for each \( k \in \mathbb{N} \cup \{0\} \) set \( J_k = \{ j : n_j = k \} \) (i.e. the set of indices for which the power of \( X \) in the term is \( k \)). We can then form the ordinary infinite series
\[
\sum_{k=0}^{\infty} \left( \sum_{j \in J_k} a_j \right) x^k
\]
provided the unordered sum \( \sum_{j \in J_k} a_j \) converges for each \( k \). This power series is called the standard revision of the unordered power series.

**Theorem.** Let
\[
\sum_{j \in J} a_j X^{n_j}
\]
be an unordered power series, and \( x_0 \) a scalar. Suppose the unordered sum
\[
\sum_{j \in J} a_j (x_1 - x_0)^{n_j}
\]
converges for some \( x_1 \neq x_0 \). Then
(i) The unordered sum
\[
f(x) = \sum_{j \in J} a_j (x - x_0)^{n_j}
\]
converges for all \( x, |x - x_0| \leq |x_1 - x_0| \).
(ii) The unordered sum
\[ \sum_{j \in J_k} a_j \]
converges for each \( k \).

(iii) The standard revision
\[
\sum_{k=0}^{\infty} \left[ \left( \sum_{j \in J_k} a_j \right) (x-x_0)^k \right]
\]
of the unordered power series converges absolutely to \( f(x) \) for all \( x, |x-x_0| \leq |x_1-x_0| \). (In particular, this power series has radius of convergence at least \( |x_1-x_0| \).)

(If \( x_1 = x_0 \), (i) and (iii) still hold, and (ii) holds for \( k = 0 \), but the result has little content in this case.)

\textbf{Proof:} By IV.2.9.9.,
\[
\sum_{j \in J} |a_j(x_1-x_0)^{n_j}| 
\]
converges, and thus for \( |x-x_0| \leq |x_1-x_0| \) the unordered sum
\[
\sum_{j \in J} |a_j(x-x_0)^{n_j}|
\]
converges by comparison, hence \( \sum_{j \in J} a_j(x-x_0)^{n_j} \) converges. (Technically we should convert the unordered sums to ordinary infinite series by taking the terms in some order; the order is irrelevant.) Also, for any \( k \), the unordered sum
\[
\sum_{j \in J_k} |a_j(x_1-x_0)^{n_j}| = |(x_1-x_0)^k| \sum_{j \in J_k} |a_j|
\]
converges since it is a subsum of a convergent nonnegative unordered sum. Thus, since \( x_1 \neq x_0 \),
\[
\sum_{j \in J_k} a_j
\]
converges for each \( k \). Finally, for any \( x \) with \( |x-x_0| \leq |x_1-x_0| \) the standard revision
\[
\sum_{k=0}^{\infty} \left[ \left( \sum_{j \in J_k} a_j \right) (x-x_0)^k \right]
\]
(with the terms from each \( J_k \) ordered in some arbitrary way) is a drastic revision (IV.2.10.14.) of the unordered sum
\[
f(x) = \sum_{j \in J} a_j(x-x_0)^{n_j}
\]
(with terms ordered in some way). Since the unordered sum converges (absolutely), the standard revision converges to the same sum by IV.2.10.16..
V.16.3.7. Conversely, any ordinary power series can be regarded as an unordered power series. The unordered power series converges exactly whenever the original power series converges absolutely, i.e. either on the interval (or disk) of convergence or the open interval (or disk) of convergence of the power series.

V.16.3.8. Corollary. Let
\[ \sum_{j \in J} a_j X^j \]
be an unordered power series. If the unordered series
\[ \sum_{j \in J} a_j x^j \]
converges for some \( x \neq 0 \), then for any \( x_0 \) the function
\[ f(x) = \sum_{j \in J} a_j(x - x_0)^j \]
is analytic on an interval (or disk) around \( x_0 \). The Taylor series for \( f \) around \( x_0 \) is the standard revision of the unordered power series.

V.16.3.9. This result can be used to justify many formal manipulations with power series. The rough general principle is:

If \( f \) and \( g \) are functions defined by infinite summations, and the summations can be formally manipulated into the same sum, and the summations converge absolutely in an appropriate sense, then the functions are equal.

We give several specific examples below (V.16.3.12.).

Products of Analytic Functions

V.16.3.10. Theorem. Let \( f \) and \( g \) be analytic functions on an open set \( U \). Then \( fg \) is analytic on \( U \).

Proof: Fix \( x_0 \in U \), and let
\[ f(x) = \sum_{k=0}^{\infty} a_k(x - x_0)^k \quad \text{and} \quad g(x) = \sum_{k=0}^{\infty} b_k(x - x_0)^k \]
be the Taylor series for \( f \) and \( g \), converging to \( f \) and \( g \) on \( \{ x : |x - x_0| < R \} \subseteq U \). If \( |x - x_0| < R \), then the above series converge absolutely, so the product series
\[ \sum_{n=0}^{\infty} \left( \sum_{k=0}^{n} a_k b_{n-k} \right)(x - x_0)^k \]
converges (absolutely) to \( f(x)g(x) \) by IV.2.11.8.
V.16.3.11. It can be readily checked by repeated applications of the Product Rule that if \( f \) and \( g \) are infinitely differentiable at \( x_0 \), then so is \( fg \), and that the Taylor series for \( fg \) around \( x_0 \) is the product of the Taylor series for \( f \) and \( g \) around \( x_0 \) (i.e. the Taylor series for \( fg \) around \( x_0 \) is the standard revision of the unordered power series obtained by multiplying the two Taylor series termwise). This observation is not needed for the proof of V.16.3.10., but it and V.16.3.10. give additional motivation for why the definition (IV.2.11.5.) of the product of two infinite series is the “right” one.

Compositions of Analytic Functions

V.16.3.12. Theorem. Let \( f \) be analytic on an open set \( U \), \( g \) analytic on an open set \( V \), and \( f(U) \subseteq V \). Then \( g \circ f \) is analytic on \( U \).

Proof: We may proceed formally as follows, ignoring convergence questions for the moment. Let \( x_0 \in U \), and \( y_0 = f(x_0) \in V \). Let

\[
f(x) = \sum_{k=0}^{\infty} a_k(x-x_0)^k \quad \text{and} \quad g(y) = \sum_{k=0}^{\infty} b_k(y-y_0)^k
\]

be the Taylor series for \( f \) around \( x_0 \) and for \( g \) around \( y_0 \). Then we can expand

\[
g(f(x)) = \sum_{k=0}^{\infty} b_k(f(x)-y_0)^k = \sum_{k=0}^{\infty} b_k \left( \sum_{j=1}^{\infty} a_j(x-x_0)^j \right)^k.
\]

If we expand the \( k \)'th power series

\[
\left( \sum_{j=1}^{\infty} a_j(x-x_0)^j \right)^k
\]

into a product series

\[
\sum_{n=0}^{\infty} c_{kn}(x-x_0)^n
\]

for each \( k \) (cf. IV.2.11.11.), we have a drastic revision

\[
\sum_{k=0}^{\infty} b_k \left( \sum_{n=1}^{\infty} c_{kn}(x-x_0)^n \right)
\]

of an unordered power series

\[
\sum_{(n,k)} b_k c_{kn}(x-x_0)^n.
\]

The drastic revision V.1 converges to \( g(f(x)) \) if \( x \) is sufficiently close to \( x_0 \); so if the unordered series V.2 converges for some \( x \neq x_0 \), we can conclude from V.16.3.8. that its sum (which is \( g \circ f \) by IV.2.10.16.) is analytic on a neighborhood of \( x_0 \). Since \( x_0 \in U \) is arbitrary, the result follows.

The two power series

\[
\sum_{k=1}^{\infty} |a_k|x^k \quad \text{and} \quad \sum_{k=0}^{\infty} |b_k|y^k
\]

694
have a positive radius of convergence, hence define analytic functions $F$ and $G$ near 0. Thus, if $t$ is a sufficiently small positive number, we can write

$$G(F(t)) = \sum_{k=0}^{\infty} |b_k| F(t)^k = \sum_{k=0}^{\infty} |b_k| \left( \sum_{j=1}^{\infty} |a_j| t^j \right)^k$$  \hspace{1cm} (V.3)

where all the series converge (absolutely). Expanding the $k$’th power series

$$\left( \sum_{j=1}^{\infty} |a_j| t^j \right)^k = \sum_{n=0}^{\infty} d_{kn} t^n$$

for each $k$ (note that $d_{kn} \geq 0$ for all $k$ and $n$) and converting V.3 to an unordered series

$$\sum_{(n,k)} |b_k| d_{kn} t^n$$

we obtain a convergent nonnegative unordered series which converges to $G(F(t))$. If we compare the series

$$\left( \sum_{j=1}^{\infty} a_j (x-x_0)^j \right)^k = \sum_{n=0}^{\infty} c_{kn} (x-x_0)^n$$

$$\left( \sum_{j=1}^{\infty} |a_j| t^j \right)^k = \sum_{n=0}^{\infty} d_{kn} t^n$$

we obtain easily from the triangle inequality that $|c_{kn}| \leq d_{kn}$ for all $k$ and $n$. Thus the unordered series

$$\sum_{(n,k)} b_k c_{kn} t^n$$

converges absolutely by comparison, which was what needed to be shown to complete the proof (taking $x = x_0 + t$).

\vspace{1cm}

**V.16.3.13. Remarks.** Here are some remarks about the theorem and its proof.

(i) In most of the proof we work with $\sum_{k=1}^{\infty} a_k (x-x_0)^k$, not $\sum_{k=0}^{\infty} a_k (x-x_0)^k$. This is because we necessarily have $a_0 = f(x_0) = y_0$ and we are working with $f(x) - y_0$.

(ii) The $k$’th power sum

$$\left( \sum_{j=1}^{\infty} a_j (x-x_0)^j \right)^k$$
when converted into the product series
\[
\sum_{n=0}^{\infty} c_{kn}(x-x_0)^n
\]
has summation starting at \(k\), i.e. \(c_{kn} = 0\) for \(n < k\). Thus the standard revision of the unordered series
\[
\sum_{(n,k)} b_k c_{kn}(x-x_0)^n
\]
representing \(g \circ f\) is an ordinary revision, not a drastic revision:
\[
\sum_{n=0}^{\infty} \left[ \left( \sum_{k=0}^{n} b_k c_{kn} \right) (x-x_0)^n \right] \quad \text{(V.4)}
\]
which must be the Taylor series for \(g \circ f\) around \(x_0\).

(iii) If \(f\) is \(C^\infty\) on \(U\), \(g\) is \(C^\infty\) on \(V\), and \(f(U) \subseteq V\), then \(g \circ f\) is \(C^\infty\) on \(U\) by the Chain Rule. If \(x_0 \in U\) and the Taylor series for \(g \circ f\) around \(x_0\) is calculated using Faà di Bruno’s formula (V.3.7.8.), the same power series as in V.4 is obtained, as can be verified by unpleasant direct calculation. In fact, the derivation of this Taylor series in the proof of V.16.3.12. is a good way to derive (but not prove!) Faà di Bruno’s formula.

Quotients of Analytic Functions

The next result is an immediate corollary of V.16.3.12., since \(g(x) = \frac{1}{x}\) is analytic on \(\{x : x \neq 0\}\):

V.16.3.14. Corollary. Let \(f\) be analytic on an open set \(U\), with \(f(x) \neq 0\) for all \(x \in U\). Then \(\frac{1}{f}\) is analytic on \(U\).

Combining this with V.16.3.10., we obtain:

V.16.3.15. Theorem. Let \(f\) and \(g\) be analytic on an open set \(U\), with \(g(x) \neq 0\) for all \(x \in U\). Then \(\frac{f}{g}\) is analytic on \(U\).

Bernstein’s Theorem

We prove a theorem obtained by S. Bernstein in 1912, which is at least interesting if not too useful. (Bernstein only stated (ii), but the more precise statements of (i) and (iii) follow easily from his proof.)

V.16.3.16. Theorem. Let \(I = (a, b)\) be an open interval in \(\mathbb{R}\) (where \(a \in \mathbb{R} \cup \{-\infty\}\) and \(b \in \mathbb{R} \cup \{+\infty\}\)), and let \(f\) be a \(C^\infty\) function on \(I\). If \(f^{(k)}(x) \geq 0\) for all \(k \geq 0\) and all \(x \in I\), then:

(i) If \(x_0 \in I\) and \(R = b - x_0\) \((R = +\infty\) if \(b = +\infty\)), then the Taylor series for \(f\) around \(x_0\) converges to \(f(x)\) for all \(x \in I\) with \(|x - x_0| < R\). (In particular, the radius of convergence of this Taylor series is at least \(R\).)

696
(ii) $f$ is analytic on $I$.

(iii) If $\ell = b - a$ ($\ell = +\infty$ if $b = +\infty$ or $a = -\infty$), then $f$ extends uniquely to an analytic function on $(a - \ell, b)$ (all of $\mathbb{R}$ if $b = +\infty$).

Note, however, that the extension in (iii) does not necessarily have $f^{(k)}(x) \geq 0$ for all $k$ and $x$: consider $f(x) = x^3$ on $(0, \infty)$.

**Proof:** (i): Fix $x_0 \in I$, and fix $x > x_0$. Let $c \in I$ with $c > x$. Let $p_n$ be the $n$th Taylor polynomial for $f$ around $x_0$. We have $p_n(x) \geq 0$ since all the terms in the Taylor series are nonnegative. By the Integral Form of the Remainder (V.10.4.2.), we have

$$R_n(x) = f(x) - p_n(x) = \frac{1}{n!} \int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n \, dt.$$ 

Using two changes of variable ($u = t - x_0$ and $s = \frac{u}{x-x_0}$), we obtain

$$R_n(x) = \frac{1}{n!} \int_{0}^{x-x_0} f^{(n+1)}(u+x_0)(x-x_0-u)^n \, du = \frac{(x-x_0)^{n+1}}{n!} \int_{0}^{1} f^{(n+1)}((x-x_0)s-x_0)(1-s)^n \, ds$$

and, since $f^{(n+1)}$ is nonnegative and nondecreasing ($f^{(n+2)}$ is also nonnegative), we have

$$0 \leq R_n(x) \leq \left(\frac{x-x_0}{c-x_0}\right)^{n+1} \left(\frac{c-x_0}{x-x_0}\right)^{n+1} \int_{0}^{1} f^{(n+1)}((c-x_0)s+x_0)(1-s)^n \, ds = \left(\frac{x-x_0}{c-x_0}\right)^{n+1} R_n(c).$$

We have $R_n(c) = f(c) - p_n(c) \leq f(c)$ since $p_n(c) \geq 0$, so we have

$$0 \leq R_n(x) \leq \left(\frac{x-x_0}{c-x_0}\right)^{n+1} f(c)$$

and thus $R_n(x) \to 0$ as $n \to \infty$, i.e. the Taylor series for $f$ around $x_0$ converges to $f(x)$.

Now suppose $x \in I$, $x < x_0$, $|x-x_0| < R$. Set $y = x_0 + |x-x_0|$, i.e. $y > x_0$ and $|y-x_0| = |x-x_0|$. By comparison of integrals we have $|R_n(x)| \leq R_n(y)$ since $f^{(n+1)}$ is nondecreasing, so $R_n(x) \to 0$ as $n \to \infty$.

(ii): Immediate from (i).

(iii): If $x_0 > a$ and $R = b - x_0$, then by (i) the Taylor series for $f$ around $x_0$ defines an analytic function on $(x_0-R, x_0+R)$ which agrees with $f$ on $I$. Let $x_0 \to a+$. (All the extensions defined this way must coincide on the intersections of their domains by V.16.2.1., which also gives the uniqueness of the extension.)

**Formal Manipulation of Power Series**

**V.16.3.17.** The results of this section can be used to justify the formal manipulation of power series. For example, suppose we are given analytic functions $f$ and $h$ taking the same value at 0 and we want to find an analytic function $g$ satisfying $g(0) = 0$ and $h = f \circ g$ (we stick to analytic functions around 0 for notational simplicity; more generally, we may take power series centered at other points than 0). Temporarily ignoring questions of convergence, we can proceed formally: write

$$f(x) = \sum_{k=0}^{\infty} a_k x^k$$

$$h(x) = \sum_{k=0}^{\infty} b_k x^k$$

$$g(x) = \sum_{k=0}^{\infty} c_k x^k$$

such that $c_k = \frac{1}{k!} \int_{0}^{x} f^{(k+1)}(t)(x-t)^k \, dt$ for all $k$. Then $g(0) = 0$, $h(x) = f(g(x))$, and $g(x) = \frac{1}{k!} \int_{0}^{x} f^{(k+1)}(t)(x-t)^k \, dt$ for all $k$. By the Integral Form of the Remainder (V.10.4.2.), we have

$$g(x) = \sum_{k=0}^{\infty} \frac{1}{k!} \int_{0}^{x} f^{(k+1)}(t)(x-t)^k \, dt = \sum_{k=0}^{\infty} \frac{1}{k!} \int_{0}^{x} a_k (x-t)^k \, dt = \sum_{k=0}^{\infty} \frac{x^k}{k!} a_k = f(x).$$

Thus $g(x) = f(x)$ for all $x$ in the domain of $f$.
\[ h(x) = \sum_{k=0}^{\infty} c_k x^k \]

with \( a_0 = c_0 \). The unknown function

\[ g(x) = \sum_{k=1}^{\infty} b_k x^k \]

can be “found” by manipulation of the power series:

\[ f(g(x)) = a_0 + a_1(b_1 x + b_2 x^2 + \cdots) + a_2(b_1 x + b_2 x^2 + \cdots)^2 + \cdots \]

We can formally multiply out the last expression to get

\[
(b_1 x + b_2 x^2 + \cdots)^2 = b_1^2 x^2 + 2b_1 b_2 x^3 + \cdots
\]

and similarly for higher powers. Collecting together like powers of \( x \), we get

\[ f(g(x)) = a_0 + a_1 b_1 x + (a_1 b_2 + a_2 b_1^2) x^2 + \cdots \]

and for this to be equal to \( h(x) \), we get the equations

\[
c_0 = a_0 \\
c_1 = a_1 b_1 \\
c_2 = a_1 b_2 + a_2 b_1^2
\]

etc., which must be solved for the \( b_k \). It is easy to see exactly when this can be done recursively, and what the (necessarily unique) solution is when it exists.

To justify the formal manipulations, the theorems of this section can be used. In fact, using (), we can show:

**V.16.3.18. Theorem.** If the system is solvable for the \( b_k \), and if the power series

\[ \sum_{k=1}^{\infty} b_k x^k \]

has a positive radius of convergence, then the function \( g \) defined by the series satisfies \( h = f \circ g \) in a neighborhood of 0. Conversely, if \( g \) is an analytic function in a neighborhood of 0 satisfying \( g(0) = 0 \) and \( h = f \circ g \) in a neighborhood of 0, and

\[ g(x) = \sum_{k=1}^{\infty} b_k x^k \]

is its Taylor series around 0, then the \( b_k \) satisfy the equations. In particular, \( g \) is the unique analytic function with these properties.

**V.16.3.19.** It is a more delicate matter to find or estimate how large a neighborhood of 0 the formula is valid in. In fact, it is valid in the largest open disk around 0 mapped by \( g \) into the open disk of convergence of the Taylor series for \( f \) around 0.
V.17. Smooth Non-Analytic Functions

A striking and important phenomenon is that a smooth ($C^1$) function on an interval $I$ need not be analytic on $I$; in fact, it can be “completely nonanalytic” on $I$. In this section we describe various ways in which a smooth function can be nonanalytic.

This is a topic in which the fairly extensive work over a rather long period of time does not seem to be widely known in the mathematical community (e.g. the author only rather recently became aware of it). It is remarkable how often the basic results and examples (with slight variations) have been rediscovered, reproved, and republished. Only variations of Cauchy’s original example, and Borel’s Theorem, have apparently become standard knowledge among mathematicians. For an extensive survey of the work done in this area, see http://mathforum.org/kb/message.jspa?messageID=387148.

V.17.1. Analyticity at a Point

The definition of an analytic function (V.16.1.1.) only defines analyticity on an open interval in $\mathbb{R}$ (or an open set in $\mathbb{C}$). It is useful to make the following definition of “analyticity at a point.”

V.17.1.1. Definition. Let $I$ be an interval in $\mathbb{R}$, and $x_0 \in I$. A real-valued $C^1$ function $f$ on $I$ is analytic at $x_0$ if the Taylor series of $f$ around $x_0$ converges to $f$ in some relatively open subinterval of $I$ containing $x_0$ (an open interval if $x_0$ is an interior point of $I$). A point where $f$ is not analytic is called a singular point of $f$.

The next result, an immediate corollary of V.16.1.6., is fundamental:

V.17.1.2. Proposition. Let $f \in C^\infty(I)$. The set

$$A_f = \{x_0 \in I : f \text{ is analytic at } x_0\}$$

is relatively open in $I$.

Thus the set of singular points of $f$ is a relatively closed subset of $I$. (It is not yet clear that this set can be nonempty.)

Types of Singular Points

V.17.1.3. If $f$ is $C^\infty$ on $I$ and $x_0 \in I$, there are two potential ways $f$ can fail to be analytic at $x_0$:

(i) The Taylor series of $f$ around $x_0$ has radius of convergence zero. Such a point is sometimes called a divergence point of $f$. In [Zah47], such points are called points of type (P), or (P)-points (for Pringsheim).

(ii) The Taylor series of $f$ around $x_0$ has positive radius of convergence, but does not represent the function on any interval around $x_0$. Such a point is sometimes called a false convergence point of $f$. The name from [Zah47] is point of type (C), or (C)-point (for Cauchy).

V.17.1.4. Thus, if $f$ is $C^\infty$ on $I$, $I$ decomposes as the disjoint union of the set $A_f$ of points where $f$ is analytic, the set $P_f$ of the (P)-points of $f$, and the set $C_f$ of the (C)-points of $f$. $A_f$ is (relatively) open by V.17.1.2., so $P_f \cup C_f$ is (relatively) closed.
V.17.2. (C)-Points and Smooth Partitions

V.17.2.1. The first example of a (C)-point, the best-known example in this subject, was given by Cauchy in 1823.

V.17.2.2. Example. Let
\[
 f(x) = \begin{cases} 
 e^{-1/x^2} & \text{if } x \neq 0 \\
 0 & \text{if } x = 0 
\end{cases}.
\]

Then \( f \) is clearly analytic on \( \mathbb{R} \setminus \{0\} \). On \( \mathbb{R} \setminus \{0\} \), it can easily be shown by induction on \( k \) that \( f^{(k)}(x) \) is a finite sum of terms of the form \( e^{-1/x^2} \) for \( c \in \mathbb{R}, \ m \in \mathbb{N} \cup \{0\} \). Thus, since \( f(0) = 0 \), we have by induction that \( f^{(k)}(0) \) exists and is 0 for all \( k \).

So \( f \) is infinitely differentiable also at 0, thus \( C^\infty \) on \( \mathbb{R} \). The Taylor series at 0 (Maclaurin series) for \( f \) has all terms zero, and thus has infinite radius of convergence. But it does not converge to \( f(x) \) for any \( x \neq 0 \) since \( f(x) > 0 \) for \( x \neq 0 \). Thus 0 is a (C)-point of \( f \) and \( f \) is not analytic at 0. See Figure (). See X.5.2.17.(ii) for an “explanation” of why the function is not analytic at 0.

V.17.2.3. This example is usually presented with an exponent of \(-\frac{1}{x^2}\), but this is just somewhat of a matter of taste. Identical arguments show that each of the following functions has the same properties:
\[
 f_2(x) = \begin{cases} 
 e^{-1/|x|} & \text{if } x \neq 0 \\
 0 & \text{if } x = 0 
\end{cases},
\]
\[
 f_3(x) = \begin{cases} 
 e^{-1/x} & \text{if } x > 0 \\
 0 & \text{if } x \leq 0 
\end{cases},
\]
\[
 f_4(x) = \begin{cases} 
 e^{-1/x^2} & \text{if } x > 0 \\
 0 & \text{if } x \leq 0 
\end{cases},
\]
i.e. each of these is \( C^\infty \) on \( \mathbb{R} \), analytic on \( \mathbb{R} \setminus \{0\} \), and has a (C)-point at 0 where its Taylor series is identically 0. Any of these can be called “Cauchy’s example.” The last two have the interesting property that they are identically zero on the negative real axis. Note that all these examples take values in \([0, 1)\), and are in particular bounded.

Smooth Bump Functions

We can combine translates of these functions to make smooth “bump functions” which are identically zero outside a specified interval:

V.17.2.4. Example. Let \( a, b \in \mathbb{R}, \ a < b \). Define
\[
 g_{a,b}(x) = f_3(x-a)f_3(b-x) = \begin{cases} 
 e^{-\frac{b-a}{x-(b-a)}} & \text{if } a < x < b \\
 0 & \text{if } x \leq a \text{ or } x \geq b 
\end{cases}.
\]

Then \( g \) is \( C^\infty \) on \( \mathbb{R} \), analytic on \( \mathbb{R} \setminus \{a, b\} \), has (C)-points at \( a \) and \( b \) where its Taylor series is identically zero, and is positive on \( (a, b) \) and identically zero on \( \mathbb{R} \setminus (a, b) \). See Figure ().
Smooth Heaviside Functions

V.17.2.5. The function

\[ \phi(x) = \int_a^x g_{a,b}(t) \, dt \]

is \( C^\infty \) on \( \mathbb{R} \), analytic on \( \mathbb{R} \setminus \{a, b\} \), identically 0 for \( x \leq a \), takes a constant positive value for \( x \geq b \), and is strictly increasing on \([a, b]\). A suitable scalar multiple \( h_{a,b} \) has the same properties and is identically 1 for \( x \geq b \). This can be regarded as a smooth version of the Heaviside function \( \chi_{[b, \infty)} \). See Figure ( ).

Smooth Mesa Functions

V.17.2.6. Let \( a < b < c < d \), and let \( f \) be the function

\[ f(x) = h_{a,b}(x)(1 - h_{c,d}(x)) . \]

Then \( f \) is \( C^\infty \) on \( \mathbb{R} \), analytic on \( \mathbb{R} \setminus \{a, b, c, d\} \), identically 0 on \( (-\infty, a] \) and on \([d, +\infty) \), identically 1 on \([b, c] \), strictly increasing on \([a, b] \), and strictly decreasing on \([c, d] \). See Figure ( ).

Smooth Partitions of Unity

Recall the definition of partition of unity ( ). We will show that on \( \mathbb{R} \), smooth partitions of unity exist:

V.17.2.7. Theorem. Let \( \mathcal{U} \) be an open cover of \( \mathbb{R} \). Then there is a countable partition of unity \( \{f_n\} \) subordinate to \( \mathcal{U} \) consisting of \( C^\infty \) functions of compact support.

Proof: This is an easy consequence of ( ). There is a collection \( \mathcal{V} = \{(a_n, b_n) : n \in \mathbb{Z}\} \) of bounded open intervals refining \( \mathcal{U} \) such that \( a_n < b_{n-1} < a_{n+1} < b_n \) for all \( n \). Take \( f_n \) to be the smooth mesa function

\[ f_n = h_{a_n,b_{n-1}}(1 - h_{a_{n+1},b_n}) \]

defined in V.17.2.6. and V.17.2.5.. \( \blacksquare \)

Smooth Functions With Prescribed Zero Sets

V.17.2.8. If \( f \) is a continuous function, then its zero set

\[ Z_f = \{ x : f(x) = 0 \} \]

is always closed. We now show that every closed subset of \( \mathbb{R} \) is the zero set of a smooth function. This is in stark contrast to the situation for analytic functions, where the zero set is either discrete or all of \( \mathbb{R} \) ( ). For a simple example of a \( C^\infty \) function with a nonisolated zero, see Exercise V.17.5.1..
V.17.2.9. **Theorem.** Let $E$ be a closed subset of $\mathbb{R}$. Then there is a nonnegative bounded $C^\infty$ function $\theta$ on $\mathbb{R}$ for which $Z_\theta = E$ and $\theta$ is analytic on $\mathbb{R} \setminus E$.

**Proof:** There is a simple and appealing argument that almost works to prove this result, which can be adapted to give a real proof. The open set $U = \mathbb{R} \setminus E$ is a countable disjoint union of open intervals $I_n$. For each $n$, we can take a smooth bump function $f_n$ which is nonzero exactly on $I_n$ (if $I_n$ is unbounded, take a translation of the $f_3$ of V.17.2.3. or its reflection instead of a bump function). Define $\theta$ to be $f_n$ on $I_n$ and $0$ on $E$. Alternately, define $\theta = \sum_{n=1}^\infty f_n$; the series converges for each $x$ since at most one term is nonzero.

However, there is a problem with this approach. Since the set of endpoints of the $I_n$ is not closed and discrete in general (e.g. consider the case where $E$ is the Cantor set), the sum of the series will not even be continuous in general unless the heights of the bumps are controlled.

We can do this as follows. The complement of $E$ consists of at most a countable number of bounded open intervals $I_n = (a_n, b_n)$, plus possibly one or two unbounded open intervals we will deal with separately at the end of the argument. We may assume there are infinitely many $I_n$ since otherwise a finite sum of bump functions works for $f$. Let $g = g_{0,1}$ be the smooth bump function of V.17.2.4., normalized so that $\max(g(x)) = 1$, and for each $k \geq 0$ let $M_k = \max |g^{(k)}|$. Let $g_n$ be $g$ translated to $I_n$, i.e.

$$g_n(x) = g \left( \frac{x - a_n}{b_n - a_n} \right)$$

for $x \in \mathbb{R}$. Set $f_n = c_ng_n$, where

$$c_n = \min(1, (b_n - a_n)^n) \frac{1}{2^n \max_{0 \leq k \leq n} M_k}.$$ 

Then $f_n$ is $C^\infty$ on $\mathbb{R}$, and $|f_n^{(k)}(x)| \leq 2^{-n}$ for $0 \leq k \leq n$ and all $x \in \mathbb{R}$. Thus

$$\sum_{n=1}^\infty f_n^{(k)}$$

converges uniformly on $\mathbb{R}$ for each $k$. If $\theta = \sum_{n=1}^\infty f_n$, then by V.15.1.2. (the simpler version suffices), $\theta$ is $C^\infty$ on $\mathbb{R}$ and $\theta^{(k)} = \sum_{n=1}^\infty f_n^{(k)}$ for all $k$. The function $\theta$ takes values in $[0, 1]$, and is clearly zero precisely on $E$ and any unbounded intervals in $E^c$. There will be such an unbounded interval if and only if $E$ has a minimum element $a$ and/or a maximal element $b$. If there is such a $b$, adding to $f$ a translated version of Cauchy’s example $f_3$ makes $\theta$ positive on $(b, +\infty)$. A similar correction can be made on $(-\infty, a)$ if $a$ exists. Since, for each $x \in E^c$, $\theta$ is equal to $f_n$ for some $n$ (or a translated Cauchy function) on a neighborhood of $x$, $\theta$ is analytic at $x$. \(\diamondsuit\)

The Smooth Urysohn Lemma

With somewhat more work along the same lines, we can get a smooth version of Urysohn’s Lemma (for $\mathbb{R}$):

V.17.2.10. **Theorem.** [SMOOTH URYSOHN LEMMA FOR $\mathbb{R}$] Let $A$ be a closed subset of $\mathbb{R}$ and $U$ an open set in $\mathbb{R}$ containing $A$. Then there is a $C^\infty$ function $\phi$ from $\mathbb{R}$ to $[0, 1]$ such that $\phi \equiv 1$ on $A$ and $\phi \equiv 0$ on $\mathbb{R} \setminus U$. We may choose $\phi$ so that $\phi^{-1}(\{1\}) = A$ and $\phi^{-1}(\{0\}) = \mathbb{R} \setminus U$. 

702
Alternatively, if $E$ and $F$ are disjoint closed subsets of $\mathbb{R}$, there is a smooth function $\phi : \mathbb{R} \to [0, 1]$ such that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$ (set $A = E$ and $U = \mathbb{R} \setminus F$).

**Proof:** We prove the alternative form. Let $E$ and $F$ be disjoint closed subsets of $\mathbb{R}$. Then the open set $V = \mathbb{R} \setminus (E \cup F)$ is a countable disjoint union of open intervals, at most two of which are unbounded. Let $\{(a_i, b_i) : i \in I\}$ be the set of bounded intervals in this union such that $a_i \in E$ and $b_i \in F$, and let $\{(c_j, d_j) : j \in J\}$ be the set of bounded intervals in the collection with $c_j \in F$ and $d_j \in E$. Then $I$ and $J$ are countable. Between any two $a_i$ there must be at least one of the intervals $(c_j, d_j)$, and similarly for the other combinations.

Let $L_E = \{a_i : i \in I\}$, $L_F = \{c_j : j \in J\}$, $R_E = \{d_j : j \in J\}$, $R_F = \{b_i : i \in I\}$. $L_E$ is countable; we claim it is closed and discrete in $\mathbb{R}$. If not, there is a sequence $(a_{i_n})$ in $L_E$ of distinct points converging to some $x \in \mathbb{R}$, and passing to a subsequence we may assume that the sequence is strictly monotone ($\uparrow$). Since each $a_{i_n} \in E$, $x \in E$. If $(a_{i_n})$ is strictly monotone, we also have $b_{i_n} \to x$ since $b_{i_n}$ is between $a_{i_n-1}$ and $a_{i_n+1}$, so $x \in F$, contradicting that $E$ and $F$ are disjoint. Thus $L_E$ is closed and discrete. By similar arguments, $L_F$, $R_E$, $R_F$ are all closed and discrete in $\mathbb{R}$. Thus $L_E \cup L_F \cup R_E \cup R_F$ is also closed and discrete, i.e. there are only finitely many points in any bounded interval.

Thus the endpoints of the intervals can be reindexed $\{(a_k, b_k), (c_k, d_k) : k \in \mathbb{Z}\}$, where

$$a_k < b_k < c_k < d_k < a_{k+1}$$

for all $k$. If $V$ contains an unbounded interval on either side, the indexing on that side will be finite. Define $\phi$ separately on each of the subintervals of $\mathbb{R}$ so defined. On $[b_k, a_{k+1}]$, let $\phi$ be the smooth Heaviside function $h_{c_k,d_k}$ (V.17.2.5), and on $[d_{k-1}, c_k]$ let $\phi$ be $1 - h_{a_k, b_k}$. Note that these definitions coincide on the overlaps of the intervals. If there is an unbounded interval to $+\infty$ in $V$, let $\phi$ be $h_{c_k,d_k}$ on $[b_k, +\infty)$ if $(c_k, d_k)$ is the last bounded interval, and $1 - h_{a_k, b_k}$ on $[d_{k-1}, +\infty)$ if $(a_k, b_k)$ is the last bounded interval; define $\phi$ analogously if $V$ has an unbounded interval to $-\infty$. Since the endpoints of the intervals form a closed discrete set, the $\phi$ so defined is a $C^\infty$ function from $\mathbb{R}$ to $[0, 1]$ which is identically 1 on $E$ and 0 on $F$.

This $\phi$, however, does not necessarily have the property that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$. We can obtain this property by a trick. Let $\theta$ be a $C^\infty$ function from $\mathbb{R}$ to $[0, 1]$ such that $\theta = 0$ precisely on $E \cup F$ (V.17.2.9.). Then $\psi_1 = \phi(1 - \theta)$ is a $C^\infty$ function from $\mathbb{R}$ to $[0, 1]$ which is 0 on $F$ and which is 1 precisely on $E$. Similarly, $\psi_2 = 1 - (1 - \phi)(1 - \theta)$ is a $C^\infty$ function from $\mathbb{R}$ to $[0, 1]$ which is 1 on $E$ and 0 precisely on $F$. Then $\psi = \frac{1}{2}(\psi_1 + \psi_2)$ is a $C^\infty$ function from $\mathbb{R}$ to $[0, 1]$ which is 1 precisely on $E$ and 0 precisely on $F$. \hfill \Box

### Multivariable Versions

**V.17.2.11.** There are versions of all these examples and results for real-valued functions on $\mathbb{R}^n$ ( ). The multivariable versions tend to be more complicated, sometimes considerably so, but also more useful.

### V.17.3. (P)-Points and Borel’s Theorem

The first example of a (P)-point was apparently obtained by CELLÈRIER in the early 1860’s, only published posthumously in 1890 (Exercise ( )). The first published example was by DU BOIS-REYMOND in 1876 (Exercise ( )).

Some specific examples of (P)-points will be given in ( ). In this subsection we will be primarily concerned with a striking theorem of E. BOREL from 1895, which provides (P)-points in profusion. The theorem can be stated in two equivalent ways.
**V.17.3.1. Theorem.** Let \((a_k)_{k \geq 0}\) be any sequence of real numbers. Then there is a \(C^\infty\) function on \(\mathbb{R}\) whose Taylor series around 0 (Maclaurin series) is

\[
\sum_{k=0}^{\infty} a_k x^k.
\]

**V.17.3.2. Corollary.** Let \((b_k)_{k \geq 0}\) be any sequence of real numbers. Then there is a \(C^\infty\) function \(f\) on \(\mathbb{R}\) such that \(f^{(k)}(0) = b_k\) for all \(k\).

The two statements are equivalent by setting \(b_k = k!a_k\).

We first give a simple proof from [Mir56] (the same proof works in the multidimensional case ()). Then we will give another argument giving an example with additional analyticity properties.

**Proof:** Fix a smooth mesa function \(g\) such that \(g\) is supported on \((-1, 1)\), and for each \(n\) let \(M_n\) be a bound on \(g^{(n)}\). Note that \(g^{(n)}(0) = 0\) for all \(n > 0\) since \(g\) is constant on an interval around 0. For \(k \geq 0\) define

\[
f_k(x) = a_k x^k g(c_k x)
\]

where \((a_k)\) is the given sequence and \((c_k)\) is a sequence of natural numbers to be chosen later. Then \(f_k\) is \(C^\infty\) on \(\mathbb{R}\), and by the higher product rule () and the chain rule for the \(g\) factors,

\[
f_k^{(n)}(x) = \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1)a_k c_k^m x^{k-n+m} g^{(m)}(c_k x)
\]

for all \(n\) (the product \(k(k-1)\cdots(k-n+m+1)\) is 1 if \(m = n\)). In particular, we have

\[
f_k^{(k)}(0) = \sum_{m=0}^{k} \binom{k}{m} k(k-1)\cdots(m+1)a_k c_k^m 0^m g^{(m)}(0) = k!a_k = b_k
\]

and \(f_k^{(n)}(0) = 0\) if \(n \neq k\) (for \(n > k\), note that \(f_k^{(n)}\) is identically 0 in a neighborhood of 0). Since \(f_k(x) = 0\) for all \(x\) with \(|x| \geq \frac{1}{c_k}\), we have

\[
|f_k^{(n)}(x)| \leq \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1)a_k c_k^m c_k^{k+n-m} M_m
\]

\[
\leq \frac{1}{c_k} \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1)a_k M_m
\]

for \(0 \leq n < k\) and all \(x\), and thus by choosing \(c_k\) sufficiently large we obtain that

\[
|f_k^{(n)}(x)| \leq 2^{-k}
\]
for $0 \leq n < k$ and all $x$. Thus the series
\[
\sum_{k=0}^{\infty} f_k^{(n)}
\]
converges uniformly on $\mathbb{R}$ for every $n \geq 0$ by the $M$-test. So the function
\[
f = \sum_{k=0}^{\infty} f_k
\]
is $C^\infty$ on $\mathbb{R}$ and
\[
f^{(n)} = \sum_{k=0}^{\infty} f_k^{(n)}
\]
for every $n$ by (). In particular, $f^{(n)}(0) = b_n$ for all $n$. 

\[V.17.3.3.\] To obtain an example of a smooth function with a (P)-point, apply the theorem with a sequence $a_k \to +\infty$ rapidly, e.g. $a_k = k!$.

\[V.17.3.4.\] Note that the example constructed in the above proof is analytic except on a countable set; in fact, if $x_0 \neq 0$ and $U$ is an open neighborhood of $x_0$ bounded away from 0, then all but finitely many of the terms in the series are identically zero on $U$. Thus this theorem is strictly a single-point result.

We can in fact improve this result to make $f$ analytic on all of $\mathbb{R} \setminus \{0\}$. This result is due to Besicovitch [Bes24].

\[V.17.3.5.\] **Theorem.** Let $(b_k)$ be an arbitrary sequence of real numbers. Then there is a real-valued function $f$, $C^\infty$ on $[0, +\infty)$ and analytic on $(0, +\infty)$, with $f^{(k)}(0) = b_k$ for all $k$ (these are one-sided derivatives).

**Proof:** [KP02b, Lemma 2.2.4] Define
\[
f_0(x) = \frac{1}{\sqrt{c_0 + x}}
\]
and for each $k \geq 1$, define
\[
f_k(x) = \int_0^x \int_0^{t_{k-1}} \cdots \int_0^{t_1} \frac{1}{\sqrt{c_k + t}} \, dt_1 \cdots dt_{k-1}
\]
where the $c_k$ are positive real numbers which will be chosen later, i.e. $f_k$ is the unique function such that $f_k^{(k)}(x) = \frac{1}{\sqrt{c_k + x}}$ and $f_k^{(n)}(0) = 0$ for $0 \leq n < k$. For each $k$, $f_k$ is $C^\infty$ on $(-c_k, +\infty)$.

Fix $N > 1$. If $0 \leq x \leq N$, we have, for $0 \leq n < k$,
\[
0 \leq f_k^{(n)}(x) \leq \int_0^N \int_0^{t_{k-n-1}} \cdots \int_0^{t_1} \frac{1}{\sqrt{t}} \, dt_1 \cdots dt_{k-n-1} = \frac{N^{k-n-1/2}}{\frac{1}{2} \cdots \frac{2(k-n)-1}{2}} < \frac{2N^{k-n}}{(k-n-1)!}
\]
and so the series
\[
\sum_{k=0}^{\infty} \sigma_k f_k^{(n)}
\]
converges uniformly on \([0, N]\) for every \(n \geq 0\) and any choice of \(\sigma_k = -1, 0, 1\) by the \(M\)-test. Thus, for any such choice of \(\sigma_k\), we have that
\[
f = \sum_{k=0}^{\infty} \sigma_k f_k
\]
is \(C^\infty\) on \([0, \infty)\) and for each \(n\),
\[
f^{(n)} = \sum_{k=0}^{\infty} \sigma_k f_k^{(n)}
\]
by (\(\cdot\)). We also have, for \(\epsilon > 0\),
\[
|f^{(n)}(x)| \leq \sum_{k=0}^{n} \frac{2}{(c_k + \epsilon)^{k+1/2}} + \sum_{k=n+1}^{\infty} \frac{2N^{k-n}}{(k-n-1)!} \leq 2(n+1) \left( \frac{1}{\epsilon} \right)^{n+1} + 2e^N
\]
for \(\epsilon \leq x \leq N\), so by (\(\cdot\)) \(f\) is analytic on \((0, \infty)\).

The above results hold for any choice of \(c_k\) and \(\sigma_k\). We need only show that these can be chosen so that \(f^{(k)}(0) = b_k\) for all \(k\). Since \(f_k^{(n)}(0) = 0\) if \(0 \leq n < k\), we must solve the equations
\[
b_0 = \sigma_0 \frac{1}{c_0^{1/2}}
\]
\[
b_1 = -\frac{1}{2} \sigma_0 \frac{1}{c_0^{1/2}} + \frac{1}{c_1^{1/2}}
\]
\[
b_2 = \frac{1}{2} \cdot 2 \cdot 3 \sigma_0 \frac{1}{c_0^{1/2}} - \frac{1}{2} \sigma_1 \frac{1}{c_1^{1/2}} + \frac{1}{2} \sigma_2 \frac{1}{c_2^{1/2}}
\]
\[
\ldots
\]
We may solve these successively. If \(b_0 = 0\), set \(\sigma_0 = 0\) and \(c_0 = 1\); otherwise, set \(\sigma_0 = \frac{b_0}{|b_0|}\) and \(c_0 = |b_0|^{-2}\). Then set \(\sigma_1\) equal to \(1, -1, 0\) depending whether
\[
b_1 + \frac{1}{2} \sigma_0 \frac{1}{c_0^{3/2}}
\]
is positive, negative, or 0, and then \(c_1\) is determined from the previous choices. This process can clearly be continued indefinitely.

As a corollary, we get a two-sided version which improves Borel’s result:
**V.17.3.6.** Corollary. Let \((b_k)\) be an arbitrary sequence of real numbers. Then there is a real-valued function \(f\), \(C^\infty\) on \(\mathbb{R}\) and analytic on \(\mathbb{R} \setminus \{0\}\), with \(f^{(k)}(0) = b_k\) for all \(k\).

**Proof:** Let \(f\) be as in the Theorem for the sequence \((b_k)\), and \(g\) as in the theorem for the sequence \((-1)^k b_k\). Extend \(f\) to \(\mathbb{R}\) by setting \(f(x) = g(-x)\) for \(x < 0\). The extended \(f\) is \(C^\infty\) at \(0\). 

We now obtain the main result of [Bes24], that the derivatives of a \(C^\infty\) function can be independently specified at two points with the function analytic in between:

**V.17.3.7.** Theorem. Let \((a_k)\) and \((b_k)\) be arbitrary sequences of real numbers. Then there is a real-valued function \(f\), \(C^\infty\) on \(\mathbb{R}\) and analytic on \(\mathbb{R} \setminus \{0, 1\}\), such that \(f^{(k)}(0) = a_k\) and \(f^{(k)}(1) = b_k\) for all \(k\).

**Proof:** Let \(h(x) = h_{0,1}\) be the smooth Heaviside function from V.17.2.5. Let \(g_0\) and \(g_1\) be functions as in V.17.3.6. for the sequences \((a_k)\) and \((b_k)\) respectively; and set

\[
\begin{align*}
f_0(x) &= g_0(x)(1 - h(x)) \\
f_1(x) &= g_1(x - 1)h(x)
\end{align*}
\]

and \(f = f_0 + f_1\). Then \(f_0, f_1, f\) are \(C^\infty\) on \(\mathbb{R}\) and analytic on \(\mathbb{R} \setminus \{0, 1\}\). We have that \(f_0^{(k)}(0) = a_k\) and \(f_0^{(k)}(1) = 0\) for all \(k\), and \(f_1^{(k)}(0) = 0\) and \(f_1^{(k)}(1) = b_k\) for all \(k\). Thus \(f\) is the desired function.

**V.17.4.** Everywhere Nonanalytic Functions

There are smooth functions on an interval \(I\) which are not analytic anywhere. By V.17.1.2., these are precisely the functions \(f\) for which the set \(C_f \cup P_f = I\). Since \(C_f \cup P_f\) is closed in \(I\), to find examples it suffices to find \(f\)'s for which \(C_f\) or \(P_f\) is dense in \(I\). (It turns out that for any such example, \(P_f\) must be dense in \(I\) (V.17.4.10.), but this is often not easy to explicitly verify.)

**V.17.4.1.** There are two standard methods for constructing smooth functions which are not analytic anywhere, both involving infinite series. The first is to take a smooth function \(g\) which is not analytic at one, or a few, points, such as Cauchy’s example, and form a sum of the form

\[
f(x) = \sum_{n=1}^{\infty} c_n g(x - r_n)
\]

where the \(r_n\) form a dense sequence (e.g. an enumeration of \(\mathbb{Q}\)) and \(c_n \to 0\) rapidly, or some variation of this involving translation and scaling of \(g\). This method is called the method of condensation of singularities.

The other method is to consider an infinite sum of the form

\[
\sum_{n=1}^{\infty} c_n \cos(\lambda_n x)
\]

(sin can be used instead of cos) where \(c_n \to 0\) rapidly, \(\lambda_n \to \infty\) rapidly, and there are appropriate relations between the \(c_n\) and \(\lambda_n\). The first example of a function with a dense set of (P)-points was a function of this second kind due to Cellèrier in the 1860’s (Exercise ()). We give a nice example of this sort in V.17.4.3. (Caution: there is an example of this kind on Wikipedia which is not correct.)
so this series therefore has infinite radius of convergence, but represents a nonanalytic function. Thus the graph of $F$ can be extended to be $C^\infty$ on all of $\mathbb{R}$ by setting $F(-t) = F(t)$ or $F(-t) = -F(t)$ (there seems to be no convention). Thus the graph of $F^{(n)}$ is similar to the graph of $F$, stretched vertically by a factor of $2^{2n}$ and compressed horizontally by a factor of $2^n$. See Figure (1) for the graph of $F$.

However, $F$ is not analytic anywhere. For if $t_0 = \frac{k}{2^n}$ is a dyadic rational number, it follows that $2^m t_0 \in \mathbb{Z}$, so $F^{(n)}(t_0) = 0$ for $n > m$. Thus the Taylor series for $F^{(n)}$ around $t_0$ has all terms zero except possibly the constant term. This series therefore has infinite radius of convergence, but represents $F^{(n)}$ only for $t = t_0$ since $F^{(n)}$ is not constant on any interval by (i). So by working backwards, the Taylor series for $F$ around $t_0$ has infinite radius of convergence (it is a polynomial), but represents $F$ only at $t_0$. So every dyadic rational is a (C)-point for $F$, i.e. $F$ has a dense set of (C)-points.
Smooth Functions With Only (P)-Points

The following example of a smooth function with only (P)-points is from [http://mathoverflow.net/questions/43462/existence-of-a-smooth-function-with-nowhere-converging-taylor-series-at-every-poi/81465#81465](http://mathoverflow.net/questions/43462/existence-of-a-smooth-function-with-nowhere-converging-taylor-series-at-every-poi/81465#81465). This is a specific instance of a general class of examples from [SZ55]. The first example of this kind was given by H. Cartan in 1940 [Car40].

**V.17.4.3. Example.** Let

\[
f_n = (n!)^{-3n+2} \cos((n!)^3 x)
\]

which has the following properties:

(i) \(|f_n^{(k)}(x)| \leq \frac{1}{n!}\) for all \(x\) if \(0 \leq k \leq n - 1\).

(ii) \(|f_n^{(n)}(x)|^2 + |f_n^{(n+1)}(x)|^2 \geq (n!)^4\) for all \(x\); hence \(\max(|f_n^{(k)}(x)|, |f_n^{(n+1)}(x)|) \geq \frac{1}{2}(n!)^2\) for all \(x\) (a better constant can be obtained, but this is sufficient for the argument).

We will inductively choose a sequence \((n_m)\) so that, if \(f = \sum_{m=1}^\infty f_{n_m}\), \(f\) is \(C^\infty\) and the Taylor series of \(f\) around any \(x_0\) has radius of convergence zero. We need to insure that at least some of the derivatives of \(f\) increase rapidly in size at every \(x\).

First note that however the \(n_m\) are chosen (strictly increasing), the resulting \(f\) will be \(C^\infty\) and its derivatives will be given by successive term-by-term differentiation. For the \(k\)'th term-by-term derived series is dominated by the series

\[
\sum_{n=1}^\infty (n!)^{-3n+2+3k}
\]

which obviously converges for each \(k\), so the \(k\)'th derived series converges uniformly on \(\mathbb{R}\) by the \(M\)-test, and we can apply \((\cdot)\). Also note that if \(r > k\), we have, for all \(x\),

\[
\sum_{n=r}^\infty |f_n^{(k)}(x)| \leq \sum_{n=r}^\infty \frac{1}{n!}
\]

which is a tail sum for the Maclaurin series for \(e^1\), so by Taylor's Theorem we have, for some \(c < 1\),

\[
\sum_{n=r}^\infty |f_n^{(k)}(x)| \leq \frac{e^c}{r!} < \frac{e}{r!}
\]

Let \(n_1 = 1\). Suppose \(n_1, \ldots, n_r\) have been chosen. Let \(g(x) = \sum_{m=1}^r f_{n_m}(x)\). Then the Maclaurin series for \(g\) converges to \(g\) on all of \(\mathbb{R}\), and \(g\) is periodic, so there is a constant \(C\) such that \(|g^{(k)}(x)| \leq C(k + 1)!\) for all \(k\) and all \(x \in \mathbb{R}\). Choose \(n \geq n_r + 2\) large enough that \(C(n + 2)! < \frac{1}{4}(n!)^2\), and set \(n_{r+1} = n\). This completes the inductive construction of the \(n_m\).

Fix \(r\) and let \(k = n_r\). Set \(g = \sum_{m=1}^{r-1} f_{n_m}\) and \(h = \sum_{m=r+1}^\infty f_{n_m}\). Then \(f = f_k + g + h\). We have

\[
\max(|f^{(k)}(x)|, |f^{(k+1)}(x)|) \geq \max(|f_k^{(k)}(x)|, |f_k^{(k+1)}(x)|) - \max(|g^{(k)}(x)|, |g^{(k+1)}(x)|) - \max(|h^{(k)}(x)|, |h^{(k+1)}(x)|)
\]

709
Proof:
We will prove the result for meager set in is compact (recall that the topology of for any polynomial polynomials are dense in point of $I$).

$\phi_{m;n}$ is closed, suppose $f_n \to f$ in $C^\infty(I)$. For each $n$ there is an $x_n$ such that $|f_n(k)(x_n)| \leq k! \cdot m \cdot n^k$. Passing to a subsequence, we may assume $x_n \to x$ for some $x \in I$ since $I$ is compact (recall that the topology of $C^\infty(I)$ is metrizable). Then $f_n^{(k)}(x_n) \to f^{(k)}(x)$ for all $k$, since $f_n^{(k)} \to f^{(k)}$ uniformly on $I$ for each $k$ (1). Thus

$|f^{(k)}(x)| = \lim_{n \to \infty} |f_n^{(k)}(x_n)| \leq k! \cdot m \cdot n^k$

for all $k$, i.e. $f \in E_{m,n}$.

Now suppose $f \in E_{m,n}$. Then there is an $x_0 \in I$ such that

$\left|\frac{f^{(k)}(x_0)}{k!}\right|^{1/k} \leq m^{1/k}n$
for all $k$, so we have
\[
\limsup_{k \to \infty} \left| \frac{f^{(k)}(x_0)}{k!} \right|^{1/k} \leq n
\]
so the radius of convergence of the Taylor series for $f$ around $x_0$ is at least $\frac{1}{n}$, and $f \notin \mathcal{P}$. Conversely, suppose $f \notin \mathcal{P}$, i.e. there is an $x_0 \in I$ where the Taylor series has positive radius of convergence. Thus there is an $n \in \mathbb{N}$ such that
\[
\limsup_{k \to \infty} \left| \frac{f^{(k)}(x_0)}{k!} \right|^{1/k} < n
\]
so there is a $k_0$ such that
\[
\left| \frac{f^{(k)}(x_0)}{k!} \right|^{1/k} < n
\]
for all $k > k_0$, i.e. $|f^{(k)}(x_0)| < k! \cdot n^k$ for $k \geq k_0$.

Let $m \in \mathbb{N}$, $m \geq \max(|f(x_0)|, |f'(x_0)|, \ldots, |f^{(k_0)}(x_0)|)$. Then $|f^{(k)}(x_0)| < k! \cdot m \cdot n^k$ for all $k$, i.e. $f \in E_{m,n}$.

**Corollary.** Let $I$ be an interval in $\mathbb{R}$. The set of all smooth functions on $I$ which are nowhere analytic is a residual subset of $C^\infty(I)$ with its usual topology. Thus the set of smooth functions which are analytic at at least one point is a meager set in $C^\infty(I)$.

**Can a Smooth Function Have Only (C)-Points?**

**Proposition.** Let $f \in C^\infty(I)$. Then each $E_{m,n}(f)$ is relatively closed in $I$, and $\bigcup_{m,n} E_{m,n}(f) = I \setminus P_f$. Thus $C_f$ is an $F_\sigma$ and $P_f$ is a $G_\delta$.

**Proof:** The proof of the first statement is a simplified form of the argument in the proof of V.17.4.4., and is left to the reader. For the second statement, note that each of the following subsets of $\mathbb{R}$ is both an $F_\sigma$ and a $G_\delta$:

$I$, an interval ($\emptyset$).

$A_f$, a relatively open subset of $I$.

$C_f \cup P_f = I \setminus A_f$, a relatively closed subset of $I$.

$E_{m,n}(f)$, a relatively closed subset of $I$.

So by the first statement $I \setminus P_f$ is an $F_\sigma$, and thus $C_f = (C_f \cup P_f) \cap (I \setminus P_f)$ is an $F_\sigma$, and $P_f = I \setminus (I \setminus P_f)$ is a $G_\delta$.

Each $E_{m,n}(f)$ is contained in $A_f \cup C_f$. The key fact is that interior points of $E_{m,n}(f)$ are in $A_f$: 711
**V.17.4.8.** Proposition. If $x_0$ is an interior point of $E_{m,n}(f)$ for some $(m,n)$, then $f$ is analytic in a neighborhood of $x_0$.

**Proof:** Let $J$ be an open interval around $x_0$ contained in $E_{m,n}(x_0)$, and let $x \in J$ with $|x - x_0| = r < \frac{1}{n}$. If $p_k$ is the $k$'th Taylor polynomial of $f$ around $x_0$, we have, by Taylor’s Theorem,

$$f(x) - p_k(x) = \frac{f^{(k+1)}(c)}{(k+1)!}(x-x_0)^{k+1}$$

for some $c \in J$, and so we have

$$|f(x) - p_k(x)| \leq \frac{(k+1)! \cdot m \cdot n^{k+1}}{(k+1)!}|x-x_0|^{k+1} = m(nr)^{k+1} \to 0$$

and so the Taylor series for $f$ around $x_0$ converges to $f$ on $J \cap (x_0 - \frac{1}{n}, x_0 + \frac{1}{n})$. 

**V.17.4.9.** Theorem. Let $f \in C^\infty(I)$. Then the set $C_f$ is a meager subset of $I$, and in particular has empty interior.

**Proof:** We have that $C_f = \cup_{m,n}[C_f \cap E_{m,n}(f)]$, and by **V.17.4.8.** $C_f \cap E_{m,n}(f)$ is nowhere dense (note that $C_f \cap E_{m,n}(f) = A_{C_f} \cap E_{m,n}(f)$ is relatively closed in $I$).

**V.17.4.10.** Corollary. Let $f \in C^\infty(I)$.

(i) If $P_f = \emptyset$, i.e. if the Taylor series for $f$ around $x_0$ has positive radius of convergence for every $x_0 \in I$, then $f$ is analytic on a dense open subset of $I$.

(ii) If $f$ is not analytic at any point of $I$, then $P_f$ is dense in $I$.

Note that a function satisfying (i) is not necessarily analytic everywhere: Cauchy’s example is a counterexample. But there is a companion result to (i), first stated by Pringsheim in 1893 with a faulty proof and proved by R. Boas in 1935 in the same paper [BJ35] where **V.17.4.9.** was first proved:

**V.17.4.11.** Theorem. Let $f \in C^\infty(I)$. If there is an $r > 0$ such the Taylor series for $f$ around $x_0$ has radius of convergence at least $r$, for every $x_0 \in I$, then $f$ is analytic at every point of $I$ (i.e. analytic on $I$ if $I$ is open).
V.17.4.12. We give only a partial argument; for the full proof see (). Since $P_f$ is empty, it follows from V.17.4.10. that $C_f$ is a nowhere dense relatively closed subset of $I$. We only show that $C_f$ has no isolated points. If $x_0$ is an isolated point of $C_f$, i.e. $f$ is analytic on a deleted open interval $J$ around $x_0$, and $r > 0$ is as in the statement, take $x_1 \in J$ with $x_0 < x_1 < x_0 + r$. The Taylor series for $f$ around $x_1$ defines an analytic function $g$ on $(x_1 - r, x_1 + r)$, and $g$ agrees with $f$ on $(x_0, x_1)$ by V.16.2.2., since they agree on an interval around $x_1$. By continuity of all derivatives of $f$ and $g$, the Taylor series for $g$ around $x_0$ agrees with the Taylor series for $f$ around $x_0$, so the Taylor series for $f$ converges to $f$ on $[x_0, x_1]$ by V.16.2.7. (applied to $g$). A similar argument shows that the Taylor series for $f$ around $x_0$ converges to $f$ on an interval $(x_2, x_0]$ for some $x_2 < x_0$. Thus $f$ is analytic at $x_0$, contradicting that $x_0 \notin C_f$.

This argument does not rule out that $C_f$ could look like a Cantor set. Showing that this is impossible requires a more delicate argument involving application of the Baire Category Theorem to the set $C_f$, which is a locally compact metrizable space.

Zahorski’s Theorem

In 1947 Z. ZAHORSKI [Zah47] gave a complete characterization of which subsets of $I$ occur as $C_f$ and $P_f$ for a smooth function $f$ on an interval $I$. We have shown that these sets have the following properties:

- $C_f$ is a meager $F_\sigma$.
- $P_f$ is a $G_\delta$ disjoint from $C_f$.
- $C_f \cup P_f$ is relatively closed in $I$.

Zahorski’s Theorem asserts that these conditions are sufficient as well as necessary:

V.17.4.13. Theorem. [Zahorski’s Theorem] Let $I$ be an interval in $\mathbb{R}$, and let $C$ and $P$ be subsets of $I$ such that

- $C$ is a meager $F_\sigma$.
- $P$ is a $G_\delta$ disjoint from $C$.
- $C \cup P$ is relatively closed in $I$.

Then there is an $f \in C^\infty(I)$ such that $C_f = C$ and $P_f = P$.

See [SZ55] for a proof which is somewhat simpler than ZAHORSKI’s, but still complicated.

V.17.4.14. Thus, although $C_f$ is always a meager set, it can be large in a measure-theoretic sense. For example, there is a smooth function $f$ on $\mathbb{R}$ for which $P_f$ is the set of Liouville numbers and $C_f$ the complement. This $f$ is not analytic anywhere, and $C_f$ is a set whose complement has Lebesgue measure zero, and in fact Hausdorff dimension zero (III.8.2.37.).

V.17.4.15. There is a multidimensional version of Zahorski’s Theorem [?].
V.17.5. Exercises

V.17.5.1. Let

\[ f(x) = \begin{cases} 
  e^{-1/x^2} \sin \frac{1}{x} & \text{if } x \neq 0 \\
  0 & \text{if } x = 0 
\end{cases}. \]

Show that \( f \) is \( C^\infty \) on all of \( \mathbb{R} \), and has a nonisolated zero at 0.
V.17.6. Exercises

1. Prove a generalization of the Ratio Test for positive series \( \sum_{k=1}^{\infty} a_k \) using \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} \), i.e. find conditions which allow the proof of (1) to work in greater generality. Note the results of Exercises (1) and (2).

2. Choose an increasing sequence \((k_n)\) such that \( \sum_{k=k_n+1}^{\infty} \frac{1}{k^2} < 4^{-n} \), and for \( k_n < k \leq k_{n+1} \) let \( a_k = \frac{2^n}{k^2} \).
   (a) Show that \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} = 2 \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} = 1 \).
   (b) Show that \( \sum_{k=1}^{\infty} a_k \) converges.
   (c) Construct a similar convergent positive series \( \sum_{k=1}^{\infty} b_k \) with \( \limsup_{k \to \infty} \frac{b_{k+1}}{b_k} = +\infty \) and \( \liminf_{k \to \infty} \frac{b_{k+1}}{b_k} = 1 \).

3. Choose an increasing sequence \((k_n)\) such that \( \sum_{k=1}^{k_n} \frac{1}{k} > 4^n \), and for \( k_n < k \leq k_{n+1} \) let \( a_k = \frac{2^n}{k} \).
   (a) Show that \( \limsup_{k \to \infty} \frac{a_{k+1}}{a_k} = 1 \) and \( \liminf_{k \to \infty} \frac{a_{k+1}}{a_k} = \frac{1}{2} \).
   (b) Show that \( \sum_{k=1}^{\infty} a_k \) diverges.
   (c) Construct a similar divergent positive series \( \sum_{k=1}^{\infty} b_k \) with \( \limsup_{k \to \infty} \frac{b_{k+1}}{b_k} = 1 \) and \( \liminf_{k \to \infty} \frac{b_{k+1}}{b_k} = 0 \).
V.17.7. The Binomial Theorem

The Binomial Theorem in algebra gives a formula for positive integral powers of a sum of two numbers. The Binomial Theorem can be extended to apply also to negative and nonintegral powers, but the extended version involves infinite series and there are convergence questions which must be addressed.

Recall that the binomial coefficients are defined by the following formula:

V.17.7.1. Definition. If \( n \in \mathbb{N}, 0 \leq k \leq n \), then

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2) \cdots (n-k+1)}{k!}
\]

V.17.7.2. The last expression makes sense for any \( n \in \mathbb{R} \) if \( k \in \mathbb{N} \), and hence we can define the generalized binomial coefficient

\[
\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2) \cdots (\alpha-k+1)}{k!}
\]

for \( \alpha \in \mathbb{R}, k \in \mathbb{N} \). We set \( \binom{\alpha}{0} = 1 \) for all \( \alpha \) (even \( \alpha = 0 \)). Note that if \( n \in \mathbb{N} \cup \{0\} \), then \( \binom{n}{k} = 0 \) for \( k > n \); but if \( \alpha \notin \mathbb{N} \cup \{0\} \), then \( \binom{\alpha}{k} \neq 0 \) for all \( k \). \( \binom{n}{k} \) is read “\( \alpha \) choose \( k \),” since if \( n \in \mathbb{N} \), \( \binom{n}{k} \) is just the number of ways of choosing \( k \) elements out of an \( n \)-element set.

Here is the standard Binomial Theorem of algebra:

V.17.7.3. Theorem. Let \( x, y \in \mathbb{C}, n \in \mathbb{N} \). Then

\[(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \]

(with the convention that \( t^0 = 1 \) for all \( t \), even \( t = 0 \).) The formula actually holds (suitably interpreted) in any commutative ring.

This theorem can easily be proved by induction on \( n \) using the following relation among binomial coefficients, which can be proved by induction on \( k \):

V.17.7.4. Proposition. Let \( \alpha, \beta \in \mathbb{R}, k \in \mathbb{N} \cup \{0\} \). Then

\[
\binom{\alpha + \beta}{k} = \sum_{j=0}^{k} \binom{\alpha}{k-j} \binom{\beta}{j}
\]

In particular, if \( k \in \mathbb{N} \),

\[
\binom{\alpha + 1}{k} = \binom{\alpha}{k} + \binom{\alpha}{k-1}
\]

The fact that the binomial coefficient \( \binom{n}{k} \) is always a natural number if \( n \) and \( k \) are natural numbers with \( k \leq n \) has an interesting consequence in number theory, which is not entirely obvious:
V.17.7.5. Corollary. If \( k \in \mathbb{N} \), then the product of any \( k \) consecutive integers is divisible by \( k! \).

Proof: Let \( n \) be the largest number in the sequence. Then \( n \geq k \), and
\[
\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \binom{n}{k} \in \mathbb{N}.
\]

V.17.7.6. Corollary. Let \( g \) be a polynomial with integer coefficients, and \( k \in \mathbb{N} \). Then all coefficients of the \( k \)'th derivative \( g^{(k)} \) are divisible by \( k! \).

Proof: It suffices to prove this for \( g(x) = x^n \). If \( n \geq k \), then \( g^{(k)}(x) = n(n-1)(n-2)\cdots(n-k+1)x^{n-k} \), and \( g^{(k)} \) is identically 0 if \( k > n \).

This result is also an easy corollary of Taylor’s Theorem.

V.17.7.7. To prove a version of the Binomial Theorem when \( n \) is not a positive integer, we restrict to the case where \( y = 1 \) and regard \( (1 + x)^n \) as a function of \( x \). This is not an essential restriction, since if \( y \neq 0 \) we have \((x+y)^n = y^n \left( \frac{x}{y} + 1 \right)^n \).

Fix \( \alpha \in \mathbb{R} \) and consider the function \( f(x) = (1 + x)^\alpha \). This function is infinitely differentiable on \((-1, \infty)\) and extends to a continuous function on \([-1, \infty)\) if \( \alpha \geq 0 \). For certain special \( \alpha \) it also extends to numbers less than \(-1\).

V.17.7.8. Proposition. For any \( k \in \mathbb{N} \cup \{0\} \),
\[
\frac{f^{(k)}(0)}{k!} = \binom{\alpha}{k}.
\]
Thus the Taylor series for \( f \) around 0 is
\[
\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k.
\]

The proof of the first formula is an easy calculation, and the second formula follows immediately.

V.17.7.9. Theorem. [Binomial Theorem] For any \( \alpha \in \mathbb{R} \), the series
\[
\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k
\]
converges on \((-1, 1)\) to \((1 + x)^\alpha\). If \( \alpha \geq 0 \), the series converges uniformly on \([-1, 1]\) to \((1 + x)^\alpha\).
Proof: The proof has three steps. (i) Show that the series converges on \((-1, 1)\). (ii) Show that the series represents the function on \((-1, 1)\). (iii) If \(\alpha \geq 0\), show the series converges uniformly on \([-1, 1]\). Since the results are obvious when \(\alpha \in \mathbb{N} \cup \{0\}\) (since then the sum is finite), we will assume \(\alpha \notin \mathbb{N} \cup \{0\}\). We set \(a_k = \binom{\alpha}{k}\); then \(a_k > 0\) for all \(k\).

(i) Using the Ratio Test, the radius of convergence of the series is

\[
\lim_{k \to \infty} \frac{a_k}{a_{k+1}} = \lim_{k \to \infty} \frac{k+1}{k-\alpha} = 1
\]

so the series converges for \(|x| < 1\).

(ii) For \(x \in (-1, 1)\), let \(f(x) = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k\). Then \(f\) is differentiable on \((-1, 1)\), and

\[
f'(x) = \sum_{k=1}^{\infty} k \binom{\alpha}{k} x^{k-1} = \sum_{k=0}^{\infty} (k+1) \binom{\alpha}{k+1} x^k
\]

on \((-1, 1)\). We have

\[
(1+x)f'(x) = \sum_{k=0}^{\infty} \left[(k+1) \binom{\alpha}{k+1} + k \binom{\alpha}{k}\right] x^k.
\]

Since

\[
(k+1) \binom{\alpha}{k+1} + k \binom{\alpha}{k} = \alpha \binom{\alpha-1}{k}
\]

we have

\[
(1+x)f'(x) = \frac{d}{dx} (1+x)^{\alpha} = \alpha (1+x)\alpha-1 f(x) = 0.
\]

Thus \((1+x)^{-\alpha} f(x)\) is constant, so \(f(x) = C(1+x)^\alpha\) for some \(C\). Since \(f(0) = 1\), \(C = 1\).

(iii) Suppose \(\alpha > 0\). We will show that \(\sum_{k=0}^{\infty} a_k\) converges. If \(k > \alpha\), then \(\frac{a_{k+1}}{a_k} = \frac{k-\alpha}{k+1}\), so \((k+1)a_{k+1} = (k-\alpha)a_k, ka_k - (k+1)a_{k+1} = \alpha a_k > 0\). Thus the sequence \((ka_k)\) is nonnegative and decreasing for \(k > \alpha\), and hence \(\lambda = \lim_{k \to \infty} ka_k\) exists. We have

\[
\sum_{k=0}^{\infty} (ka_k - (k+1)a_{k+1}) = -(n+1)a_{n+1} \to -\lambda
\]

so \(\sum_{k=0}^{\infty} (ka_k - (k+1)a_{k+1})\) converges. Since \(a_k = \alpha^{-1}(ka_k - (k+1)a_{k+1})\) for \(k > \alpha\), \(\sum_{k=0}^{\infty} a_k\) converges. Thus, by the Weierstrass M-test, \(\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k\) converges absolutely and uniformly on \([-1, 1]\). Since by (ii) the series represents \((1+x)^\alpha\) on \((-1, 1)\), it represents \((1+x)^\alpha\) on \([-1, 1]\) by continuity.

\(\diamondsuit\)
V.17.8. Exercises

V.17.8.1. (a) Prove the second formula in V.17.7.4. by induction on $k$. [Hint: show using the inductive step that
\[
\binom{\alpha + 1}{k + 1} = \binom{\alpha}{k} \left[ \frac{\alpha - k + 1}{k + 1} \right] + \binom{\alpha - k}{k - 1} \frac{\alpha - k + 1}{k + 1}
\]
and consider three terms by multiplying. Show that the first term is $(\frac{\alpha}{k+1})$ and the last is $\frac{k}{k+1} (\frac{\alpha}{k})$.]

(b) Give a combinatorial argument for this formula if $\alpha \in \mathbb{N}$, interpreting $\binom{n}{k}$ as the number of ways of choosing $k$ elements from an $n$-element set.

(c) Use this formula to prove V.17.7.3. by induction on $n$.

(d) Generalize the argument in (a) to prove the first formula in V.17.7.4.

V.17.8.2. Here is an alternate proof of V.17.7.9. (ii) (this was essentially Cauchy’s original argument). Fix $x$ with $|x| < 1$. Set $F_n(\alpha) = \sum_{k=0}^{\alpha} \binom{\alpha}{k} x^k$.

(a) Mimic the proof of the Ratio test to show that the infinite series $\sum_{k=0}^{\alpha} \binom{\alpha}{k} x^k$ converges absolutely and uniformly on $\mathbb{R}$ to a function $F(\alpha)$, i.e. $F_n(\alpha) \to F(\alpha)$ uniformly in $\alpha$. Thus $F(\alpha)$ is continuous in $\alpha$.

(b) For fixed $\alpha, \beta \in \mathbb{R}$, use the product formula for power series () to show that
\[
F(\alpha)F(\beta) = \sum_{k=0}^{\infty} \left[ \sum_{j=0}^{k} \binom{\alpha}{k-j} \binom{\beta}{j} \right] x^k.
\]

(c) Conclude from V.17.7.4. that $F(\alpha)F(\beta) = F(\alpha + \beta)$.

(d) Apply () to conclude that $F(\alpha) = b^\alpha$ for some $b > 0$. Since $F(1) = 1 + x$, $b = 1 + x$.

V.17.8.3. (a) Show that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$,
\[
(1 - x)(1 + x + x^2 + \cdots + x^n) = 1 - x^{n+1}.
\]

(b) Show that for all $x \in \mathbb{R}, x \neq -1$,
\[
\frac{1}{1+x} = 1 - x + x^2 - \cdots + (-1)^n x^n + \frac{(-1)^{n+1} x^{n+1}}{1+x}.
\]

(c) Show from (b) that if $|x| < 1$, then the series $\sum_{n=0}^{\infty} (-1)^n x^n$ converges to $\frac{1}{1+x}$. Compare with the Binomial Theorem for $\alpha = -1$.

V.17.8.4. Here is yet another proof of V.17.7.9. (ii), based on the Cauchy Remainder Formula for Taylor’s Theorem (V.10.6.7.).

(a) Fix $x$ with $|x| < 1$. By the Cauchy remainder formula, for any $n$ there is a $\theta = \theta_n$, $0 < \theta < 1$, such that
\[
R_n(x) = (1 + x)^\alpha - \sum_{k=0}^{n} \binom{\alpha}{k} x^k = \frac{f^{(n+1)}(\theta x)}{n!} (1 - \theta)^n x^{n+1}
\]
\[
\begin{align*}
\frac{\alpha(\alpha - 1) \cdots (\alpha - n)(1 + \theta x)^{\alpha - n - 1}}{n!} (1 - \theta)^n x^{n+1} \\
= \left[\frac{(\alpha - 1)(\alpha - 2) \cdots (\alpha - n)}{n!} x^n\right] \left[\alpha x(1 + \theta x)^{\alpha - 1} \left(\frac{1 - \theta}{1 + \theta x}\right)^n\right]
\end{align*}
\]

(b) If \( M = \max((1 + x)^{\alpha - 1}, (1 - x)^{\alpha - 1}) \), the second factor is bounded in absolute value by \( M|\alpha x| \) independent of \( n \) since \( \left|\frac{1 - \theta}{1 + \theta x}\right| < 1 \).

(c) The first factor is \( \binom{\alpha - 1}{n} x^n \) which goes to 0 as \( n \to \infty \) since \( \sum \binom{\alpha - 1}{n} x^n \) converges by V.17.7.9.(i).

(d) Thus \( R_n(x) \to 0 \) as \( n \to \infty \) and the series converges to \( (1 + x)^\alpha \).
Chapter VI

The Topology of \( \mathbb{R} \), \( \mathbb{R}^n \), and Metric Spaces

In this chapter we discuss the basic notions of open and closed sets, closure of a set, interior and boundary of a set, and some of the elementary results about them, beginning with the real numbers, then moving on to \( \mathbb{R}^n \) and general metric spaces. This chapter should be considered a parallel thread to Chapter ( ) on sequences and series; there is much interplay between the two chapters.

The general theory of topology and more detailed results along the lines of this chapter are the subject of Chapter ( ).

VI.1. The Topology of \( \mathbb{R} \)

We begin with the real numbers. Recall the definition of an interval in \( \mathbb{R} \) (III.1.9.12.). The open intervals and the closed intervals are most important for this section. The open intervals are the intervals of the form \((a, b)\), where \( a \in \mathbb{R} \cup \{-\infty\}, \ b \in \mathbb{R} \cup \{+\infty\}, \ \text{and} \ a < b \). The closed intervals are of the form \([a, b]\) with \( a, b \in \mathbb{R}, \ a < b \), or \([a, +\infty), \ (-\infty, b]\), or all of \( \mathbb{R} \). We generalize both open intervals and closed intervals into open sets and closed sets respectively.

VI.1.1. Open and Closed Sets in \( \mathbb{R} \)

VI.1.1.1. Definition. (i) A subset \( U \) of \( \mathbb{R} \) is an open set if, whenever \( x \in U \), there is an open interval \( I \) with \( x \in I \) and \( I \subseteq U \).
(ii) A subset \( E \) of \( \mathbb{R} \) is a closed set if \( E^c = \mathbb{R} \setminus E \) is an open set.

A slight rephrasing of the definition gives:

VI.1.1.2. Proposition. A subset of \( \mathbb{R} \) is an open set if and only if it is a union of open intervals.
VI.1.1.3. An open set is, roughly, a set which locally looks like an open interval. Any open interval is an open set. The empty set vacuously satisfies the definition of open set, hence is also an open set (it is an “empty union” of open intervals). The complement of a closed interval is either one open interval, the disjoint union of two open intervals, or the empty set, hence open; thus a closed interval is a closed set, so the terminology is consistent. The complement of the empty set is $\mathbb{R}$, which is open, so $\emptyset$ is closed as well as open. There are many nonempty closed sets which look nothing like closed intervals; for example, any finite set is closed since its complement is a union of open intervals; similarly $\mathbb{Z}$ is a closed set. There are much more complicated examples; see e.g. $\left(\right)$, $\left(\right)$.

A half-open interval of the form $[a,b)$ for $a < b$ is not an open set since it does not contain an open interval around $a$; similarly, its complement is not open, so $[a,b)$ is not a closed set either. $\mathbb{Q}$ is also neither open nor closed, since every open interval contains both rational and irrational numbers. In fact, “most” subsets of $\mathbb{R}$ are neither open nor closed.

The only subsets of $\mathbb{R}$ which are both open and closed are $\emptyset$ and $\mathbb{R}$ ($\emptyset$).

Here is a useful rephrasing of the definition of open set:

VI.1.1.4. Proposition. Let $U$ be a subset of $\mathbb{R}$. Then $U$ is an open set if and only if, for every $x \in U$, there is an $\epsilon > 0$ such that the open interval $(x-\epsilon, x+\epsilon) = \{y \in \mathbb{R} : |y-x| < \epsilon\}$ is contained in $U$. (The $\epsilon$ can depend on $x$.)

Proof: Any such set is obviously open, since $(x-\epsilon, x+\epsilon)$ is an open interval. Conversely, if $U$ is open and $x \in U$, and $I$ is an open interval with $x \in I$ and $I \subseteq U$, then there is a smaller open interval $J$ centered at $x$ contained in $I$ (and hence in $U$); $J$ will be of the form $(x-\epsilon, x+\epsilon)$ for some $\epsilon > 0$. $\Diamond$

Unions and Intersections

VI.1.1.5. Proposition. (i) An arbitrary union of open sets is open. A finite intersection of open sets is open.

(ii) An arbitrary intersection of closed sets is closed. A finite union of closed sets is closed.

Proof: (i): It is obvious from VI.1.1.2. (or just straight from the definition) that a union of any collection of open sets is open. For a finite intersection, we use VI.1.1.4.. Let $U_1, \ldots, U_n$ be a finite collection of open sets, and $U = U_1 \cap \cdots \cap U_n$. Let $x \in U$. Then $x \in U_j$ for each $j$, so for each $j$ there is an $\epsilon_j > 0$ such that $(x-\epsilon_j, x+\epsilon_j) \subseteq U_j$. Set $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$. Then $\epsilon > 0$, and $(x-\epsilon, x+\epsilon) \subseteq U$. Since this can be done for any $x \in U$, $U$ is open.

(ii) follows from (i) by taking complements and using de Morgan’s Laws ($\emptyset$).

VI.1.1.6. The intersection of an infinite number of open sets is not open in general. For example, $\cap_{n=1}^{\infty} \left(x - \frac{1}{n}, x + \frac{1}{n}\right) = \{x\}$, which is not open (in fact it is closed). Similarly, an infinite union of closed sets is not closed in general. In fact, every subset of $\mathbb{R}$ is a union of singleton sets, each of which is closed.
VI.1.1.7. **COROLLARY.** Let \( U \) be an open set and \( E \) a closed set. Then \( U \setminus E \) is open and \( E \setminus U \) is closed.

**Proof:** \( U \setminus E = U \cap E^c \) and \( E \setminus U = E \cap U^c \).

#### Neighborhoods and Deleted Neighborhoods

VI.1.1.8. Let \( x \) be a real number. An open interval containing \( x \), or more generally an open set containing \( x \), is often called a *neighborhood*, or open neighborhood, of \( x \). In topology, a neighborhood of \( x \) is generally thought of as a set of points “close to \( x \)”; this is literally true if the neighborhood is something like \((x - \epsilon, x + \epsilon)\) for a small positive \( \epsilon \), but may be somewhat of a stretch for general neighborhoods. Nonetheless, the notion of neighborhood is a useful one, as we will see. An open set is an open neighborhood of each of its points.

Sometimes we use the term “neighborhood” to mean a not necessarily open set: a neighborhood of \( x \) is a set containing an open neighborhood of \( x \). For example, it is sometimes useful to consider closed neighborhoods (e.g. \((0)\)).

VI.1.1.9. It is also often useful to consider deleted neighborhoods of a point. If \( x \in \mathbb{R} \), a deleted neighborhood of \( x \) is a subset of \( \mathbb{R} \) of the form \( U \setminus \{x\} \), where \( U \) is a neighborhood of \( x \). Deleted neighborhoods are usually taken to be open, but not necessarily.

The most commonly used deleted neighborhoods of an \( x \in \mathbb{R} \) are sets of the form

\[
(x - \epsilon, x) \cup (x, x + \epsilon) = \{y \in \mathbb{R} : 0 < |y - x| < \epsilon\}
\]

for some (typically “small”) \( \epsilon > 0 \). Every deleted neighborhood of \( x \) contains a deleted neighborhood of this form.

#### VI.1.2. Limit Points and Closures

There is an alternate way to view closed sets, via the notions of limit points and closures, which we now describe.

VI.1.2.1. **DEFINITION.** Let \( A \) be a subset of \( \mathbb{R} \), and \( x \in \mathbb{R} \). Then \( x \) is a *limit point* of \( A \) if every open neighborhood of \( x \) contains at least one point of \( A \).

The word “open” may be deleted from the definition without effect.

VI.1.2.2. Any point of \( A \) is a limit point of \( A \): if \( x \in A \), then any neighborhood of \( x \) contains at least one point of \( A \), namely \( x \) itself. But the importance of the definition is that there may be real numbers which are not in \( A \) but are limit points of \( A \). For example, \( 0 \notin (0,1) \), but 0 is a limit point of \( (0,1) \) since every neighborhood of 0 contains small positive numbers. For a more dramatic example, every real number is a limit point of \( \mathbb{Q} \), since every nonempty open set in \( \mathbb{R} \) contains rational numbers.

Some authors define “limit point” to mean what is more commonly called an *accumulation point* of \( A \). The number \( x \) is an accumulation point of \( A \) if every neighborhood of \( x \) contains a point of \( A \) other than \( x \), i.e. every deleted neighborhood of \( x \) contains a point of \( A \). Not every element of \( A \) is necessarily an accumulation point of \( A \); those which are not are called *isolated points* of \( A \). We will not need to make much use of the notion of accumulation point; limit points are more important for most purposes.

As with open sets, the notion of limit point can be phrased in terms of \( \epsilon \)'s:
VI.1.2.3. **Proposition.** Let \( A \subseteq \mathbb{R} \), \( x \in \mathbb{R} \). Then \( x \) is a limit point of \( A \) if and only if, for every \( \epsilon > 0 \), there is a \( y \in A \) with \( |x - y| < \epsilon \).

Intuitively, \( x \) is a limit point of \( A \) if and only if there are points of \( A \) arbitrarily close to \( x \).

VI.1.2.4. **Definition.** Let \( A \) be a subset of \( \mathbb{R} \). The **closure** of \( A \) is the set \( \bar{A} \) of all limit points of \( A \).

VI.1.2.5. Since every element of \( A \) is a limit point of \( A \), we have \( A \subseteq \bar{A} \). If \( A \subseteq B \), then it follows immediately from the definition of limit point that every limit point of \( A \) is a limit point of \( B \), so \( \bar{A} \subseteq \bar{B} \).

The term “closure” is justified by the next result.

VI.1.2.6. **Theorem.** Let \( A \) be a subset of \( \mathbb{R} \). Then

(i) \( \bar{A} = \bar{A} \), i.e. every limit point of \( \bar{A} \) is in \( \bar{A} \).

(ii) \( \bar{A} \) is a closed set.

(iii) \( \bar{A} \) is the smallest closed set containing \( A \), i.e. if \( E \) is a closed set containing \( A \), then \( \bar{A} \subseteq E \).

**Proof:** (i): Let \( x \) be a limit point of \( \bar{A} \), and \( U \) an open neighborhood of \( x \). Then there is a \( y \in A \) with \( y \in U \). Then \( y \) is a limit point of \( A \), and \( U \) is a neighborhood of \( y \), so there is a \( z \in A \) with \( z \in U \). This is true for every open neighborhood \( U \) of \( x \), so \( x \) is a limit point of \( A \).

(ii): Let \( x \in (\bar{A})^c \). Then \( x \) is not a limit point of \( A \), so there is an open neighborhood \( U \) of \( x \) containing no points of \( A \). Then \( U \) also cannot contain any limit points of \( A \), for if \( y \in U \) were a limit point of \( A \), then the neighborhood \( U \) of \( y \) would have to contain a point of \( A \). Thus \( U \subseteq (\bar{A})^c \). Since such a \( U \) exists for every \( x \in (\bar{A})^c \), \( (\bar{A})^c \) is open.

(iii): There is a smallest closed set \( F \) containing \( A \), namely the intersection of all closed sets containing \( A \) (cf. VI.1.1.5.; there is at least one closed set containing \( A \), \( \mathbb{R} \) itself). Since \( \bar{A} \) is a closed set containing \( A \) by (ii), we have \( F \subseteq \bar{A} \). To show the opposite inclusion, let \( x \notin F \). Since \( F \) is closed, \( F^c \) is an open neighborhood of \( x \) containing no point of \( A \). Thus \( x \) cannot be a limit point of \( A \).  

724
VI.2. Metric Spaces

Much of what was done in the last section works almost verbatim in the more general setting of metric spaces. Besides Euclidean spaces, there are many other metric spaces which are important in analysis, as well as in other parts of mathematics, so it is quite worthwhile working in this generality.

VI.2.1. Definitions

A metric space is a set \( X \) with a specified numerical distance between any two points, which behaves reasonably like the usual distance between numbers or points in Euclidean space. The distance may be regarded as a function \( \rho : X \times X \to \mathbb{R} \), with \( \rho(x, y) \) denoting the distance between \( x \) and \( y \). Formally:

VI.2.1.1. Definition. A metric on a set \( X \) is a function \( \rho : X \times X \to \mathbb{R} \) satisfying:

(i) \( \rho(x, y) \geq 0 \) for all \( x, y \in X \).

(ii) \( \rho(x, y) = 0 \) if and only if \( x = y \) (definiteness).

(iii) \( \rho(x, y) = \rho(y, x) \) for all \( x, y \in X \) (symmetry).

(iv) \( \rho(x, z) \leq \rho(x, y) + \rho(y, z) \) for all \( x, y, z \in X \) (triangle inequality).

A pair \( (X, \rho) \), where \( X \) is a set and \( \rho \) is a metric on \( X \), is called a metric space. The diameter of a metric space \( (X, \rho) \) is \( \text{diam}(X, \rho) = \sup \{ \rho(x, y) : x, y \in X \} \). The metric \( \rho \) is bounded if \( \text{diam}(X, \rho) \) is finite.

VI.2.1.2. It is customary to call a set \( X \) with a given metric a “space”, and to call elements of \( X \) “points”, even though there may be no geometric picture or interpretation of \( X \) as an actual space in the usual sense. The geometric terminology helps to motivate some aspects of the theory of metric spaces.

VI.2.1.3. Examples.

(i) On \( \mathbb{R} \), the function \( \rho(x, y) = |x - y| \) is a metric. This is called the standard metric on \( \mathbb{R} \), and we will often use this metric on \( \mathbb{R} \) without specific mention.

(ii) On \( \mathbb{R}^n \), the Euclidean distance function

\[
\rho_2(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}
\]

for \( x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \) in \( \mathbb{R}^n \) is a metric, the standard metric on \( \mathbb{R}^n \).

(iii) There are other important metrics on \( \mathbb{R}^n \). Two of them are:

\[
\rho_1(x, y) = |x_1 - y_1| + \cdots + |x_n - y_n|
\]

\[
\rho_\infty(x, y) = \max(|x_1 - y_1|, \ldots, |x_n - y_n|)
\]

for \( x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \) in \( \mathbb{R}^n \).
(iv) There are important metrics on function spaces analogous to those of (ii) and (iii). For example, let 
\[ C[a, b] \]
be the set of real-valued continuous functions on the interval \([a, b]\), and for \( f, g \in C[a, b] \) define
\[
\rho_2(f, g) = \left[ \int_a^b (f(t) - g(t))^2 \, dt \right]^{1/2}
\]
\[
\rho_1(f, g) = \int_a^b |f(t) - g(t)| \, dt
\]
\[
\rho_\infty(f, g) = \max_{a \leq t \leq b} |f(t) - g(t)|
\]
(It is not obvious that \( \rho_2 \) satisfies the triangle inequality. See Exercise (iv) or (v) for a proof.)

(v) Let \( X \) be a sphere (recall that a sphere in mathematics is a two-dimensional surface, not including the points “inside”). If \( x, y \in X \), let \( \rho(x, y) \) be the arc length of the shortest great circle path between \( x \) and \( y \). (The triangle inequality is not entirely obvious here, but can be proved.) This example has obvious practical applications: it is the appropriate “distance” to consider on the surface of the earth.

(vi) Let \( X \) be any set, \( \rho(x, y) = \begin{cases} 
1 & \text{if } x \neq y \\
0 & \text{if } x = y 
\end{cases} \). Then \( \rho \) is a metric on \( X \), called the \textit{discrete metric}.

(vii) Let \( X \) be the set of all \( n \)-tuples of 0’s and 1’s, and define \( \rho(x, y) \) to be the number of coordinates in which \( x \) and \( y \) differ. This is a metric, called the \textit{Hamming metric}, which is important in coding theory.

In these examples, only the metrics in (v)–(vii) are bounded.

VI.2.1.4. A “distance function” \( \rho \) on a set \( X \) satisfying (i), (iii), and (iv) is called a \textit{pseudometric}. Pseudometrics play an important role in more advanced parts of analysis.

VI.2.2. \textbf{Balls, Limit Points, and Closures}

VI.2.2.1. \textbf{Definition.} Let \((X, \rho)\) be a metric space, and \( x_0 \in X \). If \( r > 0 \), the \textit{(open) ball of radius} \( r \) \textit{around} \( x_0 \) (or \textit{centered at} \( x_0 \)) is
\[ B_r^o(x_0) = \{ x \in X : \rho(x_0, x) < r \} \]
(the \( \rho \) is often omitted if the metric is understood and there is no possibility of confusion.) The \textit{closed ball of radius} \( r \) \textit{around} \( x_0 \) is
\[ B_r^c(x_0) = \{ x \in X : \rho(x_0, x) \leq r \} \].

VI.2.2.2. The term “ball” is a standard term even though \( B_r(x_0) \) need not look anything like a geometric ball. (In fact, there is usually no geometric interpretation of distance in a general metric space; see Exercise (v), for example, for interpretations of \( \rho_1 \) and \( \rho_\infty \) on \( \mathbb{R}^2 \).) The term “ball” without the qualification “closed” will always mean “open ball”.

726
VI.2.2.3. Examples.

(i) In \( \mathbb{R} \) with the usual metric, the open ball of radius \( r \) around \( x_0 \) is the open interval of length \( 2r \) centered at \( x_0 \); \( B_r(x_0) \) is the closed interval of length \( 2r \) centered at \( x_0 \).

(ii) In \( (\mathbb{R}^2, \rho_2) \), \( B_r(x_0) \) is an open disk centered at \( x_0 \). In \( (\mathbb{R}^3, \rho_2) \), \( B_r(x_0) \) is the usual geometric open ball of radius \( r \) centered at \( x_0 \). The closed balls have a similar interpretation.

(iii) In \( (\mathbb{R}^2, \rho_1) \), balls (open or closed) are square diamonds; in \( (\mathbb{R}^2, \rho_\infty) \) they are squares with horizontal and vertical sides (see Exercise ().)

(iv) There is a geometric interpretation of balls in \( C[a,b] \) for the \( \rho_\infty \) metric. Fix \( f_0 \in C[a,b] \), and \( r > 0 \). Draw the graphs of the functions \( f(t) - r \) and \( f(t) + r \) (Figure ().) Then \( B^{\rho_\infty}_{r}(f_0) \) consists of all functions in \( C[a,b] \) whose graph lies within the strip strictly between the top and bottom curves. \( \bar{B}^{\rho_\infty}_{r}(f_0) \) consists of the functions whose graphs lie in the closed strip including the two boundary functions.

(v) If \( \rho \) is the discrete metric on a set \( X \), and \( x_0 \in X \), then \( B^{\rho}_{r}(x_0) = \{x_0\} \) if \( r \leq 1 \), and \( B^{\rho}_{r}(x_0) = X \) if \( r > 1 \). \( B^{\rho}_{r}(x_0) = \{x_0\} \) if \( r < 1 \) and \( \bar{B}^{\rho}_{r}(x_0) = X \) if \( r \geq 1 \). Thus \( B^{\rho}_{1}(x_0) = \{x_0\} \) but \( \bar{B}^{\rho}_{1}(x_0) = X \), so \( B_r(x_0) \) can be much larger than the “closure” of \( B_r(x_0) \) in general (this will be made precise in ().)

VI.2.2.4. Definition. Let \((X, \rho)\) be a metric space, and \( A \subseteq X \). A point \( x \in X \) is a limit point (or point of closure) of \( A \) (with respect to \( \rho \)) if every ball centered at \( x \) contains at least one point of \( A \), i.e., if, for every \( \epsilon > 0 \), there is a point \( y \in A \) with \( \rho(x,y) < \epsilon \). The closure of \( A \) (with respect to \( \rho \)), denoted \( \bar{A} \), is the set of all limit points of \( A \). The set \( A \) is closed if \( A = \bar{A} \).

When there is any possibility of confusion about which metric is being used, we will use the terms \( \rho \)-limit point and \( \rho \)-closed set.

VI.2.2.5. Note that a limit point of \( A \) is not an element of \( A \) in general: for example, 0 and 1 are limit points of \((0,1)\) in \( \mathbb{R} \). But a point of \( A \) is always a limit point of \( A \), so \( A \subseteq \bar{A} \) for any \( A \). If \( A \subseteq B \), then \( \bar{A} \subseteq B \). We also have that \( \emptyset = \emptyset \) and \( \bar{X} = X \), i.e. \( \emptyset \) and \( X \) are closed sets in \( X \).

VI.2.2.6. Proposition. Let \((X, \rho)\) be a metric space, and \( A \subseteq X \). Then every limit point of \( \bar{A} \) is a limit point of \( A \), i.e. \( \bar{A} = \bar{A} \), so \( A \) is closed.

Proof: Suppose \( x \) is a limit point of \( \bar{A} \), and \( \epsilon > 0 \). We must show that there is a \( z \in A \) with \( \rho(x,z) < \epsilon \). But \( x \) is a limit point of \( \bar{A} \), so there is a \( y \in \bar{A} \) with \( \rho(x,y) < \epsilon/2 \); and there is a \( z \in A \) with \( \rho(y,z) < \epsilon/2 \). By the triangle inequality, \( \rho(x,z) < \epsilon \).

The complement of a \( \rho \)-closed set is called an open set (with respect to \( \rho \), or \( \rho \)-open set). The next proposition gives two other useful characterizations of open sets:
VI.2.2.7. **Proposition.** Let \((X, \rho)\) be a metric space, and \(U \subseteq X\). Then the following are equivalent:

(i) \(U\) is an open set.

(ii) \(U\) is a union of open balls.

(iii) For every \(x \in U\), there is an \(\epsilon > 0\) such that \(B_\epsilon(x) \subseteq U\).

**Proof:**

(i) \(\Rightarrow\) (iii): If \(x \in U\), then \(x\) is not a limit point of \(U^c\) since \(U^c\) is closed. Thus, there is an \(\epsilon > 0\) such that \(B_\epsilon(x) \subseteq U\), i.e. \(B_\epsilon(x) \subseteq U\).

(iii) \(\Rightarrow\) (i): If \(x \in U\), then there is an \(\epsilon > 0\) such that \(B_\epsilon(x) \subseteq U\), hence \(B_\epsilon(x) \subseteq U\) and \(x\) is not a limit point of \(U^c\). Thus \(U^c\) is closed.

(ii) \(\Rightarrow\) (iii): Suppose \(x \in U\). Then \(x \in B_r(y) \subseteq U\) for some \(r\) and \(y\), since \(U\) is a union of open balls. Set \(\epsilon = r - \rho(x, y)\). Then by the triangle inequality, \(B_\epsilon(x) \subseteq B_r(y) \subseteq U\).

(iii) \(\Rightarrow\) (ii) is obvious. \(\blacksquare\)

In particular, every open ball is an open set.

VI.2.2.8. **Corollary.** An arbitrary union of open sets is open. A finite intersection of open sets is open.

An arbitrary intersection of closed sets is closed. A finite union of closed sets is closed.

**Proof:** It is obvious from either (ii) or (iii) that an arbitrary union of open sets is open. To prove that a finite intersection of open sets is open, use (iii): if \(U_1, \ldots, U_n\) are open and \(x \in U_1 \cap \cdots \cap U_n\), choose \(\epsilon_1, \ldots, \epsilon_n > 0\) with \(B_{\epsilon_k}(x) \subseteq U_k\) for each \(k\). Set \(\epsilon = \min(\epsilon_1, \ldots, \epsilon_n)\); then \(\epsilon > 0\) and \(B_\epsilon(x) \subseteq U_1 \cap \cdots \cap U_n\).

The statements about unions and intersections of closed sets are obtained by taking complements. \(\blacksquare\)

VI.2.2.9. **Note** that an intersection of infinitely many open sets is not open in general: for example, \(\cap_{n=1}^{\infty} B_{1/n}(x) = \{x\}\), which is rarely an open set (\(\{x\}\) is actually a closed set for any \(x\)). Similarly, an arbitrary union of closed sets is not closed in general: in fact, every subset of \(X\) is a union of singleton subsets, which are closed.

VI.2.2.10. **Note** that not every subset of a metric space is either open or closed; in fact, this is far from true in most metric spaces.

"Most sets, like doors, are neither open nor closed, but ajar."

C. Pugh\(^1\)

\(\uparrow\)\[Pug15, p. 60]\]

In fact, the "most" can be precisely quantified: in \(\mathbb{R}\) (and in most other metric spaces encountered in analysis), there are \(2^{2^{\aleph_0}}\) subsets in all, but only \(2^{\aleph_0}\) open sets and \(2^{\aleph_0}\) closed sets (\(\{\\}\)), so there are \(2^{2^{\aleph_0}}\) subsets which are "ajar." Although "most" of these subsets are rather bizarre and cannot be even explicitly described, many frequently encountered and easily described sets are ajar, such as half-open intervals of the form \((a, b]\) or \([a, b)\), or the rational numbers \(\mathbb{Q}\).
VI.2.3. Limits and Continuity

Convergence of sequences and limits and continuity of functions work in any metric space almost exactly as in \( \mathbb{R} \). The statements and proofs are essentially identical to the case of \( \mathbb{R} \); all that is needed is to “redecorate” using the metric distance in place of the absolute value. We will do a few examples of redecoration in proofs, and leave others as exercises.

Limits of Sequences

VI.2.3.1. Definition. Let \((X, \rho)\) be a metric space, \(x \in X\), and \((x_n)\) a sequence in \(X\). Then \((x_n)\) converges to \(x\) (with respect to \(\rho\)), or \(x\) is the limit of the sequence \((x_n)\), written \(x_n \to x\) or \(x = \lim_{n \to \infty} x_n\), if for every \(\epsilon > 0\) there is an \(n_0 \in \mathbb{N}\) such that \(\rho(x_n, x) < \epsilon\) for all \(n \geq n_0\).

Where it is necessary to explicitly specify the metric \(\rho\), or where there is any chance of confusion, we will say that \((x_n)\) \(\rho\)-converges to \(x\), or \(x\) is the \(\rho\)-limit of \((x_n)\), written \(x_n \to_\rho x\) or \(x = \rho\lim_{n \to \infty} x_n\).

As in \(\mathbb{R}\), limits of sequences are unique:

VI.2.3.2. Proposition. Let \((X, \rho)\) be a metric space, \(x, y \in X\), and \((x_n)\) a sequence in \(X\). If \(x_n \to x\) and \(x_n \to y\), then \(x = y\).

Proof: If \(x \neq y\), set \(\epsilon = \rho(x, y)/3 > 0\), and choose \(n_1, n_2 \in \mathbb{N}\) with \(\rho(x_n, x) < \epsilon\) for all \(n \geq n_1\) and \(\rho(x_n, y) < \epsilon\) for all \(n \geq n_2\). Then, if \(n \geq \max(n_1, n_2)\),

\[
\rho(x, y) \leq \rho(x, x_n) + \rho(x_n, y) < 2\epsilon = 2\rho(x, y)/3 < \rho(x, y)
\]

which is a contradiction. ☐

VI.2.3.3. Proposition. Let \((X, \rho)\) be a metric space, \(x \in X\), and \((x_n)\) a sequence in \(X\). If \(x_n \to x\), then every subsequence of \((x_n)\) also converges to \(x\).

There is also a notion of a cluster point of a sequence in a metric space:

VI.2.3.4. Definition. Let \((X, \rho)\) be a metric space, \(x \in X\), and \((x_n)\) a sequence in \(X\). Then \(x\) is a cluster point of the sequence \((x_n)\) if for every \(\epsilon > 0\) and \(n_0 \in \mathbb{N}\), \(\rho(x_n, x) < \epsilon\) for some \(n \geq n_0\). (Equivalently, \(\rho(x_n, x) < \epsilon\) for infinitely many \(n\).) When \(\rho\) needs to be specified, we say \(x\) is a \(\rho\)-cluster point of \((x_n)\).

If \(x_n \to x\), then \(x\) is a cluster point of \((x_n)\), but the converse is generally not true. A sequence can have many cluster points.

VI.2.3.5. The term “limit point” is sometimes used in place of “cluster point”, but this can lead to confusion, so we will use the term “limit point” only in reference to a set, and “cluster point” in reference to a sequence. If \(x\) is a cluster point of a sequence \((x_n)\), then \(x\) is a limit point of the set \(\{x_n : n \in \mathbb{N}\}\), but the converse is not necessarily true: each term \(x_k\) is a limit point of \(\{x_n : n \in \mathbb{N}\}\), but \(x_k\) is usually not a cluster point of \((x_n)\). In fact, we have:
VI.2.3.6. Proposition. Let \((x_n)\) be a sequence in a metric space \((X, \rho)\), and let \(x \in X\). Then \(x\) is a cluster point of \((x_n)\) if and only if \(x\) is a limit point of
\[
A_k = \{x_n : n \geq k\}
\]
for every \(k \in \mathbb{N}\).

This proposition is an immediate consequence of the definitions, and the proof is left as an exercise.

Cluster points can also be characterized using subsequences:

VI.2.3.7. Proposition. Let \((x_n)\) be a sequence in a metric space \((X, \rho)\), and let \(x \in X\). Then \(x\) is a cluster point of \((x_n)\) if and only if there is a subsequence \((x_{k_n})\) which converges to \(x\).

Proof: Suppose there is a subsequence of \((x_n)\) converging to \(x\). Then, for any \(\epsilon > 0\), infinitely many terms of the subsequence lie within \(\epsilon\) of \(x\), so \(x\) is a cluster point of \((x_n)\). Conversely, if \(x\) is a cluster point of \((x_n)\), then set \(k_1 = 1\) and inductively let \(k_{n+1} \geq k_n + 1\) for which \(\rho(x_{k_{n+1}}, x) < 1/(n+1)\). Then \(x_{k_n} \to x\).

A related result about subsequences has a somewhat complicated statement, but is frequently useful:

VI.2.3.8. Proposition. Let \((X, \rho)\) be a metric space, \(x \in X\), and \((x_n)\) a sequence in \(X\). Then \(x_n \to x\) if and only if \(x\) is a cluster point of every subsequence of \((x_n)\).

Proof: If \(x_n \to x\), then every subsequence \((x_{k_n})\) of \((x_n)\) also converges to \(x\) (), hence \(x\) is a cluster point of \((x_{k_n})\). Conversely, suppose \((x_n)\) does not converge to \(x\). Then there is an \(\epsilon > 0\) such that, for every \(n_0 \in \mathbb{N}\), there is an \(n \geq n_0\) with \(\rho(x_n, x) \geq \epsilon\). Let \(k_1 = 1\), and inductively choose \(k_{n+1} \geq k_n + 1\) with \(\rho(x_{k_{n+1}}, x) \geq \epsilon\). Then \(x\) is not a cluster point of the subsequence \((x_{k_n})\).

Limits of sequences and functions, continuity, sequential criterion, dependence on metric; cluster points; subsequences; closure in terms of sequences; uniform continuity; mean and mean-square convergence

Dominant and Equivalent Metrics

It is sometimes important to compare two or more metrics on the same set. The next definition is reminiscent of the definition of continuity, and is in fact very closely related ().

VI.2.3.9. Definition. Let \(\rho\) and \(\sigma\) be metrics on the same set \(X\). Then \(\rho\) dominates \(\sigma\) if, for every \(x_0 \in X\) and \(\epsilon > 0\), there is a \(\delta > 0\) such that \(\sigma(x, x_0) < \epsilon\) whenever \(x \in X\) and \(\rho(x, x_0) < \delta\). The metric \(\rho\) uniformly dominates \(\sigma\) if, for every \(\epsilon > 0\), there is a \(\delta > 0\) such that \(\sigma(x, y) < \epsilon\) whenever \(x, y \in X\) and \(\rho(x, y) < \delta\).

The metrics \(\rho\) and \(\sigma\) are equivalent if each dominates the other, and uniformly equivalent if each uniformly dominates the other.

The metrics \(\rho\) and \(\sigma\) are incomparable if neither one dominates the other.
VI.2.3.10. Roughly speaking, “ρ dominates σ” means that \( \sigma(x, y) \) is small whenever \( \rho(x, y) \) is small. The most common way a metric ρ uniformly dominates another metric σ is if there is a \( C > 0 \) with \( \sigma(x, y) \leq C\rho(x, y) \) for all \( x, y \in X \).

VI.2.3.11. Proposition. Let \( \rho \) and \( \sigma \) be metrics on a set \( X \). Then the following are equivalent:

(i) \( \rho \) dominates \( \sigma \).

(ii) The identity function from \((X, \rho)\) to \((X, \sigma)\) is continuous.

(iii) Whenever \((x_n)\) is a sequence in \( X \) converging to a limit \( x \) with respect to \( \rho \) (i.e. \( \lim_{n \to \infty} \rho(x_n, x) = 0 \)), \((x_n)\) also converges to \( x \) with respect to \( \sigma \) (i.e. \( \lim_{n \to \infty} \sigma(x_n, x) = 0 \)).

(iv) Every set which is \( \sigma \)-closed is also \( \rho \)-closed.

(v) Every set which is \( \sigma \)-open is also \( \rho \)-open.

Paraphrasing (iii)–(v), if \( \rho \) dominates \( \sigma \), it is easier for a set to be \( \rho \)-open (or \( \rho \)-closed) than \( \sigma \)-open (or \( \sigma \)-closed), but harder for a sequence to converge with respect to \( \rho \) than with respect to \( \sigma \).

Proof: (i) \( \iff \) (ii) is a straightforward comparison of definitions, and is left to the reader. (ii) \( \iff \) (iii) is the Sequential Criterion for continuity.

(iii) \( \Rightarrow \) (iv): Suppose \( A \subseteq X \) is \( \sigma \)-closed, and \( x \) is a \( \rho \)-limit point of \( A \). We must show that \( x \in A \). But by the Sequential Criterion for limit points, there is a sequence \((x_n)\) in \( A \) with \( x_n \to x \) with respect to \( \rho \). By (iii), \( x_n \to x \) with respect to \( \sigma \), so \( x \in A \) because \( A \) is \( \sigma \)-closed.

(iv) \( \iff \) (v) by taking complements. (v) \( \Rightarrow \) (ii) by ??.

Similarly, we obtain:

VI.2.3.12. Proposition. Let \( \rho \) and \( \sigma \) be metrics on a set \( X \). Then \( \rho \) uniformly dominates \( \sigma \) if and only if the identity function from \((X, \rho)\) to \((X, \sigma)\) is uniformly continuous.

VI.2.3.13. Examples.

(i) On \( \mathbb{R}^n \), the metrics \( \rho_2, \rho_1, \) and \( \rho_\infty \) are uniformly equivalent since, for any \( x, y \in \mathbb{R}^n \),

\[
\rho_\infty(x, y) \leq \rho_2(x, y) \leq \rho_1(x, y) \leq n\rho_\infty(x, y) \leq n\rho_2(x, y)
\]

(the inequality \( \rho_2(x, y) \leq \rho_1(x, y) \) is the only one which is not obvious; see Exercise ()).

(ii) On \( C[a, b] \), \( \rho_\infty \) uniformly dominates \( \rho_1 \) and \( \rho_2 \) since \( \rho_1(f, g) \leq (b - a)\rho_\infty(f, g) \) and \( \rho_2(f, g) \leq (b - a)^{1/2}\rho_\infty(f, g) \) for any \( f, g \). Actually, \( \rho_1(f, g) \leq (b - a)^{3/2}\rho_2(f, g) \) for any \( f, g \in C[a, b] \) (Exercise ()), so \( \rho_2 \) also uniformly dominates \( \rho_1 \).

However, none of the three are equivalent. For example, let \( f_n \) be the function whose graph is in Figure ().
Then
\[ \lim_{n \to \infty} \rho_1(f_n, 0) = \lim_{n \to \infty} \rho_2(f_n, 0) = 0, \]
but \( \rho_\infty(f_n, 0) = 1 \) for all \( n \). Apply the criterion of VI.2.3.11.(iii) to conclude that \( \rho_1 \) and \( \rho_2 \) do not dominate \( \rho_\infty \). See Exercise () for a proof that \( \rho_1 \) does not dominate \( \rho_2 \).

(iii) Let \( \rho \) be a metric on \( X \), and for \( x, y \in X \), set \( \sigma(x, y) = \min(\rho(x, y), 1) \). Then \( \sigma \) is a bounded metric on \( X \) which is uniformly equivalent to \( \rho \) [it is obvious that \( \rho \) uniformly dominates \( \sigma \); to prove that \( \sigma \) uniformly dominates \( \rho \), for any \( \epsilon > 0 \) choose \( \delta = \min(\epsilon, 1) \).

(iv) Let \( \rho \) be the usual metric on \( \mathbb{R} \), \( \rho(x, y) = |x - y| \), and let \( \sigma(x, y) = |\arctan x - \arctan y| \). Then \( \sigma \) is a metric on \( \mathbb{R} \), and \( \rho \) uniformly dominates \( \sigma \) (in fact, \( \sigma(x, y) \leq \rho(x, y) \) for all \( x, y \in \mathbb{R} \).) Conversely, \( \sigma \) dominates \( \rho \) by continuity of the tangent function, but not uniformly since \( \sigma(n, n + 1) \to 0 \) as \( n \to \infty \); so \( \rho \) and \( \sigma \) are equivalent but not uniformly equivalent.

(v) The discrete metric on a set \( X \) uniformly dominates any other metric on \( X \) [for any \( \epsilon > 0 \) we can choose \( \delta = 1 \).]

(vi) If \( \rho \) is a metric on \( X \), then \( c\rho \) is a metric which is uniformly equivalent to \( \rho \), for any \( c > 0 \).

(vii) If \( \rho \) and \( \sigma \) are metrics on a set \( X \), then \( \rho + \sigma \) (defined by \( (\rho + \sigma)(x, y) = \rho(x, y) + \sigma(x, y) \)) and \( \max(\rho, \sigma) \) are uniformly equivalent metrics which uniformly dominate both \( \rho \) and \( \sigma \). If \((\rho_n)\) is a sequence of metrics on \( X \), set \( \sigma_n = \min(\rho_n, 1) \) for all \( n \) (cf. (iii)), and define \( \rho \) by
\[ \rho(x, y) = \sum_{n=1}^{\infty} 2^{-n} \sigma_n(x, y). \]

Then \( \rho \) is a metric which uniformly dominates all the \( \rho_n \)'s.

732
VI.2.4. Extended Metrics

It turns out to be quite harmless to allow a metric to also take the value $\infty$ ($+\infty$), i.e. to allow points to be an infinite distance apart.

VI.2.4.1. **Definition.** An *extended metric* on a set $X$ is a function $\rho : X \times X \to [0, \infty]$ satisfying:

(i) $\rho(x, y) \geq 0$ for all $x, y \in X$.

(ii) $\rho(x, y) = 0$ if and only if $x = y$ (*definiteness*).

(iii) $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$ (*symmetry*).

(iv) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for all $x, y, z \in X$ (*triangle inequality*).

A pair $(X, \rho)$, where $X$ is a set and $\rho$ is a metric on $X$, is called an *extended metric space*.

In the definition, we take the usual order on $[0, \infty]$, i.e. $x < \infty$ for all $x \in \mathbb{R}$, and $x + \infty = \infty$ for any $x \in [0, \infty]$. Note that any metric is an extended metric. We may also consider extended pseudometrics which have all the properties except (ii).

VI.2.4.2. All the results of this section hold for extended metrics with no modifications. If $\rho$ is an extended metric on $X$, define $\tilde{\rho}$ by $\tilde{\rho}(x, y) = \min(\rho(x, y), 1)$; then $\tilde{\rho}$ is a metric on $X$ which is equivalent to $\rho$, so if desired any extended metric can be converted to a metric without changing the topology.

VI.2.4.3. **Definition.** Let $(X, \rho)$ be an extended metric space, and $x \in X$. The *galaxy* of $x$ in $X$ is

$$\{y \in X : \rho(x, y) < \infty\}.$$ 

The proof of the next simple proposition is left as an exercise for the reader.

VI.2.4.4. **Proposition.** Let $(X, \rho)$ be an extended metric space. Then

(i) Any two distinct galaxies in $X$ are disjoint. Thus $X$ partitions into a disjoint union of galaxies. Any points in distinct galaxies are an infinite distance apart.

(ii) Each galaxy is a clopen set in $X$.

(iii) As a topological space, $X$ is the separated union () of its galaxies.

(iv) The restriction of $\rho$ to any galaxy is a metric.

A metric space is just an extended metric space with only one galaxy.

Maximal Metrics

The next simple but important observation is most naturally phrased using extended metrics:
Proposition. Let $X$ be a set, $\{\rho_i : i \in I\}$ a collection of extended pseudometrics on $X$. For $x, y \in X$, set 
\[ \tilde{\rho}_I(x, y) = \sup_{i \in I} \rho_i(x, y). \]
Then $\tilde{\rho}_I$ is an extended pseudometric on $X$.

Proof: The only thing that needs to be checked is the triangle inequality. If $x, y, z \in X$, then, by (i), 
\[ \tilde{\rho}_I(x, z) = \sup_i \rho_i(x, z) \leq \sup_i [\rho_i(x, y) + \rho_i(y, z)] \leq \sup_i \rho_i(x, y) + \sup_i \rho_i(y, z) = \tilde{\rho}_I(x, y) + \tilde{\rho}_I(y, z). \]

VI.2.4.6. If the $\rho_i$ are separating (in particular, if at least one $\rho_i$ is an extended metric), then $\tilde{\rho}_I$ is an extended metric. The metric $\tilde{\rho}_I$ dominates each $\rho_i$; i.e., the $\tilde{\rho}_I$ topology is stronger than all the $\rho_i$ topologies, but it might not be the weakest topology stronger than all the $\rho_i$ topologies. For a simple example, let $\sigma$ be a metric on $X$, and for each $n$ let $\rho_n = n\sigma$. Then all the $\rho_n$ are equivalent to $\sigma$, but $\tilde{\rho}_I$ is the discrete extended metric. (For an actual metric version, set $\rho_n(x, y) = \min(1, n\sigma(x, y))$ for $x, y \in X$.) If $I$ is finite, then the $\tilde{\rho}_I$ topology is indeed the weakest topology stronger than all the $\rho_i$ topologies.

VI.2.4.7. A particularly important example (cf. XI.8.4.2.) comes from a bound function. Let $X$ be a set, and $\beta : X \times X \to [0, +\infty]$ a function. Let $I$ be the set of all pseudometrics $\sigma$ on $X$ such that $\sigma(x, y) \leq \beta(x, y)$ for all $x, y \in X$ ($I$ is nonempty since the zero pseudometric is in $I$; one could also allow extended pseudometrics without affecting the construction). Set $\rho_\beta = \tilde{\rho}_I$. The extended pseudometric $\rho_\beta$ is called the maximal pseudometric of $\beta$; it is the largest pseudometric $\rho$ satisfying $\rho(x, y) \leq \beta(x, y)$ for all $x, y \in X$.

VI.2.5. Cauchy Sequences

VI.2.6. Normed Vector Spaces

Many of the important metric spaces in analysis are vector spaces in which the metric comes from a norm.

VI.2.6.1. Definition. Let $X$ be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A norm on $X$ is a function $x \mapsto \|x\|$ from $X$ to $\mathbb{R}$ satisfying

(i) $\|x\| \geq 0$ for all $x \in X$.

(ii) $\|x\| = 0$ if and only if $x = 0$ (definiteness).

(iii) $\|\alpha x\| = |\alpha|\|x\|$ for all $x \in X$ and all scalars $\alpha$ (homogeneity).

(iv) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (triangle inequality).

VI.2.6.2. A norm on a vector space $X$ defines a metric on $X$ by $\rho(x, y) = \|x - y\|$. The metrics in VI.2.1.3.(i)–(iv) arise this way:
VI.2.6.3. Examples.

(i) The usual metric on \( \mathbb{R} \) comes from the norm \( \|x\| = |x| \).

(ii) The metric \( \rho_2 \) on \( \mathbb{R}^n \) comes from \( \|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \).

(iii) The metrics \( \rho_1 \) and \( \rho_\infty \) on \( \mathbb{R}^n \) also come from norms:
\[
\|x\|_1 = |x_1| + \cdots + |x_n| \\
\|x\|_\infty = \max(|x_1|, \ldots, |x_n|)
\]

(iv) The metrics \( \rho_2, \rho_1, \rho_\infty \) on \( C([a,b]) \) come from norms
\[
\|f\|_2 = \left[ \int_a^b (f(t))^2 \, dt \right]^{1/2} \\
\|f\|_1 = \int_a^b |f(t)| \, dt \\
\|f\|_\infty = \max_{a \leq t \leq b} |f(t)|
\]

(See Exercise (i) or (ii) for a proof that \( \| \cdot \|_2 \) satisfies the triangle inequality.)

(v) On \( \mathbb{C}^n \), the same formulas as in (ii) and (iii) define norms, except that the 2-norm must be written
\[
\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}.
\]

VI.2.6.4. The metric \( \rho \) induced by a norm is translation-invariant, i.e. \( \rho(x + z, y + z) = \rho(x, y) \) for all \( x, y, z \in X \), and homogeneous, i.e. \( \rho(\alpha x, \alpha y) = |\alpha| \rho(x, y) \) for all \( x, y \in X \) and all scalars \( \alpha \). A consequence of translation-invariance is that \( B_r(z) = z + B_r(0) \) for any \( r > 0, z \in X \), and \( B_{r \alpha}(0) = \alpha B_r(0) \) for any \( \alpha, r > 0 \). Thus all balls in a normed vector space “look the same” up to scaling and translation.

Another consequence of homogeneity is:

VI.2.6.5. Proposition. Let \( X \) be a vector space, \( \| \cdot \| \) and \( \| \cdot \|' \) norms on \( X \), \( \rho \) and \( \sigma \) the associated metrics. Then the following are equivalent:

(i) \( \rho \) dominates \( \sigma \).

(ii) \( \rho \) uniformly dominates \( \sigma \).

(iii) There is a \( C > 0 \) such that \( \|x\|' \leq C \|x\| \) for all \( x \in X \).
This is actually a special case of a fundamental principle about operators between normed vector spaces.

**Proof:** (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) are trivial, so we need only prove (i) $\Rightarrow$ (iii). Choose $\delta > 0$ so that $\|x\| = \rho(x, 0) < \delta$ implies $\|x\|' = \sigma(x, 0) < 1$, for any $x \in X$, and set $C = 1/\delta$. Then $\|x\| < 1$ implies that $\|x\|' < C$. If $x \in X$ and $\|x\|' > C\|x\|$, choose $\alpha \in \mathbb{R}$ with

$$\frac{C}{\|x\|'} < \alpha < \frac{1}{\|x\|}$$

and then $\|\alpha x\| < 1$ but $\|\alpha x\|' > C$, a contradiction; so there can be no such $x$, proving (iii).

[Move to later section]

**Theorem.** Let $X$ and $Y$ be vector spaces (over the same field, $\mathbb{R}$ or $\mathbb{C}$), with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively, and let $T : X \to Y$ be a linear operator. Then the following are equivalent:

(i) $T$ is continuous at 0.

(ii) $T$ is continuous.

(iii) $T$ is uniformly continuous.

(iv) There is a constant $K \geq 0$ such that $\|Tx\|_Y \leq K\|x\|_X$ for all $x \in X$.

If $T$ satisfies (iv), $T$ is said to be bounded; the smallest such $K$ is called the operator norm of $T$, denoted $\|T\|_{op}$ or usually just $\|T\|$.  

**Seminorms**

**Banach Spaces**

**VI.2.7. Exercises**

1. Here are physical interpretations of $\rho_1$ and $\rho_\infty$ on $\mathbb{R}^2$.

(a) The metric $\rho_1$ is often called the city metric or driving metric. Imagine a city in which all streets run north-south or east-west. Then the $\rho_1$-distance between two points located on streets is the shortest driving distance between the points (ignoring one-way streets, etc.)

(b) There is a similar, though perhaps less satisfactory, interpretation of $\rho_\infty$. In this city, there are buses on each street. The fare structure is unusual: the cost of riding a bus is strictly proportional to the distance traveled, say $1$ per mile. However, a rider also receives a transfer good for a free ride on a bus in the perpendicular direction, of length not exceeding the length of the paid ride. The $\rho_\infty$ distance between two points on streets is then the minimum cost of a bus ride between the two points.

2. Show that $\rho_1$, $\rho_2$, and $\rho_\infty$ are equivalent metrics on $\mathbb{R}^n$.

3. (a) If $x_1, \ldots, x_n$ are nonnegative real numbers, show that

$$\sum_{k=1}^{n} x_k^2 \leq \left(\sum_{k=1}^{n} x_k\right)^2.$$

[Hint: Multiply out the right side.] Give a careful proof by induction on $n$.  

736
(b) If $x_1, x_2$ are real numbers, show that $(x_1 + x_2)^2 \leq 2(x_1^2 + x_2^2)$. [Hint: $(x_1 - x_2)^2 \geq 0$.]

(c) If $x_1, \ldots, x_n$ are arbitrary real numbers, and $n = 2^m$, show by induction on $m$ that
\[
\left( \sum_{k=1}^{n} x_k \right)^2 \leq n \sum_{k=1}^{n} x_k^2.
\]

(d) Let $f$ be Riemann integrable on $[a, b]$ (e.g. $f \in C[a, b]$.) Show that
\[
\left( \int_a^b |f(t)| \, dt \right)^2 \leq (b - a) \int_a^b |f(t)|^2 \, dt .
\]
[Hint: approximate the integrals by Riemann sums, dividing the interval into $2^n$ subintervals of equal length, and apply (c).]

(e) Show that the inequality in (c) is also true if $n$ is not a power of 2. [Given $x_1, \ldots, x_n$, let $f$ be the function on $[0, n]$ with $f(t) = x_k$ if $k - 1 \leq x < k$. Apply (d).]

The inequalities in (b)–(e) are special cases of Jensen’s Inequality. See also VI.2.7. for an alternate proof of (c) for general $n$.

4. (a) Show that on $C[a, b]$, the $\infty$-norm dominates the 2-norm, and the 2-norm dominates the 1-norm.

(b) Let $g_n$ be the function whose graph is given in Figure (a), and $h_n = \sqrt{g_n}$.

\[\text{Figure VI.2: The graph of } g_n\]

Then $\rho_1(h_n, 0) < n^{-1/2}$, so $\lim_{n \to \infty} \rho_1(h_n, 0) \to 0$. However, for any $n$,
\[
\rho_2(h_n, 0) = \left( \int_a^b g_n(t) \, dt \right)^{1/2} = \frac{1}{\sqrt{2}}.
\]
so $\rho_2(h_n, 0) \neq 0$.

5. Give a direct proof (not involving taking complements) that the collection of closed sets in a metric space is closed under arbitrary intersections and under finite unions.

The graph of the Cantor function is given in Figure VI.4.

![The Cantor function graph](image-url)

Figure VI.3: The Cantor function

[Move to later section]

6. Let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be real numbers. Show that

$$(x_1 + \cdots + x_n)(y_1 + \cdots + y_n) \leq n \sqrt{x_1^2 + \cdots + x_n^2} \sqrt{y_1^2 + \cdots + y_n^2}.$$  

[Hint: let $\mathbf{x} = (x_1, \ldots, x_n)$, and for $1 \leq k \leq n$ let

$$y^{(k)} = (y_k, y_k+1, \ldots, y_n, y_1, \ldots, y_{k-1}).$$

Then $(x_1 + \cdots + x_n)(y_1 + \cdots + y_n) = \mathbf{x} \cdot \mathbf{y}^{(1)} + \cdots + \mathbf{x} \cdot \mathbf{y}^{(n)}$. Apply the CBS inequality.] In particular, $(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$. The $n$ in this inequality is the best possible constant in general: if all the $x_i$ and $y_j$ are equal, we have equality.
Figure VI.4: The Cantor Function
VI.2.8. How Many Separable Metric Spaces Are There?

In this subsection, we answer the following two questions:

VI.2.8.1. Question. How many separable metric spaces are there?

VI.2.8.2. Question. How many separable complete metric spaces are there?

VI.2.8.3. These questions are meaningless unless we specify a notion of equivalence. After all, there are as many one-point metric spaces as there are sets! (For a similar reason, it is meaningless to ask how many metric spaces there are, even up to isometry: any set with the discrete metric is a (complete) metric space.)

The two obvious notions of equivalence are homeomorphism and isometry. Isometry is a much finer equivalence: each homeomorphism class of metric spaces (with more than one point) includes many isometry classes. Thus we will obtain, and answer, four questions by inserting the phrase “up to homeomorphism” or “up to isometry” in each. The answers, however, will turn out not to depend on which equivalence we take: the answer to VI.2.8.1. in both cases is $2^\aleph_0$, and to VI.2.8.2. in both cases is $2^\aleph_0$.

VI.2.8.4. Proposition. There are exactly $2^\aleph_0$ functions from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{R}$.

VI.2.8.5. Corollary. There are at most $2^\aleph_0$ metrics on $\mathbb{N}$. Hence up to isometry there are at most $2^\aleph_0$ countable metric spaces.

In fact, it is easily seen that there are at least $2^\aleph_0$ metrics on $\mathbb{N}$ (the distance between 1 and 2 can be any positive number), hence exactly $2^\aleph_0$. (One should also consider finite metric spaces: there are clearly exactly $2^\aleph_0$ metrics on a finite set with more than one element.)

VI.2.8.6. Corollary. Up to isometry, there are at most $2^\aleph_0$ separable complete metric spaces.

Proof: Every separable complete metric space is isometric to the completion of a countable metric space, and the completion of a metric space is unique up to isometry.

VI.2.8.7. There are $2^\aleph_0$ topologically distinct compact subsets of $\mathbb{R}^2$ (XI.18.4.1.; there are even $2^\aleph_0$ topologically distinct compact subsets of $\mathbb{R}$ – see XI.18.4.2.), each of which is a separable complete metric space. Thus up to homeomorphism, there are at least $2^\aleph_0$ compact metric spaces, each of which is separable and complete (). Thus we obtain the answer to Question VI.2.8.2.:
VI.2.8.8. THEOREM. (i) Up to homeomorphism, there are exactly $2^{\aleph_0}$ compact metric spaces.
(ii) Up to isometry, there are exactly $2^{\aleph_0}$ compact metric spaces.
(iii) Up to homeomorphism, there are exactly $2^{\aleph_0}$ separable complete metric spaces.
(iv) Up to isometry, there are exactly $2^{\aleph_0}$ separable complete metric spaces.

Recall () that a Polish space is a topological space which is homeomorphic to a separable complete metric space. Thus we can rephrase (iii):

VI.2.8.9. COROLLARY. Up to homeomorphism, there are exactly $2^{\aleph_0}$ Polish spaces.

We can now answer Question VI.2.8.1.:

VI.2.8.10. THEOREM. (i) Up to homeomorphism, there are exactly $2^{2^{\aleph_0}}$ separable metric spaces.
(ii) Up to isometry, there are exactly $2^{2^{\aleph_0}}$ separable metric spaces.

Proof: Every separable metric space is isometric to a subset of a separable complete metric space. Every separable complete metric space has cardinality $\leq 2^{\aleph_0}$ (XI.7.8.5.(c)) (in fact, exactly $2^{\aleph_0}$ if uncountable; cf. XII.3.1.10.). Thus a separable complete metric space has at most $2^{2^{\aleph_0}}$ subsets. Since there are $2^{\aleph_0}$ separable complete metric spaces up to isometry (VI.2.8.8.), there are up to isometry at most $2^{\aleph_0} \cdot 2^{2^{\aleph_0}} = 2^{2^{\aleph_0}}$ subsets of complete metric spaces. Hence there are at most $2^{2^{\aleph_0}}$ separable metric spaces up to isometry.

Conversely, there are $2^{2^{\aleph_0}}$ topologically distinct subsets of $\mathbb{R}$ (XI.18.4.42.), each of which is a separable metric space. Thus there are at least $2^{2^{\aleph_0}}$ separable metric spaces up to homeomorphism.

\textcopyright
VI.2.9. The Banach Fixed-Point Theorem

There are many fixed-point theorems in topology, which assert that under certain conditions a function from a topological space to itself sends at least one, or exactly one, point to itself. Perhaps the simplest example is (); other examples are XI.20.1.2. Here we discuss a fairly elementary but very useful example, the Banach fixed-point theorem, sometimes called the Contraction Mapping Theorem.

VI.2.9.1. Definition. Let \((X, \rho)\) be a metric space. A function \(T : X \to X\) is a strict contraction if there is an \(\alpha, 0 < \alpha < 1\), such that \(\rho(T(x), T(y)) \leq \alpha \rho(x, y)\) for all \(x, y \in X\).

A strict contraction is obviously a (uniformly) continuous function from \(X\) to \(X\).

VI.2.9.2. Caution: For \(T\) to be a strict contraction, it is not sufficient that \(\rho(T(x), T(y)) < \rho(x, y)\) whenever \(x \neq y\); the distance must decrease by at least a constant factor less than 1. See VI.2.9.6.

VI.2.9.3. Theorem. [Banach Fixed-Point Theorem] Let \((X, \rho)\) be a complete metric space, and \(T : X \to X\) a strict contraction. Then \(T\) has a unique fixed point: there is a unique point \(p \in X\) with \(T(p) = p\). If \(x\) is any point of \(X\), then the sequence \((T^n(x))\) converges to \(p\).

Proof: Fix \(x \in X\). Inductively set \(x_0 = x\) and \(x_{n+1} = T(x_n)\) for all \(n\). We claim \((x_n)\) is a Cauchy sequence. By induction, we have \(\rho(x_n, x_{n+1}) \leq \alpha^n \rho(x_0, x_1)\) for all \(n\). Thus, if \(n < m\), by the triangle inequality we have

\[
\rho(x_n, x_m) \leq \rho(x_0, x_1) \sum_{k=n}^{m-1} \alpha^k.
\]

The sum is a tail sum of the convergent geometric series \(\sum_{k=1}^{\infty} \alpha^k\), hence is arbitrarily small if \(n\) is large enough. Thus \((x_n)\) is a Cauchy sequence and therefore converges to a point \(p \in X\). Then \(T(x_n) \to T(p)\) since \(T\) is continuous; but \(T(x_n) = x_{n+1}\), so \(T(x_n) \to p\), i.e. \(T(p) = p\) by uniqueness of limits in a metric space.

If \(q\) is any fixed point of \(T\), then

\[
\rho(p, q) = \rho(T(p), T(q)) \leq \alpha \rho(p, q)
\]

which can only hold if \(\rho(p, q) = 0\), i.e. \(q = p\). Thus \(p\) is the only fixed point of \(T\).

VI.2.9.4. The proof of the theorem is constructive: the sequence of iterates of any starting point converges to the fixed point, i.e. the fixed point is a “universal attractor.” Thus the fixed point can be found by an explicit iterative algorithm of successive approximation.

VI.2.9.5. If \((X, \rho)\) is not complete, a strict contraction on \(X\) need not have a fixed point (VI.2.10.1). However, if it does, the argument of the proof shows that the fixed point is unique (this part of the proof does not require completeness). And if \(p\) is a fixed point of \(T\), then for any \(x \in X\) we have \(\rho(T^n(x), p) = \rho(T^n(x), T^n(p)) \leq \alpha^n \rho(x, p) \to 0\), so \(T^n(x) \to p\). Alternatively, any strict contraction on \((X, \rho)\) extends uniquely to a strict contraction \(\bar{T}\) on the completion \((\bar{X}, \bar{\rho})\) of \((X, \rho)\) by uniform continuity (); if \(T\) has a fixed point, it must be the unique fixed point of \(\bar{T}\) and we have \(T^n(x) \to p\) for any \(x \in X\).
VI.2.9.6. Example. To obtain a fixed point, it does not suffice to simply have \( \rho(T(x), T(y)) < \rho(x, y) \) for \( x \neq y \); this condition is not enough to insure that \( (T^n(x)) \) is a Cauchy sequence for any \( x \). For example, let \( T : [1, +\infty) \to [1, +\infty) \) be defined by \( T(x) = x + \frac{1}{x} \). Then \( T \) is differentiable everywhere and \( 0 < T'(x) < 1 \) for \( 1 < x < +\infty \), so by the MVT \( |T(x) - T(y)| < |x - y| \) for \( x, y \in [1, +\infty) \), \( x \neq y \) (this also has an elementary proof not requiring calculus). But \( T \) has no fixed point.

Continuity of Fixed Points

If two strict contractions are uniformly close, their fixed points must be close:

VI.2.9.7. Proposition. Let \( (X, \rho) \) be a metric space, and \( S, T \) strict contractions on \( X \), i.e. there are \( 0 \leq \alpha, \beta < 1 \) such that \( \rho(S(x), S(y)) \leq \alpha \rho(x, y) \) and \( \rho(T(x), T(y)) \leq \beta \rho(x, y) \) for all \( x, y \in X \). Let \( p \) and \( q \) be (unique) fixed points of \( S \) and \( T \) respectively. If \( S \) and \( T \) are uniformly close within \( \epsilon > 0 \), i.e. \( \rho(S(x), T(x)) < \epsilon \) for all \( x \in X \), then \( \rho(p, q) < \frac{\epsilon}{1 - \gamma} \), where \( \gamma = \min(\alpha, \beta) \).

Proof: We have \( \lim_{n \to \infty} S^n(q) = p \), and by the triangle inequality we have, for each \( n \),

\[
\rho(q, S^n(q)) = \rho(q, S(q)) + \rho(S(q), S^2(q)) + \cdots + \rho(S^{(n-1)}(q), S^n(q)) \\
\leq (1 + \alpha + \cdots + \alpha^{n-1})\rho(q, S(q)) = \frac{1 - \alpha^n}{1 - \alpha}\rho(q, S(q)) < \frac{1}{1 - \alpha}\rho(q, S(q))
\]

and thus, letting \( k \to \infty \),

\[
\rho(q, p) \leq \frac{1}{1 - \alpha}\rho(q, S(q)) = \frac{1}{1 - \alpha}\rho(T(q), S(q)) < \frac{\epsilon}{1 - \alpha}.
\]

Similarly, we have \( \rho(p, q) = \lim_{n \to \infty} \rho(p, T^n(p)) < \frac{\epsilon}{1 - \gamma} \).

Strict Power Contractions and the Extended Banach Fixed-Point Theorem

There is a substantial extension of the Banach Fixed-Point Theorem.

VI.2.9.8. Definition. Let \( (X, \rho) \) be a metric space. A function \( T : X \to X \) is a strict power contraction if \( T^n \) is a strict contraction for some \( n \in \mathbb{N} \).

VI.2.9.9. Example. A strict power contraction is not necessarily a strict contraction. It is not necessarily even continuous! Let \( X = \mathbb{R} \) with its usual metric, and define \( T : \mathbb{R} \to \mathbb{R} \) by

\[
T(x) = \begin{cases} 
\frac{x}{2} & \text{if } x \in \mathbb{Q} \\
-x/2 & \text{if } x \notin \mathbb{Q}
\end{cases}
\]

Then \( T \) is not continuous, but \( T^2 \) is a strict contraction, so \( T \) is a strict power contraction. Note that \( T \) has a unique fixed point 0, and is continuous at 0 (and only at 0).
VI.2.9.10.  **Theorem.** [Extended Banach Fixed-Point Theorem] Let \((X, \rho)\) be a complete metric space, and \(T : X \to X\) a strict power contraction. Then \(T\) has a unique fixed point \(p\). If \(x\) is any point of \(X\), then the sequence \((T^n(x))\) converges to \(p\).

**Proof:** Suppose \(T^n\) is a strict contraction. Then \(T^n\) has a unique fixed point \(p\) by Theorem VI.2.9.3. We have
\[
T^n(T(p)) = T^{n+1}(p) = T(T^n(p)) = T(p)
\]
so \(T(p)\) is also a fixed point of \(T^n\); thus \(T(p) = p\) by uniqueness, and \(p\) is a fixed point of \(T\). If \(q\) is any fixed point of \(T\), then \(q\) is also a fixed point of \(T^n\), hence \(q = p\) by uniqueness, so \(T\) has a unique fixed point.

Fix \(x \in X\). Then \(\lim_{k \to \infty} T^{kn}(x) = p\). Also, for each \(j\), \(T^{kn+j}(x) = T^{kn}(T^j(x)) \to p\) as \(k \to \infty\). It follows that \(\lim_{m \to \infty} T^m(x) = p\). \(\blacksquare\)

VI.2.9.11.  If \((X, \rho)\) is not complete, a strict power contraction on \(X\) need not have a fixed point (VI.2.10.1). However, if it does, the fixed point is unique (VI.2.9.5), and the proof shows that \(T^n(x) \to p\) for any \(x \in X\) (this part of the proof does not require completeness). The alternate argument of VI.2.9.5 does not work, however, since in general there is no reasonable way to extend a power contraction to the completion.

VI.2.10.  **Exercises**

VI.2.10.1.  Let \(X = \mathbb{Q} \cap [1, 2]\) with the usual metric, and define \(T : X \to X\) by
\[
T(x) = \frac{1}{2} \left( x + \frac{2}{x} \right).
\]

(a) Show that \(T\) is a strict contraction on \(X\).
(b) Show that \(T\) has no fixed point in \(X\).
(c) Identify \(X\) with \([1, 2]\) with the usual metric, and \(\bar{T}\) with the function from \(X\) to \(\bar{X}\) defined by the same formula as \(T\). Find the fixed point of \(\bar{T}\).

VI.2.10.2.  A strict power contraction does not have to be continuous (VI.2.9.9). If it has a fixed point (e.g. if the metric space is complete), does it have to be continuous at the fixed point? Give a proof or counterexample.

VI.2.10.3.  Discuss the difficulties in trying to extend VI.2.9.7. to power contractions. Is there any reasonable result?
VI.3. Euclidean Space

VI.3.1. Metric Axiomatization of Euclidean Space

There are several approaches one can take to defining \( n \)-dimensional Euclidean space. We can just define it to be \( \mathbb{R}^n \) (with the standard metric \( \rho_2 \)). We will eventually take this approach, but it is worth noting that Euclidean space can be defined in a more abstract manner and shown to be effectively the same as \( \mathbb{R}^n \).

There are two such approaches possible, characterizing \( \mathbb{R}^n \) directly as a metric space, and giving a set of purely geometric axioms. Since this is a book about analysis, we will take the metric space approach, which may be rather unsatisfying to geometers. The metric approach is, however, quite elegant (although some of the proofs are rather involved), and not as well known as it should be. Our exposition is adapted from [Blu70] (cf. also [Blu75]), where the history of various metric approaches is discussed. For the purely geometric approach, see [Gre93], [Har00], and/or [Moi90].

VI.3.1.1. One easy metric approach is to define \( n \)-dimensional Euclidean space to be a metric space isometric to \( \mathbb{R}^n \). Since the isometries of \( \mathbb{R}^n \) can be nicely described (XV.1.2.4.), and preserve the affine structure, an isometry from an abstract metric space \( E \) to \( \mathbb{R}^n \) is almost unique (up to a choice of origin and an ordered orthonormal basis) and can be used to transfer the structure of \( \mathbb{R}^n \) to \( E \). We will regard a choice of an isometry of \( E \) with \( \mathbb{R}^n \) to be a choice of a coordinate system in \( E \). By choosing a coordinate system, we can identify \( E \) with \( \mathbb{R}^n \) and all analysis we do on \( E \) can be described using the chosen coordinates; it will all turn out to be independent of the coordinate system chosen in an appropriate sense ( ).

VI.3.1.2. We can, however, characterize \( n \)-dimensional Euclidean space abstractly as a metric space by three axioms (VI.3.1.43.), or four simpler axioms plus completeness (VI.3.1.44.). These axioms will presuppose having \( \mathbb{R} \) available by some previous construction; this is undesirable from a geometric point of view, but necessary from the metric standpoint (indeed, the definition of a metric space itself depends on already having \( \mathbb{R} \), or at least \( \mathbb{R}_+ \)). \( \mathbb{R} \) is always assumed to have its usual metric.

We will give the three axioms for a metric space \( (E, \rho) \), and discuss some variations and equivalent versions of them.

The Line Axiom

VI.3.1.3. Definition. A subset \( L \) of a metric space \( (E, \rho) \) is a line if \( L \) is isometric to \( \mathbb{R} \). A subset of \( E \) isometric to a closed bounded interval \( [\alpha, \beta] \) in \( \mathbb{R} \) is called a line segment in \( E \), with the points corresponding to \( \alpha \) and \( \beta \) the endpoints. A subset of \( E \) isometric to \( [0, \infty) \) is called a ray in \( E \), beginning at (or with endpoint) the point corresponding to \( 0 \).

VI.3.1.4. Line Axiom: For any two points \( a, b \in E \) there is a line in \( E \) containing \( a \) and \( b \).

Such a subset is called a line through \( a \) and \( b \). Note that no uniqueness is required as part of the axiom (uniqueness will follow from the second axiom; cf. VI.3.1.27.). A metric space in which each pair of distinct points lies on a unique line will be said to satisfy the Unique Line Axiom. Similarly:

VI.3.1.5. Line Segment Axiom: For any two (distinct) points \( a, b \in E \) there is a line segment in \( E \) with endpoints \( a \) and \( b \).
A metric on a set which satisfies the Line Segment Axiom is called a *strictly intrinsic metric*. See () for an explanation of this terminology from metric geometry. The Line Axiom is more important than the Line Segment Axiom in this section, but the Line Segment Axiom is central in metric geometry ().

**VI.3.1.6. Examples.** (i) \((\mathbb{R}^2, \rho_\infty)\) () satisfies the Line Axiom (as does any normed real vector space viewed as a metric space). But lines through two given points are not unique in general: for example, the graph of any Lipschitz function from \(\mathbb{R}\) to \(\mathbb{R}\) () with Lipschitz constant \(\leq 1\) (contractive function), regarded as a subset of \(\mathbb{R}^2\) in the usual way, is a line in this metric space.

(ii) Any convex set in \(\mathbb{R}^n\) (or any normed vector space) satisfies the Line Segment Axiom. Again the line segment need not be unique.

(iii) A circle with the arc length metric satisfies the Line Segment Axiom, as does a 2-sphere with the great circle arc length metric. The line segment is not unique for diametrically opposed points. They do not satisfy the Line Axiom.

If the metric space \((E, \rho)\) is assumed to be complete, an apparently much weaker assumption implies the Line Axiom. We first make a definition:

**VI.3.1.7. Definition.** Let \(a, c \in E\). A point \(b \in E\) is *between* \(a\) and \(c\) if \(b \neq a, b \neq c\, \text{and} \, \rho(a, c) = \rho(a, b) + \rho(b, c)\).

In \(\mathbb{R}\), \(b\) is between \(a\) and \(c\) in this sense if and only if \(b\) is between \(a\) and \(c\) with respect to the usual ordering on \(\mathbb{R}\). More generally, if \(L\) is a line in a metric space, then the various isometries of \(L\) with \(\mathbb{R}\) give two orderings on \(L\), and if \(a, b, c \in L\), then \(b\) is between \(a\) and \(c\) in the sense of VI.3.1.7. if and only if \(b\) is between \(a\) and \(c\) with respect to either (both) of these orderings. In particular, if \(a, b, c\) are distinct points on a line in a metric space, exactly one of the points is between the other two.

Note, however, that the set of points between two given points in a metric space need not look anything like a line segment (cf. Exercise VI.3.3.1); indeed, this set can easily be empty or finite.

There is a limited sense in which betweenness is transferable:

**VI.3.1.8. Proposition.** Let \(a, b, c, d\) be points in a metric space \((E, \rho)\). If \(b\) is between \(a\) and \(d\) and \(c\) is between \(b\) and \(d\), then \(c\) is between \(a\) and \(d\) and \(b\) is between \(a\) and \(c\).

**Proof:** We have
\[
\rho(a, d) = \rho(a, b) + \rho(b, d) = \rho(a, b) + \rho(b, c) + \rho(c, d)
\]
and since \(\rho(a, d) \leq \rho(a, c) + \rho(c, d)\) by the triangle inequality, we have
\[
\rho(a, c) \geq \rho(a, b) + \rho(b, c)
\]
by subtraction; the opposite inequality holds by the triangle inequality, so
\[
\rho(a, c) = \rho(a, b) + \rho(b, c)
\]
and \(c\) is between \(a\) and \(b\). Then we have
\[
\rho(a, d) = \rho(a, b) + \rho(b, c) + \rho(c, d) = \rho(a, c) + \rho(c, d)
\]
so \(c\) is between \(a\) and \(d\).  

746
VI.3.1.9. Care must be exercised in combining betweenness relations, however; overlapping relations cannot be combined in general (betweenness is not “transitive”). For example, if $(E;\rho) = ((0,0), (1,1), (2,0), (1,-1))$, then $b$ is between $a$ and $c$ and $c$ is between $b$ and $d$, but neither $b$ nor $c$ is between $a$ and $d$.

There is a useful consequence of VI.3.1.8.: 

VI.3.1.10. Corollary. Let $(E;\rho)$ be a metric space, $a,b,c \in E$, $S_1$ a line segment with endpoints $a$ and $b$, and $S_2$ a line segment with endpoints $b$ and $c$. Then $S = S_1 \cup S_2$ is a line segment with endpoints $a$ and $c$ if and only if $b$ is between $a$ and $c$.

Proof: If $S$ is a line segment, then $b$ is clearly between $a$ and $c$. Conversely, if $b$ is between $a$ and $c$, to show $S$ is a line segment it suffices to show that if $x \in S_1$, $y \in S_2$, $x \neq b$, $y \neq b$, then $\rho(x,y) = \rho(x,b) + \rho(b,y)$, i.e. that $b$ is between $x$ and $y$. Apply VI.3.1.8. to $a,b,y,c$ to conclude that $b$ is between $a$ and $y$, and then apply VI.3.1.8. again to $y,b,x,a$ to conclude that $b$ is between $y$ and $x$. ☐

VI.3.1.11. Between Axiom: If $a,c \in E$, $a \neq c$, then there is a $b \in E$ which is between $a$ and $c$.

VI.3.1.12. Extension Axiom: If $a,b \in E$, $a \neq b$, then there is a $c \in E$ such that $b$ is between $a$ and $c$.

In metric geometry (e.g. in [Blu70]), a metric space satisfying the Between Axiom is often called metrically convex, and a metric space satisfying the Extension Axiom is called externally convex. We have not used these terms since they do not seem appropriate in general: for example, the subset $[0,1) \cup \{2\}$ of $\mathbb{R}$ satisfies the Between Axiom, and $\mathbb{Z}$ satisfies the Extension Axiom, but neither resembles a convex set in the usual sense (). However, the two axioms together, along with completeness, are quite powerful:

VI.3.1.13. Theorem. Let $(E,\rho)$ be a complete metric space satisfying the Between Axiom and the Extension Axiom. Then $(E,\rho)$ satisfies the Line Axiom.

Completeness is essential for this result: the subsets $\mathbb{Q}$ and $(0,1)$ of $\mathbb{R}$ each satisfy both the Between Axiom and the Extension Axiom, but not the Line Axiom. Note also that even for complete metric spaces, neither axiom implies the other: consider $[0,1]$ and $\mathbb{Z}$.

The proof of VI.3.1.13. consists of three lemmas.

VI.3.1.14. Lemma. Let $(E,\rho)$ be a complete metric space satisfying the Between Axiom. Then $(E,\rho)$ satisfies the Line Segment Axiom.

Proof: We give a simple proof using Zorn’s Lemma. A somewhat more complicated proof can be given which requires only the Axiom of Dependent Choice; cf. [Blu70, Theorem 14.1] and Exercise VI.3.3.8..

Let $\mathcal{X}$ be the set of all subsets of $E$ containing $a$ and $b$ and isometric to a subset of $[0,\lambda]$, where $\lambda = \rho(a,b)$, ordered by inclusion. (Note that $\{a,b\} \in \mathcal{X}$, so $\mathcal{X}$ is nonempty.) For every set $X \in \mathcal{X}$, there is a unique
isometry \( \phi_X \) from \( X \) to \([0, \lambda]\) with \( \phi_X(a) = 0 \) and \( \phi_X(b) = \lambda \). If \( \{X_i\} \) is a chain in \( X \), and \( X = \bigcup X_i \), define \( \phi_X : X \to [0, 1] \) by \( \phi_X(x) = \phi_X(x') \) for \( x \in X_i \); this is well defined by uniqueness of the \( \phi_X \): if \( X_i \subseteq X_j \), then \( \phi_X|_{X_i} = \phi_X |_{X_i} \). Then \( \phi_X \) is an isometry since if \( x, y \in X \), then \( x, y \in X_i \) for some \( i \).

If \( X \in \mathcal{X} \), then since \( \phi_X \) is an isometry it extends to an isometry from the closure \( \overline{X} \) in \( E \) to \([0, \lambda]\), and hence \( \overline{X} \in \mathcal{X} \).

Let \( X \) be a maximal element of \( \mathcal{X} \). Then \( X \) is closed in \( E \), and hence complete; thus \( K = \phi_X(X) \) is complete, hence closed in \([0, \lambda]\). If \( K \neq [0, \lambda] \), there is an interval \([\alpha, \beta] \subseteq [0, \lambda] \) with \( \alpha, \beta \in K \) and \((\alpha, \beta) \cap K = \emptyset \). Let \( c, d \in X \) with \( \phi_X(c) = \alpha \), \( \phi_X(d) = \beta \). There is a \( p \in E \) between \( c \) and \( d \). Since \( c \) and \( d \) are between \( a \) and \( b \), by VI.3.1.8. we have \( p \) between \( a \) and \( b \). Then \( p \notin X \) since \( \rho(a, p) \in (\alpha, \beta) \). If \( x \) is any element of \( X \), then \( x \) is either between \( a \) and \( c \) or between \( d \) and \( b \); it follows from VI.3.1.8. that if \( x, y \in X \), then one of \( x, y, p \) is between the other two. Thus \( \phi_X \) extends to an isometry of \( X \cup \{p\} \) to \([0, \lambda]\), contradicting maximality of \( X \). Thus \( K = [0, \lambda] \) and \( X \) is a line segment with endpoints \( a \) and \( b \). \( \diamond \)

Now we turn to the procedure of extending line segments.

**VI.3.1.15. Lemma.** Let \((E, \rho)\) be a complete metric space satisfying the Extension Axiom, and let \( a, b \in E \). Then

\[
\sup \{\rho(a, c) : c \in E, b \text{ is between } a \text{ and } c \} = +\infty.
\]

**Proof:** Suppose \( \lambda_1 = \sup \{\rho(a, c) : c \in E, b \text{ is between } a \text{ and } c \} \) is finite. Let \( c_1 \in E \) with \( b \) between \( a \) and \( c_1 \) and \( \rho(a, c_1) > \lambda_1 - 1 \). If \( c \in E \) and \( c_1 \) is between \( a \) and \( c \), then \( b \) is between \( a \) and \( c \) by VI.3.1.8. Thus \( \lambda_2 = \sup \{\rho(a, c) : c \in E, c_1 \text{ is between } a \text{ and } c \} \) is finite and \( \lambda_2 \leq \lambda_1 \). Inductively let

\[
\lambda_{n+1} = \sup \{\rho(a, c) : c \in E, c_n \text{ is between } a \text{ and } c \} \leq \lambda_n
\]

and let \( c_{n+1} \) be a point of \( E \) with \( c_n \) between \( a \) and \( c_{n+1} \) (hence by VI.3.1.8. \( b \) is between \( a \) and \( c_{n+1} \)) and \( \rho(a, c_{n+1}) > \lambda_{n+1} - \frac{1}{n+1} \).

Set \( \lambda = \inf \lambda_n = \lim \lambda_n \). If \( n < m \), we have \( c_n \) between \( a \) and \( c_m \), hence

\[
\lambda_n - \frac{1}{n} < \rho(a, c_n) < \rho(a, c_m) \leq \lambda_n
\]

so \( \rho(c_n, c_m) < \frac{1}{n} \) for \( n < m \) and thus \( (c_n) \) is a Cauchy sequence in \( E \). By completeness, \( (c_n) \) converges to some \( c \in E \), and we have

\[
\lambda_n - \frac{1}{n} < \rho(a, c) \leq \lambda_n
\]

for all \( n \); thus \( \rho(a, c) = \lambda \) by the Squeeze Theorem (). By continuity, \( b \) and \( c_n \) are between \( a \) and \( c \) for all \( n \).

Choose \( d \in E \) with \( c \) between \( a \) and \( d \). Then \( \rho(a, d) > \lambda \), so \( \rho(a, d) > \lambda_n \) for some \( n \). But \( c_n \) is between \( a \) and \( d \), so this contradicts the definition of \( \lambda_n \). Thus the assumption that \( \lambda_1 < +\infty \) is false. \( \diamond \)
VI.3.1.16.  **Lemma.** Let \((E, \rho)\) be a complete metric space satisfying the Between and Extension Axioms. Then

(i) \((E, \rho)\) satisfies the Line Axiom.

(ii) Every line segment in \(E\) can be extended to a line in \(E\).

(iii) If \(a, b \in E\), and \(x_1, \ldots, x_n\) are points of \(E\) between \(a\) and \(b\), with \(x_k\) between \(a\) and \(x_{k+1}\) for \(1 \leq k \leq n-1\), then there is a line in \(E\) containing \(a\), \(b\), and all the \(x_k\).

**Proof:** For (i), let \(a, b \in E\). By Lemma VI.3.1.14., there is a line segment \(S_0\) with endpoints \(a\) and \(b\). For (ii), let \(S_0\) be a line segment in \(E\), and let \(a\) and \(b\) be its endpoints. For (iii), there is a line segment with endpoints \(a\) and \(x_1\), for each \(k < n\) a line segment with endpoints \(x_k\) and \(x_{k+1}\), and a line segment with endpoints \(x_n\) and \(b\). By repeated use of VI.3.1.10., the union of these segments is a line segment \(S_0\) with endpoints \(a\) and \(b\) containing all the \(x_k\).

In each case, begin with the segment \(S_0\). By Lemma VI.3.1.15. there is a \(c_1 \in E\) with \(b\) between \(a\) and \(c_1\) and \(\rho(a, c_1) \geq 2\rho(a, b)\). By Lemma VI.3.1.14. there is a line segment \(S_1\) in \(E\) with endpoints \(b\) and \(c_1\). Since \(b\) is between \(a\) and \(c_1\), \(S_0 \cup S_1\) is a line segment with endpoints \(a\) and \(c_1\) by VI.3.1.10.. Now let \(c_2 \in E\) with \(a\) between \(c_1\) and \(c_2\) and \(\rho(c_1, c_2) \geq 2\rho(a, c_1)\), and let \(S_2\) be a line segment with endpoints \(a\) and \(c_2\). Then \(S_0 \cup S_1 \cup S_2\) is a line segment in \(E\) with endpoints \(c_1\) and \(c_2\). Now let \(c_3 \in E\) with \(c_1\) between \(c_2\) and \(c_3\) and \(\rho(c_2, c_3) \geq 2\rho(c_1, c_2)\), and let \(S_3\) be a line segment between \(c_1\) and \(c_3\). Continue inductively placing \(c_{n+1}\) at the \(c_{n-1}\) end of the already constructed segment with endpoints \(c_{n-1}\) and \(c_n\). Then \(\cup_{k=0}^{n+1} S_k\) is a line segment between \(c_n\) and \(c_{n+1}\) extending the previous segment, and the length of this segment is at least \(2^{n+1}\rho(a, b)\). The union of the \(S_n\) is a line in \(E\) through \(a\) and \(b\). 

This completes the proof of Theorem VI.3.1.13.

Here are variations of the Between and Extension Axioms which are sometimes useful:

VI.3.1.17. **Definition.** If \(a, b \in E\), then \((a, b)\) has the Interpolation Property if for each \(t \in [0, 1]\) there is a point \(c \in E\) such that

\[\rho(a, c) = t\rho(a, b) \quad \text{and} \quad \rho(b, c) = (1-t)\rho(a, b)\ .\]

If \(a, b \in E\), and \(t \in \mathbb{R}\), then \((a, b)\) has the Extrapolation Property if for each \(t \geq 1\) there is a point \(c \in E\) such that

\[\rho(a, c) = t\rho(a, b) \quad \text{and} \quad \rho(b, c) = (t-1)\rho(a, b)\ .\]

The pair \((a, b)\) has the Unique Interpolation Property (resp. the Unique Extrapolation Property) if the \(c\) is unique.

The Interpolation Property is symmetric in \(a\) and \(b\), but the Extrapolation Property is not.
VI.3.1.18. If there is a line segment in $E$ with endpoints $a$ and $b$, then $(a, b)$ has the Interpolation Property. But even the Unique Interpolation Property for $(a, b)$ does not imply that there is a line segment in $E$ with endpoints $a$ and $b$ (Exercise VI.3.3.3.). Similarly, if there is a ray from $a$ through $b$, then $(a, b)$ has the Extrapolation Property; but even the Unique Extrapolation Property does not imply existence of a ray.

A metric space with the Interpolation Property satisfies the Between Axiom; the converse is false in general if the space is incomplete (e.g. $\mathbb{Q}$). Similarly, the Extrapolation Property implies the Extension Axiom, but not conversely (e.g. $\mathbb{Z}$).

Here is a result which, among other things, can substitute for completeness in yielding the Line Axiom:

VI.3.1.19. Theorem. (i) If every pair $(a, b)$ of points in $E$ has the Unique Interpolation Property, then for every $a, b \in E$, $a \neq b$, there is a unique line segment in $E$ with endpoints $a$ and $b$.

(ii) If every pair $(a, b)$ of points in $E$ has the Unique Interpolation Property and the Unique Extrapolation Property, then through any two distinct points of $E$ there is a unique line.

See Exercise VI.3.3.7. for the proof.

The Pythagorean Axiom

The second axiom for Euclidean space has several equivalent forms, at least in the presence of the Line Axiom.

VI.3.1.20. Weak Euclidean Four-Point Axiom: If $a, b, c, d \in E$, and $b, c, d$ lie on a line in $E$, then $\{a, b, c, d\}$ is isometric to a set of four points in $\mathbb{R}^2$ (with the usual Euclidean metric).

The stronger Euclidean Four-Point Axiom is the same without the requirement that three of the points lie on a line, and $\mathbb{R}^2$ replaced by $\mathbb{R}^3$. These axioms trivially hold in $\mathbb{R}^n$ (four points in $\mathbb{R}^n$ generate an affine subspace of dimension $\leq 3$).

One can similarly define the Euclidean $n$-Point Axiom for any $n$. Note that every metric space satisfies the Euclidean Three-Point Axiom, since if $\alpha, \beta, \gamma$ are any positive real numbers with each less than or equal to the sum of the other two, there is a (possibly degenerate) triangle in $\mathbb{R}^2$ with sides $\alpha, \beta, \gamma$. It turns out that the Weak Euclidean Four-Point Axiom along with the Line Axiom implies the Euclidean $n$-Point Axiom for all $n$ (VI.3.1.27., VI.3.1.35.).

VI.3.1.21. Definition. Let $(E, \rho)$ be a metric space and $a, b$ distinct points of $E$. A point $m$ in $E$ is a midpoint for $a$ and $b$ if $\rho(a, m) = \rho(m, b) = \frac{1}{2}\rho(a, b)$. A midpoint $m$ is a linear midpoint if there is a line in $E$ containing $a$, $b$, and $m$.

A midpoint for $a$ and $b$ is obviously between $a$ and $b$. Each line through $a$ and $b$ contains a unique midpoint of $a$ and $b$; however, if lines are not unique, (linear) midpoints may not be unique either (Exercise VI.3.3.2.).

If $(E, \rho)$ is complete and satisfies the Between and Extension Axioms, then any midpoint $m$ for $a$ and $b$ automatically lies on a line through $a$ and $b$ (VI.3.1.16.(iii)) and is thus a linear midpoint. The Line Axiom is not enough to guarantee this, however (?).
VI.3.1.22. **Parallelogram Axiom:** If \( a, b, c \in E \), and \( m \) is a linear midpoint of \( b \) and \( c \), then

\[
2 \left[ \rho(a, b)^2 + \rho(a, c)^2 \right] = \rho(b, c)^2 + 4\rho(a, m)^2.
\]

This law expresses the familiar geometric fact that the sum of the squares of the lengths of the diagonals of a parallelogram equals the sum of the squares of the lengths of the sides. (The fourth vertex of the “parallelogram” would be a point \( d \) for which \( m \) is also a midpoint for \( a \) and \( d \). In the presence of the Line Axiom, it will follow that there is a unique such \( d \), and opposite sides of the resulting figure are automatically equal length.) This axiom holds in \( \mathbb{R}^n \) (XV.9.2.6.).

VI.3.1.23. **Definition.** Let \( S \) be a subset of a metric space \((E, \rho)\), and \( a \) a point of \( E \). A foot of \( a \) on \( S \) is a point \( f \) of \( S \) such that \( \rho(a, f) = \rho(a, S) = \min\{\rho(a, c) : c \in S\} \).

If \( a \in S \), then \( a \) itself is a foot of \( a \) on \( S \) (and is the only foot of \( a \) on \( S \)). The concept of foot is only interesting if \( \rho(a, S) > 0 \).

VI.3.1.24. **Proposition.** Let \((E, \rho)\) be a metric space, \( S \) a subset of \( E \) with the Bolzano-Weierstrass property () (e.g. a subset of \( E \) isometric to \( \mathbb{R}^n \) for some \( n \)), and \( a \) a point of \( E \). Then there exists a foot of \( a \) on \( S \).

**Proof:** Let \( c \) be an element of \( S \). Let

\[
K = \{ x \in S : \rho(x, c) \leq 2\rho(a, c) \}.
\]

Then the triangle inequality implies that if \( e \) is a point of \( L \) which is not in \( K \), then \( \rho(a, e) > \rho(a, c) \). So we have \( \rho(a, S) = \rho(a, K) \), and \( K \) is closed and bounded in \( E \), hence compact, so the distance is minimized at some point of \( K \). \( \square \)

VI.3.1.25. In particular, if \( L \) is a line in \( E \), and \( a \) a point of \( E \), then \( a \) has a foot on \( L \). A foot of a point \( a \) on a line \( L \) need not be unique in general. For example, in \((\mathbb{R}^2, \rho_\infty)\), if \( L \) is the \( x \)-axis and \( a = (0, 1) \), then any point \( (t, 0) \) with \( |t| \leq 1 \) is a foot of \( a \) on \( L \).

VI.3.1.26. **Pythagorean Axiom:** If \( L \) is a line in \( E \), and \( a \) a point of \( E \), and \( f \) is a foot of \( a \) on \( L \), then for any \( c \in L \) we have

\[
\rho(a, c)^2 = \rho(a, f)^2 + \rho(f, c)^2.
\]

The Pythagorean Axiom holds in \( \mathbb{R}^n \). It obviously implies that the foot of a point on a line is unique.

VI.3.1.27. **Theorem.** Let \((E, \rho)\) be a metric space satisfying the Line Axiom. The following are equivalent:

(i) \((E, \rho)\) satisfies the Weak Euclidean Four Point Axiom.

(ii) \((E, \rho)\) satisfies the Parallelogram Axiom.

(iii) \((E, \rho)\) satisfies the Pythagorean Axiom.
(iv) If \( L \) is a line in \( E \) and \( a \) a point of \( E \) not on \( L \), then \( L \cup \{a\} \) can be isometrically embedded in \( \mathbb{R}^2 \) (with the usual metric).

These conditions imply that there is a unique line through any two distinct points of \( E \).

**Proof:** (iv) \( \Rightarrow \) (i) \( \Rightarrow \) (ii) and (iv) \( \Rightarrow \) (iii) are trivial. We will prove (iii) \( \Rightarrow \) (ii) \( \Rightarrow \) (iv). (In fact, the Line Axiom is only needed for (iv) \( \Rightarrow \) (i) and for the last assertion.) \( \diamondsuit \)

**VI.3.1.28. Lemma.** Let \((E, \rho)\) be a metric space satisfying the Pythagorean Axiom. Then \( E \) satisfies the Parallelogram Axiom.

**Proof:** Let \( a, b, c \in E \) and \( L \) a line through \( b \) and \( c \). Let \( m \) be the midpoint of \( b \) and \( c \) on \( L \), and let \( f \) be the foot of \( a \) on \( L \). Set \( \alpha = \rho(a,c), \beta = \rho(a,b), \gamma = \rho(b,c) \). There are two cases to consider: (1) \( f = b, f = c, \) or \( f \) is between \( b \) and \( c \), and (2) \( c \) is between \( b \) and \( f \) (the case where \( b \) is between \( f \) and \( c \) is symmetric and handled by interchanging \( b \) and \( c \)). See Figure (), although the reader is cautioned that since the points are in \( E \) and not in \( \mathbb{R}^2 \), the accuracy of the picture cannot (yet) be justified and it should be used only as a visual aid in following the proof (this is usually a good attitude to take with any picture or diagram).

Consider case (1). By interchanging \( b \) and \( c \) if necessary, we may assume that \( f = c, f = m, \) or \( f \) is between \( m \) and \( c \). Then there is a \( t, 0 \leq t \leq \frac{1}{2}, \) such that \( \rho(f,c) = t \gamma, \rho(m,f) = (\frac{1}{2} - t) \gamma, \rho(b,f) = (1-t) \gamma. \) Set \( \eta = \rho(a,f) \) and \( \zeta = \rho(a,m) \). By the Pythagorean Axiom we have

\[
\alpha^2 = \eta^2 + t^2 \gamma^2 \\
\beta^2 = \eta^2 + (1-t)^2 \gamma^2 = \eta^2 + \gamma^2 - 2t \gamma^2 + t^2 \gamma^2 \\
\zeta^2 = \eta^2 + \left(\frac{1}{2} - t\right)^2 \gamma^2 = \eta^2 + \frac{1}{4} \gamma^2 - t \gamma^2 + t^2 \gamma^2
\]

so we have

\[
2\alpha^2 + 2\beta^2 = 2\eta^2 + 2t^2 \gamma^2 + 2\eta^2 + 2\gamma^2 - 4t \gamma^2 + 2t^2 \gamma^2 = 4\eta^2 + 2\gamma^2 - 4t \gamma^2 + 4t^2 \gamma^2 \\
4\zeta^2 + \gamma^2 = 4\eta^2 + \gamma^2 - 4t \gamma^2 + 4t^2 \gamma^2 + \gamma^2 = 2\alpha^2 + 2\beta^2
\]

which is the Parallelogram formula.

Now consider case (2). There is a \( t \geq 0 \) such that \( \rho(f,c) = t \gamma, \rho(f,m) = (\frac{1}{2} + t) \gamma, \rho(f,b) = (1+t) \gamma. \) Again by the Pythagorean Axiom

\[
\alpha^2 = \eta^2 + t^2 \gamma^2 \\
\beta^2 = \eta^2 + (1+t)^2 \gamma^2 = \eta^2 + \gamma^2 + 2t \gamma^2 + t^2 \gamma^2 \\
\zeta^2 = \eta^2 + \left(\frac{1}{2} + t\right)^2 \gamma^2 = \eta^2 + \frac{1}{4} \gamma^2 + t \gamma^2 + t^2 \gamma^2
\]

so we have

\[
2\alpha^2 + 2\beta^2 = 2\eta^2 + 2t^2 \gamma^2 + 2\eta^2 + 2\gamma^2 + 4t \gamma^2 + 2t^2 \gamma^2 = 4\eta^2 + 2\gamma^2 + 4t \gamma^2 + 4t^2 \gamma^2 \\
4\zeta^2 + \gamma^2 = 4\eta^2 + \gamma^2 + 4t \gamma^2 + 4t^2 \gamma^2 + \gamma^2 = 2\alpha^2 + 2\beta^2
\]

which is the Parallelogram formula. \( \diamondsuit \)
VI.3.1.29. Lemma. Let \((E, \rho)\) be a metric space satisfying the Parallelogram Axiom. Let \(L\) be a line in \(E, b, c\) points on \(L\), and \(a\) a point not on \(L\). Set \(\alpha = \rho(a, c), \beta = \rho(a, b), \gamma = \rho(b, c)\). Then for any \(t \in \mathbb{R}\), if \(d_t\) is the unique point on \(L\) with \(\rho(b, d_t) = |t|\gamma\) and \(\rho(d, c) = |1 - t|\gamma\), we have
\[
\rho(a, d_t)^2 = t\alpha^2 + (1 - t)\beta^2 - t(1 - t)\gamma^2.
\]

Proof: By continuity, it suffices to prove the formula if \(t\) is a dyadic rational number. The formula is trivially true if \(t = 0\) or \(t = 1\) (and the case \(t = \frac{1}{2}\) is the Pythagorean Axiom, although we do not explicitly need this case).

We first show that if the formula holds for \(t\), it also holds for \(-t\). Set \(\rho(a, d_t) = \eta\). By assumption,
\[
\eta^2 = t\alpha^2 + (1 - t)\beta^2 - t(1 - t)\gamma^2.
\]
Since \(\rho(d_t, d_{-t}) = 2|t|\gamma\) and \(b\) is the midpoint of \(d_t\) and \(d_{-t}\) on \(L\), by the Parallelogram Axiom we have
\[
2\rho(a, d_{-t})^2 + 2\eta^2 = 4\beta^2 + 4t^2\gamma^2
\]
so the formula holds for \(-t\).

Similarly, we show that if the formula holds for \(t\), it holds for \(2t\). Again let \(\eta = \rho(a, d_t)\). Since \(\rho(b, d_{2t}) = 2|t|\gamma\) and \(d_t\) is the midpoint of \(b\) and \(d_{2t}\) on \(L\), by the Parallelogram Axiom we have
\[
2\rho(a, d_{2t})^2 + 2\eta^2 = 4\beta^2 + 4t^2\gamma^2
\]
so the formula holds for \(2t\).

Finally, suppose \(s, t \in \mathbb{R}\) and the statement is true for \(s\) and \(t\). We will show that the statement is also true for \(r = \frac{s + t}{2}\). Let \(d_s, d_t, d_r\) be the points of \(L\) corresponding to \(s, t, r\) respectively, and set \(\eta = \rho(a, d_t), \zeta = \rho(a, d_s)\). Then by assumption
\[
\eta^2 = t\alpha^2 + (1 - t)\beta^2 - t(1 - t)\gamma^2
\]
and, since \(\rho(d_t, d_s) = |s - t|\gamma\) and \(d_r\) is the midpoint of \(d_t\) and \(d_s\) on \(L\), by the Parallelogram Axiom we have
\[
\rho(a, d_r)^2 = \frac{1}{2}[(\zeta^2 + \eta^2) - \frac{1}{4}(s - t)^2\gamma^2]
\]
so the formula holds for \(r\) too. Therefore the formula holds for all dyadic rational \(t\).
VI.3.1.30. Corollary. Let \((E, \rho)\) be a metric space satisfying the Parallelogram Axiom. Let \(L\) be a line in \(E\), and \(a\) a point of \(E\). Then there is an isometry from \(L \cup \{a\}\) to \(\mathbb{R}^2\). If \(b, c \in L\) and \(\{a', b', c'\}\) is a subset of \(\mathbb{R}^2\) isometric to \(\{a, b, c\}\), then there is an isometry \(\phi\) from \(L \cup \{a\}\) to \(\mathbb{R}^2\) with \(\phi(a) = a'\), \(\phi(b) = b'\), \(\phi(c) = c'\).

Proof: Let \(b\) and \(c\) be two distinct points on \(L\). Let \(\{a', b', c'\}\) be three points in \(\mathbb{R}^2\) isometric to \(\{a, b, c\}\), and \(L'\) the line in \(\mathbb{R}^2\) through \(b'\) and \(c'\). If \(d\) is any point on \(L\) and \(d'\) the point on \(L'\) with \(\rho_2(b', d') = \rho(b, d)\) and \(\rho_2(c', d') = \rho(c, d)\), then applying Lemma VI.3.1.29. in both \(E\) and \(\mathbb{R}^2\) we obtain \(\rho_2(a', d') = \rho(a, d)\). Thus \(L\) and \(\{a\}\) isometric to \(L' \cup \{a'\}\) under an isometry \(\phi\) with \(\phi(a) = a'\), \(\phi(b) = b'\), \(\phi(c) = c'\). \(\blacksquare\)

VI.3.1.31. Lemma. Let \((E, \rho)\) be a metric space satisfying the Parallelogram Axiom. Let \(a, b, c\) be points of \(E\) with \(a \neq b\) and \(a \neq c\), and let \(L_1\) be a line in \(E\) through \(a\) and \(b\), and \(L_2\) a line in \(E\) through \(a\) and \(c\). If \(\phi\) is an isometry of \(\{a, b, c\}\) with \(\{a', b', c'\}\) \(\subseteq \mathbb{R}^2\), then \(\phi\) extends uniquely to an isometry from \(L_1 \cup L_2\) to \(\mathbb{R}^2\).

Proof: Let \(L_1'\) be the (unique) line in \(\mathbb{R}^2\) through \(a'\) and \(b'\), and \(L_2'\) the line in \(\mathbb{R}^2\) through \(a'\) and \(c'\). There is a unique isometry \(\phi_1\) from \(L_1\) to \(L_1'\) with \(\phi_1(a) = a'\) and \(\phi_1(b) = b'\), since if \(x \in L_1\), \(x\) is uniquely determined by \(\rho(x, a)\) and \(\rho(x, b)\), and similarly for \(L_1'\). In the same way, there is a unique isometry \(\phi_2\) from \(L_2\) to \(L_2'\) with \(\phi_2(a) = a'\), \(\phi_2(c) = c'\). It suffices to show that if \(x \in L_1\) and \(y \in L_2\), and \(x' = \phi_1(x), y' = \phi_2(y)\), then \(\rho_2(x', y') = \rho(x, y)\). We may assume \(x \neq a\) and \(y \neq a\).

By VI.3.1.30., there is an isometry \(\psi\) from \(L_1 \cup \{a\}\) to \(\mathbb{R}^2\) with \(\psi(a) = a', \psi(b) = b', \psi(c) = c'\), and we must have \(\psi|_{L_1} = \phi_1\), so \(\psi(x) = x'\). Thus \(\rho_2(x', c') = \rho(x, c)\), so by VI.3.1.30. there is an isometry \(\theta\) from \(L_2 \cup \{x\}\) to \(\mathbb{R}^2\) with \(\theta(a) = a', \theta(c) = c', \theta(x) = x'\). We must have \(\theta|_{L_2} = \phi_2\), so \(\theta(y) = y'\) and we have \(\rho_2(x', y') = \rho(x, y)\). \(\blacksquare\)

VI.3.1.32. Corollary. Let \((E, \rho)\) be a metric space satisfying the Parallelogram Axiom, and let \(a\) and \(b\) be distinct points of \(E\). Then there is at most one line in \(E\) through \(a\) and \(b\).

Proof: If \(L_1\) and \(L_2\) are lines in \(E\) through \(a\) and \(b\), set \(c = b\) and apply VI.3.1.31.. \(\blacksquare\)

This completes the proof of Theorem VI.3.1.27..

Affine Subspaces

It would seem to make sense to define an \(n\)-dimensional affine subspace of a metric space to be a subset isometric to \(\mathbb{R}^n\). But it will turn out to be more convenient to make the following unorthodox inductive definition, which is equivalent (VI.3.1.35.) in the presence of the Line and Parallelogram Axioms:

VI.3.1.33. Definition. Let \((E, \rho)\) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom.

(i) A zero-dimensional affine subspace of \(E\) is a single point.
(ii) If \( n \geq 1 \), an \( n \)-dimensional affine generator in \( E \) is a subset \( G(L, a) \) consisting of all points \( x \) lying on a line through \( a \) and a point of \( L \), where \( L \) is an \( (n-1) \)-dimensional affine subspace of \( E \) and \( a \) is a point of \( E \) not in \( L \).

(iii) If \( n \geq 1 \), an \( n \)-dimensional affine subspace of \( E \) is the closure of an \( n \)-dimensional affine generator in \( E \).

(iv) A set \( \{x_0, \ldots, x_n\} \) of \( n+1 \) points of \( E \) is independent if it is not contained in any \( (n-1) \)-dimensional affine subspace of \( E \).

It is easy to verify that the one-dimensional affine subspaces of \( E \) are precisely the lines in \( E \).

**VI.3.1.34.** Proposition. Let \((E, \rho)\) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and \( L \) a subset of \( E \) isometric to \( \mathbb{R}^n \). Then \( L \) is an \( n \)-dimensional affine subspace of \( E \).

Proof: This is easily seen by induction, since if \( L \) is an \((n-1)\)-dimensional affine subspace of \( \mathbb{R}^n \) (in the usual sense, hence by the inductive hypothesis in the sense of VI.3.1.33.), and \( a \) is a point of \( \mathbb{R}^n \) not on \( L \), then \( G(L, a) \) consists of all points of \( \mathbb{R}^n \) not in the affine subspace through \( a \) parallel to \( L \), along with \( a \), and this set is dense in \( \mathbb{R}^n \).

The next theorem, essentially the converse of VI.3.1.34., is perhaps the most important result of this section.

**VI.3.1.35.** Theorem. Let \((E, \rho)\) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and \( n \in \mathbb{N} \). Then

(i) Every \( n \)-dimensional affine subspace of \( E \) is isometric to \( \mathbb{R}^n \).

(ii) If \( \{x_0, \ldots, x_n\} \) is an independent set of \( n+1 \) points of \( E \), then there is a unique \( n \)-dimensional subspace of \( E \) containing \( \{x_0, \ldots, x_n\} \).

Proof: The proof is by induction on \( n \). The case \( n = 1 \) is obvious from VI.3.1.27.. For the inductive step, we suppose that the statement is true for \( n \) for any metric space satisfying the hypotheses. Then let \((E, \rho)\) be a metric space satisfying the hypotheses. Proofs of (i) and (ii) for \( n + 1 \) will be done in parallel. To set up the proof, we make notational definitions for the two cases.

To prove (i), let \( L_{n+1} \) be an \((n+1)\)-dimensional affine subspace of \( E \), and let \( L_n \) be an \( n \)-dimensional affine subspace of \( E \) and \( a \) a point of \( E \) such that \( L_{n+1} \) is the closure of \( G(L_n, a) \). To prove (ii), let \( \{x_0, \ldots, x_{n+1}\} \) be an independent set of \( n+2 \) points of \( E \). If \( \{x_0, \ldots, x_n\} \) is contained in an affine \((n-1)\)-dimensional subspace \( L_{n-1} \), then \( \{x_0, \ldots, x_{n+1}\} \) is contained in the \( n \)-dimensional affine subspace which is the closure of \( G(L_{n-1}, x_{n+1}) \), a contradiction. Thus \( \{x_0, \ldots, x_n\} \) is independent and contained in a unique \( n \)-dimensional affine subspace \( L_n \) of \( E \). Let \( a = x_{n+1} \) and let \( L_{n+1} \) be the closure of \( G(L_n, a) \).

In both cases we have an \((n+1)\)-dimensional affine subspace \( L_{n+1} \) which is the closure of an \((n+1)\)-dimensional generator \( G(L_n, a) \), where \( L_n \) is an \( n \)-dimensional affine subspace of \( E \) and \( a \) a point of \( E \) not in
$L_n$. By the inductive hypothesis, $L_n$ is isometric to $\mathbb{R}^n$, hence is complete and has the Bolzano-Weierstrass property, and satisfies the Line Axiom, i.e. if $b, c \in L_n$, then the unique line in $E$ through $b$ and $c$ is contained in $L_n$.

By VI.3.1.24., there is a foot $f$ of $a$ on $L_n$. If $g$ is another foot of $a$ on $L_n$, let $L$ be the line in $L_n$ through $f$ and $g$; then both $f$ and $g$ are feet of $a$ on $L$, impossible by the Pythagorean Axiom if $f \neq g$. Thus $f$ is the unique foot of $a$ on $L_n$, and $f$ is the foot of $a$ on any line in $L_n$ through $f$. So if $b$ is any point of $L_n$, we have

$$\rho(a, b)^2 = \rho(a, f)^2 + \rho(f, b)^2.$$  

In $\mathbb{R}^{n+1}$, let $L'_n$ be the $n$-dimensional subspace of points with last coordinate $0$, and let $a' = (0, \ldots, 0, \lambda)$, where $\lambda = \rho(a, f)$. Let $\phi$ be an isometry of $L_n$ onto $L'_n$ sending $f$ to the origin. Then if $b \in L_{n+1}$, we have

$$\rho_2(a', \phi(b))^2 = \lambda^2 + \rho(0, \phi(b))^2$$

so by the previous calculation $\rho_2(a', \phi(b)) = \rho(a, b)$, and $\phi$ extends to an isometry from $L_n \cup \{a\}$ to $L'_n \cup \{a'\}$, also denoted $\phi$.

We next show that $\phi$ can be extended to an isometry from $G(L_n, a)$ to $G(L'_n, a')$, which consists of all the points of $\mathbb{R}^{n+1}$ whose last coordinate is not $\lambda$, along with $a'$ (i.e. $\mathbb{R}^{n+1}$ with the horizontal hyperplane through $a'$, except $a'$ itself, removed).

Let $x \in G(L_n, a), x \neq a$. Then there is a unique $b \in L_n$ such that $x$ is on the line through $a$ and $b$. Let $\phi(x)$ be the corresponding point on the line through $a'$ and $\phi(b)$ in $\mathbb{R}^{n+1}$. If $y$ is another point in $G(L_n, a), y \neq a$, let $y$ be on the line through $a$ and $c \in L_n$. If the plane through $a', \phi(b), \phi(c)$ in $\mathbb{R}^{n+1}$ is identified with $\mathbb{R}^2$, then by VI.3.1.31. the isometry $\phi$ from $\{a, b, c\}$ to $\{a', \phi(b), \phi(c)\}$ extends uniquely to the lines through $a$ and $b$ and through $a$ and $c$; hence in particular $\rho_2(\phi(x), \phi(y)) = \rho(x, y)$ and the $\phi$ we have defined is an isometry from $G(L_n, a)$ into $G(L'_n, a')$. To see that it is surjective, let $x' \in G(L'_n, a') x' \neq a'$, and let $b'$ be the point of $L'_n$ on the line through $a'$ and $x'$ (a' and $x'$ have different last coordinates, so there is a unique point on this line with last coordinate $0$). Then $b' = \phi(b)$ for some $b \in L_n$, and there is a (unique) point $x$ on the line in $E$ through $a$ and $b$ with $x' = \phi(x)$, and $\phi$ maps $G(L_n, a)$ onto $G(L'_n, a')$.

Thus $\phi$ extends to an isometry from $L_{n+1}$ into $\mathbb{R}^{n+1}$, also denoted $\phi$. We show that this $\phi$ is surjective. (This would be immediate if $(E, \rho)$ is complete, but $(E, \rho)$ need not be complete.) We must show that if

$$m' = (\alpha_1, \ldots, \alpha_n, \lambda) \in \mathbb{R}^n \setminus G(L'_n, a')$$

then there is an $m \in E$ with $\phi(m) = m'$. Let $b' = (\alpha_1, \ldots, \alpha_n, 0)$ and $c' = (\alpha_1, \ldots, \alpha_n, 2\lambda)$ in $\mathbb{R}^{n+1}$, and let $b = \phi^{-1}(b')$, $c = \phi^{-1}(c')$, and let $m$ be the midpoint of the line segment in $E$ with endpoints $a$ and $b$. It suffices to show that $m \in L_{n+1}$, for then necessarily $\phi(m) = m'$. But if $t \in (0, 1)$, $t \neq \frac{1}{2}$, then the point

$$d'_t = (\alpha_1, \ldots, \alpha_n, 2t\lambda) \in G(L'_n, a')$$

is between $b'$ and $c'$, hence $d_t = \phi^{-1}(d'_t) \in G(L_n, a)$ is between $b$ and $c$; hence all the points of the line segment between $b$ and $c$ other than $m$ are in $G(L_n, a)$, so $m$ is in the closure of $G(L_n, a)$.

We have now proved (i) for $n + 1$, and proved all of (ii) except uniqueness. For uniqueness, let $\hat{L}_{n+1}$ be any $(n + 1)$-dimensional affine subspace of $E$ containing $\{x_0, \ldots, x_{n+1}\}$. Then by (i) $\hat{L}_{n+1}$ is isometric to $\mathbb{R}^{n+1}$ and hence satisfies the Line Axiom. Any $k$-dimensional affine subspace of $\hat{L}_{n+1}$ ($k \leq n + 1$) is a $k$-dimensional affine subspace of $E$ by VI.3.1.34.. Thus $\{x_0, \ldots, x_n\}$ is independent as a subset of $\hat{L}_{n+1}$, so by the inductive hypothesis applied to $\hat{L}_{n+1}$, $\{x_0, \ldots, x_n\}$ is contained in a unique $n$-dimensional affine subspace of $\hat{L}_{n+1}$. But $L_n$ is the unique $n$-dimensional affine subspace of $E$ containing $\{x_0, \ldots, x_n\}$, so $L_n \subseteq \hat{L}_{n+1}$. Then $G(L_n, x_{n+1}) \subseteq \hat{L}_{n+1}$ since $\hat{L}_{n+1}$ contains the line in $E$ through any two of its points.

756
Thus the closure $L_{n+1}$ of $G(L_n, x_{n+1})$ is contained in $\hat{L}_{n+1}$ since $\hat{L}_{n+1}$ is closed in $E$. But this implies that $L_{n+1} = \hat{L}_{n+1}$ since $\mathbb{R}^{n+1}$ has no proper $(n + 1)$-dimensional affine subspaces.

This completes the proof of Theorem VI.3.1.35.

VI.3.1.36. **Corollary.** Let $(E, \rho)$ be a metric space satisfying the Line and Parallelogram (or Pythagorean) Axioms, and $S = \{x_0, \ldots, x_m\}$ a subset of $E$ with $n + 1$ points. Then, for some $m \leq n$, there is a (unique) $m$-dimensional affine subspace of $E$ containing $S$ which is the smallest affine subspace of $E$ containing $S$, called the affine subspace of $E$ spanned by $S$, denoted $\text{span}(S)$; $m = n$ if and only if $S$ is independent. Any subset of an independent set is independent.

Any real inner product space with the induced metric satisfies all the axioms considered so far. Conversely:

VI.3.1.37. **Theorem.** Let $(E, \rho)$ be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and let $a \in E$. Then

(i) There is a unique real vector space structure on $E$ with $a$ the zero vector, and a unique inner product on this vector space, such that $\rho$ is the metric induced by the inner product.

(ii) Every subset of $E$ satisfying the Line Axiom (in particular, every $n$-dimensional affine subspace in the sense of VI.3.1.33.) is an affine subspace in the linear algebra sense.(

(iii) If $S = \{x_0, \ldots, x_n\}$ is a finite subset of $E$, then the span of $S$ in the sense of VI.3.1.36. is the linear span of $S$ in the vector space structure on $E$ with zero vector $x_0$. The set $S$ is independent in the sense of VI.3.1.33. if and only if $\{x_1, \ldots, x_n\}$ is linearly independent in this vector space structure on $E$.

Combining this with VI.3.1.13., we obtain:

VI.3.1.38. **Corollary.** Let $(E, \rho)$ be a complete metric space satisfying the Between, Extension, and Parallelogram (or Pythagorean) Axioms. Then $(E, \rho)$ is isometric to a real Hilbert space.(

**Proof:** The proof of VI.3.1.37. follows easily from VI.3.1.35. Fix $a \in E$. If $b, c \in E$, there is an isometry from the span of $\{a, b, c\}$ to $\mathbb{R}^n$ for some $n \leq 2$ sending $a$ to 0, and we define vector addition, scalar multiplication, norm, and inner product for $b, c$ via this isometry. Since an isometry of $\mathbb{R}^n$ fixing the origin is linear and orthogonal, these operations are well defined. All except the inner product have simple intrinsic characterizations. (The inner product can be recovered from the norm by polarization.) We have $\|b\| = \rho(a, b)$. If $a \in \mathbb{R}$ and $b \neq a$, then $ab$ is the unique element $c$ on the line in $E$ through $a$ and $b$ with $\rho(a, c) = |a|\rho(a, b)$ and $a$ between $c$ and $b$ if and only if $a < 0$. If $m$ is the midpoint of $b$ and $c$ in $E$, then $b + c = 2m$. Thus this structure is the only possible vector space structure and norm compatible with $\rho$.

To verify that $E$ is a normed vector space with these operations, the only axiom which is not obviously satisfied is associativity of vector addition. But if $b, c, d \in E$, then the span of $\{a, b, c, d\}$ is isometric to $\mathbb{R}^n$ for some $n \leq 3$, and the vector addition in $E$ restricted to this subspace agrees with vector addition in this $\mathbb{R}^n$ if an isometry is chosen sending $a$ to the origin, and associativity of vector addition follows.

Parts (ii) and (iii) are immediate from (i).
The Dimension Axiom

Thus, to characterize $\mathbb{R}^n$ as a metric space, we need only one additional axiom insuring that the space is $n$-dimensional. The simplest way to do this is

**VI.3.1.39. Dimension $n$ Axiom:** There is an independent set of $n + 1$ elements of $E$ which spans $E$.

The Dimension $n$ Axiom can be rephrased in a purely metric way not requiring reference to the Unique Line Axiom.

**VI.3.1.40. Definition.** Let $(E, \rho)$ be a metric space, and $(a_0, \ldots, a_m)$ an $(m + 1)$-tuple of points in $E$. The Cayley-Menger determinant $D(a_0, \ldots, a_m)$ of $(a_0, \ldots, a_m)$ is the determinant of the $(m + 2) \times (m + 2)$ matrix

$$
\begin{vmatrix}
0 & 1 & 1 & \cdots & 1 \\
1 & 0 & \rho(a_0, a_1)^2 & \cdots & \rho(a_0, a_n)^2 \\
1 & \rho(a_1, a_0)^2 & 0 & \cdots & \rho(a_1, a_m)^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \rho(a_m, a_0)^2 & \rho(a_m, a_1)^2 & \cdots & 0
\end{vmatrix}
$$

**VI.3.1.41.** If $E = \mathbb{R}^n$, there is a constant $c_m = \frac{(-1)^{n+1}}{2^n(n!)^2}$ depending only on $m$ such that $c_m D(a_0, \ldots, a_m)$ is the square of the $m$-dimensional volume of the parallelepiped with edges $a_1 - a_0, \ldots, a_m - a_0$. In particular, it is nonzero if and only if the set $\{a_1 - a_0, \ldots, a_m - a_0\}$ is linearly independent. (The constant is not relevant for our uses of the Cayley-Menger determinant in this section, since we will be concerned only whether it is 0.)

**VI.3.1.42. $n$-Determinant Axiom:** The space $E$ has an $(n+1)$-tuple $(a_0, \ldots, a_n)$ with $D(a_0, \ldots, a_n) \neq 0$, but $D(b_0, \ldots, b_{n+1}) = 0$ for any $(n + 2)$-tuple $(b_0, \ldots, b_{n+1})$ in $E$.

The Characterization Theorems

The main result of this section is:

**VI.3.1.43. Theorem.** Let $(E, \rho)$ be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom. Then the following are equivalent:

(i) $(E, \rho)$ satisfies the Dimension $n$ axiom.

(ii) $(E, \rho)$ satisfies the $n$-Determinant Axiom.

(iii) $(E, \rho)$ is isometric to $(\mathbb{R}^n, \rho_2)$.

**Proof:** (i) $\Rightarrow$ (iii) is VI.3.1.35., (iii) $\Rightarrow$ (ii) and the equivalence of (i) and (ii) for $\mathbb{R}^n$ is VI.3.1.41.. $\diamond$

Combining this result with VI.3.1.13., we obtain perhaps the nicest metric characterization of $n$-dimensional Euclidean space:

758
VI.3.1.44. **Corollary.** Let \((E, \rho)\) be a complete metric space satisfying the Between, Extension, Parallelogram (or Pythagorean), and \(n\)-Determinant Axioms. Then \((E, \rho)\) is isometric to \((\mathbb{R}^n, \rho_2)\).

In fact, the Parallelogram Axiom can be replaced with a more complicated condition involving Cayley-Menger determinants of smaller order. This was the original metric characterization of \(\mathbb{R}^n\) obtained by K. Menger in 1928. See [Blu70, Theorem 48.2].

VI.3.2. \(\mathbb{R}^n\) is \(n\)-Dimensional

Everyone “knows” that \(\mathbb{R}^n\) is \(n\)-dimensional. But what does this mean? Specifically, what does “dimension” mean? And how can \(\mathbb{R}^n\) be distinguished from \(\mathbb{R}^m\) for \(m \neq n\) by its properties?

\(\mathbb{R}^n\) is a set with several types of structure: algebraic, geometric, topological, and analytic. Efforts have been made to define and/or describe the “dimension” of \(\mathbb{R}^n\) in terms of each of these types of structure, and various theories of dimension have been studied in different contexts; the only thing these various theories have in common is that they all assign the same number \(n\) to \(\mathbb{R}^n\), justifying the use of the name “dimension” for all of them.

VI.3.2.1. Specifically, we have:

(i) The vector space dimension \((\cdot)\) of \(\mathbb{R}^n\) as a real vector space is \(n\).

(ii) The topological dimension \((\cdot)\) of \(\mathbb{R}^n\) as a topological space is \(n\). (In fact, there are several notions of topological dimension, all of which give dimension \(n\) to \(\mathbb{R}^n\).)

(iii) The Hausdorff dimension \((\cdot)\) of \(\mathbb{R}^n\) is \(n\).

VI.3.2.2. There is one respect in which all Euclidean spaces are the same, however: cardinality. Contrary to Cantor’s expectations (he first thought the cardinality of \(\mathbb{R}^n\) should be \(\aleph_n\)), the cardinality of \(\mathbb{R}^n\) is \(2^{\aleph_0}\) for all \(n\). Thus some additional structure beyond set theory is needed to distinguish \(\mathbb{R}^n\) from \(\mathbb{R}^m\) for \(n \neq m\).

VI.3.3. Exercises

VI.3.3.1. Let \(X\) be a set, and \(a, b \in X\). Define a metric \(\rho\) on \(X\) by \(\rho(a, b) = 2\) and \(\rho(x, y) = 1\) for all other distinct pairs \(x, y\). Then every element of \(X \setminus \{a, b\}\) is between \(a\) and \(b\).

VI.3.3.2. In \((\mathbb{R}^2, \rho_\infty)\), the set of points between \(a = (0, 0)\) and \(b = (2, 0)\) is

\[\{(t, s) : 0 < t < 2, |s| \leq \min(t, 2 - t)\}\]

which is the “diamond” with corners \(a, b, (1, 1)\), and \((1, -1)\).

VI.3.3.3. Let \((E, \rho)\) be the subset

\[\{(t, t) : 0 \leq t \leq 1\} \cup \{(t, t - 2) : 1 < t \leq 2\}\]

of \((\mathbb{R}^2, \rho_\infty)\). If \(a = (0, 0)\) and \(b = (2, 0)\), then for every \(t, 0 \leq t \leq 2\), there is a unique \(c_t\) between \(a\) and \(b\) with \(\rho(a, c_t) = t\) and \(\rho(c_t, b) = 2 - t\) (i.e. \(E\) has the Unique Interpolation Property), but there is no line segment in \(E\) with endpoints \(a\) and \(b\).
VI.3.3.4. Let $E$ be the unit sphere in $\mathbb{R}^3$.

(a) If $E$ is given the induced metric from $\mathbb{R}^3$ (sometimes called the chord metric), and $a, b \in E$, then there are no points of $E$ between $a$ and $b$.

(b) If the distance between points $a$ and $b$ is defined to be the length of the shortest great circle arc through $a$ and $b$, a metric is obtained called the arc metric. If $a$ and $b$ are not diametrically opposed, then there is a unique line segment with endpoints $a$ and $b$. If $a$ and $b$ are diametrically opposed, every other point of $E$ is between $a$ and $b$. $E$ is complete and satisfies the Between Axiom, but $E$ contains no lines. It almost satisfies the Extension Axiom, except for pairs of points diametrically opposed.

The same results hold in $\mathbb{R}^n$ for any $n > 1$ (in particular, for the unit circle in $\mathbb{R}^2$).

VI.3.3.5. Let $(E, \rho)$ be a metric space satisfying the Line Axiom. A subset $S$ of $E$ is collinear if there is a line in $E$ containing $S$.

(a) Prove the following theorem (cf. [Bli70, Theorem 21.3]):

**Theorem.** $(E, \rho)$ satisfies the Unique Line Axiom if and only if $(E, \rho)$ has the Two-Triple Property: whenever $a, b, c, d \in E$, $\{a, b, c\}$ is collinear, and $\{b, c, d\}$ is collinear, then $\{a, b, d\}$ is collinear and $\{a, c, d\}$ is collinear.

It is then a consequence of the theorem that the four lines necessarily coincide, and therefore $\{a, b, c, d\}$ is collinear.

(b) Show directly that if $E$ satisfies the Weak Euclidean Four-Point Axiom, then $E$ has the Two-Triple Property.

VI.3.3.6. Show that $(\mathbb{R}^2, \rho_p)$ has the Unique Line Property for $1 < p < \infty$, but not the Parallelogram Property (or other properties of VI.3.1.27.) if $p \neq 2$. More generally, any strictly convex normed real vector space $(\cdot)$ has the Unique Line Property.

VI.3.3.7. Let $(E, \rho)$ be a metric space satisfying the Unique Interpolation and Unique Extrapolation Axioms. Prove that $E$ has the Unique Line Property, as follows. Let $a, b \in E$ with $a \neq b$, and set $\lambda = \rho(a, b)$.

(a) Show the following:

(i) For each $s \geq \lambda$, there is a unique $c \in E$ with $\rho(a, c) = s$ and $\rho(b, c) = s - \lambda$.

(ii) For each $s$, $0 \leq s \leq \lambda$, there is a unique $c \in E$ with $\rho(a, c) = s$ and $\rho(c, b) = \lambda - s$.

(iii) For each $s < 0$, there is a unique $c \in E$ with $\rho(a, c) = -s$ and $\rho(b, c) = \lambda - s$.

If $s \in \mathbb{R}$, write $\phi(s)$ for the corresponding point $c$.

(b) Show that $\rho(\phi(r), \phi(s)) = |s - r|$ for all $r, s \in \mathbb{R}$; we may assume $r \leq s$. Consider six separate cases:

(i) $d \leq r \leq s$.

(ii) $0 \leq r \leq d \leq s$.

(iii) $0 \leq r \leq s \leq d$.

(iv) $r \leq 0 < d \leq s$. 

760
(v) $r \leq 0 \leq s \leq d$.

(vi) $r \leq s \leq 0$.

For case (i), consider the unique $b \in E$ with $\rho(a_1, b) = r - \lambda$ and $\rho(b, \phi(s)) = s - r$ by (a). Use the triangle inequality twice to show that $\rho(a_0, b) = r$, so $b = \phi(r)$. Give similar (or simpler) arguments in the other five cases.

VI.3.3.8. In metric geometry (), a stronger version of the Between Axiom is often used. **Midpoint Axiom:** Every pair of points in $(E, \rho)$ has a midpoint (VI.3.1.21).

No uniqueness is assumed. A metric space satisfying the Midpoint Axiom obviously satisfies the Between Axiom.

(a) Prove directly that a complete metric space satisfying the Midpoint Axiom satisfies the Line Segment Axiom, i.e. is strictly intrinsic. (Your proof should not need the AC, but will need the Axiom of Dependent Choice since midpoints are not necessarily unique.)

(b) [Blu70, Theorem 14.1] Adapt this proof to give a proof of Lemma VI.3.1.14. using only DC.

It is known that every Peano space () has a metric satisfying the Midpoint Axiom, or, equivalently, the Between Axiom or the Line Segment Axiom ([Bin49], [Moi49]; note that this result is incorrectly stated in [HY88, p. 130]).
Figure VI.5: A favorite subject in Croatia
VI.4. Length Structures and Metric Geometry

Metric spaces can be regarded as a first approximation to geometry (this is of course not the only interpretation or use for them!) True geometry has at least two other features in addition to a metric structure, i.e. a notion of distance between points:

(i) A notion of “shortest path,” curves in the space which at least locally determine distance, i.e. the distance between nearby points is measured along such a path. Such paths, generalizing lines in Euclidean space or great circles on a sphere, are called geodesics. For a given point \( p \), any other point \( q \) sufficiently close to \( p \) should be connected to \( p \) by a unique geodesic giving the metric distance between \( p \) and \( q \), and geodesics should be extendible.

(ii) A notion of “angle” between geodesics intersecting at a point. Implicit in this is the idea that each geodesic through a point \( p \) determines a well-defined “direction” (“tangent”) at \( p \). Angles should locally give well-defined notions of area and higher-dimensional volumes, which can be extended globally.

There are at least three approaches possible to defining the additional structure, of increasing generality. First, for Euclidean space there is an “elementary” interpretation of lines and angles (although a fully rigorous treatment is more complicated than it may at first appear; cf. ()), so for “nice” subsets of Euclidean space, e.g. for smooth surfaces in \( \mathbb{R}^3 \), geodesics and angles can be defined analytically. This approach is described in (). Secondly, the case of smooth surfaces can be abstracted into the notion of smooth manifolds with a Riemannian or similar structure; cf. (). Finally, the notions can be directly developed straight from the general notion of metric space. This is the approach taken in this section.

As one might expect, abstract metric spaces are too general to support geometry with reasonable properties, and additional conditions on the space must be assumed. These additional conditions will be seen as the theory is developed. Ultimately, the additional conditions force the space to essentially be the type of structure from the second approach above; it is significant that the same mathematical structures can be arrived at in these two ways, reinforcing the naturality and importance of the resulting structures.

We will only cover the basic aspects of this approach. A far more extensive and detailed treatment can be found in [BBI01] and in [Gro99] (read and study the first before tackling the second!)

VI.4.1. Arc Length in Metric Spaces

The notion of curves and arc length in Euclidean space () has a straightforward generalization to metric spaces. Throughout this section, we will fix a metric space \((X, \rho)\).

Parametrized Curves

VI.4.1.1. Definition. A parametrized curve in \( X \) is a continuous function \( \gamma : [a, b] \to X \) for some closed bounded interval \([a, b]\) in \( \mathbb{R} \). The points \( p = \gamma(a) \) and \( q = \gamma(b) \) are the initial and final endpoints of \( \gamma \) respectively, and \( \gamma \) is a parametrized curve from \( p \) to \( q \).

Note that the interval \([a, b]\) is allowed to be degenerate, i.e. we allow \( a = b \). (To be sure, parametrized curves defined on degenerate intervals are not very interesting!)

VI.4.1.2. A parametrized curve has a direction. If \( \gamma : [a, b] \to X \) is a parametrized curve from \( p \) to \( q \), its negative or reversal is \( -\gamma : [a, b] \to X \) defined by \((-\gamma)(t) = \gamma(a + b - t)\) is a parametrized curve from \( q \) to \( p \).
VI.4.1.3. We may think of the range of a parametrized curve as a “curve” in $X$ (although it need not look anything like a geometric curve), and the parametrization as giving “motion” along this curve. This motion may “retrace” parts of the curve in a complicated way.

Restriction and Concatenation

VI.4.1.4. Definition. Let $\gamma : [a, b] \to X$ be a parametrized curve, and $[c, d]$ a closed subinterval of $[a, b]$, i.e. $a \leq c \leq d \leq b$. The parametrized curve $\gamma|_{[c,d]}$ is the restriction of $\gamma$ to $[c, d]$; it is a parametrized curve from $\gamma(c)$ to $\gamma(d)$.

VI.4.1.5. Definition. If $\gamma_1 : [a, b] \to X$ is a parametrized curve from $p$ to $q$, and $\gamma_2 : [b, c] \to X$ is a parametrized curve from $q$ to $r$, the concatenation is the parametrized curve $\gamma = \gamma_1 + \gamma_2 : [a, c] \to X$ defined by

$$\gamma(t) = \begin{cases} 
\gamma_1(t) & \text{if } a \leq t \leq b \\
\gamma_2(t) & \text{if } b \leq t \leq c
\end{cases}$$

The concatenation $\gamma$ is a parametrized curve from $p$ to $r$.

We may extend the definition to $\gamma_1 + \cdots + \gamma_n$ if the curves line up properly.

VI.4.1.6. Note the conditions under which the concatenation is defined are rather severe: $\gamma_1$ and $\gamma_2$ must be parametrized on intervals in $\mathbb{R}$ which concatenate, and the final endpoint of $\gamma_1$ must be the initial endpoint of $\gamma_2$. Concatenation can be slightly generalized using reparametrizations (VI.4.1.16). Also note that although additive notation is used, concatenation is not commutative: if $\gamma_1 + \gamma_2$ is defined, then $\gamma_2 + \gamma_1$ is not even defined unless one of the intervals is degenerate.

VI.4.1.7. If $\gamma = \gamma_1 + \gamma_2$ is defined, then the restriction of $\gamma$ to $[a, b]$ (resp. to $[b, c]$) is $\gamma_1$ (resp. $\gamma_2$). Conversely, if $\gamma : [a, b]$ is a parametrized curve and $c \in [a, b]$, then $\gamma = \gamma|_{[a,c]} + \gamma|_{[c,b]}$. More generally, if $P = \{a = t_0, t_1, \ldots, t_n = b\}$ is a partition of $[a, b]$, we have

$$\gamma = \gamma|_{[a,t_1]} + \gamma|_{[t_1,t_2]} + \cdots + \gamma|_{[t_{n-1},b]}.$$

Reparametrization

VI.4.1.8. Definition. Let $\gamma : [a, b] \to X$ and $\tilde{\gamma} : [c, d] \to X$ be parametrized curves. Then $\tilde{\gamma}$ is a reparametrization of $\gamma$ if $\tilde{\gamma} = \gamma \circ \phi$, where $\phi$ is a nondecreasing continuous function from $[c, d]$ onto $[a, b]$ (hence $\phi(c) = a$ and $\phi(d) = b$). If $\phi$ is strictly increasing, $\tilde{\gamma}$ is a strict reparametrization of $\gamma$.

VI.4.1.9. A reparametrization of a parametrized curve has the same initial and final endpoints, and the same range. But two parametrized curves with the same range, even also with the same initial and final endpoints, need not be reparametrizations. A reparametrization of a parametrized curve describes “motion” along the range “curve” in the same sense, with the same retracing, but possibly over a different interval of time at a different and varying speed.
VI.4.1.10. Various special kinds of reparametrizations can be considered. The simplest is linear reparametrizations, where \( \phi \) is a linear (affine) function from \([c, d]\) onto \([a, b]\), i.e.

\[
\phi(t) = a + \frac{b - a}{d - c}(t - c) .
\]

One can also consider \( C^r \) reparametrizations, where \( \phi \) is \( C^r \), and strict \( C^r \) reparametrizations, where both \( \phi \) and \( \phi^{-1} \) are \( C^r \) (these are also called smooth reparametrizations). There are also piecewise linear and piecewise smooth reparametrizations.

VI.4.1.11. Reparametrization is not an equivalence relation: it is reflexive and transitive, but not symmetric since the function \( \phi \) need not be one-to-one. Strict reparametrization is symmetric, thus an equivalence relation on parametrized curves. Parametrization does generate an important equivalence relation, which can be described thus: two parametrized curves are equivalent if they are both reparametrizations of the same parametrized curve (VI.4.2.7.).

Nonstationary Curves

VI.4.1.12. Definition. Let \( \gamma : [a, b] \to X \) be a parametrized curve, and \([c, d]\) a closed subinterval of \([a, b]\). Then \( \gamma \) is stationary on \([c, d]\) if \( \gamma(t) = \gamma(c) \) for all \( t \in [c, d] \). The curve \( \gamma \) is nonstationary if it is not stationary on any (nondegenerate) subinterval of \([a, b]\).

VI.4.1.13. A nonstationary parametrized curve need not be one-to-one; in fact, it can be very noninjective. For example, the space-filling curve of XIV.12.14. is nonstationary (the one in XIV.12.13. is not nonstationary). For a simpler example, a nonstationary parametrized curve can wrap around a circle in \( \mathbb{R}^2 \) several times.

VI.4.1.14. The maximal intervals on which a parametrized curve \( \gamma \) is stationary form a collection of disjoint closed subintervals of \([a, b]\). Every parametrized curve is a reparametrization of a nonstationary parametrized curve, which is unique up to strict reparametrization (Exercise ()), obtained by collapsing each maximal stationary subinterval to a point (there may be infinitely many such intervals!) Two parametrized curves are equivalent in the equivalence relation generated by reparametrization (VI.4.1.11.) if and only if their nonstationary versions are strict reparametrizations of each other.

Paths

VI.4.1.15. Definition. Let \( p, q \in X \). A path from \( p \) to \( q \) is an equivalence class of parametrized curves from \( p \) to \( q \) (using the equivalence relation of VI.4.1.11. or VII.1.2.7.).

We will write \( \gamma_1 \approx \gamma_2 \) if \( \gamma_1 \) and \( \gamma_2 \) are equivalent, and write \( \langle \gamma \rangle \) for the equivalence class of \( \gamma \) (the usual notation \( [\gamma] \) for an equivalence class is normally reserved for the homotopy class of \( \gamma \), which is a quite different thing ()).

VI.4.1.16. Paths can be concatenated: if \( \langle \gamma_1 \rangle \) is a path from \( p \) to \( q \) and \( \langle \gamma_2 \rangle \) is a path from \( q \) to \( r \), then \( \gamma_1 \) can be concatenated with a suitable reparametrization \( \hat{\gamma}_2 \) of \( \gamma_2 \) (e.g. a linear reparametrization), and the equivalence class of \( \gamma_1 + \hat{\gamma}_2 \) is well defined and depends only on the classes of \( \gamma_1 \) and \( \gamma_2 \) (Exercise ()). Thus the concatenation of the paths \( \langle \gamma_1 \rangle + \langle \gamma_2 \rangle \) is well defined as \( \langle \gamma_1 + \hat{\gamma}_2 \rangle \).

Restriction of paths is a much dicier matter ().
Arc Length

So far we have not used the metric ρ (except to give the topology on X, i.e. to determine continuity of parametrized curves). We now bring ρ into the picture.

VI.4.1.17. Definition. Let γ : [a, b] → X be a parametrized curve, and 𝒫 = {a = t₀, t₁, . . . , tₙ = b} a partition of [a, b]. The polygonal approximate length of γ from 𝒫 is

$$\ell_ρ(γ, 𝒫) = \sum_{k=1}^{n} ρ(γ(t_{k-1}), γ(t_k)).$$

The arc length of γ is

$$\ell_ρ(γ) = \sup_{𝒫} \ell_ρ(γ, 𝒫).$$

We often suppress the subscript if ρ is understood and there is no possibility of confusion.

VI.4.1.18. If Q is a partition of [a, b] which refines 𝒫, we have $\ell(γ, 𝒫) ≤ \ell(γ, Q)$ by the triangle inequality. Since any two partitions have a common refinement, the polygonal approximations approach the arc length as the partitions become arbitrarily fine. The term “polygonal approximation” should not be taken too literally: there is usually not an actual “polygon” in X approximating the curve. The term is suggestive that $(p;q)$ should be the “straight-line” distance between $p$ and $q$.

VI.4.1.19. The arc length of a curve can easily be infinite (e.g. VI.4.1.20). The curve γ is rectifiable if $\ell(γ)$ is finite. Arc length and rectifiability depend crucially on $\rho (\text{VI.4.1.20}(ii)).$

VI.4.1.20. Examples. (i) Let $X = [0, 1]$ with the ordinary metric $ρ(x, y) = |x - y|$. A parametrized curve in $X$ is a continuous function $γ : [a, b] → [0, 1]$. If 𝒫 is a partition of [a, b], then $\ell(γ, 𝒫)$ is the total variation $V(γ, 𝒫)$ of $γ$ over 𝒫 (XIV.16.1.3.), and $\ell(γ)$ is the total variation $V_{[a,b]}(γ)$ of $γ$ over [a, b] (XIV.16.1.6.). The curve γ is rectifiable if and only if $γ$ has bounded variation. If $γ$ is a smooth function, then $γ$ is rectifiable and

$$\ell(γ) = \int_{a}^{b} |γ'(t)| \, dt.$$

(ii) Let $X = [0, 1]$, and let $γ : [0, 1] → X$ be the identity function $γ(t) = t$. If we give $X$ the usual metric $ρ$, we have $\ell_ρ(γ) = 1$. But if $0 < s < 1$ and we give $X$ the snowflake metric $ρ^s$, where $ρ^s(x, y) = |x - y|^s$, then for the partition $𝒫 = \{\frac{k}{n} : 0 ≤ k ≤ n\}$ of [0, 1], we have

$$\ell_{ρ^s}(γ, 𝒫) = \sum_{k=1}^{n} \frac{1}{n^s} = \frac{n}{n^s} = n^{1-s}$$

so $\ell_{ρ^s}(γ) = +∞$, i.e. γ is not rectifiable. In fact, in the metric space $(X, ρ^s)$, no nonconstant parametrized curve is rectifiable.

We now develop basic properties of arc length. First note that arc length is independent of parametrization:
VI.4.1.21. Proposition. Let \( \tilde{\gamma} : [c, d] \to X \) be a reparametrization of \( \gamma : [a, b] \to X \), i.e. \( \tilde{\gamma} = \gamma \circ \phi \), \( \phi : [c, d] \to [a, b] \). Then \( \ell(\tilde{\gamma}) = \ell(\gamma) \).

Proof: Let \( \mathcal{P} = \{ c = t_0, \ldots, t_n = d \} \) be a partition of \([c, d]\). Then \( \{ a = \phi(t_0), \ldots, \phi(t_n) = b \} \) is not quite a partition of \([a, b]\) since there may be repetitions if \( \phi \) is not injective. But repeated points contribute nothing to the sum \( \ell(\tilde{\gamma}, \mathcal{P}) \), and if \( \mathcal{Q} \) is the partition of \([a, b]\) obtained by deleting repetitions, we have \( \ell(\gamma, \mathcal{Q}) = \ell(\tilde{\gamma}, \mathcal{P}) \). Every partition of \([a, b]\) arises in this manner.

VI.4.1.22. Thus arc length of paths is well defined.

Although the reversal \(-\gamma\) of a parametrized curve is not a reparametrization of \(\gamma\), we nonetheless have:

VI.4.1.23. Proposition. If \( \gamma : [a, b] \to X \) is a parametrized curve in \( X \), then \( \ell(-\gamma) = \ell(\gamma) \).

Proof: If \( \mathcal{P} \) is a partition of \([a, b]\), let \( \tilde{\mathcal{P}} \) be the partition consisting of \( \{ a + b - t : t \in \mathcal{P} \} \). Then \( \tilde{\mathcal{P}} \) is also a partition of \([a, b]\), and every partition of \([a, b]\) arises in this manner. We have \( \ell(-\gamma, \tilde{\mathcal{P}}) = \ell(\gamma, \mathcal{P}) \).

VI.4.1.24. If \( \gamma : [a, b] \to X \) is a parametrized curve from \( p \) to \( q \) and \( \mathcal{P} = \{ a, b \} \) is the trivial partition, we have
\[
\rho(p, q) = \ell(\gamma, \mathcal{P}) \leq \ell(\gamma).
\]
More generally, we have:

VI.4.1.25. Proposition. Let \( \gamma : [a, b] \to X \) be a parametrized curve in \( X \), and \( a \leq c \leq d \leq b \), then
\[
\rho(\gamma(c), \gamma(d)) \leq \ell(\gamma|_{[c, d]}) \leq \ell(\gamma).
\]

Proof: The first inequality has already been noted. For the second, any partition \( \mathcal{P} \) of \([c, d]\) can be expanded to a partition \( \mathcal{Q} \) of \([a, b]\) by adding \( a \) and \( b \) if necessary. We have
\[
\ell(\gamma|_{[c, d]}, \mathcal{P}) \leq \ell(\gamma, \mathcal{Q}) \leq \ell(\gamma)
\]
and taking the supremum over all \( \mathcal{P} \) gives the inequality.

If \( \gamma \) is a constant path, obviously \( \ell(\gamma) = 0 \). Conversely:

VI.4.1.26. Corollary. Let \( \gamma \) be a parametrized curve in \( X \). If \( \ell(\gamma) = 0 \), then \( \gamma \) is constant.

Proof: If \( \gamma \) is not constant, there are \( c, d \in [a, b], c < d \), with \( \gamma(c) \neq \gamma(d) \). Then \( 0 < \rho(\gamma(c), \gamma(d)) \leq \ell(\gamma) \).

Here is a more precise version of VI.4.1.25:
**VI.4.1.27.** Proposition. Let $\gamma : [a, b] \to X$ be a parametrized curve, and $c \in [a, b]$. Then
\[
\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}) .
\]

Proof: If $P$ is a partition of $[a, c]$ and $Q$ a partition of $[c, b]$, then $P \cup Q$ is a partition of $[a, b]$, and
\[
\ell(\gamma, P \cup Q) = \ell(\gamma|_{[a,c]}, P) + \ell(\gamma|_{[c,b]}, Q) .
\]
Every partition of $[a, b]$ containing $c$ (i.e. every sufficiently fine partition) arises in this manner. 

Iterating, we get:

**VI.4.1.28.** Corollary. Let $\gamma : [a, b] \to X$ be a parametrized curve, and $P = \{a = x_0, \ldots, x_n = b\}$ a partition of $[a, b]$. Then
\[
\ell(\gamma) = \sum_{k=1}^{n} \ell(\gamma|_{[x_{k-1}, x_k]}) .
\]

Rephrasing **VI.4.1.27.**, we get:

**VI.4.1.29.** Proposition. Let $\gamma_1 : [a, b] \to X$ and $\gamma_2 : [b, c] \to X$ with $\gamma_1(b) = \gamma_2(b)$, so that the concatenation $\gamma_1 + \gamma_2$ is defined. Then
\[
\ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2) .
\]

This can also be iterated to apply to arbitrary finite concatenations.

A crucial property of arc length is lower semicontinuity:

**VI.4.1.30.** Proposition. For each $n$, let $\gamma_n : [a, b] \to X$ be a parametrized curve from $p_n$ to $q_n$. Suppose $(\gamma_n)$ converges pointwise to a parametrized curve $\gamma : [a, b] \to X$ from $p$ to $q$, i.e. $\gamma_n(t) \to \gamma(t)$ for each $t \in [a, b]$ (so $p_n \to p$, $q_n \to q$). Then
\[
\ell(\gamma) \leq \liminf_{n \to \infty} \ell(\gamma_n) .
\]

We can have strict inequality, even if $\gamma_n \to \gamma$ uniformly on $[a, b]$ (**I.5.1.11.**).
Proof: First note that if \( P = \{a = t_0, \ldots, t_m = b\} \) is any partition of \([a, b]\), we have

\[
\ell(\gamma_n, P) = \sum_{k=1}^{m} \rho(\gamma_n(t_{k-1}), \gamma_n(t_k)) = \sum_{k=1}^{m} \rho(\gamma(t_{k-1}), \gamma(t_k)) = \ell(\gamma, P)
\]

by continuity of \( \rho \).

Suppose \( \ell(\gamma) = +\infty \). Fix \( M > 0 \). Then there is a partition \( P \) of \([a, b]\) with \( \ell(\gamma, P) > M \). Then

\[
M < \ell(\gamma, P) = \lim_{n \to \infty} \ell(\gamma_n, P) \leq \liminf_{n \to \infty} \ell(\gamma_n).
\]

This holds for all \( M > 0 \), so \( \liminf_{n \to \infty} \ell(\gamma_n) = +\infty \) and the inequality holds.

Now suppose \( \gamma \) is rectifiable. Let \( \epsilon > 0 \). Then there is a partition \( P \) of \([a, b]\) with \( \ell(\gamma, P) > \ell(\gamma) - \epsilon \). Then

\[
\ell(\gamma) - \epsilon < \ell(\gamma, P) = \lim_{n \to \infty} \ell(\gamma_n, P) \leq \liminf_{n \to \infty} \ell(\gamma_n).
\]

This holds for all \( \epsilon > 0 \), so \( \liminf_{n \to \infty} \ell(\gamma_n) \geq \ell(\gamma) \).

VI.4.1.31. Note for future reference that the definitions of \( \ell(\gamma, P) \) and \( \ell(\gamma) \) make perfect sense for a general function \( \gamma : [a, b] \to X \), not necessarily continuous, so “arc length” is defined even for discontinuous functions (although it is typically infinite for very discontinuous functions). None of the proofs of VI.4.1.21.–VI.4.1.30. use that \( \gamma \) is continuous, so these properties hold also for discontinuous \( \gamma \). (But VI.4.1.33. requires continuity of \( \gamma \).)

Continuity of Arc Length

VI.4.1.32. Let \( \gamma : [a, b] \to X \) be a rectifiable parametrized curve. Define a function \( \phi : [a, b] \to [0, \ell(\gamma)] \) by setting

\[
\phi(t) = \ell(\gamma|_{[a,t]}).
\]

Then \( \phi \) is nondecreasing, \( \phi(a) = 0 \), \( \phi(b) = \ell(\gamma) \). If \( \gamma \) is nonstationary, \( \phi \) is strictly increasing.

VI.4.1.33. Proposition. The function \( \phi \) is continuous.

Proof: The following lemma will provide left continuity on \([a, b]\). The simplest way to get right continuity is to apply the lemma also to \(-\gamma\) and use that \( \ell(\gamma|_{[a,t_0]}) = \ell(\gamma) - \ell(\gamma|_{[t_0,b]}) \).

VI.4.1.34. Lemma. Let \( \gamma : [a, b] \to X \) be a parametrized curve, not necessarily rectifiable. If \((t_n)\) is an increasing sequence in \([a, b]\) converging to \( t_0 \), then

\[
\ell(\gamma|_{[a,t_0]}) = \lim_{n \to \infty} \ell(\gamma|_{[a,t_n]}) = \sup_n \ell(\gamma|_{[a,t_n]})
\]

769
Proof: Define a sequence \((\gamma_n)\) of parametrized curves on \([a,t_0]\) by setting \(\gamma_n(t) = \gamma(t)\) if \(a \leq t \leq t_n\) and \(\gamma_n(t) = \gamma(t_n)\) if \(t_n \leq t \leq t_0\) (i.e. \(\gamma_n\) agrees with \(\gamma\) on \([a,t_n]\) and is stationary on \([t_n,t_0]\)). We have
\[
\ell(\gamma_n) = \ell(\gamma|_{[a,t_n]}) \leq \ell(\gamma|_{[a,t_0]})
\]
for any \(n\) by monotonicity (VI.4.1.25). On the other hand, since \(\gamma\) is continuous at \(t_0\), \(\gamma_n \to \gamma\) pointwise on \([a,t_0]\). So applying VI.4.1.30, we get
\[
\sup\limits_n \ell(\gamma|_{[a,t_n]}) \leq \ell(\gamma|_{[a,t_0]}) \leq \liminf_{n \to \infty} \ell(\gamma_n) = \liminf_{n \to \infty} \ell(\gamma|_{[a,t_n]})
\]
and the result follows.

Parametrization by Arc Length

VI.4.1.35. Let \(\gamma : [a, b] \to X\) be a nonstationary rectifiable parametrized curve, and let \(\phi\) be the above function. Then \(\phi\) is strictly increasing and continuous, so has a continuous strictly increasing inverse function \(\phi^{-1} : [0, \ell(\gamma)] \to [a, b]\). Set \(\tilde{\gamma} = \gamma \circ \phi^{-1} : [0, \ell(\gamma)] \to X\). Then \(\tilde{\gamma}\) is a strict reparametrization of \(\gamma\) with the property that \(\ell(\tilde{\gamma}|_{[0,t]}) = t\) for \(0 \leq t \leq \ell(\gamma)\), i.e. \(\tilde{\gamma}\) “traverses the path \((\gamma)\) at uniform unit speed.” The parametrized curve \(\tilde{\gamma}\) is called the reparametrization of \(\gamma\) by arc length. If \(\gamma\) is replaced by any strict reparametrization, the \(\tilde{\gamma}\) is unchanged; thus it depends only on \((\gamma)\), i.e. any rectifiable path has a unique canonical parametrization by arc length.

Length of Arcs

If \(C\) is a “curve” in \(X\), we would like to be able to say what the intrinsic “arc length” of \(C\) is. This is a tricky problem. The first question is what kind of subsets should be called “curves.” It is too broad to say that the range of any parametrized curve is a “curve” (). One reasonable possibility is:

VI.4.1.36. Definition. An arc in \(X\) is a subset of \(X\) homeomorphic to \([0,1]\).

VI.4.1.37. A subset \(C\) of \(X\) is an arc if and only if it is the range of an injective parametrized curve: a homeomorphism \(\gamma\) from \([0,1]\) to \(C\) is such a parametrized curve, and conversely an injective parametrized curve \(\gamma : [a, b] \to X\) with \(C = \gamma([a, b])\) is a homeomorphism by (). An arc thus has a unique equivalence class of oriented injective parametrizations, thus two in all; the arc length of all these parametrized curves is the same and can be taken to be the arc length of \(C\).

VI.4.1.38. Slightly more generally, if \(C\) can be written as a finite union of arcs with only endpoints in common, the arc length of \(C\) can be taken to be the sum of the arc lengths of these arcs. For more complicated subsets, there are grave difficulties in defining arc length this way.

VI.4.1.39. There is a way to unambiguously define “arc length” for general (Borel) subsets of \(X\): one-dimensional Hausdorff measure (). This “arc length” has the nice properties of a measure, generalizing VI.4.1.28. Subsets whose one-dimensional Hausdorff measure is positive and finite can be regarded as “one-dimensional” (cf. (); not every “one-dimensional” subset will have these properties).
VI.4.2. Length Metrics

VI.4.2.1. The metric distance between two points in a metric space may not have much relationship to the “traveling distance.” Here are some concrete examples, in part adapted from [BB101].

(1) What is the distance between New York and Sydney? The straight-line distance is about 11,000 km. This number is relevant if one is, say, triangulating observational data to compute the distance from Earth to Mars, but is useless for travel since it would involve tunneling through the earth. The great circle distance, which is about 16,000 km, is much more relevant.

(2) What is the distance between the top of the Eiger and the top of the Jungfrau? It is about 5 km “as the crow flies.” But a hiker/climber would have to go down into the valley and around or over the Mönch to get from one to the other, a much longer path. And the proper measure of the length of a route might be the time it takes to traverse it rather than the distance traveled.

(3) What is the distance between Boston and San Francisco? Even the great circle distance is not too relevant if one is driving; the road distance is the important consideration. So only certain paths are allowable for driving. And driving time might again be more important than distance – a longer route on interstate highways might be faster.

(4) On a more abstract level, if one lived in the Cantor set, it would not be possible to “move” or “travel” from one point to another even if these points were very nearby, if time is interpreted as a real number and motion must be continuous. Something like this could happen in the real world if people lived on a chain of small islands and did not have boats or know how to swim.

We will address (1), (2), and (4) in this section, and (3) in the next section.

Intrinsic Metrics

VI.4.2.2. We would like to describe when a metric actually represents “traveling distance” and show that every metric defines a new associated “traveling distance” metric. It will be convenient to work with extended metrics, i.e. ones allowed to take values in \([0, +\infty]\); in some of the literature on metric geometry, “metric” means “extended metric” (we will not make this convention). The next definition is reasonable:

VI.4.2.3. Definition. An extended metric \(\rho\) on a set \(X\) is strictly intrinsic, or a strict length metric, if, for every \(p, q \in X\), there is a path \(\gamma\) from \(p\) to \(q\) in \(X\) with \(\ell(\gamma) = \rho(p, q)\).

For technical reasons, it is convenient to slightly broaden the definition:

VI.4.2.4. Definition. An extended metric \(\rho\) on a set \(X\) is intrinsic, or a length metric, if, for every \(p, q \in X\), we have

\[
\rho(p, q) = \inf_{\gamma} \ell(\gamma)
\]

where the infimum is over all paths \(\gamma\) from \(p\) to \(q\) in \(X\) (with the convention that \(\inf(\emptyset) = +\infty\)).

771
VI.4.2.5. A strict length metric is a length metric (VI.4.1.24.), but not conversely. If \( \rho \) is a strict length metric, then \((X, \rho)\) is path-connected. If \( \rho \) is a length metric, then \((X, \rho)\) is not necessarily path-connected; but if \( p, q \in X \) and \( \rho(p, q) = +\infty \), then \( p \) and \( q \) are in different connected components of \((X, \rho)\) and hence there is no path in \( X \) from \( p \) to \( q \). So if \( \rho \) is an extended metric which is a length metric, the following are equivalent:

(i) \( \rho \) is a metric.

(ii) \((X, \rho)\) is path-connected.

(iii) \((X, \rho)\) is rectifiably path-connected, i.e. for any \( p, q \in X \), there is a rectifiable path from \( p \) to \( q \) in \( X \).

In particular, a strictly intrinsic metric must be an actual metric, not an extended metric. The notions of length and strict length metrics are designed for and almost exclusively used in the (rectifiably) path-connected case. In fact, see VI.4.2.8.

VI.4.2.6. The term “intrinsic” is intended to mean that the metric represents “traveling distance within the space.” For example, the arc-length metric on a circle is (strictly) intrinsic, but the Euclidean distance on a circle embedded in \( \mathbb{R}^2 \) is not intrinsic since one must travel “outside the space” to realize the Euclidean distance. Similarly, if the space is the surface of the earth (assumed to be a perfect sphere), the great-circle metric is intrinsic but not the straight-line (Euclidean) metric, which involves traveling through the earth and not within the space.

VI.4.2.7. Examples. (i) The Euclidean metric on \( \mathbb{R}^n \) is strictly intrinsic.

(ii) More generally, the norm metric on any normed vector space is strictly intrinsic, as is the restriction to any convex subset.

(iii) As noted above, the arc length metric on a circle, or the great circle metric on a sphere, is strictly intrinsic. But the restriction of the Euclidean metric to an embedded circle or sphere is not intrinsic.

(iv) Let \( X = \mathbb{R}^2 \setminus \{(0,0)\} \). Then the restriction of the usual Euclidean metric to \( X \) is intrinsic but not strictly intrinsic: \((1,0)\) and \((-1,0)\) cannot be connected by a path of minimum length in \( X \). See VI.4.5.1. for an example of a complete metric which is intrinsic but not strictly intrinsic.

(v) The discrete extended metric \( \rho \) on any set, where \( \rho(p, q) = +\infty \) if \( p \neq q \), is intrinsic.

VI.4.2.8. Proposition. Let \( \rho \) be an intrinsic metric on a set \( X \). Then \((X, \rho)\) is locally path-connected. In fact, every open ball in \((X, \rho)\) is path-connected.

Proof: Let \( p \in X \) and \( \epsilon > 0 \). Let \( q \in B_\epsilon(p) \). Since \( \rho(p, q) < \epsilon \), there is a parametrized curve \( \gamma : [a, b] \to X \) from \( p \) to \( q \) with \( \ell(\gamma) < \epsilon \). If \( c \in [a, b] \), then by VI.4.1.25.

\[
\rho(p, \gamma(c)) \leq \ell(\gamma|_{[a,c]}) \leq \ell(\gamma) < \epsilon
\]

so the range of \( \gamma \) is contained in \( B_\epsilon(p) \). Thus \( B_\epsilon(p) \) is path-connected. 

772
The Length Metric of a Metric

Now suppose \( \rho \) is a general metric (or extended metric) on a space \( X \). We want to define a new (extended) metric on \( X \) which describes “traveling distance” in \( (X, \rho) \).

VI.4.2.9. Definition. Let \((X, \rho)\) be a metric space. If \( p, q \in X \), define

\[
\rho_1(p, q) = \inf_{\gamma} \ell(\gamma)
\]

where the infimum is over all paths \( \gamma \) from \( p \) to \( q \) in \( X \) (with the convention that \( \inf(\emptyset) = +\infty \)).

The function \( \rho_1 \) is called the length metric of \( \rho \); the terminology will be justified below.

VI.4.2.10. Proposition. The function \( \rho_1 \) is an extended metric on \( X \).

Proof: We need only verify the triangle inequality. Let \( p, q, r \in X \); we show \( \rho_1(p, r) \leq \rho_1(p, q) + \rho_1(q, r) \). If \( \rho_1(p, q) \) or \( \rho_1(q, r) \) is \(+\infty\), there is nothing to prove, so assume both are finite. Let \( \epsilon > 0 \). There are paths \( \gamma_1 \) from \( p \) to \( q \) and \( \gamma_2 \) from \( q \) to \( r \) with \( \ell(\gamma_1) < \rho_1(p, q) + \epsilon \) and \( \ell(\gamma_2) < \rho_1(q, r) + \epsilon \). Reparametrizing if necessary, we can concatenate \( \gamma_1 \) and \( \gamma_2 \) (VI.4.1.16.) to get a path \( \gamma_1 + \gamma_2 \) from \( p \) to \( r \). Thus

\[
\rho_1(p, r) \leq \ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2) < \rho_1(p, q) + \rho_1(q, r) + 2\epsilon
\]

Since \( \epsilon > 0 \) is arbitrary, the desired inequality follows.

VI.4.2.11. Even if \( \rho \) is a metric, \( \rho_1 \) is an extended metric in general. For example, if \((X, \rho)\) is not path-connected, then \( \rho_1 \) is necessarily extended: \( \rho_1(p, q) = +\infty \) if \( p \) and \( q \) are in different path components. Even if \((X, \rho)\) is path-connected, \( \rho_1 \) can be an extended metric. In fact, \( \rho_1 \) is a metric if and only if \((X, \rho)\) is rectifiably path-connected.

VI.4.2.12. We have \( \rho(p, q) \leq \rho_1(p, q) \) for any \( p, q \) by VI.4.1.24., i.e. \( \rho_1 \) dominates \( \rho \), so the topology of \( \rho_1 \) is stronger than the topology of \( \rho \). The metrics \( \rho \) and \( \rho_1 \) are equivalent in good cases, but not in general, i.e. the \( \rho_1 \) topology can be strictly stronger than the \( \rho \)-topology. If \( \rho \) is intrinsic, then \( \rho_1 = \rho \); the converse is also true (VI.4.2.14.).

VI.4.2.13. Proposition. Let \((X, \rho)\) be a metric space, and \( \rho_1 \) the length metric of \( \rho \). Then for any parametrized curve \( \gamma : [a, b] \rightarrow X \) in \((X, \rho)\), we have \( \ell_{\rho_1}(\gamma) = \ell_\rho(\gamma) \). 

Note that there is a subtlety: a continuous function from \([a, b]\) to \((X, \rho)\) is not obviously continuous as a function from \([a, b]\) to \((X, \rho_1)\) (and is not continuous in general). But by VII.3.1.17. arc length is also defined for discontinuous \( \gamma \). In fact, we will not use continuity of \( \gamma \) in the proof.

Proof: Fix \( \gamma \). Since \( \rho \leq \rho_1 \), for any partition \( \mathcal{P} \) of \([a, b]\) we have \( \ell_\rho(\gamma, \mathcal{P}) \leq \ell_{\rho_1}(\gamma, \mathcal{P}) \), and hence \( \ell_\rho(\gamma) \leq \ell_{\rho_1}(\gamma) \). Conversely, let \( \epsilon > 0 \). If \( \mathcal{P} = \{a = t_0, \ldots, t_n = b\} \) is a partition of \([a, b]\), set \( p_k = \gamma(t_k) \) for \( 0 \leq k \leq n \). Then

\[
\ell_{\rho_1}(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho_1(p_{k-1}, p_k) \leq \sum_{k=1}^{n} \ell_\rho(\gamma|_{[t_{k-1}, t_k]}) = \ell_\rho(\gamma)
\]
VI.4.1.28.}, and taking the supremum over all $\mathcal{P}$ we obtain $\ell_{\rho_i}(\gamma) \leq \ell_\rho(\gamma)$.

VI.4.2.14. Corollary. Let $(X, \rho)$ be a metric space, and $\rho_l$ the length metric of $\rho$. Then $\rho_l$ is intrinsic. Thus $(\rho_l)_l = \rho_l$, and $\rho = \rho_l$ if and only if $\rho$ is intrinsic.

VI.4.2.15. It can be shown using VI.4.2.13. and VI.4.1.33. that if $\gamma$ is a rectifiable parametrized curve in $(X, \rho)$, i.e. $\gamma$ is a continuous function from $[a, b]$ to $(X, \rho)$ with $\ell_\rho(\gamma) < +\infty$, then $\gamma$ is also continuous as a function from $[a, b]$ to $(X, \rho_l)$. This can be false if $\gamma$ is not rectifiable (VI.4.2.16.(iv)).

VI.4.2.16. Examples. (i) If $X$ is a circle embedded in $\mathbb{R}^2$ and $\rho$ is the restriction of the Euclidean metric, then $\rho_l$ is the arc length metric. Similarly, if $X$ is a sphere embedded in $\mathbb{R}^3$ and $\rho$ is the Euclidean metric, then $\rho_l$ is the great circle metric. In these cases $\rho_l \neq \rho$, but the metrics are equivalent.

(ii) Let $X$ be the Warsaw circle (XI.18.4.4.), and $\rho$ the restriction of the Euclidean metric in $\mathbb{R}^2$. Since $(X, \rho)$ is not locally path-connected and $(X, \rho_l)$ is (VI.4.2.8.), the metrics cannot be equivalent. In fact, $(X, \rho_l)$ is homeomorphic to a half-open interval.

(iii) Let $(X, \rho)$ be a totally disconnected metric space, e.g. the Cantor set with the usual metric. The metric $\rho_l$ is the discrete extended metric (VI.4.2.7.(v)), so the topology from $\rho_l$ is the discrete topology.

(iv) Let $X$ be $[0, 1]$ with the snowflake metric $\rho^s$ for some $s$, $0 < s < 1$ (VI.4.1.20.(ii)). Then $X$ is path-connected and locally path-connected, but has no nonconstant rectifiable paths. The metric $(\rho^s)_l$ is the discrete extended metric, so the topology from $\rho_l$ is the discrete topology. Thus the topology from $\rho_l$ can depend on the metric $\rho$ and not just on the topology from $\rho$.

Local Characterization of Intrinsic Metrics

Intrinsic metrics are characterized locally, and are in an appropriate sense the only locally determined metrics.

VI.4.2.17. Definition. Let $\rho$ and $\sigma$ be extended metrics on a set $X$. The metric $\rho$ locally dominates $\sigma$ if, for each $x \in X$, there is a $\rho$-open ball $B_x$ centered at $x$ such that $\sigma(y, z) \leq \rho(y, z)$ for all $y, z \in B_x$. The metric $\rho$ uniformly locally dominates $\sigma$ if there is an $\epsilon > 0$ such that, whenever $B$ is a ball of $\rho$-radius $\epsilon$ in $X$, $\sigma \leq \rho$ on $B$.

The extended metric $\rho$ is locally determined if, whenever $\sigma$ is an extended metric which is locally dominated by $\rho$, then $\sigma \leq \rho$.

VI.4.2.18. Examples. (i) The usual metric on $\mathbb{R}$, or any subset of $\mathbb{R}$ (e.g. $\mathbb{Q}$), is locally determined.

(ii) The arc length metric $\sigma$ on a circle in $\mathbb{R}^2$ is locally determined. But the Euclidean metric $\rho$ is not, since $\rho$ locally dominates $(1 - \epsilon)\sigma$ for every $\epsilon > 0$ but does not dominate if $\epsilon$ is small enough.

(iii) If a small open arc (less than half the circumference) is removed from a circle, the restriction of arc length to the remaining space is not locally determined.

(iv) The discrete metric $\sigma$ on any set $X$ is locally determined.

774
VI.4.2.19. **Proposition.** Let \((X, \rho)\) be a metric space. If \(\rho\) is intrinsic, it is locally determined.

**Proof:** Suppose \(\rho\) is intrinsic, and \(\sigma\) is a pseudometric which is locally dominated by \(\rho\). Let \(x, y \in X\); we show \(\sigma(x, y) \leq \rho(x, y)\). We may assume \(\rho(x, y) < +\infty\). Fix \(\epsilon > 0\). Let \(\gamma : [0, 1] \to X\) be a path from \(x\) to \(y\) with \(\ell(\gamma) < \rho(x, y) + \epsilon\). For \(z \in \gamma([0, 1])\), there is an open neighborhood \(U_z\) of \(z\) with \(\sigma(u, v) \leq \rho(u, v)\) for all \(u, v \in U_z\). Since \(\gamma([0, 1])\) is compact, there are finitely many such \(U_z\) which cover it, so there is a partition \(\{0 = t_0, \ldots, t_n = 1\}\) of \([0, 1]\) such that, if \(x_k = \gamma(t_k)\), consecutive \(x_k\) are in the same \(U_z\). Thus we have

\[
\sigma(x, y) \leq \sum_{k=1}^n \sigma(x_{k-1}, x_k) \leq \sum_{k=1}^n \rho(x_{k-1}, x_k) \leq \ell(\gamma) < \rho(x, y) + \epsilon
\]

and since \(\epsilon > 0\) is arbitrary, \(\sigma(x, y) \leq \rho(x, y)\).

VI.4.2.20. The converse is not true, even if \(\rho\) is complete (VI.4.2.18. (iv)). To obtain an appropriate converse, we make the following construction. Let \(\rho\) be an extended metric on a set \(X\) and \(\epsilon > 0\). Set

\[
\rho_{\epsilon}(x, y) = \inf \sum_{k=1}^n \rho(x_{k-1}, x_k)
\]

where the infimum is over all finite chains \(\{x = x_0, x_1, \ldots, x_n = y\}\) where \(\rho(x_{k-1}, x_k) \leq \epsilon\) for all \(k\) (set \(\rho_{\epsilon}(x, y) = +\infty\) if there is no such chain). It is easily proved by an argument almost identical to () that \(\rho_{\epsilon}\) is an extended metric, that \(\rho \leq \rho_{\epsilon} \leq \rho_{\delta}\) if \(\delta < \epsilon\). We have that \(\rho_{\epsilon}(x, y) = \rho(x, y)\) if \(\rho(x, y) \leq \epsilon\), and in particular \(\rho\) uniformly locally dominates \(\rho_{\epsilon}\) for any \(\epsilon > 0\).

VI.4.2.21. By (), \(\rho_{\text{loc}} = \sup_{\epsilon > 0} \rho_{\epsilon}\) is an extended metric, called the **local metric** of \(\rho\). Comparing definitions, we have \(\rho_{\text{loc}} \leq \rho_\epsilon\). Since \((\rho_\epsilon)_\delta = \rho_\delta\) if \(\delta \leq \epsilon\), it follows that

\[
(\rho_{\epsilon})_{\text{loc}} = (\rho_{\text{loc}})_\epsilon = (\rho_{\text{loc}})_{\text{loc}} = \rho_{\text{loc}}
\]

for any \(\epsilon > 0\).

The next proposition is an immediate corollary of VI.4.2.19., or it can be easily proved directly.

VI.4.2.22. **Proposition.** Let \((X, \rho)\) be a metric space. If \(\rho\) is intrinsic, then \(\rho = \rho_{\epsilon}\) for all \(\epsilon > 0\).

VI.4.2.23. **Definition.** Let \(\rho\) be a metric on a set \(X\). Then \(\rho\) is **uniformly locally determined** if \(\rho = \rho_{\epsilon}\) for all \(\epsilon > 0\), i.e. if \(\rho = \rho_{\text{loc}}\).

VI.4.2.24. **Examples.**

(i) The usual metric on \(\mathbb{Q}\) (or any other subset of \(\mathbb{R}\)) is uniformly locally determined.

(ii) The discrete metric on a set with more than one element is locally determined but not uniformly locally determined.

(iii) For any metric \(\rho\), \(\rho_{\text{loc}}\) is uniformly locally determined (VI.4.2.21.).

775
VI.4.2.25. Theorem. Let \((X, \rho)\) be a complete metric space. Then \(\rho_1 = \rho_{loc}\).

VI.4.2.26. Corollary. Let \((X, \rho)\) be a complete metric space. If \(\rho\) is uniformly locally determined, it is intrinsic.

For the proof of the theorem, we will use a variant of the Between Axiom (VI.3.1.11.) which is of independent interest:

VI.4.2.27. Definition. A metric space \((X, \rho)\) has the approximate midpoint property if, for every \(x, y \in X\) and \(\epsilon > 0\) there is a \(z \in X\) with

\[
\rho(x, z) - \frac{\rho(x, y)}{2} < \epsilon \quad \text{and} \quad \rho(z, y) - \frac{\rho(x, y)}{2} < \epsilon.
\]

Such a \(z\) is called an \(\epsilon\)-midpoint of \(x\) and \(y\).

VI.4.2.28. Proposition. A uniformly locally determined metric has the approximate midpoint property.

Proof: If \((X, \rho)\) is locally determined, let \(x, y \in X\) and \(\epsilon > 0\). If \(\{x = x_0, \ldots, x_n = y\}\) is a finite chain with \(\rho(x_{k-1}, x_k) < \frac{\epsilon}{2}\) for all \(k\) and \(\sum_k \rho(x_{k-1}, x_k) < \rho(x) + \frac{\epsilon}{2}\), then one of the \(x_k\) (for \(k \approx \frac{n}{2}\)) will be an \(\epsilon\)-midpoint of \(x\) and \(y\).

VI.4.2.29. Theorem. Let \((X, \rho)\) be a complete metric space with the approximate midpoint property. Then \(\rho\) is intrinsic.

Proof: This is somewhat similar to ( ), but no form of choice is needed. Let \(x, y \in X\) and \(\epsilon > 0\). Let \(x_{1/2}\) be an \(\frac{\epsilon}{4}\)-midpoint of \(x\) and \(y\). Let \(x_{1/4}\) be an \(\frac{\epsilon}{6}\)-midpoint of \(x = x_0\) and \(x_{1/2}\), and \(x_{3/4}\) an \(\frac{\epsilon}{16}\)-midpoint of \(x_{1/2}\) and \(y = x_1\). Continue inductively in this manner to define \(x_r\) for dyadic rational \(r \in [0, 1]\), using a denominator of \(4^n\) if the denominator of \(r\) is \(2^n\). It is easy to check that \(r \mapsto x_r\) is uniformly continuous and thus extends to a map \(\gamma : [0, 1] \to X\) since \(\rho\) is complete. Then \(\gamma\) is a path from \(x\) to \(y\), and \(\ell(\gamma) \leq \rho(x, y) + \epsilon\).

Theorem VI.4.2.25., and hence Corollary VI.4.2.26., follow immediately.

VI.4.2.30. The example of \(\mathbb{Q}\) (VI.4.2.24.(i)) shows that completeness is necessary in the theorem.

VI.4.3. Length Structures

We now change point of view. Instead of working with arc length defined by a metric, we assume directly that we have a notion of length for certain (“admissible”) paths in the space \(X\) and from it derive a length metric and thus a topology. We must start with a metric, or at least a topology, on \(X\), in order to have a notion of continuous path in \(X\), but the initial topology will be inconsequential at the end. In order to do this, there must be enough admissible paths, and the length function must have reasonable properties. Our exposition is largely adapted from [BBI01].
VI.4.3.1. **Definition.** Let \((X, \sigma)\) be a metric space. A length structure on \((X, \sigma)\) is a pair \(L = (A, \ell)\), where \(A\) is a set of parametrized curves in \(X\) (called admissible paths for \(L\)) such that

1. If \(\gamma : [a, b] \to X\) is in \(A\), then \(\gamma|_{[c, d]} \in L\) for every subinterval \([c, d]\) of \([a, b]\).
2. If \(\gamma_1, \gamma_2 \in A\) and \(\gamma_1 + \gamma_2\) is defined, then \(\gamma_1 + \gamma_2 \in A\).
3. Any linear reparametrization of a parametrized curve in \(A\) is also in \(A\). (Combined with (3), any piecewise-linear reparametrization of a parametrized curve in \(A\) is then also in \(A\).) Also, if \(\gamma \in A\), then its reversal \(-\gamma\) is in \(A\) too.

and \(\ell : A \to [0, +\infty]\) (called the length function of \(L\)) satisfying

1. If \(\gamma : [a, b] \to X\) is in \(A\) and \(c \in [a, b]\), then \(\ell(\gamma) = \ell(\gamma|_{[a, c]}) + \ell(\gamma|_{[c, b]})\) (i.e. \(\ell\) is additive on concatenations).
2. If \(\gamma : [a, b] \to X\) is in \(A\), then the function \(\phi(t) = \ell(\gamma|_{[a, t]}\) is continuous on \([a, b]\).
3. If \(\gamma \in A\) and \(\tilde{\gamma}\) is a linear reparametrization of \(\gamma\), then \(\ell(\tilde{\gamma}) = \ell(\gamma)\). (From (i) it then follows that \(\ell\) is invariant under piecewise-linear reparametrizations, i.e. \(\ell\) is really defined on equivalence classes of parametrized curves.) Also, \(\ell(-\gamma) = \ell(\gamma)\) for any \(\gamma \in A\).
4. If \(p \in X\) and \(U\) is an open neighborhood of \(p\) (in the \(\sigma\)-topology), then there is a \(\delta > 0\) such that any \(\gamma \in A\) from \(p\) to a point \(q \in U^\circ\) satisfies \(\ell(\gamma) \geq \delta\).

The set \(A\) of admissible paths is technically encoded in the length function \(\ell\) as its domain, and some of the literature does not distinguish between \(L\) and \(\ell\); but it seems a little clearer to make a notational distinction.

VI.4.3.2. **Condition (iv) is an additional compatibility condition of \(L\) with the \(\sigma\)-topology on \(X\) besides the requirement that the curves in \(A\) are continuous.** These two compatibility conditions are the only place the metric \(\sigma\) appears, and only depend on the \(\sigma\)-topology; the metric \(\sigma\) could be dispensed with entirely and just replaced by an initial topology on \(X\) (which should be assumed Hausdorff for a reasonable theory, although it could be dispensed with entirely by taking it to be the indiscrete topology).

VI.4.3.3. **Examples.** (i) Let \((X, \rho)\) be a metric space, \(A\) the collection of all parametrized curves in \(X\), and \(\ell(\gamma) = \ell_\rho(\gamma)\). This is the length structure induced by \(\rho\), denoted \(L_\rho\).

(ii) More generally, let \((X, \rho)\) be a metric space and \(Y \subseteq X\). Let \(A\) be the set of parametrized curves in \(X\) whose range is contained in \(Y\), and for \(\gamma \in A\) set \(\ell(\gamma) = \ell_\rho(\gamma)\). Then \((A, \ell)\) is a length structure on \((X, \rho)\), as well as on \((Y, \rho)\). This was the main motivating example for the theory of length structures, with \(X = \mathbb{R}^n\), \(\rho\) the Euclidean metric, and \(Y\) a smooth submanifold.

(iii) Let \((X, \rho)\) be \(\mathbb{R}^n\) with the Euclidean metric, and take \(A\) to be either the set of piecewise-linear curves or piecewise-smooth curves in \(\mathbb{R}^n\), with \(\ell = \ell_\rho|_A\). These length structures are not essentially different from the length structure using all parametrized curves. One can also consider the restrictions to suitable subsets as in (ii).

(iv) Let \((X, \rho)\) be \(\mathbb{R}^n\) with the Euclidean metric, and take \(A\) to be the set of all parametrized curves in \(\mathbb{R}^n\). Set \(\ell(\gamma) = \ell_\rho(\gamma)\) if \(\gamma\) is piecewise-linear and \(\ell(\gamma) = +\infty\) otherwise.

(v) Let \(X = \mathbb{R}^2\), \(\sigma\) the Euclidean metric. Let \(A\) be the collection of all parametrized curves whose image consists of a union of finitely many horizontal and vertical line segments. For \(\gamma \in A\), set \(\ell(\gamma) = \ell_\sigma(\gamma)\).
Here is a concrete example, stated somewhat imprecisely. Let $X$ be the United States, and $A$ the set of driving routes (so all points, including endpoints, are points on roads). The length $\ell(\gamma)$ of a driving route is the driving time for the route. (Driving routes are taken as paths, i.e. equivalence classes of parametrized curves, and driving time is not connected with parametrization.) This must be idealized to become precise and give a length structure: driving time must be well defined and be independent of direction, and it must be assumed that a vehicle can reverse direction instantaneously to make driving time additive on curves which retrace routes.

In this example, not all points of the space are connected by admissible paths. To rectify this, we could add to $A$ the set of all routes traversable by surface transportation, including walking (assuming that every point of the country is within walking distance of a road).

As a precise theoretical version of (vi), let $X$ be $\mathbb{R}^n$ (or a smooth submanifold) and $f : X \to (0, +\infty)$ a continuous function (continuity can be somewhat relaxed). Let $A$ be the set of piecewise-smooth parametrized curves in $X$, and for $\gamma : [a, b] \to X$ in $A$, set

$$\ell(\gamma) = \int_a^b f(\gamma(t)) \gamma'(t) \, dt$$

(this is just the integral of $f$ along the path defined by $\gamma$). In the case of (v), $X$ could be the set of all points on roads, and $f(x)$ the reciprocal of the maximum speed one can (or is allowed to) drive at the point $x$.

As a generalization, let $X$ be $\mathbb{R}^n$ (or a smooth submanifold) and $f : X \times \mathbb{R}^n \to (0, +\infty)$ a continuous function (continuity can be somewhat relaxed). Let $A$ be the set of piecewise-smooth parametrized curves in $X$, and for $\gamma : [a, b] \to X$ in $A$, set

$$\ell(\gamma) = \int_a^b f(\gamma(t), \gamma'(t)) \, dt$$

To make $\ell$ invariant under linear reparametrizations, we need $f(x, \alpha y) = |\alpha| f(x, y)$ for all $x, y, \alpha$. Such a length structure is called a Finslerian length structure. The idea is that the speed of travel may depend not only on the point, but also on the direction of travel (although the speed in opposite directions must be the same, not necessarily a realistic assumption in applications).

The Length Metric of a Length Structure

**Definition.** Let $L = (A, \ell)$ be a length structure on a metric space $(X, \sigma)$. Define an extended metric $\rho_L$ by

$$\rho_L(p, q) = \inf_{\gamma} \ell(\gamma)$$

for $p \neq q$, where the infimum is over all paths $\gamma \in A$ from $p$ to $q$ (with the convention that $\inf(\emptyset) = +\infty$).

The proof that $\rho_L$ is an extended metric is essentially identical to the proof of VI.4.2.10., using that a concatenation of admissible paths is admissible. The extended metric $\rho_L$ is a metric if and only if any two points of $X$ can be connected by an admissible path of finite length. The $\rho_L$-topology on $X$ is stronger than the initial $\sigma$-topology by condition (iv); it is often strictly stronger.
VI.4.3.5. **Proposition.** Let \( L = (A, \ell) \) be a length structure on \((X, \sigma)\), and \( \rho_L \) the corresponding extended metric. Then \( \rho_L \) is a length metric, and \( \ell_{\rho_L}(\gamma) \leq \ell(\gamma) \) for all \( \gamma \in A \).

We can have strict inequality (VI.4.3.6.(iv)).

**Proof:** For the second statement, let \( \gamma : [a, b] \to X \) be in \( A \). If \( \mathcal{P} = \{a = t_0, \ldots, t_n = b\} \) is any partition of \([a, b]\), we have

\[
\ell_{\rho_L}(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho_L(\gamma(t_{k-1}), \gamma(t_k)) \leq \sum_{k=1}^{n} \ell(\gamma|_{[t_{k-1}, t_k]}) = \ell(\gamma)
\]

since \( \gamma|_{[t_{k-1}, t_k]} \in A \) by (1), and taking the supremum over all \( \mathcal{P} \) we obtain the inequality.

To show \( \rho_L \) is a length metric, let \( p, q \in X \). We show that

\[
\rho_L(p, q) = \lambda := \inf_{\gamma} \ell_{\rho_L}(\gamma)
\]

with the infimum over all parametrized curves from \( p \) to \( q \). We need only show \( \rho_L(p, q) \geq \lambda \), since \( \leq \) is automatic (VI.4.1.24). If \( \rho_L(p, q) = +\infty \), there is nothing to prove, so we may suppose \( \rho_L(p, q) \) is finite.

Let \( \epsilon > 0 \). There is a \( \gamma \in A \) from \( p \) to \( q \) with \( \ell(\gamma) < \rho_L(p, q) + \epsilon \). Then

\[
\lambda \leq \ell_{\rho_L}(\gamma) \leq \ell(\gamma) < \rho_L(p, q) + \epsilon .
\]

Since \( \epsilon > 0 \) is arbitrary, the inequality follows.

VI.4.3.6. **Examples.** (i) If \((X, \rho)\) is a metric space and \( L_\rho \) is the induced length structure (VI.4.3.3.(i)), then \( \rho_{L_\rho} = \rho \).

(ii) In the situation of VI.4.3.3.(ii), if \( L \) is regarded as a length structure on \((Y, \rho)\), then \( \rho_L = \rho \) as in (i). If \( L \) is regarded as a length structure on \((X, \rho)\), then \( \rho_L \) is an extended metric on \( X \), with \( \rho_L(p, q) = +\infty \) unless both \( p \) and \( q \) are in \( Y \). In the \( \rho_L \) topology on \( X \), \( Y \) is a clopen set and each point of \( X \setminus Y \) is an isolated point.

(iii) In VI.4.3.3.(iii), the induced metric \( \rho_L \) on \( \mathbb{R}^n \) is just the Euclidean metric, since the set of admissible curves includes a curve of minimal length between any two points. Thus restricting the set of admissible curves does not necessarily change the length metric, so long as the admissible curves include arbitrarily short paths (i.e. close to the minimum for all paths) between any pair of points.

(iv) In VI.4.3.3.(iv), the induced metric \( \rho_L \) on \( \mathbb{R}^n \) is again just the Euclidean metric. Thus, even if the set of admissible curves is the same, changing the length function \( \ell \) does not necessarily change the length metric, if there are short paths of (approximately) the same length under the two length functions.

(v) In VI.4.3.3.(v), the length metric is not the Euclidean metric \( \rho = \rho_2 \), but rather the 1-metric \( \rho_1 \) (\( \rho_1 \)). Thus if the set of admissible curves is changed, the length metric can change if the set of admissible curves does not contain arbitrarily short paths between points.
Recovering a Length Structure from its Metric

VI.4.3.7. If $\rho$ is a length metric, then the induced metric from the length structure $L_\rho$ is $\rho$ itself (cf. VI.4.3.6.(i)). How about the converse: if $L = (A, \ell)$ is a length structure and $\rho_L$ the induced length metric, is $L = L_{\rho_L}$? In one sense, the answer is usually no: the set of admissible paths for $L_{\rho_L}$ is often larger than $A$ (cf. Examples VI.4.3.6.(iii) and (v)). But $\ell_{\rho_L}$ may not agree with $\ell$ even on $A$ (VI.4.3.6.(iv); the problem is lower semicontinuity (VI.4.3.9)). It would be useful to know when $\ell_{\rho_L}$ agrees with $\ell$ on $A$, since $\ell$ is often given by a formula making it easier to calculate than $\ell_{\rho_L}$ (e.g. in VI.4.3.3.(vii)–(viii)).

VI.4.3.8. The length function $\ell_{\rho_L}$ is always lower semicontinuous (VI.4.1.30.), so a necessary condition that $\ell = \ell_{\rho_L}|_A$ for a length structure $L = (A, \ell)$ is that $\ell$ is lower semicontinuous on $A$. But $\ell$ is not necessarily lower semicontinuous, e.g. the $\ell$ of Example VI.4.3.3.(iv) (the $\ell$ of Example VI.4.3.3.(viii) is also not necessarily lower semicontinuous, cf. (i)). Lower semicontinuity is the only constraint:

VI.4.3.9. **Proposition.** Let $L = (A, \rho)$ be a length structure on $(X, \sigma)$, and $\rho_L$ the corresponding length metric. Then $\ell = \ell_{\rho_L}|_A$ if and only if $\ell$ is lower semicontinuous on $A$.

VI.4.4. Metric Geometry

VI.4.5. Exercises

VI.4.5.1. For each $n$ let $I_n$ be a closed interval in $\mathbb{R}$ of length $1 + \frac{1}{n}$ with endpoints $a_n$ and $b_n$. Glue the $I_n$ together by identifying all the $a_n$ to a point $a$, and all the $b_n$ to a point $b$. Let $X$ be the quotient space, and $\rho$ the length metric on $X$.

(a) Show that $(X, \rho)$ is a complete metric space.

(b) Show that $\rho(a, b) = 1$, but there is no path from $a$ to $b$ of length 1. Thus $(X, \rho)$ is not complete as a length space, i.e. $\rho$ is not strictly intrinsic.

(c) Realize $X$ concretely as a subset of $\mathbb{R}^2$ consisting of a union of circular arcs in the upper half plane through $(0, 0)$ and $(1, 0)$. Show that $\rho$ is the length metric defined by the restriction $\sigma$ of the Euclidean metric on $\mathbb{R}^2$, and that $\rho$ is equivalent (but not uniformly equivalent) to the incomplete metric $\sigma$.

VI.4.5.2. Let $(X, \rho)$ be a metric space, and $\rho_l$ the associated length metric on $X$.

(a) Show that if $\rho$ is complete, then $\rho_l$ is also complete.

(b) If $\rho_l$ is complete, then $\rho$ is not necessarily complete, even if $(X, \rho)$ is rectifiably path-connected and locally path-connected. [Let $X$ be the topologist’s sine curve (XI.18.4.3.) with the points on the $y$-axis removed, with $\rho$ the ordinary metric from $\mathbb{R}^2$. $\mathbb{Q}$ with the ordinary metric is an example if the space is not required to be path-connected.]
Chapter VII

Curves

“Everyone knows what a curve is, until he has studied enough mathematics to become confused through the countless number of possible exceptions.”

*Felix Klein*

What is a curve? Almost everyone has a good intuitive idea of what a curve should be, but it is a surprisingly delicate matter to make a precise definition which includes all the things which should be curves and excludes everything that should not.

At a minimum, a curve should be a one-dimensional subset of a Euclidean space. Most people would agree that the following are curves:

- Graphs of continuous functions, or at least graphs of differentiable functions, from $\mathbb{R}$ to $\mathbb{R}$.
- Lines, circles, and polygons.
- A figure 8, figure 6, figure $\alpha$, and the like.
- Finite separated unions of the above.

On the other hand, the following are clearly not curves:

- Anything zero-dimensional, e.g. the Cantor set $\mathcal{C}$.
- Anything more than one-dimensional, e.g. a (solid) disk or square, or a ball or sphere in $\mathbb{R}^n$, $n > 2$.
- Anything not at least locally compact.

In addition, there are many things which are in a gray area, e.g. a figure $X$, figure $Y$, figure $\otimes$, and many of the the one-dimensional subsets of $\mathbb{R}^2$ and $\mathbb{R}^3$ discussed in (), such as the Topologist’s Sine Curve, the Hawaiian Earring, the Sierpiński gasket and carpet, solenoids, the pseudo-arc, even wildly embedded arcs.

We will begin by defining curves via parametrizations, although we will distinguish between curves and parametrized curves. Parametrizations are natural since they describe “motion along the curve.” But defining a curve as just a continuous image of an interval is too broad: for example, there are “space-filling curves,” continuous functions from $[0,1]$ onto a square $[0,1]^2$. In fact, the Banach-Mazurkiewicz Theorem () says
that every compact metrizable topological space which is connected and locally connected, and in particular any connected, locally connected closed bounded subset of a Euclidean space, is a continuous image of $[0, 1]$. Allowing continuous images of non-closed intervals, e.g. $\mathbb{R}$, admits even more pathology: a continuous image of $\mathbb{R}$, even a one-to-one continuous image, need not even be locally compact (). On the other hand, requiring the parametrization to be a homeomorphism eliminates such a natural curve as a circle, and to relax the requirement to being a local homeomorphism rules out curves like a figure 8. So we will have to be more careful.

VII.1. Curves and Parametrized Curves in $\mathbb{R}^n$

VII.1.1. Parametrized Curves

**Definition.** A parametrized curve in $\mathbb{R}^n$ is a continuous function $\gamma : [a, b] \rightarrow \mathbb{R}^n$ for some closed bounded interval $[a, b]$ in $\mathbb{R}$. The points $p = \gamma(a)$ and $q = \gamma(b)$ are the initial and final endpoints of $\gamma$ respectively, and $\gamma$ is a parametrized curve from $p$ to $q$.

Note that the interval $[a, b]$ is allowed to be degenerate, i.e. we allow $a = b$. (To be sure, parametrized curves defined on degenerate intervals are not very interesting!)

**Definition.** A parametrized curve has a direction. If $\gamma : [a, b] \rightarrow \mathbb{R}^n$ is a parametrized curve from $p$ to $q$, its negative or reversal is $-\gamma : [a, b] \rightarrow \mathbb{R}^n$ defined by $(-\gamma)(t) = \gamma(a + b - t)$ is a parametrized curve from $q$ to $p$.

**Definition.** We may think of the range of a parametrized curve as a “curve” in $\mathbb{R}^n$ (although it need not look anything like a geometric curve), and the parametrization as giving “motion” along this curve. This motion may “retrace” parts of the curve in a complicated way.

**Restriction and Concatenation**

**Definition.** Let $\gamma : [a, b] \rightarrow \mathbb{R}^n$ be a parametrized curve, and $[c, d]$ a closed subinterval of $[a, b]$, i.e. $a \leq c \leq d \leq b$. The parametrized curve $\gamma|_{[c, d]}$ is the restriction of $\gamma$ to $[c, d]$; it is a parametrized curve from $\gamma(c)$ to $\gamma(d)$.

**Definition.** If $\gamma_1 : [a, b] \rightarrow \mathbb{R}^n$ is a parametrized curve from $p$ to $q$, and $\gamma_2 : [b, c] \rightarrow \mathbb{R}^n$ is a parametrized curve from $q$ to $r$, the concatenation is the parametrized curve $\gamma = \gamma_1 + \gamma_2 : [a, c] \rightarrow \mathbb{R}^n$ defined by

$$
\gamma(t) = \begin{cases} 
\gamma_1(t) & \text{if } a \leq t \leq b \\
\gamma_2(t) & \text{if } b \leq t \leq c
\end{cases}
$$

The concatenation $\gamma$ is a parametrized curve from $p$ to $r$.

We may extend the definition to $\gamma_1 + \cdots + \gamma_n$ if the curves line up properly.

782
VII.1.1.6. Note the conditions under which the concatenation is defined are rather severe: $\gamma_1$ and $\gamma_2$ must be parametrized on intervals in $\mathbb{R}$ which concatenate, and the final endpoint of $\gamma_1$ must be the initial endpoint of $\gamma_2$. Concatenation can be slightly generalized using reparametrizations (VII.4.1.16.). Also note that although additive notation is used, concatenation is not commutative: if $\gamma_1 + \gamma_2$ is defined, then $\gamma_2 + \gamma_1$ is not even defined unless one of the intervals is degenerate.

VII.1.1.7. If $\gamma = \gamma_1 + \gamma_2$ is defined, then the restriction of $\gamma$ to $[a, b]$ (resp. to $[b, c]$) is $\gamma_1$ (resp. $\gamma_2$). Conversely, if $\gamma : [a, b]$ is a parametrized curve and $c \in [a, b]$, then $\gamma = \gamma|_{[a,c]} + \gamma|_{[c,b]}$. More generally, if $P = \{a = t_0, t_1, \ldots, t_n = b\}$ is a partition of $[a, b]$, we have

$$\gamma = \gamma|_{[a,t_1]} + \gamma|_{[t_1,t_2]} + \cdots + \gamma|_{[t_{n-1},b]}.$$

VII.1.2. Reparametrization

VII.1.2.1. Definition. Let $\gamma : [a, b] \to \mathbb{R}^n$ and $\tilde{\gamma} : [c, d] \to \mathbb{R}^n$ be parametrized curves. Then $\tilde{\gamma}$ is a reparametrization of $\gamma$ if $\tilde{\gamma} = \gamma \circ \phi$, where $\phi$ is a nondecreasing continuous function from $[c, d]$ onto $[a, b]$ (hence $\phi(c) = a$ and $\phi(d) = b$). If $\phi$ is strictly increasing, $\tilde{\gamma}$ is a strict reparametrization of $\gamma$.

VII.1.2.2. A reparametrization of a parametrized curve has the same initial and final endpoints, and the same range. But two parametrized curves with the same range, even also with the same initial and final endpoints, need not be reparametrizations. A reparametrization of a parametrized curve describes “motion” along the range “curve” in the same sense, with the same retracing, but possibly over a different interval of time at a different and varying speed.

VII.1.2.3. Various special kinds of reparametrizations can be considered. The simplest is linear reparametrizations, where $\phi$ is a linear (affine) function from $[c, d]$ onto $[a, b]$, i.e.

$$\phi(t) = a + \frac{b - a}{d - c}(t - c).$$

One can also consider $C^r$ reparametrizations, where $\phi$ is $C^r$, and strict $C^r$ reparametrizations, where both $\phi$ and $\phi^{-1}$ are $C^r$ (these are also called smooth reparametrizations). There are also piecewise linear and piecewise smooth reparametrizations.

VII.1.2.4. Reparametrization is not an equivalence relation: it is reflexive and transitive, but not symmetric since the function $\phi$ need not be one-to-one. Strict reparametrization is symmetric, thus an equivalence relation on parametrized curves. Parametrization does generate an important equivalence relation, which can be described thus: two parametrized curves are equivalent if they are both reparametrizations of the same parametrized curve (VII.1.2.7.).

Nonstationary Curves

VII.1.2.5. Definition. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a parametrized curve, and $[c, d]$ a closed subinterval of $[a, b]$. Then $\gamma$ is stationary on $[c, d]$ if $\gamma(t) = \gamma(c)$ for all $t \in [c, d]$. The curve $\gamma$ is nonstationary if it is not stationary on any (nondegenerate) subinterval of $[a, b]$. 

783
VII.1.2.6. A nonstationary parametrized curve need not be one-to-one; in fact, it can be very noninjective. For example, the space-filling curve of XIV.12.1.14. is nonstationary (the one in XIV.12.1.13. is not nonstationary). For a simpler example, a nonstationary parametrized curve can wrap around a circle in \( \mathbb{R}^2 \) several times.

VII.1.2.7. The maximal intervals on which a parametrized curve \( \gamma \) is stationary form a collection of disjoint closed subintervals of \([a, b]\). Every parametrized curve is a reparametrization of a nonstationary parametrized curve, which is unique up to strict reparametrization (Exercise ()), obtained by collapsing each maximal stationary subinterval to a point (there may be infinitely many such intervals!) Two parametrized curves are equivalent in the equivalence relation generated by reparametrization (VII.1.2.4.) if and only if their nonstationary versions are strict reparametrizations of each other.

VII.1.3. Paths

VII.1.3.1. Definition. Let \( p, q \in \mathbb{R}^n \). A path from \( p \) to \( q \) is an equivalence class of parametrized curves from \( p \) to \( q \) (using the equivalence relation of VI.4.1.11. or VII.1.2.7.).

We will write \( \gamma_1 \approx \gamma_2 \) if \( \gamma_1 \) and \( \gamma_2 \) are equivalent, and write \( \langle \gamma \rangle \) for the equivalence class of \( \gamma \) (the usual notation \([\gamma]\) for an equivalence class is normally reserved for the homotopy class of \( \gamma \), which is a quite different thing ()).

VII.1.3.2. Paths can be concatenated: if \( \langle \gamma_1 \rangle \) is a path from \( p \) to \( q \) and \( \langle \gamma_2 \rangle \) is a path from \( q \) to \( r \), then \( \gamma_1 \) can be concatenated with a suitable reparametrization \( \tilde{\gamma}_2 \) of \( \gamma_2 \) (e.g. a linear reparametrization), and the equivalence class of \( \gamma_1 + \tilde{\gamma}_2 \) is well defined and depends only on the classes of \( \gamma_1 \) and \( \gamma_2 \) (Exercise ()). Thus the concatenation of the paths \( \langle \gamma_1 \rangle + \langle \gamma_2 \rangle \) is well defined as \( \langle \gamma_1 + \tilde{\gamma}_2 \rangle \).

Restriction of paths is a much dicier matter ().

VII.2. Smooth and Piecewise Smooth Curves in \( \mathbb{R}^n \)

VII.3. Rectifiable Curves and Arc Length

Throughout this section, \( \rho \) will denote the Euclidean distance function on \( \mathbb{R}^d \). (We use \( d \) instead of \( n \) so that \( n \) is available as a counting parameter for partitions.)

VII.3.1. Arc Length

VII.3.1.3. Definition. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a parametrized curve, and \( \mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\} \) a partition of \([a, b]\). The polygonal approximate length of \( \gamma \) from \( \mathcal{P} \) is

\[
\ell(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho(\gamma(t_{k-1}), \gamma(t_k)) .
\]

The arc length of \( \gamma \) is

\[
\ell(\gamma) = \sup_{\mathcal{P}} \ell(\gamma, \mathcal{P}) .
\]
VII.3.1.4. If \( Q \) is a partition of \([a, b]\) which refines \( P \), we have \( \ell(\gamma, P) \leq \ell(\gamma, Q) \) by the triangle inequality. Since any two partitions have a common refinement, the polygonal approximations approach the arc length as the partitions become arbitrarily fine. The term “polygonal approximation” should not be taken too literally: there is usually not an actual “polygon” in \( \mathbb{R}^d \) approximating the curve. The term is suggestive that \( \rho(p, q) \) should be the “straight-line” distance between \( p \) and \( q \).

The arc length of a curve can easily be infinite (e.g. VI.4.1.20.).

VII.3.1.5. Definition. The curve \( \gamma \) is rectifiable if \( \ell(\gamma) \) is finite.

VII.3.1.6. Examples. (i) A parametrized curve in \( \mathbb{R}^1 \) is a continuous function \( \gamma : [a, b] \to \mathbb{R} \). If \( P \) is a partition of \([a, b]\), then \( \ell(\gamma, P) \) is the total variation \( V(\gamma, P) \) of \( \gamma \) over \( P \) (XIV.16.1.3.), and \( \ell(\gamma) \) is the total variation \( V[a, b](\gamma) \) of \( \gamma \) over \([a, b]\) (XIV.16.1.6.). The curve \( \gamma \) is rectifiable if and only if \( \gamma \) has bounded variation. If \( \gamma \) is a smooth function, then \( \ell(\gamma) = \int_a^b |\gamma'(t)| \, dt \).

We now develop basic properties of arc length. First note that arc length is independent of parametrization:

VII.3.1.7. Proposition. Let \( \tilde{\gamma} : [c, d] \to \mathbb{R}^d \) be a reparametrization of \( \gamma : [a, b] \to \mathbb{R}^d \), i.e. \( \tilde{\gamma} = \gamma \circ \phi \), \( \phi : [c, d] \to [a, b] \). Then \( \ell(\tilde{\gamma}) = \ell(\gamma) \).

Proof: Let \( P = \{c = t_0, \ldots, t_n = d\} \) be a partition of \([c, d]\). Then \( \{a = \phi(t_0), \ldots, \phi(t_n) = b\} \) is not quite a partition of \([a, b]\) since there may be repetitions if \( \phi \) is not injective. But repeated points contribute nothing to the sum \( \ell(\tilde{\gamma}, P) \), and if \( Q \) is the partition of \([a, b]\) obtained by deleting repetitions, we have \( \ell(\gamma, Q) = \ell(\tilde{\gamma}, P) \). Every partition of \([a, b]\) arises in this manner.

VII.3.1.8. Thus arc length of paths is well defined.

Although the reversal \(-\gamma\) of a parametrized curve is not a reparametrization of \( \gamma \), we nonetheless have:

VII.3.1.9. Proposition. If \( \gamma : [a, b] \to \mathbb{R}^d \) is a parametrized curve in \( X \), then \( \ell(-\gamma) = \ell(\gamma) \).

Proof: If \( P \) is a partition of \([a, b]\), let \( \tilde{P} \) be the partition consisting of \( \{a + b - t : t \in P\} \). Then \( \tilde{P} \) is also a partition of \([a, b]\), and every partition of \([a, b]\) arises in this manner. We have \( \ell(-\gamma, \tilde{P}) = \ell(\gamma, P) \).

VII.3.1.10. If \( \gamma : [a, b] \to X \) is a parametrized curve from \( p \) to \( q \) and \( P = \{a, b\} \) is the trivial partition, we have
\[
\rho(p, q) = \ell(\gamma, P) \leq \ell(\gamma) .
\]
More generally, we have:
VII.3.1.11. PROPOSITION. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve in $\mathbb{R}^d$, and $a \leq c \leq d \leq b$, then

$$\rho(\gamma(c), \gamma(d)) \leq \ell(\gamma|_{[c,d]}) \leq \ell(\gamma).$$

**Proof:** The first inequality has already been noted. For the second, any partition $\mathcal{P}$ of $[c, d]$ can be expanded to a partition $\mathcal{Q}$ of $[a, b]$ by adding $a$ and $b$ if necessary. We have

$$\ell(\gamma|_{[c,d]}, \mathcal{P}) \leq \ell(\gamma, \mathcal{Q}) \leq \ell(\gamma)$$

and taking the supremum over all $\mathcal{P}$ gives the inequality. $\blacksquare$

If $\gamma$ is a constant path, obviously $\ell(\gamma) = 0$. Conversely:

VII.3.1.12. COROLLARY. Let $\gamma$ be a parametrized curve in $\mathbb{R}^d$. If $\ell(\gamma) = 0$, then $\gamma$ is constant.

**Proof:** If $\gamma$ is not constant, there are $c, d \in [a, b]$, $c < d$, with $\gamma(c) \neq \gamma(d)$. Then $0 < \rho(\gamma(c), \gamma(d)) \leq \ell(\gamma)$. $\blacksquare$

Here is a more precise version of VII.3.1.11.: 

VII.3.1.13. PROPOSITION. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve, and $c \in [a, b]$. Then

$$\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}).$$

**Proof:** If $\mathcal{P}$ is a partition of $[a, c]$ and $\mathcal{Q}$ a partition of $[c, b]$, then $\mathcal{P} \cup \mathcal{Q}$ is a partition of $[a, b]$, and

$$\ell(\gamma, \mathcal{P} \cup \mathcal{Q}) = \ell(\gamma|_{[a,c]}, \mathcal{P}) + \ell(\gamma|_{[c,b]}, \mathcal{Q}).$$

Every partition of $[a, b]$ containing $c$ (i.e. every sufficiently fine partition) arises in this manner. $\blacksquare$

Iterating, we get:

VII.3.1.14. COROLLARY. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve, and $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ a partition of $[a, b]$. Then

$$\ell(\gamma) = \sum_{k=1}^{n} \ell(\gamma|_{[x_{k-1},x_k]}).$$

Rephrasing VII.3.1.13., we get:
VII.3.1.15. **Proposition.** Let \( \gamma_1 : [a, b] \to \mathbb{R}^d \) and \( \gamma_2 : [b, c] \to \mathbb{R}^d \) with \( \gamma_1(b) = \gamma_2(b) \), so that the concatenation \( \gamma_1 + \gamma_2 \) is defined. Then

\[
\ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2).
\]

This can also be iterated to apply to arbitrary finite concatenations.

A crucial property of arc length is lower semicontinuity:

VII.3.1.16. **Proposition.** For each \( n \), let \( \gamma_n : [a, b] \to \mathbb{R}^d \) be a parametrized curve from \( p_n \) to \( q_n \). Suppose \( \gamma_n \) converges pointwise to a parametrized curve \( \gamma : [a, b] \to X \) from \( p \) to \( q \), i.e. \( \gamma_n(t) \to \gamma(t) \) for each \( t \in [a, b] \) (so \( p_n \to p \), \( q_n \to q \)). Then

\[
\ell(\gamma) \leq \liminf_{n \to \infty} \ell(\gamma_n).
\]

We can have strict inequality, even if \( \gamma_n \to \gamma \) uniformly on \( [a, b] \) (I.5.1.11.).

**Proof:** First note that if \( \mathcal{P} = \{a = t_0, \ldots, t_m = b\} \) is any partition of \( [a, b] \), we have

\[
\ell(\gamma_n, \mathcal{P}) = \sum_{k=1}^{m} \rho(\gamma_n(t_{k-1}), \gamma_n(t_k)) \to \sum_{k=1}^{m} \rho(\gamma(t_{k-1}), \gamma(t_k)) = \ell(\gamma, \mathcal{P})
\]

by continuity of \( \rho \).

Suppose \( \ell(\gamma) = +\infty \). Fix \( M > 0 \). Then there is a partition \( \mathcal{P} \) of \( [a, b] \) with \( \ell(\gamma, \mathcal{P}) > M \). Then

\[
M < \ell(\gamma, \mathcal{P}) = \lim_{n \to \infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n \to \infty} \ell(\gamma_n).
\]

This holds for all \( M > 0 \), so \( \liminf_{n \to \infty} \ell(\gamma_n) = +\infty \) and the inequality holds.

Now suppose \( \gamma \) is rectifiable. Let \( \epsilon > 0 \). Then there is a partition \( \mathcal{P} \) of \( [a, b] \) with \( \ell(\gamma, \mathcal{P}) > \ell(\gamma) - \epsilon \). Then

\[
\ell(\gamma) - \epsilon \leq \ell(\gamma, \mathcal{P}) = \lim_{n \to \infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n \to \infty} \ell(\gamma_n).
\]

This holds for all \( \epsilon > 0 \), so \( \ell(\gamma) \leq \liminf_{n \to \infty} \ell(\gamma_n) \). 

VII.3.1.17. Note for future reference that the definitions of \( \ell(\gamma, \mathcal{P}) \) and \( \ell(\gamma) \) make perfect sense for a general function \( \gamma : [a, b] \to \mathbb{R}^d \), not necessarily continuous, so “arc length” is defined even for discontinuous functions (although it is typically infinite for very discontinuous functions). None of the proofs of VII.3.1.7.–VII.3.1.16. use that \( \gamma \) is continuous, so these properties hold also for discontinuous \( \gamma \). (But VII.3.2.2. requires continuity of \( \gamma \).)
VII.3.2. Rectifiable Curves

Continuity of Arc Length

VII.3.2.1. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a rectifiable parametrized curve. Define a function \( \phi : [a, b] \to [0, \ell(\gamma)] \) by setting

\[
\phi(t) = \ell(\gamma|_{[a,t]}) .
\]

Then \( \phi \) is nondecreasing, \( \phi(a) = 0 \), \( \phi(b) = \ell(\gamma) \). If \( \gamma \) is nonstationary, \( \phi \) is strictly increasing.

VII.3.2.2. Proposition. The function \( \phi \) is continuous.

Proof: The following lemma will provide left continuity on \([a, b]\). The simplest way to get right continuity is to apply the lemma also to \( \gamma(t_0) \) and use that \( \ell(\gamma|_{[a,t_0]}) = \ell(\gamma|_{[a,t_0]}) \).

VII.3.2.3. Lemma. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a parametrized curve, not necessarily rectifiable. If \( (t_n) \) is an increasing sequence in \([a, b]\) converging to \( t_0 \), then

\[
\ell(\gamma|_{[a,t_0]}) = \lim_{n \to \infty} \ell(\gamma|_{[a,t_n]}) = \sup_n \ell(\gamma|_{[a,t_n]}) .
\]

Proof: Define a sequence \( (\gamma_n) \) of parametrized curves on \([a, t_0]\) by setting \( \gamma_n(t) = \gamma(t) \) if \( a \leq t \leq t_n \) and \( \gamma_n(t) = \gamma(t_n) \) if \( t_n \leq t \leq t_0 \) (i.e. \( \gamma_n \) agrees with \( \gamma \) on \([a, t_n]\) and is stationary on \([t_n, t_0]\)). We have

\[
\ell(\gamma_n) = \ell(\gamma|_{[a,t_n]}) \leq \ell(\gamma|_{[a,t_0]})
\]

for any \( n \) by monotonicity (VII.3.1.11.). On the other hand, since \( \gamma \) is continuous at \( t_0 \), \( \gamma_n \to \gamma \) pointwise on \([a, t_0]\). So applying VII.3.1.16. we get

\[
\sup_n \ell(\gamma|_{[a,t_n]}) \leq \ell(\gamma|_{[a,t_0]}) \leq \lim_{n \to \infty} \inf \ell(\gamma_n) = \lim_{n \to \infty} \inf \ell(\gamma|_{[a,t_n]})
\]

and the result follows.

VII.3.2.4. Corollary. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a rectifiable curve. Then for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that, whenever \( a \leq c < d \leq b \) and \( d - c < \delta \), then \( \ell(\gamma|_{[c,d]}) < \epsilon \).

Proof: The function \( \phi \) is uniformly continuous on \([a, b]\).

Estimation of Arc Length

The next result is strongly reminiscent of the equivalence of Definitions 1A and 1B of the Riemann integral (XIV.2.3.1.), and proved in a similar way.
THEOREM. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a rectifiable curve. Then for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that \( \ell(\gamma, P) > \ell(\gamma) - \epsilon \) for any partition \( P \) of \([a, b] \) with \( \|P\| < \delta \).

PROOF: Let \( Q = \{a = t_0, \ldots, t_n = b\} \) be a partition of \([a, b] \) with \( \ell(\gamma, Q) > \ell(\gamma) - \frac{\epsilon}{2} \). Since \( \gamma \) is uniformly continuous (), there is a \( \delta > 0 \) such that if \( a \leq c < d \leq b \) and \( d - c < \delta \), \( \rho(\gamma(c), \gamma(d)) < \frac{\epsilon}{3n} \). We may assume \( 3\delta < \|Q\| \).

Let \( P = \{a = s_0, \ldots, s_m = b\} \) be a partition of \([a, b] \) with \( \|P\| < \delta \). For \( 1 \leq k \leq n - 1 \) let \( j_k \) be the largest \( j \) such that \( s_j \leq t_k \). We then have

\[
\ell(\gamma, P) = \sum_{j=1}^{m} \rho(\gamma(s_{j-1}), \gamma(s_j))
\]

\[
= \sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}), \gamma(s_j)) + \sum_{k=1}^{n-2} \sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}), \gamma(s_j)) + \sum_{j=n-1+2}^{m} \rho(\gamma(s_{j-1}), \gamma(s_j)) + \sum_{k=1}^{m-1} \rho(\gamma(s_{j_k}), \gamma(s_{j_k+1})).
\]

By the triangle inequality, we have

\[
\sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}), \gamma(s_j)) \geq \rho(\gamma(a), \gamma(t_1)) - \rho(\gamma(s_{j_1}), \gamma(t_1)) > \rho(\gamma(a), \gamma(t_1)) - \frac{\epsilon}{4n}.
\]

Similarly, for \( 1 \leq k \leq n - 2 \), we have

\[
\sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}), \gamma(s_j)) \geq \rho(\gamma(t_k), \gamma(t_{k+1})) - \rho(\gamma(s_{j_{k+1}}), \gamma(t_{k+1})) > \rho(\gamma(t_k), \gamma(t_{k+1})) - \frac{\epsilon}{2n}
\]

\[
\sum_{j=n-1+2}^{m} \rho(\gamma(s_{j-1}), \gamma(s_j)) > \rho(\gamma(t_{n-1}), \gamma(b)) - \frac{\epsilon}{4n}.
\]

Thus we have

\[
\ell(\gamma, P) = \sum_{j=1}^{m} \rho(\gamma(s_{j-1}), \gamma(s_j)) \geq \sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}), \gamma(s_j)) + \sum_{k=1}^{n-2} \sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}), \gamma(s_j)) + \sum_{j=n-1+2}^{m} \rho(\gamma(s_{j-1}), \gamma(s_j)) \geq \sum_{k=1}^{n} \rho(\gamma(t_{k-1}), \gamma(t_k)) - (n-2) \frac{\epsilon}{2n} - 2 \frac{\epsilon}{4n} = \ell(\gamma, Q) - \frac{\epsilon}{2} > \ell(\gamma) - \epsilon.
\]

\( \square \)
Parametrization by Arc Length

VII.3.2.6. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a nonstationary rectifiable parametrized curve, and let $\phi$ be the above function. Then $\phi$ is strictly increasing and continuous, so has a continuous strictly increasing inverse function $\phi^{-1} : [0, \ell(\gamma)] \to [a, b]$. Set $\tilde{\gamma} = \gamma \circ \phi^{-1} : [0, \ell(\gamma)] \to X$. Then $\tilde{\gamma}$ is a strict reparametrization of $\gamma$ with the property that $\ell(\tilde{\gamma}|_{[0, t]}) = t$ for $0 \leq t \leq \ell(\gamma)$, i.e. $\tilde{\gamma}$ “traverses the path $\langle \gamma \rangle$ at uniform unit speed.” The parametrized curve $\tilde{\gamma}$ is called the reparametrization of $\gamma$ by arc length. If $\gamma$ is replaced by any strict reparametrization, the $\tilde{\gamma}$ is unchanged; thus it depends only on $\langle \gamma \rangle$, i.e. any rectifiable path has a unique canonical parametrization by arc length.

Differentiability of Rectifiable Curves

A rectifiable curve has “bounded variation” and hence should be differentiable almost everywhere, and in fact is:

VII.3.2.7. Theorem. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a rectifiable curve. Then $\gamma$ is differentiable a.e.

Proof: Let $\gamma_1, \ldots, \gamma_n$ be the coordinate functions of $\gamma$. Then each $\gamma_k$ has bounded variation $V_{[a, b]}(\gamma_k) \leq \ell(\gamma)$ (VII.3.4.1.(a)). Thus $\gamma_k$ is differentiable on $[a, b]$ except on a set $A_k$ of measure 0 (XIV.16.2.4.). So $\gamma$ is differentiable on $[a, b]$ except on $A_1 \cup \cdots \cup A_n$, a set of measure 0.

A rectifiable curve can fail to be differentiable at a dense set of points, however. If $\gamma$ is rectifiable, one might expect that its arc length can be computed by integration of the length of the derivative. However, this fails in general since $\gamma$ is not necessarily absolutely continuous. For an extreme example, if $\gamma$ is parametrized on $[0, 1]$ and $g$ is the Cantor function, then $\gamma \circ g$ is a reparametrization of $\gamma$, and $(\gamma \circ g)' = 0$ a.e. But if $\gamma$ is parametrized by an absolutely continuous function, it is true. This result is surprisingly hard to prove and find in books perhaps because it falls through the crack between real analysis and differential geometry. Our proof is adapted from [Pel77], where a more general result is proved.

VII.3.2.9. Theorem. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a (necessarily rectifiable) curve, with absolutely continuous coordinate functions. Then $\gamma$ is differentiable a.e., and

$$\ell(\gamma) = \int_a^b ||\gamma'(t)|| \, d\lambda(t)$$

(where the integral is a Lebesgue integral). The length function $\phi(x) = \ell(\gamma|_{[a,x]})$ is absolutely continuous, hence differentiable a.e., and $\phi'(t) = ||\gamma'(t)||$ a.e.

The proof will mostly consist of three lemmas. The first is of independent interest, that a rectifiable curve can be “unwound” to a curve with nondecreasing coordinate functions without changing the arc length or length of the derivative (a.e.):
VII.3.2.10. **Lemma.** Let $\gamma : [a, b] \to \mathbb{R}^n$ be a rectifiable curve with coordinate functions $\gamma_1, \ldots, \gamma_n$ of bounded variation. For $1 \leq k \leq n$ let $v_k = T_{\gamma_k}$ be the total variation function (XIV.16.2.1.) of $\gamma_k$ on $[a, b]$, i.e. $v_k(t) = V_{[a,t]}(\gamma_1)$ (the $v_k$ are continuous by XIV.16.2.6.). Let $\tilde{\gamma} : [a, b] \to \mathbb{R}^n$ have coordinate functions $v_1, \ldots, v_n$. Then $\ell(\tilde{\gamma}) = \ell(\gamma)$.

**Proof:** We have, for any $k$ and any $s, t \in [a, b]$,

$$|\gamma_k(t) - \gamma_k(s)| \leq |v_k(t) - v_k(s)|$$

and it follows that for any partition $P$ of $[a, b]$, $\ell(\gamma, P) \leq \ell(\tilde{\gamma}, P)$. Taking the supremum over all $P$, we obtain $\ell(\gamma) \leq \ell(\tilde{\gamma})$.

For the opposite inequality, suppose $a \leq c < d \leq b$ and $P = \{c = t_0, \ldots, t_m = d\}$ is a partition of $[c, d]$. Using the triangle inequality in $\mathbb{R}^n$ we obtain

$$\ell(\gamma|_{[c,d]}, P) = \sum_{j=1}^{m} \left( \sum_{k=1}^{n} (\gamma_k(t_j) - \gamma_k(t_{j-1})) \right)^2 \geq \sum_{k=1}^{n} \left( \sum_{j=1}^{m} |\gamma_k(t_j) - \gamma_k(t_{j-1})| \right)^2 = \sum_{k=1}^{n} [V(\gamma_k, P)]^2.$$ 

Taking the supremum over all $P$ we obtain

$$\ell(\gamma|_{[c,d]}) \geq \sqrt{\sum_{k=1}^{n} [V_{[c,d]}(\gamma_k)]^2} = \sqrt{\sum_{k=1}^{n} (v_k(d) - v_k(c))^2}.$$ 

If $Q = \{a = s_0, \ldots, s_r = b\}$ is any partition of $[a, b]$, applying this inequality over each subinterval $[s_{j-1}, s_j]$, we obtain

$$\ell(\gamma) = \sum_{j=1}^{r} \ell(\gamma|_{[s_{j-1}, s_j]}) \geq \sum_{j=1}^{r} \sqrt{\sum_{k=1}^{n} (v_k(s_j) - v_k(s_{j-1}))^2} = \ell(\tilde{\gamma}, Q)$$

and it follows by taking the supremum over all $Q$ that $\ell(\gamma) \geq \ell(\tilde{\gamma})$.

VII.3.2.11. If $\gamma_1, \ldots, \gamma_n$ are absolutely continuous, so are $v_1, \ldots, v_n$, and $v'_k = |\gamma'_k|$ a.e. for each $k$, so $\|\gamma'\| = \|\gamma'_k\|$ a.e. Thus to prove VII.3.2.9. we may work with $\tilde{\gamma}$, i.e. we may assume that $\gamma'_k \geq 0$ a.e. for each $k$.

The other two lemmas are very general:
VII.3.2.12. Lemma. Let \((X, \mathcal{A}, \mu)\) be a measure space and \(f_1, \ldots, f_n\) nonnegative measurable functions. Then, for any \(A \in \mathcal{A},\)
\[
\sqrt{\sum_{k=1}^{n} \left[ \int_A f_k \, d\mu \right]^2} \leq \int_A \sqrt{\sum_{k=1}^{n} f_k^2} \, d\mu .
\]

Proof: This is a good example of bootstrapping. First suppose the \(f_k\) are simple functions, and are all constant on sets \(E_1, \ldots, E_m,\) with \(f_k\) taking value \(\alpha_{k,j}\) on \(E_j.\) Then by the triangle inequality on \(\mathbb{R}^n,\)
\[
\sqrt{\sum_{k=1}^{n} \left[ \int_A f_k \, d\mu \right]^2} = \sqrt{\sum_{k=1}^{n} \left[ \sum_{j=1}^{m} \int_{A \cap E_j} f_k \, d\mu \right]^2} = \sqrt{\sum_{k=1}^{n} \left[ \sum_{j=1}^{m} \alpha_{k,j} \mu(A \cap E_j) \right]^2}
\]
\[
\leq \sum_{j=1}^{m} \sqrt{\sum_{k=1}^{n} \alpha_{k,j} (A \cap E_j)^2} = \sum_{j=1}^{m} \left[ \sum_{k=1}^{n} \alpha_{k,j}^2 \mu(A \cap E_j) \right] = \sum_{j=1}^{m} \left[ \int_{A \cap E_j} \sqrt{\sum_{k=1}^{n} f_k^2} \, d\mu \right] = \int_A \sqrt{\sum_{k=1}^{n} f_k^2} \, d\mu .
\]
For general \(f_k,\) take increasing sequences of nonnegative simple functions and apply the Monotone Convergence Theorem. \(\boxdot\)

VII.3.2.13. Lemma. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f, f_1, \ldots, f_n\) nonnegative measurable functions on \(X.\) If
\[
\left[ \int_A f \, d\mu \right]^2 \geq \sum_{k=1}^{n} \left[ \int_A f_k \, d\mu \right]^2
\]
for all \(A \in \mathcal{A},\) then \(f \geq \sqrt{\sum_{k=1}^{n} f_k^2}\) a.e.

Proof: Suppose \(f_1, \ldots, f_n\) are simple and all constant on \(E_1, \ldots, E_m,\) and \(f_k\) takes the constant value \(\alpha_{k,j}\) on \(E_j.\) If \(A \subseteq E_j,\) then
\[
\int_A f \, d\mu \geq \sqrt{\sum_{k=1}^{n} \left[ \int_A f_k \, d\mu \right]^2} = \sqrt{\sum_{k=1}^{n} \left[ \sum_{j=1}^{m} \alpha_{k,j} \mu(A \cap E_j) \right]^2} = \mu(A) \sqrt{\sum_{k=1}^{n} \alpha_{k,j}^2}.
\]
This is true for all \(A \subseteq E_j,\) which implies that
\[
f \geq \sqrt{\sum_{k=1}^{n} \alpha_{k,j}^2} = \sqrt{\sum_{k=1}^{n} f_k^2}
\]
a.e. on \(E_j.\) Since \(\cup_{j=1}^{m} E_j = X,\) we have \(f \geq \sqrt{\sum_{k=1}^{n} f_k^2}\) a.e. on \(X.\)

For general \(f_k,\) take increasing sequences of nonnegative simple functions converging to the \(f_k.\) \(\boxdot\)
VII.3.2.14. We now give the proof of Theorem VII.3.2.9.

Proof: We may assume $\gamma_k'$ is nonnegative for all $k$ (VII.3.2.11.). If $\mathcal{P} = \{a = t_0, \ldots, t_m = b\}$ is a partition of $[a, b]$, we have

$$
\ell(\gamma, \mathcal{P}) = \sum_{j=1}^{m} \|\gamma(t_j) - \gamma(t_{j-1})\| = \sum_{j=1}^{m} \sqrt{\sum_{k=1}^{n} \left[ \int_{t_{j-1}}^{t_j} \gamma_k' \, d\lambda \right]^2}
$$

$$
\leq \sum_{j=1}^{m} \left[ \int_{t_{j-1}}^{t_j} \sqrt{\sum_{k=1}^{n} (\gamma_k')^2} \, d\lambda \right] = \sum_{j=1}^{m} \left[ \int_{t_{j-1}}^{t_j} \|\gamma'\| \, d\lambda \right] = \int_{a}^{b} \|\gamma'\| \, d\lambda
$$

by Lemma VII.3.2.12. Taking the supremum over all $\mathcal{P}$, we obtain

$$
\ell(\gamma) \leq \int_{a}^{b} \|\gamma'\| \, d\lambda.
$$

For the reverse inequality, note that, if $\phi(t) = \ell(\gamma|_{[a, t]})$ for $a \leq t \leq b$, then $\phi$ is absolutely continuous (VII.3.4.1.(b)). If $a \leq c < d \leq b$, we have

$$
\left[ \int_{c}^{d} \phi' \, d\lambda \right]^2 = [\phi(d) - \phi(c)]^2 = [\ell(\gamma|_{[c, d]})]^2
$$

$$
\geq \|\gamma(d) - \gamma(c)\|^2 = \sum_{k=1}^{n} \left[ \int_{c}^{d} \gamma_k' \, d\lambda \right]^2
$$

and thus $(\cdot)$, for every measurable $A$, we have

$$
\left[ \int_{A} \phi' \, d\lambda \right]^2 \geq \sum_{k=1}^{n} \left[ \int_{A} \gamma_k' \, d\lambda \right]^2.
$$

So by Lemma VII.3.2.13, we have

$$
\phi' \geq \sqrt{\sum_{k=1}^{n} (\gamma_k')^2} = \|\gamma'\| \text{ a.e.}
$$

and hence

$$
\ell(\gamma) = \int_{a}^{b} \phi' \, d\lambda \geq \int_{a}^{b} \|\gamma'\| \, d\lambda.
$$

We have, for each $x \in [a, b]$,

$$
\phi(x) = \int_{a}^{x} \phi' \, d\lambda = \int_{a}^{x} \|\gamma'\| \, d\lambda
$$

so if $f = \phi' - \|\gamma'\|$, then $f$ is an integrable function on $[a, b]$ and $\int_{a}^{x} f \, d\lambda = 0$ for all $x \in [a, b]$, so $f = 0$ a.e. (XIV.4.8.12.).

The following special case of VII.3.2.9. is especially noteworthy. The result is probably not surprising, but it is definitely nontrivial (cf. VII.3.2.16.). See $(\cdot)$ for a direct proof valid in general metric spaces.
VII.3.2.15. **Corollary.** Let \( \gamma : [0, b] \to \mathbb{R}^n \) be a rectifiable curve, parametrized by arc length. Then \( \gamma \) is absolutely continuous and hence is differentiable a.e., and \( \|\gamma'(t)\| = 1 \) for almost all \( t \).

**Proof:** Since \( \gamma \) is parametrized by arc length, it is a Lipschitz function with constant 1, so the coordinate functions are absolutely continuous and VII.3.2.9. applies; and \( \phi' \) is the constant function 1.

VII.3.2.16. **Example.** One might naïvely think that if \( \gamma \) is a rectifiable curve, parametrized by arc length, then \( \|\gamma'(t)\| = 1 \) whenever it is defined. But this is false. Here is an example, adapted from ().

Let \( c \) be the curve in \( \mathbb{R}^2 \) consisting of semicircular arcs between \( \frac{1}{n} \) and \( \frac{1}{n+1} \) for each \( n \in \mathbb{N} \), alternating above and below the \( x \)-axis to make the curve differentiable on \((0, 1)\), reflected on the interval \([1, 0)\), along with the origin. See Figure (). Then \( \gamma \) is the graph of a function on \([\frac{1}{n}, \frac{1}{n+1}]\). Parametrize \( c \) by arc length, starting at \((\frac{1}{n}, 0)\). Then \( \ell(t) = \frac{1}{n+1} \leq t \leq \frac{1}{n} \) for almost all \( t \), and thus

\[
\frac{2}{\pi} \cdot \frac{n}{n+1} = \frac{1}{\frac{n+1}{\pi}} \leq \frac{\gamma_1(t) - \gamma_1(\frac{\pi}{2})}{t - \frac{\pi}{2}} \leq \frac{1}{\frac{n}{\pi}} \leq \frac{2}{\pi} \cdot \frac{n+1}{n}
\]

and so by the squeeze theorem

\[
\lim_{t \to \frac{\pi}{2}} \frac{\gamma_1(t) - \gamma_1(\frac{\pi}{2})}{t - \frac{\pi}{2}} = \frac{2}{\pi}
\]

exists and equals \( \frac{2}{\pi} \). The limit from the left similarly equals \( \frac{2}{\pi} \). So \( \gamma_1(\frac{\pi}{2}) \) exists and equals \( \frac{2}{\pi} \).

Thus \( \gamma \) is differentiable at \( \frac{\pi}{2} \) and \( \gamma'(\frac{\pi}{2}) = (\frac{2}{\pi}, 0) \), so

\[
\left\|\gamma'(\frac{\pi}{2})\right\| = \frac{2}{\pi} < 1
\]

Replacing the semicircles by curves \( c_n \) of arc length \( \frac{2}{\sqrt{n}} \) lying in the square

\[
\left[\frac{1}{n+1}, 1\right] \times \left[\frac{1}{n^2}, \frac{1}{n^2}\right]
\]

(and similarly on the left) we can get a similar example where \( \gamma' = 0 \) at the origin. Much more pathological examples, where \( \|\gamma'(t)\| < 1 \) for many \( t \), can be constructed similarly.

VII.3.3. **Length of Arcs**

If \( C \) is a “curve” in \( \mathbb{R}^n \), we would like to be able to say what the intrinsic “arc length” of \( C \) is. This is a tricky problem. The first question is what kind of subsets should be called “curves.” It is too broad to say that the range of any parametrized curve is a “curve” (). One reasonable possibility is:
VII.3.3.17. Definition. An arc in \( \mathbb{R}^n \) is a subset of \( \mathbb{R}^n \) homeomorphic to \([0, 1]\).

VII.3.3.18. A subset \( C \) of \( \mathbb{R}^n \) is an arc if and only if it is the range of an injective parametrized curve: a homeomorphism \( \gamma \) from \([0, 1]\) to \( C \) is such a parametrized curve, and conversely an injective parametrized curve \( \gamma : [a, b] \to \mathbb{R}^n \) with \( C = \gamma([a, b]) \) is a homeomorphism by \((\). An arc thus has a unique equivalence class of oriented injective parametrizations, thus two in all; the arc length of all these parametrized curves is the same and can be taken to be the arc length of \( C \).

VII.3.3.19. Slightly more generally, if \( C \) can be written as a finite union of arcs with only endpoints in common, the arc length of \( C \) can be taken to be the sum of the arc lengths of these arcs. For more complicated subsets, there are grave difficulties in defining arc length this way.

VII.3.3.20. There is a way to unambiguously define “arc length” for general (Borel) subsets of \( \mathbb{R}^n \): one-dimensional Hausdorff measure \((\). This “arc length” has the nice properties of a measure, generalizing VII.3.1.14.. Subsets whose one-dimensional Hausdorff measure is positive and finite can be regarded as “one-dimensional” (cf. \((\); such a set need not be topologically one-dimensional, and not every “one-dimensional” subset will have these properties).

VII.4. Exercises

VII.3.4.1. (a) Let \( \gamma : [a, b] \to \mathbb{R}^n \) be a parametrized curve, with coordinate functions \( \gamma_1, \ldots, \gamma_n \). Show that the total variation \((\text{XIV.16.1.6.})\) satisfies \( V_{[a, b]}(\gamma_k) \leq \ell(\gamma) \leq \sum_{j=1}^n V_{[a, b]}(\gamma_j) \) for any \( k \), and thus \( \gamma \) is rectifiable if and only if all coordinate functions of \( \gamma \) have bounded variation. [For each \( k \) and each partition \( P \) of \([a, b]\), we have \( V(\gamma_k, P) \leq \ell(\gamma, P) \leq \ell(\gamma) \).]

(b) Let \( \gamma : [a, b] \to \mathbb{R}^n \) be a rectifiable parametrized curve, and \( \phi \) the function defined in VII.3.2.2.. Show that \( \phi \) is absolutely continuous if and only if each coordinate function of \( \gamma \) is absolutely continuous. [If each \( \gamma_k \) is absolutely continuous, and \( \epsilon > 0 \), for each \( k \) choose \( \delta_k \) which works in the definition of absolute continuity for \( \gamma_k \) for \( \xi \). Show that \( \delta = \min_k (\delta_k) \) works in the definition for \( \phi \) and \( \epsilon \).]

VII.3.4.2. Prove the following result complementary to VII.3.2.5.:

Theorem. Let \( \gamma : [a, b] \to \mathbb{R}^d \) be a parametrized curve. If \( \ell(\gamma) = \infty \), then for any \( M > 0 \) there is a \( \delta > 0 \) such that \( \ell(\gamma, P) > M \) for any partition \( P \) of \([a, b]\) with \( \|P\| < \delta \).

VII.3.4.3. Carry out the details of the last construction in VII.3.2.16..
Chapter VIII

Multivariable and Vector Calculus

In this chapter, we discuss calculus for functions from one Euclidean space $\mathbb{R}^n$ (or a subset) to another Euclidean space $\mathbb{R}^m$. The terms multivariable calculus and vector calculus used are roughly synonymous, but “vector calculus” is slightly more general: “multivariable calculus” refers to the case where $n > 1$ (and $m \geq 1$), while “vector calculus” also includes the case where $n = 1$ and $m > 1$, which is exactly the case of curves in $\mathbb{R}^m$ discussed in Chapter (). (In developing the theories, we work with general $n$ and $m$, not even excluding the case $n = m = 1$.) There is also a slight difference in point of view: multivariable calculus is generally thought of as consideration of real-valued functions, or systems of real-valued functions, of several real variables, while vector calculus is thought of as being the study of functions whose input and output are vectors in Euclidean spaces, not necessarily of the same dimension. (The two points of view are complementary and both are needed to get the full picture of the subject.)

Consideration of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ has distinct and complicating features over the case $n = m = 1$ arising from having either $n > 1$ or $m > 1$ (or both). The complications caused by considering the case $m > 1$ are quite modest, however. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function, write

$$f(x) = (f_1(x), \ldots, f_m(x))$$

where $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are the coordinate functions of $f$ (technically, $f_k = \pi_k^{(m)} \circ f$, where $\pi_k^{(m)} : \mathbb{R}^m \to \mathbb{R}$ is projection onto the $k$'th coordinate, i.e. $\pi_k^{(m)}(x_1, \ldots, x_m) = x_k$), then analysis of $f$ can be largely reduced to analysis of the $f_k$. There are situations where it is undesirable to reduce to the coordinate functions, though, such as in the study of curves, and in results like the Chain Rule and the Implicit and Inverse Function Theorems, where the interplay of coordinates is crucial. Roughly speaking, the multivariable calculus approach is to reduce to coordinate functions and the vector calculus approach is to avoid doing this.

The complications arising from having $n > 1$ are essential, however, and mostly cannot be handled by considering coordinates separately. Thus multivariable calculus (and vector calculus) has a quite different nature than one-variable calculus.

VIII.1. Limits and Continuity for Vector Functions

Defining limits and continuity for functions of more than one variable can be very easy or very tricky, depending whether one takes the proper point of view.
VIII.1.1. Graphs of Functions of Several Variables

VIII.1.2. Limits of Sequences in $R^n$

VIII.1.2.1. The first observation is that there is much more “room” in $R^n$, $n \geq 2$, for sequences to behave in a complicated way than in $R = R^1$. A sequence in $R$ can approach a limit only from the left or right, or more generally jump back and forth from left or right, i.e. can be decomposed into at most two subsequences which approach from the left and right respectively. But even in $R^2$, a sequence can approach a limit in a very complicated way: it can approach from any direction or along any curve, can spiral around the limit wildly, etc., and cannot generally be decomposed into subsequences which approach in a controlled way (things can be even wilder in $R^3$). Trying to make a definition which covers all the possibilities may seem like a daunting task.

VIII.1.2.2. However, there is a very simple solution. The definition of limit of a sequence in $R$ can be used almost verbatim in $R^n$ and gives exactly the right thing. The definition only needs slight redecoration to work perfectly:

VIII.1.2.3. Definition. Let $(x_k)$ be a sequence in $R^n$, and $x \in R^n$. Then $x_k \to x$, or $\lim_{k \to \infty} x_k = x$, if for every $\epsilon > 0$ there is a $K$ such that

$$
\|x_k - x\| < \epsilon
$$

for all $k \geq K$.

This is a special case of redecorating the definition in $R$ to one valid in a general metric space $(\cdot)$.

The Coordinatewise Criterion

There is one basic fact about limits in $R^n$ which will appear in various forms, which allows a reduction to the case of limits in $R$ in some cases, including sequences:

VIII.1.2.4. Proposition. [Coordinatewise Criterion for Sequences in $R^n$] Let $(x_k)$ be a sequence in $R^n$, and $x \in R^n$. Write

$$
x_k = (x_k^{(1)}, \ldots, x_k^{(n)}), \quad x = (x^{(1)}, \ldots, x^{(n)})
$$

in coordinates. Then $x_k \to x$ in $R^n$ if and only if $x_k^{(j)} \to x^{(j)}$ in $R$ for each $j$, $1 \leq j \leq n$.

Proof: Note that for each $m$ we have

$$
|x_k^{(m)} - x^{(m)}| \leq \|x_k - x\| \leq \sum_{j=1}^{n} |x_k^{(j)} - x^{(j)}|.
$$

Suppose $x_k \to x$, and fix $j$. Let $\epsilon > 0$. There is a $K$ such that

$$
|x_k^{(j)} - x^{(j)}| \leq \|x_k - x\| < \epsilon
$$
for all $k \geq K$; thus $x_k^{(j)} \to x^{(j)}$. Conversely, suppose $x_k^{(j)} \to x^{(j)}$ for all $j$. Let $\epsilon > 0$. For each $j$ choose $K_j$ such that

$$|x_k^{(j)} - x^{(j)}| < \frac{\epsilon}{n}$$

for all $k \geq K_j$. Set $K = \max(K_1, \ldots, K_n)$. Then, if $k \geq K$,

$$|x_k - x| \leq \sum_{j=1}^{n} |x_k^{(j)} - x^{(j)}| < n \cdot \frac{\epsilon}{n} = \epsilon.$$

Since $\epsilon > 0$ is arbitrary, $x_k \to x$. 

VIII.1.3. Multivariable Limits

We now look at limits for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. We first consider the case $m = 1$.

Limits for Real-valued Functions of Several Variables

VIII.1.3.1. Suppose $f$ is a real-valued function defined on a deleted neighborhood $U$ of $a \in \mathbb{R}^n$. We want to say what $\lim_{x \to a} f(x)$ means. We first consider some examples with $n = 2$, where we change notation slightly by writing $(x,y)$ for $x$.

VIII.1.3.2. Examples. (i) Let

$$f(x,y) = \frac{xy}{x^2 + y^2}$$

for $(x,y) \neq (0,0)$. Then $f$ is defined in a deleted neighborhood of $(0,0)$ in $\mathbb{R}^2$. What should $\lim_{(x,y) \to (0,0)} f(x,y)$ be?

If $(0,0)$ is approached along either the $x$- or $y$-axis, the function is identically 0, so the limit could only reasonably be 0. On the other hand, $f(t,t) = \frac{1}{2}$ for all $t \neq 0$, so if we approach the origin along the line $x = y$ the function value approaches $\frac{1}{2}$. Since the limit should be unique and we approach different values from different directions, we should say the limit does not exist.

Geometrically, the graph is so wrinkled near the origin that no point can be added there to make the graph a reasonable unbroken surface (Figure VIII.1).
Figure VIII.1: Maple plot of $f(x, y) = \frac{xy}{x^2 + y^2}$
(ii) Here is a more striking version of the same phenomenon: Let

\[ f(x, y) = \frac{x^2y}{x^4 + y^2} \]

for \((x, y) \neq (0, 0)\). Then \(f\) is defined in a deleted neighborhood of \((0, 0)\) in \(\mathbb{R}^2\). What should \(\lim_{(x,y) \to (0,0)} f(x, y)\) be?

Again, if \((0, 0)\) is approached along either the \(x\)- or \(y\)-axis, the function is identically 0, so the limit could only reasonably be 0. And if we approach along the line \(y = mx\), we have

\[
\lim_{t \to 0} f(t, mt) = \lim_{t \to 0} \frac{mt^3}{t^4 + m^2t^2} = 0
\]

thus the function approaches 0 as \((0, 0)\) is approached radially from any direction. Does this mean that we should say the limit is 0?

Suppose we approach along the parabola \(y = x^2\). We then have

\[
\lim_{t \to 0} f(t, t^2) = \lim_{t \to 0} \frac{t^4}{2t^4} = \frac{1}{2}
\]

so the limit cannot exist. This time the crinkle in the graph is curved (Figure VIII.2).
Figure VIII.2: Maple plot of $f(x, y) = \frac{x^2 y}{x^3 + y^2}$
VIII.2. Partial and Directional Derivatives

If \( f \) is a function of several variables, one simple way to partially analyze the behavior of \( f \) is to hold all but one of the variables fixed, so that \( f \) becomes simply a function of the one remaining variable. This leads to the notion of partial derivative. Partial derivatives turn out to give considerably more information about the function than might be at first expected, at least in good cases. The notion can then be generalized to directional derivatives in any direction.

VIII.2.1. Partial Derivatives

**VIII.2.1.1. Definition.** Let \( U \) be an open set in \( \mathbb{R}^n \), and \( f : U \to \mathbb{R}^m \) a function. Let \( \mathbf{a} = (a_1, \ldots, a_n) \in U \). For \( 1 \leq k \leq n \), we define the partial derivative

\[
\frac{\partial f}{\partial x_k}(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{h}
\]

where

\[
\frac{1}{h} \left[ f(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n) - f(a_1, \ldots, a_n) \right]
\]

if the limit exists.

In other words, \( \frac{\partial f}{\partial x_k} \) is the derivative of \( f \) when it is thought of as only a function of \( x_k \) by holding the other variables fixed: if \( g(x) = f(a_1, \ldots, a_{k-1}, x, a_{k+1}, \ldots, a_n) \)

then

\[
\frac{\partial f}{\partial x_k}(\mathbf{a}) = g'(a_k)
\]

**VIII.2.1.2.** If \( \frac{\partial f}{\partial x_k}(\mathbf{a}) \) exists, it is a vector in \( \mathbb{R}^m \). Partial derivatives are most commonly discussed in the case \( m = 1 \), in which case \( \frac{\partial f}{\partial x_k}(\mathbf{a}) \in \mathbb{R} \). The relation with the general case is that if \( (f_1, \ldots, f_m) \) are the coordinate functions of \( f \), then \( \frac{\partial f}{\partial x_k}(\mathbf{a}) \) exists if and only if \( \frac{\partial f_j}{\partial x_k}(\mathbf{a}) \) exists for all \( 1 \leq j \leq m \), in which case

\[
\frac{\partial f}{\partial x_k}(\mathbf{a}) = \left( \frac{\partial f_1}{\partial x_k}(\mathbf{a}), \ldots, \frac{\partial f_m}{\partial x_k}(\mathbf{a}) \right)
\]

**VIII.2.1.3.** The partial derivative symbol \( \partial \) appears to be an italic Cyrillic \( d \). However, the origin of the notation is somewhat obscure. Alternate notations are sometimes used, especially if the variables are denoted \( x, y, z, \ldots \): if \( f \) is a function of, say, \( x, y \), and \( z \), \( \frac{\partial f}{\partial x} \) is sometimes denoted \( f_x \) or \( f_1 \), \( \frac{\partial f}{\partial y} \) is denoted \( f_y \) or \( f_2 \), etc. We will mostly stick to the \( \frac{\partial}{\partial x} \), etc., notation, except in (\).

**VIII.2.1.4.** It is easy to get caught in confusing or ambiguous notation concerning partial derivatives if one is not careful, and many references use sloppy notation in places. For example, suppose \( u = f(x, y, z) \) and \( z = g(x, y) \); then what does \( \frac{\partial u}{\partial x} \) mean? As a variation, if \( u = f(x, y) \) and \( y = g(x) \), then one can make sense of both \( \frac{\partial u}{\partial x} \) and \( \frac{du}{dx} \). To avoid such nasty ambiguities, we will sometimes seem to be overly pedantic in our notation.
VIII.2.1.5. Like ordinary derivatives, partial derivatives are defined pointwise for each fixed \( a \in U \). However, after the fact each partial derivative \( \frac{\partial f}{\partial x_k} \) may be thought of as a function from \( U \), or the subset on which it is defined, to \( \mathbb{R}^m \).

VIII.2.1.6. When a function \( f \) of two or more variables is defined by a formula, partial derivatives can be calculated by the usual rules of differentiation, treating the remaining variables as constants. For example, if
\[
f(x, y, z) = x^2y^3 + \frac{\cos yz}{z^2 + 1} - e^{xyz}
\]
we have
\[
\frac{\partial f}{\partial x}(x, y, z) = 2xy^3 - yze^{xyz}
\]
\[
\frac{\partial f}{\partial y}(x, y, z) = 3x^2y^2 - \frac{z \sin yz}{z^2 + 1} - xze^{xyz}
\]
\[
\frac{\partial f}{\partial z}(x, y, z) = \left( z^2 + 1 \right) \left( -y \sin yz - (\cos yz)(2z) \right) - yze^{xyz}.
\]

Showing that such calculations are valid takes a little (easy) work, though, a point overlooked in many references; see Exercise VIII.2.4.1.

Higher-Order Partials

VIII.2.1.7. If \( f : U \to \mathbb{R}^m \) is a function, and the partial derivatives \( \frac{\partial f}{\partial x_k} \) are defined on \( U \), then partial derivatives of these partial derivatives are also (potentially) defined. For example, if \( \frac{\partial f}{\partial x_k} \) is defined in a neighborhood of \( a \in U \), then for \( 1 \leq j \leq n \) we can define
\[
\frac{\partial^2 f}{\partial x_j \partial x_k}(a) = \lim_{h \to 0} \frac{\frac{\partial f}{\partial x_k}(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - \frac{\partial f}{\partial x_k}(a_1, \ldots, a_n)}{h}
\]
if the limit exists.

Note the order of differentiation: \( \frac{\partial^2 f}{\partial x_j \partial x_k}(a) \) and \( \frac{\partial^2 f}{\partial x_k \partial x_j}(a) \) have different definitions if \( j \neq k \). It turns out (VIII.2.2.2.) that under mild continuity hypotheses these two partial derivatives are numerically equal.

VIII.2.1.8. We may keep going, defining third-order partial derivatives \( \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} \), etc. The \( i, j, k \) need not be distinct.

VIII.2.1.9. In the definition, we may have \( j = k \) or \( j \neq k \). Thus, if \( f \) is a function \( f(x, y) \) of two variables, there are four second-order partials: \( \frac{\partial^2 f}{\partial x^2} \) (usually denoted \( \frac{\partial^2 f}{\partial x^2} \)), \( \frac{\partial^2 f}{\partial x \partial y} \), \( \frac{\partial^2 f}{\partial y \partial x} \), and \( \frac{\partial^2 f}{\partial y^2} \). The middle two, called mixed partials, are often but not always equal. The variety increases rapidly with higher derivatives and/or more variables.

VIII.2.1.10. We may also use the alternate notation \( f_{xx} \) or \( f_{11} \) for \( \frac{\partial^2 f}{\partial x^2} \), etc. When the alternate notation is used, it is conventional to take derivatives from left to right, unlike in the Leibniz notation where derivatives go from right to left: thus \( f_{xy} \) or \( f_{12} \) denotes \( \frac{\partial^2 f}{\partial y \partial x} \), i.e. the function is first differentiated with respect to \( x \) and then with respect to \( y \), not \( \frac{\partial^2 f}{\partial x \partial y} \) which is denoted \( f_{yx} \) or \( f_{21} \). (The distinction is academic in good cases.)
VIII.2.2. Equality of Mixed Partials

In general, the order in which the derivatives are taken in a higher-order partial derivative makes a difference. But under a mild continuity hypothesis, the order does not matter.

VIII.2.2.1. Example. Let

\[ f(x, y) = \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \]

Then \( f \) is \( C^1 \) on \( \mathbb{R}^2 \): we have

\[ \frac{\partial f}{\partial x}(x, y) = \begin{cases} \frac{y(x^4 + 4x^2y^2 - y^4)}{(x^2 + y^2)^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \]

\[ \frac{\partial f}{\partial y}(x, y) = \begin{cases} \frac{-x(x^4 + 4x^2y^2 - x^4)}{(x^2 + y^2)^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \]

The absolute value of the numerators are bounded by \( 6\| (x, y) \|^5 \), so the partials are continuous at \((0, 0)\).

Also, \( \frac{\partial^2 f}{\partial x \partial y} \) and \( \frac{\partial^2 f}{\partial y \partial x} \) are defined everywhere on \( \mathbb{R}^2 \), and

\[ \frac{\partial^2 f}{\partial x \partial y}(0, 0) = \lim_{h \to 0} \frac{\frac{\partial f}{\partial y}(h, 0) - \frac{\partial f}{\partial y}(0, 0)}{h} = 1 \]

\[ \frac{\partial^2 f}{\partial y \partial x}(0, 0) = \lim_{k \to 0} \frac{\frac{\partial f}{\partial x}(0, k) - \frac{\partial f}{\partial x}(0, 0)}{k} = -1 \]

so the second-order mixed partials exist at \((0, 0)\) are defined but unequal.

However, \( \frac{\partial^2 f}{\partial x \partial y} \) and \( \frac{\partial^2 f}{\partial y \partial x} \) are not continuous at \((0, 0)\).

Actually, \( f \) has continuous partials of all orders everywhere except at the origin, i.e. \( f \) is \( C^\infty \) on \( \mathbb{R}^2 \setminus \{(0, 0)\} \).

The graph has a subtle crinkle near the origin (Figure VIII.3).

VIII.2.2. Theorem. Let \( U \) be an open set in \( \mathbb{R}^n \), \( f: U \to \mathbb{R}^m \), \( a \in U \). Suppose, for some \( i \) and \( j \), \( \frac{\partial f}{\partial x_i} \) and \( \frac{\partial f}{\partial x_j} \) exist everywhere on \( U \), and \( \frac{\partial^2 f}{\partial x_i \partial x_j} \) exists everywhere on \( U \) and is continuous at \( a \). Then \( \frac{\partial^2 f}{\partial x_i \partial x_j}(a) \) is defined and

\[ \frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a) \]

Proof: There is no content to the statement if \( i = j \), so we assume \( i \neq j \).

We first make two reductions. By considering the coordinate functions, it suffices to assume \( m = 1 \). And, since only the coordinates \( i \) and \( j \) are varied in computing the partials, we may assume \( n = 2 \). So we suppose
Figure VIII.3: Maple plot of graph of $f(x, y) = \frac{xy(x^2 - y^2)}{x^2 + y^2}$

$U \subseteq \mathbb{R}^2$, $f : U \rightarrow \mathbb{R}$, $(a, b) \in U$, that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist on $U$, and $\frac{\partial^2 f}{\partial x \partial y}$ is defined on $U$ and continuous at $(a, b)$. We want to show that $\frac{\partial^2 f}{\partial y \partial x}(a, b)$ is defined and equal to $c := \frac{\partial^2 f}{\partial x \partial y}(a, b)$.

Let $\epsilon > 0$, and fix $\delta > 0$ such that if $|h| < \delta$ and $|k| < \delta$, then $(a + h, b + k) \in U$ and

$$\left| \frac{\partial^2 f}{\partial x \partial y}(a + h, b + k) - c \right| < \epsilon .$$

Set $V = \{(h, k) : |h| < \delta, |k| < \delta\}$. Then for $(h, k) \in V$, set

$$\Delta(h, k) = f(a + h, b + k) - f(a, b + k) - f(a + h, b) + f(a, b) .$$

If $(h, k) \in V$ is fixed, for $0 \leq t \leq 1$ set

$$\phi(t) = f(a + h, b + tk) - f(a, b + tk) .$$

Then $\phi$ is differentiable on $[0, 1]$ and by the Chain Rule

$$\phi'(t) = \frac{\partial f}{\partial y}(a + h, b + tk) \cdot k - \frac{\partial f}{\partial y}(a, b + tk) \cdot k .$$
Apply the MVT to get a \( t \in (0,1) \) such that
\[
k \frac{\partial f}{\partial y}(a + h, b + tk) - k \frac{\partial f}{\partial y}(a, b + tk) = \phi'(t)(1 - 0) = \phi(1) - \phi(0) = \Delta(h, k).
\]
Fix such a \( t \), and for \( 0 \leq s \leq 1 \) set
\[
\psi(s) = k \frac{\partial f}{\partial y}(a + sh, b + tk).
\]
Then \( \psi \) is differentiable on \([0,1]\) and
\[
\psi'(s) = k \frac{\partial^2 f}{\partial x \partial y}(a + sh, b + tk) \cdot h.
\]
Again use the MVT to obtain an \( s \in (0,1) \) such that
\[
hk \frac{\partial^2 f}{\partial x \partial y}(a + sh, b + tk) = \psi'(s)(1 - 0) = \psi(1) - \psi(0) = \Delta(h, k).
\]
For every \((h,k) \in V\) there is such a pair \((s,t) \in (0,1)^2\) (which depends on \((h,k)\)), and \((sh,tk) \in V\), so if \((h,k) \in V\), \( h \neq 0, k \neq 0 \), we have
\[
\left| \frac{\Delta(h,k)}{hk} - c \right| < \epsilon.
\]
Now fix \( k, 0 < |k| < \delta \). We have
\[
\lim_{h \to 0} \frac{\Delta(h,k)}{hk} = \frac{1}{k} \lim_{h \to 0} \left[ \frac{f(a + h, b + k) - f(a, b + k)}{h} - \frac{f(a + h, b) - f(a, b)}{h} \right] = \frac{1}{k} \left[ \frac{\partial f}{\partial x}(a, b + k) - \frac{\partial f}{\partial x}(a, b) \right]
\]
so the limit exists and we have that
\[
\left| \frac{1}{k} \left[ \frac{\partial f}{\partial x}(a, b + k) - \frac{\partial f}{\partial x}(a, b) \right] - c \right| \leq \epsilon
\]
if \( 0 < |k| < \delta \). Since \( \epsilon > 0 \) is arbitrary, we conclude that
\[
\frac{\partial^2 f}{\partial y \partial x}(a,b) = \lim_{k \to 0} \frac{\partial f}{\partial x}(a, b + k) - \frac{\partial f}{\partial x}(a, b)
\]
exists and equals \( c \).

**Remark.** (i) In **VIII.2.2.2.**, the hypothesis that \( \frac{\partial f}{\partial x} \) exists on \( U \) is implicit and automatic in the hypothesis that \( \frac{\partial^2 f}{\partial x \partial y} \) exists on \( U \).

(ii) Let \( g \) be a continuous but nowhere differentiable function from \( \mathbb{R} \) to \( \mathbb{R} \), and let \( f : \mathbb{R}^2 \to \mathbb{R} \) be defined by \( f(x,y) = g(x) \). Then \( f \) is continuous, and \( \frac{\partial f}{\partial y} \) exists everywhere and is identically zero; thus \( \frac{\partial^2 f}{\partial x \partial y} \) also exists everywhere and is identically zero, hence continuous. But \( \frac{\partial f}{\partial x} \) does not exist anywhere, so \( \frac{\partial^2 f}{\partial y \partial x} \) does not exist anywhere. So the hypothesis that \( \frac{\partial f}{\partial x} \) exists, which is necessary for the existence of \( \frac{\partial^2 f}{\partial x \partial y} \), is not automatic under the other hypotheses.
Corollary. Let $U$ be an open set in $\mathbb{R}^n$, and $f$ a $C^2$ function from $U$ to $\mathbb{R}^m$. Then, for any $i$ and $j$,
\[
\frac{\partial^2 f}{\partial x_j \partial x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j}
\]
on $U$.

 VIII.2.2.5. VIII.2.2.4. is called Clairaut’s Theorem. The version in VIII.2.2.2. can be found in [Str81, p. 377]. Statements of the result in some real analysis texts have unnecessary extra hypotheses, and the “proofs” in many books have flaws. See VIII.4.1.7. for a variation, and XIV.3.11.7. for a simple alternate proof of VIII.2.2.4. based on Fubini’s Theorem.

 VIII.2.2.6. Corollary. Let $U$ be an open set in $\mathbb{R}^n$, and $f$ a $C^r$ function from $U$ to $\mathbb{R}^m$ ($r > 1$). Then any two mixed partials of $f$ of total order $\leq r$ in which each coordinate direction appears the same number of times are equal on $U$.

To show this, apply VIII.2.2.4. repeatedly to $f$ and its partial derivatives. For example, if $n \geq 5$ and $r \geq 4$, to show that
\[
\frac{\partial^4 f}{\partial x_2 \partial x_5 \partial x_3 \partial x_1} = \frac{\partial^4 f}{\partial x_2 \partial x_3 \partial x_5 \partial x_1}
\]
apply VIII.2.2.4. to $\frac{\partial f}{\partial x_1}$ to conclude that
\[
\frac{\partial^3 f}{\partial x_5 \partial x_3 \partial x_1} = \frac{\partial^3 f}{\partial x_3 \partial x_5 \partial x_1}
\]
amd then take the partial derivative of this function with respect to $x_2$. Any permutation of the order of derivatives can be built up from interchanges of successive derivatives, which can be justified in this way.

 VIII.2.2.7. This result leads to some convenient notation. A multi-index of dimension $n$ is an $n$-tuple $\alpha = (r_1, \ldots, r_n)$, where the $r_k$ are in $\mathbb{N} \cup \{0\}$. The total degree of $\alpha$ is
\[
|\alpha| = r_1 + \cdots + r_n.
\]
If $\alpha$ is a multi-index of dimension $n$ and total degree $r$, $U$ an open subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}^m$ is $C^r$, write
\[
\frac{\partial^\alpha f}{\partial x^\alpha} = \frac{\partial^r f}{\partial x_1^{r_1} \partial x_2^{r_2} \cdots \partial x_n^{r_n}}.
\]
By VIII.2.2.6., the successive partial derivatives may be taken in any order.

There is a “converse” to Clairaut’s Theorem which is important in vector field theory and in differential equations (compare with XIV.3.11.7.).
Theorem. Let $U$ be an open rectangle or open disk in $\mathbb{R}^2$, and $g$ and $h$ continuous functions from $U$ to $\mathbb{R}$ such that $\frac{\partial g}{\partial y}$ and $\frac{\partial h}{\partial x}$ are defined, continuous, and equal everywhere on $U$. Then there is a $C^1$ function $f$ on $U$, unique up to adding a constant, with $g = \frac{\partial f}{\partial x}$ and $h = \frac{\partial f}{\partial y}$.

Proof: For uniqueness, if $f$ and $F$ satisfy the conclusion, then $\phi = F - f$ is $C^1$ and $D\phi = 0$, so $\phi$ is constant.

For existence, let $(a, b)$ be the center point of $U$. For $(x, y) \in U$, define

$$f_1(x, y) = \int_b^y h(a, s) \, ds + \int_a^x g(t, y) \, dt$$
$$f_2(x, y) = \int_a^x g(t, b) \, dt + \int_b^y h(x, s) \, ds$$

(note that the paths lie in $U$). By the Fundamental Theorem of Calculus (FTC1), $\frac{\partial f_1}{\partial x} = g$ and $\frac{\partial f_2}{\partial y} = h$. So it suffices to show $f_1 = f_2$.

Fix $(x, y) \in U$. Then, by the Fundamental Theorem of Calculus (FTC2),

$$f_1(x, y) - f_2(x, y) = \int_a^x [g(t, y) - g(t, b)] \, dt - \int_b^y [h(x, s) - h(a, s)] \, ds$$

$$= \int_a^x \int_a^b \frac{\partial g}{\partial y}(t, s) \, ds \, dt - \int_b^y \int_a^x \frac{\partial h}{\partial x}(t, s) \, dt \, ds = 0$$

by Fubini’s Theorem (XIV.3.10. or XIV.3.5.2.) and the fact that $\frac{\partial g}{\partial y} = \frac{\partial h}{\partial x}$.

The equality of $f_1$ and $f_2$ is a special case of Green’s Theorem ()

VIII.2.2.9. The result can be extended to hold on a general simply connected open set (). But it fails in general on open sets which are not simply connected (). Using more advanced integration, the continuity hypotheses can be relaxed, but it is clumsy to make a precise statement. Of course, if $g$ and $h$ are $C^r$, the $f$ is $C^{r+1}$ (but the uniqueness is still among $C^1$ functions).
VIII.2.3. Directional Derivatives

VIII.2.3.1. The definition of partial derivatives can be rephrased in vector notation. We write \( \{e_1, \ldots, e_n\} \) for the standard basis vectors of \( \mathbb{R}^n \), i.e. \( e_k \) has 1 in the \( k \)'th place and 0’s elsewhere (the notation is a little ambiguous since \( n \) does not appear; we could write \( e_k^{(n)} \) if it is necessary to specify). If \( n = 3, e_1, e_2, e_3 \) are usually written \( i, j, k \) respectively.

With this notation, if \( U \subseteq \mathbb{R}^n \) and \( f : U \to \mathbb{R}^m \) is a function (we normally only consider the case \( m = 1 \)), then we have, for \( a \in U \) and \( 1 \leq k \leq n \),

\[
\frac{\partial f}{\partial x_k}(a) = \lim_{h \to 0} \frac{f(a + he_k) - f(a)}{h}
\]

(provided the limit exists).

We may generalize this definition (we only write the generalization in the case \( m = 1 \), although it can be done for general \( m \)):

VIII.2.3.2. Definition. Let \( U \subseteq \mathbb{R}^n \), and \( f : U \to \mathbb{R} \) a function. Let \( u \) be a nonzero vector in \( \mathbb{R}^n \) (which is almost always taken to be a unit vector, i.e. \( \|u\| = 1 \)). The directional derivative of \( f \) at \( a \in U \) with respect to \( u \), or in the direction \( u \) if \( u \) is a unit vector, is

\[
\frac{\partial f}{\partial u}(a) = \lim_{h \to 0} \frac{f(a + hu) - f(a)}{h}
\]

(provided the limit exists).

Partial derivatives are the special case \( u = e_k \) for some \( k \) (note that \( e_k \) is a unit vector):

\[
\frac{\partial f}{\partial x_k}(a) = \frac{\partial f}{\partial e_k}(a).
\]

VIII.2.3.3. There is a nice geometric interpretation of directional derivatives (and in particular partial derivatives). If \( f : \mathbb{R}^n \to \mathbb{R} \), the graph of \( f \) can be regarded as a hypersurface in \( \mathbb{R}^n \times \mathbb{R} \cong \mathbb{R}^{n+1} \). If \( u \) is a unit vector in \( \mathbb{R}^n \), there is a vertical plane through \((a, f(a))\) parallel to the plane spanned by \( u \) and \( e_{n+1} \). The intersection of this plane with the graph is a curve. If the plane is identified with \( \mathbb{R}^2 \) by taking \((a, f(a))\) as the origin and \( u \) and \( e_{n+1} \) as the coordinate directions, the curve becomes a plane curve in \( \mathbb{R}^2 \) through the origin, and \( \frac{\partial f}{\partial u}(a) \) is just the slope of the tangent line to this curve at the origin (Figure ()). It also has a more elementary interpretation as the rate of increase of \( f \) as one moves from \( a \) with unit speed along the line in the direction \( u \).

VIII.2.3.4. In the rarely used case where \( u \) is not a unit vector, let \( v = \frac{u}{\|u\|} \) be a unit vector in the direction of \( u \). Then we have, writing \( t = \|u\|h \),

\[
\frac{\partial f}{\partial u}(a) = \lim_{h \to 0} \frac{f(a + hu) - f(a)}{h} = \|u\| \lim_{t \to 0} \frac{f(a + tv) - f(a)}{t} = \|u\| \frac{\partial f}{\partial v}(a).
\]

809
VIII.2.4. Exercises

VIII.2.4.1. (a) Prove the following rules of partial differentiation, carefully stating each as a theorem.

For each, $U$ is an open set in $\mathbb{R}^n$, $f$ and $g$ are real-valued functions on $U$ which are differentiable at $a \in U$, $h : \mathbb{R} \to \mathbb{R}$ is differentiable at $f(a)$, $c \in \mathbb{R}$ (also denoting the constant function $c$ from $\mathbb{R}^n$ to $\mathbb{R}$), and $1 \leq j, k \leq n$. $\pi_j$ denotes the $j$'th coordinate function, i.e. $\pi_j(x_1, \ldots, x_n) = x_j$.

$$\frac{\partial (c)}{\partial x_k} (a) = 0 .$$

$$\frac{\partial (\pi_j)}{\partial x_k} (a) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases} .$$

$$\frac{\partial (f + g)}{\partial x_k} (a) = \frac{\partial f}{\partial x_k} (a) + \frac{\partial g}{\partial x_k} (a) .$$

$$\frac{\partial (cf)}{\partial x_k} (a) = c \frac{\partial f}{\partial x_k} (a) .$$

$$\frac{\partial (fg)}{\partial x_k} (a) = \frac{\partial f}{\partial x_k} (a)g(a) + f(a) \frac{\partial g}{\partial x_k} (a) .$$

$$\frac{\partial (f/g)}{\partial x_k} (a) = \frac{\frac{\partial f}{\partial x_k} (a)g(a) - f(a) \frac{\partial g}{\partial x_k} (a)}{|g(a)|^2} \quad \text{if } g(a) \neq 0 .$$

$$\frac{\partial (h \circ f)}{\partial x_k} (a) = h'(f(a)) \frac{\partial f}{\partial x_k} (a) .$$

(b) Use these rules to justify calculating partial derivatives as in VIII.2.1.6.
VIII.3. Differentiability of Multivariable Functions

VIII.3.1. Differentiability and the Derivative

VIII.3.1.1. What should it mean for a function from one Euclidean space to another to be differentiable at a point, and what kind of object should the derivative be? In the one-dimensional case, \( f \) is differentiable at \( a \) if the limit
\[
\lim_{h \to 0} \frac{f(a + h) - f(a)}{h}
\]
exists; if it does, the limit is called \( f'(a) \). But if \( f : \mathbb{R}^n \to \mathbb{R}^m \) and \( a \in \mathbb{R}^n, n > 1 \), the analogous expression
\[
\lim_{h \to 0} \frac{f(a + h) - f(a)}{h}
\]
does not even make sense since division by a vector is not defined.

VIII.3.1.2. We can change the one-dimensional definition into an equivalent one which does make sense when redecorated in this way, however. The following statements are equivalent by trivial calculation:
\[
\lim_{h \to 0} \frac{f(a + h) - f(a)}{h} = f'(a)
\]
\[
\lim_{h \to 0} \left[ \frac{f(a + h) - f(a)}{h} - f'(a) \right] = 0
\]
\[
\lim_{h \to 0} \frac{f(a + h) - f(a) - f'(a)h}{h} = 0
\]
\[
\lim_{h \to 0} \frac{f(a + h) - f(a) - f'(a)h}{|h|} = 0
\]
(for the last equivalence, consider left and right limits).

VIII.3.1.3. This last expression is almost in the form needed to make a reasonable definition in the vector case. However, it may be unclear how to modify the term \( f'(a)h \) in the numerator. A key to the proper approach comes from differential approximation: the best affine approximation to \( f \) near \( a \) is the function \( \phi(a + h) = f(a) + df(a)(h) = f(a) + f'(a)h \), so the numerator is \( f(a + h) - \phi(a + h) \), and the statement that the limit is 0 states that \( f \) can be closely approximated by this affine function.

A reasonable geometric criterion for differentiability at \( a \) in the vector case is that the function must be closely approximable near \( a \) by an affine function, i.e. there is an affine function \( \phi : \mathbb{R}^n \to \mathbb{R}^m \) (if \( f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m \)) such that
\[
\lim_{h \to 0} \frac{f(a + h) - \phi(a + h)}{|h|} = 0
\]
where the fraction is interpreted as the vector
\[
\frac{1}{|h|} \left[ f(a + h) - \phi(a + h) \right] \in \mathbb{R}^m.
\]
The affine function \( \phi \) is of the form \( \phi(x) = c + T(x - a) \), where \( T : \mathbb{R}^n \to \mathbb{R}^m \) is a linear function and \( c \) is a constant. Since we must have \( \phi(a) = f(a) \), \( c = f(a) \). So we conclude that the following definition is reasonable:
VIII.3.1.4. **Definition.** Let $U \subseteq \mathbb{R}^n$ be open, and $f : U \rightarrow \mathbb{R}^m$, $a \in U$. Then $f$ is differentiable at $a$ if there is a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ for which

$$\lim_{h \rightarrow 0} \frac{f(a + h) - f(a) - T(h)}{\|h\|} = 0.$$ 

VIII.3.1.5. We will see that the linear transformation $T$, if it exists, is unique (this can be proved directly; cf. Exercise ()). This linear transformation will be called the **total differential** of $f$ at $a$, denoted $df(a)$, and its matrix the **total derivative matrix** of $f$ at $a$, denoted $Df(a)$. Note that $df(a)$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $Df(a)$ is an $m \times n$ matrix. $Df(a)$ has a simple formula in terms of partial derivatives of the coordinate functions of $f$ (VIII.3.1.9).

VIII.3.1.6. Since a linear transformation from $\mathbb{R}$ to $\mathbb{R}$ is of the form $T(x) = cx$ for some constant $c$, we recover the usual one-dimensional definition as modified in VIII.3.1.2., and $df(a)$ is the linear function $df(a)(x) = f'(a)x$ and $Df(a)$ is the $1 \times 1$ matrix $[f'(a)]$. So our definitions are (up to rephrasing) consistent.

VIII.3.1.7. Note that the definition of differentiability can be rephrased to require that

$$\lim_{h \rightarrow 0} \frac{\|f(a + h) - f(a) - T(h)\|}{\|h\|} = 0$$

since a limit $\lim_{x \rightarrow b} g(x) = 0$ of a vector function $g$ holds if and only if $\lim_{x \rightarrow b} \|g(x)\| = 0$. It will sometimes be more convenient to use this version.

**Uniqueness and a Formula**

VIII.3.1.8. Suppose $f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable at $a$, with a linear transformation $T$ which works in the definition. Fix $j$, $1 \leq j \leq n$. As a special case of the limit in the definition, we have

$$\lim_{t \rightarrow 0} \frac{f(a + te_j) - f(a) - T(te_j)}{|t|} = 0.$$ 

By considering left and right limits, we may eliminate the absolute value from the denominator:

$$\lim_{t \rightarrow 0} \frac{f(a + te_j) - f(a) - T(te_j)}{t} = 0$$

$$\lim_{t \rightarrow 0} \left[ \frac{f(a + te_j) - f(a)}{t} - T(e_j) \right] = 0$$

$$\lim_{t \rightarrow 0} \frac{f(a + te_j) - f(a)}{t} = T(e_j).$$

In particular, the limit on the left exists and equals the vector $T(e_j)$, which is the $j$'th column of the matrix of $T$. But the left side is the definition of $\frac{\partial f}{\partial x_j}(a)$. So we conclude:

812
VIII.3.1.9. PROPOSITION. Let \( f = (f_1, \ldots, f_m) : U \subseteq \mathbb{R}^n \to \mathbb{R}^m \) be differentiable at \( a \in U \). Then

(i) All partial derivatives \( \frac{\partial f_i}{\partial x_j}(a) \) exist.

(ii) The linear transformation \( T \) in the definition of differentiability at \( a \) is unique, and the \( ij \)'th entry of \( [T] \) is \( \frac{\partial f_i}{\partial x_j}(a) \) for each \( i \) and \( j \).

Thus the total differential \( df(a) \) and total derivative \( Df(a) \) are well defined, and

\[
Df(a) = \begin{bmatrix}
\frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) \\
\vdots & & \vdots \\
\frac{\partial f_m}{\partial x_1}(a) & \cdots & \frac{\partial f_m}{\partial x_n}(a)
\end{bmatrix}.
\]

VIII.3.1.10. The matrix

\[
\begin{bmatrix}
\frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) \\
\vdots & & \vdots \\
\frac{\partial f_m}{\partial x_1}(a) & \cdots & \frac{\partial f_m}{\partial x_n}(a)
\end{bmatrix}
\]

is defined whenever all partial derivatives exist at \( a \), and existence of this matrix is a necessary condition for differentiability at \( a \). But existence of this matrix is not sufficient for differentiability (cf. VIII.3.6.1–VIII.3.6.4.). We will not call this matrix the total derivative matrix or use the notation \( Df(a) \) unless \( f \) is differentiable at \( a \).

VIII.3.1.11. EXAMPLE. Let \( T : \mathbb{R}^n \to \mathbb{R}^m \) be linear. If \( a \) and \( h \) are any vectors in \( \mathbb{R}^n \), we have

\[
T(a + h) - T(a) - T(h) = 0
\]

so, fixing \( a \), we certainly have

\[
\lim_{h \to 0} \frac{T(a + h) - T(a) - T(h)}{\|h\|} = 0
\]

and thus \( T \) is differentiable at \( a \) with \( dT(a) = T \) and \( DT(a) = [T] \). So linear functions are differentiable everywhere; the differential of a linear function at any point is the function itself, and the total derivative matrix of \( T \) at any point is \( [T] \) (i.e. “the derivative of a linear function is constant”).

Coordinatewise Differentiability

We can examine differentiability coordinatewise in the range space:
VIII.3.1.12. Proposition. Let $U \subseteq \mathbb{R}^n$ be open, $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$, $a \in U$. Then $f$ is differentiable at $a$ if and only if each $f_i$ is differentiable at $a$. In this case, the total differential $df(a)$ has coordinate functions $df_i(a)$.

Proof: This is an immediate consequence of the Coordinatewise Criterion for limits (): if $T = (T_1, \ldots, T_m)$ is linear, then
\[
\lim_{h \to 0} \frac{f(a + h) - f(a) - T(h)}{\|h\|} = 0
\]
if and only if
\[
\lim_{h \to 0} \frac{f_i(a + h) - f_i(a) - T_i(h)}{\|h\|} = 0
\]
for $1 \leq i \leq m$. 

Differentiability and Continuity

Here is a simple but important observation:

VIII.3.1.13. Proposition. Let $U \subseteq \mathbb{R}^n$ be open, $f : U \to \mathbb{R}^m$, $a \in U$. If $f$ is differentiable at $a$, then $f$ is continuous at $a$.

Proof: From the assumption
\[
\lim_{h \to 0} \frac{f(a + h) - f(a) - df(a)(h)}{\|h\|} = 0
\]

it follows that
\[
\lim_{h \to 0} [f(a + h) - f(a) - df(a)(h)] = 0
\]
and, since linear functions are continuous (), $\lim_{h \to 0} df(a)(h) = 0$, so $\lim_{h \to 0} [f(a + h) - f(a)] = 0$ also. 

VIII.3.1.14. Note that this result relies on the continuity of linear functions. This is automatic on finite-dimensional normed spaces, but not infinite-dimensional ones; so continuity is an issue if differentiability of functions between Banach spaces is considered ().

VIII.3.2. A Criterion for Differentiability

For a function of one variable, differentiability and existence of the derivative are the same thing by definition. But for a function of more than one variable, existence of all first-order partials is not sufficient to insure differentiability, or even continuity; cf. VIII.3.6.1–VIII.3.6.4. However, continuity of first-order partials is sufficient:
VIII.3.2.1. **Theorem.** Let $U$ be an open subset of $\mathbb{R}^n$, and $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$. If all partial derivatives $\frac{\partial f}{\partial x_j}$ exist on $U$ and are continuous at $a \in U$, then $f$ is differentiable at $a$.

**Proof:** First note that by VIII.3.1.12, it suffices to assume $m = 1$, i.e. $f : U \to \mathbb{R}$ and $\frac{\partial f}{\partial x_j}$ exists on $U$ and is continuous at $a$ for each $j$.

We prove the result by induction on $n$. If $n = 1$, the statement is a tautology (and the hypotheses are even unnecessarily strong!). Now assume the statement is true for $n$, and let $U$ be an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}$ with $\frac{\partial f}{\partial x_j}$ defined on $U$ and continuous at $c = (a, b) \in U$ for $1 \leq j \leq n + 1$. We need to show that $f$ is differentiable at $c$. Set

$$V = \{x \in \mathbb{R}^n : (x, b) \in U\}.$$

Then $V$ is an open set in $\mathbb{R}^n$ and $a \in V$. Define $g : V \to \mathbb{R}$ by $g(x) = f(x, b)$. For $1 \leq j \leq n$, we have $\frac{\partial g}{\partial x_j}(x) = \frac{\partial f}{\partial x_j}(x, b)$, so for $1 \leq j \leq n$, $\frac{\partial g}{\partial x_k}$ is defined on $V$ and continuous at $a$. Thus $g$ is differentiable at $a$ by the inductive hypothesis, i.e.

$$\lim_{h \to 0} \frac{g(a + h) - g(a) - Dg(h)}{\|h\|} = 0$$

$$\lim_{h \to 0} \frac{f(a + h, b) - f(a, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j}{\|h\|} = 0.$$

If $h \in \mathbb{R}^n$, $k \in \mathbb{R}$, write $(h, k) \in \mathbb{R}^{n+1}$. Since $\|h\| \leq \|(h, k)\|$ for any $k$, we also have

$$\lim_{(h, k) \to (0, 0)} \frac{f(a + h, b) - f(a, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j}{\|(h, k)\|} = 0 \quad (\text{VIII.1})$$

If $h$ and $k$ are small, by the Mean Value Theorem there is a $d_{h, k}$ between $b$ and $b + k$ such that

$$f(a + h, b + k) - f(a + h, b) = \frac{\partial f}{\partial x_{n+1}}(a + h, d_{h, k})k.$$

Now

$$\lim_{(h, k) \to (0, 0)} \frac{f(a + h, b + k) - f(a, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j - \frac{\partial f}{\partial x_{n+1}}(a, b)k}{\|(h, k)\|}$$

$$= \lim_{(h, k) \to (0, 0)} \frac{f(a + h, b) - f(a, b) + f(a + h, b + k) - f(a + h, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j - \frac{\partial f}{\partial x_{n+1}}(a, b)k}{\|(h, k)\|}$$

$$= \lim_{(h, k) \to (0, 0)} \left[ \frac{f(a + h, b) - f(a, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j}{\|(h, k)\|} + \frac{f(a + h, b + k) - f(a + h, b) - \frac{\partial f}{\partial x_{n+1}}(a, b)k}{\|(h, k)\|} \right]$$

$$= \lim_{(h, k) \to (0, 0)} \left[ \frac{f(a + h, b) - f(a, b) - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(a, b)h_j}{\|(h, k)\|} + \frac{\frac{\partial f}{\partial x_{n+1}}(a + h, d_{h, k}) - \frac{\partial f}{\partial x_{n+1}}(a, b)}{\|(h, k)\|} k \right].$$

The limit of the first term is 0 by VIII.1, and the limit of the second term is also 0 by continuity of $\frac{\partial f}{\partial x_{n+1}}$ at $c$ (using $|k| \leq \|(h, k)\|$). Thus $f$ is differentiable at $c$. 815
VIII.3.2.2. Remark. Examining the proof carefully, we do not need all the partial derivatives to be continuous at \( a \): the partial with respect to one variable can be discontinuous. In fact, the proof shows that if, for some \( r \), the restriction of \( f \) to an \( r \)-dimensional coordinate subspace is differentiable at \( a \), and the additional partial derivatives are continuous at \( a \), then \( f \) is differentiable at \( a \); cf. Exercise VIII.3.6.8.

VIII.3.2.3. The hypotheses of the theorem give sufficient conditions for \( f \) to be differentiable at \( a \). But these conditions are not necessary: cf. V.3.1.15., V.12., VIII.3.6.6–VIII.3.6.7. The main significance of the theorem is that the continuity conditions are often satisfied and easy to check in cases of interest, and the theorem is frequently by far the easiest way to establish differentiability.

The next special case of VIII.3.2.1. is the most important and commonly used one. We first make a definition:

VIII.3.2.4. Definition. Let \( U \) be an open set in \( \mathbb{R}^n \), and \( f : U \to \mathbb{R}^m \) a function. If \( r \in \mathbb{N} \), then \( f \) is a \( C^r \) function if \( f \) and all its partial derivatives of order \( \leq r \) exist and are continuous on \( U \).

If \( n = 1 \), this definition agrees with the definition in (); and the agreement goes deeper (VIII.4.1.5.). We extend the notation to say that \( f \) is \( C^0 \) on \( U \) if it is continuous on \( U \), and \( C^\infty \) on \( U \) if it is \( C^r \) on \( U \) for all \( r \), i.e. all partials of \( f \) of all orders exist and are continuous on \( U \).

VIII.3.2.5. Corollary. Let \( U \) be an open set in \( \mathbb{R}^n \), and \( f : U \to \mathbb{R}^m \) a \( C^1 \) function. Then \( f \) is differentiable at all points of \( U \).

VIII.3.2.6. Corollary. Let \( U \) be an open set in \( \mathbb{R}^n \), \( f : U \to \mathbb{R}^m \). If \( r \in \mathbb{N} \), suppose all partial derivatives of \( f \) of order \( r \) are continuous on \( U \). Then \( f \) itself and all partial derivatives of \( f \) of order \( < r \) are differentiable on \( U \), hence \textit{a fortiori} continuous on \( U \). In particular, \( f \) is \( C^r \) on \( U \).

\[ \text{Proof:} \] Working backwards, any partial derivative of \( f \) of order \( r - 1 \) has continuous first-order partials on \( U \), hence is differentiable (thus continuous) on \( U \). The same is then true of the partials of \( f \) of order \( r - 2 \), etc.

VIII.3.3. The Chain Rule

VIII.3.3.1. The Chain Rule concerns differentiability and the derivative of a composition of functions. Two previous results suggest that the derivative of a composition of two functions should be the product of the derivatives:

(i) If \( f, g : \mathbb{R} \to \mathbb{R} \) and \( a \in \mathbb{R} \), then \( (g \circ f)'(a) = g'(f(a))f'(a) \) (cf. V.3.4.3.).

(ii) If \( S : \mathbb{R}^n \to \mathbb{R}^m \) and \( T : \mathbb{R}^m \to \mathbb{R}^p \) are linear, then \( T \circ S : \mathbb{R}^n \to \mathbb{R}^p \) is linear and \( [T \circ S] = [T][S] \).

The Chain Rule of vector calculus generalizes these two facts. Note that, as always, the Chain Rule is a theorem, not a formula: it has hypotheses (that \( f \) and \( g \) are differentiable) and a conclusion with a theoretical part (that \( g \circ f \) is differentiable) as well as a formula for the total derivative of \( g \circ f \).

The Chain Rule is a prime example of the utility of vector calculus notation. The corresponding statement written in coordinates is unwieldy, although this form is often needed in applications.
VIII.3.3.2. **Theorem.** [Chain Rule] Let $U$ and $V$ be open sets in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, $f : U \to \mathbb{R}^m$ and $g : V \to \mathbb{R}^p$ with $f(U) \subseteq V$, $a \in U$, $b = f(a) \in V$. If $f$ is differentiable at $a$ and $g$ is differentiable at $b$, then $g \circ f$ is differentiable at $a$ and

$$D(g \circ f)(a) = Dg(b)Df(a) = Dg(f(a))Df(a).$$

Note that under the hypotheses, $Dg(b)$ is $p \times m$, $Df(a)$ is $m \times n$, and $D(g \circ f)(a)$ is $p \times n$, so the product of the matrices on the right is defined and is the same size as the matrix on the left (this is of course not a proof of the theorem, only an indication that the statement makes sense!)

It is easily seen that the two examples in VIII.3.3.1. are special cases of this result (cf. VIII.3.3.11.).

**Proof:** The proof is largely a matter of choosing good notation, which can be a little complicated to keep straight.

For $h$ in a neighborhood of $0$ in $\mathbb{R}^n$, set

$$\phi(h) = f(a + h) - f(a) - Df(a)h$$

so we have

$$\lim_{h \to 0} \frac{\phi(h)}{\|h\|} = \lim_{h \to 0} \frac{f(a + h) - f(a) - Df(a)h}{\|h\|} = 0.$$

Similarly, for $k$ in a neighborhood of $0$ in $\mathbb{R}^m$, set

$$\psi(k) = g(b + k) - g(b) - Dg(b)k$$

so we have

$$\lim_{k \to 0} \frac{\psi(k)}{\|k\|} = \lim_{k \to 0} \frac{g(b + k) - g(b) - Dg(b)k}{\|k\|} = 0.$$

Set $\omega(k) = \frac{\psi(k)}{\|k\|}$ for $k \neq 0$ and $\omega(0) = 0$. Then $\omega$ is continuous at $0$.

If $h$ is sufficiently close to $0$, then $f(a + h)$ is defined; and if we set

$$k = f(a + h) - f(a) = Df(a)h + \phi(h)$$

we have that $k \to 0$ as $h \to 0$ since $f$ is continuous at $a$, and

$$g(f(a + h)) = g(f(a)) = g(b + k) - g(b) = Dg(b)k + \psi(k)$$

so we have

$$g(f(a + h)) - g(f(a)) - Dg(b)Df(a)h = \frac{Dg(b)\phi(h)}{\|h\|} + \frac{\psi(k)}{\|h\|}.$$

Since $\|Dg(b)\phi(h)\| \leq \|Dg(b)\|\|\phi(h)\|$ and $\lim_{h \to 0} \frac{\phi(h)}{\|h\|} = 0$, the first term has limit $0$. We show the second term also has limit $0$. We have

$$\|k\| = \|f(a + h) - f(a)\| = \|Df(a)h + \phi(h)\| \leq \|Df(a)\|\|h\| + \|\phi(h)\|.$$
∥k∥ ≤ ∥Df(a)∥ + ∥ϕ(h)∥ ∥h∥

and since the second term goes to 0 as h → 0, it is bounded in a neighborhood of 0. Thus

\[ \lim_{h \to 0} \frac{\psi(k)}{\|h\|} = \lim_{h \to 0} \frac{\|k\|}{\|h\|} \omega(k) = 0 \]

since the first factor is bounded and the second factor tends to 0. Thus

\[ \lim_{h \to 0} \frac{g(f(a + h)) - g(f(a)) - Dg(b)Df(a)h}{\|h\|} = 0. \]

VIII.3.4. The Gradient and Tangent Planes

For a real-valued differentiable function f on a region in \( \mathbb{R}^n \), there is an alternate representation or interpretation of the total derivative \( Df \) which has a nice geometric interpretation.

VIII.3.4.1. If \( U \subseteq \mathbb{R}^n \), \( a \in U \), and \( f : U \to \mathbb{R} \) is differentiable at \( a \), then \( Df(a) \) is a \( 1 \times n \) matrix:

\[ Df(a) = \left[ \frac{\partial f}{\partial x_1}(a) \quad \frac{\partial f}{\partial x_2}(a) \quad \cdots \quad \frac{\partial f}{\partial x_n}(a) \right]. \]

A \( 1 \times n \) matrix can be interpreted as a row vector in \( \mathbb{R}^n \), and thus \( Df(a) \) can be alternately interpreted as a vector in \( \mathbb{R}^n \). When it is, it is given a different name to emphasize the change of point of view:

VIII.3.4.2. Definition. Let \( U \subseteq \mathbb{R}^n \), \( a \in U \), and \( f : U \to \mathbb{R} \) a real-valued function. If \( f \) is differentiable at \( a \), the gradient of \( f \) at \( a \) is the vector \( \nabla f(a) \) in \( \mathbb{R}^n \) defined by

\[ \nabla f(a) = \left( \frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), \cdots, \frac{\partial f}{\partial x_n}(a) \right). \]

Note that, as usual, \( \nabla f \) is regarded as a single symbol. Some references use the alternate notation \( \text{Grad}(f) \).

VIII.3.4.3. Just as the total derivative matrix of a function at \( a \) is technically defined when all first-order partials exist at \( a \), even if the function is not differentiable at \( a \), the formula for the gradient makes sense if all first-order partials are just defined at \( a \). But just as we only use the notation \( Df(a) \) if \( f \) is differentiable at \( a \), we only use the notation \( \nabla f(a) \) and say the gradient exists at \( a \) if \( f \) is actually differentiable at \( a \) (e.g. if all first-order partials of \( f \) are continuous at \( a \)).
Directional Derivatives Revisited

If \( f : U \to \mathbb{R} \) is differentiable at \( a \in U \subseteq \mathbb{R}^n \), then the directional derivatives can be easily calculated using the gradient. Let \( u \) be a nonzero vector in \( \mathbb{R}^n \), and define \( \gamma : \mathbb{R} \to \mathbb{R}^n \) by

\[
\gamma(t) = a + tu .
\]

Then \( \gamma'(0) = u \), or in matrix terms \( D\gamma(0) \) is the \( n \times 1 \) matrix \( u \), regarded as a column vector. Set \( \phi(t) = f(\gamma(t)) \). If \( U \) is open, then \( \phi \) is defined in an interval around 0, and by definition we have

\[
\frac{\partial f}{\partial u}(a) = \phi'(0)
\]

and by the Chain Rule we have

\[
\phi'(0) = D\phi(0) = Df(\gamma(0))\gamma'(0) = Df(a)u
\]

where \( u \) is regarded as an \( n \times 1 \) matrix (column vector). However, observe that the scalar \( Df(a)u \) is exactly the dot product \( \nabla f(a) \cdot u \). Thus we conclude:

VIII.3.4.4. Proposition. Let \( U \) be an open set in \( \mathbb{R}^n \), \( a \in U \), and \( f : U \to \mathbb{R} \) a real-valued function. If \( f \) is differentiable at \( a \) and \( \nabla f(a) \neq 0 \), then for any nonzero vector \( u \in \mathbb{R}^n \), the directional derivative of \( f \) at \( a \) with respect to \( u \) is defined and

\[
\frac{\partial f}{\partial u}(a) = \nabla f(a) \cdot u .
\]

VIII.3.4.5. This simple relationship does not necessarily hold if \( f \) is not differentiable at \( a \), even if all directional derivatives (and in particular all first-order partials) of \( f \) exist at \( a \). For example, if \( f : \mathbb{R}^2 \to \mathbb{R} \) is the function of VIII.3.6.5., then \( \frac{\partial f}{\partial x}(0) \) exists for all \( u \). But \( \frac{\partial f}{\partial x}(0) = \frac{\partial f}{\partial y}(0) = 0 \), so “\( \nabla f(0) \)” is \( 0 \), although \( \frac{\partial f}{\partial u}(0) \neq 0 \) for some \( u \)’s. This shows again that \( f \) cannot be differentiable at \( 0 \) (and, in particular, \( \nabla f(0) \) is actually undefined).

Geometric Interpretation of the Gradient

Returning to the case where \( f \) is differentiable at \( a \), we can interpret VIII.3.4.4. geometrically:

VIII.3.4.6. Corollary. Let \( U \) be an open set in \( \mathbb{R}^n \), \( a \in U \), and \( f : U \to \mathbb{R} \) a real-valued function. If \( f \) is differentiable at \( a \) and \( \nabla f(a) \neq 0 \), then for any unit vector \( u \in \mathbb{R}^n \), the directional derivative of \( f \) at \( a \) in the direction \( u \) is defined and

\[
\frac{\partial f}{\partial u}(a) = \|\nabla f(a)\| \cos \theta
\]

where \( \theta \) is the angle between \( \nabla f(a) \) and \( u \).

The conclusion also holds if \( \nabla f(a) = 0 \) (although \( \theta \) is not defined): then \( \frac{\partial f}{\partial u}(a) = 0 \) for any \( u \).
VIII.3.4.7. Thus $\frac{\partial f}{\partial u}(a)$ is maximum when $\theta = 0$, i.e. if and only if $u$ is a unit vector in the direction of $\nabla f(a)$ (in the case $\nabla f(a) \neq 0$). So if $\nabla f(a) \neq 0$, it points in the direction of greatest increase of $f$ at $a$, and the length of the gradient gives the maximum rate of increase. In the case $n = 2$, the graph of $f$ can be regarded as a surface in $\mathbb{R}^3$, and at any point of the surface where the function is differentiable with nonzero gradient the gradient points “straight uphill” (see Figure ()).

The Tangent Hyperplane

VIII.3.4.8. If $\nabla f(a) \neq 0$, then for a unit vector $u \in \mathbb{R}^n$ we have $\frac{\partial f}{\partial u}(a) = 0$ if and only if $u \perp \nabla f(a)$. The set of all such vectors spans an $(n - 1)$-dimensional subspace of $\mathbb{R}^n$.  

820
VIII.3.5. Functional Independence

VIII.3.5.1. Let $U$ be an open set in $\mathbb{R}^n$, and $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $C^1$ function. We say \{f_1, \ldots, f_m\} is functionally dependent on $U$ if there is a $C^1$ function $g$ on an open set $V$ in $\mathbb{R}^{n-1}$ such that, for some $k$, $1 \leq k \leq m$, we have

$$(f_1(x), \ldots, f_{k-1}(x), f_{k+1}(x), \ldots, f_m(x)) \in V$$

$$f_k(x) = g(f_1(x), \ldots, f_{k-1}(x), f_{k+1}(x), \ldots, f_m(x))$$

for all $x \in U$. Otherwise \{f_1, \ldots, f_m\} is functionally independent on $U$. (The $C^1$ requirement can be relaxed in various ways, but this is the most useful form of the definition.)

Geometrically, functional dependence says that $f(x)$ lies on the graph of $g$ in $\mathbb{R}^m = \mathbb{R}^m \setminus \mathbb{R}$ for all $x \in U$. Note that functional independence implies, but is much stronger than, linear independence (Exercise ()).

VIII.3.5.2. An apparently more general functional dependence is to just require that there is a function $F$, $C^1$ on an open set in $\mathbb{R}^m$ containing $f(U)$ with nonvanishing derivative, such that $F(f(x)) = 0$ for all $x \in U$. But this is actually not really more general, at least locally, by the Implicit Function Theorem.

VIII.3.5.3. In general, it is a difficult problem to determine whether a set of functions is functionally dependent. We will obtain from the Chain Rule a simple necessary condition in the case $n \geq m$, although this condition is by no means sufficient.

VIII.3.5.4. Proposition. Let $U$ be an open set in $\mathbb{R}^n$, and $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $C^1$ function, with $n \geq m$. If \{f_1, \ldots, f_m\} is functionally dependent, then for every $\{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$ and every $x \in U$ we have

$$\det \begin{bmatrix} \frac{\partial f_1}{\partial x_{j_1}}(x) & \cdots & \frac{\partial f_1}{\partial x_m}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{j_1}}(x) & \cdots & \frac{\partial f_m}{\partial x_m}(x) \end{bmatrix} = 0.$$ 

Proof: By reordering $f_1, \ldots, f_m$, we may assume the $k$ in the definition of functional dependence is $m$, i.e. there is a $C^1$ function $g$ such that $f_m(x) = g(f_1(x), \ldots, f_{m-1}(x))$ for all $x \in U$, to simplify notation. Fix $x \in U$, and write $y = (f_1(x), \ldots, f_{m-1}(x)) \in \mathbb{R}^{m-1}$. Then by the Chain Rule, for each $j$ we have

$$\frac{\partial f_m}{\partial x_j} = \sum_{k=1}^{m-1} \frac{\partial g}{\partial y_k}(y) \frac{\partial f_k}{\partial x_j}(x)$$

so the last row in the matrix is a linear combination of the first $m - 1$ rows (the $k$'th row is multiplied by $\frac{\partial g}{\partial y_k}(y)$), hence the matrix has zero determinant. $\blacksquare$
VIII.3.5.5. **Corollary.** Let $U$ be an open set in $\mathbb{R}^n$, and $f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $C^1$ function, with $n \geq m$. If there is an $a \in U$ and $\{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$ such that

$$
\begin{vmatrix}
\frac{\partial f_1}{\partial x_{j_1}}(a) & \cdots & \frac{\partial f_1}{\partial x_{j_m}}(a) \\
\vdots & \ddots & \vdots \\
\frac{\partial f_m}{\partial x_{j_1}}(a) & \cdots & \frac{\partial f_m}{\partial x_{j_m}}(a)
\end{vmatrix} \neq 0
$$

then $\{f_1, \ldots, f_m\}$ is functionally independent.

VIII.3.5.6. In fact, under the hypotheses of VIII.3.5.5., a stronger consequence holds: the Inverse Function Theorem (applied to a suitable $m$-dimensional affine subspace of $\mathbb{R}^n$) implies that the range of $f$ has nonempty interior (which clearly implies functional independence).

VIII.3.5.7. The converse of VIII.3.5.4. does not hold, and there are no reasonable sufficient conditions for functional dependence along these lines. If the conclusion of VIII.3.5.4. holds, or if $n < m$, then the Straightening Theorem says that generically $\{f_1, \ldots, f_m\}$ is *locally* functionally dependent.
VIII.3.6. Exercises

VIII.3.6.1. Define \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) by
\[
f(x, y) = \begin{cases} 
\frac{xy}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]
(a) Show that \( f \) is not continuous at \( 0 \).
(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) exist everywhere.
(c) Show that \( D_u f(0, 0) \) does not exist unless \( u = e_1 \) or \( u = e_2 \).

VIII.3.6.2. Define \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) by
\[
f(x, y) = \begin{cases} 
\sqrt{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]
(a) Show that \( f \) is continuous at \( 0 \).
(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) exist everywhere.
(c) Show that \( D_u f(0, 0) \) does not exist unless \( u = e_1 \) or \( u = e_2 \).
(d) Show that \( f \) is not differentiable at \( 0 \). [Use ().]

VIII.3.6.3. Define \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) by
\[
f(x, y) = \begin{cases} 
\frac{x^2y}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\
0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]
(a) Show that \( f \) is not continuous at \( 0 \). [Consider \( \lim_{t \to 0} f(t, t^2) \).] Then \( f \) is a fortiori not differentiable at \( 0 \).
(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) exist everywhere.
(c) Show that \( D_u f(x, y) \) exists for all \( u \) and for all \( (x, y) \) (and in particular for \( (x, y) = (0, 0) \)).

VIII.3.6.4. Define \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) by
\[
f(x, y) = (x^{1/3} + y^{1/3})^3.
\]
(a) Show that \( f \) is continuous everywhere, and in particular at \( 0 \).
(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) exist everywhere.
(c) Show that \( D_u f(0, 0) \) exists and equals \( f(u) \) for all \( u \).
(d) Show that \( f \) is not differentiable at \( 0 \). [Use ().]
VIII.3.6.5. Define \( f : \mathbb{R}^2 \to \mathbb{R} \) by

\[
f(x, y) = \begin{cases} 
  \frac{x^2 y}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\
  0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]

(a) Show that \( f \) is continuous everywhere, and in particular at \( 0 \).

(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) exist everywhere.

(c) Show that \( \frac{\partial f}{\partial u}(0, 0) \) exists for all \( u \).

(d) Show that \( f \) is not differentiable at \( 0 \).

VIII.3.6.6. Define \( f : \mathbb{R}^2 \to \mathbb{R} \) by

\[
f(x, y) = \begin{cases} 
  x^2 + y^2 & \text{if } x^2 + y^2 \text{ is rational} \\
  0 & \text{if } x^2 + y^2 \text{ is irrational}
\end{cases}
\]

(a) Show that \( f \) is continuous only at \( 0 \).

(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) only exist at \( 0 \).

(c) Show that \( f \) is differentiable at \( 0 \).

VIII.3.6.7. Define \( f : \mathbb{R}^2 \to \mathbb{R} \) by

\[
f(x, y) = \begin{cases} 
  (x^2 + y^2) \sin \frac{1}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\
  0 & \text{if } (x, y) = (0, 0)
\end{cases}
\]

(a) Show that \( f \) is differentiable everywhere, and in particular at \( 0 \).

(b) Show that \( \frac{\partial f}{\partial x} \) and \( \frac{\partial f}{\partial y} \) are not continuous at \( 0 \).

VIII.3.6.8. Modify the proof of Theorem VIII.3.2.1. to give the following precise version of Remark VIII.3.2.2.:

**Theorem.** Let \( U \) be an open subset of \( \mathbb{R}^n \), and \( f = (f_1, \ldots, f_m) : U \to \mathbb{R}^m \). Suppose that, for some \( r \), \( 1 \leq r < n \), and \( c = (a, b) \in U \), where \( a \in \mathbb{R}^r \) and \( b \in \mathbb{R}^{n-r} \), we have that

(1) The function \( g(x) = f(x, b) \) is differentiable at \( a \).

(2) All partial derivatives \( \frac{\partial f_i}{\partial x_j} \), \( 1 \leq i \leq n, r+1 \leq j \leq n \), exist on \( U \) and are continuous at \( c \).

Then \( f \) is differentiable at \( c \).
VIII.4. Higher-Order Derivatives and Critical Point Analysis

VIII.4.1. Higher-Order Differentiability

VIII.4.1.1. If $U$ is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ is differentiable on $U$ (i.e. at every point of $U$), the total derivative $df$ is a function from $U$ to the set $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a vector space over $\mathbb{R}$ of dimension $nm$ and can thus be identified with $\mathbb{R}^{nm}$ in a natural way (Exercise VIII.4.1.1). Alternatively, $Df$ is a function from $U$ to the $m \times n$ matrices $M_{m \times n}$, which can evidently be identified with $\mathbb{R}^{nm}$. Thus $df$ or, equivalently, $Df$, may be regarded as a function from $U$ to $\mathbb{R}^{nm}$.

VIII.4.1.2. Definition. If $U$ is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ is differentiable on $U$, then $f$ is second-order differentiable at $a \in U$ if $df$ (or, equivalently, $Df$) is differentiable at $a$. Write $d^2f(a)$ for $d(df)(a)$ and $D^2f(a)$ for $D(Df)(a)$.

VIII.4.1.3. We can go on. If $df$ ($Df$) is differentiable at $a$, then $D^2f(a)$ is an $nm \times n$ matrix, identifiable with $\mathbb{R}^{nm}$. Thus if $f$ is second-order differentiable on $U$ and $D^2f$ is differentiable at $a \in U$, then $f$ is third-order differentiable at $a$. Continuing, we can define what it means for $f$ to be $r$th order differentiable at a point or on the set $U$, for any $r \in \mathbb{N}$. We also say $f$ is $r$th order differentiable on $U$ if it is $r$th order differentiable on $U$ for every $r$. We can also write $d^rf(a)$ and $D^rf(a)$ if $f$ is differentiable to order $r$ at $a$; $D^rf(a)$ is an $n^{r-1} m \times n$ matrix, which can be identified with a point in $\mathbb{R}^{nm}$.

The coordinate functions of $D^rf$ are just the $r$th order partials of the coordinate functions of $f$ (Exercise VIII.4.4.1.). By VIII.3.1.2., we have:

VIII.4.1.4. Proposition. If $U$ is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$, then $f$ is $r$th order differentiable at $a \in U$ if and only if each partial derivative of $f$ of order $r-1$ is defined in a neighborhood of $a$ and is differentiable at $a$.

We thus have the following immediate corollary of VIII.3.1.13. and VIII.3.2.1.:

VIII.4.1.5. Corollary. Let $U$ be an open set in $\mathbb{R}^n$, and $f : U \to \mathbb{R}^m$, and $r \in \mathbb{N}$. If $f$ is $r$th order differentiable on $U$, then $f$ is $C^{r-1}$ on $U$. If $f$ is $C^r$ on $U$, then it is $r$th order differentiable on $U$. The function $f$ is $C^\infty$ on $U$ if and only if it is infinitely differentiable on $U$.

Note that an $r$th order differentiable function does not have to be $C^r$ (VIII.3.2.3.).

VIII.4.1.6. There is a little more structure to higher derivatives: $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ has more structure than $\mathbb{R}^{nm}$. If $f$ is second-order differentiable at $a$, then $d^2f(a)$ is actually a linear transformation from $\mathbb{R}^n$ to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, which can be thought of as a bilinear () function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^m$. Similarly, for any $r$, $d^rf(a)$ can be regarded as an $r$-multilinear function from $\mathbb{R}^n \times \cdots \times \mathbb{R}^n$ to $\mathbb{R}^m$. This point of view is important in formulating the multivariable version of Taylor’s Theorem (), for example. (The two points of view are actually equivalent, since bilinear functions from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^m$ correspond naturally to linear functions from $\mathbb{R}^n \otimes \mathbb{R}^n \cong \mathbb{R}^{n^2}$ to $\mathbb{R}^m$, etc.)
Equality of Mixed Partial Again

There is a variation of Theorem VIII.2.2. involving differentiability:

VIII.4.1.7. \textbf{Theorem.} Let $U$ be an open set in $\mathbb{R}^n$, $f: U \to \mathbb{R}^m$, $a \in U$. If $f$ is second-order differentiable at $a$, then, for $1 \leq i, j \leq n$,

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(a) = \frac{\partial^2 f}{\partial x_i \partial x_j}(a).$$

For the proof, which is similar to the proof of VIII.2.2. and of comparable difficulty, see [Pug15].

VIII.4.1.8. \textbf{Theorems VIII.2.2.2. and VIII.4.1.7.} are complementary: neither directly implies the other. (Clairaut’s Theorem VIII.2.2.4., the version of the equality of mixed partials used most frequently, is a special case of both VIII.2.2.2. and VIII.4.1.7., so the distinction between these results is mostly academic.) In VIII.4.1.7. there is no assumption of continuity for any second-order partial, so it might seem to be more general than VIII.2.2.2. However, the hypotheses of VIII.4.1.7. imply that \textit{all} second-order partials of $f$ exist at $a$ (and even more), which is not necessary for VIII.2.2.2., as the next example shows.

VIII.4.1.9. \textbf{Example.} Let $g: \mathbb{R} \to \mathbb{R}$ be a function which is differentiable everywhere, but does not have a second derivative anywhere; $g'$ may be continuous, i.e. $g$ is an antiderivative of a continuous but non-differentiable function, or it may be quite discontinuous. Define $f: \mathbb{R}^2 \to \mathbb{R}$ by $f(x, y) = g(x)$. Then $\frac{\partial f}{\partial x}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$, and $\frac{\partial^2 f}{\partial y^2}$ all exist everywhere, and the last four are identically 0, hence continuous, so the hypotheses (and conclusion!) of VIII.2.2.2. are satisfied everywhere. But $f$ is not second-order differentiable anywhere: in fact, $\frac{\partial^2 f}{\partial x \partial y}$ does not exist anywhere.

Theorem VIII.4.1.7. has an extension which is a true generalization of VIII.2.2.6.:

VIII.4.1.10. \textbf{Corollary.} Let $U$ be an open set in $\mathbb{R}^n$, and $f$ an $r$'th order differentiable function from $U$ to $\mathbb{R}^m$ ($r > 1$). Then any two mixed partials of $f$ of total order $\leq r$ in which each coordinate direction appears the same number of times are equal on $U$.

VIII.4.1.11. \textbf{From the point of view of VIII.4.1.6.} (cf. VIII.4.4.1.), VIII.4.1.7. has a nice rephrasing: the bilinear form $d^2 f(a)$, when defined, is \textit{symmetric}: $d f(a)(x, y) = d f(a)(y, x)$ for all $x, y \in \mathbb{R}^n$. Similarly, the $r$-multilinear map $d^r f(a)$, when defined, is symmetric in the sense that it is invariant under any permutation of the arguments.

VIII.4.2. \textbf{Taylor’s Theorem}

VIII.4.3. \textbf{Critical Point Analysis}

VIII.4.4. \textbf{Exercises}
(a) Let \( f : U \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m \) be second-order differentiable at \( a \in U \). Show that

\[
d^2 f(a)(e_j)(e_k) = \frac{\partial^2 f}{\partial x_j \partial x_k}(a)
\]

for any \( j,k \). Thus the second-order partials of \( f \) are the natural coordinates for the identification of \( d^2 f \) regarded as a function from \( U \) to \( \mathbb{R}^{nm} \).

(b) More generally, if \( f \) is \( r \)'th order differentiable at \( a \), and \( d^r f(a) \) is regarded as an \( r \)-multilinear map from \( \mathbb{R}^n \times \cdots \times \mathbb{R}^n \) to \( \mathbb{R}^m \), then for any \( i_1, \ldots, i_r \) we have

\[
d^r f(a)(e_{i_1}, \ldots, e_{i_r}) = \frac{\partial^r f}{\partial x_{i_1} \cdots \partial x_{i_r}}(a) .
\]

(c) Give a similar description of \( D^r f(a) \).

VIII.5. Vector Analysis

VIII.5.1. Surfaces in \( \mathbb{R}^3 \)

VIII.5.2. Divergence and Curl
VIII.6. The Implicit and Inverse Function Theorems

The Implicit and Inverse Function Theorems are the deepest results from differential calculus, and among the most important; they are the primary means of passing from the *infinitesimal* to the *local* in analysis and geometry. Although they are distinct results, each is a relatively easy corollary of the other, so one only needs to carry out the (rather difficult) proof of one of them to obtain both (and in fact the two theorems have a common generalization, cf. VIII.6.2.1.). They can both be regarded as aspects of the general problem of solving systems of functional equations; existence and uniqueness theorems for solutions to differential equations fall into the same class of results (see IX.1.9.10.).

These results are stated in less generality than they have to be in many references, with proofs more difficult than they have to be (and at that not even always entirely complete or correct). There are several standard proofs, and some of these proofs arguably give better insight into why the theorems are true than the simplest ones. We will give proofs of each theorem which illustrate two common approaches; each approach gives versions not obtainable from the others.

These theorems have a long and complicated history, going back at least to DESCARTES (although for a long time it was not always appreciated that they had to be proved), and there are many versions. We will give only a few. The Smooth (C¹) Implicit Function Theorem (VIII.6.1.7.) was first proved by Dini in the 1870’s, although Dini’s contributions were not properly recognized for a long time since he published them only in his lecture notes, which had a limited circulation. For continuous and higher-order smooth versions, see [Gou03] and [You09]. The book [KP02a] contains the most thorough treatment available of the various forms of the Implicit and Inverse Function Theorems.

VIII.6.1. Statements of the Theorems

VIII.6.1.1. The idea of the theorems is that if one has n equations in m unknowns, then there is “usually” a unique solution if n = m, many solutions if n < m, and no solutions if n > m; if n < m, it should be possible to solve for n of the variables as functions of the others (specifying values for m−n of the variables yields a system of n equations in the remaining n variables). These statements are literally true if the equations are linear (cf. ()) unless there is degeneracy in the system. Since smooth functions are “approximately linear”, it might also be possible (and in fact is) to do the same locally in the smooth case.

VIII.6.1.2. Vector notation is a convenient way to rephrase systems of equations. An equation in unknowns x₁, . . . , xₘ can, by moving everything to one side of the equation, be written as $F(x₁, . . . , xₘ) = 0$, where $F : \mathbb{R}^m \to \mathbb{R}$ is a function. A set of n equations

$$F_1(x₁, . . . , xₘ) = 0$$
$$F_2(x₁, . . . , xₘ) = 0$$
$$\ldots$$
$$F_n(x₁, . . . , xₘ) = 0$$

can be written as $F(x₁, . . . , xₘ) = 0$, or $F(x) = 0$, where $F$ is the function from $\mathbb{R}^m$ (or a subset) to $\mathbb{R}^n$ with coordinate functions $(F₁, . . . , Fₙ)$, and $x = (x₁, . . . , xₘ)$.  

828
In the case \( m > n \), it is convenient to change notation slightly by setting \( p = m - n \). We would like to be able to solve the system \( \mathbf{F}(\mathbf{x}) = \mathbf{0} \) for \( n \) of the variables in terms of the other \( p \) variables. (Here the relative sizes of \( n \) and \( p \) do not matter; we could have \( n < p \), \( n = p \), or \( n > p \).) We can in general choose any \( n \) of the variables to solve for in terms of the other ones; we will notationally choose the last \( n \) and rename them \( y_1, \ldots, y_n \), i.e. our equation to solve is of the form

\[
\mathbf{F}(x_1, \ldots, x_p, y_1, \ldots, y_n) = \mathbf{0}
\]

or \( \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \), where \( \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p \) and \( \mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n \). A solution on a subset \( W \) of \( \mathbb{R}^p \) is a function \( f : W \to \mathbb{R}^n \) such that the pair \( (\mathbf{x}, \mathbf{y}) \) for \( \mathbf{y} = f(\mathbf{x}) \) satisfies the equation for all \( \mathbf{x} \in W \), i.e. \( \mathbf{F}(\mathbf{x}, f(\mathbf{x})) = \mathbf{0} \) for all \( \mathbf{x} \in W \). We say the function \( f \) is implicitly defined by the equation \( \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \). It is rarely possible to find an explicit formula for an implicitly defined function in the nonlinear case.

Examples with \( n = p = 1 \) from elementary calculus (e.g. the one in VIII.6.1.10.) show that one can only expect the equation \( \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \) to define \( \mathbf{y} \) as a continuous function of \( \mathbf{x} \) locally, i.e. in some neighborhood of a point \( \mathbf{a} \) where \((\mathbf{a}, \mathbf{b})\) satisfies the equation, and then only around points where there is some nondegeneracy condition on \( \mathbf{F} \). If \( n = p = 1 \) and \( \mathbf{F} \) is smooth, the graph of the equation \( \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \) is a curve in \( \mathbb{R}^2 \), and this curve is in general only locally the graph of a smooth function, and not even necessarily around every point of the curve; points where the curve crosses itself or has a vertical tangent must be excluded (see Exercise VIII.6.9.2.).

In the statements of the theorems, we will use the following notation. Let \( \mathbf{F} \) be a function from an open set \( U \) in \( \mathbb{R}^{p+n} \) to \( \mathbb{R}^n \). Write the coordinates of \( \mathbb{R}^{p+n} \) as \( (x_1, \ldots, x_p, y_1, \ldots, y_n) \), or \( (\mathbf{x}, \mathbf{y}) \) with \( \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p \) and \( \mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n \) as above, and let \( F_1, \ldots, F_n \) be the coordinate functions of \( \mathbf{F} \). So, in coordinates,

\[
\mathbf{F}(\mathbf{x}, \mathbf{y}) = (F_1(\mathbf{x}, \mathbf{y}), \ldots, F_n(\mathbf{x}, \mathbf{y})) = (F_1(x_1, \ldots, x_p, y_1, \ldots, y_n), \ldots, F_n(x_1, \ldots, x_p, y_1, \ldots, y_n))
\]

Let \( (\mathbf{x}, \mathbf{y}) \in U \), and suppose all partial derivatives of the form \( \frac{\partial F_i}{\partial y_j}(\mathbf{x}, \mathbf{y}) \) exist. Write

\[
D_y \mathbf{F}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix}
\frac{\partial F_1}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_1}{\partial y_n}(\mathbf{x}, \mathbf{y}) \\
\vdots & \ddots & \vdots \\
\frac{\partial F_n}{\partial y_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_n}{\partial y_n}(\mathbf{x}, \mathbf{y})
\end{bmatrix}
\]

and, if all partial derivatives of the form \( \frac{\partial F_i}{\partial x_j}(\mathbf{x}, \mathbf{y}) \) exist, write

\[
D_x \mathbf{F}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix}
\frac{\partial F_1}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_1}{\partial x_p}(\mathbf{x}, \mathbf{y}) \\
\vdots & \ddots & \vdots \\
\frac{\partial F_n}{\partial x_1}(\mathbf{x}, \mathbf{y}) & \cdots & \frac{\partial F_n}{\partial x_p}(\mathbf{x}, \mathbf{y})
\end{bmatrix}
\]

(note that \( D_y \mathbf{F}(\mathbf{x}, \mathbf{y}) \) is \( n \times n \) and \( D_x \mathbf{F}(\mathbf{x}, \mathbf{y}) \) is \( n \times p \) if they are defined). Then if \( \mathbf{F} \) is differentiable at \( (\mathbf{x}, \mathbf{y}) \), we have that the \( n \times (p + n) \) matrix \( D \mathbf{F}(\mathbf{x}, \mathbf{y}) \) partitions as

\[
D \mathbf{F}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix}
\vdots & D_y \mathbf{F}(\mathbf{x}, \mathbf{y}) & \vdots \\
D_x \mathbf{F}(\mathbf{x}, \mathbf{y}) & \vdots & \vdots
\end{bmatrix}
\]

829
VIII.6.1.6. Theorem. [Implicit Function Theorem, Continuous Version] Let $U$ be an open set in $\mathbb{R}^{p+n}$ and $(a, b) \in U$. Let $F : U \to \mathbb{R}^n$ with $F(a, b) = 0$. Suppose $F$ is continuous on $U$ and that $D_y F$ exists and is continuous on $U$, i.e. that $\frac{\partial F}{\partial y}$ exists and is continuous on $U$ for $1 \leq j \leq n$. If $D_y F(a, b)$ is invertible, i.e. $\det(D_y F(a, b)) \neq 0$, then there are:

(i) An open set $V$ in $\mathbb{R}^{p+n}$ containing $(a, b)$, with $V \subseteq U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^{p}$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}^n$ such that $f(a) = b$ and, for all $x \in W$, $(x, f(x)) \in V$ and

$$F(x, f(x)) = 0.$$ 

Additionally, $f$ is continuous on $W$.

VIII.6.1.7. Theorem. [Implicit Function Theorem, Smooth Version] Let $U$ be an open set in $\mathbb{R}^{p+n}$ and $(a, b) \in U$. Let $F : U \to \mathbb{R}^n$ with $F(a, b) = 0$. Suppose $F$ is $C^r$ on $U$ for some $r$, $1 \leq r \leq \infty$. If $D_y F(a, b)$ is invertible, i.e. $\det(D_y F(a, b)) \neq 0$, then there are:

(i) An open set $V$ in $\mathbb{R}^{p+n}$ containing $(a, b)$, with $V \subseteq U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^{p}$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}^n$ such that $f(a) = b$ and, for all $x \in W$, $(x, f(x)) \in V$ and

$$F(x, f(x)) = 0.$$ 

Additionally, $f$ is $C^r$ on $W$, and for any $x \in W$, we have

$$Df(x) = -D_y F(x, f(x))^{-1}D_x F(x, f(x))$$

(in particular, $D_y F(x, f(x))$ is invertible for all $x \in W$).

VIII.6.1.8. Note that the difference between the continuous version and the $C^1$ version (smooth version with $r = 1$) is merely in the differentiability of $F$ in the $x$-coordinates; continuity of $F$ and continuous differentiability in the $y$-coordinates are necessary in both cases (there is a version with this weakened: see VIII.6.3.9. and VIII.6.6.7.). To obtain continuous differentiability of higher order for the implicitly defined function, higher-order continuous differentiability of $F$ in both $x$- and $y$-coordinates is needed.

VIII.6.1.9. The uniqueness of the implicitly defined function $f$ precisely means that there is an open neighborhood $V'$ of $(a, b)$ (specifically, $V' = \{(x, y) \in V : x \in W\}$) such that the part of the level set $S = \{(x, y) \in U : F(x, y) = 0\}$ in $V'$ is exactly the graph of $f$, i.e. $S$ is locally the graph of a (continuous or smooth) function.

The use of $0$ in the statements is arbitrary: it could be replaced by any constant $c \in \mathbb{R}^n$ by applying the theorems to $F = F - c$. 830
VIII.6.1.10. In the statement of the theorems, one would expect the neighborhood $V$ to be contained in $U$ (this of course does not preclude the possibility that the conclusion could hold on a larger $V$). To show the necessity in general of passing to a smaller neighborhood $V$ of $(a, b)$ as well as the (expected) passage to a small neighborhood $W$ of $a$, consider the basic example of implicit differentiation every calculus class begins with: the case $p = n = 1$ and

$$F(x, y) = x^2 + y^2 - 25.$$ 

The set \{(x, y) : F(x, y) = 0\} is, of course, the circle of radius 5 around the origin in $\mathbb{R}^2$. Let $(a, b) = (3, 4)$. In this case the set $U$ could be taken to be all of $\mathbb{R}^2$, but no matter how small an open interval $W$ around $a = 3$ in $\mathbb{R}$ is taken, there is not a unique function $f$ on $W$ giving a solution, since we could have $f(x) = \pm \sqrt{25 - x^2}$ with the sign chosen arbitrarily for each $x$ (so long as $f(3) = 4$). There is a unique \textit{continuous solution} for $W$ small, but to obtain a unique \textit{solution}, one must cut down to a smaller neighborhood $V$ of $(3, 4)$. Actually $V$ does not have to be all that small in this case: it only needs not to contain any points on the lower half-circle with $x$-coordinate close to 3.

In this case, if $f$ is the unique solution on an interval $W$ (which must be contained in the interval $(-5, 5)$), then by the theorem

$$f'(x) = -\frac{\partial F}{\partial x}(x, f(x)) = -\frac{2x}{2f(x)} = -\frac{x}{f(x)},$$

for any $x \in W$. Here we can find an explicit formula for $f$: $f(x) = \sqrt{25 - x^2}$, and verify the formula for $f'$ directly. In most examples it is impossible to find an explicit formula for the implicitly defined function.

VIII.6.1.11. To obtain a reasonable Implicit Function Theorem, the function $F$ must be continuous. This in itself is not enough to guarantee an implicitly defined function, even a nonunique one, even if $p = n = 1$ (Exercise VIII.6.9.4.). But if there is a (unique) implicitly defined function, it is automatically continuous.

VIII.6.1.12. \textbf{Proposition.} Let $F$ be a continuous function on a subset $U$ of $\mathbb{R}^{p+n}$, with values in $\mathbb{R}^n$, and $K$ a compact subset of $U$. Suppose there is an open set $W$ in $\mathbb{R}^p$ on which there is a unique function $f : W \to \mathbb{R}^n$ implicitly defined by the equation $F(x, y) = 0$ with respect to $K$, i.e. for each $x \in W$ there is a unique $y \in \mathbb{R}^n$ such that $(x, y) \in K$ and $F(x, y) = 0$. Then $f$ is continuous.

\textbf{Proof:} Let $(x_k)$ be a sequence in $W$ with $x_k \to a \in W$. Suppose $f(x_k) \not= f(a)$. Then by compactness of $K$ there is a subsequence $(x_{k_j})$ such that $(f(x_{k_j}))$ converges to a point $y \in \mathbb{R}^n$, $y \not= f(a)$, with $(a, y) \in K$. But then $0 = F(x_{k_j}, f(x_{k_j})) \to F(a, y)$ by continuity of $F$, so $F(a, y) = F(a, f(a)) = 0$, contradicting the uniqueness assumption on $F$ and $K$.

VIII.6.1.13. The Implicit Function Theorem is strictly a local result, i.e. the conclusion just holds on some neighborhood $W$ of $a$. In some situations it is possible to determine an explicit $W$ on which the conclusion holds (cf. IX.1.9.10.).

\textbf{The Inverse Function Theorem}

VIII.6.1.14. If $f$ is a function defined on an open subset $U$ of $\mathbb{R}^n$, with values in $\mathbb{R}^m$, we would like to know when $f$ has an inverse function with the same continuity or differentiability properties as $f$. An
obvious necessary condition is that \( f \) be one-to-one (injective) on \( U \), and to discuss differentiability we also need that \( f(U) \) is open in \( \mathbb{R}^m \).

The Invariance of Domain Theorem \( (\ref{VIII.6.1.16}) \) gives a complete answer to the continuous version of the question: if \( f \) is continuous and one-to-one, then \( f(U) \) is open if and only if \( n = m \), and in this case \( f^{-1} \) is automatically continuous on \( f(U) \).

The answer to the differentiability question is more elementary. The next result does not depend on the Invariance of Domain Theorem:

\textbf{VIII.6.1.15. Proposition.} Let \( U \) be an open set in \( \mathbb{R}^n \), \( a \in U \), \( V \) an open set in \( \mathbb{R}^m \), and \( f \) a one-to-one function from \( U \) onto \( V \). Set \( b = f(a) \). Suppose \( f \) is differentiable at \( a \). Then \( f^{-1} : V \to U \) is differentiable at \( b \) if and only if \( m = n \) and \( Df(a) \) is invertible. If these conditions hold, then \( D(f^{-1})(b) = [Df(a)]^{-1} \).

\textbf{Proof:} One direction is almost immediate. If \( f^{-1} \) is differentiable at \( b \), then by the Chain Rule we have
\[
I_n = D(f^{-1} \circ f)(a) = D(f^{-1})(b)Df(a) \\
I_m = D(f \circ f^{-1})(b) = Df(a)D(f^{-1})(b)
\]
and so \( m = n \), \( Df(a) \) is invertible, and \( D(f^{-1})(b) = Df(a)^{-1} \) by \( (\ref{VIII.6.1.16}) \).

For the converse direction, we need a lemma:

\textbf{VIII.6.1.16. Lemma.} Let \( U \) be an open set in \( \mathbb{R}^n \), \( a \in U \), and \( f \) a one-to-one function from \( U \) to \( \mathbb{R}^n \). Suppose \( f \) is differentiable at \( a \) and \( Df(a) \) is invertible. Then there are strictly positive constants \( c, d \) and a neighborhood \( W \) of \( a \) contained in \( U \), such that
\[
c\|x - a\| < \|f(x) - f(a)\| < d\|x - a\|
\]
for all \( x \in W \setminus \{a\} \). In fact, we may choose \( c \) to be any positive number less than \( \|Df(a)^{-1}\|^{-1} \) and \( d \) any number greater than \( \|Df(a)\| \).

\textbf{Proof:} Fix \( c, d \) with \( 0 < c < \|Df(a)^{-1}\|^{-1} \) and \( d > \|Df(a)\| \), and set
\[
\epsilon = \min \left( \|Df(a)^{-1}\|^{-1} - c, d - \|Df(a)\| \right) > 0.
\]
Since \( f \) is differentiable at \( a \), there is a \( \delta > 0 \) such that, whenever \( 0 < \|x - a\| < \delta \), \( x \in U \) and
\[
\left| \frac{\|f(x) - f(a)\|}{\|x - a\|} - \frac{\|Df(a)(x - a)\|}{\|x - a\|} \right| \leq \frac{\|f(x) - f(a) - Df(a)(x - a)\|}{\|x - a\|} < \epsilon
\]
\[
-\epsilon < \frac{\|f(x) - f(a)\|}{\|x - a\|} - \frac{\|Df(a)(x - a)\|}{\|x - a\|} < \epsilon.
\]
But we have \( \|Df(a)(x - a)\| \leq \|Df(a)\|\|x - a\| \) by definition of the operator norm, so
\[
\frac{\|f(x) - f(a)\|}{\|x - a\|} < \frac{\|Df(a)(x - a)\|}{\|x - a\|} + \epsilon \leq \|Df(a)\| + \epsilon \leq d
\]
and similarly
\[ \|x - a\| = \|Df(a)^{-1}(Df(a)(x - a))\| \leq \|Df(a)^{-1}\| \|Df(a)(x - a)\| \]
\[ c \leq \|Df(a)^{-1}\|^{-1} - \epsilon \leq \frac{\|Df(a)(x - a)\|}{\|x - a\|} - \epsilon < \frac{\|f(x) - f(a)\|}{\|x - a\|} \]
so we may take \( W \) to be the open ball of radius \( \delta \) around \( a \).

**VIII.6.1.17.** We now prove the other direction of VIII.6.1.15. Suppose \( m = n \) and \( Df(a) \) is invertible. Since \( f \) is a bijection between \( U \) and \( V \), for each \( x \in U \) there is a unique \( y \in V \) with \( y = f(x) \), and conversely for each \( y \in V \) there is a unique \( x \in U \) with \( x = f^{-1}(y) \). We will suppress the function names and just use \( x \) and \( y \) with the understanding that they are related in this manner.

We must show that for any \( \epsilon > 0 \), there is a \( \delta > 0 \) such that, whenever \( 0 < \|y - b\| < \delta \), \( y \in V \) and
\[ \frac{\|x - a - Df(a)^{-1}(y - b)\|}{\|y - b\|} < \epsilon . \]
By VIII.6.1.16., reducing \( U \) if necessary we may assume there are positive constants \( c \) and \( d \) such that
\[ c\|x - a\| < \|f(x) - f(a)\| < d\|x - a\| \]
\[ \frac{1}{d} < \frac{\|x - a\|}{\|y - b\|} < \frac{1}{c} \]
for all \( x \in U, x \neq a \). Since \( f \) is differentiable at \( a \), there is an \( \eta > 0 \) such that, whenever \( 0 < \|x - a\| < \eta \), \( x \in U \) and
\[ \frac{\|y - b - Df(a)(x - a)\|}{\|x - a\|} < \frac{\epsilon c}{\|Df(a)^{-1}\|} . \]
Since
\[ x - a - Df(a)^{-1}(y - b) = Df(a)^{-1}[Df(a)(x - a) - (y - b)] \]
we have, for \( \|y - b\| < c\eta, \|x - a\| < \eta \) and
\[ \frac{\|x - a - Df(a)^{-1}(y - b)\|}{\|y - b\|} \leq \frac{\|Df(a)^{-1}\| \|Df(a)(x - a) - (y - b)\|}{\|y - b\|} \]
\[ = \frac{\|x - a\| \|Df(a)^{-1}\| \|y - b - Df(a)(x - a)\|}{\|x - a\| \|y - b\|} \leq \frac{1}{c} \|Df(a)^{-1}\| \frac{\epsilon c}{\|Df(a)^{-1}\|} = \epsilon \]
so we may take \( \delta = c\eta \). \( \diamondsuit \)

**VIII.6.1.18.** Corollary. Let \( U \) be an open set in \( \mathbb{R}^n \), \( V \) an open set in \( \mathbb{R}^m \), and \( f \) a one-to-one function from \( U \) onto \( V \). If \( f \) is differentiable at every point of \( U \), then \( f^{-1} : V \to U \) is differentiable at every point of \( V \) if and only if \( m = n \) and, for every \( x \in U \), \( Df(x) \) is invertible. If these conditions hold, then \( D(f^{-1})(y) \) is invertible for every \( y \in V \) and \( D(f^{-1})(y) = Df(f^{-1}(y))^{-1} \).

Actually, if \( f : U \to V \) is one-to-one, onto, and differentiable, then \( m \) must equal \( n \) even without the assumption that \( Df(x) \) is invertible for every \( x \in U \), by Invariance of Domain; but we do not need to use this fact for the Corollary.
VIII.6.1.19. There are two principal drawbacks to these results, however. First, the Invariance of Domain is a deep theorem requiring a considerable excursion into topology to prove. And second, even for the differentiable result which is more elementary, it must be verified that the function is locally one-to-one, with open range, which is often not easy to do directly. The most useful result would be one in which (local) injectivity can be deduced simply from nondegeneracy of the derivative. This is the principal content of the Inverse Function Theorem:

VIII.6.1.20. \textbf{Theorem. [Inverse Function Theorem]} Let $U$ be an open set in $\mathbb{R}^n$, $a \in U$, and $f : U \to \mathbb{R}^n$ a $C^r$ function for some $r, 1 \leq r \leq \infty$. If $Df(a)$ is invertible, then there is a neighborhood $V$ of $a$, $V \subseteq U$, such that

(i) $f$ is one-to-one on $V$.

(ii) $f(V)$ is open in $\mathbb{R}^n$.

(iii) $Df(x)$ is invertible for all $x \in V$.

(iv) $f^{-1}$ is $C^r$ on $f(V)$ and, for all $y \in f(V)$,

$$D(f^{-1})(y) = [Df(f^{-1}(y))]^{-1}.$$ 

VIII.6.1.21. If $Df(x)$ is not invertible everywhere on $U$, it will obviously be necessary to restrict to a smaller neighborhood $V$ of $a$ to get (iii) and (iv). But this may be necessary even if $Df(x)$ is invertible for all $x \in U$ to make $f$ injective.

If $n = 1$ and $I$ is an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$ is differentiable on $I$ with $f'(x) \neq 0$ for all $x \in I$, then $f$ is strictly monotone on $I$ (and thus one-to-one; so if $U = I$ we may take $V = U$). But if $n > 1$ invertibility of $Df$ on $U$ does not guarantee that $f$ is one-to-one on $U$, even if $U$ is connected. A simple and important example is the following function, which is the transformation from rectangular coordinates to polar coordinates in $\mathbb{R}^2$ (hence we use $(r, \theta)$ for the coordinates). Set $U = \{(r, \theta) : r > 0\}$, and define $f : U \to \mathbb{R}^2$ by

$$f(r, \theta) = (r \cos \theta, r \sin \theta).$$

Then $Df(r, \theta)$ is invertible everywhere on $U$; but $f$ is not one-to-one on all of $U$, only locally one-to-one.

VIII.6.2. \textbf{Twin Theorems and a Generalization}

The Inverse Function Theorem and Implicit Function Theorem (smooth version) are twin theorems in the sense that each is easily derivable from the other; the arguments are much easier than proving either theorem from scratch. In fact, either argument can be used to prove a common generalization:

VIII.6.2.1. \textbf{Theorem. [General Implicit/Inverse Function Theorem]} Let $n \in \mathbb{N}$, $p \in \mathbb{N} \cup \{0\}$, $a \in \mathbb{R}^p$, $b, c \in \mathbb{R}^n$, $U$ an open neighborhood of $(a, b)$ in $\mathbb{R}^{p+n}$, $F : U \to \mathbb{R}^n$ a $C^r$ function $(1 \leq r \leq \infty)$ with $F(a, b) = c$ and $D_y F(a, b)$ invertible. Then there are:

(i) An open neighborhood $V$ of $(a, b)$ in $\mathbb{R}^{p+n}$, contained in $U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;
(ii) An open neighborhood $W$ of $(a, c)$ in $\mathbb{R}^{p+n}$;

(iii) A unique function $G : W \to \mathbb{R}^n$ satisfying $G(a, c) = b$ and, for all $(x, z) \in W$, $(x, G(x, z)) \in V$ and

$$F(x, G(x, z)) = z.$$  

Additionally, $G$ is $C^r$ on $W$, and for all $(x, z) \in W$,

$$DG(x, z) = \begin{bmatrix} -D_y F(x, G(x, z))^{-1} D_x F(x, G(x, z)) & D_y F(x, G(x, z))^{-1} \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}.$$  

Note that the Inverse Function Theorem is the case $p = 0$ (so the $a$ is absent). The Smooth Implicit Function Theorem is a corollary by taking $c = 0$ and setting $f(x) = G(x, 0)$. In fact, the Smooth Implicit Function Theorem says that for certain fixed $z$ near $c$, there is a function $f_z$ implicitly defined near $a$ by $F(x, y) = z$, which is $C^r$ in $x$; we have $G(x, z) = f_z(x)$, and the additional content of the above theorem is that the $f_z$ are defined for all $z$ near $c$ and “vary smoothly” in $z$. Thus Theorem VIII.6.2.1. is a “parametrized version” of both the Implicit and Inverse Function Theorems: holding $x$ fixed gives the Inverse Function Theorem and holding $z$ fixed gives the (Smooth) Implicit Function Theorem.

The situation is symmetric in $F$ and $G$. $G$ satisfies the hypotheses of the theorem, so there is a $C^r$ function $H$ defined near $(a, b)$ for which $H(a, b) = c$ and $G(x, H(x, y)) = y$ for $(x, y)$ near $(a, b)$. By the formula, $DH(x, y) = DF(x, y)$ for all $(x, y)$ near $(a, b)$, and since $H(a, b) = F(a, b)$, we have $H = F$ on a neighborhood of $(a, b)$, i.e. $H = F$ (perhaps restricted to a smaller neighborhood of $(a, b)$). In particular, we have $G(x, F(x, y)) = y$ for $(x, y)$ near $(a, b)$ (this can be verified directly).

We will give simple proofs of VIII.6.2.1. from both the Inverse Function Theorem and the Smooth Implicit Function Theorem.

**Proof From Inverse Function Theorem**

**VIII.6.2.2.** Assume the Inverse Function Theorem VIII.6.1.20. We will deduce VIII.6.2.1.

Suppose $U \subseteq \mathbb{R}^{p+n}$ is open, $(a, b) \in U$, and $F : U \to \mathbb{R}^n$ is a $C^r$ function with $F(a, b) = c$ and $D_y F(a, b)$ invertible. Define

$$\tilde{F}(x, y) = (x, F(x, y))$$

for $(x, y) \in U$. Then $\tilde{F}$ is a $C^r$ function from $U$ to $\mathbb{R}^{p+n}$. We have

$$D \tilde{F}(x, y) = \begin{bmatrix} I_p & 0 \\ D_x F(x, y) & D_y F(x, y) \end{bmatrix}$$

for $(x, y) \in U$; thus $D \tilde{F}(a, b)$ is invertible ($\tilde{F}(a, b) = (a, c)$). We have $(a, c) \in W$ and $\Phi(a, c) = (a, b)$. 

835
Write $\Phi(x, z) = (H(x, z), G(x, z))$, where $H: W \to \mathbb{R}^p$ and $G: W \to \mathbb{R}^n$. Then $H$ and $G$ are $C^r$, and $G(a, c) = b$. Since for $(x, z) \in W$ we have
\[(x, z) = \tilde{F}(\Phi(x, z)) = \tilde{F}(H(x, z), G(x, z)) = (H(x, z), F(H(x, z), G(x, z)))\]
we obtain that $H(x, z) = x$ and
\[F(H(x, z), G(x, z)) = F(x, G(x, z)) = z\]
for $(x, z) \in W$.

Uniqueness of $G$ follows from the fact that $\tilde{F}$ is one-to-one: if $\tilde{G}$ is a function from $W$ to $\mathbb{R}^n$ with $(x, \tilde{G}(x, z)) \in V$ and $F(x, \tilde{G}(x, z)) = z$ for all $(x, z) \in W$, then for $(x, z) \in W$ we have
\[\tilde{F}(x, \tilde{G}(x, z)) = (x, z) = \tilde{F}(x, G(x, z))\]
so $\tilde{G}(x, z) = G(x, z)$.

Thus VIII.6.2.1. is proved (see Exercise VIII.6.9.5. for the derivative formula). \(\square\)

**Proof From Implicit Function Theorem**

**VIII.6.2.3.** Now assume the Smooth Implicit Function Theorem VIII.6.1.7. holds. We will deduce VIII.6.2.1.

Suppose $U \subseteq \mathbb{R}^{p+n}$ is open, $(a, b) \in U$, and $F: U \to \mathbb{R}^n$ is as in the statement of VIII.6.2.1.. Set $c = F(a, b)$. Write $\mathbb{R}^{p+2n} \cong \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n$, and set
\[U' = \{(x, z, y) \in \mathbb{R}^{p+2n} : (x, y) \in U\}.

Then $U'$ is open in $\mathbb{R}^{p+2n}$, and $(a, c, b) \in U'$. Define $\tilde{F}: U' \to \mathbb{R}^n$ by
\[\tilde{F}(x, z, y) = F(x, y) - z.\]

Then $\tilde{F}$ is $C^r$ on $U'$, $\tilde{F}(a, c, b) = 0$, and
\[D\tilde{F}(x, z, y) = [D_x F(x, y) - I_n \quad D_y F(x, y)]\]
for $(x, z, y) \in U'$, and in particular $D_y \tilde{F}(a, c, b) = D_y F(a, b)$ is invertible.

We can thus solve the equation $\tilde{F}(x, z, y) = 0$ implicitly for $y$ as a function of $(x, z)$, i.e. there are open neighborhoods $W$ of $(a, c)$ and $V$ of $(a, b)$ and a unique function $G : W \to \mathbb{R}^n$ such that $G(x, z) \in V$ and $F(x, z, G(x, z)) = 0$ for all $(x, z) \in W$, and $G$ is $C^r$. But $\tilde{F}(x, z, G(x, z)) = 0$ is equivalent to $F(x, G(x, z)) = z$.

Thus VIII.6.2.1. is proved (cf. Exercise VIII.6.9.5.). \(\square\)

**VIII.6.2.4.** There are versions of the Continuous and Mixed Implicit Function Theorems similar to VIII.6.2.1. (Exercise VIII.6.9.15.).

836
VIII.6.3. Proof of the Implicit Function Theorem, Case $n = 1$

It is much easier to prove the Implicit Function Theorem in the case where we are only solving for one variable in terms of the others. We also obtain the desired results under somewhat weaker hypotheses. It does not make any difference in the proof how many other variables there are, i.e. we will prove the case where $n = 1$ and $p$ is arbitrary.

VIII.6.3.1. The general setup for this case is a function $F : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Write $(x, y) = (x_1, \ldots, x_p, y)$ for the coordinates of a general point of $U$. We will need to assume at a minimum that $F$ is continuous on $U$ and that $\frac{\partial F}{\partial y}(a, b)$ exists and is nonzero. To obtain most of the conclusions we will need a little more. Of course, the conditions need only hold on some neighborhood of $(a, b)$, since for the conclusion we will have to restrict to a smaller neighborhood of $(a, b)$ anyway in general, i.e. the conclusion will be only a local one.

The Continuous Version

VIII.6.3.2. Theorem. [One-Dimensional Weak Implicit Function Theorem, Continuous Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Suppose $F$ is continuous on $U$ and $\frac{\partial F}{\partial y}$ exists and is nonzero at $(a, b)$. Then there are:

(i) An open set $V$ in $\mathbb{R}^{p+1}$ containing $(a, b)$, with $V \subseteq U$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

(iii) Functions $\overline{f}, \underline{f} : W \to \mathbb{R}$ such that $\overline{f}(a) = \underline{f}(a) = b$, $(x, \overline{f}(x))$ and $(x, \underline{f}(x))$ are in $V$ and $F(x, \overline{f}(x)) = F(x, \underline{f}(x)) = 0$

for all $x \in W$, and such that if $f$ is any function from $W$ to $\mathbb{R}$ with $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$, then $\underline{f} \leq f \leq \overline{f}$.

Additionally, $\overline{f}$ is upper semicontinuous and $\underline{f}$ is lower semicontinuous on $W$, and $\overline{f}$ and $\underline{f}$ are continuous at $a$.

Proof: By passing to a smaller neighborhood, we may assume $U$ is convex. Suppose $\frac{\partial F}{\partial y}(a, b) > 0$. Then there are $y_1 > b$ and $y_2 < b$ such that $(a, y_1)$ and $(a, y_2)$ are in $U$ and $\frac{F(a, z) - F(a, b)}{z - b} > 0$ for $y_2 \leq z \leq y_1$ and $z \neq b$ (in fact, this will be true for all such $y_1, y_2$ sufficiently close to $b$). It follows that $F(a, y_1) > 0$ and $F(a, y_2) < 0$ (and in fact, $F(a, z) > 0$ for $b < z \leq y_1$ and $F(a, z) < 0$ for $y_2 \leq z < b$). Then there is an open neighborhood $W$ of $a$ in $\mathbb{R}^p$ such that, for all $x \in W$, $(x, y_1)$ and $(x, y_2)$ are in $U$, $F(x, y_1) > 0$, and $F(x, y_2) < 0$. So for each fixed $x \in W$, there is at least one $y$ between $y_2$ and $y_1$ for which $F(x, y) = 0$ by the Intermediate Value Theorem, since $F$ is continuous on $U$. Among these, there is a largest $\overline{f}$ and a smallest $\underline{f}$. Set $\overline{f}(x) = y_1$ and $\underline{f}(x) = y_2$. This defines $\overline{f}$ and $\underline{f}$ on $W$. We obviously have $\overline{f}(a) = \underline{f}(a) = b$.

Set $V = W \times (y_2, y_1) \subseteq U$. If $f$ is any real-valued function on $W$ such that $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$, we obviously have $\underline{f}(x) \leq f(x) \leq \overline{f}(x)$ for all $x \in W$.

It remains to show that $\overline{f}$ and $\underline{f}$ are semicontinuous on $W$. We give the argument for $\overline{f}$; the one for $\underline{f}$ is essentially identical. Let $x_n \in \overline{W}$, $x_n \to x_0$. Passing to a subsequence we may assume $\overline{f}(x_n) \to z$ for some $z$ between $y_2$ and $y_1$. But then $0 = F(x_n, \overline{f}(x_n)) \to F(x_0, z)$ by continuity of $F$, so $F(x_0, z) = 0$ and
and \( F \) (contradicting the assumption that By continuity of Proof: Corollary VIII.6.3.5.

\[
\begin{align*}
\text{Proof: Let } F \text{ function Under the hypotheses of VIII.6.3.4. we have that An open set } V \text{ except the uniqueness of An open set } V = \text{A unique function } V \text{ with } z \text{ are continuous at } \text{An open set } V \text{ and } F. \text{Fix } x \in W. \text{Then there are:}
\end{align*}
\]

(i) An open set \( V \) in \( \mathbb{R}^{p+1} \) containing \((a, b)\), with \( V \subset U \);

(ii) An open set \( W \) in \( \mathbb{R}^p \) containing \( a \);

(iii) A unique function \( f : W \to \mathbb{R} \) such that \( f(a) = b \), and \((x, f(x)) \in V \) and \( F(x, f(x)) = 0 \) for all \( x \in W \).

Additionally, \( f \) is continuous on \( W \).

**Proof:** Everything follows immediately from VIII.6.3.2, except the uniqueness of \( f \), i.e. that \( \overline{F} = f \). Define \( V \) and \( W \) as in the proof of VIII.6.3.2. Fix \( x \in W \). If there were \( y \) and \( z \), \( y \neq z \), with \((x, y)\) and \((x, z)\) in \( V \) and \( F(x, y) = F(x, z) = 0 \), then by Rolle’s Theorem there would be a \( c \) between \( y \) and \( z \) with \( \frac{\partial F}{\partial y}(x, c) = 0 \), contradicting the assumption that \( \frac{\partial F}{\partial y} \) is never zero on \( U \). Thus there is a unique \( y \) such that \((x, y) \in V \) and \( F(x, y) = 0 \), and we may set \( f(x) = y \). This \( f \) is the unique real-valued function defined on \( W \) with \((x, f(x)) \in V \) and \( F(x, f(x)) = 0 \) for all \( x \in W \).

**VIII.6.3.5. Corollary.** Let \( F(x_1, \ldots, x_p, y) : U \to \mathbb{R}, \) where \( U \subset \mathbb{R}^{p+1} \) is open, and \((a, b) = (a_1, \ldots, a_p, b) \in U \) with \( F(a, b) = 0 \). Suppose \( F \) is continuous on \( U \) and \( \frac{\partial F}{\partial y} \) exists at all points of \( U \), and is continuous at \((a, b)\), with \( \frac{\partial F}{\partial y}(a, b) \neq 0 \). Then there are:

(i) An open set \( V \) in \( \mathbb{R}^{p+1} \) containing \((a, b)\), with \( V \subset U \);

(ii) An open set \( W \) in \( \mathbb{R}^p \) containing \( a \);

(iii) A unique function \( f : W \to \mathbb{R} \) such that \( f(a) = b \), and \((x, f(x)) \in V \) and \( F(x, f(x)) = 0 \) for all \( x \in W \).

Additionally, \( f \) is continuous on \( W \).

**Proof:** By continuity of \( \frac{\partial F}{\partial y} \) at \((a, b)\), there is a neighborhood \( U' \) of \((a, b)\) contained in \( U \) such that \( \frac{\partial F}{\partial y}(x, y) \neq 0 \) for all \((x, y) \in U' \). Apply VIII.6.3.4, with \( U' \) in place of \( U \).
A Pictorial Version

VIII.6.3.6. We illustrate the argument in the case \( p = n = 1 \) (the picture represents the case of general \( p \) if the horizontal axis is regarded as multidimensional). Suppose \( U \) is an open set in \( \mathbb{R}^2 \) large enough to include the region pictured in the diagrams (which may actually be very small, and the proportions may vary), and \( F \) is a continuous function from \( U \) to \( \mathbb{R} \) such that \( \frac{\partial F}{\partial y} \) exists and is nonzero on \( U \), and let \( (a, b) \in U \) with \( F(a, b) = 0 \). We assume that \( \frac{\partial F}{\partial y}(a, b) > 0 \); the proof in the case \( \frac{\partial F}{\partial y}(a, b) < 0 \) is identical with red and blue reversed. Because of the shapes in the diagrams, this proof can be called the “T proof.”

Since \( F(a, b) = 0 \) and \( \frac{\partial F}{\partial y}(a, b) > 0 \), \( g(y) = F(a, y) \) is increasing at \( b \) (V.3.6.8.), i.e. there are \( y_1 < b \) and \( y_2 > b \) such that \( F(a, y) > F(a, b) = 0 \) for \( b < y \leq y_2 \), and \( F(a, y) < 0 \) for \( y_1 \leq y < b \), i.e. \( F(x, y) > 0 \) for all points \((x, y)\) in red in Figure VIII.4 and \( F(x, y) < 0 \) for all \((x, y)\) in blue:

![Figure VIII.4: Implicit Function Theorem I](image)

We may assume \( y_2 - b = b - y_1 \) if desired by taking a minimum; although this is not necessary for the proof, it will be convenient later since this can be called \( \epsilon \).
Since $F(a, y_2) > 0$ and $F$ is continuous, there are $x_1 < a$ and $x_2 > a$ such that $F(x, y_2) > 0$ for $x_1 \leq x \leq x_2$. Similarly, since $F(a, y_1) < 0$, there are $x_3 < a$ and $x_4 > a$ with $F(x, y_1) < 0$ for $x_3 \leq x \leq x_4$. As before, we may assume $x_3 = x_1$ and $x_4 = x_2$ and that $x_2 - a = a - x_1$ by taking a minimum; this can be called $\delta$. Thus all red points $(x, y)$ in Figure VIII.5 have $F(x, y) > 0$, and all blue points have $F(x, y) < 0$:

![Diagram of red and blue points with $F(x, y)$ values]

Figure VIII.5: Implicit Function Theorem II

In the case of general $p$, the top and bottom horizontal segments become balls in $\mathbb{R}^p$ centered at $a$, which can be taken to have the same radius which we can call $\delta$. 
Now fix $x_0$, $x_1 \leq x_0 \leq x_2$. We have that $F(x_0, y_1) < 0$ and $F(x_0, y_2) > 0$. Since $F(x_0, y)$ is a continuous function of $y$, there must be a $y$, $y_1 < y < y_2$, with $F(x_0, y) = 0$. If there were more than one such $y$, by Rolle’s Theorem there would be a $y$ in between with $\frac{\partial F}{\partial y}(x_0, y) = 0$, contradicting that $\frac{\partial F}{\partial y}$ is nonzero on $U$. Thus there is a unique $y$ between $y_1$ and $y_2$ with $F(x_0, y) = 0$, and we call this $y \ f(x_0)$. This procedure uniquely defines a function $f$ on the interval $[x_1, x_2]$ with values in $(y_1, y_2)$, which satisfies $F(x, f(x)) = 0$ for $x_1 \leq x \leq x_2$. We also have $F(x, y) > 0$ for $f(x) < y \leq y_2$ and $F(x, y) < 0$ if $y_1 \leq y < f(x)$, for any $x \in [x_1, x_2]$. See Figure VIII.6.

Figure VIII.6: Implicit Function Theorem III
To show $f$ is continuous, fix $x_0$, $x_1 < x_0 < x_2$. Let $\epsilon > 0$. We may assume $\epsilon \leq \min(y_2 - f(x_0), f(x_0) - y_1)$ without loss of generality. We repeat the first part of the proof replacing $a$ and $b$ by $x_0$ and $f(x_0)$, and $y_1$ and $y_2$ by $f(x_0) \pm \epsilon$: $F(x_0, y) > 0$ for $f(x_0) < y \leq f(x_0) + \epsilon$ and $F(x_0, y) < 0$ for $f(x_0) - \epsilon \leq y < f(x_0)$, so there is a $\delta, 0 < \delta \leq \min(x_2 - x_0, x_0 - x_1)$, such that $F(x, y) > 0$ for all red points in Figure VIII.7 and $F(x, y) < 0$ for all blue points.

![Figure VIII.7: Implicit Function Theorem IV](image)

If $|x - x_0| < \delta$, there is thus a unique $y$, $f(x_0) - \epsilon < y < f(x_0) + \epsilon$, with $F(x, y) = 0$; this $y$ must be $f(x)$, i.e. $|f(x) - f(x_0)| < \epsilon$. Thus $f$ is continuous at $x_0$. 

842
The Differentiable Version

The next lemma contains the most delicate part of the entire proof of the Implicit Function Theorem. It is needed for the differentiable and smooth versions of the theorem if $n = 1$, and even for the continuous version for general $n$. If the first-order partials of $F$ are continuous at $(a, b)$, the proof can be simplified (Exercise VIII.6.9.8).

VIII.6.3.7. Lemma. Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Suppose $F$ is continuous on $U$ and differentiable at $(a, b)$, and that $\frac{\partial F}{\partial y}(a, b) \neq 0$. Then there are:

(i) An open set $V$ in $\mathbb{R}^{p+1}$ containing $(a, b)$, with $V \subseteq U$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

such that whenever $f$ is a real-valued function on $W$ with $f(a) = b$ and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$ (there exist such functions by VIII.6.3.2.), then $f$ is differentiable at $a$ and, for $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(a) = -\frac{\frac{\partial F}{\partial x_j}(a, b)}{\frac{\partial F}{\partial y}(a, b)}.$$

Proof: The little-o notation $()$ is very helpful in this proof. As $h = (\Delta x, \Delta y) = (\Delta x_1, \ldots, \Delta x_p, \Delta y) \to 0$, we have

$$\Delta F = F(a + \Delta x, b + \Delta y) - F(a, b) = \frac{\partial F}{\partial x_1}(a, b)\Delta x_1 + \cdots + \frac{\partial F}{\partial x_p}(a, b)\Delta x_p + o(1)\|\Delta x\| + \frac{\partial F}{\partial y}(a, b)\Delta y + o(1)\Delta y$$

and if we set $y = f(x)$, so $\Delta y = f(a + \Delta x) - f(a)$, we have $\Delta F = 0$ since $F(a + \Delta x, b + \Delta y) = F(a, b) = 0$, i.e.

$$\Delta y \left[\frac{\partial F}{\partial y}(a, b) + o(1)\right] = -\frac{\partial F}{\partial x_1}(a, b)\Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(a, b)\Delta x_p + o(1)\|\Delta x\|.$$

By VIII.6.3.2., we have $f \leq f \leq \overline{f}$, and hence by the Squeeze Theorem $f$ is continuous at $a$; thus $\Delta y = o(1)$ as $\Delta x \to 0$, and hence $\overline{h} = o(1)$ as $\Delta x \to 0$. Thus we have that the $o(1)$’s in the last expression are still $o(1)$ as $\Delta x \to 0$. Also, if $\Delta x$ is sufficiently small, we can divide by the expression in brackets on the left (using that $\frac{\partial F}{\partial y}(a, b) \neq 0$ by assumption) to obtain

$$\Delta y = \frac{-\frac{\partial F}{\partial x_1}(a, b)\Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(a, b)\Delta x_p + o(1)\|\Delta x\|}{\frac{\partial F}{\partial y}(a, b) + o(1)}$$

as $\Delta x \to 0$. By $()$, the last expression is

$$\Delta y = \|\Delta x\|\left[\frac{-\frac{\partial F}{\partial x_1}(a, b)\Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(a, b)\Delta x_p + o(1)}{\frac{\partial F}{\partial y}(a, b) + o(1)}\right]$$

$$= \|\Delta x\|\left[\frac{-\frac{\partial F}{\partial x_1}(a, b)\Delta x_1}{\frac{\partial F}{\partial y}(a, b)\|\Delta x\|} - \cdots - \frac{\partial F}{\partial x_p}(a, b)\Delta x_p}{\frac{\partial F}{\partial y}(a, b)\|\Delta x\|} + o(1)\right]$$

843
\[
= - \frac{\partial F}{\partial x_1}(a, b) \Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(a, b) \Delta x_p + o(1)\|\Delta x\|
\]
as \(\Delta x \to 0\); thus \(f\) is differentiable at \(a\) and

\[
\frac{\partial f}{\partial x_j}(a) = - \frac{\partial F}{\partial x_j}(a, b) \cdot \frac{\partial F}{\partial y}(a, b).
\]

The following technical strengthening will be used in the inductive proof of the general Implicit Function Theorem. It can be proved in the same way, or it follows immediately from VIII.6.3.7.

VIII.6.3.8. **Lemma.** Let \(F(x_1, \ldots, x_p, y) : U \to \mathbb{R}\), where \(U \subseteq \mathbb{R}^{p+1}\) is open, and \((a, b) = (a_1, \ldots, a_p, b) \in U\) with \(F(a, b) = 0\). Let \(S = \{x_{k_1}, \ldots, x_{k_m}\}\) be a subset of \(\{x_1, \ldots, x_p\}\). Suppose \(F\) is continuous on \(U\) and differentiable with respect to \(S \cup \{y\}\) at \((a, b)\), and that \(\frac{\partial F}{\partial y}(a, b) \neq 0\). Then there are:

(i) An open set \(V\) in \(\mathbb{R}^{p+1}\) containing \((a, b)\), with \(V \subseteq U\);

(ii) An open set \(W\) in \(\mathbb{R}^p\) containing \(a\);

such that whenever \(f\) is a real-valued function on \(W\) with \(f(a) = b\) and \((x, f(x)) \in V\) and \(F(x, f(x)) = 0\) for all \(x \in W\) (there exist such functions by VIII.6.3.2.), then \(f\) is differentiable with respect to \(S\) at \(a\) and, for \(x_j \in S\),

\[
\frac{\partial f}{\partial x_j}(a) = - \frac{\partial F}{\partial x_j}(a, b) \cdot \frac{\partial F}{\partial y}(a, b).
\]

VIII.6.3.9. **Theorem.** [One-Dimensional Implicit Function Theorem, Differentiable Version] Let \(F(x_1, \ldots, x_p, y) : U \to \mathbb{R}\), where \(U \subseteq \mathbb{R}^{p+1}\) is open, and \((a, b) = (a_1, \ldots, a_p, b) \in U\) with \(F(a, b) = 0\). Suppose \(F\) is differentiable at all points of \(U\) and \(\frac{\partial F}{\partial y}\) is nonzero at all points of \(U\). Then there are:

(i) An open set \(V\) in \(\mathbb{R}^{p+1}\) containing \((a, b)\), with \(V \subseteq U\);

(ii) An open set \(W\) in \(\mathbb{R}^p\) containing \(a\);

(iii) A unique function \(f : W \to \mathbb{R}\) such that \(f(a) = b\), and \((x, f(x)) \in V\) and \(F(x, f(x)) = 0\) for all \(x \in W\).

Additionally, \(f\) is differentiable at every point of \(W\) and, for all \(x \in W\) and \(1 \leq j \leq p\),

\[
\frac{\partial f}{\partial x_j}(x) = - \frac{\partial F}{\partial x_j}(x, f(x)) \cdot \frac{\partial F}{\partial y}(x, f(x)).
\]
Proof: The existence of $V, W, \text{and } f$, and the uniqueness (and continuity) of $f$, follow immediately from VIII.6.3.4., and if $x_0 \in V$, differentiability of $f$ at $x_0$ follows from VIII.6.3.7. applied at the point $(x_0, f(x_0))$. 

In the case $p = 1$, we get the version of the Implicit Function Theorem which justifies implicit differentiation in first-semester calculus (note that there is no truly “one-dimensional” version of the Implicit Function Theorem – one-variable implicit differentiation must be justified by the “two-dimensional” version):

**VIII.6.3.10. Corollary. [One-Variable Implicit Function Theorem, Differentiable Version]**

Let $F(x, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^2$ is open, and $(a, b) \in U$ with $F(a, b) = 0$. Suppose $F$ is differentiable at all points of $U$ and $\frac{\partial F}{\partial y}$ is nonzero at all points of $U$. Then there are:

(i) An open set $V$ in $\mathbb{R}^2$ containing $(a, b)$, with $V \subseteq U$;

(ii) An open set $W$ in $\mathbb{R}$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, $f$ is differentiable at every point of $W$ and, for all $x \in W$,

$$f'(x) = -\frac{\frac{\partial F}{\partial x}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))}.$$ 

**VIII.6.3.11.** The fact that the slope of the tangent line to the level curve at $(x, y)$ is $-\frac{\frac{\partial F}{\partial x}(x, y)}{\frac{\partial F}{\partial y}(x, y)}$ should come as no surprise, since this level curve should be perpendicular to $\nabla F(x, y) = \left(\frac{\partial F}{\partial x}(x, y), \frac{\partial F}{\partial y}(x, y)\right)$.

But what is not obvious is that the level curve through $(x, y)$ is a smooth curve with a tangent line, or indeed that the level set is a “curve” at all; this is the theoretical conclusion of the theorem (cf. VIII.6.1.9.).

The formula for $f'(x)$ given in the theorem is equivalent to but more efficient than the one obtained by first-semester calculus methods before partial derivatives are introduced (VIII.6.9.7.); indeed, the partial derivative version is the “right” way to do one-variable implicit differentiation.

We also get a mixed version in which differentiability with respect to only some of the variables is assumed and concluded:

**VIII.6.3.12. Theorem. [One-Dimensional Implicit Function Theorem, Mixed Version]**

Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Let $S = \{x_{k_1}, \ldots, x_{k_p}\}$ be a subset of $\{x_1, \ldots, x_p\}$. Suppose $F$ is differentiable with respect to $S \cup \{y\}$ at all points of $U$ and $\frac{\partial F}{\partial y}$ is nonzero at all points of $U$. Then there are:

(i) An open set $V$ in $\mathbb{R}^{p+1}$ containing $(a, b)$, with $V \subseteq U$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;
(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, $f$ is continuous on $W$ and differentiable with respect to $S$ at every point of $W$ and, for all $x \in W$ and $x_j \in S$,

$$\frac{\partial f}{\partial x_j}(x) = -\frac{\frac{\partial F}{\partial x_j}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))}.$$

The continuous version (VIII.6.5.) and the differentiable version (VIII.6.9.) are the special cases $S = \emptyset$ and $S = \{x_1, \ldots, x_p\}$ respectively.

**The Smooth Version**

**Theorem.** [One-Dimensional Implicit Function Theorem, Smooth Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Suppose $F$ is $C^r$ on $U$ ($1 \leq r \leq \infty$) and $\frac{\partial F}{\partial y}(a, b) \neq 0$. Then there are:

(i) An open set $V$ in $\mathbb{R}^{p+1}$ containing $(a, b)$, with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(x, y) \neq 0$ for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, $f$ is $C^r$ on $W$ and, for all $x \in W$ and $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(x) = -\frac{\frac{\partial F}{\partial x_j}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))}.$$

**Proof:** By continuity, there is an open neighborhood $U'$ of $(a, b)$ contained in $U$ such that $\frac{\partial F}{\partial y}$ is nonzero everywhere on $U'$. Everything then follows immediately from VIII.6.9. (applied with $U'$ in place of $U$) except the fact that $f$ is $C^r$. We prove this by induction on $r$.

First suppose $F$ is $C^1$. Then, for $1 \leq j \leq p$, $\frac{\partial F}{\partial x_j}$ is continuous on $V$, and $f$ is continuous on $W$, so $g_j(x) = \frac{\partial F}{\partial x_j}(x, f(x))$ is continuous on $W$. Similarly, $g(x) = \frac{\partial F}{\partial y}(x, f(x))$ is continuous and nonzero on $W$, so

$$\frac{\partial f}{\partial x_j}(x) = -\frac{g_j(x)}{g(x)}$$

is also continuous on $W$ for each $j$. Thus $f$ is $C^1$ on $W$.

Now assume the result for $r$, and suppose $F$ is $C^{r+1}$. Then by the inductive hypothesis $f$ is $C^r$, so $g$ and each $g_j$ are $C^r$ by the higher-order Chain Rule (1). We then obtain that $\frac{\partial F}{\partial x_j}$ is $C^r$ for each $j$ since it is a quotient of $C^r$ functions with nonvanishing denominator. Thus $f$ is $C^{r+1}$.
**VIII.6.3.14.** Corollary. [One-Variable Implicit Function Theorem, Smooth Version] Let $F(x, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^2$ is open, and $(a, b) \in U$ with $F(a, b) = 0$. Suppose $F$ is $C^r$ on $U$ ($1 \leq r \leq \infty$) and $\frac{\partial F}{\partial y}(a, b) \neq 0$. Then there are:

(i) An open set $V$ in $\mathbb{R}^2$ containing $(a, b)$, with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(x, y) \neq 0$ for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, $f$ is $C^r$ on $W$ and, for all $x \in W$,

$$f'(x) = -\frac{\frac{\partial F}{\partial x}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))}.$$

We also get a mixed version which will be used in the inductive proof of the general theorem:

**VIII.6.3.15.** Theorem. [One-Dimensional Implicit Function Theorem, Mixed Smooth Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(a, b) = (a_1, \ldots, a_p, b) \in U$ with $F(a, b) = 0$. Let $S = \{x_{k_1}, \ldots, x_{k_m}\}$ be a subset of $\{x_1, \ldots, x_p\}$. Suppose $F$, $\frac{\partial F}{\partial y}$, and $\frac{\partial F}{\partial x_j}$ are continuous on $U$ ($1 \leq r \leq \infty$) for all $x_j \in S$, and $\frac{\partial F}{\partial y}(a, b) \neq 0$. Then there are:

(i) An open set $V$ in $\mathbb{R}^{p+1}$ containing $(a, b)$, with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(x, y) \neq 0$ for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, $f$ and $\frac{\partial f}{\partial x_j}$ are continuous on $W$ for all $x_j \in S$ and, for all $x \in W$, $x_j \in S$, and $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(x) = \frac{\frac{\partial F}{\partial y}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))}.$$

**Proof:** The conclusion follows from VIII.6.3.12. just as VIII.6.3.13. follows from VIII.6.3.9.. 

**VIII.6.4.** Proof of the General Implicit Function Theorem

**VIII.6.4.1.** The most straightforward proof of the general Implicit Function Theorem is by induction on the number $n$ of variables being solved for (this was also the first proof). The inductive proof requires a stronger continuity hypothesis on the partial derivatives of $F$ than the one-dimensional version, as explained in VIII.6.4.3.

We will prove the following mixed version of the Implicit Function Theorem, which includes both the continuous version (VIII.6.1.6.) and the smooth ($C^1$) version (VIII.6.1.7.):
Theorem. [Implicit Function Theorem, Mixed Version] Let $U$ be an open set in $\mathbb{R}^{p+n}$ and $(a, b) \in U$. Let $F : U \to \mathbb{R}^n$ with $F(a, b) = 0$. Suppose $F$ is continuous on $U$ and that $D_y F$ exists and is continuous on $U$ for $1 \leq j \leq n$. If $D_y F(a, b)$ is invertible, i.e. $\det(D_y F(a, b)) \neq 0$, then there are:

(i) An open set $V$ in $\mathbb{R}^{p+n}$ containing $(a, b)$, with $V \subseteq U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}^n$ such that $f(a) = b$ and, for all $x \in W$, $(x, f(x)) \in V$ and $F(x, f(x)) = 0$.

Additionally, $f$ is continuous on $W$. If additionally $S = \{x_k_1, \ldots, x_{k_m}\}$ is a subset of $\{x_1, \ldots, x_p\}$ and $\frac{\partial F}{\partial x_{k_i}}$ exists and is continuous on $U$ for each $x_{k_i} \in S$, then $\frac{\partial F}{\partial x_{k_i}}$ exists and is continuous on $W$ for each $x_{k_i} \in S$ and these partials satisfy

$$D_S f(x) = -D_y F(x, f(x))^{-1}D_S F(x, f(x))$$

for all $x \in W$, where $D_S F(x, y)$ is the $n \times m$ matrix

$$D_S F(x, y) = \begin{bmatrix}
\frac{\partial F_1}{\partial x_{k_1}}(x, y) & \cdots & \frac{\partial F_n}{\partial x_{k_1}}(x, y) \\
\vdots & \ddots & \vdots \\
\frac{\partial F_1}{\partial x_{k_m}}(x, y) & \cdots & \frac{\partial F_n}{\partial x_{k_m}}(x, y)
\end{bmatrix}.$$

Proof: The case $n = 1$ is VIII.6.3.15. We now assume the theorem is true for $n$, and show it is true for $n+1$. So suppose $F(x, y) = F(x_1, \ldots, x_p, y_1, \ldots, y_{n+1})$ is a continuous function from an open set $U \subseteq \mathbb{R}^{p+n+1}$ to $\mathbb{R}^{n+1}$, $(a, b) = (a_1, \ldots, a_p, b_1, \ldots, b_{n+1}) \in U$, that $D_y F$ and $D_S F$ are continuous on $U$ for some subset $S$ of $\{x_1, \ldots, x_p\}$, and $D_y F(a, b)$ is invertible. The strategy for finding the implicit function $f : \mathbb{R}^p \to \mathbb{R}^{n+1}$ will be to show that for some $i$, $1 \leq i \leq n+1$, the $y_k$ for $2 \leq k \leq n+1$ can be written as implicit functions $g_k$ of $(x_1, \ldots, x_p, y_1)$ satisfying

$$F_j(x_1, \ldots, x_p, y_1, g_2(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1)) = 0$$

for $j \neq i$ and $(x_1, \ldots, x_p, y_1)$ close to $(a_1, \ldots, a_p, b_1)$ by the inductive hypothesis, and that $y_1$ can then be written as an implicit function $h$ of $(x_1, \ldots, x_p)$ satisfying

$$F_i(x_1, \ldots, x_p, h(x), g_2(x, h(x)), \ldots, g_{n+1}(x, h(x))) = 0$$

for $x$ near $a$ by an application of VIII.6.3.15. Then $f$ can be defined by

$$f(x) = (h(x), g_2(x, h(x)), \ldots, g_{n+1}(x, h(x))).$$

For the details, we will use the following notation. Write $J(x, y)$ for the determinant of the $(n+1) \times (n+1)$ matrix $D_y F(x, y)$. For each $i$, let $J_i(x, y)$ be the determinant of the $i$’th minor in the first column, i.e.

$$J_i(x, y) = (-1)^{i+1} \det \begin{bmatrix}
\frac{\partial F_1}{\partial y_2}(x, y) & \cdots & \frac{\partial F_1}{\partial y_{n+1}}(x, y) \\
\vdots & \ddots & \vdots \\
\frac{\partial F_{n+1}}{\partial y_2}(x, y) & \cdots & \frac{\partial F_{n+1}}{\partial y_{n+1}}(x, y)
\end{bmatrix}.$$
with the $i$'th row (partials of $F_i$) deleted. Then

$$J(x, y) = J_1(x, y) \frac{\partial F_1}{\partial y_1}(x, y) + J_2(x, y) \frac{\partial F_2}{\partial y_1}(x, y) + \cdots + J_{n+1}(x, y) \frac{\partial F_{n+1}}{\partial y_1}(x, y)$$

and so, since $J(a, b) \neq 0$, there is an $i$, $1 \leq i \leq n+1$, such that $J_i(a, b) \neq 0$ and $\frac{\partial F_i}{\partial y_1}(a, b) \neq 0$. Then there is an open neighborhood $U'$ of $(a, b)$ such that $J_i(x, y) \neq 0$ and $\frac{\partial F_i}{\partial y_1}(x, y) \neq 0$ for all $(x, y) \in U'$. Apply the inductive hypothesis with $S' = S \cup \{y_1\}$ to obtain an open neighborhood $V_1$ of $(a, b)$ in $\mathbb{R}^{p+n+1}$, an open neighborhood $W_1$ of $(a_1, \ldots, a_p, b_1)$ in $\mathbb{R}^{p+1}$, and unique functions $g_k : W_1 \to \mathbb{R} \ (2 \leq k \leq n+1)$ which satisfy, for $(x_1, \ldots, x_p, y_1) \in W_1$,

$$(x_1, \ldots, x_p, y_1, g_1(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1)) \in V_1$$

$$F_j(x_1, \ldots, x_p, y_1, g_1(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1)) = 0$$

for $j \neq i$, and $g_k, \frac{\partial g_k}{\partial y_1}$, and $\frac{\partial g_j}{\partial x_j}$ for $x_j \in S$ continuous on $W_1$ for each $k$.

Now, for $(x_1, \ldots, x_p, y_1) \in W_1$, set

$$G(x_1, \ldots, x_p, y_1) = F_i(x_1, \ldots, x_p, y_1, g_1(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1)).$$

Then $G$ is continuous on $W_1$ and $G(a_1, \ldots, a_p, b_1) = 0$. We show that $G$ satisfies the hypotheses of VIII.6.3.15. on $W_1$. By the Chain Rule, $\frac{\partial G}{\partial y_1}$ exists on $W_1$ and

$$\frac{\partial G}{\partial y_1}(x, y_1) = \frac{\partial F}{\partial y_1}(x, y_1, g_1(x, y_1), \ldots, g_{n+1}(x, y_1)) + \frac{\partial F_i}{\partial y_1}(x, y_1, g_2(x, y_1), \ldots, g_{n+1}(x, y_1)) \frac{\partial g_2}{\partial y_1}(x, y_1) + \cdots + \frac{\partial F_i}{\partial y_1}(x, y_1, g_2(x, y_1), \ldots, g_{n+1}(x, y_1)) \frac{\partial g_{n+1}}{\partial y_1}(x, y_1).$$

Thus $\frac{\partial G}{\partial y_1}$ is continuous on $W_1$. It remains to show that $\frac{\partial G}{\partial y_1}(a_1, b_1) \neq 0$. If we set $G_i = G$ and define, for $j \neq i$,

$$G_j(x_1, \ldots, x_p, y_1) = F_j(x_1, \ldots, x_p, y_1, g_1(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1))$$

we similarly have

$$\frac{\partial G_j}{\partial y_1}(x, y_1) = \frac{\partial F}{\partial y_1}(x, y_1, g_1(x, y_1), \ldots, g_{n+1}(x, y_1)) + \frac{\partial F_i}{\partial y_1}(x, y_1, g_2(x, y_1), \ldots, g_{n+1}(x, y_1)) \frac{\partial g_2}{\partial y_1}(x, y_1) + \cdots + \frac{\partial F_i}{\partial y_1}(x, y_1, g_2(x, y_1), \ldots, g_{n+1}(x, y_1)) \frac{\partial g_{n+1}}{\partial y_1}(x, y_1).$$

But $G_j$ is identically 0, so abbreviating notation we have

$$\frac{\partial G_i}{\partial y_1} = \frac{\partial F_i}{\partial y_1} + \frac{\partial F_i}{\partial y_2} \frac{\partial g_2}{\partial y_1} + \cdots + \frac{\partial F_i}{\partial y_{n+1}} \frac{\partial g_{n+1}}{\partial y_1} = 0$$

849
for \( j \neq i \). Multiplying the \( j \)’th equation by \( J_j \) and adding the equations together, we get

\[
J_i \frac{\partial G_i}{\partial y_1} = \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_1} + \frac{\partial g_2}{\partial y_2} \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_2} + \cdots + \frac{\partial g_{n+1}}{\partial y_{n+1}} \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_{n+1}}.
\]

The first sum is \( J \), and the second sum is the determinant of the matrix

\[
\begin{bmatrix}
\frac{\partial F_1}{\partial y_2} & \frac{\partial F_1}{\partial y_2} & \cdots & \frac{\partial F_1}{\partial y_{n+1}} \\
\frac{\partial F_2}{\partial y_2} & \frac{\partial F_2}{\partial y_2} & \cdots & \frac{\partial F_2}{\partial y_{n+1}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial F_{n+1}}{\partial y_2} & \frac{\partial F_{n+1}}{\partial y_2} & \cdots & \frac{\partial F_{n+1}}{\partial y_{n+1}}
\end{bmatrix}
\]

expanded by minors on the first column. This matrix has two identical columns, hence has determinant 0 ()), so the second sum is 0. Similarly, the other sums are 0. Thus

\[
J_i(x, y) \frac{\partial G}{\partial y_1}(x, y_1) = J(x, y)
\]

for all \((x, y)\); so \( \frac{\partial G}{\partial y_1}(a, b_1) \neq 0 \) since \( J(a, b) \neq 0 \).

Then by VIII.6.3.15, there is an open set \( W \) in \( \mathbb{R}^p \) containing \( a \) and an open set \( V \) in \( \mathbb{R}^{p+1} \) containing \((a, b_1), V_2 \subseteq W_1, \) and a unique function \( h : W \to \mathbb{R} \), such that \((x, h(x)) \in V_2 \) and \( G(x, h(x)) = 0 \) for all \( x \in W \). Set

\[
V = \{(x_1, \ldots, x_p, y_1, \ldots, y_n) \in V_1 : (x_1, \ldots, x_p, y_1) \in V_2\}
\]

and define

\[
f(x) = (h(x), g_2(x, h(x)), \ldots, g_{n+1}(x, h(x)))
\]

for \( x \in W \). Then, for \( x \in W \), we have \((x, f(x)) \in V \), and \( F_i(x, f(x)) = G(x, h(x)) = 0 \). And if \( j \neq i \),

\[
F_j(x, f(x)) = F_j(x, h(x), g_2(x, h(x)), \ldots, g_{n+1}(x, h(x))) = 0.
\]

Thus \( F(x, f(x)) = 0 \).

The uniqueness of \( f \) follows immediately from the uniqueness of the \( g_k \) and of \( h \), and continuity of \( f \) follows from continuity of the \( g_k \) and \( h \). Continuity of \( \frac{\partial f}{\partial x_j} \) on \( W \) for \( x_j \in S \) then follows from continuity of the partials of the \( g_k \) and \( h \) for \( x_j \in S \).

The formula for \( D_S f \) now follows easily from the Chain Rule. We write the case \( S = \{x_1, \ldots, x_p\} \); the case of general \( S \) is identical with notational complications (or can be obtained immediately by restricting coordinates). Define \( H : W \to \mathbb{R}^{p+n} \) by \( H(x) = (x, f(x)) \). Then \( H \) is \( C^1 \) on \( W \), and

\[
DH(x) = \begin{bmatrix} I_p \\ Df(x) \end{bmatrix}
\]

for \( x \in W \), where \( I_p \) is the \( p \times p \) identity matrix and \( Df(x) \) is \( n \times p \). Then by the Chain Rule we have

\[
D(F \circ H)(x) = D(F(H(x)))DH(x) = \begin{bmatrix} D_x F(x, f(x)) & D_y F(x, f(x)) \end{bmatrix} \begin{bmatrix} I_p \\ Df(x) \end{bmatrix} = D_x F(x, f(x)) + D_y F(x, f(x))Df(x)
\]

850
for \( x \in W \). But \( F \circ H \) is identically zero on \( W \), so \( D(F \circ H)(x) = 0 \) for all \( x \in W \). Since \( D_y F(x, f(x)) \) is invertible for \( x \in W \), we have, for \( x \in W \),
\[
D_y F(x, f(x))^{-1} D_x F(x, f(x)) + Df(x) = 0.
\]

This completes the proof of VIII.6.4.2., and hence of the continuous version VIII.6.1.6. (the case \( S = \emptyset \)) and the \( C^1 \) smooth version VIII.6.1.7. (the case \( S = \{x_1, \ldots, x_p\} \)).

** VIII.6.4.3. ** The reason this proof requires continuity of the partials, and not just that the partials exist and \( J(x, y) \neq 0 \) in a neighborhood of \( (a, b) \), is that not only must \( J \) be nonzero, but at least one of the minors \( J_i \) must also be nonzero in a neighborhood of \( (a, b) \) to apply the inductive assumption. The continuity can be replaced by an assumption that not only is the \( J \) nonzero in a neighborhood of \( (a, b) \), but also a decreasing chain of minors of smaller and smaller size are nonzero in a neighborhood of \( (a, b) \). This version was explicitly stated in [You09], but is quite clumsy so we have stuck to the cleaner statement of VIII.6.4.2., which is adequate for almost all applications. For a full extension of VIII.6.3.10. to the \( n \)-dimensional case, see VIII.6.6.7.. It does not appear to be known whether the assumption of the decreasing sequence of nonzero minors is automatic (cf. Exercise VIII.6.9.13.(b)).

** VIII.6.4.4. ** To finish the proof of the smooth version, it remains only to show that if \( F \) is \( C^r \), then \( f \) is also \( C^r \). The entries of \( D_x F(x, f(x)) \) and \( D_y F(x, f(x)) \) are \( C^{r-1} \) by assumption. Since the determinant of \( D_y F(x, f(x)) \) is a polynomial in the entries, it is also \( C^{r-1} \) on \( W \), and since the determinant is never 0 on \( W \), it follows that the reciprocal of the determinant is also \( C^{r-1} \) on \( W \). Thus the entries of \( D_y F(x, f(x))^{-1} \) are also \( C^{r-1} \) on \( W \) by the formula in III.11.7.45., and therefore so are the entries of
\[
Df(x) = -D_y F(x, f(x))^{-1} D_x F(x, f(x))
\]
which is equivalent to \( f \) being \( C^r \) on \( W \).

This completes the full proof of the smooth version VIII.6.1.7..

** VIII.6.5. ** **Proof of the Inverse Function Theorem**

The Inverse Function Theorem can now be proved from the Implicit Function Theorem as in ( ). In this section we give an alternate direct proof, which “explains” better why the theorem is true and which generalizes to other settings, such as Banach spaces ( ). This proof uses the Contraction Mapping Principle ( ). This argument first appeared, in somewhat disguised form, in [Gou03]; cf. Exercise VIII.6.9.10. .

To prove the Inverse Function Theorem, it suffices the prove the following theorem. Differentiability of the inverse function then follows from VIII.6.1.15., and the formula there shows that the inverse function is \( C^1 \); and if \( f \) is \( C^r \), the inverse function is also \( C^r \) by the argument in VIII.6.4.4..
VIII.6.5.1.  **Theorem.** Let $U$ be an open set in $\mathbb{R}^n$, $a \in \mathbb{R}^n$, and $f : U \to \mathbb{R}^n$ a $C^1$ function. If $Df(a)$ is invertible, then there is a neighborhood $V$ of $a$, $V \subseteq U$, such that

(i) $f$ is one-to-one on $V$.

(ii) $W = f(V)$ is open in $\mathbb{R}^n$.

(iii) $Df(x)$ is invertible for all $x \in V$.

VIII.6.5.2.  The first reduction is to make an affine change of variables to make $a = 0$, $f(0) = 0$, and $Df(0) = I_n$, which greatly simplifies the notation of the rest of the proof. We define

$$T(x) = Df(a)^{-1}(x).$$

Then $T$ is an invertible linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$; $T$ is $C^1$ and $DT(x) = Df(a)^{-1}$ for all $x$. Define

$$g(x) = T(f(x + a)) - c$$

for $x \in U - a$, where $c = T(f(a))$; then by the Chain Rule $g$ is $C^1$ on $U - a$ and

$$Dg(x) = DT(f(x + a))Df(x + a) = Df(a)^{-1}Df(x + a)$$

for $x \in U - a$. In particular, $Dg(0) = I_n$.

It suffices to prove VIII.6.5.1. for $g$ on $U - a$, since we would then have

$$f^{-1}(y) = g^{-1}(T(y) - c) + a$$

on $T^{-1}(W + c)$ with values in $V + a$.

So, changing notation, in VIII.6.5.1. we will assume $a = 0$, $f(0) = 0$, and $Df(0) = I_n$.

VIII.6.5.3.  Now define $g(x) = x - f(x)$. Then $g$ is $C^r$. Since $Dg(0) = 0$, there is an $\epsilon > 0$ such that $\|Dg_k(x)\| < \frac{1}{2\epsilon}$ for all $k$ whenever $\|x\| < \epsilon$. Set $W_1 = B_\epsilon(0)$. If $x, z \in W_1$, then by the MVT (1) for each $k$ there is a $c_k$ between $x$ and $z$ with $g_k(x) - g_k(z) = Dg_k(c_k)(x - z)$; $c_k \in W_1$ since $W_1$ is convex. Thus

$$\|g(x) - g(z)\| \leq \sum_{k=1}^n |g_k(x) - g_k(z)| = \sum_{k=1}^n |Dg_k(c_k)(x - z)| \leq \sum_{k=1}^n \|Dg_k(c_k)\| \|x - z\| < \frac{1}{2} \|x - z\|$$

for $x, z \in W_1$. In particular, $\|g(x)\| < \frac{1}{2}\|x\|$ for all $x \in W_1$.

VIII.6.5.4.  Now fix $\eta, 0 < \eta < \epsilon$, and fix $y \in \mathbb{R}^n$ with $\|y\| \leq \frac{\eta}{\epsilon}$. Define $g_y : \mathbb{R}^n \to \mathbb{R}^n$ by

$$g_y(x) = y + x - f(x) = y + g(x).$$

If $\|x\| \leq \eta$, then $\|g(x)\| < \frac{\eta}{\epsilon}$, so $\|g_y(x)\| < \eta$; thus $g_y$ maps $B_\eta(0)$ into itself. If $x, z \in B_\eta(0)$, then

$$\|g_y(x) - g_y(z)\| = \|g(x) - g(z)\| < \frac{1}{2} \|x - z\|$$

852
and so $g_y$ is a contractive mapping of the complete metric space $\tilde{B}_\eta(0)$ to itself. By the Contractive Mapping Principle (), $g_y$ has a unique fixed point $x \in \bar{B}_\eta(0)$, i.e. there is a unique $x$ with $\|x\| \leq \eta$ and $f(x) = y$. Thus, if $W = B_{\epsilon/2}(0)$, for any $y \in W$ there is a unique $x \in B_\epsilon(0)$ with $f(x) = y$ (such a $y$ is in $\tilde{B}_{\eta/2}(0)$ for $2\|y\| \leq \eta < \epsilon$). Set $V = B_\epsilon(0) \cap f^{-1}(W)$; then $V$ is an open set in $\mathbb{R}^n$ containing 0 and $f$ maps $V$ bijectively to $W$.

This completes the proof of VIII.6.5.1.

VIII.6.5.5. Of course, this argument is not a full proof of the Inverse Function Theorem; the arguments of VIII.6.1.15. and VIII.6.4.4. are also needed to finish the proof.

VIII.6.6. How Much Can the Theorems be Generalized?

VIII.6.6.1. One might hope that the hypotheses of the Inverse Function Theorem can be relaxed to only require differentiability at $a$, or at least nondegeneracy of the derivative at $a$. But this is not enough to force the function to be locally injective around $a$, even in one dimension:

VIII.6.6.2. Example. Let

$$f(x) = \begin{cases} x + x^2 \sin \frac{1}{x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

for $x \in \mathbb{R}$. Then $f$ is differentiable on $\mathbb{R}$ and $f'(0) = 1$, but $f$ is not injective (strictly increasing) on any neighborhood of 0. In fact, $f'$ takes all real number values (in particular, both positive and negative values) in any neighborhood of 0.

VIII.6.6.3. Suppose we assume $f$ is injective in an open neighborhood $V$ of $a$, with $f(V)$ open, and that $f$ is differentiable at $a$ and $Df(a)$ is invertible. Then there is an inverse function $g$ defined on $f(V)$. By VIII.6.1.15., $g$ is differentiable at $f(a)$.

VIII.6.6.4. However, one of the main points of the Inverse Function Theorem is that nondegeneracy of the derivative of $f$ at $a$ forces $f$ to be injective in a neighborhood $V$ of $a$, with $f(V)$ open. Continuity of the partial derivatives is seemingly used mainly to insure that the derivative is nondegenerate in an entire neighborhood of $a$ (this can fail if the derivative is discontinuous; cf. VIII.6.6.2.). So what if this is just assumed? In the one-dimensional case, if the derivative of a function is nonzero in an open interval around $a$, then $f$ must be strictly monotone in that interval by () and hence locally injective. Remarkably, this remains true in higher dimensions:

VIII.6.6.5. Theorem. Let $U$ be an open set in $\mathbb{R}^n$, $a \in U$, and $f : U \rightarrow \mathbb{R}^n$. Suppose $f$ is differentiable at every point of $U$ and that $Df(x)$ is invertible for all $x \in U$. Then there is a neighborhood $V$ of $a$ such that $f$ is one-to-one on $V$, $f(V)$ is open in $\mathbb{R}^n$, and the inverse function $g : f(V) \rightarrow V$ is differentiable at every point of $f(V)$ with

$$Dg(f(x)) = (Df(x))^{-1}$$

for all $x \in V$. 

853
VIII.6.6.6. Curiously, this theorem appears to have been first stated and proved by J. Saint Raymond in 2002 [Ray02], although it is implicit in an earlier result of Chernavskii in 1964 [Čer64]; see also [Hal74], [Cri92], and http://terrytao.wordpress.com/2011/09/12/the-inverse-function-theorem-for-everywhere-differentiable-maps/. It should be pointed out, however, that as a practical matter this result is of limited use beyond the case covered in the usual Inverse Function Theorem if \( n > 1 \), since in most cases the only reasonable way to verify that a function is differentiable is to show that the partial derivatives are continuous (cf. ()), although there are cases where the partial derivatives are continuous except at one point \( a \) and it can be shown by brute force that the function is differentiable at \( a \) (cf. ()), and this version but not the usual version can be applied in such a situation (provided, of course, that the nondegeneracy condition is satisfied).

There is a corresponding version of the Implicit Function Theorem, proved from VIII.6.6.5. the same way VIII.6.1.7. is deduced from VIII.6.1.20. () The one-dimensional version of this result is VIII.6.3.10.. I am not aware of any analogous version of the Continuous or Mixed Implicit Function Theorems; cf. Exercises VIII.6.9.12. and VIII.6.9.13.(b). There is also a corresponding version of VIII.6.2.1.

VIII.6.6.7. **Theorem.** Let \( U \) be an open set in \( \mathbb{R}^{p+n} \) and \((a, b) \in U \). Let \( F : U \to \mathbb{R}^n \) with \( F(a, b) = 0 \). Suppose \( F \) is differentiable at every point of \( U \). If \( D_y F(x, y) \) is invertible, i.e. \( \det(D_y F(x, y)) \neq 0 \), for all \((x, y) \in U \), then there is an open set \( V \) in \( \mathbb{R}^{p+n} \) containing \((a, b) \), an open set \( W \) in \( \mathbb{R}^p \) containing \( a \), and a unique function \( f : W \to \mathbb{R}^n \) such that \( f(a) = b \) and, for all \( x \in W \), \((x, f(x)) \in V \) and \( F(x, f(x)) = 0 \). Additionally, \( f \) is differentiable at every point of \( W \), and for any \( x \in W \), we have

\[
Df(x) = -D_y F(x, f(x))^{-1} D_x F(x, f(x)).
\]

VIII.6.6.8. It should be noted that the results of this section are technically not “generalizations” of the usual Implicit and Inverse Function Theorems – they are distinct results (but obviously closely related) since both the hypotheses and conclusions are weaker. A true generalization of a theorem is a new result for which the original theorem is a special case, generally consisting of obtaining the same, or stronger, conclusions under weaker hypotheses (here “weaker” means hypotheses which hold in actually or potentially more situations); obtaining stronger conclusions under the same hypotheses could also be considered a type of generalization, although such a result is usually just called a stronger theorem and not a generalization. Thus VIII.6.3.12. is a true generalization of both VIII.6.3.4. and VIII.6.3.9., and VIII.6.4.2. a generalization of VIII.6.1.6. and VIII.6.1.7. (\( C^1 \) version). The only other true generalizations I know of the Implicit and Inverse Function Theorems are extensions to maps between more general spaces like Banach spaces (cf. ()), or results like VIII.6.2.1.

VIII.6.7. **The Domain and Range Straightening Theorems**

From the Implicit and Inverse Function Theorems we can get a simple general local picture of a differentiable function, especially one of constant rank: up to a smooth change of coordinates, such a function of constant rank \( k \) from \( \mathbb{R}^n \) to \( \mathbb{R}^m \) locally consists simply of projection onto a \( k \)-dimensional subspace of \( \mathbb{R}^n \) followed by a linear embedding of this subspace into \( \mathbb{R}^m \).

The first result, a variation of VIII.6.2.1., is called the **Domain-Straightening Theorem** in some references, e.g. [?]. If \( f \) is a smooth function from \( U \subseteq \mathbb{R}^n \) to \( \mathbb{R}^m \) and \( a \in U \), then \( a \) is called a **regular point** of \( f \) if the
rank of $Df(a)$ is $m$. (If $f$ has a regular point, then necessarily $n \geq m$.) The set of regular points of $f$ is an open set.

**VIII.6.7.1. Theorem.** [Domain-Straightening Theorem] Let $f$ be a $C^r$ function ($1 \leq r \leq \infty$) from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, $n \geq m$, and $a$ a regular point of $f$. Then there is an open neighborhood $U$ of $a$ and a $C^r$ diffeomorphism $g$ from a neighborhood $U'$ of $(b, 0)$ in $\mathbb{R}^n \cong \mathbb{R}^m \times \mathbb{R}^{n-m}$ to $U$ with $g(b, 0) = a$ and $f \circ g(x, y) = x$ for all $(x, y) \in U'$, i.e.

$$f \circ g(x_1, \ldots, x_n) = (x_1, \ldots, x_m)$$

for all $(x_1, \ldots, x_n) \in U'$.

The name “Domain-Straightening Theorem” is appropriate since the theorem says that local coordinates can be chosen around $a$ so that $f$ just picks out the first $m$ coordinates of a point, and thus the preimage of any point in $\mathbb{R}^n$ in a neighborhood of $b$ is a piece of a vertical $(n-m)$-dimensional hyperplane. The case $n = m$ is just the Inverse Function Theorem.

**Proof:** By precomposing $f$ with a translation, there is no loss of generality in assuming $a = 0$. By assumption, the column rank of $Df(0)$ is $m$, so the columns of $Df(0)$ can be permuted so that the first $m$ columns are linearly independent. Thus $f$ can be precomposed with a linear transformation $T$ (permutation of coordinates) so that, if we write $\mathbb{R}^n$ as $\mathbb{R}^m \times \mathbb{R}^{n-m}$ and points in $\mathbb{R}^n$ as $(x, y)$, with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^{n-m}$, we have $D_x(f \circ T)(0, 0)$ invertible. Thus applying VIII.6.2.1. with the roles of $x$ and $y$ interchanged, there is an open neighborhood $V$ of $(b, 0)$ in $\mathbb{R}^n$ and a $C^r$ function $h : V \to \mathbb{R}^m$ such that $h(b, 0) = (0, 0)$ and $(f \circ T)(h(x, y), y) = x$ for all $(x, y) \in V$. The function $\phi : V \to \mathbb{R}^n$ defined by $\phi(x, y) = (h(x, y), y)$ has $D\phi(0, 0)$ invertible, so by the Inverse Function Theorem there is an open neighborhood $U'$ of $(b, 0)$ for which $\phi(U') \subseteq V$ and $\phi|_{U'}$ is a $C^r$ diffeomorphism from $U'$ to $\phi(U')$. Set $U = T(\phi(U'))$ and $g = T \circ \phi$.

**VIII.6.7.2. Corollary.** Let $f$ be a $C^r$ function ($r \geq 1$) from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ ($n \geq m$), $a$ a regular point of $f$, and $b = f(a)$. Then $a$ has an open neighborhood $U$ such that every open neighborhood of $a$ contained in $U$ maps onto an open neighborhood of $b$, and $f^{-1}(b) \cap U$ is an $(n-m)$-dimensional $C^r$ hypersurface in $U$. In particular, the range of $f$ contains an open set around $f(a)$.

**Proof:** Let $U$ and $g$ be as in VIII.6.7.1. Then

$$f^{-1}(b) \cap U = g((f \circ g)^{-1}(b)) = g \left(\{(b_1, \ldots, b_m, x_{m+1}, \ldots, x_n) : x_{m+1}, \ldots, x_n \in \mathbb{R}\} \cap U'\right)$$

where $b = (b_1, \ldots, b_m)$.

**VIII.6.7.3.** Conversely, it is a remarkable fact (??, a consequence of Sard’s Theorem) that if the range of $f$ contains a nonempty open set in $\mathbb{R}^m$, and $f$ is sufficiently smooth, then $f$ has “many” regular points.

By straightening only some of the coordinates in the domain, we obtain: 855
VIII.6.7.4. **Theorem.** Let \( f \) be a \( C^r \) function \((1 \leq r < \infty)\) from an open subset \( \Omega \) of \( \mathbb{R}^n \) to \( \mathbb{R}^m \), and \( a \) a point of the domain of \( f \). Suppose \( Df(a) \) has rank \( \geq k \), \( k < m \). Then there is an open neighborhood \( U \) of \( a \), a point \( b \in \mathbb{R}^k \), and a \( C^r \) diffeomorphism \( g \) from a neighborhood \( U' \) of \((b,0)\) in \( \mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}\) to \( U \) with \( g(b,0) = a \), an invertible linear transformation \( T : \mathbb{R}^m \rightarrow \mathbb{R}^m \) (permutation of coordinates) with \( T \circ f(a) = (b,c) \) for some \( c \in \mathbb{R}^{m-k} \), and a \( C^r \) function \( f_0 : \mathbb{R}^n \rightarrow \mathbb{R}^{m-k} \) defined on a neighborhood of \((b,0)\) such that
\[
T \circ f \circ g(x,y) = (x,f_0(x,y))
\]
for all \((x,y) \in U' \subseteq \mathbb{R}^k \times \mathbb{R}^{n-k} \), i.e.
\[
T \circ f \circ g(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) = (x_1, \ldots, x_k, f_0(x_1, \ldots, x_n))
\]
for all \((x_1, \ldots, x_n) \in U' \). We have \( f_0(b,0) = c \).

**Proof:** The matrix \( Df(a) \) has rank at least \( k \), so by permuting rows we may make the first \( k \) rows linearly independent. Let \( T \) be such a coordinate permutation, and let \( \phi \) be the function from \( \Omega \) to \( \mathbb{R}^k \) whose coordinate functions are the first \( k \) coordinate functions of \( T \circ f \). Set \( b = \phi(a) \). Then \( \phi \) is \( C^r \) and satisfies the hypotheses of VIII.6.7.1., so there is a \( C^r \) diffeomorphism \( g \) from a neighborhood \( U' \) of \((b,0)\) in \( \mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k} \) to a neighborhood \( U \) of \( a \) with \( \phi \circ g(x,y) = x \) for all \((x,y) \in U' \). Then \( T \circ f \circ g(x,y) \) is of the form
\[
(\phi \circ g(x,y), f_0(x,y)) = (x,f_0(x,y))
\]
for \((x,y) \in U' \), for some \( C^r \) function \( f_0 : U' \rightarrow \mathbb{R}^{m-k} \).

VIII.6.7.5. Suppose \( f \) is as in the statement of the theorem. If the rank of \( Df(z) \) is exactly \( k \) for all \( z \) in a neighborhood of \( a \), we can say more. Set \( h = T \circ f \circ g \). Then \( h \) has rank exactly \( k \) in a neighborhood of \((b,0)\). The derivative of \( h \) at \((x,y)\) is of the form
\[
\begin{bmatrix}
I_k & 0 \\
B & A
\end{bmatrix}
\]
where \( Df_0(x,y) = [B \; A] \), \( B \) is \((m-k) \times k \), and \( A \) is \((m-k) \times (n-k) \). For such a matrix to have rank \( k \), \( A \) must be 0: the first \( k \) columns are linearly independent, and if \( A \) has a nonzero entry the column it appears in will be linearly independent from the first \( k \) columns, a contradiction. Thus the partial derivatives of \( f_0 \) in the \( y \) variables are identically zero near \((b,0)\), i.e. \( f_0(x,y) \) depends only on the \( x \) variables in a neighborhood of \((b,0)\). Thus we obtain:

VIII.6.7.6. **Theorem.** [Rank Theorem] Let \( f \) be a \( C^r \) function \((1 \leq r < \infty)\) from an open subset of \( \mathbb{R}^n \) to \( \mathbb{R}^m \), and \( a \) a point of the domain of \( f \). Suppose \( Df \) has constant rank \( k \) in a neighborhood of \( a \). Then there is an open neighborhood \( U \) of \( a \), a point \( b \in \mathbb{R}^k \), a \( C^r \) diffeomorphism \( g \) from an open neighborhood \( U' \) of \((b,0)\) in \( \mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k} \) to \( U \), an open neighborhood \( W \) of \( b \) in \( \mathbb{R}^k \) such that \( x \in W \) whenever \((x,y) \in U' \), and a \( C^r \) function \( \tilde{f} : W \rightarrow \mathbb{R}^m \) such that
\[
f \circ g(x,y) = \tilde{f}(x)
\]
for all \((x,y) \in U' \).
In fact, since $f_0(x, y) = f_0(x, 0)$ for any $y$, we can take
\[ \tilde{f}(x) = T^{-1}(x, f_0(x, 0)) \]

In the case $k = n < m$ in VIII.6.7.4., the $T$ is unnecessary and can be eliminated, and the Rank Theorem applies (the case $k = n = m$ is just the Inverse Function Theorem). In this case a better result holds, called the Range-Straightening Theorem in [?]:

\[ \text{VIII.6.7. Theorem. [Range-Straightening Theorem]} \]

Let $f$ be a $C^r$ function $(1 \leq r \leq \infty)$ from an open subset $\Omega$ of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $a$ a point of the domain of $f$. Suppose $Df(a)$ has rank $n$ (then necessarily $n \leq m$). Set $b = f(a)$. Then there are open neighborhoods $U$ of $a$ in $\mathbb{R}^n$ and $V$ of $b$ in $\mathbb{R}^m$, and a $C^r$ diffeomorphism $h$ from $V$ to a neighborhood $V'$ of $(a, 0)$ in $\mathbb{R}^m$ with $h(b) = (a, 0)$, such that

\[ h \circ f(x) = (x, 0) \]

for all $x \in U$, i.e.

\[ h \circ f(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0) \]

for all $(x_1, \ldots, x_n) \in U$.

The name “Range-Straightening Theorem” means that a smooth function from $\mathbb{R}^n$ to $\mathbb{R}^m$ $(n < m)$ which has rank $n$ at a point $a$ has local range which up to a smooth change of coordinates is a piece of a horizontal $n$-dimensional hyperplane in $\mathbb{R}^m$. See VIII.6.7.10. for a precise and more general result.

**Proof:** Since $Df(a)$ has row rank $n$, there is an invertible linear transformation $T : \mathbb{R}^m \to \mathbb{R}^m$ (permutation of coordinates) so that the first $n$ rows of $D(T \circ f)(a)$ are linearly independent. Define $F : \Omega \times \mathbb{R}^{m-n} \subseteq \mathbb{R}^m \to \mathbb{R}^m$ by

\[ F(x, z) = T(f(x)) + (0, z) \]

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^{m-n}$, and the last expression uses $\mathbb{R}^m \cong \mathbb{R}^n \times \mathbb{R}^{m-n}$. Then $DF(a, 0)$ is of the form

\[
\begin{bmatrix}
A & 0 \\
B & I_{m-n}
\end{bmatrix}
\]

where the first column is $D(T \circ f)(a)$; since $A$ is invertible, $DF(a, 0)$ is invertible. Thus by the Inverse Function Theorem there is an inverse function $G$ defined on a neighborhood $V''$ of $F(a, 0) = T(f(a))$ to a neighborhood $V'$ of $(a, 0)$. Set $V = T^{-1}(V'')$ and $h = G \circ T$. If $U = \{ x \in \Omega : (x, 0) \in V' \}$, then

\[ h \circ f(x) = G \circ T \circ f(x) = G(T(f(x))) + (0, 0) = G(F(x, 0)) = (x, 0) \]

for all $x \in U$. \hfill \Box

Putting together the Range-Straightening Theorem with the Rank Theorem (which depends on the Domain-Straightening Theorem), we obtain the ultimate result (this result is called the Rank Theorem or the Constant Rank Theorem in some references):

\[ \text{VIII.6.7.8. Theorem. [Straightening Theorem]} \]

Let $f$ be a $C^r$ function $(1 \leq r \leq \infty)$ from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $a$ a point of the domain of $f$. Suppose $Df$ has constant rank $k$ in a neighborhood of $a$. Then there is an open neighborhood $U$ of $a$ and an open neighborhood $V$ of $b = f(a)$ with $f(U) \subseteq V$, and $C^r$ diffeomorphisms $g$ from a neighborhood $U'$ of $(0, 0)$ in $\mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}$ to $U$ and $h$ from $V$ to a neighborhood $V'$ of $(0, 0)$ in $\mathbb{R}^m \cong \mathbb{R}^k \times \mathbb{R}^{m-k}$ with $g(0, 0) = a$, $h(b) = (0, 0)$, and

\[ h \circ f \circ g(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) = (x_1, \ldots, x_k, 0, \ldots, 0) \]
Proof: By pre- and post-composing \( f \) with translations, there is no loss of generality in assuming that \( a = 0 \) and that \( b = f(0) = 0 \). By the Rank Theorem, there is a \( C^r \) diffeomorphism \( g \), defined on an open neighborhood \( U'' \) of \( 0 \) in \( \mathbb{R}^n \), an open neighborhood \( W \) of \( 0 \) in \( \mathbb{R}^k \), and a \( C^r \) function \( \tilde{f} : W \to \mathbb{R}^m \), such that \( f \circ g(x, y) = \tilde{f}(x) \) for all \( (x, y) \in U'' \). Then \( \tilde{f} \) satisfies the hypotheses of the Range-Straightening Theorem, so there is a neighborhood \( W' \) of \( 0 \) in \( \mathbb{R}^k \), contained in \( W \), an open neighborhood \( V \) of \( 0 \) in \( \mathbb{R}^m \), and a \( C^r \) diffeomorphism \( h \) from \( V \) to an open neighborhood \( V' \) of \( 0 \) such that

\[
\tilde{f} \circ h(x) = (x, 0)
\]

for all \( x \in W' \). Set

\[
U' = \{(x, y) \in U'' : x \in W'\}
\]

Then \( U' \) is an open neighborhood of \( 0 \) in \( \mathbb{R}^n \), and

\[
h \circ f \circ g(x, y) = (x, 0)
\]

for all \( (x, y) \in U' \). Set \( U = g(U') \).

VIII.6.7.9. Note that this theorem automatically applies if the rank of \( Df(a) \) is \( m \) (i.e. \( a \) is a regular point of \( f \)), since it is rank \( \leq m \) everywhere and at least \( m \) in some neighborhood of \( a \). In this case the \( h \) can be simply be taken to be a translation. Thus the Straightening Theorem is a generalization of the Domain-Straightening Theorem VIII.6.7.1. A similar observation holds if \( Df(a) \) has rank \( n \), so the Range-Straightening Theorem VIII.6.7.7. is also a special case of the Straightening Theorem. (And the Inverse Function Theorem is the case \( k = n = m \).)

VIII.6.7.10. Corollary. Let \( f \) be a \( C^r \) function \((1 \leq r \leq \infty)\) from an open subset of \( \mathbb{R}^n \) to \( \mathbb{R}^m \), and \( a \) a point of the domain of \( f \). Suppose \( Df \) has constant rank \( k \) in a neighborhood of \( a \). Then there is an open neighborhood \( U \) of \( a \) such that \( f(U) \) is a \( k \)-dimensional \( C^r \) hypersurface in \( \mathbb{R}^m \) passing through \( b = f(a) \).

Although the Straightening Theorem only applies at points where the function is locally of constant rank (i.e. has constant rank in a neighborhood of the point), this set of points is “generic”:

VIII.6.7.11. Proposition. Let \( f \) be a \( C^r \) function \((1 \leq r \leq \infty)\) from an open subset \( U \) of \( \mathbb{R}^n \) to \( \mathbb{R}^m \). Then the set of points where \( f \) is locally of constant rank is a dense open subset of \( U \).

Proof: For \( 0 \leq k \leq n \) let \( V_k \) be the set of points of \( U \) where \( f \) has rank \( \geq k \). Then each \( V_k \) is open, and the set \( W \) of points where \( f \) is locally of constant rank is

\[
W = \bigcup_{k=0}^{n} (V_k \setminus \bar{V}_{k+1})
\]

858
Thus $W$ is open, and $U \setminus W$ is

$$U \setminus W = U \cap \left[ \bigcup_{k=0}^{n} (\tilde{V}_k \setminus V_k) \right].$$

Each $\tilde{V}_k \setminus V_k = \partial V_k$ is closed and nowhere dense, hence its complement is dense and open; and a finite intersection of dense open sets is dense and open (XI.2.3.10.).

VIII.6.7.12. Remarks. (i) The only result in this section which applies to an $f$ which is not necessarily of locally constant rank around $a$ is VIII.6.7.4. But the Straightening Theorem applies at “almost every” point by VIII.6.7.11.

(ii) The Domain-Straightening Theorem and Range-Straightening Theorem are stated in such a way that only composition of $f$ on one side with a diffeomorphism is necessary. For the ultimate Straightening Theorem, composition on both sides is necessary in general.

(iii) Much of the notational complication of the results and proofs in this section comes from the need to carefully specify and name exact neighborhoods on which functions are defined and formulas are valid. It is probably a good idea for a reader to at first ignore these precise neighborhoods, replacing each instance simply by “on some neighborhood of $a$ (etc.)” or just “near $a$” (as is done in the text in a few places where precise specification of the neighborhood is unnecessary) to follow the essence of the argument without distraction; at the end of the day, the exact neighborhoods are unimportant, since the real conclusion is merely that there are some neighborhoods on which the results hold.

VIII.6.8. Analytic Versions

There are analytic versions of most of the results of this section:

VIII.6.8.1. Theorem. [Implicit Function Theorem, Real Analytic Version] Let $U$ be an open set in $\mathbb{R}^{p+n}$ and $(a, b) \in U$. Let $F : U \to \mathbb{R}^n$ with $F(a, b) = 0$. Suppose $F$ is analytic on $U$. If $D_y F(a, b)$ is invertible, i.e. $\det(D_y F(a, b)) \neq 0$, then there are:

(i) An open set $V$ in $\mathbb{R}^{p+n}$ containing $(a, b)$, with $V \subseteq U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{R}^p$ containing $a$;

(iii) A unique function $f : W \to \mathbb{R}^n$ such that $f(a) = b$ and, for all $x \in W$, $(x, f(x)) \in V$ and

$$F(x, f(x)) = 0.$$

Additionally, $f$ is analytic on $W$, and for any $x \in W$, we have

$$Df(x) = -D_y F(x, f(x))^{-1} D_x F(x, f(x))$$

(in particular, $D_y F(x, f(x))$ is invertible for all $x \in W$).
VIII.6.8.2. **Theorem.** [Real Analytic Inverse Function Theorem] Let $U$ be an open set in $\mathbb{R}^n$, $a \in U$, and $f : U \to \mathbb{R}^n$ an analytic function. If $Df(a)$ is invertible, then there is a neighborhood $V$ of $a$, $V \subseteq U$, such that

(i) $f$ is one-to-one on $V$.

(ii) $f(V)$ is open in $\mathbb{R}^n$.

(iii) $Df(x)$ is invertible for all $x \in V$.

(iv) $f^{-1}$ is analytic on $f(V)$ and, for all $y \in f(V)$,

$$D(f^{-1})(y) = [Df(f^{-1}(y))]^{-1}.$$ 

VIII.6.8.3. **Theorem.** [General Real Analytic Implicit/Inverse Function Theorem] Let $n \in \mathbb{N}$, $p \in \mathbb{N} \cup \{0\}$, $a \in \mathbb{R}^p$, $b, c \in \mathbb{R}^n$, $U$ an open neighborhood of $(a, b)$ in $\mathbb{R}^{p+n}$, $F : U \to \mathbb{R}^n$ an analytic function with $F(a, b) = c$ and $D_y F(a, b)$ invertible. Then there are:

(i) An open neighborhood $V$ of $(a, b)$ in $\mathbb{R}^{p+n}$, contained in $U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open neighborhood $W$ of $(a, c)$ in $\mathbb{R}^{p+n}$;

(iii) A unique function $G : W \to \mathbb{R}^n$ satisfying $G(a, c) = b$ and, for all $(x, z) \in W$, $(x, G(x, z)) \in V$ and

$$F(x, G(x, z)) = z.$$ 

Additionally, $G$ is analytic on $W$, and for all $(x, z) \in V$,

$$DG(x, z) = \begin{bmatrix} \vdots \\ -D_y F(x, G(x, z))^{-1}D_x F(x, G(x, z)) & D_y F(x, G(x, z))^{-1} \\ \vdots \end{bmatrix}.$$ 

VIII.6.8.4. **Theorem.** [Implicit Function Theorem, Complex Analytic Version] Let $U$ be an open set in $\mathbb{C}^{p+n}$ and $(a, b) \in U$. Let $F : U \to \mathbb{C}^n$ with $F(a, b) = 0$. Suppose $F$ is analytic on $U$. If $D_y F(a, b)$ is invertible, i.e. $det(D_y F(a, b)) \neq 0$, then there are:

(i) An open set $V$ in $\mathbb{C}^{p+n}$ containing $(a, b)$, with $V \subseteq U$, such that $D_y F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set $W$ in $\mathbb{C}^p$ containing $a$;
(iii) A unique function \( f : W \to \mathbb{C}^n \) such that \( f(a) = b \) and, for all \( x \in W \), \((x, f(x)) \in V \) and
\[
F(x, f(x)) = 0.
\]
Additionally, \( f \) is analytic on \( W \), and for any \( x \in W \), we have
\[
Df(x) = -DyF(x, f(x))^{-1}DxF(x, f(x))
\]
(in particular, \( D_yF(x, f(x)) \) is invertible for all \( x \in W \)).

VIII.6.8.5. Theorem. [Complex Analytic Inverse Function Theorem] Let \( U \) be an open set in \( \mathbb{C}^n \), \( a \in U \), and \( f : U \to \mathbb{C}^n \) an analytic function. If \( Df(a) \) is invertible, then there is a neighborhood \( V \)
of \( a \), \( V \subseteq U \), such that
(i) \( f \) is one-to-one on \( V \).
(ii) \( f(V) \) is open in \( \mathbb{C}^n \).
(iii) \( Df(x) \) is invertible for all \( x \in V \).
(iv) \( f^{-1} \) is analytic on \( f(V) \) and, for all \( y \in f(V) \),
\[
D(f^{-1})(y) = [Df(f^{-1}(y))]^{-1}.
\]

VIII.6.8.6. Theorem. [General Complex Analytic Implicit/Inverse Function Theorem] Let \( n \in \mathbb{N} \), \( p \in \mathbb{N} \cup \{0\} \), \( a \in \mathbb{C}^p \), \( b, c \in \mathbb{C}^n \), \( U \) an open neighborhood of \((a, b)\) in \( \mathbb{C}^{p+n} \), \( F : U \to \mathbb{C}^n \) an analytic function with \( F(a, b) = c \) and \( D_yF(a, b) \) invertible. Then there are:
(i) An open neighborhood \( V \) of \((a, b)\) in \( \mathbb{C}^{p+n} \), contained in \( U \), such that \( D_yF(x, y) \) is invertible for all \((x, y) \in V \);
(ii) An open neighborhood \( W \) of \((a, c)\) in \( \mathbb{C}^{p+n} \);
(iii) A unique function \( G : W \to \mathbb{C}^n \) satisfying \( G(a, c) = b \) and, for all \((x, z) \in W \), \((x, G(x, z)) \in V \) and
\[
F(x, G(x, z)) = z.
\]
Additionally, \( G \) is analytic on \( W \), and for all \((x, z) \in V \),
\[
DG(x, z) = \begin{bmatrix}
-D_yF(x, G(x, z))^{-1}D_xF(x, G(x, z)) & \cdots & D_yF(x, G(x, z))^{-1} \\
\vdots & \ddots & \vdots \\
\vdots & \cdots & \end{bmatrix}.
\]
The complex analytic versions can be efficiently proved as follows. We outline the proof of VIII.6.8.6., of which VIII.6.8.4. and VIII.6.8.5. are corollaries. So begin with an analytic function $F : U \to \mathbb{C}^n$ satisfying the hypotheses.

1. Regard $F$ as a $C^\infty$ function from $\mathbb{R}^{2p+2n}$ to $\mathbb{R}^{2n}$.

2. Apply VIII.6.2.1. to obtain the unique $G : \mathbb{R}^{2p+2n} \to \mathbb{R}^{2n}$ with the right properties. Then $G$ is $C^\infty$.

3. Check from the derivative formula of VIII.6.2.1. (cf. III.11.11.2.) that the partial derivatives of $G$ with respect to any consecutive odd/even pair of coordinates satisfy the Cauchy-Riemann equations (X.8.1.7.).

4. Conclude from X.8.1.6. that $G$, regarded as a function from $\mathbb{C}^{p+n}$ to $\mathbb{C}^n$, is complex analytic.

The Taylor series of $G$ around $(a, c)$ can be computed by brute force algebraic manipulation of power series (there is a unique multivariable power series which formally satisfies the identity $F(x, G(x, z)) = z$ for all $x, z$; cf. VIII.6.9.9.), although the calculation becomes rather ghastly. This calculation can be used to give an alternate proof of most of VIII.6.8.6.. But the calculation is not sufficient by itself: it must also be shown that the resulting series converges in a neighborhood of $(a, c)$. Then the function it converges to must satisfy the identity in the conclusion, and is the only analytic solution. Thus there is a unique analytic function satisfying the conclusion. But to conclude there is a unique function satisfying the identity, additional argument is needed (such as an application of VIII.6.2.1.).

There is an alternate approach which avoids all use of VIII.6.2.1. (or the other smooth results). Begin by proving the Complex Analytic Inverse Function Theorem VIII.6.8.5., where uniqueness comes along essentially for free, using the method of VIII.6.8.8. Then conclude VIII.6.8.6. as in VIII.6.2.2.

For the real analytic versions, the following procedure seems to suffice (we again outline the proof of VIII.6.8.3.). Let $F : U \to \mathbb{R}^n$ satisfy the hypotheses.

1. Complexify $F$ to a complex analytic function $F_C$ from a neighborhood of $(a, b)$ in $\mathbb{C}^{p+n}$ to $\mathbb{C}^n$, using the same power series centered at $(a, b)$.

2. Apply VIII.6.8.6. to $F_C$ to obtain a complex analytic function $G_C : \mathbb{C}^{p+n} \to \mathbb{C}^n$.

3. Restrict $G_C$ to $\mathbb{R}^{p+n}$ to get $G$.

There is, however, a technical subtlety: it must be shown that $G$ takes values in $\mathbb{R}^n$, not just in $\mathbb{C}^n$. This can be done by computing the Taylor series for $G_C$ around $(a, c)$ as in VIII.6.8.8. and verifying that all coefficients (partial derivatives) are real.

There is a better approach: as in VIII.6.8.9., first prove the Real Analytic Inverse Function Theorem VIII.6.8.2.. This can be done as follows. Suppose $f$ satisfies the hypotheses of VIII.6.8.2..

1. Apply the Inverse Function Theorem to show that $f$ is one-to-one on a neighborhood $V_r$ of $a$ with $C^\infty$ inverse function $f^{-1}$ on the neighborhood $W_r = f(V_r)$ of $b = f(a)$ in $\mathbb{R}^n$. 

862
(2) Complexify \( f \) to \( f_C \) on a neighborhood of \( a \) in \( \mathbb{C}^n \) with the same Taylor series. Verify that \( f_C \) satisfies the hypotheses of the Complex Analytic Inverse Function Theorem, hence is one-to-one on a neighborhood \( V_\epsilon \) of \( a \) in \( \mathbb{C}^n \) with complex analytic inverse function \( f_C^{-1} \) on the neighborhood \( W_\epsilon = f_C(V_\epsilon) \) of \( b \) in \( \mathbb{C}^n \).

(3) Use that \( f_C \) is one-to-one and extends \( f \) to conclude that the restriction of \( f_C^{-1} \) to the neighborhood \( W = W_\epsilon \cap W_\epsilon' \) of \( b \) in \( \mathbb{R}^n \) agrees with the restriction of \( f^{-1} \) to \( W \). Thus \( f^{-1} \) is real analytic on \( W \).

Theorem VIII.6.8.2. can also be proved by the method of VIII.6.8.8. without any use of the usual Inverse Function Theorem.

Thus the real and complex analytic theorems are all proved (see Exercise VIII.6.9.19. for details). \( \square \)

VIII.6.8.12. There are then real and complex analytic versions of the Straightening Theorems and the Rank Theorem; see Exercise VIII.6.9.19.

VIII.6.9. Exercises

VIII.6.9.1. Show that the hypothesis that \( D_yF(a, b) \) is invertible is not necessary for the conclusion of the Smooth Implicit Function Theorem, i.e. that the converse of the theorem does not hold. [Consider \( F(x, y) = x^3 - y^3 \).]

VIII.6.9.2. (a) If \( F \) is a function from \( \mathbb{R}^2 \) to \( \mathbb{R} \), and the graph of the equation \( F(x, y) = 0 \) is a curve that crosses itself at \( a \) with different tangent lines, and \( F \) is differentiable at \( a \), show that \( DF(a) = 0 \). [Consider directional derivatives.] This “explains” why the Implicit Function Theorem cannot apply at \( a \).

(b) If \( F \) is a \( C^1 \)-function from \( \mathbb{R}^2 \) to \( \mathbb{R} \), and the graph of the equation \( F(x, y) = 0 \) is a curve with two segments passing through \( a \) with the same tangent line, not identically equal in any neighborhood of \( a \), show that \( DF(a) = 0 \). [By making a linear change of variables, it may be assumed the tangent line at \( a \) is horizontal. Show that the Implicit Function Theorem cannot apply at \( a \).]

VIII.6.9.3. Let \( W \) be an open set in \( \mathbb{R}^p \), and \( f \) a continuous function from \( U \) to \( \mathbb{R}^n \). Define \( F : W \times \mathbb{R}^n \to \mathbb{R}^n \) by \( F(x, y) = y - f(x) \). Let \( a \in W \), and set \( b = f(a) \).

(a) Show that \( F, a, b \) satisfy the hypotheses of VIII.6.1.6..

(b) Show that the implicitly defined function is \( f \) (possibly restricted to a smaller neighborhood of \( a \)).

Thus every continuous function is defined implicitly by a function as in VIII.6.1.6. Since \( f \) need not be differentiable anywhere, no stronger conclusion can be obtained in VIII.6.1.6. without additional hypotheses, even in the case \( p = n = 1 \).

VIII.6.9.4. Show that under the hypotheses of VIII.6.3.2., even if \( p = 1 \), there is not necessarily a continuous function \( f \) on a neighborhood \( W \) of \( a \) with \( f(a) = b \) and \( F(x, f(x)) = 0 \) for all \( x \in W \). [Construct a pseudo-arc \( P \) in \( \{0, 1\} \times (0, 1) \subseteq \mathbb{R}^2 \) containing only one point \((0, c)\) on the left edge and one point on the right edge, whose complement in \([0, 1]^2\) has exactly two components, and let \( F \) be a continuous function from \([0, 1]^2\) to \( \mathbb{R} \) with \( F(0, y) = y - c \), \( F(x, 1) = 1 - c \), \( F(x, 0) = -c \), and which is zero precisely on \( P \). Note that \( \frac{\partial f}{\partial y}(0, c) \) exists and equals 1.]
VIII.6.9.5. (a) Assuming the Implicit Function Theorem, use the formula for $DF$ in VIII.6.2.3. to deduce the formula for the total derivative of the inverse function to $f$.

(b) Assuming the Inverse Function Theorem, use the formula for $D\tilde{F}$ in VIII.6.2.2. and () to deduce the formula for the total derivative $Df = -(D_yF)^{-1}D_xF$ of the implicitly defined function $f$.

VIII.6.9.6. Use VIII.6.3.9. and the argument of VIII.6.2.3. to obtain the version of the one-dimensional Inverse Function Theorem described in VIII.6.6.4.

VIII.6.9.7. Recall the way the formula for $f'(x)$ is found by implicit differentiation in first-semester calculus from an equation $F(x,y) = 0$: differentiate $G(x) = F(x,f(x))$ by the Chain Rule and other differentiation rules, set it equal to 0, and solve for $f'(x)$ (streamlined notation is usually used which amounts to this). Show that the same formula as the one in VIII.6.3.10. is always obtained.

VIII.6.9.8. In VIII.6.3.7., assume that the first-order partials of $F$ are defined on $U$ and continuous at $(a,b)$ (which implies that $F$ is differentiable at $(a,b)$), and that $\partial F/\partial y$ is nonzero on $U$ (which can be automatically arranged in the continuous partial case if $\partial F/\partial y(a,b) \neq 0$ by replacing $U$ by a smaller neighborhood of $(a,b)$). Give a simplified proof as follows (note that the $f$ is unique and continuous in this case by VIII.6.3.4.):

(a) For notational simplicity, assume $p = 1$. If $\Delta x$ is sufficiently small, write $\Delta y = f(a+\Delta x) - f(a) = f(a+\Delta x) - b$. Then

\[ 0 = F(a+\Delta x, b+\Delta y) = F(a, b) = F(a+\Delta x, b+\Delta y) - F(a, b) \]

\[ = [F(a+\Delta x, b + \Delta y) - F(a, b + \Delta y)] + [F(a, b + \Delta y) - F(a, b)] . \]

(b) By two applications of the MVT, there are $0 < \alpha, \beta < 1$ (depending on $\Delta x$) such that

\[ F(a + \Delta x, b + \Delta y) - F(a, b + \Delta y) = \partial F/\partial x(a + \alpha \Delta x, b + \Delta y) \Delta x \]

\[ F(a, b + \Delta y) - F(a, b) = \partial F/\partial y(a, b + \beta \Delta y) \Delta y . \]

(c) From (a) and (b),

\[ \partial F/\partial x(a + \alpha \Delta x, b + \Delta y) \Delta x + \partial F/\partial y(a, b + \beta \Delta y) \Delta y = 0 \]

so

\[ \frac{\Delta y}{\Delta x} = -\frac{\partial F/\partial x(a + \alpha \Delta x, b + \Delta y)}{\partial F/\partial y(a, b + \beta \Delta y)} . \]

(d) By continuity,

\[ \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} -\frac{\partial F/\partial x(a + \alpha \Delta x, b + \Delta y)}{\partial F/\partial y(a, b + \beta \Delta y)} \]

exists and equals $\partial F/\partial y(a,b)$.

(e) Adapt the argument notationally to the case $p > 1$. 

864
VIII.6.9.9. Use the Chain Rule to find formulas for the higher-order derivatives of implicitly defined functions.

VIII.6.9.10. Give a direct argument proving the Continuous Implicit Function Theorem along the lines of the proof of VIII.6.5.1. by proving the following version first proved in [Gou03]:

THEOREM. Let $U$ be an open set in $\mathbb{R}^{p+n}$ containing $(0, 0)$, and $G : U \to \mathbb{R}^n$ a continuous function with $G(0, 0) = 0$, such that $D_yG$ is continuous on $U$ and $D_yG(0, 0) = 0$. Then there are open sets $V$ in $\mathbb{R}^{p+n}$, $V \subseteq U$, $(0, 0) \in V$, and $W$ in $\mathbb{R}^p$, $0 \in W$, and $Y$ in $\mathbb{R}^n$, $0 \in Y$, such that for every $y \in Y$ there is a unique $x \in W$ with $(x, y) \in V$ and $G(x, y) = y$.

(a) Prove this theorem by using the Contraction Mapping Principle.
(b) If $U \subseteq \mathbb{R}^{p+n}$, $(0, 0) \in U$, and $F : U \to \mathbb{R}^n$ is a continuous with $F(0, 0) = 0$, such that $D_yF$ is continuous on $U$ and $D_yF(0, 0) = I$, show that there is a unique implicitly defined function $f$ on a neighborhood $W$ of $0$ in $\mathbb{R}^p$, taking values in a neighborhood $Y$ of $0$ in $\mathbb{R}^n$, such that $F(x, f(x)) = 0$ for all $x \in W$. [Set $G(x, y) = y - F(x, y)$]
(c) Use an affine change of variables to deduce the existence of an implicit function under the hypotheses of VIII.6.1.6. Conclude continuity of the implicit function from VIII.6.1.12.

See [?] for a nice geometric interpretation of this argument.

VIII.6.9.11. Define $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$ F(x, y) = \begin{cases} y + x^2 \cos \left( \frac{y}{x^2} \right) & \text{if } x \neq 0 \\ y & \text{if } x = 0 \end{cases} $$

(a) Show that $F$ is differentiable everywhere. In particular, $F$ is continuous, and $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ exist everywhere.
(b) Show that $\frac{\partial F}{\partial y}$ is continuous in $y$ for fixed $x$, and that $\frac{\partial F}{\partial y}(0, 0) = 1$.
(c) Show that the equation $F(x, y) = 0$ cannot be uniquely solved for $y$ as a differentiable function of $x$ in any small neighborhood of $(0, 0)$ (the smaller the neighborhood, the more differentiable solutions there are).

VIII.6.9.12. (a) Write a statement which is a common generalization of VIII.6.3.12. and VIII.6.6.7., along the lines of VIII.6.4.2.

***(b) Is this statement true? [cf. VIII.6.9.13.(b).]

VIII.6.9.13. (a) Show that under the hypotheses of VIII.6.6.5., if $U$ is connected, then $\text{det}(Df(x))$ does not change sign on $U$, i.e. $f$ either preserves or reverses orientation on all of $U$.

***(b) Is the assumption of a decreasing chain of minors as in VIII.6.4.3. automatic? (If so, an inductive proof of VIII.6.6.5. and the statement in VIII.6.9.12. can be given.)

VIII.6.9.14. Let $U$ and $V$ be open sets in $\mathbb{R}^n$, and $f : U \to V$ a bijection which is $C^r$ ($1 \leq r \leq \infty$). If $f^{-1}$ is differentiable everywhere on $V$, show that $f^{-1}$ is necessarily $C^r$ on $V$.

VIII.6.9.15. Formulate and prove generalizations of the Continuous and Mixed Implicit Function Theorem similar to VIII.6.2.1.
VIII.6.9.16. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be given by $f(x) = x^2$. Around which points can $f$ be straightened?

VIII.6.9.17. Let $f : \mathbb{R} \rightarrow \mathbb{R}^2$ be given by $f(x) = (x^3, x^2)$. Then $f$ is $C^\infty$. Show that the range of $f$ can be topologically straightened but not smoothly straightened around 0.

VIII.6.9.18. Define $f : \mathbb{R} \rightarrow \mathbb{R}^2$ by

$$f(x) = \left( e^{-1/x^2} \left[ 1 - \cos \frac{2\pi}{x} \right], e^{-1/x^2} \sin \frac{2\pi}{x} \right)$$

for $x \neq 0$, and $f(0) = (0, 0)$.

(a) Show that $f$ is a $C^\infty$ function.

(b) Show that the image of $\left( -\frac{1}{n}, \frac{1}{n} \right)$ is topologically a Hawaiian earring (XI.18.4.6.) for any $n \in \mathbb{N}$. Thus the range cannot even be topologically straightened around 0.

(c) Discuss straightening of the function $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by

$$g(x, y) = \left( e^{-1/x^2} \left[ 1 - \cos \frac{2\pi}{x} \right], y + e^{-1/x^2} \sin \frac{2\pi}{x} \right)$$

for $x \neq 0$ and $g(0, y) = (0, y)$.

VIII.6.9.19. (a) Write out the details of the proofs of the real and complex analytic versions of the Implicit and Inverse Function Theorems.

(b) Compute enough terms of the Taylor series of the implicitly defined function to indicate the general procedure and the proof of uniqueness.

(c) Verify by calculation the derivative formulas for the complex analytic versions (for the real analytic versions, these formulas are trivial from the smooth versions).

(d) Formulate and prove real and complex analytic versions of the Straightening Theorems and the Rank Theorem.

VIII.6.9.20. Formulate and prove precise versions, both real and complex, of the statement that the roots of an $n$'th degree polynomial are continuous, and in fact analytic, functions of the coefficients, except at multiple roots. [Fix $n$, and consider the function

$$F(a_0, \ldots, a_n, x) = a_0 + a_1 x + \cdots + a_n x^n$$

and solve $F(a_0, \ldots, a_n, x) = 0$ for $x$.]

This suggests there are analytic formulas for the roots of a general $n$'th degree polynomial, although explicit expressions for the formulas may not be possible in general. There cannot be explicit formulas just involving finitely many algebraic operations, including roots, if $n \geq 5$. 

866
VIII.6.9.21. Let $U$ be an open set in $\mathbb{R}^3$, and $F : U \to \mathbb{R}$ be a $C^1$ function with $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, and $\frac{\partial F}{\partial z}$ nonzero on $U$. Then the equation $F(x, y, z) = 0$ can be locally solved for each of $x, y, z$ as functions of the other two. Show that

$$\frac{\partial x}{\partial y} \frac{\partial y}{\partial z} \frac{\partial z}{\partial x} = -1.$$ 

(Give a careful explanation of what this equation means.) Thus algebraic manipulations of Leibniz notation must be done carefully.


VIII.7. Power Series and Analytic Functions of Several Variables

The theory of power series and analytic functions in several variables is similar to the theory in one variable with some technical complications. In this section, we will take the path of least resistance to the most important aspects of the theory, avoiding some of the (sometimes interesting) technicalities that arise. As in the one-variable case, we can do the real and complex theories in parallel; we let \( \mathbb{F} \) stand for either \( \mathbb{R} \) or \( \mathbb{C} \).

VIII.7.1. Multivariable Power Series

VIII.7.1.1. In general, we want to consider power series of the form

\[
\sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} \cdots \sum_{k_n=0}^{\infty} a_{k_1 \ldots k_n} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}
\]

where the \( a_{k_1 \ldots k_n} \) are scalars (real or complex) and \( x_1, \ldots, x_n \) are independent variables. The notation can become oppressive working with a general \( n \), although multiindex notation (VIII.7.1.5.) can help, so sometimes we will notationally stick to the case \( n = 2 \), writing \( x \) and \( y \) (or \( z \) and \( w \) in the complex case) for the independent variables. Just as in multivariable calculus, the case \( n = 2 \) exhibits all the essential features of the theory for general \( n \) in contrast with the case \( n = 1 \). Thus a power series in two variables will be written

\[
\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} a_{jk} x^j y^k
\]

or more generally

\[
\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} a_{jk} (x-x_0)^j (y-y_0)^k.
\]

VIII.7.1.2. Even in the two-variable case, there is no natural ordering on the terms, so we will regard the infinite series as an unordered sum

\[
\sum_{(j,k)} a_{jk} (x-x_0)^j (y-y_0)^k
\]

where the sum is over \( \mathbb{N}_0^3 \) (recall the notation \( \mathbb{N}_0 = \mathbb{N} \cup \{0\} \)). In general, we will have an unordered sum over \( \mathbb{N}_0^n \). Regarding multivariable power series as unordered sums means that we will only consider absolute convergence, and leads to technical simplifications in the theory without essential loss of generality.

VIII.7.1.3. Since we are considering multivariable power series as unordered sums, there is no harm in allowing more than one term with the same powers \( (x-x_0)^j (y-y_0)^k \); we may even have infinitely many terms with the same pair \( (j,k) \) of powers, and it is convenient to allow this. Any such sum can be converted to standard form by first summing together all terms with powers a fixed \( (j,k) \), and then summing over \( (j,k) \); the standard form is a drastic revision of the original power series. As long as the original power series converges (as an unordered sum, i.e. converges absolutely), the standard form conversion and series will also converge (IV.2.10.20.). So we can make a formal definition:
VIII.7.1.4. Definition. Let $\Omega$ be an index set, and $\phi = (\phi_1, \ldots, \phi_n) : \Omega \to \mathbb{N}_0^n$ a function. A multivariable power series in variables $X_1, \ldots, X_n$ indexed by $\Omega$ is an unordered sum

$$
\sum_{\omega \in \Omega} a_{\omega} X_1^{\phi_1(\omega)} X_2^{\phi_2(\omega)} \cdots X_n^{\phi_n(\omega)}
$$

where the $a_{\omega}$ are scalars. If $(c_1, \ldots, c_n)$ is an $n$-tuple of scalars, a multivariable power series in $x_1, \ldots, x_n$ centered at $(c_1, \ldots, c_n)$, indexed by $\Omega$, is an unordered sum

$$
\sum_{\omega \in \Omega} a_{\omega} (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_n - c_n)^{\phi_n(\omega)} .
$$

If $n = 2$, a multivariable power series in $x$ and $y$ centered at $(x_0, y_0)$, indexed by $\Omega$, is an unordered sum

$$
\sum_{\omega \in \Omega} a_{\omega} (x - x_0)^{\phi_1(\omega)} (y - y_0)^{\phi_2(\omega)} .
$$

The multivariable power series is in standard form if $\phi$ is one-to-one; in this case, we may take $\Omega = \mathbb{N}_0^n$ and write the power series as

$$
\sum_{(k_1, \ldots, k_n) \in \mathbb{N}_0^n} a_{k_1, \ldots, k_n} (x_1 - c_1)^{k_1} \cdots (x_n - c_n)^{k_n} .
$$

Any multivariable power series has a unique drastic revision which is in standard form.

VIII.7.1.5. It is convenient to use multi-index notation: if $\alpha = (m_1, \ldots, m_n) \in \mathbb{N}_0^n$, and $x = (x_1, \ldots, x_n)$, write $x^\alpha = x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}$. We can then write the series

$$
\sum_{\omega \in \Omega} a_{\omega} (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_n - c_n)^{\phi_n(\omega)}
$$

centered at $c = (c_1, \ldots, c_n)$ as

$$
\sum_{\omega \in \Omega} a_{\omega} (x - c)^{\phi(\omega)}
$$

and a standard form multivariable power series as

$$
\sum_{\alpha \in \mathbb{N}_0^n} a_{\alpha} (x - c)^\alpha .
$$

VIII.7.1.6. If $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, the power series

$$
\sum_{\omega \in \Omega} a_{\omega} (x - c)^{\phi(\omega)}
$$

converges at $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$ if the unordered sum

$$
\sum_{\omega \in \Omega} a_{\omega} (b - c)^{\phi(\omega)}
$$

converges (i.e. converges absolutely). We then have by comparison:
VIII.7.1.7. **Proposition.** Let
\[ \sum_{\omega \in \Omega} a_{\omega}(x - c)^{\phi(\omega)} \]
be a multivariable power series. If the series converges at \( b = (b_1, \ldots, b_n) \), then it converges at \( x = (x_1, \ldots, x_n) \) for any \( (x_1, \ldots, x_n) \in \mathbb{F}^n \) with \( |x_1 - c_1| \leq |b_1 - c_1|, \ldots, |x_n - c_n| \leq |b_n - c_n| \).

VIII.7.1.8. Thus, if the series converges for some \( b \) with \( b_k \neq c_k \) for all \( k \), then it converges at all points of the rectangle (or polydisk if \( \mathbb{F} = \mathbb{C} \))
\[ R = \{ x : |x_k - c_k| \leq |b_k - c_k| \text{ for all } 1 \leq k \leq n \} \]
centered at \( c \), and thus defines a function from \( R \) to \( \mathbb{F} \).

VIII.7.1.9. **Example.** A one-variable power series converges precisely on an interval (or disk in the complex case). But the situation is more complicated with multivariable power series: the set of convergence need not be a rectangle (or polydisk). For example, consider the real 2-variable power series
\[ \sum_{k=0}^{\infty} x^k y^k. \]
This series converges (to \( \frac{1}{1-xy} \)) if and only if \( |xy| < 1 \), i.e. on the open subset of \( \mathbb{R}^2 \) containing \( (0,0) \) bounded by the hyperbolas \( y = \pm \frac{1}{x} \).

The best simple general statement we can make is that for any fixed \( y \), a 2-variable power series converges at \( (x,y) \) exactly for all \( x \) in a (possibly degenerate or empty) interval (disk) centered at \( x_0 \), and similarly for fixed \( x \). Thus if \( S \) is the set of convergence, every horizontal cross section is an interval (disk) centered at \( x_0 \), and every vertical cross section is an interval (disk) centered at \( y_0 \). These statements are true since if one variable is fixed, the series is an ordinary power series in the other variable. A similar statement holds in higher dimensions (fixing all but one variable). Thus, in general, the set of convergence is a union of rectangles (polydisks) centered at \( c \).

VIII.7.1.10. **Proposition.** Let
\[ \sum_{\omega \in \Omega} a_{\omega}(x - c)^{\phi(\omega)} \]
be a multivariable power series which converges on a rectangle (polydisk)
\[ R = \{ x : |x_k - c_k| \leq r_k \text{ for all } 1 \leq k \leq n \} \]
centered at \( c \), where \( r_1, \ldots, r_n > 0 \). Let \( f \) be the function on \( R \) defined by the power series. For each finite \( E \subseteq \Omega \), define
\[ f_E(x) = \sum_{\omega \in E} a_{\omega}(x - c)^{\phi(\omega)} \]

for $\mathbf{x} \in \mathbb{R}$. Then $f_E \to f$ uniformly on $\mathbb{R}$, i.e. for every $\epsilon > 0$ there is a finite subset $F$ of $\Omega$ such that, whenever $E$ is a finite subset of $\Omega$ containing $F$,

$$|f(\mathbf{x}) - f_E(\mathbf{x})| < \epsilon$$

for all $\mathbf{x} \in \mathbb{R}$.

**Proof:** This is essentially the Weierstrass $M$-test (). Set $\mathbf{r} = (r_1, \ldots, r_n)$. The unordered sum

$$\sum_{\omega \in \Omega} a_{\omega} \mathbf{r}^{\phi(\omega)}$$

converges; hence for $\epsilon > 0$ there is a finite subset $F \subseteq \Omega$ such that

$$\sum_{\omega \in \Omega \setminus F} |a_{\omega}| r^{\phi(\omega)} < \epsilon .$$

Thus, if $E$ is a finite subset of $\Omega$ containing $F$, we have

$$|f(\mathbf{x}) - f_E(\mathbf{x})| = \left| \sum_{\omega \in \Omega \setminus E} a_{\omega} (\mathbf{x} - \mathbf{c})^{\phi(\omega)} \right|$$

$$\leq \sum_{\omega \in \Omega \setminus E} |a_{\omega}| |x_1 - c_1|^{\phi_1(\omega)} |x_2 - c_2|^{\phi_2(\omega)} \cdots |x_n - c_n|^{\phi_n(\omega)} \leq \sum_{\omega \in \Omega \setminus F} |a_{\omega}| r^{\phi(\omega)} < \epsilon .$$

\[ \blacksquare \]

**VIII.7.1.11. Corollary.** If a multivariable power series converges on a rectangle (polydisk) $\mathbb{R}$, it defines a continuous function on $\mathbb{R}$.

We now show that a multivariable power series can be differentiated term-by-term.

**VIII.7.1.12. Theorem.** Let

$$\sum_{\omega \in \Omega} a_{\omega} (\mathbf{x} - \mathbf{c})^{\phi(\omega)}$$

be a multivariable power series which converges on a rectangle (polydisk)

$$\mathbb{R} = \{ \mathbf{x} : |x_k - c_k| \leq r_k \text{ for all } 1 \leq k \leq n \}$$

centered at $\mathbf{c}$, where $r_1, \ldots, r_n > 0$. Let $f$ be the function on $\mathbb{R}$ defined by the power series. Then, for each $k, 1 \leq k \leq n$, the unordered sum

$$\sum_{\omega \in \Omega} a_{\omega} \frac{\partial}{\partial x_k} ((\mathbf{x} - \mathbf{c})^{\phi(\omega)}) = \sum_{\omega \in \Omega} \phi_k(\omega) a_{\omega} (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_k - c_k)^{\phi_k(\omega)-1} \cdots (x_n - c_n)^{\phi_n(\omega)}$$

(VIII.2)
converges u.c. to \( \frac{\partial f}{\partial x_k} \) on the open rectangle (polydisk) \( R^0 = \{ x : |x_k - c_k| < r_k \text{ for all } 1 \leq k \leq n \} \).

**Proof:** This can be quickly reduced to the one-variable case. Fix \( k \). For each \( j \neq k \), fix \( x_j \) with \( |x_j - c_j| \leq r \). For any \( x_k, |x_k - c_k| \leq r_k \), the unordered sum
\[
\sum_{\omega \in \Omega} a_\omega (x - c)^{\phi(\omega)}
\]
converges; hence it can be computed by summing in any manner. Form the drastic revision
\[
\sum_{m \in \mathbb{N}_0} b_m (x_k - c_k)^m
\]
obtained by first summing all terms with factor \( (x_k - c_k)^m \) for fixed \( m \). This is an unordered power series in one variable \( x_k - c_k \) which converges for \( |x_k - c_k| \leq r_k \); thus the associated ordinary power series
\[
\sum_{m=0}^{\infty} b_m (x_k - c_k)^m \tag{VIII.3}
\]
has radius of convergence at least \( r_k \), and converges to \( f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) \) for \( |x - c_k| \leq r_k \). Thus, by V.15.3.23., the series of term-by-term derivatives converges absolutely to
\[
\frac{d}{dx} \left( f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) \right) = \frac{\partial f}{\partial x_k} (x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)
\]
for \( |x - c_k| < r_k \), i.e. VIII.3 converges to \( \frac{\partial f}{\partial x_k} \) on \( R^0 \).

We have not yet shown that the original multivariable series VIII.2 converges. But this can be shown by a majorization argument. The above argument can be used starting with the series of absolute values
\[
\sum_{\omega \in \Omega} |a_\omega| (|x - c|)^{\phi(\omega)}
\]
which converges on \( R \) to a function \( F \). If we choose the \( x_j \) so that \( x_j - c_j \geq 0 \) for all \( j \), then the unordered sum
\[
\sum_{\omega \in \Omega} \phi_k(\omega)|a_\omega| (x_1 - c_1)^{\phi_1(\omega)}(x_2 - c_2)^{\phi_2(\omega)} \cdots (x_k - c_k)^{\phi_k(\omega)-1} \cdots (x_n - c_n)^{\phi_n(\omega)} \tag{VIII.4}
\]
has a drastic revision which converges (to \( \frac{\partial F}{\partial x_k} \)); thus the series VIII.4 also converges since all the terms are nonnegative. This implies that VIII.2 converges on \( R^0 \) by comparison. Since it has a drastic revision which converges to \( \frac{\partial f}{\partial x_k} \) on \( R^0 \), VIII.2 also converges to \( \frac{\partial F}{\partial x_k} \) on \( R^0 \).

This argument shows pointwise convergence of VIII.2 on \( R^0 \); u.c. convergence is then automatic by VIII.7.1.10. \( \blacklozenge \).
VIII.7.1.13. **Corollary.** Let
\[ \sum_{\omega \in \Omega} a_\omega (x - c)^{\phi(\omega)} \]
be a multivariable power series which converges on an open rectangle (polydisk)
\[ R^o = \{ x : |x_k - c_k| < r_k \text{ for all } 1 \leq k \leq n \} \]
centered at \( c \), where \( r_1, \ldots, r_n > 0 \). Let \( f \) be the function on \( R^o \) defined by the power series. Then \( f \) is \( C^\infty \) on \( R^o \), and the partial derivatives of all orders of \( f \) are obtained by term-by-term partial differentiation, i.e.
for any \( m_1, \ldots, m_n \geq 0 \) the unordered sum
\[ \sum_{\omega \in \Omega} a_\omega \frac{\partial^{m_1 + \cdots + m_n}}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}} ( (x - c)^{\phi(\omega)} ) \]
converges u.c. to \( \frac{\partial^{m_1 + \cdots + m_n} f}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}} \) on \( R^o \).

**Proof:** Every point of \( R^o \) is contained in the interior of a closed polydisc centered at \((c_1, \ldots, c_n)\) contained in \( R^o \). Apply VIII.7.1.12. repeatedly, shrinking the polydisk slightly each time.

VIII.7.1.14. We extend the multi-index notation: if \( \alpha = (m_1, \ldots, m_n) \in \mathbb{N}^n_0 \), write
\[ |\alpha| = m_1 + \cdots + m_n \]
\[ \alpha! = m_1! m_2! \cdots m_n! \]
\[ \frac{\partial^\alpha}{\partial x_\alpha} = \frac{\partial^{|\alpha|}}{\partial x_\alpha} = \frac{\partial^{m_1 + \cdots + m_n}}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}. \]
We can then rephrase VIII.7.1.13.: under the hypotheses, for every \( \alpha \) the series
\[ \sum_{\omega \in \Omega} a_\omega \frac{\partial^\alpha}{\partial x_\alpha} ( (x - c)^{\phi(\omega)} ) \]
converges u.c. on \( R^o \) to \( \frac{\partial^\alpha f}{\partial x_\alpha} \).

VIII.7.1.15. Equality of mixed partials (VIII.2.2.2.) for functions defined by convergent power series is an immediate corollary of VIII.7.1.13., giving a simpler alternate proof in this case.

VIII.7.1.16. Note that when \( F = \mathbb{C} \), these partial derivatives are interpreted in the complex sense: fix the remaining variables and take the derivative as in X.1.1.2.. Thus a complex-valued function defined on a polydisk by a multivariable power series has partial derivatives of all orders in the complex sense, which are also given by multivariable power series obtained by term-by-term differentiation.
VIII.7.2. Analytic Functions of Several Variables

In this section, we describe some of the basics of the theory of real and complex analytic functions of several variables. Some aspects of the theory are virtually identical to the one-variable case, but there are also significant differences.

VIII.7.2.1. Definition. Let $U$ be an open set in $\mathbb{F}^n$. A function $f : U \rightarrow \mathbb{F}$ is analytic if, for every $c \in U$, $f$ is represented in a neighborhood of $c$ by a multivariable power series centered at $c$. A function $f = (f_1, \ldots, f_m) : U \rightarrow \mathbb{F}^m$ is analytic if each $f_k$ is analytic.

VIII.7.2.2. As in the one-variable case, this definition is a local one: a function $f$ on an open set $U$ is analytic on $U$ if and only if it is analytic on a neighborhood of $c$ for each $c \in U$. In particular, if $U = \cup_j U_j$ is a union of open sets, then $f$ is analytic on $U$ if and only if it is analytic on each $U_j$.

An immediate corollary of VIII.7.1.13. is:

VIII.7.2.3. Corollary. Let $U$ be an open set in $\mathbb{F}^n$, and $f$ analytic on $U$. Then $f$ is $C^\infty$ on $U$, and for any multi-index $\alpha$, $\frac{\partial^\alpha f}{\partial x^\alpha}$ is analytic on $U$.

VIII.7.2.4. An analytic function $f$ is “jointly analytic” in its variables. If all but one variable is fixed, it is analytic in the remaining variable in the usual sense of $(\cdot)$, since the multivariable power series for $f$ around any point has a drastic revision obtained by collecting together all terms of like power in the remaining variable, which will converge in a neighborhood of the point in $\mathbb{F}$. (More generally, if some of the variables are fixed, the function is analytic in the remaining variables.) Thus an analytic function is “separately analytic” in its variables. But a separately real analytic function is not (jointly) analytic in general. For example,

$$f(x, y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

is separately analytic, but not even continuous at $(0, 0)$ $(\cdot)$. This function is analytic on $\mathbb{R}^2 \setminus \{(0,0)\}$. The function

$$g(x, y) = \begin{cases} xye^{-1/(x^2+y^2)} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

is separately analytic, $C^\infty$ on $\mathbb{R}^2$, and analytic on $\mathbb{R}^2 \setminus \{(0,0)\}$, but not analytic on any neighborhood of $(0,0)$.

The situation is completely different in the complex case: a complex-valued function on an open subset of $\mathbb{C}^n$ which is separately analytic is in fact analytic (X.8.1.5.).

As in the one-variable case, the definition of analytic function is hard to check, since the condition of the definition must be verified for every $c \in U$. Thus the next result is crucial in obtaining a useful theory:
VIII.7.2.5. **Theorem.** Let $U$ be an open set in $\mathbb{F}^n$. Then any multivariable power series which converges on $U$ defines an analytic function on $U$.

**Proof:** Let

$$\sum_{\omega \in \Omega} a_\omega (x - c)^{\phi(\omega)}$$

be a multivariable power series which converges on $U$. We may assume that $U$ is the interior of the set on which the series converges, which is a union of polydisks centered at $c$ (since this is an open set containing $U$), and in particular that $c \in U$. $\Diamond$
Chapter IX

Differential Equations

“The most widely and successfully used concept in mathematical descriptions of scientific problems has been the differential equation, even though success has often been impeded by the fact that differential equations can be extremely difficult to analyze. These difficulties, rather than thwarting progress, have instead stimulated more research and better understanding. As a result, the subject of differential equations contains a wealth of deep theories and subtle methods of analysis.”

E. A. Herman

The subject of Differential Equations consists of the study of differential equations, how to solve them, and the properties the solutions must have. A differential equation is an equation involving an unknown function and its derivatives. A solution to a differential equation is a sufficiently differentiable function satisfying the equation, usually also satisfying some specified initial or boundary conditions. The unknown function may be a function of one independent variable or more than one. A differential equation involving a function of one independent variable is an ordinary differential equation or ODE; a differential equation involving a function of more than one independent variable is a partial differential equation or PDE. One can also consider systems of ODEs or PDEs involving several unknown functions. In general, as might be easily predicted, the theory of PDEs is much more difficult than the theory of ODEs, and most active research in Differential Equations today involves PDEs.

Differential equations are of fundamental importance in both pure and applied mathematics. For example, the basic laws of physics, beginning with $F = ma$, are really differential equations [if $y(t)$ is position at time $t$, and $F(y)$ is the force at position $y$, then the equation is

$$my''(t) = F(y(t)).$$

The force could also be a function of both $y$ and $t$. There are variations possible such as taking both $y$ and $F$ to be vector-valued.]

The subject of Differential Equations is one of the largest areas of mathematics, bridging pure and applied mathematics. Entire books are written on the solutions to important specific differential equations. A considerable part of Numerical Analysis involves computing numerical approximations to solutions of

---

1[? p. 161].

876
differential equations. We will only treat the most basic aspects of the theory, with an emphasis on the analysis tools needed to obtain them, and we will concentrate on establishing general theorems about existence, uniqueness, and properties of solutions rather than the elaborate techniques used to actually find them explicitly.

“Newton’s *Principia* was impressive, with its revelation of deep mathematical laws underlying natural phenomena. But what happened next was even more impressive. Mathematicians tackled the entire panoply of physics – sound, light, heat, fluid flow, gravitation, electricity, magnetism. In every case, they came up with differential equations that described the physics, often very accurately.

The long-term implications have been remarkable. Many of the most important technological advances, such as radio, television, and commercial jet aircraft depend, in numerous ways, on the mathematics of differential equations. The topic is still the subject of intense research activity, with new applications emerging almost daily. It is fair to say that Newton’s invention of differential equations, fleshed out by his successors in the 18th and 19th centuries, is in many ways responsible for the society in which we now live. This only goes to show what is lurking just behind the scenes, if you care to look.”

*Ian Stewart*²

---

IX.1. Existence and Uniqueness of Solutions to Ordinary Differential Equations

IX.1.1. Elementary Considerations and Terminology

IX.1.1.1. It is customary (although by no means universal) in the theory of ODEs to denote the independent variable by $t$, since in many applications the independent variable represents time. The unknown function is typically denoted $x$ or $y$; we will standardize on using $y$ as being less confusing to those of us used to using $x$ as an independent variable. While $t$ is always assumed to be a real variable, usually ranging over an interval, $y$ can be real-valued or vector-valued (we will use $x$’s to denote elements of the range space and $y$’s to denote functions from $\mathbb{R}$ to the range space).

IX.1.1.2. An ODE is thus an equation involving $t, y, y', \ldots, y^{(n)}$. Such an equation is called an ordinary differential equation of order $n$ (most differential equations arising in physics and other applications have order 1 or 2). For the most part, we consider only differential equations of the form

$$y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t))$$

which is an abbreviation for

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$$

where $f$ is a specified real-valued (or, more generally, vector-valued) function with reasonable properties (usually at least continuous) and the equation is to hold for all $t$ in a suitable subset of $\mathbb{R}$. Most reasonable ODEs can in principle be put into this form using the Implicit Function Theorem (Exercise IX.1.9.2.), and most ODE’s arising in applications are already of this form or can easily be put into it explicitly (with the exception of certain linear ODEs; cf. ()).

If $f$ is a function of only $y, y', \ldots, y^{(n-1)}$ (i.e. not explicitly of $t$), the differential equation is called autonomous; when $t$ is interpreted as time, an autonomous differential equation is “time-independent.”

Many, but by no means all, differential equations in applications are autonomous.

A solution $y$ to an $n$’th order ODE is necessarily a function which is $n$-times differentiable; if $f$ is continuous, $y$ must be $C^n$ since the right side will be a continuous function of $t$. If $f$ is $C^1$, then $y^{(n)}$ is $C^1$ by the Chain Rule, so $y$ is $C^{(n+1)}$.

IX.1.1.3. We almost always want to consider only solutions to differential equations on intervals, generally open intervals; sometimes solutions given by formulas have domains consisting of a disjoint union of open intervals, in which case the restriction to each of the subintervals is considered a separate solution. Most commonly we want a solution $y(t)$ on an interval $I$ taking a specified value $x_0$ at a specified number $t_0 \in I$ (our use of $x_0$ for the initial value instead of $y_0$ is nonstandard, but the reasons for using this notation will be apparent later, and it is consistent with our convention described in IX.1.1.1.). A differential equation with a specified value at an “initial” point $t_0$ is called an initial value problem or IVP (for a higher-order differential equation, initial values for lower-order derivatives are also needed, or values at more than one point, called a boundary value problem or BVP). While differential equations typically have many solutions, with various intervals of definition, in reasonable cases initial value problems have unique solutions.

IX.1.1.4. In fact, the general principle of the theory of differential equations is that, subject to some mild technical restrictions, the general solution to an $n$’th order differential equation is a family of functions
determined by \( n \) more or less independent constants (parameters), which can be expressed in various ways. (Conversely, a smooth family of functions determined by \( n \) independent parameters is the general solution set to an \( n' \)th order differential equation, again subject to mild technical restrictions, cf. Exercise IX.1.9..)

However, it is usually impossible to find an explicit or closed formula for the general solution to even simple differential equations. Implicit relations can be found in somewhat greater generality, but even these can be difficult or impossible to come by in many cases. Power-series solutions can often be found if the differential equation is sufficiently regular. But in many cases the best one can do is generate approximate numerical solutions, and there are now powerful and efficient techniques for doing this.

**IX.1.1.5. Examples.** Here are some simple examples of first- and second-order differential equations that illustrate the general nature of the subject.

(i) Let \( f \) be a real-valued continuous function on an interval \( I \). A solution to the first-order differential equation

\[
y' = f(t)
\]

is simply an antiderivative for \( f \) on \( I \). There is thus a solution by V.8.5.9., in fact infinitely many, all differing by constants. If \( t_0 \in I \) is fixed and \( x_0 \) is a fixed real number, then there is a unique solution \( y \) on \( I \) satisfying the differential equation (i.e. an antiderivative for \( f \) on \( I \)) and also the initial condition \( y(t_0) = x_0 \).

(ii) Consider the autonomous first-order differential equation

\[
y' = ky
\]
on \( \mathbb{R} \), where \( k \in \mathbb{R} \) is a constant. A solution is \( y(t) = Ae^{kt} \), where \( A \) is an arbitrary real constant. To see that these are all the solutions, suppose \( y \) is a solution which is not identically 0, and fix \( t_0 \) with \( y(t_0) \neq 0 \). Then \( y \) is nonzero in an interval around \( t_0 \), and we have, for \( t \) in this interval,

\[
\frac{d}{dt}(\log |y(t)|) = \frac{y'(t)}{y(t)} = k
\]

so \( |y(t)| = kt + C \) for some constant \( C \). Thus \( |y(t)| = e^{C}e^{kt} \) for some \( C \), at least for \( t \) near enough to \( t_0 \). But it follows easily that \( y(t) = \pm e^{C}e^{kt} \) for \( t \) near \( t_0 \) (with constant sign depending on the sign of \( y(t_0) \)), and this solution extends uniquely to a solution on all of \( \mathbb{R} \). If we specify the initial condition \( y(t_0) = x_0 \), we have a unique solution with \( A = x_0e^{-kt_0} \), valid on all of \( \mathbb{R} \).

(iii) Let \( k > 0 \) be a constant, and consider the autonomous second-order differential equation

\[
y'' = k^2y
\]
The function

\[
y(t) = Ae^{kt} + Be^{-kt}
\]
is a solution on all of \( \mathbb{R} \) for arbitrary real constants \( A \) and \( B \). The differential equation

\[
y'' = -k^2y
\]
has as solution

\[
y(t) = A \cos kt + B \sin kt
\]
for constants $A$, $B$. It can be shown that in both cases, every solution is of this form ($\cdot$). Simple calculations show that if we fix $t_0 \in \mathbb{R}$ and real constants $x_0$ and $x'_0$, then in either case there is a unique solution (i.e. unique choice of $A$ and $B$) with $y(t_0) = x_0$ and $y'(t_0) = x'_0$.

Here are some more examples which illustrate other characteristics of solutions to differential equations:

**IX.1.1.6. Examples.** (i) Consider the autonomous first-order differential equation

$$y' = |y|.$$ 

The constant function $f_0(t) = 0$ is a solution, as is $f_A(t) = Ae^t$ for $A > 0$ and $f_A(t) = Ae^{-t}$ for $A < 0$. It can be shown (Exercise ($\cdot$)) that every solution on an interval $I$ is the restriction of one of these solutions to $I$. These solutions are $C^\infty$ on all of $\mathbb{R}$ despite the fact that $|y|$ is not a differentiable function of $y$. It is easily checked that for every $t_0$ and $x_0$ there is a unique choice of $A \in \mathbb{R}$ for which $f_A$ is a solution to the IVP

$$y' = |y|, \quad y(t_0) = x_0.$$ 

The function $f(t, y) = |y|$ is continuous on $\mathbb{R}^2$; although it is not differentiable in $y$ for $y = 0$, it satisfies a uniform Lipschitz condition in $y$: $|f(t, y_2) - f(t, y_1)| \leq |y_2 - y_1|$ for all $t, y_1, y_2 \in \mathbb{R}$.

(ii) Consider the autonomous first-order differential equation

$$y' = y^2.$$ 

The constant function 0 is a solution, as is

$$f_c(t) = \frac{1}{c-t}$$

for $c \in \mathbb{R}$. It can be shown (Exercise ($\cdot$)) that every solution on an interval $I$ is the restriction of one of these solutions to $I$. For any $t_0$ and $x_0$, the initial-value problem

$$y' = y^2, \quad y(t_0) = x_0$$

has a unique solution in an interval around $t_0$: if $x_0 = 0$, the solution is the constant function 0, and if $x_0 \neq 0$, the solution is the restriction of $f_c$, where $c = t_0 + \frac{1}{x_0}$. This solution is only defined on the interval ($-\infty, c$) or ($c, +\infty$) around $t_0$, not on all of $\mathbb{R}$. The (maximum) domain of the solution depends on $t_0$ and $x_0$.

(iii) Consider the autonomous first-order differential equation

$$y' = \frac{3}{2}y^{1/3}$$

(the constant $\frac{3}{2}$ just makes subsequent calculations easy). The constant function 0 is a solution, as are

$$\phi_c^+(t) = \begin{cases} 0 & \text{if } t < c \\ (t - c)^{3/2} & \text{if } t \geq c \end{cases}$$

$$\phi_c^-(t) = \begin{cases} 0 & \text{if } t < c \\ -(t - c)^{3/2} & \text{if } t \geq c \end{cases}$$
for any \( c \in \mathbb{R} \). Note that these solutions are defined and \( C^1 \) (but not \( C^2 \)) on all of \( \mathbb{R} \). It can be shown (Exercise ()) that every solution on an interval \( I \) is the restriction of one of these solutions to \( I \). Note that \( f(t, y) = \frac{3}{2} y^{1/3} \) is continuous on \( \mathbb{R}^2 \), but does not satisfy a Lipschitz condition in \( y \) around \( y = 0 \).

If \( t_0, x_0 \in \mathbb{R} \), the initial-value problem

\[
y' = \frac{3}{2} y^{1/3}, \quad y(t_0) = x_0
\]

always has a solution. If \( x_0 \neq 0 \), the unique solution is \( \phi_c^\pm \), where \( c = t_0 - x_0^{2/3} \) and the sign is the sign of \( x_0 \). If \( x_0 = 0 \), there are many solutions: the constant function 0 is a solution, as are \( \phi_c^+ \) and \( \phi_c^- \) for any \( c \geq t_0 \). The solutions 0, \( \phi_{t_0}^+ \), and \( \phi_{t_0}^- \) do not agree on any interval around \( t_0 \), so the solutions are not locally unique.

(iv) Consider the autonomous first-order differential equation

\[
y' = 3y^{2/3}.
\]

The constant function 0 is a solution, as are

\[
\phi_c(t) = (t - c)^3
\]

\[
\psi_c^+(t) = \begin{cases} 0 & \text{if } t < c \\ (t - c)^3 & \text{if } t \geq c \end{cases}
\]

\[
\psi_c^-(t) = \begin{cases} (t - c)^3 & \text{if } t < c \\ 0 & \text{if } t \geq c \end{cases}
\]

for any \( c \in \mathbb{R} \), and

\[
\psi_{c,d}(t) = \begin{cases} (t - c)^3 & \text{if } t < c \\ 0 & \text{if } c \leq t \leq d \\ (t - d)^3 & \text{if } t > d \end{cases}
\]

for any \( c, d \in \mathbb{R} \), \( c < d \) (we could write \( \psi_{c,c} = \phi_c \)). Note that these solutions are defined on all of \( \mathbb{R} \); \( \phi_c \) is \( C^\infty \) (even analytic) on \( \mathbb{R} \), and \( \psi_c^+ \) and \( \psi_{c,d} \) for \( c < d \) are \( C^2 \) but not \( C^3 \) on \( \mathbb{R} \). It can be shown (Exercise ()) that every solution on an interval \( I \) is the restriction of one of these solutions to \( I \). Note that \( f(t, y) = 3y^{2/3} \) is continuous on \( \mathbb{R}^2 \), but does not satisfy a Lipschitz condition in \( y \) around \( y = 0 \).

If \( t_0, x_0 \in \mathbb{R} \), the initial-value problem

\[
y' = 3y^{2/3}, \quad y(t_0) = x_0
\]

always has a solution. If \( x_0 \neq 0 \), there is locally a unique solution \( \phi_c \), where \( c = t_0 - x_0^{1/3} \). However, this solution is not globally unique: \( \psi_c^\pm \) (where the sign is the sign of \( x_0 \)) is also a solution, as is \( \psi_{c,d} \) for any \( d > c \) if \( x_0 < 0 \) and \( \psi_{d,c} \) for any \( d < c \) if \( x_0 > 0 \). If \( x_0 = 0 \), there are many solutions: the constant function 0 is a solution, as are \( \phi_{t_0}^+ \), \( \psi_{t_0}^+ \) for any \( c \geq t_0 \), \( \psi_{t_0}^- \) for any \( c \leq t_0 \), and \( \psi_{c,d} \) for \( c \leq t_0 \leq d \). The solutions 0, \( \phi_{t_0}^+ \), \( \psi_{t_0}^+ \), and \( \psi_{t_0}^- \) do not agree on any interval around \( t_0 \), so the solutions are not locally unique.
IX.1.2. Existence and Uniqueness Theorem for First-Order ODEs

In this subsection, we prove a general theorem about existence and uniqueness of local solutions to a first-order initial-value problem

\[ y' = f(t, y), \quad y(t_0) = x_0 \]

where \( y \) is an unknown function from an interval \( I \) in \( \mathbb{R} \) to a (real) Banach space \( \mathcal{X} \). In most elementary applications \( \mathcal{X} \) will be \( \mathbb{R} \), or \( \mathbb{R}^n \) for some \( n \). This fundamental theorem is closely related to the Implicit and Inverse Function Theorems, at least in spirit (cf. IX.1.9.10.). In particular, it is a local result, but we have control over the size of the interval on which the solution is valid, which in some key special cases gives global solutions. This result was first obtained by E. Picard in 1890; our exposition is adapted from [?]. The result is called the Picard Existence Theorem, Picard-Lindelöf Theorem, or Cauchy-Lipschitz Theorem in various references (the Cauchy-Lipschitz Theorem is properly the crucial special case IX.1.2.7., obtained before Picard’s general result).

IX.1.2.1. The setup of this theorem is as follows. Let \( \mathcal{X} \) be a Banach space and \( x_0 \in \mathcal{X} \). Let \( \tilde{B}_r(x_0) \) be the closed ball in \( \mathcal{X} \) centered at \( x_0 \) with radius \( r > 0 \). Let \( t_0 \in \mathbb{R} \), and \( I \) an interval in \( \mathbb{R} \) with \( t_0 \in I \). Let \( f \) be a continuous function from \( I \times \tilde{B}_r(x_0) \) to \( \mathcal{X} \). We suppose in addition there is a constant \( M \) such that \( \|f(t, x)\| \leq M \) for all \((t, x) \in I \times \tilde{B}_r(x_0)\), and that \( f \) satisfies a uniform Lipschitz condition in \( x \) (or, as we usually write, in \( y \)) on \( I \times \tilde{B}_r(x_0) \), i.e. there is a constant \( K \) such that for every \( t \in I \) and \( x_1, x_2 \in \tilde{B}_r(x_0) \) we have

\[ \|f(t, x_2) - f(t, x_1)\| \leq K \|x_2 - x_1\| . \]

We seek an interval \( I_0 \subseteq I \), \( t_0 \in I_0 \), and a differentiable function \( y : I_0 \rightarrow \mathcal{X} \) such that \( y'(t) = f(t, y(t)) \) for \( t \in I_0 \) and \( y(t_0) = x_0 \), and to show that any two such functions agree on \( I_0 \).

IX.1.2.2. Theorem. [Existence and Uniqueness Theorem for First-Order ODEs] Using the above notation, set

\[ I_0 = I \cap \left[ t_0 - \frac{r}{M}, t_0 + \frac{r}{M} \right] . \]

Then there is a unique differentiable function \( y : I_0 \rightarrow \mathcal{X} \) such that \( y'(t) = f(t, y(t)) \) for \( t \in I_0 \) and \( y(t_0) = x_0 \). Additionally, \( y \) is \( C^1 \) on \( I_0 \). If \( \bar{I} \) is any subinterval of \( I_0 \) containing \( t_0 \), the restriction of \( y \) to \( \bar{I} \) is the unique solution to the IVP on \( \bar{I} \).

Proof: It suffices to show there is a unique solution on any closed subinterval \( J \) of \( I_0 \) containing \( t_0 \). The strategy of the proof is to use the (Extended) Banach Fixed-Point Theorem (VI.2.9.10.) to obtain the solution on \( J \) as the unique fixed point of a suitable strict power contraction on a complete metric space. To this end, set \( \epsilon = \frac{r}{4M} \) and let \( Y \) be the set of continuous functions from \( J \) to the closed ball \( \tilde{B}_r(x_0) \) in \( \mathcal{X} \) of radius \( r \) centered at \( x_0 \). Define a metric \( \rho \) on \( Y \) by

\[ \rho(\phi, \psi) = \max_{t \in J} \|\phi(t) - \psi(t)\| . \]

Then \( Y \) is a complete metric space under \( \rho \) (it is a closed subset of \( C(J, \mathcal{X}) \)). Define a map \( T : Y \rightarrow Y \) by

\[ [T(\phi)](t) = x_0 + \int_{t_0}^t f(u, \phi(u)) \, du . \]
It is obvious from (i) that $T(\phi)$ is a continuous function from $J$ to $\mathcal{X}$. We must show that $T(\phi) \in Y$ if $\phi \in Y$, and that there is an $n$ and a constant $\alpha < 1$ such that $\rho(T^n(\phi), T^n(\psi)) \leq \alpha \rho(\phi, \psi)$ for all $\phi, \psi \in Y$.

We have, for $t \in J$ and $\phi \in Y$,

$$\|T(\phi)(t) - x_0\| = \left\| \int_{t_0}^{t} f(u, \phi(u)) \, du \right\| \leq \int_{t_0}^{t} \|f(u, \phi(u))\| \, du \leq M|t - t_0| \leq M \epsilon = r$$

so $T(\phi) \in Y$. And we have, for $\phi, \psi \in Y$ and $t \in J$,

$$\|T(\phi)(t) - T(\psi)(t)\| = \left\| \int_{t_0}^{t} [f(u, \phi(u)) - f(u, \psi(u))] \, du \right\|$$

$$\leq \int_{t_0}^{t} \|f(u, \phi(u)) - f(u, \psi(u))\| \, du \leq K \int_{t_0}^{t} \|\phi(u) - \psi(u)\| \, du$$

$$\leq K \epsilon \max_{u \in J} \|\phi(u) - \psi(u)\| = K \epsilon \rho(\phi, \psi).$$

Similarly, for each $n$ we have

$$\|T^n(\phi) - T^n(\psi)\| \leq K \left| \int_{t_0}^{t} \|T^{n-1}(\phi)(u_1) - T^{n-1}(\psi)(u_1)\| \, du_1 \right|$$

$$\leq K^2 \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \|T^{n-2}(\phi)(u_2) - T^{n-2}(\psi)(u_2)\| \, du_2 \, du_1 \right|$$

$$\leq \cdots \leq K^n \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \cdots \int_{t_0}^{u_{n-1}} \|\phi(u_n) - \psi(u_n)\| \, du_n \cdots du_1 \right|$$

$$\leq K^n \rho(\phi, \psi) \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \cdots \int_{t_0}^{u_{n-1}} \, du_n \cdots du_1 \right| = \frac{K^n |t - t_0|^n}{n!} \rho(\phi, \psi) \leq \frac{(K \epsilon)^n}{n!} \rho(\phi, \psi).$$

Since $t$ is arbitrary, we have

$$\rho(T^n(\phi), T^n(\psi)) \leq \frac{(K \epsilon)^n}{n!} \rho(\phi, \psi).$$

If $n$ is sufficiently large, then

$$\frac{(K \epsilon)^n}{n!} < 1$$

and thus $T^n$ is a strict contraction from $Y$ to $Y$.

Thus, by VI.2.9.10., $T$ has a unique fixed point $y$. We thus have

$$y(t) = t_0 + \int_{t_0}^{t} f(u, y(u)) \, du$$

for $t \in J$, and since $g(u) = f(u, y(u))$ is continuous, by (i) we have that $y'(t) = f(t, y(t))$ for $t \in J$. Also obviously $y(t_0) = x_0$. Conversely, any solution on $J$ must be a fixed point of $T$, and thus $y$ is the unique solution on $J$.

If $J'$ is a closed subinterval of $I_0$ containing $t_0$ different from $J$, say $J \subseteq J'$, the unique solution on $J'$ restricted to $J$ must be the unique solution on $J$.

Since $g(t) = f(t, y(t))$ is continuous, $y'$ is continuous, i.e. $y$ is $C^1$. \nonumber
IX.1.2.3.  The restriction of the domain of the solution to \([t_0 - \frac{r}{M}, t_0 + \frac{r}{M}]\) simply insures that the unique solution near \(t_0\), when continued to the endpoints of the interval, does not leave the range of \(y\)-values for which \(f\) is defined (the restriction is a real one even if \(r\) can be chosen arbitrarily large; cf. Example IX.1.1.6.(ii)); the constant \(\frac{r}{M}\) is conservative in general (cf. Exercise IX.1.9.6.). Note that the size of the Lipschitz constant \(K\) does not affect the length of the interval on which the existence of a unique solution is guaranteed by the theorem. Note also that if \(t_0\) is in the interior of \(I\) (the usual situation in which the theorem is applied), it is also in the interior of \(I_0\).

IX.1.2.4.  The proof of Theorem IX.1.2.2. is constructive: it not only shows the existence of a unique solution, but actually gives an iterative algorithm for finding it. In fact, if \(\phi\) is any continuous function from \(I_0\) to \(X\) with \(\phi(t_0) = x_0\) (e.g. the constant function \(x_0\)), then the sequence \((T^n(\phi))\) converges uniformly on \(I_0\) to the solution. This “method of successive approximations,” called Picard iteration, is a practical computational method for finding approximations to the solution in many cases, although there are usually better numerical techniques.

IX.1.2.5.  In fact, we have the following error estimate for the successive approximation: if \(\phi_0\) is the constant function \(x_0\) and \(\phi_n = T^n(\phi_0)\) for \(n \in \mathbb{N}\), then, for \(t \in I_0\),

\[
\|\phi_0(t) - \phi_1(t)\| = \left\| \int_{t_0}^{t} f(u, y_0) \, du \right\| \leq \left\| \int_{t_0}^{t} \|f(t, y_0)\| \, du \right\| \leq M|t - t_0|
\]

and by induction

\[
\|\phi_n(t) - \phi_{n+1}(t)\| \leq \frac{MK^n}{(n+1)!}|t - t_0|^{n+1} \leq \frac{M (Kr/M)^{n+1}}{K (n+1)!}
\]

\[
\|\phi_n(t) - y(t)\| \leq \frac{M}{K} \sum_{k=n+1}^{\infty} \frac{(Kr/M)^k}{k!}
\]

The sum is the error in the \(n\)’th Taylor approximation to \(e^{Kr/M}\), hence by ()

\[
\rho(\phi_n, y) \leq \frac{Me^{Kr/M}}{K(n+1)!} .
\]

Thus the \(\phi_n\) converge uniformly quite rapidly to \(y\).

We then get a local version with a cleaner statement but less precise conclusion:

IX.1.2.6.  Corollary.  [Existence and Uniqueness Theorem for First-Order ODEs, Local Version] Let \(\mathcal{X}\) be a Banach space, \(x_0 \in \mathcal{X}, t_0 \in \mathbb{R}\), and \(f(t, y)\) a continuous function from a neighborhood \(U\) of \((t_0, x_0)\) in \(\mathbb{R} \times \mathcal{X}\) to \(\mathcal{X}\). Suppose \(f\) satisfies a uniform Lipschitz condition in \(y\) on \(U\). Then there is a \(\delta > 0\) such that the first-order initial value problem

\[
y' = f(t, y), \quad y(t_0) = x_0
\]

has a unique solution on \([t_0 - \delta, t_0 + \delta]\). Additionally, \(y\) is \(C^1\) on \([t_0 - \delta, t_0 + \delta]\).
Proof: An $r$ and a closed bounded interval $I$ containing $t_0$ in its interior can be chosen such that $I \times \bar{B}_r(x_0)$ is contained in $U$. By compactness, $f(t,x_0)$ is bounded, say by $N$, on $I$. If $K$ is the Lipschitz constant, then for $y \in \bar{B}_r(x_0)$ we have
\[
\|f(t,y)\| \leq \|f(t,x_0)\| + \|f(t,y) - f(t,x_0)\| \leq N + Kr
\]
so $f$ is bounded on $I \times \bar{B}_r(y_0)$. Apply the theorem with this $I$ and $r$.

As a special case, we get the following version which covers most standard applications. This is the version commonly stated and (sometimes) proved in Differential Equations texts.

**IX.1.2.7. Corollary. [Existence and Uniqueness Theorem for First-Order ODEs, Standard Version]** Let $R$ be a rectangle in $\mathbb{R}^2$, and $(t_0,x_0)$ an interior point of $R$. Let $f(t,y)$ be a function from $R$ to $\mathbb{R}$ with $f$ and $\frac{\partial f}{\partial y}$ continuous on $R$. Then there is a $\delta > 0$ such that the initial-value problem
\[
y' = f(t,y), \quad y(t_0) = x_0
\]
has a unique solution on $(t_0 - \delta, t_0 + \delta)$. Additionally, $y$ is $C^1$ on $(t_0 - \delta, t_0 + \delta)$.

Proof: Since $\frac{\partial f}{\partial y}$ is continuous, it is bounded in a neighborhood of $(t_0,x_0)$, and thus by the MVT (V.8.2.11.) $f$ satisfies a uniform Lipschitz condition in $y$ on this neighborhood.

Using the theorem, one can get an explicit estimate for the size $\delta$ can be. Actually $R$ can be any neighborhood of $(t_0,x_0)$.

**Global Solutions**

**IX.1.2.8.** The above results only give a solution in a (usually small) interval around $t_0$. Even if $f$ is defined on all of $I \times \mathcal{X}$, i.e. the hypotheses of IX.1.2.2. are satisfied for all $r$ (with the Lipschitz constant $K$ depending on $r$), Example IX.1.1.6.(ii) shows that the solution may not be defined on all of $I$. The maximum interval on which the solution is defined can even depend on the choice of $x_0$ (this might be predicted since the constant $M$ can depend on $x_0$ as well as on $r$).

But if $f$ satisfies a uniform Lipschitz condition in $y$ on all of $I \times \mathcal{X}$, the control on the domain in IX.1.2.2. allows us to cobble together local solutions into a solution defined on all of $I$.

**IX.1.2.9. Theorem. [Existence and Uniqueness Theorem for First-Order ODEs, Global Version]** Let $\mathcal{X}$ be a Banach space, $I$ an interval in $\mathbb{R}$, and $t_0 \in I$. Let $f(t,y)$ be a continuous function from $I \times \mathcal{X}$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in $y$ on $I$. Then, for any $x_0 \in \mathcal{X}$, the initial-value problem
\[
y' = f(t,y), \quad y(t_0) = x_0
\]
has a unique solution defined on all of $I$. Additionally, the solution is $C^1$ on $I$.

Proof: It suffices to show that if $J = [a,b]$ is any closed bounded subinterval of $I$ containing $t_0$ in its interior, then there is a unique solution to the IVP $y' = f(t,y)$, $y(t_0) = x_0$ on $J$. Fix such a $J$. If $K$ is the Lipschitz constant, also fix $L$ with $0 < L < \frac{1}{K}$. 885
First let $t_1 \in J$ and $x_1 \in \mathcal{X}$, and consider the IVP $y' = f(t, y)$, $y(t_1) = x_1$. The function $g(t) = f(t, x_1)$ is bounded on $J$ by compactness, say by $N$. Then for any $r$, $f$ is bounded on $J \times B_r(x_1)$ by $N + Kr$ (cf. the proof of IX.1.2.6.). Thus one bound on the length of the interval on which a unique solution is guaranteed by Theorem IX.1.2.2. is $\frac{r}{N + Kr}$. If $r$ is taken large enough, this is greater than $L$. Thus Theorem IX.1.2.2. gives a unique solution to this IVP on $J \cap [t_1 - L, t_1 + L]$.

Now return to the original IVP $y' = f(t, y)$, $y(t_0) = x_0$. Theorem IX.1.2.2. gives a unique solution $y = \phi_0(t)$ on $J_0 = J \cap [t_0 - L, t_0 + L]$. We now successively extend this unique solution to the right until it extends to $b$. If $t_0 + L \geq b$, we are done. Otherwise, fix $t_1$ with $t_0 + \frac{L}{2} < t_1 < t_0 + L$. Set $x_1 = \phi_0(t_1)$. Theorem IX.1.2.2. gives a unique solution $y = \phi_1(t)$ to the IVP $y' = f(t, y)$, $y(t_1) = x_1$ on $J_1 = J \cap [t_1 - L, t_1 + L]$. We must have $\phi_1 = \phi_0$ on $J_0 \cap J_1$ since both are solutions to the IVP $y' = f(t, y)$, $y(t_1) = x_1$ on $J_0 \cap J_1$, and thus we have extended $\phi_0$ to a solution on $J_0 \cup J_1$.

If $t_1 + L \geq b$, we are done extending to the right. If not, we have $t_1 + L \geq (t_0 + L) + \frac{L}{2}$, i.e. we have extended to the right by at least $\frac{L}{2}$. Repeat the process, choosing $t_2$ with $t_1 + \frac{L}{2} < t_2 < t_1 + L$, $x_2$, $J_2$, $\phi_2$, etc. Since at each stage we extend to the right by at least $\frac{L}{2}$, we reach $b$ in a finite number of steps.

Extend $\phi_0$ to the left of $t_0$ to $a$ in the same manner. We obtain a solution $\phi$ on $J$ in this manner. Any solution to the original IVP must agree with $\phi$ on a subset of $J$ which is both relatively open and closed, hence must agree with $\phi$ everywhere on $J$.

### IX.1.3. Systems of First-Order Differential Equations

The result IX.1.2.2. is general enough to immediately give solutions not only to a single differential equation, but also to systems of differential equations: we want differentiable functions $y_1, \ldots, y_n$ on an interval $I$ satisfying the $n$ equations

$$ y_k' = f_k(t, y_1, \ldots, y_n) \quad (1 \leq k \leq n). $$

We also want the $y_k$ to have specified values at some $t_0 \in I$.

#### IX.1.3.1. The setup is similar to the one for single equations, but we change notation slightly. We have a Banach space $\mathcal{X}$, an interval $I \subseteq \mathbb{R}$, a point $t_0 \in I$, $n$ points $x_0(1), \ldots, x_0(n)$ in $\mathcal{X}$, $r > 0$, $M > 0$, $K > 0$, and $n$ continuous functions

$$ f_1, \ldots, f_n : I \times B_r(x_0(1)) \times \cdots \times B_r(x_0(n)) \to \mathcal{X} $$

with the property that $\|f_k(t, x_1, \ldots, x_n)\| \leq M$ for $t \in I$ and $\|x_j - x_j(0)\| \leq r$ (1 \leq j, k \leq n), and

$$ \|f_k(t, x_1^{(2)}, \ldots, x_n^{(2)}) - f_k(t, x_1^{(1)}, \ldots, x_n^{(1)})\| \leq K \max_{1 \leq j \leq n} \|x_j^{(2)} - x_j^{(1)}\| $$

for $1 \leq k \leq n$ whenever $t \in I$ and $x_j^{(1)}$, $x_j^{(2)} \in B_r(x_j(0))$ for all $j$ (the $f_k$ satisfy a uniform Lipschitz condition in $x_1, \ldots, x_n$ on $I$).

#### IX.1.3.2. Theorem. [Existence and Uniqueness Theorem for Systems of First-Order ODEs] Using the above notation, set

$$ I_0 = I \cap \left[ t_0 - \frac{r}{M}, t_0 + \frac{r}{M} \right]. $$

886
Then there are unique differentiable functions \( y_1, \ldots, y_n : I_0 \to \mathcal{X} \) such that
\[
y'_k(t) = f_k(t, y_1(t), \ldots, y_n(t))
\]
(\( 1 \leq k \leq n \)) for \( t \in I_0 \) and \( y_k(t_0) = x_k^{(0)} \). Additionally, the \( y_k \) are \( C^1 \) on \( I_0 \). If \( \tilde{I} \) is any subinterval of \( I_0 \) containing \( t_0 \), the restriction of the \( y_k \) to \( \tilde{I} \) is the unique solution to the IVP on \( \tilde{I} \).

**Proof:** This is an almost immediate corollary of IX.1.2.2. Set \( \tilde{X} = X^n \) with the \( \infty \)-norm \( (\cdot) \), and \( x_0 = (x_1^{(0)}, \ldots, x_n^{(0)}) \). Define \( f : I \times \tilde{B}_r(x_0) \to \tilde{X} \) by
\[
f(t, (x_1, \ldots, x_n)) = (f_1(t, x_1, \ldots, x_n), \ldots, f_n(t, x_1, \ldots, x_n))
\]
Then the hypotheses of IX.1.2.2. are satisfied, and the coordinate functions of the solution \( y \) give the \( y_k \).

As in the single equation case, we get a local version with a cleaner statement but less precise conclusion:

**IX.1.3.3. Corollary.** [Existence and Uniqueness Theorem for Systems of First-Order ODEs, Local Version] Let \( \mathcal{X} \) be a Banach space, \( x_1^{(0)}, \ldots, x_n^{(0)} \in \mathcal{X}, t_0 \in \mathbb{R} \), and \( f_1, \ldots, f_n \) continuous functions from a neighborhood \( U \) of \( (t_0, x_1^{(0)}, \ldots, x_n^{(0)}) \) in \( \mathbb{R} \times \mathcal{X}^n \) to \( \mathcal{X} \). Suppose each \( f_k \) satisfies a uniform Lipschitz condition in \( x_1, \ldots, x_n \) on \( U \). Then there is a \( \delta > 0 \) such that the first-order initial value problem
\[
y'_k = f_k(t, y_1, \ldots, y_n), \quad y_k(t_0) = x_k^{(0)} \quad (1 \leq k \leq n)
\]
has a unique solution on \( (t_0 - \delta, t_0 + \delta) \). Additionally, the \( y_k \) are \( C^1 \) on \( (t_0 - \delta, t_0 + \delta) \).

We also get a global version from IX.1.2.9.:

**IX.1.3.4. Theorem.** [Existence and Uniqueness Theorem for Systems of First-Order ODEs, Global Version] Let \( \mathcal{X} \) be a Banach space, \( I \) an interval in \( \mathbb{R} \), and \( t_0 \in I \). Let \( f_1, \ldots, f_n \) be continuous functions from \( I \times \mathcal{X}^n \) to \( \mathcal{X} \) satisfying a uniform Lipschitz condition in \( x_1, \ldots, x_n \) on \( I \). Then, for any \( x_1^{(0)}, \ldots, x_n^{(0)} \in \mathcal{X} \), the initial-value problem
\[
y'_k = f_k(t, y_1, \ldots, y_n), \quad y_k(t_0) = x_k^{(0)} \quad (1 \leq k \leq n)
\]
has a unique solution defined on all of \( I \). Additionally, the solutions are \( C^1 \) on \( I \).
IX.1.4. Higher-Order Differential Equations and Systems

We can use the results for systems of first-order equations to obtain similar existence and uniqueness theorems for higher-order differential equations, by a simple but clever trick. We want to solve the initial-value problem

\[ y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x'_0, \ldots, y^{(n-1)}(t_0) = x^{(n-1)}_0 \]

where \( y \) is an \( n \)-times differentiable function from an interval in \( \mathbb{R} \) to a Banach space \( \mathcal{Y} \) and \( f \) is a suitably nice function from a neighborhood of \( (t_0, x_0, x'_0, \ldots, x^{(n-1)}_0) \) in \( \mathbb{R} \times \mathcal{X}^n \) to \( \mathcal{X} \).

The trick is to define new unknown functions \( y_1, \ldots, y_{n-1} \) which satisfy the system of equations

\[ y' = y_1, y'_1 = y_2, \ldots, y'_{n-2} = y_{n-1}, y'_{n-1} = f(t, y_1, \ldots, y_{n-1}) \]

so \( y_k = y^{(k)} \). Then we have a system of \( n \) first-order equations in \( n \) unknown functions \( y, y_1, \ldots, y_{n-1} \), and we can apply the various existence and uniqueness theorems for such systems. The statements are:

IX.1.4.1. Theorem. [Existence and Uniqueness Theorem for Higher-Order ODE’s] Let \( \mathcal{X} \) be a Banach space, \( I \) an interval in \( \mathbb{R} \), \( t_0 \in I \), \( x_0, x'_0, \ldots, x^{(n-1)}_0 \in \mathcal{X} \), \( r > 0 \), and

\[ f : I \times \bar{B}_r(x_0) \times \bar{B}_r(x'_0) \times \cdots \times \bar{B}_r(x^{(n-1)}_0) \to \mathcal{X}. \]

Suppose \( f \) is continuous and bounded by \( N \) on \( I \times \bar{B}_r(x_0) \times \bar{B}_r(x'_0) \times \cdots \times \bar{B}_r(x^{(n-1)}_0) \) and satisfies a uniform Lipschitz condition in the \( y \)-variables. Set

\[ M = \max\{N, \|y_0\| + r, \ldots, \|y^{(n-1)}_0\| + r\} \]

\[ I_0 = I \cap \left[ t_0 - \frac{r}{M}, t_0 + \frac{r}{M} \right]. \]

Then there is a unique \( n \)-times differentiable function \( y : I_0 \to \mathcal{X} \) with \( y(t_0) = x_0, y'(t_0) = x'_0, \ldots, y^{(n-1)}(t_0) = x^{(n-1)}_0 \), and \( y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t)) \) for all \( t \in I_0 \). Additionally, \( y \) is \( C^n \) on \( I_0 \).

Proof: Define \( f_k(t, x_0, x_1, \ldots, x_{n-1}) = x_{k+1} \) for \( 0 \leq k < n - 1 \) and \( f_{n-1} = f \), and apply IX.1.3.2. \( \diamond \)

The interval on which the existence of a unique solution is guaranteed can sometimes be shown to be larger than \( I_0 \); see Exercise IX.1.9.8.

IX.1.4.2. Corollary. [Existence and Uniqueness Theorem for Higher-Order ODEs, Local Version] Let \( \mathcal{X} \) be a Banach space, \( x_0, x'_0, \ldots, x^{(n-1)}_0 \in \mathcal{X} \), \( t_0 \in \mathbb{R} \), and \( f \) a continuous function from a neighborhood \( U \) of \( (t_0, x_0, x'_0, \ldots, x^{(n-1)}_0) \) in \( \mathbb{R} \times \mathcal{X}^n \) to \( \mathcal{X} \). Suppose \( f \) satisfies a uniform Lipschitz condition in the \( y \)-variables on \( U \). Then there is a \( \delta > 0 \) such that the first-order initial value problem

\[ y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x'_0, \ldots, y^{(n-1)}(t_0) = x^{(n-1)}_0 \]

has a unique solution on \( (t_0 - \delta, t_0 + \delta) \). Additionally, \( y \) is \( C^n \) on \( (t_0 - \delta, t_0 + \delta) \).
IX.1.4.3. Theorem. [Existence and Uniqueness Theorem for Higher-Order ODEs, Global Version] Let $\mathcal{X}$ be a Banach space, $I$ an interval in $\mathbb{R}$, and $t_0 \in I$. Let $f$ be a continuous function from $I \times \mathcal{X}^n$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in the $y$-variables on $I$. Then, for any $x_0, x'_0, \ldots, x_0^{(n-1)} \in \mathcal{X}$, the initial-value problem

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x'_0, \ldots, y^{(n-1)}(t_0) = x_0^{(n-1)}$$

has a unique solution defined on all of $I$. Additionally, the solution is $C^n$ on $I$.

IX.1.4.4. There are similar extensions to systems of higher-order differential equations. The equations in the system do not all need to be of the same order. The extensions are straightforward, but the notation is oppressive, and it is left to the reader to formulate and prove the results.

IX.1.5. Continuous Dependence on Initial Conditions

IX.1.5.1. We now examine what happens when the initial conditions are varied. To do this we will treat the initial value as a parameter. Thus, if $f(t, y)$ is a suitably nice function (e.g. continuous and satisfying a uniform Lipschitz condition in $y$) defined near $(t_0, x_0)$, the IVP

$$y' = f(t, y), \quad y(t_0) = x_0$$

has a unique solution, and the IVP

$$y' = f(t, y), \quad y(t_0) = x_1$$

even has a unique solution for any $x_1$ sufficiently close to $x_0$.

IX.1.5.2. What we would hope is that if $x_1$ is close to $x_0$, the solution with initial value $x_1$ should be uniformly close to the solution for initial value $x_0$ for $t$ close to $t_0$, i.e. that the solution for initial value $x$ should “vary continuously” (or even “smoothly”) in $x$.

To be precise, what we might hope for is that there is a continuous (or even smooth) function $u(t, x)$ of two variables, defined near $(t_0, x_0)$, such that if $x_1$ is close to $x_0$, the function $y(t) = u(t, x_1)$ is the unique solution to the IVP with initial value $x_1$. We get continuity:

IX.1.5.3. Theorem. Let $\mathcal{X}, f, r, K, M$ be as in IX.1.2.2. Fix $\epsilon$, $0 < \epsilon < \min \left( \frac{1}{rM}, \frac{1}{K} \right)$. Let $y_0$ be the unique solution to the IVP

$$y' = f(t, y), \quad y(t_0) = x_0$$

in the interval $J = [t_0 - \epsilon, t_0 + \epsilon]$. Fix $x_1$, $\|x_1 - x_0\| \leq \frac{\epsilon}{2}$, and let $y_1$ be the unique solution to the IVP

$$y' = f(t, y), \quad y(t_0) = x_1$$
on $J$. Then, for any $t \in J$, we have

$$\|y_1(t) - y_0(t)\| \leq \frac{\|x_1 - x_0\|}{1 - K \epsilon}.$$
IX.1.5.4. Some initial observations:

(i) The $\epsilon$ is not the same as the one in the proof of IX.1.2.2. (it is smaller).

(ii) The interval $J$ is shorter than the maximum length interval on which $y_0$ is guaranteed to exist by IX.1.2.2. and is not the same as the $J$ in that proof, but $t_0$ is still in the interior of $J$.

(iii) It is easily checked that IX.1.2.2. guarantees the existence of the solution $y_1$ as well as $y_0$ on $J$.

The reason $J$ is taken shorter in this theorem than in IX.1.2.2. is that we need the $T$ defined in the proof to be an actual strict contraction, not just a strict power contraction, in order to apply VI.2.9.7.

IX.1.5.5. Corollary. Under the hypotheses and notation of IX.1.5.3., if $u(t, x)$ is the function from $I \times \bar{B}_{r/2}(x_0)$ to $Y$ such that, for fixed $x$, $u(t, x)$ is the unique solution to the initial-value problem

$$y' = f(t, y), \quad y(t_0) = x$$

then $u$ is uniformly continuous on $I \times \bar{B}_{r/2}(x_0)$.

We now give the proof of IX.1.5.3.

Proof: We proceed exactly as in the proof of IX.1.2.2. Let $Y$ be the set of continuous functions from $J$ to the closed ball $\bar{B}_r(x_0)$ in $\mathcal{X}$ of radius $r$ centered at $x_0$. Define a metric $\rho$ on $Y$ by

$$\rho(\phi, \psi) = \max_{t \in J} \|\phi(t) - \psi(t)\|.$$ 

Then $Y$ is a complete metric space under $\rho$ (it is a closed subset of $C(J, \mathcal{X})$). Define maps $T, S : Y \to Y$ by

$$[T(\phi)](t) = x_0 + \int_{t_0}^{t} f(u, \phi(u)) \, du$$

$$[S(\phi)](t) = x_1 + \int_{t_0}^{t} f(u, \phi(u)) \, du.$$ 

It is obvious from () that $T(\phi)$ and $S(\phi)$ are continuous functions from $J$ to $\mathcal{X}$. We have, for $t \in J$ and $\phi \in Y$,

$$\|T(\phi)(t) - x_0\| = \int_{t_0}^{t} f(u, \phi(u)) \, du \leq \int_{t_0}^{t} \|f(u, \phi(u))\| \, du \leq M|t - t_0| \leq M\epsilon = \frac{r}{2}$$ 

so $T(\phi) \in Y$. Also,

$$\|S(\phi)(t) - x_0\| \leq \|[S(\phi)](t) - x_1\| + \|x_1 - x_0\| = \int_{t_0}^{t} f(u, \phi(u)) \, du + \|x_1 - x_0\|$$

$$\leq \int_{t_0}^{t} \|f(u, \phi(u))\| \, du + \|x_1 - x_0\| \leq M|t - t_0| + \|x_1 - x_0\| \leq M\epsilon + \frac{r}{2} = \frac{r}{2} + \frac{r}{2} = r$$ 

so $S(\phi) \in Y$. And we have, for $\phi, \psi \in Y$ and $t \in J$,

$$\|T(\phi)(t) - T(\psi)(t)\| = \int_{t_0}^{t} [f(u, \phi(u)) - f(u, \psi(u))] \, du$$
\[
\leq \int_{t_0}^{t} \| f(u, \phi(u)) - f(u, \psi(u)) \| du \leq K \int_{t_0}^{t} \| \phi(u) - \psi(u) \| du
\]

and similarly \( \| S(\phi)(t) - S(\psi)(t) \| \leq K \epsilon \rho(\phi, \psi) \). Thus \( T \) and \( S \) are strict contractions on \( Y \) with constant \( K \epsilon < 1 \). Since \( T(\phi) - S(\phi) = x_0 - x_1 \) for all \( \phi \), we have \( \| T(\phi) - S(\phi) \| = \| x_1 - x_0 \| \) for all \( \phi \). The fixed points for \( T \) and \( S \) are \( y_0 \) and \( y_1 \) respectively. The result now follows from an application of VI.2.9.7.

**IX.1.5.6.** The technicalities of the proof only yield that uniform continuity holds on a small neighborhood of each point, but it follows that continuity holds throughout the region in which unique local solutions exist. Uniform continuity can be cobbled together to hold on somewhat larger regions; details are left to the interested reader.

**IX.1.5.7.** The problem in IX.1.5.2. can be converted to a first-order quasilinear PDE problem (Cauchy problem): we seek a smooth function \( u(t, x) \) in a neighborhood of \( (t_0, x_0) \) satisfying (writing \( u_t \) for \( \frac{\partial u}{\partial t} \))

\[
u_t = f(t, u)
\]

with boundary condition \( u(t_0, x) = x \). () then shows the solution \( u(t, x) \) of IX.1.5.5. is not only continuous, but even smooth.

**IX.1.5.8.** We might even want to allow the function \( f \) to depend on the \( x \), i.e. to allow the differential equation itself to vary as the initial condition is varied, or in other words to consider the PDE

\[
u_t = f(t, x, u)
\]

with boundary condition \( u(t_0, x) = x \), where \( f \) is a suitable function defined in a neighborhood of \( (t_0, x_0, x_0) \). Under reasonable hypotheses () shows that there is also a smooth solution in this case.

**IX.1.6. Peano’s Existence Theorem for First-Order ODEs**

If \( f \) does not satisfy a uniform Lipschitz condition in \( y \), we cannot expect a locally unique solution to the IVP \( y' = f(t, y), y(t_0) = x_0 \) in general (Examples IX.1.1.6.(iii)–(iv)). But we can still show existence of a local solution. The first step is to uniformly approximate \( f \) by functions which do satisfy a uniform Lipschitz condition in \( y \); this can only done in general if the target Banach space is finite-dimensional. An application of the Arzela-Ascoli Theorem (which also needs finite-dimensionality of the target space) then gives a solution. In fact, the conclusion can fail if the Banach space \( \mathcal{X} \) is infinite-dimensional, even if \( \mathcal{X} \) is a separable Hilbert space [Yor70]. See [?] for other proofs of the theorem and some generalizations.

**IX.1.6.1. Theorem.** [Peano’s Existence Theorem for First-Order ODEs] Let \( \mathcal{X} \) be a finite-dimensional Banach space, \( I \) a closed bounded interval in \( \mathbb{R} \), \( t_0 \in I \), \( x_0 \in \mathcal{X} \). Let \( f : I \times \bar{B}_r(x_0) \to \mathcal{X} \) be a continuous function, and

\[
M = \max \{ \| f(t, y) \| : (t, y) \in I \times \bar{B}_r(x_0) \}.
\]
Set
\[ I_0 = I \cap \left[ t_0 - \frac{r}{M}, t_0 + \frac{r}{M} \right]. \]
Then there is a \( C^1 \) function \( y : I_0 \to B_r(y_0) \subseteq \mathcal{X} \) satisfying \( y(t_0) = x_0 \) and \( y'(t) = f(t, y(t)) \) for all \( t \in I_0 \).

**Proof:** We may assume \( \mathcal{X} = \mathbb{R}^n \) with a norm equivalent to the Euclidean norm. By the Weierstrass Approximation Theorem, we may approximate \( f \) uniformly on \( I \times B_r(y_0) \) by functions \( f_n \) whose coordinate functions are polynomials. Thus each \( f_n \) satisfies a uniform Lipschitz condition in \( y \) on \( I \times B_r(x_0) \).

Thus by Theorem IX.1.2.2., there is a (unique) solution \( \phi_n \) to the IVP \( y' = f_n(t, y) \), \( y(t_0) = x_0 \) on the closed bounded interval \( I_0 \). We have, for all \( n \) and all \( t \in I_0 \),
\[ \phi_n(t) = x_0 + \int_{t_0}^{t} f_n(u, \phi_n(u)) \, du \]
and thus
\[ \|\phi_n(t)\| \leq \|x_0\| + \left| \int_{t_0}^{t} \|f_n(u, \phi_n(u))\| \, du \right| \leq \|x_0\| + M|t - t_0| \leq \|x_0\| + r \]
so the \( \phi_n \) are uniformly bounded on \( I_0 \). Also, \( \|\phi'_n(t)\| = \|f_n(t, \phi_n(t))\| \leq M \) for all \( t \in I_0 \), so the set \( \{ \phi_n : n \in \mathbb{N} \} \) is equicontinuous on \( I_0 \) ( ). Thus there is a uniformly convergent subsequence by the Arzela-Ascoli Theorem ( ). Passing to a subsequence, we may assume \( (\phi_n) \) converges uniformly on \( I_0 \) to a (necessarily continuous) function \( \phi \). Then, for each \( t \in \mathbb{R} \),
\[ x_0 + \int_{t_0}^{t} f_n(u, \phi_n(u)) \, du \to x_0 + \int_{t_0}^{t} f(u, \phi(u)) \, du \]
by ( ). Since the left side is \( \phi_n(t) \), we obtain
\[ \phi(t) = x_0 + \int_{t_0}^{t} f(u, \phi(u)) \, du \]
for all \( t \in I_0 \), and thus \( \phi \) is a solution to the original IVP on \( I_0 \). The equation \( \phi'(t) = f(t, \phi(t)) \) shows that \( \phi \) is \( C^1 \), since the right side is a continuous function of \( t \).

IX.1.7. Power Series and Analytic Solutions

IX.1.8. Linear ODEs

IX.1.9. Exercises

IX.1.9.1. Let \( I \) be an interval, \( t_0 \in I \), and \( f : I \times \mathbb{R} \to \mathbb{R} \) a continuous function. Consider the differential equation
\[ y' = f(t, y) \]
and the integral equation
\[ y(t) = y(t_0) + \int_{t_0}^{t} f(u, y(u)) \, du . \]
The aim of this problem is to show that a function on \( I \) satisfies the differential equation if and only if it satisfies the integral equation.

(a) Suppose \( y \) is a differentiable function on \( I \) satisfying the differential equation. Show that \( y \) is necessarily \( C^1 \) on \( I \). [Use continuity of \( f \).]

(b) Suppose \( y \) is a differentiable function on \( I \) satisfying the differential equation. Integrate both sides of the differential equation to show that \( y \) also satisfies the integral equation. [Use the Fundamental Theorem of Calculus II.]

(c) Conversely, suppose \( y \) is a function on \( I \) such that the integral in the integral equation exists as a Riemann, Lebesgue, or H-K integral for each \( t \in I \), and that \( y \) satisfies the integral equation. Show that \( y \) is necessarily continuous on \( I \), and hence the integral in the integral equation exists as a Riemann integral for each \( t \in I \). [Use (.).]

(d) Under the assumptions of (c), show that \( y \) is differentiable on \( I \) and satisfies the differential equation. [Use the Fundamental Theorem of Calculus I.]

(e) Extend the results to the case where the range space \( \mathbb{R} \) is replaced by a Banach space.

**IX.1.9.2. Implicit Differential Equations.** (a) Prove the following theorem about existence and uniqueness of solutions to a differential equation of the form \( F(t,y,y') = 0 \).

**Theorem.** Let \( U \) be an open set in \( \mathbb{R}^3 \), and \((t_0,x_0,z_0) \in U \). Let \( F : U \to \mathbb{R} \) be a continuous function with \( F(t_0,x_0,z_0) = 0 \). Suppose \( \frac{\partial F}{\partial z} \) exists and is continuous on \( U \) and that \( \frac{\partial F}{\partial z}(t_0,x_0,z_0) \neq 0 \). Then

(i) There is an open interval \( I \) around \( t_0 \) and a \( C^1 \) function \( y : I \to \mathbb{R} \) such that \( y(t_0) = x_0 \), \( y'(t_0) = z_0 \), and \( F(t,y(t),y'(t)) = 0 \) for all \( t \in I \).

(ii) If \( \frac{\partial F}{\partial y} \) also exists and is continuous on \( U \), then there is an open interval \( I \) around \( t_0 \) and a unique differentiable function \( y : I \to \mathbb{R} \) such that \( y(t_0) = x_0 \), \( y'(t_0) = z_0 \), and \( F(t,y(t),y'(t)) = 0 \) for all \( t \in I \), and additionally \( y \) is \( C^1 \) on \( I \).

(iii) If \( F \) is analytic on \( U \), then there is an open interval \( I \) around \( t_0 \) and a unique differentiable function \( y : I \to \mathbb{R} \) such that \( y(t_0) = x_0 \), \( y'(t_0) = z_0 \), and \( F(t,y(t),y'(t)) = 0 \) for all \( t \in I \), and additionally \( y \) is analytic on \( I \).

[First use the Implicit Function Theorem to rewrite the equation \( F(t,y,z) = 0 \) as \( z = f(t,y) \). Apply VIII.6.3.4. and IX.1.6.1. in case (i), VIII.6.3.12. and IX.1.2.7. in case (ii), and (i) and (ii) in case (iii).]

However, such an implicit differential equation need not have any solutions, i.e. there may be no \((t_0,x_0,z_0)\) satisfying the hypotheses: see Exercise IX.1.9.3.

(b) Generalize to systems and to higher-order implicit equations.

**IX.1.9.3.** Show that the (autonomous) implicit differential equation

\[(y' - 1)^2 + y^2 = 0\]

has no (real) solutions.
IX.1.9.4. Implicit Solutions. (a) Carefully define what it means for a relation \( G(t, y) = 0 \) to be a “solution” to the differential equation \( F(t, y, y') = 0 \) (or a higher-order equation).

(b) [p. 3] Show that \( t^2 = 2y^2 \log y \) is an implicit solution to the differential equation
\[
y' = \frac{ty}{t^2 + y^2}.
\]

Implicit solutions are particularly easily obtained for exact differential equations (ones which can be put in the form \( g(t, y)y' + h(t, y) = 0 \) with \( \frac{\partial g}{\partial t} = \frac{\partial h}{\partial y} \)), cf. VIII.2.2.8., and in particular for separable differential equations (ones which can be put in the form \( g(y)y' = h(t) \)).

IX.1.9.5. This is a sort of converse to IX.1.9.4. Let \( G(t, y, c_1, \ldots, c_n) \) be a \( C^n \) function with \( \frac{\partial G}{\partial y} \neq 0 \). Regard \( t \) as the independent variable, \( y \) the dependent variable, and \( c_1, \ldots, c_n \) as “constants” (parameters). By the Implicit Function Theorem, the equation
\[
G(t, y, c_1, \ldots, c_n) = 0
\]
nimplicitly defines \( y \) as a function of \( t \) and the parameters \( c_1, \ldots, c_n \), i.e. \( y = g(t, c_1, \ldots, c_n) \).

(a) Successively differentiate \( y \) with respect to \( t \) to obtain \( n \) additional equations
\[
\frac{\partial G}{\partial t} + \frac{\partial G}{\partial y} y' = 0
\]
\[
\frac{\partial^2 G}{\partial t^2} + \frac{\partial^2 G}{\partial t \partial y} y' + \frac{\partial^2 G}{\partial y^2} (y')^2 + \frac{\partial G}{\partial y} y'' = 0
\]
\[
\ldots
\]
\[
\frac{\partial^n G}{\partial t^n} + \ldots + \frac{\partial G}{\partial y} y^{(n)} = 0
\]
(where we really mean \( y^{(k)} = \frac{\partial^k y}{\partial t^k} \), but we symbolically regard the \( y^{(k)} \) as new variables).

(b) Show that if \( c = (c_1, \ldots, c_n) \) is independent in \( G \) in a suitable sense, then by the Implicit Function Theorem the first \( n \) equations can (in principle) be solved for \( c_1, \ldots, c_n \) as functions of \( t, y, y', \ldots, y^{(n-1)} \) and substituted into the last equation to obtain a differential equation
\[
F(t, y, y', \ldots, y^{(n-1)}, y^{(n)}) = 0
\]
for which \( G(t, y, c_1, \ldots, c_n) = 0 \) is a solution (in the sense of IX.1.9.4.) for any suitable \( c \) (references such as [Gou59, V. 2 Part 2] and [Inc44] state that \( c_1, \ldots, c_n \) “can be eliminated” from the \( n + 1 \) equations without elaboration; in [?] it is at least acknowledged there is a nontrivial issue, but that “a precise statement of the conditions under which this is true is too complicated to be stated here.”) Use the existence and uniqueness theorem to show that, under appropriate conditions, every solution is of this form.

(c) Generalize to the case where there is more than one \( G, y, \) and/or \( t \) (the same number of \( G \)'s and \( y \)'s) to obtain a system of ordinary or partial differential equations from an equation or set of equations with an arbitrary (finite) set of constants.
Consider the IVP \( y' = y^2, y(t_0) = x_0 \) (IX.1.1.6.iii). Theorem IX.1.2.2. guarantees a unique solution on
\[
\left[ t_0 - \frac{r}{(|x_0| + r)^2}, t_0 + \frac{r}{(|x_0| + r)^2} \right]
\]
for any \( r > 0 \). If \( x_0 = 0 \), the supremum of the lengths of these intervals for all \( r \) is \( +\infty \), and the solution to the IVP \( (y = 0) \) is defined on all of \( \mathbb{R} \). If \( x_0 \neq 0 \), show that the maximum of \( \frac{r}{(|x_0| + r)^2} \) is \( \frac{1}{4|x_0|} \). In this case, the unique solution \( y = \frac{1}{c-t}, c = t_0 + \frac{1}{x_0} \), has domain \((-\infty, c)\) if \( x_0 > 0 \) and \((c, +\infty)\) if \( x_0 < 0 \), so the maximum symmetric interval around \( t_0 \) on which the solution exists is \( \left( t_0 - \frac{1}{|x_0|}, t_0 + \frac{1}{|x_0|} \right) \).

**IX.1.9.7.** Show that Theorem IX.1.3.2. can be generalized as follows to give a unique solution on a generally larger interval. Suppose we have a Banach space \( \mathcal{X} \), an interval \( I \subseteq \mathbb{R} \), a point \( t_0 \in I \), \( n \) points \( x_1^{(0)}, \ldots, x_n^{(0)} \) in \( \mathcal{X} \), \( r_1, \ldots, r_n > 0 \), \( M > 0 \), \( K > 0 \), and \( n \) continuous functions
\[
f_1, \ldots, f_n : I \times \bar{B}_{r_1}(x_1^{(0)}) \times \cdots \times \bar{B}_{r_n}(x_n^{(0)}) \to \mathcal{X}
\]
with the property that \( \|f_k(t, x_1, \ldots, x_n)\| \leq M_k \) for \( t \in I \) and \( \|x_j - x_j^{(0)}\| \leq r_j (1 \leq j, k \leq n) \), and
\[
\|f_k(t, x_1^{(2)}, \ldots, x_n^{(2)}) - f_k(t, x_1^{(1)}, \ldots, x_n^{(1)})\| \leq K \max_{1 \leq j \leq n} (\|x_j^{(2)} - x_j^{(1)}\|)
\]
for \( 1 \leq k \leq n \) whenever \( t \in I \) and \( x_j^{(1)}, x_j^{(2)} \in \bar{B}_{r_j}(x_j^{(0)}) \) for all \( j \) (the \( f_k \) satisfy a uniform Lipschitz condition in \( x_1, \ldots, x_n \) on \( I \)). Slightly adapt the proof of Theorem IX.1.2.2. to show the following:

**Theorem. [Existence and Uniqueness Theorem for Systems of First-Order ODEs]** Using the above notation, set
\[
L = \min \left\{ \frac{r_1}{M_1}, \ldots, \frac{r_n}{M_n} \right\}
\]
\[
I_0 = I \cap [t_0 - L, t_0 + L] .
\]
Then there are unique differentiable functions \( y_1, \ldots, y_n : I_0 \to \mathcal{X} \) such that
\[
y_k(t) = f_k(t, y_1(t), \ldots, y_n(t))
\]
\( (1 \leq k \leq n) \) for \( t \in I_0 \) and \( y_k(t_0) = x_k^{(0)} \). Additionally, the \( y_k \) are \( C^1 \) on \( I_0 \). If \( \tilde{I} \) is any subinterval of \( I_0 \) containing \( t_0 \), the restriction of the \( y_k \) to \( \tilde{I} \) is the unique solution to the IVP on \( \tilde{I} \).

**IX.1.9.8.** Use the theorem in IX.1.9.7. to obtain the following generalization of Theorem IX.1.4.1. which guarantees a unique solution on a larger interval in general:

**Theorem. [Existence and Uniqueness Theorem for Higher-Order ODE’s]** Let \( \mathcal{X} \) be a Banach space, \( I \) an interval in \( \mathbb{R} \), \( t_0 \in I \), \( x_0, x_0', \ldots, x_0^{(n-1)} \in \mathcal{X} \), \( r_0, \ldots, r_{n-1} > 0 \), and
\[
f : I \times \bar{B}_{r_0}(x_0) \times \bar{B}_{r_1}(x_0') \times \cdots \times \bar{B}_{r_{n-1}}(x_0^{(n-1)}) \to \mathcal{X} .
\]
Suppose $f$ is continuous and bounded by $N$ on $I \times \bar{B}_r(x_0) \times \bar{B}_r(x_0') \times \cdots \times \bar{B}_r(x_0^{(n-1)})$ and satisfies a uniform Lipschitz condition in the $y$-variables. Set

$$L = \max \left\{ \frac{r_0}{\|x_0\| + r_1}, \ldots, \frac{r_{n-2}}{\|x_0^{(n-1)}\| + r_{n-1}}, \frac{r_{n-1}}{N} \right\}$$

$$I_0 = I \cap [t_0 - L, t_0 + L] .$$

Then there is a unique $n$-times differentiable function $y : I_0 \to Y$ with $y(t_0) = x_0$, $y'(t_0) = x_0'$, $\ldots$, $y^{(n-1)}(t_0) = x_0^{(n-1)}$, and $y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t))$ for all $t \in I_0$. Additionally, $y$ is $C^n$ on $I_0$.

**IX.1.9.9.** Let $I$ be an interval in $\mathbb{R}$, $\mathcal{X}$ a Banach space, and $f : I \to \mathcal{X}$ a continuous function. Then $f$ may be regarded as a continuous function from $I \times \mathcal{X}$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in $\mathcal{X}$ with Lipschitz constant 0. Show that one step of Picard iteration gives the exact (unique) solution to the IVP $y' = f(t)$, $y(t_0) = x_0$ for any $t_0 \in I$, $x_0 \in \mathcal{X}$.

**IX.1.9.10.** (a) Use the Existence and Uniqueness Theorem IX.1.2.7. to give an alternate proof of the one-variable $C^r$ Implicit Function Theorem VIII.6.3.14. for $r \geq 2$ by considering the IVP (with $x$ the independent variable)

$$y' = -\frac{\partial F(x, y)}{\partial x}, \quad y(a) = b.$$ Use the Chain Rule to show that the unique solution $y = f(x)$ satisfies $F(x, f(x)) = 0$ for all $x$ near $a$. (One does not obtain uniqueness of the implicitly defined function this way, only uniqueness of a differentiable implicitly defined function.) What can one conclude from this argument if $F$ is only $C^1$? (cf. IX.1.6.1.)

(b) Generalize to the case of VIII.6.1.7. for $p = 1$ and $r \geq 2$, for general $n$, by considering a system of differential equations.

(c) Give a similar proof of the general $C^r$ Implicit Function Theorem VIII.6.1.7. ($r \geq 2$) by considering a system of partial differential equations

$$Dy = -D_y F(x, y)^{-1} D_x F(x, y), \quad y(a) = b.$$ Use these results to give explicit descriptions of sets on which the implicitly defined functions exist.
IX.2. Partial Differential Equations

"PDEs constitute one of the most dynamic and varied domains of the mathematical sciences, defying all attempts at unification. They are found in every phenomenon studied by the physics of continuous systems, involving all states of matter (gases, fluids, solids, plasmas) and all physical theories (classical, relativistic, quantum, and so on).

But partial differential equations also lurk behind many geometric problems. Geometric PDEs, as they are called, make it possible to deform geometric objects in accordance with well-established laws. The application of methods of analysis to problems in other fields of mathematics is an example of the sort of cross-fertilization that became increasingly common in the course of the twentieth century."

C. Villani

IX.2.1. Elementary Theory

It is common to denote the unknown function in a partial differential equation by \( u \) (or \( v, w \)), and the independent variables by \( x, y, z \), and/or \( t \). Of course, a PDE can have any number of independent variables, but PDE’s arising in physics often describe an unknown function of three (or fewer) spatial coordinates \((x, y, z)\) and/or one time coordinate \( t \). It is also notationally economical to use the subscript notation for partial derivatives, e.g. \( u_x \) for \( \frac{\partial u}{\partial x} \) or \( u_{xy} \) for \( \frac{\partial^2 u}{\partial y \partial x} \) (note the order of differentiation, which is often but not always irrelevant; cf. VIII.2.2.2).

IX.2.1.1. Example. Consider the apparently extremely simple PDE which can be variously written

\[
\frac{\partial u}{\partial x} = \frac{\partial u}{\partial t} \\
\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0 \\
u_x - u_t = 0
\]

where \( u \) is an unknown function of \( x \) and \( t \). It is easy to check using the Chain Rule that

\[
u(x, t) = f(x + t)\]

is a solution on \( \Omega = \{(x, t) : x + t \in I\} \), for any real-valued function \( f \) which is differentiable on an open interval \( I \) in \( \mathbb{R} \) (it turns out that these are all the solutions (i)). There are several observations to take away from this example:

(i) There is not a clean standard form for PDEs such as in IX.1.1.2..

(ii) Even a homogeneous linear first-order PDE with constant coefficients can have solutions which are not even \( C^1 \), much less analytic (cf. (i)).

\[^3[Vil16, \text{p. } 107].\]
(iii) The general solution to such a PDE does not just contain an arbitrary constant, but an arbitrary (differentiable) function which must be determined by initial or boundary conditions (the fact that there is one equation and two independent variables suggests that there is an arbitrary function of $2 - 1 = 1$ variables which comes into the general solution, which is correct when properly formulated).

IX.2.2. Quasilinear Partial Differential Equations


IX.2.3. The Cauchy-Kovalevskaya Theorem
Chapter X

Fundamentals of Complex Analysis

Complex Analysis is calculus done on the complex numbers. Although at first it bears superficial resemblance to ordinary calculus, it quickly assumes a much different character.

Complex Analysis is one of the most beautiful parts of mathematics. Many of the theorems seem almost too good to be true, especially in comparison with the situation in Real Analysis. Complex Analysis is also extraordinarily important throughout mathematics, with applications ranging from number theory to geometry to mathematical physics (and almost everything else).

It is beyond the scope of this book to give a full treatment of this vast subject. We will only describe some of the most fundamental aspects and their connections with real analysis. There is an overabundant supply of texts on Complex Analysis, which are of varying quality but mostly adequate or better. Four of the best are the classic [Ahl78], [GK06], [Gam01], and [SS03].

Throughout this chapter, we will always use the standard model of the complex numbers $\mathbb{C}$, the complex plane (III.6.2.). Thus we will freely regard $\mathbb{C}$ as $\mathbb{R}^2$ with the additional operation of complex multiplication. Thus any function from $\mathbb{C}$ (or a subset) to $\mathbb{C}$ can be equally well thought of as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$. In discussing limits and continuity in $\mathbb{C}$, we always use the topology on $\mathbb{C}$ obtained by the identification with $\mathbb{R}^2$. This topology is induced by the metric $\rho(z, w) = |z - w|$.

X.1. Elementary Theory

X.1.1. Complex Differentiation and Holomorphic Functions

We first discuss the complex analog of ordinary differentiation. Many basic definitions and results bear a close formal resemblance to corresponding notions in usual calculus of one variable.

X.1.1.1. We will use the following standard notation. The independent variable in complex analysis is usually denoted $z = x + iy$, where $z \in \mathbb{C}$ and $x$ and $y$ are the real and imaginary parts of $z$. Then $z$ corresponds to the point $(x, y) \in \mathbb{R}^2$. If $\Omega$ is a (usually open) subset of $\mathbb{C}$, and $f : \Omega \to \mathbb{C}$ is a function, write

$$f(z) = f(x + iy) = u(x, y) + iv(x, y)$$

where $u, v : \Omega \to \mathbb{R}$ are the real and imaginary parts of $f$ (the coordinate functions of $f$ when $f$ is regarded as a function from $\Omega$ to $\mathbb{R}^2$).
X.1.1.2. Definition. Let $\Omega$ be an open subset of $\mathbb{C}$, $f : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$. Then $f$ is differentiable in the complex sense (or just complex differentiable) at $z_0$ if

$$
\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}
$$

exists. If the limit exists, it is called the complex derivative of $f$ at $z_0$, denoted $f'(z_0)$.

The function $f$ is complex differentiable on $\Omega$, or holomorphic on $\Omega$, if it is complex differentiable at every $z_0 \in \Omega$.

X.1.1.3. In Complex Analysis, the term analytic is often used as a synonym for holomorphic. We will not initially use analytic in this way since it has a different official definition (V.16.1.1.). One of the major fundamental theorems of Complex Analysis is that the notions of being analytic and holomorphic on an open set $\Omega$ coincide (X.4.4.7.).

Here is the relation between complex differentiation and ordinary differentiation on $\mathbb{R}^2$:

X.1.1.4. Proposition. Let $\Omega$ be an open subset of $\mathbb{C}$, $f : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$, $c = a + ib \in \mathbb{C}$. The following are equivalent:

(i) $f$ is complex differentiable at $z_0$ with $f'(z_0) = c$.

(ii) For $z \in \Omega$, $f(z) = f(z_0) + c(z - z_0) + o(|z - z_0|)$.

(iii) $f$ (regarded as a function from $\Omega$ to $\mathbb{R}^2$) is differentiable at $z_0 = (x_0, y_0)$ in the vector calculus sense () and $Df(x_0, y_0) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$.

This result gives motivation and “explanation” for the model of $\mathbb{C}$ in Exercise III.6.3.1..

Proof: The argument is very similar to the calculation motivating the definition of vector calculus differentiation (). The following statements are equivalent:

$$
\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = c
$$

$$
\lim_{z \to z_0} \frac{f(z) - f(z_0) - c(z - z_0)}{z - z_0} = 0
$$

$$
\lim_{z \to z_0} \frac{|f(z) - f(z_0) - c(z - z_0)|}{|z - z_0|} = 0
$$

$$
\lim_{(x,y) \to (x_0,y_0)} \frac{||f(x,y) - f(x_0,y_0) - T(h)||}{||h||} = 0
$$

where $h = (x, y) - (x_0, y_0)$ and $T$ is the linear transformation of $\mathbb{R}^2$ given by multiplication by $c$. The argument is completed by noting that the matrix of $T$ is $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$.

$\square$
X.1.1.5. **Corollary.** Let $\Omega$ be an open subset of $\mathbb{C}$, $f = u + iv : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$. Then $f$ is complex differentiable at $z_0$ if and only if $u$ and $v$ are differentiable in the vector calculus sense at $(x_0, y_0)$ and satisfy the Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0), \quad \frac{\partial u}{\partial y}(x_0, y_0) = -\frac{\partial v}{\partial x}(x_0, y_0).$$

The complex derivative, when it exists, is given by

$$f'(x_0 + iy_0) = \frac{\partial u}{\partial x}(x_0, y_0) + i \frac{\partial v}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0) - i \frac{\partial u}{\partial y}(x_0, y_0).$$

So $f$ is holomorphic on $\Omega$ if and only if $f$ is differentiable in the vector calculus sense on $\Omega$ and the Cauchy-Riemann equations hold everywhere on $\Omega$. In particular, if $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are continuous on $\Omega$ and satisfy the Cauchy-Riemann equations, then $f$ is holomorphic on $\Omega$.

X.1.1.6. It turns out that in X.1.1.5., the assumption that $f$ be differentiable in the vector calculus sense on $\Omega$ can be relaxed. If the first-order partials of $u$ and $v$ exist everywhere on $\Omega$ and satisfy the Cauchy-Riemann equations, and if $f$ is continuous on $\Omega$ (Looman-Menchoff Theorem) or just locally bounded (Montel-Tolstoff Theorem), then $f$ is automatically holomorphic on $\Omega$. Some additional condition on $f$ besides the Cauchy-Riemann equations is necessary: cf. Exercise X.8.2.1. The question of when a function satisfying the Cauchy-Riemann equations is holomorphic is a subtle one, which is discussed thoroughly in [GM78]. (And a strong version of the converse of the last statement of X.1.1.5. also holds (X.3.3.6.).)

X.1.1.7. X.1.1.4. suggests that it is qualitatively harder for a function from $\mathbb{C}$ to $\mathbb{C}$ to be complex differentiable than it is for a function from $\mathbb{R}$ to $\mathbb{R}$ to be differentiable in the usual sense. This turns out to be correct, and complex differentiability implies all sorts of remarkable properties that do not necessarily hold for usual differentiable functions on $\mathbb{R}$.

There are nonetheless many examples of holomorphic functions, enough to make a rich theory. Here are some basic ones:

X.1.1.8. **Examples.** (i) If $f$ is a polynomial in $z$ with complex coefficients, then by a calculation using the product formula in $\mathbb{C}$ the real and imaginary parts of $f$ are polynomials in $x$ and $y$ with real coefficients. Thus $f$ is differentiable (in fact, $C^\infty$) on $\mathbb{C}$. A rather ugly calculation shows that the Cauchy-Riemann equations are satisfied, so $f$ is holomorphic on $\mathbb{C}$. A simpler argument which shows this, and that $f'$ is given by the usual formula, is in X.1.2.5.

(ii) Similarly, any rational function in $z$ is holomorphic on its domain (X.1.2.8).

(iii) **[Complex Exponential Function]** For $z = x + iy \in \mathbb{C}$, set

$$e^z = e^{x+iy} = e^x \cos y + ie^x \sin y = e^x (\cos y + i \sin y).$$

Then $f(z) = e^z$ is differentiable (in fact, $C^\infty$) on $\mathbb{C}$, and it is easily checked that the Cauchy-Riemann equations hold; hence $f$ is holomorphic on $\mathbb{C}$. By X.1.1.4.,

$$f'(x + iy) = \frac{\partial u}{\partial x}(x, y) + i \frac{\partial v}{\partial x}(x, y) = f(x + iy).$$
i.e. $f' = f$. The complex exponential function also satisfies $e^{z+w} = e^z e^w$, $e^{-z} = \frac{1}{e^z}$, and $e^{z-w} = \frac{e^z}{e^w}$ for all $z, w \in \mathbb{C}$, and agrees with the usual exponential function for $z$ real (and is the only holomorphic function extending the real exponential function, cf. X.5.1.4. and Exercise X.8.2.6.).

(iv) **[Complex Trigonometric Functions]** For $z = x + iy \in \mathbb{C}$, set

$$
sin z = \sin x \cosh y + i \cos x \sinh y
$$

$$
cos z = \cos x \cosh y - i \sin x \sinh y
$$

Then $f(z) = \sin z$ and $g(z) = \cos z$ are differentiable (in fact, $\mathcal{C}^\infty$) on $\mathbb{C}$, and the Cauchy-Riemann equations hold, so $f$ and $g$ are holomorphic on $\mathbb{C}$. By a calculation as in (iii), $f' = g$ and $g' = -f$. These functions satisfy the usual trigonometric identities, and reduce to the usual $\sin$ and $\cos$ when $z$ is real. The other trigonometric functions can be defined from these in the usual way, and give holomorphic functions on their domains extending the usual ones and satisfying the same identities and differentiation rules.

(v) **[Complex Hyperbolic Functions]** For $z = x + iy \in \mathbb{C}$, set

$$
\sinh z = \sinh x \cosh y + i \cosh x \sinh y = \frac{e^z - e^{-z}}{2}
$$

$$
cosh z = \cosh x \cosh y + i \sinh x \sinh y = \frac{e^z + e^{-z}}{2}
$$

Then similarly $f(z) = \sinh z$ and $g(z) = \cosh z$ are holomorphic on $\mathbb{C}$, and $f' = g$, $g' = f$. These functions satisfy the usual identities, and reduce to the usual $\sinh$ and $\cosh$ when $z$ is real. The other hyperbolic functions can be defined from these in the usual way, and give holomorphic functions on their domains extending the usual ones and satisfying the same identities and differentiation rules. We have

$$
cos z = \cosh iz = \frac{e^{iz} + e^{-iz}}{2}
$$

$$
sin z = -i \sinh iz = \frac{e^{iz} - e^{-iz}}{2i}
$$

for all $z \in \mathbb{C}$.

These examples show the intimate connections between the exponential, trigonometric, and hyperbolic functions which become clear only when the extensions to the complex numbers are considered.

Defining inverse functions (log, arcsin, etc., even $\sqrt{z}$) is a trickier matter, primarily due to complications of domain.

**X.1.1.9.** In this case, all these functions are $\mathcal{C}^\infty$ on their domains (as functions from $\mathbb{R}^2$ to $\mathbb{R}^2$), and their derivatives of all orders are also holomorphic. This in fact is true for any holomorphic function on an open set in $\mathbb{C}$ (X.3.3.6.), one of the remarkable properties of holomorphic functions.

**X.1.2. Rules of Differentiation**

There are complex analogs of the usual rules of differentiation. These can be proved by an easy adaptation of the arguments in the real case, or by rather tediously using the real results, the formulas for complex multiplication and division, and X.1.1.4.
X.1.2.1. **Proposition.** Let \( f \) be a function from an open set \( \Omega \subseteq \mathbb{C} \) to \( \mathbb{C} \). If \( f \) is complex differentiable at \( z_0 \in \Omega \), then \( f \) is continuous at \( z_0 \).

X.1.2.2. **Theorem.** [Power Rule] Let \( n \in \mathbb{N} \), and \( f : \mathbb{C} \to \mathbb{C} \) be defined by \( f(z) = z^n \). Then \( f \) is complex differentiable at any \( z_0 \in \mathbb{C} \), and

\[
f'(z_0) = nz_0^{n-1}.
\]

X.1.2.3. **Proposition.** [Sum Rule] Let \( f \) and \( g \) be functions from an open set \( \Omega \subseteq \mathbb{C} \) to \( \mathbb{C} \), and \( h = f + g : \Omega \to \mathbb{C} \). If \( f \) and \( g \) are complex differentiable at \( z_0 \in \Omega \), then \( h \) is complex differentiable at \( z_0 \) and

\[
h'(z_0) = f'(z_0) + g'(z_0).
\]

X.1.2.4. **Proposition.** [Constant Multiple Rule] Let \( f \) be a function from an open set \( \Omega \subseteq \mathbb{C} \) to \( \mathbb{C} \), \( c \in \mathbb{C} \), and \( h = cf : \Omega \to \mathbb{C} \). If \( f \) is complex differentiable at \( z_0 \in \Omega \), then \( h \) is complex differentiable at \( z_0 \) and

\[
h'(z_0) = cf'(z_0).
\]

X.1.2.5. **Corollary.** Any sum or constant multiple of functions holomorphic on \( \Omega \) is holomorphic on \( \Omega \). In particular, a polynomial with complex coefficients is holomorphic on \( \mathbb{C} \).

X.1.2.6. **Theorem.** [Product Rule] Let \( f \) and \( g \) be functions from an open set \( \Omega \subseteq \mathbb{C} \) to \( \mathbb{C} \), and \( h = fg : \Omega \to \mathbb{C} \). If \( f \) and \( g \) are complex differentiable at \( z_0 \in \Omega \), then \( h \) is complex differentiable at \( z_0 \) and

\[
h'(z_0) = f'(z_0)g(z_0) + f(z_0)g'(z_0).
\]

X.1.2.7. **Theorem.** [Quotient Rule] Let \( f \) and \( g \) be functions from an open set \( \Omega \subseteq \mathbb{C} \) to \( \mathbb{C} \). If \( f \) and \( g \) are complex differentiable at \( z_0 \in \Omega \) and \( g(z_0) \neq 0 \), then \( h = \frac{f}{g} \) is defined in an open neighborhood of \( z_0 \) and differentiable at \( z_0 \), and

\[
h'(z_0) = \frac{g(z_0)f'(z_0) - f(z_0)g'(z_0)}{|g(z_0)|^2}.
\]
X.1.2.8. Corollary. If \( f \) and \( g \) are holomorphic on \( \Omega \), then \( fg \) is holomorphic on \( \Omega \), and if \( g \) is nonzero on \( \Omega \), then \( \frac{1}{g} \) is also holomorphic on \( \Omega \). In particular, every rational function with complex coefficients is holomorphic on its domain.

X.1.2.9. Theorem. [Chain Rule] Let \( f \) be defined on a neighborhood \( U \) of \( z_0 \) in \( \mathbb{C} \) and complex differentiable at \( z_0 \), with \( f(z_0) = w_0 \), and let \( g \) be defined on a neighborhood \( V \) of \( w_0 \) in \( \mathbb{C} \) and complex differentiable at \( w_0 \), with \( f(U) \subseteq V \). Set \( h = g \circ f \) on \( U \). Then \( h \) is complex differentiable at \( z_0 \), and
\[
h'(z_0) = (g \circ f)'(z_0) = g'(w_0)f'(z_0) = g'(f(z_0))f'(z_0).
\]

X.1.2.10. Theorem. [Inverse Function Theorem] Let \( \Omega \subseteq \mathbb{C} \) be open, \( f \) holomorphic on \( \Omega \), and \( z_0 \in \Omega \). If \( f'(z_0) \neq 0 \), then there is an open neighborhood \( V \) of \( z_0 \) contained in \( \Omega \) for which \( f \) maps bijectively to an open neighborhood \( W \) of \( f(z_0) \), and the inverse function \( f^{-1} \) is holomorphic on \( W \). If \( z \in V \), then
\[
(f^{-1})'(f(z)) = \frac{1}{f'(z)}.
\]

Proof: There is one aspect of this theorem not yet established: it follows automatically that \( f' \) is continuous (in fact, holomorphic) on \( \Omega \) (X.3.3.6.). The result now follows from the Inverse Function Theorem, since if \( f \) is regarded as a function from \( \Omega \) to \( \mathbb{R}^2 \), we have \( J_f(z) = det(Df(z)) = |f'(z)|^2 \). The inverse function is holomorphic, since if \( z \in V \) and \( f'(z) = a + ib \), then
\[
D(f^{-1})(f(z)) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}^{-1} = \frac{1}{a^2 + b^2} \begin{bmatrix} a & b \\ -b & a \end{bmatrix}
\]
and thus \( f^{-1} \) satisfies the Cauchy-Riemann equations; then
\[
(f^{-1})'(f(z)) = \frac{a}{a^2 + b^2} - i \frac{b}{a^2 + b^2} = \frac{1}{f'(z)}.
\]

Although we have stated this result in maximum generality, we have so far only given a complete proof of the following slightly weaker version:

X.1.2.11. Corollary. Let \( \Omega \subseteq \mathbb{C} \) be open, \( f \) holomorphic on \( \Omega \), and \( z_0 \in \Omega \). If \( f' \) is continuous on \( \Omega \) and \( f'(z_0) \neq 0 \), then there is an open neighborhood \( V \) of \( z_0 \) contained in \( \Omega \) which \( f \) maps bijectively to an open neighborhood \( W \) of \( f(z_0) \), and the inverse function \( f^{-1} \) is holomorphic on \( W \). If \( z \in V \), then
\[
(f^{-1})'(f(z)) = \frac{1}{f'(z)}.
\]
X.1.2.12. In particular, if $f$ is holomorphic on $\Omega$ and $f'$ never vanishes, then $f$ is an open mapping. If $\Omega$ is connected, this result extends to the case where $f$ is just assumed to be not constant (). In fact, $f$ is locally an open mapping around any point where it is not locally constant ()

X.1.3. Branch Cuts and Branches

X.1.3.1. A holomorphic function which is locally one-to-one (i.e. whose derivative does not vanish) is not globally one-to-one in general. The fundamental example is the complex exponential function: $e^{z+2\pi i} = e^z$ for any $z$. If we want to define an inverse function for such a function, we must restrict to regions where the function gives a bijection.

X.1.3.2. The prototype problem is how to define a complex logarithm. Since we have $e^{x+iy} = e^x e^{iy}$, we have

$$|e^{x+iy}| = e^x, \quad \text{arg}(e^{x+iy}) = y$$

so we may try to define the complex logarithm using polar coordinates by

$$\log(re^{i\theta}) = \log r + i\theta.$$ 

The real part is well defined as long as $r \neq 0$ (obviously $\log 0$ cannot make sense since $e^z$ is never zero); but the problem is that the argument $\theta$ is only defined up to adding a multiple of $2\pi$.

X.1.3.3. The standard solution is to restrict the exponential function to an open horizontal strip of width $2\pi$, say the strip $S_\alpha = \{x + iy : \alpha < y < \alpha + 2\pi\}$ for a fixed real number $\alpha$. The exponential function maps this open strip one-one onto the plane with a ray out from 0 at angle $\alpha$ (measured in the usual way from the positive real axis counterclockwise) removed. An inverse function can then be defined on this open set $\Omega_\alpha$ taking values in the strip, which is a reasonable definition of a complex logarithm function. This function is called a branch of the complex logarithm function with the ray as branch cut. There is thus one branch of the complex logarithm for each real number $\alpha$. Note that for a given branch cut, there are infinitely many branches: the branch cuts for the branches corresponding to $\alpha$ and $\beta$ are the same if (and only if) $\alpha$ and $\beta$ differ by a multiple of $2\pi$.

A branch of the complex logarithm cannot be continuously extended to any point on the branch cut, since the limits as the point is approached from opposite sides of the cut differ by $2\pi i$.

X.1.3.4. The branch corresponding to $\alpha = -\pi$ is called the principal branch of the complex logarithm. The branch cut for the principal branch is the negative real axis, and the values are in $S_{-\pi} = \{x + iy : -\pi < y < \pi\}$.

The principal branch of the complex logarithm agrees with the usual (real) logarithm function on the positive real axis.

X.1.3.5. Although branch cuts are usually taken to be straight rays, any curve beginning at 0 and going to $\infty$ without crossing itself can be used as a branch cut, and a branch of the complex logarithm (actually infinitely many) can be defined on the complement. (The inverse image of such a curve under the exponential function gives a sequence of parallel curves in the plane a vertical distance $2\pi$ apart, and the exponential function is bijective on the region between consecutive curves.)
X.1.4. The Fundamental Theorem of Algebra

The Fundamental Theorem of Algebra, that every complex polynomial has a root in $\mathbb{C}$, has many essentially different proofs. There is a nice one due to J. Milnor [Mil97], based on the results of X.1.1., which we describe here. See X.8.2.3. and X.8.2.4. for two more proofs using Complex Analysis.

X.1.4.1. Lemma. Let $p$ be a nonconstant polynomial with complex coefficients,

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$$

with $n \geq 1$ and $a_n \neq 0$. Then for any $M > 0$ there is a $K > 0$ such that $|p(z)| > M$ for all $z$ with $|z| > K$.

Proof: For all $z$ with $|z| \geq 1$, we have

$$|p(z)| = |z^n|(|a_0 z^{-n} + \cdots + a_n|z|^{-1} + a_n|) \geq |z|^n \left( |a_n| - |z|^{-1} \sum_{k=0}^{n-1} |a_k| \right)$$

If $|z|$ is sufficiently large, the second factor is greater than $\left| \frac{a_n}{z} \right|$, and $|z|^n > \frac{2M}{|a_n|}$. So

$$K = \max \left( 1, \frac{2 \sum_{k=0}^{n-1} |a_k|}{|a_n|}, \left[ \frac{2M}{|a_n|} \right]^{1/n} \right)$$

will do.

X.1.4.2. Theorem. [Fundamental Theorem of Algebra] Let $p$ be a nonconstant polynomial with coefficients in $\mathbb{C}$. Then there is at least one $c \in \mathbb{C}$ with $p(c) = 0$.

Proof: Let $R \subseteq \mathbb{C}$ be the range of $p$, i.e.

$$R = \{p(z) : z \in \mathbb{C}\}$$

and set

$$F = \{p(z) : z \in \mathbb{C}, p'(z) = 0\}$$

and $S = R \setminus F$. Then $F$ is a finite subset of $R$ since $p'$ is not identically zero. We will show that $R$ is closed and $S$ is open in $\mathbb{C}$. This will prove the theorem since if $\mathbb{C} \setminus R$ is nonempty, then $S$ and $\mathbb{C} \setminus R$ would be disjoint nonempty open sets whose union is the connected open set $\mathbb{C} \setminus F$, a contradiction. (Note that $S$ is nonempty since otherwise $R = F$, and $F$ would have to be a singleton by continuity, so $p$ would be constant.) Thus $R = \mathbb{C}$, and in particular $0 \in R$.

To show that $R$ is closed, it suffices to show that $R \cap \bar{B}_M(0)$ is closed for all $M > 0$. Fix $M$. There is a $K$ such that $|p(z)| > M$ for all $z$ with $|z| > K$ (Lemma X.1.4.1.). Thus

$$R \cap \bar{B}_M(0) = p(\bar{B}_K(0)) \cap \bar{B}_M(0)$$

which is compact since $p$ is continuous (\).

We now show $S$ is open. Let $w_0 \in S$. Fix $z_0 \in \mathbb{C}$ with $p(z_0) = w_0$, $p'(z_0) \neq 0$. So by X.1.2.11. there is a neighborhood $V$ of $z_0$ which $p$ maps homeomorphically onto a neighborhood $W$ of $w_0$; thus $W \subseteq R$. Since $F$ is finite and $w_0 \notin F$, $W \setminus F$ is a neighborhood of $w_0$ contained in $S$. \(\blacksquare\)
X.2. Contour Integration

Contour integration is one of the most fundamental tools of Complex Analysis. Cauchy’s Theorem, which roughly says that the contour integral of a holomorphic function around a closed contour is zero, is the key to all the deeper properties of holomorphic functions.

X.2.1. Integration along a Contour

X.2.1.1. Definition. Let $p, q \in \mathbb{C}$. A contour from $p$ to $q$ is a continuous rectifiable curve ($\gamma$) in $\mathbb{C} \cong \mathbb{R}^2$ from $p$ to $q$.

X.2.1.2. A contour from $p$ to $q$ by definition has a continuous parametrization $\zeta : [a, b] \rightarrow \mathbb{C}$ with $\zeta(a) = p$, $\zeta(b) = q$. However, the parametrization is not unique, and is not intrinsic to $\gamma$ (only the equivalence class of the parametrization is part of the definition). But since $\gamma$ is rectifiable, it has an intrinsic parametrization by arc length starting at $p$ ($\gamma$). More importantly, $\gamma$ has an orientation, or specified direction of motion, with initial point $p$ and final point $q$. There is an opposite contour $-\gamma$, which is $\gamma$ traversed from $q$ to $p$. If $\zeta : [a, b] \rightarrow \mathbb{C}$ is a parametrization of $\gamma$, then $\tilde{\zeta} : [0, b - a] \rightarrow \mathbb{C}$ defined by $\tilde{\zeta}(t) = \zeta(b - t)$ is a parametrization of $-\gamma$. The interval a parametrization is defined on is not important, and can be changed to any other (closed bounded) interval by a linear change of variables.

We will often consider closed contours, ones whose initial and final points coincide. A closed contour still has an orientation, and if $\gamma$ is a simple closed contour, i.e. the image is homeomorphic to a circle, it also makes sense to say how many (net) times $\gamma$ is traversed in a counterclockwise direction ($\gamma$).

We will almost always concern ourselves only with the special case where $\gamma$ is piecewise-smooth, i.e. has a piecewise-smooth parametrization ($\gamma$). If $\gamma$ is piecewise-smooth, parametrization by arc length is a piecewise-smooth parametrization ($\gamma$).

X.2.1.3. The goal of this section is to show that if $\gamma$ is a contour in $\mathbb{C}$, and $f$ is a continuous complex-valued function on $\mathbb{C}$, then there is a well-defined complex number called the (contour) integral of $f$ along $\gamma$, denoted

$$\int_{\gamma} f = \int_{\gamma} f(\zeta) d\zeta$$

with the usual properties of an integral. The integral does not depend on a parametrization of $\gamma$, although it is usually calculated using a parametrization. It does, however, depend on the orientation of $\gamma$: we will have

$$\int_{-\gamma} f = -\int_{\gamma} f.$$

Although contour integrals of somewhat more general functions can be defined, we will restrict to the case of continuous functions for simplicity, and because this is the crucial case for applications. In fact, almost all applications involve only contour integrals of holomorphic functions.
Integration on an Interval

Definition of the Contour Integral

Properties of Contour Integrals

The Fundamental Inequality

The next fundamental estimate is an easy consequence of the definition:

\textbf{X.2.1.4. Proposition.} Let \( \gamma : [a, b] \to \mathbb{C} \) be a rectifiable curve, and \( f \) a continuous function on \( \gamma \). If \( |f(z)| \leq M \) for all \( z \in \gamma \), then

\[
\left| \int_{\gamma} f \right| \leq M \cdot \ell(\gamma) .
\]

\textbf{Proof:} Let \( \mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\} \) be a partition of \([a, b]\). Then by the triangle inequality we have

\[
\left| \sum_{k=1}^{n} f(\gamma(t_k))(\gamma(t_k) - \gamma(t_{k-1})) \right| \leq \sum_{k=1}^{n} |f(\gamma(t_k))||\gamma(t_k) - \gamma(t_{k-1})| \leq M \sum_{k=1}^{n} |\gamma(t_k) - \gamma(t_{k-1})| .
\]

As the mesh of \( \mathcal{P} \) goes to zero, the left side approaches \( \int_{\gamma} f \) and the right side approaches \( M \cdot \ell(\gamma) \). \( \square \)

An important consequence is:

\textbf{X.2.1.5. Corollary.} Let \( \gamma : [a, b] \to \mathbb{C} \) be a rectifiable curve, and \((f_n)\) a sequence of continuous functions on \( \gamma \) converging uniformly on \( \gamma \) to a (continuous) function \( f \). Then

\[
\int_{\gamma} f_n \to \int_{\gamma} f .
\]

Calculation of Contour Integrals

On piecewise-smooth curves, contour integrals can be computed by an ordinary integral:

\textbf{X.2.1.6. Theorem.} If \( \gamma : [a, b] \to \mathbb{C} \) is a piecewise-smooth curve in \( \mathbb{C} \) and \( f \) is a complex-valued function on \( \gamma \), then

\[
\int_{\gamma} f = \int_{a}^{b} f(\gamma(t))\gamma'(t) \, dt .
\]
X.2.1.7. Example. Here is perhaps the most important example of a contour integral. Let $\gamma$ be the unit circle in $\mathbb{C}$, parametrized once around counterclockwise. Let $f(z) = \frac{1}{z}$ for $|z| = 1$. Then

$$\int_{\gamma} f = \int_{0}^{2\pi} \frac{1}{e^{it}} e^{it} dt = 2\pi i .$$

This integral will appear over and over in many parts of complex analysis, and many of the fundamental theorems and formulas of the subject are based on it.

The Fundamental Theorem of Calculus

The contour integral of a function with an antiderivative is easy to compute:

X.2.1.8. Proposition. If $f$ is a continuous function on an open set $\Omega$ in $\mathbb{C}$ with an antiderivative $F$ on $\Omega$, and $\gamma : [a, b] \to \Omega$ is a piecewise-smooth contour from $p$ to $q$, then

$$\int_{\gamma} f = F(q) - F(p) .$$

In particular, the contour integral of $f$ along any contour depends only on the endpoints. The contour integral of $f$ around any closed contour is 0.

Proof: This just looks like the Fundamental Theorem of Calculus, but it does not follow immediately from the real FTC but requires a bit of calculation using the Chain Rule of vector calculus and the Cauchy-Riemann equations. By subdividing $[a, b]$, we may assume $\gamma$ is smooth. Write

$$F(x + iy) = U(x, y) + iV(x, y)$$

so that

$$f(x + iy) = \frac{\partial U}{\partial x}(x, y) + i \frac{\partial V}{\partial x}(x, y) .$$

Write

$$\gamma(t) = \gamma_1(t) + i\gamma_2(t)$$
$$G(t) = U(\gamma_1(t), \gamma_2(t))$$
$$H(t) = V(\gamma_1(t), \gamma_2(t))$$

so that $F(\gamma(t)) = G(t) + iH(t)$. By the vector calculus Chain Rule,

$$G'(t) = \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t))\gamma_1'(t) + \frac{\partial U}{\partial y}(\gamma_1(t), \gamma_2(t))\gamma_2'(t)$$
$$H'(t) = \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t))\gamma_1'(t) + \frac{\partial V}{\partial y}(\gamma_1(t), \gamma_2(t))\gamma_2'(t) .$$

Then

$$\int_{\gamma} f = \int_{a}^{b} f(\gamma(t)) \gamma'(t) dt$$
\[\int_a^b \left[ \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t)) + i \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t)) \right] \left[ \gamma_1'(t) + i \gamma_2'(t) \right] dt\]

\[= \int_a^b \left[ \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_1'(t) - \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_2'(t) \right] dt + i \int_a^b \left[ \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_1'(t) + \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_2'(t) \right] dt\]

\[= \int_a^b \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_1'(t) dt + i \int_a^b \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t)) \gamma_1'(t) dt + \int_a^b \frac{\partial U}{\partial y}(\gamma_1(t), \gamma_2(t)) \gamma_2'(t) dt + i \int_a^b \frac{\partial V}{\partial y}(\gamma_1(t), \gamma_2(t)) \gamma_2'(t) dt\]

\[= \int_a^b G'(t) dt + i \int_a^b H'(t) dt = (G(b) - G(a)) + i(H(b) - H(a)) = F(q) - F(p).
\]

The Complex Mean Value Inequality

The next result is a complex version of the Mean Value Inequality (V.8.2.10):

**X.2.1.9. PROPOSITION.** Let \( f \) be a holomorphic function on an open set \( \Omega \) in \( \mathbb{C} \) (with \( f' \) continuous), and \( A \) a convex subset of \( \Omega \). If \( |f'(z)| \leq M \) for all \( z \in A \), then

\[ |f(z) - f(w)| \leq M|z - w| \]

for all \( z, w \in A \).

The requirement that \( f' \) be continuous turns out to be automatic (X.3.3.6.).

**PROOF:** Let \( \gamma \) be the straight line segment from \( w \) to \( z \); then \( \gamma \subseteq A \) since \( A \) is convex.

\[ |f(z) - f(w)| = \left| \int_\gamma f' \right| \leq M|\ell(\gamma)| = M|z - w| . \]

Relation with Line Integrals

Green’s Theorem for Complex Functions

**X.3. The Big Theorems**

In this section, we use contour integration to prove some of the big theorems of elementary Complex Analysis: Cauchy’s Theorem, the Cauchy Integral Formula, Morera’s Theorem, Cauchy’s Inequalities, and Liouville’s Theorem.

These are, of course, not all the theorems of Complex Analysis which can be called “big theorems.” A few more big theorems concerning power series representation and analysis of singularities will be discussed in the next sections. Other big theorems such as the Riemann Mapping Theorem (and many others) are discussed later or are beyond the scope of this book.
X.3.1. Goursat’s Theorem

We begin by proving a special case of Cauchy’s Theorem, due to E. Goursat in 1883. This result is a considerable technical improvement over Cauchy’s original theorem, since it assumes nothing about the properties of the derivative of a holomorphic function, and in particular makes no continuity assumption about the derivative (in fact, it will turn out later that the derivative of a holomorphic function is not only automatically continuous, but even itself holomorphic, but we do not want to assume anything about the derivative in developing the theory).

The form of Goursat’s Theorem we give (originally due to A. Pringsheim in 1901) uses triangles. It can be equally well stated and proved for rectangles, or for any other polygonal region that can be subdivided into a finite number of congruent pieces of smaller diameter and perimeter (Goursat used squares). The use of triangles gives a more flexible result which simplifies succeeding proofs.

X.3.1.1. Theorem. [Goursat’s Theorem] Let $\Omega$ be an open set in $\mathbb{C}$, and $f$ a holomorphic function on $\Omega$. If $T$ is a (solid) triangle contained in $\Omega$, and $\gamma$ is the boundary of $T$ parametrized once around counterclockwise, then

$$\int_{\gamma} f = 0.$$ 

Proof: Let $d$ and $p$ be the diameter and perimeter of $T$ respectively. Divide $T$ into four subtriangles $T_{1}^{(1)}, \ldots, T_{1}^{(4)}$ whose vertices are the vertices of $T$ and the midpoints of the edges of $T$. Then each $T_{1}^{(k)}$ is similar to $T$. If $\gamma_{1}^{(k)}$ is the perimeter of $T_{1}^{(k)}$, parametrized once around counterclockwise, then we have

$$\int_{\gamma} f = \int_{\gamma_{1}^{(1)}} f + \int_{\gamma_{1}^{(2)}} f + \int_{\gamma_{1}^{(3)}} f + \int_{\gamma_{1}^{(4)}} f$$

since all interior edges are traversed once in each direction, so those contour integrals cancel. Thus there is a $k$ such that

$$\left| \int_{\gamma} f \right| \leq 4 \left| \int_{\gamma_{1}^{(k)}} f \right|.$$ 

Set $T_{1} = T_{1}^{(k)}$ and $\gamma_{1} = \gamma_{1}^{(k)}$ for this $k$. Repeat the process inductively using $T_{n}$ in place of $T$ to obtain a decreasing sequence $(T_{n})$ of triangles with boundary $\gamma_{n}$, with

$$\left| \int_{\gamma} f \right| \leq 4^{n} \left| \int_{\gamma_{n}} f \right|$$

for all $n$. We have that the diameter $d_{n}$ and perimeter $p_{n}$ of $T_{n}$ satisfy $d_{n} = 2^{-n}d$ and $p_{n} = 2^{-n}p$. We have that

$$\cap_{n}T_{n} = \{ z_{0} \}$$

for some $z_{0} \in T$.

Since $f$ is differentiable at $z_{0}$, for $z \in \Omega \setminus \{ z_{0} \}$ we may set

$$g(z) = \frac{f(z) - f(z_{0})}{z - z_{0}} - f'(z_{0}).$$


911
If we set $g(z_0) = 0$, then $g$ is continuous on $\Omega$ and we may write
\[
f(z) = f(z_0) + f'(z_0)(z - z_0) + g(z)(z - z_0)
\]
for $z \in \Omega$. Thus we have
\[
\int_{\gamma_n} f(\zeta) \, d\zeta = \int_{\gamma_n} f(z_0) \, d\zeta + \int_{\gamma_n} f'(z_0)(\zeta - z_0) \, d\zeta + \int_{\gamma_n} g(\zeta)(\zeta - z_0) \, d\zeta
\]
for any $n$. The first two integrals on the right are always zero since the integrands have antiderivatives on $\Omega$ (they are a constant function and a linear function respectively); thus for any $n$ we have
\[
\int_{\gamma_n} f(\zeta) \, d\zeta = \int_{\gamma_n} g(\zeta)(\zeta - z_0) \, d\zeta.
\]

Let $\epsilon > 0$. Choose $\delta > 0$ such that $|g(z)| < \epsilon$ whenever $|z - z_0| < \delta$. Then for any $n$ large enough that $d_n < \delta$, we have
\[
\left| \int_{\gamma_n} f(\zeta) \, d\zeta \right| = \left| \int_{\gamma_n} g(\zeta)(\zeta - z_0) \, d\zeta \right| \leq \epsilon d_n p_n.
\]
Combining the inequalities, for $n$ sufficiently large we have
\[
\left| \int_{\gamma} f \right| \leq \epsilon \left| \int_{\gamma_n} f \right| \leq 4^n \epsilon d_n p_n = \epsilon dp.
\]
Since $\epsilon > 0$ is arbitrary, we obtain $\int_{\gamma} f = 0$. \(\square\)

It will be convenient to have a technical strengthening of this theorem:

**X.3.1.2. COROLLARY.** Let $\Omega$ be an open set in $\mathbb{C}$, \{z$_1$, \ldots, z$_n$\} a finite set of points in $\Omega$, and $f$ a holomorphic function on $\Omega \setminus \{z_1, \ldots, z_n\}$. Suppose
\[
\lim_{z \to z_k} f(z)(z - z_k) = 0
\]
for $1 \leq k \leq n$. If $T$ is a (solid) triangle contained in $\Omega$, and the boundary $\gamma$ of $T$ (parametrized once around counterclockwise) contains no $z_k$, then
\[
\int_{\gamma} f = 0.
\]

**PROOF:** $T$ can be subdivided into smaller triangles, with some of the subtriangles $T_1, \ldots, T_m$ each containing exactly one of the $z_k$ in its interior, and the rest of the subtriangles contained in $\Omega \setminus \{z_1, \ldots, z_n\}$. Let $\gamma_j$ be the boundary of $T_j$ ($1 \leq j \leq m$). Then
\[
\int_{\gamma} f = \sum_{j=1}^{n} \int_{\gamma_j} f.
\]
since the integrals around the other triangles are 0 and all edges not part of \( \gamma \) or one of the \( \gamma_j \) are traversed once in each direction. Thus it suffices to show that \( \int_{\gamma_j} f = 0 \) for each \( j \), i.e. it suffices to show the result in the case where \( T \) contains exactly one \( z_k \) in its interior.

Changing notation, we assume \( z_j \) is the only \( z_k \) in the interior of \( T \). By a similar subdivision argument, the integral of \( f \) around any other triangle in \( \Omega \) containing only \( z_j \) in its interior is the same. Thus we may assume \( T \) has arbitrarily small diameter \( d \) and circumference \( p \).

Let \( \epsilon > 0 \), and choose \( > 0 \) such that \( |f(z)(z-z_j)| < \epsilon \) whenever \( 0 < |z-z_j| < \delta \). Choose \( T \) with \( p < \delta \) (and hence \( d < \delta \) also). If \( \zeta \in \gamma \), we then have \( |f(\zeta)| < \frac{\epsilon}{\delta} \). We then have

\[
\left| \int_{\gamma} f \right| \leq \frac{\epsilon}{\delta} d < \epsilon.
\]

Since the integral is independent of which \( T \) is chosen, and \( \epsilon > 0 \) is arbitrary, we have \( \int_{\gamma} f = 0 \).

X.3.2. The Antiderivative Theorem and Cauchy’s Theorem for Convex Sets

**X.3.2.1. Theorem.** [Antiderivative Theorem for Convex Sets] Let \( \Omega \) be a convex open set in \( \mathbb{C} \). Let \( f \) be a complex-valued continuous function on \( \Omega \). If

\[
\int_{\gamma} f = 0
\]

for every triangular contour \( \gamma \) in \( \Omega \), then \( f \) has an antiderivative on \( \Omega \): there is a holomorphic function \( F \) on \( \Omega \) with \( F' = f \) on \( \Omega \).

**Proof:** Fix \( z_0 \in \Omega \). For any \( z \in \Omega \), define

\[
F(z) = \int_{\gamma} f
\]

where \( \gamma \) is the straight line segment from \( z_0 \) to \( z \). If \( z \) and \( w \) are in \( \Omega \), then

\[
F(w) - F(z) = \int_{\gamma} f
\]

where \( \gamma \) is the straight line segment from \( z \) to \( w \), since the integral of \( f \) around the triangle with vertices \( z_0 \), \( z \), and \( w \) is 0 by assumption.

Now fix \( z \in \Omega \), and fix \( \eta > 0 \) such that \( B_{\eta}(z) \subseteq \Omega \). If \( h \in \mathbb{C} \), \( |h| < \eta \), then

\[
\frac{F(z+h) - F(z)}{h} = \frac{1}{h} \int_{\gamma_h} f
\]

where \( \gamma_h \) is the straight line segment from \( z \) to \( z+h \). We have

\[
\int_{\gamma_h} f(\zeta) d\zeta = \int_{\gamma_h} [f(\zeta) - f(z)] d\zeta + \int_{\gamma_h} f(z) d\zeta.
\]
The second integral is the contour integral of the constant function \( f(z) \) (\( z \) is fixed), hence equals \( f(z)h \).

Since \( f \) is continuous at \( z \), for any \( \epsilon > 0 \) there is a \( \delta > 0 \), \( \delta \leq \eta \), such that \( |f(\zeta) - f(z)| < \epsilon \) whenever \( |\zeta - z| < \delta \). Thus if \( |h| < \delta \), we have

\[
\left| \int_{\gamma_h} [f(\zeta) - f(z)] \, d\zeta \right| < \epsilon |h|.
\]

Thus

\[
\lim_{h \to 0} \frac{F(z + h) - F(z)}{h} = \lim_{h \to 0} \frac{1}{h} \int_{\gamma_n} f = \lim_{h \to 0} \frac{1}{h} \int_{\gamma_h} [f(\zeta) - f(z)] \, d\zeta + \lim_{h \to 0} \frac{1}{h} \int_{\gamma_h} f(z) \, d\zeta = f(z)
\]

since the first limit on the right is 0 and the second expression is exactly \( f(z) \) for all \( h \). Thus \( F'(z) \) exists and equals \( f(z) \) for any \( z \in \Omega \).

The proof actually shows the following slightly more general version:

**X.3.2.2. Theorem.** Let \( \Omega \) be a convex open set in \( \mathbb{C} \), \( \{z_1, \ldots, z_n\} \) a finite set of points in \( \Omega \), and \( f \) a continuous function on \( \Omega \setminus \{z_1, \ldots, z_n\} \). If

\[
\int_\gamma f = 0
\]

for every triangular contour \( \gamma \) in \( \Omega \) not passing through one of the \( z_k \), then \( f \) has an antiderivative on \( \Omega \setminus \{z_1, \ldots, z_n\} \): there is a holomorphic function \( F \) on \( \Omega \setminus \{z_1, \ldots, z_n\} \) with \( F' = f \) on \( \Omega \setminus \{z_1, \ldots, z_n\} \).

**Proof:** There are only a few slight modifications needed to the proof. Take \( z_0 \in \Omega \setminus \{z_1, \ldots, z_n\} \), and for \( z \in \Omega \setminus \{z_1, \ldots, z_n\} \) we take

\[
F(z) = \int_\gamma f
\]

where \( \gamma \) is a path from \( z_0 \) to \( z \) consisting of at most two line segments in \( \Omega \setminus \{z_1, \ldots, z_n\} \) (by hypothesis, \( F \) is well defined; such angled paths will sometimes be necessary since the straight line segment from \( z_0 \) to \( z \) may pass through a \( z_k \)). If \( z \in \Omega \setminus \{z_1, \ldots, z_n\} \), choose \( \eta > 0 \) such that \( B_\eta(z) \subseteq \Omega \setminus \{z_1, \ldots, z_n\} \). The rest of the proof is identical.

There is a version for general open sets, whose proof is essentially identical:

**X.3.2.3. Theorem.** [Antiderivative Theorem] Let \( \Omega \) be an open set in \( \mathbb{C} \). Let \( f \) be a complex-valued continuous function on \( \Omega \). If

\[
\int_\gamma f = 0
\]

for every closed contour \( \gamma \) in \( \Omega \), then \( f \) has an antiderivative on \( \Omega \): there is a holomorphic function \( F \) on \( \Omega \) with \( F' = f \) on \( \Omega \).

**Proof:** We may assume \( \Omega \) is connected since \( F \) may be defined separately on each component. Fix \( z_0 \in \Omega \), and for \( z \in \Omega \) define

\[
F(z) = \int_\gamma f
\]

914
where $\gamma$ is any contour in $\Omega$ from $z_0$ to $z$. If $\gamma_1$ and $\gamma_2$ are two such contours, then

$$\int_{\gamma_1} f - \int_{\gamma_2} f = \int_{\gamma_1 - \gamma_2} f = 0$$

by assumption, so $F$ is well defined. If $z$ and $w$ lie in a convex open subset of $\Omega$ (e.g. a small disk), then

$$F(w) - F(z) = \int_{\gamma} f$$

where $\gamma$ is the straight line segment from $z$ to $w$, since $F(w)$ can be calculated by taking a contour from $z_0$ to $z$ and appending this line segment.

The rest of the proof is identical to the proof of X.3.2.1.

The advantage of X.3.2.1. on convex sets is that only triangular contours need to be checked. Putting X.3.2.1. together with Goursat’s Theorem and (1), we obtain:

**X.3.2.4.** [Cauchy’s Theorem for Convex Sets] Let $\Omega$ be a convex open set in $\mathbb{C}$, and $f$ a holomorphic function on $\Omega$. Then $f$ has an antiderivative on $\Omega$, and if $\gamma$ is any piecewise-smooth closed curve in $\Omega$, then

$$\int_{\gamma} f = 0.$$  

**X.3.2.5.** Example. This result can fail if $\Omega$ is not convex. For example, let $\Omega = \mathbb{C} \setminus \{0\}$, and $f(z) = \frac{1}{z}$. Then $f$ is holomorphic on $\Omega$, but the integral of $f$ around the unit circle is nonzero (1) and $f$ does not have an antiderivative on $\Omega$ (i.e. $\log z$ cannot be defined as a holomorphic function on all of $\Omega$).

Cauchy’s Theorem is true if $\Omega$ is simply connected (1).

We do get a slight extension by using X.3.2.2. together with X.3.1.2.:

**X.3.2.6.** Theorem. Let $\Omega$ be a convex open set in $\mathbb{C}$, $\{z_1, \ldots, z_n\}$ a finite set of points in $\Omega$, and $f$ a holomorphic function on $\Omega \setminus \{z_1, \ldots, z_n\}$. Suppose

$$\lim_{z \to z_k} f(z)(z - z_k) = 0$$

for $1 \leq k \leq n$. Then $f$ has an antiderivative on $\Omega \setminus \{z_1, \ldots, z_n\}$, and if $\gamma$ is any piecewise-smooth closed curve in $\Omega \setminus \{z_1, \ldots, z_n\}$, then

$$\int_{\gamma} f = 0.$$  

The next result can be generalized using the general form of Cauchy’s Theorem, but this form is already very useful. Circles can be replaced by convex or, more generally, starlike polygonal arcs such as triangles or rectangles with the same proof (which even simplifies in this case).
X.3.2.7. Theorem. Let $f$ be holomorphic on an open set $\Omega$, and $\gamma_1$ and $\gamma_2$ circles in $\Omega$ (not necessarily concentric), parametrized once around counterclockwise, with $\gamma_2$ contained in the interior of $\gamma_1$. If the region between $\gamma_1$ and $\gamma_2$ is contained in $\Omega$, then

$$\int_{\gamma_1} f = \int_{\gamma_2} f.$$

Note that the integrals around $\gamma_1$ and $\gamma_2$ are not necessarily 0!

Proof: The inner circle $\gamma_2$ can be subdivided into arcs with consecutive endpoints $z_1, \ldots, z_n$ such that the line segments between pairs of consecutive points, and the region between the segment and the circle, are contained in $\Omega$ (cover the circle with finitely many open disks contained in $\Omega$). Thus, if radial line segments are drawn outward from each $z_k$ to $\gamma_1$, each wedge is contained in a convex open subset of $\Omega$, and hence the contour integral of $f$ around each, once around counterclockwise, is 0:

The sum of these integrals is

$$\int_{\gamma_1} f - \int_{\gamma_2} f$$

since each radial segment is traversed once in each direction, so those contour integrals cancel. Thus

$$\int_{\gamma_1} f - \int_{\gamma_2} f = 0.$$


X.3.3. The Cauchy Integral Formula

X.3.3.1. A remarkable consequence of the theorems obtained so far is that if a function $f$ is holomorphic on and inside a circle (or more general simple closed curve) $\gamma$, then the values of $f$ on $\gamma$ completely determine the values of $f$ at all points inside $\gamma$, and in fact there is a simple formula involving a contour integral.

X.3.3.2. Theorem. [Cauchy Integral Formula] Let $f$ be holomorphic on an open set $\Omega$ in $\mathbb{C}$. Then for each $z \in \Omega$ we have

$$f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta,$$

where $\gamma$ is any circle in $\Omega$ bounding an open disk contained in $\Omega$ and containing $z$, parametrized once around counterclockwise.

Proof: Fix $z$ inside $\gamma$. If $\delta > 0$ is small enough that the circle $\gamma_\delta$ of radius $\delta$ centered at $z$ is in the interior of $\gamma$, we have

$$\int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta = \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta,$$

by X.3.2.7. Thus it suffices to prove the result for the $\gamma_\delta$, in fact to show that

$$f(z) = \lim_{\delta \to 0} \frac{1}{2\pi i} \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta.$$
Let $\epsilon > 0$. Since $f$ is continuous, if $\delta$ is sufficiently small then $|f(\zeta) - f(z)| < \epsilon$ for $\zeta \in \gamma_\delta$. We then have, for $\gamma_\delta$ inside $\gamma$,

$$\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta = \int_{\gamma_\delta} \frac{f(\zeta) - f(z)}{\zeta - z} \, d\zeta + \int_{\gamma_\delta} \frac{f(z)}{\zeta - z} \, d\zeta .$$

If $\delta$ is small enough that $|f(\zeta) - f(z)| < \epsilon$ for $\zeta \in \gamma_\delta$, the first integral satisfies

$$\left| \int_{\gamma_\delta} \frac{f(\zeta) - f(z)}{\zeta - z} \, d\zeta \right| \leq \frac{\epsilon}{\delta} \cdot 2\pi \delta = 2\pi \epsilon$$

and thus the first integral approaches 0 as $\delta \to 0$. The second integral can be computed by taking $\zeta(t) = z + \delta e^{it}$, $0 \leq t \leq 2\pi$; thus

$$\int_{\gamma_\delta} \frac{f(z)}{\zeta - z} \, d\zeta = f(z) \int_{0}^{2\pi} \frac{1}{\delta e^{it}} i\delta e^{it} \, dt = 2\pi i f(z) .$$

An interesting reformulation is that if $D$ is a closed disk and $f$ is holomorphic on and inside $D$, then the value of $f$ at the center of the disk is the average of the values on the boundary circle:

**X.3.3.3. Corollary.** Let $f$ be holomorphic on an open set $\Omega$, and $D$ a closed disk in $\Omega$ of radius $r$, centered at $z_0$. Then

$$f(z_0) = \frac{1}{2\pi} \int_{0}^{2\pi} f(z_0 + r e^{it}) \, dt .$$

**Proof:** Parametrize the boundary circle by $\zeta(t) = z_0 + r e^{it}$, $0 \leq t \leq 2\pi$. Then

$$f(z_0) = \frac{1}{2\pi i} \int_{0}^{2\pi} \frac{f(z_0 + r e^{it})}{r e^{it}} i\delta e^{it} \, dt .$$

**The Cauchy Integral Formula for Derivatives**

**X.3.3.4.** We now obtain another remarkable consequence of the Cauchy Integral Formula. The integrand in the formula is a function of $z$ and $\zeta$; since we integrate with respect to $\zeta$, it should be possible to differentiate with respect to $z$ under the integral sign, in analogy with XIV.3.5.1., to obtain an integral formula for $f'$. In fact, since the integrand is infinitely differentiable in $z$, we could expect that the function defined by the integral, namely $f$, is also infinitely differentiable and all its derivatives are obtained by successive differentiation under the integral sign. This expectation is valid (X.3.3.6.). In fact, a more general result along the same lines holds, which is also useful for other purposes:
X.3.3.5. **Theorem.** Let $\gamma$ be a contour in $\mathbb{C}$ (not necessarily a closed contour), and $\phi$ a complex-valued continuous function on $\gamma$. For $z \in \mathbb{C} \setminus \gamma$, define

$$f(z) = \int_\gamma \frac{\phi(\zeta)}{\zeta - z} \, d\zeta.$$ 

Then $f$ is infinitely differentiable on $\mathbb{C} \setminus \gamma$, and for each $z \in \mathbb{C} \setminus \gamma$ and each $n$ we have

$$f^{(n)}(z) = n! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z)^{n+1}} \, d\zeta.$$ 

**Proof:** By induction on $n$. The case $n = 0$ holds by assumption. Assume the result is true for $n - 1$, i.e. for any $z \in \mathbb{C} \setminus \gamma$,

$$f^{(n-1)}(z) = (n - 1)! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z)^n} \, d\zeta.$$ 

Fix $z \in \mathbb{C} \setminus \gamma$, and fix $\delta > 0$ such that the closed disk of radius $\delta$ around $z$ is contained in $\mathbb{C} \setminus \gamma$. If $|h| < \delta$, then

$$\frac{f^{(n-1)}(z + h) - f^{(n-1)}(z)}{h} = (n - 1)! \int_\gamma \frac{\phi(\zeta)}{h} \left[ \frac{1}{(\zeta - z - h)^n} - \frac{1}{(\zeta - z)^n} \right] \, d\zeta.$$ 

Set

$$a = \frac{1}{\zeta - z - h}, \quad b = \frac{1}{\zeta - z}$$

and recall the formula

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1})$$

(the second factor is 1 if $n = 1$). We have

$$a - b = \frac{1}{\zeta - z - h} - \frac{1}{\zeta - z} = \frac{h}{(\zeta - z - h)(\zeta - z)}$$

so we have

$$\frac{f^{(n-1)}(z + h) - f^{(n-1)}(z)}{h} = (n - 1)! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z - h)(\zeta - z)} \left[ a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1} \right] \, d\zeta.$$ 

Since the disk of radius $\delta$ around $z$ is bounded away from $\gamma$, as $h \to 0$ we have that $\frac{1}{\zeta - z - h}$ converges uniformly in $\zeta$ to $\frac{1}{\zeta - z}$ on $\gamma$ (i.e. $a \to b$ uniformly on $\gamma$). Since the other factors $\phi(\zeta)$ and $\frac{1}{\zeta - z}$ are uniformly bounded on $\gamma$ and do not depend on $h$, as $h \to 0$ we have

$$\frac{\phi(\zeta)}{(\zeta - z - h)(\zeta - z)} \left[ a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1} \right] \to \frac{\phi(\zeta)}{(\zeta - z)^2} \cdot nb^{n-1} = n \cdot \frac{f(\zeta)}{(\zeta - z)^{n+1}}$$

uniformly in $\zeta$ on $\gamma$. Thus the integrals around $\gamma$ also converge. So

$$f^{(n)}(z) = \lim_{h \to 0} \frac{f^{(n-1)}(z + h) - f^{(n-1)}(z)}{h}$$

918
exists and equals
\[
\lim_{h \to 0} (n-1)! \int_{\gamma} \frac{\phi(\zeta)}{(\zeta - z - h)(\zeta - z)} \left[ a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1} \right] d\zeta = n! \int_{\gamma} \frac{\phi(\zeta)}{(\zeta - z)^{n+1}} d\zeta.
\]

**X.3.3.6. Corollary. [Cauchy Integral Formula for Derivatives]** Let \( f \) be holomorphic on an open set \( \Omega \) in \( \mathbb{C} \). Then \( f \) is infinitely differentiable on \( \Omega \), and for each \( z \in \Omega \) and each \( n \) we have
\[
f^{(n)}(z) = \frac{n!}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta
\]
where \( \gamma \) is any circle in \( \Omega \) bounding an open disk \( D \) contained in \( \Omega \) and containing \( z \), parametrized once around counterclockwise.

**Proof:** For \( \zeta \in \gamma \), set \( \phi(\zeta) = \frac{f(\zeta)}{2\pi i} \). Then, for \( z \) inside \( \gamma \), by X.3.3.2.
\[
f(z) = \int_{\gamma} \frac{\phi(\zeta)}{\zeta - z} d\zeta.
\]

**X.3.3.7.** Note that in the special case of X.3.3.6., for \( \zeta \in \gamma \) we have that \( \phi(\zeta) \) is, up to a factor of \( 2\pi i \), the limit of \( f(z) \) as \( z \to \zeta \) from inside the circle. But in the general situation of X.3.3.5., there is no such simple relationship (for example, in the situation of X.3.3.6., the function defined by the integral is identically 0 outside \( \gamma \) by (i)). In fact, there is no way to recover \( \phi \) from \( f \) in general, although there is in many cases (i).

**Morera’s Theorem**

Morera’s Theorem is a converse to Cauchy’s Theorem:

**X.3.3.8. Theorem. [Morera’s Theorem]** Let \( \Omega \) be an open set in \( \mathbb{C} \), and \( f \) a complex-valued continuous function on \( \Omega \). If the contour integral of \( f \) around the boundary of every solid triangle contained in \( \Omega \) is zero, then \( f \) is holomorphic on \( \Omega \).

**Proof:** It suffices to show that \( f \) is holomorphic in every open disk contained in \( \Omega \), i.e. we may assume \( \Omega \) is a disk. Then \( f \) has an antiderivative \( F \) on \( \Omega \) by X.3.2.1.. But \( F \) is infinitely differentiable on \( \Omega \), hence so is \( F' = f \).

**Cauchy’s Inequalities**

Cauchy’s Inequalities are an immediate consequence of the Cauchy Integral Formula and the usual estimate for the size of a contour integral. They say that the higher derivatives of a holomorphic do not grow too rapidly at any point.
X.3.3.9. **Theorem.** [Cauchy’s Inequalities] Let $f$ be holomorphic on an open set $\Omega$ in $\mathbb{C}$, and $z_0 \in \Omega$. If the closed disk of radius $R$ centered at $z_0$ is contained in $\Omega$, then for all $n$

$$|f^{(n)}(z_0)| \leq \frac{n!}{R^n} \max_{|\zeta - z_0| = R} |f(\zeta)| .$$

Note that the quantity $\max_{|\zeta - z_0| = R} |f(\zeta)|$, the maximum of $|f|$ on the boundary circle, depends on $R$ but not on $n$.

**Liouville’s Theorem**

An immediate consequence of Cauchy’s Inequality for $n = 1$ is the following remarkable result:

X.3.3.10. **Theorem.** [Liouville’s Theorem] A bounded entire function is constant.

**Proof:** Let $f$ be a bounded entire function, and fix $M$ with $|f(z)| \leq M$ for all $z \in \mathbb{C}$. Fix $z_0 \in \mathbb{C}$. Then for any $R > 0$ we have

$$|f'(z_0)| \leq \frac{M}{R}$$

and thus $f'(z_0) = 0$. Since $z_0 \in \mathbb{C}$ is arbitrary, $f'$ is identically 0.

This can be used, among many other things, to give a proof of the Fundamental Theorem of Algebra (Exercise X.8.2.3.).

The result can be sharpened:

X.3.3.11. **Corollary.** Let $f$ be a nonconstant entire function. Then the range of $f$ is dense in $\mathbb{C}$.

**Proof:** Suppose the range is not dense. Then there is a $w \in \mathbb{C}$ and an $\epsilon > 0$ such that $|f(z) - w| \geq \epsilon$ for all $z \in \mathbb{C}$. Set

$$g(z) = \frac{1}{f(z) - w} .$$

Then $g$ is an entire function, and $|g(z)| \leq \frac{1}{\epsilon}$ for all $z$. Thus $g$, and hence $f$, is constant, a contradiction.

In fact, the following stronger result holds. See e.g. [] for a proof.

X.3.3.12. **Theorem.** [Little Picard Theorem] Let $f$ be an entire function. Then the range of $f$ is all of $\mathbb{C}$ except possibly one point.

This theorem is the best possible: $f(z) = e^z$ is never 0. See X.5.2.19. for the “Big Picard Theorem” which is an even stronger statement.
X.3.4. The Homotopy Theorem for Contour Integrals

Cauchy’s Theorem roughly states that the contour integral of a holomorphic function along a path \( \gamma \) depends only on the endpoints of \( \gamma \) and not on which path is chosen, or alternatively that the contour integral around a closed contour is 0. But this is not literally true in general ().

To prove, or even state, the general version of Cauchy’s Theorem, we will show that if \( f \) is a holomorphic function on a region \( \Omega \) and the path \( \gamma \) is “continuously deformed” within \( \Omega \) leaving the endpoints fixed, the contour integral along \( \gamma \) does not change. This will imply that the integral around any closed curve in \( \Omega \) which can be “continuously shrunk” to the initial point within \( \Omega \) is zero.

The careful formulation of a “continuous deformation” is a homotopy:

X.3.4.1. Definition. Let \( \Omega \) be an open set in \( \mathbb{C} \), \( p,q \in \Omega \). Let \( \gamma_0 \) and \( \gamma_1 \) be contours from \( p \) to \( q \) in \( \Omega \) with parametrizations \( \zeta_0(t), \zeta_1(t) \) defined on the same interval \([a,b]\). A homotopy from \( \gamma_0 \) to \( \gamma_1 \) in \( \Omega \) is a continuous function \( h : [0,1] \times [a,b] \to \Omega \) such that \( h(0,t) = \zeta_0(t) \) and \( h(1,t) = \zeta_1(t) \) for all \( t \in [a,b] \), and \( h(s,0) = p \) and \( h(s,1) = q \) for all \( s \in [0,1] \). We say \( \gamma_0 \) and \( \gamma_1 \) are homotopic (relative to \( p \) and \( q \) ) if there is a homotopy from \( \gamma_0 \) to \( \gamma_1 \).

Technically, what we have defined is a homotopy between the parametrizations \( \zeta_0 \) and \( \zeta_1 \) relative to \( p \) and \( q \), but by slight abuse of terminology we will call it a homotopy from \( \gamma_0 \) to \( \gamma_1 \).

X.3.4.2. If \( s \) is fixed, then \( \zeta_s(t) = h(s,t) \) defines a parametrized curve \( \gamma_s \) in \( \Omega \) from \( p \) to \( q \), and as \( s \) varies these curves “move continuously” from \( \gamma_0 \) to \( \gamma_1 \), and conversely if “continuous deformation” is defined reasonably, the function \( h(s,t) = \zeta_s(t) \) is a homotopy from \( \gamma_0 \) to \( \gamma_1 \). Thus \( \gamma_0 \) and \( \gamma_1 \) are homotopic in \( \Omega \) if and only if \( \gamma_0 \) can be “continuously deformed” to \( \gamma_1 \) within \( \Omega \).

X.3.4.3. A closed contour \( \gamma \) in an open set \( \Omega \subseteq \mathbb{C} \) is contractible in \( \Omega \) if it is homotopic in \( \Omega \) to a constant contour, i.e. if \( \gamma \) is parametrized by \( \zeta : [a,b] \to \mathbb{C} \) with \( \zeta(a) = \zeta(b) = p \), there is a continuous \( h : [0,1] \times [a,b] \to \Omega \) with \( h(0,t) = \zeta(t) \) and \( h(1,t) = h(1,0) \) for all \( t \) and \( h(s,0) = h(s,1) \) for all \( s \). If \( \gamma \) is contractible, there is such a homotopy with \( h(s,0) = h(s,1) = h(1,t) \) for all \( t \), i.e. the point \( p \) is fixed as the base point of each contour \( \gamma_s \).

X.3.4.4. Theorem. [Homotopy Theorem for Contour Integrals] Let \( f \) be holomorphic on an open set \( \Omega \), and \( p,q \in \Omega \). If \( \gamma_0 \) and \( \gamma_1 \) are homotopic contours in \( \Omega \) from \( p \) to \( q \), then

\[
\int_{\gamma_0} f = \int_{\gamma_1} f.
\]

Proof: We first give a simple argument which “almost” proves the theorem but which has a serious flaw. (This argument is represented as a proof of the theorem in some Complex Analysis texts.)

Suppose \( \zeta_0 \) and \( \zeta_1 \) are parametrizations of \( \gamma_0 \) and \( \gamma_1 \) on the same interval, which we may take to be \([0,1]\) to simplify notation, and \( h : [0,1]^2 \to \Omega \) a homotopy from \( \zeta_0 \) to \( \zeta_1 \) relative to \( p,q \). Fix partitions \( \mathcal{P} = \{0 = s_0, s_1, \ldots, s_m = 1\} \) and \( \mathcal{Q} = \{0 = t_0, t_1, \ldots, t_n = 1\} \) of \([0,1]\). For each \( j \) and \( k \) define the curvilinear “rectangle”

\[ R_{jk} = \{ h(s,t) : s_{j-1} \leq s \leq s_j, t_{k-1} \leq t \leq t_k \} \]
(this curvilinear “rectangle” is a curvilinear “triangle” if \( k = 1 \) or \( k = n \)). We have that the image of \( h \) is compact, so there is an \( \epsilon > 0 \) such that the ball of radius \( \epsilon \) around any point of the image of \( h \) is contained in \( \Omega \). Since \( h \) is uniformly continuous on \([0,1]^2\), there is a \( \delta > 0 \) such that \( |h(s',t') - h(s,t)| < \epsilon \) whenever \( s,s',t,t' \in [0,1] \), \( |s - s'| < \delta \), and \( |t - t'| < \delta \). Thus if \( P \) and \( Q \) are sufficiently fine, for each \( j \) and \( k \) the “rectangle” \( R_{jk} \) is contained within an open disk in \( \Omega \). So if we integrate \( f \) around the boundary curve \( \gamma_{jk} \) of \( R_{jk} \) obtained by setting \( s = s_{j-1} \) and letting \( t \) vary from \( t_{k-1} \) to \( t_k \), and similarly for the other edges, we have that

\[
\int_{\gamma_{jk}} f = 0
\]

by X.3.2.4. But

\[
\sum_{j,k} \int_{\gamma_{jk}} f = \int_{\gamma_0} f - \int_{\gamma_1} f
\]

since all the other edges are traversed once each direction, so those contour integrals cancel.

The problem with this argument is that while the \( \gamma_{jk} \) are continuous curves, they are not piecewise-smooth or even rectifiable in general, so \( \int_{\gamma_{jk}} f \) is not even defined in general. (In fact, this “curve” does not even have to look anything like an ordinary curve topologically; it is just a continuous image of \([0,1]\).) To make this proof work as written, the homotopy \( h \) must be “piecewise-smooth,” and it must be shown that homotopic piecewise-smooth contours are actually homotopic by a piecewise-smooth homotopy. This can be shown to be true (the argument below can be developed into a proof).

We will proceed differently, avoiding the difficulty by a trick (see [?], p. 338 for a different way around the problem). Choose the partitions and define the \( R_{jk} \) as above. However, replace \( \gamma_{jk} \) by the curve \( \tilde{\gamma}_{jk} \) consisting of the straight line segments from \( h(s_{j-1},t_{k-1}) \) to \( h(s_j,t_{k-1}) \), then to \( h(s_j,t_k) \), then to \( h(s_{j-1},t_k) \), then back to \( h(s_{j-1},t_{k-1}) \) (some of these line segments may be degenerate). If \( j = 1 \) or \( j = m \), one of the edge curves of \( R_{jk} \) is a piece of \( \gamma_0 \) or \( \gamma_1 \), and we use this piece in \( \tilde{\gamma}_{jk} \) instead of the straight line segment.

Note that since \( R_{jk} \) is contained in a disk in \( \Omega \), which is convex, these line segments are also contained in the same disk. Then for each \( j \) and \( k \), \( \tilde{\gamma}_{jk} \) is a piecewise-smooth closed contour contained in a disk inside \( \Omega \). We thus have

\[
\int_{\tilde{\gamma}_{jk}} f = 0
\]

by X.3.2.4. But

\[
\sum_{j,k} \int_{\tilde{\gamma}_{jk}} f = \int_{\gamma_0} f - \int_{\gamma_1} f
\]

since all the other edges are traversed once each direction, so those contour integrals cancel.

\[
\text{The Winding Number of a Closed Curve}
\]

If \( \gamma \) is a closed contour and \( z \) is a point of \( \mathbb{C} \) not on \( \gamma \), then intuitively \( \gamma \) “winds around \( z \)” some net number of times in the counterclockwise sense (this number can be zero or negative as well as positive). If \( \gamma \) is a circle parametrized once around counterclockwise, this winding number should be 1 if \( z \) is inside \( \gamma \) and 0 if \( z \) is outside. We can make this notion precise:
**X.3.4.5.** Proposition. Let $\gamma$ be a closed piecewise-smooth contour in $\mathbb{C}$, and $z$ a point of $\mathbb{C}$ not on $\gamma$. Then
\[ \int_{\gamma} \frac{1}{\zeta - z} \, d\zeta \]
is an integer multiple of $2\pi i$.

**Proof:** [Ahl78] Let $\zeta : [a, b] \to \mathbb{C}$ be a piecewise-smooth parametrization of $\gamma$ with $\zeta(a) = \zeta(b) = p$. Then
\[ \int_{\gamma} \frac{1}{\zeta - z} \, d\zeta = \int_{a}^{b} \frac{\zeta'(t)}{\zeta(t) - z} \, dt . \]
For $a \leq x \leq b$, set
\[ \phi(x) = \int_{a}^{x} \frac{\zeta'(t)}{\zeta(t) - z} \, dt . \]
Then $\phi$ is differentiable at $t$ and
\[ \phi'(t) = \frac{\zeta'(t)}{\zeta(t) - z} \]
whenever $a \leq t \leq b$ and $\zeta'$ is continuous at $t$, i.e. for all but possibly finitely many $t \in [a, b]$. The function
\[ \psi(t) = e^{-\phi(t)}(\zeta(t) - z) \]
is continuous on $[a, b]$, and satisfies
\[ \psi'(t) = -e^{-\phi(t)}\phi'(t)(\zeta(t) - z) + e^{-\phi(t)}\zeta'(t) = 0 \]
whenever $\zeta'$ is continuous. Thus $\psi$ is constant on $[a, b]$. We have $\psi(a) = p - z$, so
\[ e^{\phi(t)} = \frac{\zeta(t) - z}{p - z} \]
for $a \leq t \leq b$. We have $e^{\phi(b)} = 1$, so $\phi(b) = 2\pi in$ for some $n \in \mathbb{Z}$. \hfill \(\Box\)

**X.3.4.6.** Definition. Let $\gamma$ be a closed piecewise-smooth contour in $\mathbb{C}$, and $z \in \mathbb{C} \setminus \gamma$. The **winding number** of $\gamma$ around $z$ is
\[ w(\gamma, z) = \frac{1}{2\pi i} \int_{\gamma} \frac{1}{\zeta - z} \, d\zeta . \]

By X.3.4.5., $w(\gamma, z)$ is an integer.

**X.3.4.7.** Proposition. Fix a closed piecewise-smooth contour $\gamma$. The function $\omega(z) = w(\gamma, z)$ is constant on connected components of $\mathbb{C} \setminus \gamma$ and is zero on the unbounded component.

**Proof:** The function $\omega$ is clearly a continuous function of $z$, and
\[ \lim_{z \to \infty} |\omega(z)| = 0 . \]
Since $\omega$ takes only integer values, the result follows. \hfill \(\Box\)
There is another interpretation of $w(\gamma, z)$. Let $T$ be the unit circle in $\mathbb{C}$. Given a parametrization $\zeta$ of $\gamma$, which we may assume is on $[0, 1]$, define

$$\tilde{\gamma}(t) = \frac{\zeta(t) - z}{|\zeta(t) - z|}$$

(this closed contour is essentially the radial projection of $\gamma$ onto a circle centered at $z$, translated to 0). Then $\tilde{\gamma}$ is a loop in $T$ as defined in (), and $w(\gamma, z)$ is precisely the index of this loop.

If $\gamma_0$ and $\gamma_1$ are closed contours not passing through $z$, and $\gamma_0$ and $\gamma_1$ are homotopic in $\mathbb{C} \setminus \{z\}$ (not necessarily with endpoint fixed), then the corresponding loops $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ are homotopic as loops in $T$. Thus their indices are the same (). So we obtain:

**Proposition.** If $\gamma_0$ and $\gamma_1$ are closed contours not passing through $z$, and $\gamma_0$ and $\gamma_1$ are homotopic in $\mathbb{C} \setminus \{z\}$ (not necessarily with endpoint fixed), then

$$w(\gamma_0, z) = w(\gamma_1, z).$$

We may restate the Cauchy Integral Formula:

**Theorem.** Let $f$ be holomorphic on a convex open set $\Omega$, $\gamma$ a closed piecewise-smooth contour in $\Omega$, and $z \in \Omega$ not on $\gamma$. Then

$$f(z)w(\gamma, z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta.$$

**Homology of Contours**

There is another useful relation on closed contours called *homology*, which is closely related to homotopy (and in fact slightly more general). Although we will primarily use this notion for closed contours, it is convenient to define it more generally for cycles:

**Definition.** Let $\Omega$ be an open set in $\mathbb{C}$. Two cycles $\gamma_0$ and $\gamma_1$ in $\Omega$ are *homologous in $\Omega$* if $w(\gamma_0, z) = w(\gamma_1, z)$ for all $z \in \mathbb{C} \setminus \Omega$. A cycle $\gamma$ is *homologous to zero in $\Omega$* if $w(\gamma, z) = 0$ for all $z \in \mathbb{C} \setminus \Omega$.

The notion of homology of cycles depends on $\Omega$. Indeed, every cycle is homologous to zero in $\mathbb{C}$. Cycles $\gamma_0$ and $\gamma_1$ in $\Omega$ are homologous in $\Omega$ if and only if $\gamma_1 - \gamma_0$ is homologous to zero in $\Omega$. If $\gamma_0$ is homologous to zero in $\Omega$, so is $-\gamma_0$; and $\gamma_0 + \gamma_1$ is homologous to $\gamma_1$ in $\Omega$ for any cycle $\gamma_1$ in $\Omega$. If $\gamma_1, \check{\gamma}_1, \check{\gamma}_2$, and $\check{\gamma}_2$ are cycles in $\Omega$, and $\gamma_j$ and $\check{\gamma}_j$ are homologous in $\Omega$ for $j = 1, 2$, then $\gamma_1 + \gamma_2$ and $\check{\gamma}_1 + \check{\gamma}_2$ are homologous in $\Omega$.
X.3.5.3. If two closed contours $\gamma_0$ and $\gamma_1$ in $\Omega$ are homotopic in $\Omega$, then it follows immediately from X.3.4.10. that $\gamma_0$ and $\gamma_1$ are homologous in $\Omega$. In particular, if $\gamma$ is contractible in $\Omega$, then $\gamma$ is homologous to zero in $\Omega$. The following example shows that the converse is false in general:

X.3.5.4. Example. Let $\Omega = \mathbb{C} \setminus \{-1, 1\}$. Let $C_1$ be the circle of radius 1 centered at 1, and $C_{-1}$ the circle of radius 1 centered at $-1$. Let $\gamma_1$ be the circle $C_1$ parametrized once around counterclockwise starting at 0, and $\gamma_2$ the circle $C_{-1}$ parametrized once around counterclockwise starting at 0. Set

$$\gamma = \gamma_1 + \gamma_2 - \gamma_1 - \gamma_2 .$$

Then $\gamma$ is a piecewise-smooth closed contour in $\Omega$, and it is easily checked that $\gamma$ is homologous to zero in $\Omega$. However, $\gamma$ is not contractible in $\Omega$; this fact, while fairly evident intuitively, is not so easy to prove rigorously (see [?, p. 356-358] for a nice proof).

This example shows that addition of closed contours is not commutative up to homotopy. In fact, the closed contours $\gamma_1 + \gamma_2$ and $\gamma_2 + \gamma_1$ are not homotopic in $\Omega$.

The underlying reason for this example is that the fundamental group $\pi_1(\Omega)$ is not abelian. In fact, it is a free group generated by the classes $[\gamma_1]$ and $[\gamma_2]$.

X.3.6. Cauchy’s Theorem, General Versions

The first general version of Cauchy’s Theorem is an immediate corollary of X.3.4.4:

X.3.6.1. Corollary. [Cauchy’s Theorem, Version 1] Let $f$ be holomorphic on an open set $\Omega$, and $\gamma$ a closed contour in $\Omega$ which is contractible in $\Omega$. Then

$$\int_{\gamma} f = 0 .$$

Using the Jordan-Schönflies Theorem, it can be shown that if $\gamma$ is a simple closed contour in $\Omega$ whose interior is also in $\Omega$, then $\gamma$ is contractible in $\Omega$. Thus we obtain another standard version of Cauchy’s Theorem:

X.3.6.2. Corollary. [Cauchy’s Theorem, Version 2] Let $f$ be holomorphic on an open set $\Omega$, and $\gamma$ a simple closed contour in $\Omega$ whose interior is also in $\Omega$. Then

$$\int_{\gamma} f = 0 .$$

X.3.6.3. Definition. An open set $\Omega$ in $\mathbb{C}$ is simply connected if it is connected and every closed contour in $\Omega$ is contractible in $\Omega$.

The same definition works (path-connectedness is required in general, not just connectedness), and is important, in a general topological space.
Informally, a connected open set is simply connected if it has no “holes.” In particular, any convex or star-shaped open set is simply connected, as is $\mathbb{C}$ with a branch cut removed (this set is actually star-shaped). But a plane or open disk with a point removed, or an annulus, is not simply connected. Any set homeomorphic to a simply connected open set is simply connected (and open).

For open sets in the plane, there are many equivalent characterizations of simple connectedness, such as the following (there are many more):

**X.3.6.5. THEOREM.** Let $\Omega$ be a connected open set in the plane. The following are equivalent:

(i) $\Omega$ is simply connected.

(ii) Whenever $\gamma$ is a simple closed contour in $\Omega$, the interior of $\gamma$ is also contained in $\Omega$.

(iii) Whenever $\gamma$ is a closed contour in $\Omega$, the winding number $\text{w}_\gamma$ of $\gamma$ around any point of $\mathbb{C} \setminus \Omega$ is 0.

If $\Omega$ is bounded, these are also equivalent to

(iv) The boundary $\partial \Omega$ is connected.

(v) The complement $\mathbb{C} \setminus \Omega$ is connected.

Condition (iii) may be rephrased:

(iii') Every cycle in $\Omega$ is homologous to zero in $\Omega$.

**X.3.6.6.** If $\Omega$ is unbounded, (iv) and (v) are still equivalent and imply (i)–(iii), but not conversely in general: consider a horizontal strip in $\mathbb{C}$. (If we work in the Riemann Sphere, the one-point compactification of $\mathbb{C}$, (i)–(v) are equivalent even if $\Omega$ is unbounded.) Also, (ii)–(v) can hold if $\Omega$ is not connected: consider a figure 8. It is not true that if $\Omega$ is simply connected, then the interior of the complement of $\Omega$ is necessarily connected, e.g. if $\Omega$ is one of the Lakes of Wada (). This example shows that simply connected open sets in $\mathbb{C}$ can be quite complicated. However, it is a (deep) theorem that every nonempty simply connected open set in $\mathbb{C}$ is homeomorphic to a disk (or, equivalently, to $\mathbb{C}$ itself). In fact, much more is true: every nonempty simply connected open proper subset of $\mathbb{C}$ is conformally equivalent to the open unit disk by the Riemann Mapping Theorem ( ).

We then get the most commonly stated version of Cauchy’s Theorem:

**X.3.6.7. COROLLARY. [CAUCHY’S THEOREM, SIMPLY CONNECTED VERSION]** Let $f$ be holomorphic on a simply connected open set $\Omega$. Then $f$ has an antiderivative on $\Omega$, and if $\gamma$ is any closed contour in $\Omega$, then

$$\int_\gamma f = 0.$$ 

The most general version of Cauchy’s Theorem which can be reasonably stated is:
**X.3.6.8. Theorem.** [Cauchy’s Theorem, Homology Version] Let $f$ be holomorphic on an open set $\Omega$. Then

$$\int_{\gamma} f = 0$$

for every cycle $\gamma$ in $\Omega$ homologous to zero in $\Omega$.

This version is somewhat complicated to prove rigorously. See e.g. [Ahl78] or [?].

We get as a special case a general version of **X.3.2.7.**:

**X.3.6.9. Corollary.** Let $f$ be holomorphic on an open set $\Omega$, and $\gamma_1$ and $\gamma_2$ simple closed contours in $\Omega$, parametrized once around counterclockwise, with $\gamma_2$ contained in the interior of $\gamma_1$. If the region between $\gamma_1$ and $\gamma_2$ is contained in $\Omega$, then

$$\int_{\gamma_1} f = \int_{\gamma_2} f .$$

**Proof:** It is easily checked that $\gamma_1 - \gamma_2$ is homologous to zero in $\Omega$.

Note that the integrals around $\gamma_1$ and $\gamma_2$ are not necessarily 0!

**X.3.7. The Cauchy Integral Formula, General Versions**

927
X.4. Sequences and Series of Holomorphic Functions; Analytic Functions

In this section, we show the remarkable fact that holomorphic functions can be expanded in power series around any point in the domain.

X.4.1. Sequences and Series of Holomorphic Functions

We begin by establishing some general results about sequences and series of holomorphic functions.

X.4.1.1. In complex analysis, u.c. convergence (IV.3.2.1) of sequences of holomorphic functions on an open set in \( \mathbb{C} \) is usually called normal convergence.

The first result is an immediate corollary of Cauchy’s Theorem and Morera’s Theorem:

X.4.1.2. Proposition. Let \( \Omega \) be an open set in \( \mathbb{C} \), and \((f_n)\) a sequence of functions from \( \Omega \) to \( \mathbb{C} \) converging normally on \( \Omega \) to a function \( f \). If each \( f_n \) is holomorphic on \( \Omega \), then \( f \) is also holomorphic on \( \Omega \).

Proof: It suffices to show that \( f \) is holomorphic on each open disk contained in \( \Omega \); so we may assume \( \Omega \) is a disk. If \( \gamma \) is any closed contour in \( \Omega \), then

\[
\int_{\gamma} f_n \to \int_{\gamma} f
\]

by (\( \ast \)). But \( \int_{\gamma} f_n = 0 \) for all \( n \) by Cauchy’s Theorem for Convex Sets (X.3.2.4.), so \( \int_{\gamma} f = 0 \). This is true for every \( \gamma \), so \( f \) is holomorphic by Morera’s Theorem.

X.4.1.3. This is one more surprising property of holomorphic functions. The real analog of this statement fails spectacularly: any continuous function on a closed bounded interval, even one not differentiable anywhere (\( \ast \)), is a uniform limit of polynomials.

Actually, the situation is even nicer for holomorphic functions, again in stark contrast to the real case (\( \ast \)):

X.4.1.4. Theorem. Let \( \Omega \) be an open set in \( \mathbb{C} \), and \((f_n)\) a sequence of functions from \( \Omega \) to \( \mathbb{C} \) converging normally on \( \Omega \) to a function \( f \). If each \( f_n \) is holomorphic on \( \Omega \), then \( f \) is also holomorphic on \( \Omega \), and \( f'_n \to f' \) normally on \( \Omega \).

In this result, normal convergence cannot be replaced by uniform convergence in general: even if \( f_n \to f \) uniformly on \( \Omega \), \( f'_n \) does not converge uniformly on \( \Omega \) to \( f' \) in general, only u.c. (Exercise X.8.2.2.).

Proof: This is a fairly easy, although slightly delicate, consequence of the Cauchy Integral Formula for derivatives (X.3.3.6.); the argument also gives another proof of X.4.1.2. By IV.3.2.5., it suffices to show that for any \( z_0 \in \Omega \), there is an open disk \( D \) centered at \( z_0 \) and contained in \( \Omega \) on which \( f \) is holomorphic and such that \( f'_n \to f' \) uniformly on \( D \). In fact, let \( D \) be any open disk centered at \( z_0 \) whose closure \( \overline{D} \) is contained in \( \Omega \) (there is always such a disk for any \( z_0 \in \Omega \)). Let \( R \) be the radius of \( D \). For sufficiently small
δ > 0, the closed disk of radius R + δ around z₀ is contained in Ω. Fix such a δ, and let γ be the circle centered at z₀ of radius R + δ, parametrized once around counterclockwise. For \( z \in \mathbb{C} \setminus \gamma \), set

\[
g(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta
\]

and note that g is holomorphic on \( C \setminus \gamma \) by X.3.3.5. By the Cauchy Integral Formula, if z is inside γ,

\[
f_n(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f_n(\zeta)}{\zeta - z} \, d\zeta
\]

for each n. We have that \( f_n \to f \) uniformly on γ. As \( n \to \infty \), the sequence

\[
\frac{f_n(\zeta)}{\zeta - z}
\]

converges uniformly on γ (for fixed \( z \) inside γ) to

\[
\frac{f(\zeta)}{\zeta - z}
\]

since the numerators converge uniformly and the absolute value of the denominator is always \( \geq \delta \). Thus we have

\[
f_n(z) = \frac{1}{2\pi i} \int_{\gamma} f_n(\zeta) \, d\zeta \to \frac{1}{2\pi i} \int_{\gamma} f(\zeta) \, d\zeta = g(z)
\]

and since \( f_n \to f \) pointwise on Ω, we conclude that \( f(z) = g(z) \) for all z inside γ, and \( f \) is holomorphic on \( D \). Since \( z_0 \) and \( D \) are arbitrary, \( f \) is holomorphic on Ω.

By a similar but slightly more delicate argument, if \( z_0, D, \) and \( \gamma \) are as above, we have that

\[
\frac{f_n(\zeta)}{\zeta - z} \to \frac{f(\zeta)}{\zeta - z} \quad \text{uniformly for } \zeta \in \gamma \text{ and } z \in \bar{D},
\]

since the numerators converge uniformly and the absolute value of the denominator is always \( \geq \delta^2 \), i.e. for any \( \epsilon > 0 \) there is an \( N \) such that

\[
\left| \frac{f_n(\zeta)}{(\zeta - z)^2} - \frac{f(\zeta)}{(\zeta - z)^2} \right| < \epsilon
\]

for all \( n \geq N \) and all \( \zeta \in \gamma \) and \( z \in \bar{D} \). Thus, for \( z \in \bar{D} \),

\[
f_n(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f_n(\zeta)}{(\zeta - z)^2} \, d\zeta \to \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta - z)^2} \, d\zeta = f'(z)
\]

uniformly on \( \bar{D} \) by the Cauchy Integral Formula for derivatives.

X.4.1.5. The proof of this theorem is a bit delicate. However, the “converse,” the complex analog of V.8.5.5., is simpler:
X.4.1.6. THEOREM. Let \( \Omega \) be a connected open set in \( \mathbb{C} \), and \((f_n)\) a sequence of functions from \( \Omega \) to \( \mathbb{C} \). If

1. Each \( f_n \) is holomorphic on \( \Omega \).
2. The sequence \((f'_n)\) converges normally on \( \Omega \) to a function \( g \).
3. There is a \( z_0 \in \Omega \) such that the sequence \((f_n(z_0))\) converges.

Then the sequence \((f_n)\) converges normally on \( \Omega \) to a function \( f \); \( f \) is holomorphic on \( \Omega \), and \( f' = g \) on \( \Omega \).

PROOF: The proof of V.8.5.5. uses the Mean Value Theorem, which is unavailable in \( \mathbb{C} \). However, we may mimic the simpler proof of the special case \((f)\). We first show that \((f_n)\) converges uniformly on a neighborhood of \( z_0 \) (in fact, on any closed disk centered at \( z_0 \) contained in \( \Omega \)) to an antiderivative for \( g \). Fix a closed disk \( D \) of radius \( r \) centered at \( z_0 \) and contained in \( \Omega \). Suppose \( f_n(z_0) \to c \). If \( z \in D \), let \( \gamma_z \) be the line segment from \( z_0 \) to \( z \). Then

\[
\int_{\gamma_z} f'_n = f_n(z) - f_n(z_0)
\]

for any \( n \) by \((f)\); and if we define

\[
f(z) = c + \int_{\gamma_z} g
\]

then \( f' = g \) on the interior of \( D \). For any \( \epsilon > 0 \) there is an \( N \) such that, for all \( n \geq N \), \(|f'_n(z) - g(z)| < \frac{\epsilon}{2r}\) for all \( z \in D \) and \(|f_n(z_0) - c| < \frac{\epsilon}{2r}\). Then, for \( n \geq N \) and \( z \in D \),

\[
|f_n(z) - f(z)| = |f_n(z_0) + \int_{\gamma_z} f'_n - c - \int_{\gamma_z} g| \leq |f_n(z_0) - c| + \left| \int_{\gamma_z} (f'_n - g) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2r}r = \epsilon.
\]

Thus \( f_n \to f \) uniformly on \( D \).

Now let \( U \) be the set of all \( z \in \Omega \) such that \((f_n)\) converges uniformly to an antiderivative for \( g \) on a neighborhood of \( z \). Then \( U \) is obviously open, and \( z_0 \in U \). If \( w \in U \), then \((f_n)\) converges uniformly on any closed disk centered at \( w \) contained in \( U \) to an antiderivative for \( g \) by the first part of the proof (with \( w \) in place of \( z_0 \)). If \( w \in \Omega \) is in the closure of \( U \), there is a sequence \((z_n)\) in \( U \) with \( z_n \to z \). There is an \( \epsilon > 0 \) such that the disk of radius \( \epsilon \) around \( w \) is contained in \( \Omega \). Fix \( n \) with \(|z_n - w| < \frac{\epsilon}{2}\); then the closed disk around \( z_n \) of radius \( \frac{\epsilon}{2} \) is contained in \( \Omega \), and is a neighborhood of \( w \) on which \((f_n)\) converges uniformly to an antiderivative for \( g \). Thus \( w \in U \). So \( U \) is relatively closed in \( \Omega \), and since \( \Omega \) is connected, \( U = \Omega \). Thus \((f_n)\) converges uniformly to an antiderivative for \( g \) on any closed disk contained in \( \Omega \); so \((f_n)\) converges pointwise on \( \Omega \) to a function \( f \) which is an antiderivative for \( g \). Since any compact subset of \( \Omega \) is contained in a finite union of closed disks in \( \Omega \), \( f_n \to f \) u.c. on \( \Omega \).

As usual, the sequence results have series versions, proved by applying the sequence results to the partial sums:
X.4.1.7. PROPOSITION. Let \( \Omega \) be an open set in \( \mathbb{C} \), and

\[ \sum_{k=1}^{\infty} f_k \]

an infinite series of functions from \( \Omega \) to \( \mathbb{C} \) converging normally on \( \Omega \) to a function \( f \). If each \( f_n \) is holomorphic on \( \Omega \), then \( f \) is also holomorphic on \( \Omega \).

X.4.1.8. THEOREM. Let \( \Omega \) be an open set in \( \mathbb{C} \), and

\[ \sum_{k=1}^{\infty} f_k \]

an infinite series of functions from \( \Omega \) to \( \mathbb{C} \) converging normally on \( \Omega \) to a function \( f \). If each \( f_n \) is holomorphic on \( \Omega \), then \( f \) is also holomorphic on \( \Omega \), and

\[ \sum_{k=1}^{\infty} f_k' \]

converges normally to \( f' \) on \( \Omega \).

X.4.1.9. THEOREM. Let \( \Omega \) be a connected open set in \( \mathbb{C} \), and

\[ \sum_{k=1}^{\infty} f_k \]

an infinite series of functions from \( \Omega \) to \( \mathbb{C} \). If

1. Each \( f_k \) is holomorphic on \( \Omega \).
2. The infinite series

\[ \sum_{k=1}^{\infty} f_k' \]

converges normally on \( \Omega \) to a function \( g \).
3. There is a \( z_0 \in \Omega \) such that the series

\[ \sum_{k=1}^{\infty} f_k(z_0) \]

converges.

Then the infinite series

\[ \sum_{k=1}^{\infty} f_k \]

converges normally on \( \Omega \) to a function \( f \); \( f \) is holomorphic on \( \Omega \), and \( f' = g \) on \( \Omega \).
X.4.2. Power Series

We recall some of the important general facts about power series from V.15.3. in the complex case.

X.4.2.1. Definition. If \( z \) is a complex number, a (complex) power series centered at \( z_0 \) is an expression of the form

\[
\sum_{k=0}^{\infty} a_k (z - z_0)^k
\]

where the \( a_k \) are complex numbers.

X.4.2.2. A power series is regarded as an infinite series of functions. It is conventional when working with power series to regard \((z - z_0)^0 = 1\) no matter how \( z - z_0 \) is interpreted; thus the \( k = 0 \) term of the power series is just the constant function \( a_0 \).

X.4.2.3. A power series is, at first, only a formal expression, but it can be regarded as defining a function \( f \) of \( z \), whose domain is the set of all \( z \) for which the series converges. Because of our convention, \( z_0 \) is always in the domain of the defined function \( f \), with \( f(z_0) = a_0 \). It can happen that the domain of \( f \) consists only of \( \{z_0\} \) (such power series are not very interesting as functions).

The most important basic property of power series is X.4.2.5., which says that the domain of the defined function is a disk, or degenerate disk, centered at \( z_0 \).

X.4.2.4. Definition. Let \( \sum_{k=0}^{\infty} a_k (z - z_0)^k \) be a power series, and define

\[
L = \limsup_{k \to \infty} |a_k|^{1/k} = \limsup_{k \to \infty} \sqrt[k]{|a_k|}
\]

and let \( R = \frac{1}{L} \) if \( 0 < L < +\infty \), \( R = 0 \) if \( L = +\infty \), and \( R = +\infty \) if \( L = 0 \). \( R \) is called the radius of convergence of the power series.

Note that the radius of convergence of a power series depends only on the coefficients \( a_k \), not on the center \( z_0 \), of the power series. The slower the coefficients grow in size, or the faster they approach zero, the larger the radius of convergence.

The term “radius of convergence” is justified by the next result:

X.4.2.5. Theorem. Let \( \sum_{k=0}^{\infty} a_k (z - z_0)^k \) be a power series, with radius of convergence \( R \). Then

(i) If \( z \in \mathbb{C} \) and \( |z - z_0| < R \), then the series \( \sum_{k=0}^{\infty} a_k (z - z_0)^k \) converges absolutely.

(ii) If \( z \in \mathbb{C} \) and \( |z - z_0| > R \), then the series \( \sum_{k=0}^{\infty} a_k (z - z_0)^k \) diverges.
(iii) If \( r \in \mathbb{R} \) and \( 0 < r < R \), then the series \( \sum_{k=0}^{\infty} a_k(z - z_0)^k \) converges uniformly on the closed disk \( \{z : |z - z_0| \leq r\} \).
(Of course, (i) and (iii) are vacuous if \( R = 0 \) and (ii) is vacuous if \( R = +\infty \).)

**Proof:** In cases (i) and (ii), apply the Root Test (\( \star \)) to the series \( \sum_{k=0}^{\infty} |a_k(z - z_0)^k| \). We have

\[
\limsup |a_k(z - z_0)^k|^{1/k} = |z - z_0| \limsup |a_k|^{1/k} = \frac{|z - z_0|}{R}
\]

(it is \( +\infty \) if \( z \neq z_0 \) and \( R = 0 \), and 0 if \( R = +\infty \) or \( z = z_0 \)). Thus the series converges absolutely if \( |z - z_0| < R \) or \( z = z_0 \) and diverges since the terms do not go to zero if \( |z - z_0| > R \).

For (iii), use the Weierstrass M-test (\( \star \)). Set \( M_k = |a_k|r^k \). Then \( |a_k(z - z_0)^k| \leq M_k \) for all \( k \) and for all \( x \), \( |z - z_0| \leq r \). We have

\[
\limsup M_k^{1/k} = r \limsup |a_k|^{1/k} = \frac{r}{R} < 1
\]

if \( 0 < R < \infty \), and \( \limsup M_k^{1/k} = 0 \) if \( R = +\infty \). Thus \( \sum_{k=0}^{\infty} M_k \) converges by the Root Test. \( \star \)

**X.4.2.6.** Thus the set on which a power series centered at \( z_0 \) converges is a disk centered at \( z_0 \), possibly of radius zero, i.e. the degenerate disk \( \{z_0\} \), or of infinite radius, i.e. all of \( \mathbb{C} \), possibly including some or all of the points on the boundary, called the **disk of convergence** of the power series.

**X.4.2.7.** Convergence behavior of a complex series on the boundary circle of the disk of convergence is quite subtle. The series can diverge everywhere on this circle (e.g. \( \sum_{k=0}^{\infty} (z - z_0)^k \)) or it can converge absolutely at all points of the circle (e.g. \( \sum_{k=1}^{\infty} \frac{1}{k^2} (z - z_0)^k \)), or it can converge conditionally on some subset \( S \) of the boundary circle. But \( S \) cannot be an arbitrary subset of the circle (it must be an \( F_{\sigma\delta} \) (\( \star \)), and not even all of these occur; see [Kör83]). The study of boundary behavior of analytic functions is a big topic in complex analysis, with important applications.

For the purposes of basic complex analysis, however, we want to restrict attention to open sets, so we generally consider only the **open disk of convergence** \( \{z : |z - z_0| < R\} \) (provided \( R > 0 \)).

Combining X.4.2.5. with X.4.1.7., we obtain:

**X.4.2.8.** **Corollary.** Let \( \sum_{k=0}^{\infty} a_k(z - z_0)^k \) be a power series, with radius of convergence \( R > 0 \). Then the series converges normally on the open disk of convergence, and the sum is holomorphic on this disk.

Using X.4.1.8., we can extend this:
**X.4.2.9. Corollary.** Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series, with radius of convergence $R > 0$. Then the series converges normally on the open disk of convergence $\{ z : |z - z_0| < R \}$, and the sum $f$ is holomorphic on this disk. The power series

$$\sum_{k=1}^{\infty} k a_k (z - z_0)^{k-1}$$

of term-by-term derivatives converges normally on $\{ z : |z - z_0| < R \}$ to $f'$.

This can be also be proved by noting that the derived series has the same radius of convergence as the original series (V.15.3.21.) and applying X.4.1.9.

By iterating, we obtain, for each $n$, that the $n$’th derived power series

$$\sum_{k=n}^{\infty} k(k-1) \cdots (k-n+1) a_k (z - z_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} a_k (z - z_0)^{k-n}$$

converges u.c. to $f^{(n)}$ on $\{ z : |z - z_0| < R \}$.

The coefficients $a_k$ are thus uniquely determined by the function $f$ defined by the power series:

**X.4.2.10. Corollary.** Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series, with radius of convergence $R > 0$, and let $f$ be the holomorphic function defined by the series on the open disk of convergence. Then, for any $n \geq 0$,

$$a_n = \frac{f^{(n)}(z_0)}{n!}.$$ 

The coefficients $a_k$ can also be computed by contour integration:

**X.4.2.11. Proposition.** Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series with radius of convergence $R > 0$, and $f$ the function defined by the series on $\{ z : |z - z_0| < R \}$. Then, for any $k \geq 0$,

$$a_k = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta$$

where $\gamma$ is any circle centered at $z_0$ of radius less than $R$.

**Proof:** By (), we have

$$\int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta = \int_{\gamma} (\zeta - z_0)^{-k-1} \left[ \sum_{j=0}^{\infty} a_j (\zeta - z_0)^j \right] d\zeta$$

934
All the terms in this sum are 0 except the term with $j = k$, which is $2\pi i a_k$.

This result can also be concluded from X.4.2.10. and the Cauchy Integral Formula for Derivatives (X.3.3.6.).

**X.4.2.12.** The radius of convergence of a power series $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ can always be computed as

$$R = \frac{1}{\limsup |a_k|^{1/k}}$$

(with the usual conventions if the lim sup is 0 or $+\infty$), but it is not always easy to evaluate this expression. Sometimes the radius of convergence can be computed more simply by the following:

**X.4.2.13.** Proposition. Let $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ be a power series, with all $a_k$ nonzero. If

$$\lim_{k \to \infty} \left| \frac{a_k}{a_{k+1}} \right|$$

exists (in the usual or extended sense), it equals the radius of convergence of the power series.

**Proof:** Let $S$ be this limit. Apply the Ratio Test () to show that $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ converges absolutely if $|z - z_0| < S$ and diverges if $|z - z_0| > S$. Thus $S$ must be exactly the radius of convergence.

Note that in the limit of X.4.2.13., the $a_k$ and $a_{k+1}$ are reversed from the way they appear in the Ratio Test ().

Another convenient characterization of the radius of convergence is the following:

**X.4.2.14.** Proposition. Let $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ be a power series. The power series has radius of convergence $R > 0$ if and only if, for all $r$, $0 < r < R$, there is a constant $M$ such that $|a_k| \leq \frac{M}{r^k}$ for all $k$, and there is no such $M$ if $r > R$.

**Proof:** Suppose the series has radius of convergence $R > 0$, and $0 < r < R$. Then

$$\limsup |a_k|^{1/k} < \frac{1}{r}$$
so there is a $k_0$ such that $|a_k| < \frac{1}{r^k}$ for all $k \geq k_0$. There is an $M \geq 1$ such that $|a_k| \leq \frac{M}{r^k}$ for $0 \leq k < k_0$. This $M$ works for all $k$.

Conversely, if the condition holds for all $r$ with $0 < r < R$ and no $r > R$, we have, for each such $r$ with $0 < r < R$,

$$\limsup |a_k|^{1/k} \leq \limsup \frac{M^{1/k}}{r} = \frac{1}{r}$$

for some constant $M$, so the radius of convergence $R'$ is $\geq R$. If $R' > R$, choose $r$ so that $R < r < R'$; then the condition holds for $r$ by the first half of the proof, a contradiction. 

\section{Inverse Power Series}

It will be convenient to extend the results about power series to closely related series we call inverse power series:

\subsection{Definition}

An inverse power series centered at $z_0 \in \mathbb{C}$ is an infinite series of the form

$$\sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k}$$

where the $b_k \in \mathbb{C}$. (We could also allow a constant term $b_0$.)

\subsection{An inverse power series defines a complex-valued function whose domain is the set of $z$ for which the series converges. This set is (almost) nicely describable, as in the case of power series:}

\subsection{Theorem}

Let

$$\sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k}$$

be an inverse power series, and set

$$S = \limsup_{k \to \infty} |b_k|^{1/k}.$$ 

Then

(i) If $S < \infty$ and $|z - z_0| > S$, then the series converges absolutely.

(ii) If $S > 0$ and $|z - z_0| < S$, then the series diverges.

(iii) If $S < \infty$ and $s > S$, then the series converges uniformly on

$$\{z : |z - z_0| \geq s\}$$

and in particular the series converges normally on

$$\{z : |z - z_0| > S\}.$$ 

936
Proof: This follows almost immediately from X.4.2.5. by considering the power series
\[ \sum_{k=1}^{\infty} b_k w^k \]
which has radius of convergence \( R = \frac{1}{S} \). If \( w = \frac{1}{z-z_0} \), then \(|w| < R\) if and only if \(|z-z_0| > S\), etc. ✐

X.4.3.4. The number \( S \) is called the inverse radius of convergence and the set
\[ \{z : |z-z_0| > S\} \]
is called the open disk complement of convergence of the inverse power series. If \( S = 0 \), this set is the entire complex plane with \( z_0 \) removed. (The series never converges for \( z = z_0 \). If \( S = \infty \), the series does not converge for any \( z \in \mathbb{C} \). If \( 0 < S < \infty \), the series may converge for some or all \( z \) with \(|z-z_0| = S\), but not necessarily; cf. X.4.2.7.)

X.4.3.5. Proposition. Let
\[ \sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k} \]
be an inverse power series with finite inverse radius of convergence. Then the series defines a holomorphic function \( f \) on its open disk complement of convergence \( \Omega \), and the series
\[ \sum_{k=1}^{\infty} \frac{-kb_k}{(z-z_0)^{k+1}} \]
of term-by-term derivatives converges normally to \( f' \) on \( \Omega \).

The fact that \( f \) is holomorphic is immediate from X.4.1.7., and that term-by-term differentiation is valid follows from X.4.1.8.. Alternatively, both parts follow from X.4.1.9. and the fact that the derived series has the same inverse radius of convergence (cf. V.15.3.21.) and thus converges u.c. on \( \Omega \).

The coefficients \( b_k \) can be computed by contour integration:

X.4.3.6. Proposition. Let
\[ \sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k} \]
be an inverse power series with finite inverse radius of convergence \( S \), and \( f \) the function defined by the series on \( \Omega = \{z : |z-z_0| > S\} \). Then, for any \( k \geq 1 \),
\[ b_k = \frac{1}{2\pi i} \oint_{\gamma} f(\zeta)(\zeta-z_0)^{k-1} d\zeta \]
where \( \gamma \) is any circle centered at \( z_0 \) of radius greater than \( S \).
Proof: By (), we have

\[
\int_\gamma f(\zeta)(\zeta - z_0)^{k-1} d\zeta = \int_\gamma (\zeta - z_0)^{k-1} \left[ \sum_{j=1}^{\infty} \frac{b_j}{(\zeta - z_0)^j} \right] d\zeta \\
= \sum_{j=1}^{\infty} \left[ \int_\gamma b_j(\zeta - z_0)^{k-1-j} d\zeta \right] .
\]

All the terms in this sum are 0 except the term with \( j = k \), which is \( 2\pi ib_k () \).

X.4.4. Analytic Functions

X.4.4.1. Definition. Let \( \Omega \) be an open set in \( \mathbb{C} \), and \( f \) a complex-valued function on \( \Omega \). Then \( f \) is analytic on \( \Omega \) if for every \( z_0 \in \Omega \) there is a power series centered at \( z_0 \) which converges to \( f \) on some open disk centered at \( z_0 \).

Note that the power series representation of \( f \) around \( z_0 \) need not converge to \( f \) on all of \( \Omega \), only on some open disk centered at \( z_0 \). Thus analyticity is a local property: if for each \( z_0 \in \Omega \) there is some neighborhood of \( z_0 \) in \( \Omega \) on which \( f \) is analytic, then \( f \) is analytic on \( \Omega \).

The following facts are immediate from X.4.2.9.:

X.4.4.2. Corollary. Let \( \Omega \) be an open set in \( \mathbb{C} \), and \( f \) a complex-valued function on \( \Omega \). If \( f \) is analytic on \( \Omega \), then \( f \) is holomorphic on \( \Omega \) and \( f' \) is also analytic on \( \Omega \).

X.4.4.3. Conversely, it follows easily from X.4.1.9. that if \( f \) is holomorphic on \( \Omega \) and \( f' \) is analytic on \( \Omega \), then \( f \) is analytic on \( \Omega \). We omit details since we will presently prove a much stronger result (X.4.4.7.).

We also obtain as a consequence of X.4.2.10.:

X.4.4.4. Corollary. Let \( \Omega \) be an open set in \( \mathbb{C} \), and \( f \) an analytic function on \( \Omega \). For any \( z_0 \in \Omega \), the power series centered at \( z_0 \) representing \( f \) in an open disk around \( z_0 \) is unique, and is

\[
\sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k
\]

which is called the Taylor series for \( f \) around \( z_0 \).
**X.4.4.5.** For real functions on an interval $I$, the following classes of functions are successively considerably larger:

(i) The class of analytic functions on $I$.

(ii) The class of $C^\infty$ functions on $I$.

(iii) The class of $C^1$ functions on $I$.

(iv) The class of differentiable functions on $I$.

The following spectacular theorem, which is the main result of this section, implies that for complex functions on an open set $\Omega$, the above four classes of functions coincide.

**X.4.4.6.** Theorem. Let $\Omega$ be an open set in $\mathbb{C}$ and $f$ a holomorphic function on $\Omega$. If $z_0 \in \Omega$ and $D$ is an open disk centered at $z_0$ contained in $\Omega$, then the Taylor series for $f$ around $z_0$ converges to $f(z)$ for every $z \in D$.

Combining this with X.4.4.2., we obtain:

**X.4.4.7.** Corollary. Let $\Omega$ be an open set in $\mathbb{C}$ and $f$ a complex-valued function on $\Omega$. Then $f$ is analytic on $\Omega$ if and only if it is holomorphic on $\Omega$.

**Proof:** Suppose $f$ is holomorphic on $\Omega$. Fix $z_0 \in \Omega$ and an open disk $D$ of radius $R$ centered at $z_0$ contained in $\Omega$. Fix $z \in D$, and let $\gamma$ be a circle centered at $z_0$ with radius $r$ for some $r$, $|z-z_0| < r < R$, parametrized once around counterclockwise. Thus

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z} d\zeta$$

by the Cauchy Integral Formula.

Since $\left| \frac{z-z_0}{\zeta - z_0} \right| = \frac{|z-z_0|}{r} < 1$ for all $\zeta \in \gamma$, the infinite series

$$\sum_{k=0}^{\infty} \left( \frac{z-z_0}{\zeta - z_0} \right)^k$$

converges uniformly on $\gamma$ to

$$\frac{1}{1 - \frac{z-z_0}{\zeta - z_0}} = \frac{\zeta - z_0}{\zeta - z}$$

i.e. for $\zeta \in \gamma$ we have

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} \sum_{k=0}^{\infty} \left( \frac{z-z_0}{\zeta - z_0} \right)^k = \sum_{k=0}^{\infty} \left( \frac{z-z_0}{\zeta - z_0} \right)^k$$

where the series converges uniformly on $\gamma$. Thus the series

$$\sum_{k=0}^{\infty} \left( \frac{z-z_0}{\zeta - z_0} \right)^k \int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta = \int_\gamma f(\zeta) \left[ \sum_{k=0}^{\infty} \left( \frac{z-z_0}{\zeta - z_0} \right)^k \right] d\zeta$$

939
converges to

\[ \int f(\zeta) \frac{d\zeta}{\zeta - z} = 2\pi i f(z). \]

In other words, the power series

\[ \sum_{k=0}^{\infty} a_k(z - z_0)^k \]

converges to \( f(z) \), where

\[ a_k = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta \]

for each \( k \).

**X.4.4.8.** The proof of the theorem as stated is then completed by either identifying \( a_k \) with \( \frac{f^{(k)}(z_0)}{k!} \) by the Cauchy Integral Formula for Derivatives (X.3.3.6.) or by using X.4.2.10.. In fact, by combining the above proof with X.4.2.10., we can obtain an alternate proof of the Cauchy Integral Formula for Derivatives. Note that the proof of X.4.4.6. makes crucial use of the ordinary Cauchy Integral Formula.

**X.4.4.9.** Note that X.4.4.6. says more than just that \( f \) is analytic on \( \Omega \): if \( z_0 \in \Omega \), then the Taylor series for \( f \) around \( z_0 \) does not merely converge to \( f \) on some open disk centered at \( z_0 \), but actually converges to \( f \) on the largest open disk centered at \( z_0 \) contained in \( \Omega \) (if \( \Omega \neq \mathbb{C} \)). If \( \Omega = \mathbb{C} \), i.e. if \( f \) is an entire function, this Taylor series converges to \( f \) on all of \( \mathbb{C} \).

This result is again in stark contrast to the real case (cf. (i)).

**Holomorphic Functions on an Annulus**

We can also obtain a similar series representation for a holomorphic function on an annulus, involving both a power series and an inverse power series in general.

**X.4.4.10.** Definition. Let \( z_0 \in \mathbb{C} \), and fix \( r \) and \( R \) with \( 0 \leq r < R \leq +\infty \). The (open) annulus centered at \( z_0 \) with inner radius \( r \) and outer radius \( R \) is the set

\[ \{ z \in \mathbb{C} : r < |z - z_0| < R \}. \]

If \( 0 < r < R < +\infty \), this annulus is the open region between the circles of radius \( r \) and \( R \) centered at \( z_0 \). If \( r = 0 \) and \( R < +\infty \), the annulus is the punctured open disk of radius \( R \) centered at \( z_0 \) with \( z_0 \) removed. If \( r = 0 \) and \( R = +\infty \), it is the punctured plane with \( z_0 \) removed; if \( r > 0 \) and \( R = +\infty \), it is the plane with the closed disk of radius \( r \) centered at \( z_0 \) removed. (It is convenient to include these degenerate cases.)
**X.4.4.11. Theorem.** Let $\Omega$ be an open annulus centered at $z_0$ with inner radius $r$ and outer radius $R$, and $f$ a holomorphic function on $\Omega$. Then there are a unique power series

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

and a unique inverse power series

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

which both converge normally on $\Omega$, and for which

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

for all $z \in \Omega$. For each $k \geq 0$ we have

$$a_k = \frac{1}{2\pi i} \int_{\gamma} f(\zeta) (\zeta - z_0)^{k+1} d\zeta$$

and for each $k \geq 1$ we have

$$b_k = \frac{1}{2\pi i} \int_{\gamma} f(\zeta)(\zeta - z_0)^{k-1} d\zeta$$

where $\gamma$ is any circle centered at $z_0$ whose radius $\rho$ satisfies $r < \rho < R$.

**Proof:** The argument is similar to the previous ones. Fix $z \in \Omega$, and let $\gamma_1$ and $\gamma_0$ be circles centered at $z_0$ of radius $S$ and $s$ respectively, where $0 < s < |z - z_0| < S < R$, parametrized once around counterclockwise. Then

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_1} f(\zeta) d\zeta = \frac{1}{2\pi i} \int_{\gamma_1} \frac{f(\zeta)}{\zeta - z} d\zeta - \frac{1}{2\pi i} \int_{\gamma_0} \frac{f(\zeta)}{\zeta - z} d\zeta$$

by (). We analyze the two integrals on the right separately.

Since $|\frac{z - z_0}{\zeta - z_0}| = \frac{|z - z_0|}{S} < 1$ for all $\zeta \in \gamma_1$, the infinite series

$$\sum_{k=0}^{\infty} \left( \frac{z - z_0}{\zeta - z_0} \right)^k$$

converges uniformly on $\gamma_1$ to

$$\frac{1}{1 - \frac{z - z_0}{\zeta - z_0}} = \frac{\zeta - z_0}{\zeta - z}$$

i.e. for $\zeta \in \gamma_1$ we have

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} \sum_{k=0}^{\infty} \left( \frac{z - z_0}{\zeta - z_0} \right)^k = \sum_{k=0}^{\infty} \frac{(z - z_0)^k}{(\zeta - z_0)^{k+1}}$$

941
where the series converges uniformly on $\gamma_1$. Thus the series

$$\sum_{k=0}^{\infty} \left[ (z - z_0)^k \int_{\gamma_1} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta \right] = \int_{\gamma_1} f(\zeta) \left[ \sum_{k=0}^{\infty} \frac{(z - z_0)^k}{(\zeta - z_0)^{k+1}} \right] d\zeta$$

collapses to

$$\int_{\gamma_1} \frac{f(\zeta)}{\zeta - z} d\zeta.$$

Similarly, since $\left| \frac{\zeta - z_0}{z - z_0} \right| = \frac{s}{|z - z_0|} < 1$ for all $\zeta \in \gamma_0$, the infinite series

$$\sum_{j=0}^{\infty} \left( \frac{\zeta - z_0}{z - z_0} \right)^j$$

collapses uniformly on $\gamma_0$ to

$$\frac{1}{1 - \frac{\zeta - z_0}{z - z_0}} = \frac{z - z_0}{\zeta - z},$$

i.e. for $\zeta \in \gamma_0$ we have

$$\frac{1}{\zeta - z} = -\frac{1}{z - z_0} \sum_{j=0}^{\infty} \left( \frac{\zeta - z_0}{z - z_0} \right)^j = -\sum_{j=0}^{\infty} \frac{(\zeta - z_0)^j}{(z - z_0)^{j+1}}$$

where the series converges uniformly on $\gamma_0$. Thus the series

$$\sum_{k=1}^{\infty} \left[ \frac{1}{(z - z_0)^k} \int_{\gamma_0} f(\zeta)(\zeta - z_0)^{k-1} d\zeta \right]$$

$$= \sum_{j=0}^{\infty} \left[ \frac{1}{(z - z_0)^{j+1}} \int_{\gamma_0} f(\zeta)(\zeta - z_0)^j d\zeta \right] = \int_{\gamma_0} f(\zeta) \left[ \sum_{j=0}^{\infty} \frac{(\zeta - z_0)^j}{(z - z_0)^{j+1}} \right] d\zeta$$

collapses to

$$-\int_{\gamma_0} \frac{f(\zeta)}{\zeta - z} d\zeta.$$

So we have that both series converge and that

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

where

$$a_k = \frac{1}{2\pi i} \int_{\gamma_1} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta$$

$$b_k = \frac{1}{2\pi i} \int_{\gamma_0} f(\zeta)(\zeta - z_0)^{k-1} d\zeta.$$
The circles $\gamma_1$ and $\gamma_0$ depend on $z$. But note that if $\gamma$ is any circle centered at $z_0$ contained in $\Omega$, then, for any $k$,

$$\int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta = \int_{\gamma_1} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta$$

$$\int_{\gamma} f(\zeta) (\zeta - z_0)^{-k-1} d\zeta = \int_{\gamma_0} f(\zeta) (\zeta - z_0)^{-k-1} d\zeta$$

so the $a_k$ and $b_k$ are independent of the choice of $\gamma_1$ and $\gamma_0$, i.e. the same series are obtained for all $z \in \Omega$, and these series converge pointwise on $\Omega$.

Since the power series converges pointwise on $\Omega$, its radius of convergence is at least $R$, hence it converges u.c. on $\Omega$. Similarly, the inverse radius of convergence of the inverse power series is at most $r$, so it converges u.c. on $\Omega$.

It remains to show uniqueness. If there are series with coefficients $c_k$ and $d_k$, converging u.c. on $\Omega$ with

$$f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k + \sum_{k=1}^{\infty} \frac{d_k}{(z - z_0)^k}$$

then for any $k \geq 0$ and any circle $\gamma$ in $\Omega$ centered at $z_0$ we have

$$2\pi i a_k = \int_{\gamma} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} d\zeta = \sum_{j=0}^{\infty} \left[ \int_{\gamma} c_j (\zeta - z_0)^{j-k-1} d\zeta \right] + \sum_{j=1}^{\infty} \left[ \int_{\gamma} \frac{d_k}{(\zeta - z_0)^{j+k+1}} d\zeta \right]$$

where term-by-term integration is justified by u.c. convergence. But all terms in the second sum are zero, and all terms in the first are also zero except the term with $j = k$, which is $2\pi i c_k$. Thus $c_k = a_k$. Similarly, if $k \geq 1$ we have

$$2\pi i b_k = \int_{\gamma} f(\zeta) (\zeta - z_0)^{-k-1} d\zeta = \sum_{j=0}^{\infty} \left[ \int_{\gamma} c_j (\zeta - z_0)^{j+k-1} d\zeta \right] + \sum_{j=1}^{\infty} \left[ \int_{\gamma} \frac{d_k}{(\zeta - z_0)^{j-k+1}} d\zeta \right]$$

and the term in the second sum with $j = k$ is $2\pi i d_k$ and all other terms are zero.

The uniqueness statement in this theorem can be extended:

**X.4.4.12. Corollary.** Let $\Omega$ be an open annulus centered at $z_0$ with inner radius $r$ and outer radius $R$, and $f$ a holomorphic function on $\Omega$, with expansion

$$f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

as in X.4.4.11.. Suppose there are a power series

$$\sum_{k=0}^{\infty} c_k(z - z_0)^k$$

943
and an inverse power series
\[ \sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k} \]
converging pointwise on \( \Omega \) with
\[ f(z) = \sum_{k=0}^{\infty} c_k(z-z_0)^k + \sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k} \]
for all \( z \in \Omega \). Then \( c_k = a_k \) and \( d_k = b_k \) for all \( k \).

**Proof:** Since
\[ \sum_{k=0}^{\infty} c_k(z-z_0)^k \]
converges pointwise on \( \Omega \), its radius of convergence is at least \( R \), so the series converges u.c. on \( \Omega \). Similarly, the inverse power series
\[ \sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k} \]
converges pointwise on \( \Omega \), so its inverse radius of convergence is at most \( r \), so this series converges u.c. on \( \Omega \). The result now follows from X.4.4.11.

**X.4.4.13.** If we set \( a_{-k} = b_k \) for \( k \geq 1 \), we can more compactly write the expansion in X.4.4.11. as
\[ f(z) = \sum_{k=0}^{\infty} a_k(z-z_0)^k \]
which is called the *Laurent expansion* of \( f \) around \( z_0 \) (it is technically only a Laurent series (III.5.3.3.) if only finitely many \( b_k \) are nonzero, which turns out to often be the case if \( r = 0 \)). But the proper phrasing of convergence of such a series is somewhat tricky, so it is better to consider the positive and negative powers separately in general.

**X.4.4.14.** One important difference between X.4.4.11. and X.4.4.6. is that there is no obvious interpretation of the coefficients \( a_k \) (or \( b_k \)) as values of derivatives of \( f \) at \( z_0 \), which of course do not make sense if \( f \) is defined only on an annulus around \( z_0 \). However, in at least some cases (e.g. if only finitely many \( b_k \) are nonzero) there is an analogous interpretation ().
X.4.5. Normal Families and Montel’s Theorem

There is an important compactness-like property for families of holomorphic functions which we discuss in this section.

Normal Families

X.4.5.1. Definition. Let $\Omega$ be an open set in $\mathbb{C}$, and $F$ a family of holomorphic functions on $\Omega$. Then $F$ is a normal family if every sequence in $F$ has a subsequence which converges normally (u.c.) on $\Omega$.

X.4.5.2. Normal families turn out to exist in abundance in Complex Analysis, and this fact is the key to many advanced results: they are used to construct holomorphic functions with an exact set of specified properties from families of functions approximately satisfying the properties.

Here is a nice technical simplification used in proving normality for families of holomorphic functions:

X.4.5.3. Proposition. Let $\Omega$ be an open set in $\mathbb{C}$, and $F$ a family of holomorphic functions on $\Omega$. Suppose there is a sequence $(D_j)$ of compact subsets of $\Omega$ such that

(i) Every compact subset of $\Omega$ is contained in the union of finitely many $D_j$.

(ii) For every $j$, every sequence in $F$ has a subsequence converging uniformly on $D_j$.

Then $F$ is a normal family.

Proof: If $(f_n)$ is a sequence in $F$, then there is a subsequence $(f_{n_k,1})$ which converges uniformly on $D_1$. This subsequence has a subsequence $(f_{n_k,2})$ which converges uniformly on $D_2$. Continue inductively to get, for each $j$, a subsequence $f_{n_k,j}$ converging uniformly on $D_j$. The diagonal subsequence $(f_{n_k,k})$ then converges uniformly on $D_j$ for each $j$, and hence on any finite union of the $D_j$; thus it converges normally on $\Omega$.

X.4.5.4. Corollary. Let $\Omega$ be an open set in $\mathbb{C}$, and $F$ a family of holomorphic functions on $\Omega$. Then $F$ is a normal family if and only if, for every closed disk $D$ in $\Omega$, any sequence in $F$ has a subsequence which converges uniformly on $D$.

The proof of this Corollary really involves no Complex Analysis, only topology. The key is the following topological lemma:

X.4.5.5. Lemma. Let $\Omega$ be an open set in $\mathbb{C}$. Then there is a sequence $(D_j)$ of closed disks in $\Omega$ such that every compact subset of $\Omega$ is contained in the union of finitely many of the $D_j$.

Proof: First note that there is a sequence $(K_n)$ of compact sets whose union is $\Omega$: for example, set

$$K_n = \left\{ z \in \Omega : d(z, \mathbb{C} \setminus \Omega) \geq \frac{1}{n}, |z| \leq n \right\}.$$
For each $z \in \Omega$, there is an open disk $B_z$ centered at $z$ whose closure is contained in $\Omega$. For each $n$, finitely many of the $B_z$ cover $K_n$; thus countably many of the $B_z$ cover $\Omega$, say \( \{ B_{z_1}, B_{z_2}, \ldots \} \). If $K$ is any compact subset of $\Omega$, finitely many of the $B_{z_j}$ cover $K$. Take $D_j = B_{z_j}$.

This proof works essentially verbatim for any open set in any locally compact, $\sigma$-compact metric space ().

**Montel’s Theorem**

**X.4.5.6. Theorem.** [Montel’s Theorem] Let $\Omega$ be an open set in $\mathbb{C}$, and $F$ a family of holomorphic functions on $\Omega$. If $F$ is uniformly bounded on every closed disk in $\Omega$, then $F$ is a normal family.

**Proof:** We will use X.4.5.3. for a carefully chosen sequence $(D_j)$ of closed disks. The result will follow immediately from the Arzela-Ascoli Theorem () if $F$ is equicontinuous on each $D_j$. We proceed similarly to the proof of X.4.5.5.: for each $z \in \Omega$ there is an $r > 0$ (depending on $z$) such that the closed disk of radius $2r$ around $z$ is contained in $\Omega$. Let $B_z$ be the open disk of radius $r$ centered at $z$. As in the proof of X.4.5.5., there is a sequence $(B_{z_j})$ such that every compact subset of $\Omega$ is covered by finitely many $B_{z_j}$. Set $D_j = B_{z_j}$.

Equicontinuity of $F$ on $D_j$ follows from the first Cauchy Inequality, as follows. Fix $j$. If $r$ is the radius of $D_j$, then the closed disk $D$ of radius $2r$ centered at $z_j$ is contained in $\Omega$, and if $z \in D_j$, then the closed disk of radius $r$ around $z$ is contained in $D$. So if $M$ is a bound for $F$ on $D$, we have $|f'(z)| \leq \frac{M}{r}$ for all $f \in F$ and $z \in D_j$. Then, since $D_j$ is convex, for any $z, w \in D_j$ we have

$$|f(z) - f(w)| \leq \frac{M}{r} |z - w|$$

for all $f \in F$ by X.2.1.9., and equicontinuity of $F$ on $D_j$ follows.

The most commonly used special case is called Montel’s Theorem in many references:

**X.4.5.7. Corollary.** Let $\Omega$ be an open set in $\mathbb{C}$, and $F$ a family of holomorphic functions on $\Omega$. If $F$ is uniformly bounded on $\Omega$, then $F$ is a normal family.

**X.4.5.8. These are not the most general theorems proved by Montel. In [Gam01], X.4.5.7. is called the “thesis grade” Montel Theorem since Montel proved it in his Ph.D. thesis. He later proved a spectacular generalization of X.4.5.7. (not quite a generalization of X.4.5.6.) related to the Picard Theorems, which is beyond the scope of this book; see e.g. [Gam01].**

946
X.5. Zeroes and Singularities

The existence of local power series representations for holomorphic functions has some surprising and pleasant rigidity consequences, and the closely related existence of Laurent expansions around isolated singularities yields more consequences which shed great light on the global properties of holomorphic functions. In fact, it was the philosophy of CAUCHY and, especially, RIEMANN that a holomorphic (analytic) function should be primarily understood through its zeroes and singularities.

X.5.1. Zeroes of Holomorphic Functions

The first result is immediate from X.4.4.6., and underlies the rest of the results of this section:

X.5.1.1. Corollary. Let \( \Omega \) be an open set in \( \mathbb{C} \), \( f \) holomorphic on \( \Omega \), and \( z_0 \in \Omega \) with \( f(z_0) = 0 \). Then either

(i) \( f \) is identically zero in a neighborhood of \( z_0 \)

or

(ii) There is a unique \( n \in \mathbb{N} \) and a unique function \( \phi \) which is holomorphic on \( \Omega \), with \( \phi(z_0) \neq 0 \), for which

\[
 f(z) = (z - z_0)^n \phi(z)
\]

for all \( z \in \Omega \); and there is a neighborhood \( U \) of \( z_0 \) containing no other zeroes of \( f \).

Proof: Let

\[
 \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z - z_0)^k
\]

be the Taylor series for \( f \) around \( z_0 \) and \( D \) an open disk centered at \( z_0 \) on which this series converges to \( f \). Either \( f^{(k)}(z_0) = 0 \) for all \( k \), in which case this series is identically zero and \( f \) is identically zero on \( D \), or else there is a unique \( n \geq 1 \) such that \( f^{(n)}(z_0) \neq 0 \) and \( f^{(k)}(z_0) = 0 \) for all \( k < n \). In this case, the power series

\[
 \sum_{k=n}^{\infty} \frac{f^{(k)}(z_0)}{k!} (z - z_0)^{k-n}
\]

has the same radius of convergence as the Taylor series, and thus converges to a holomorphic function \( \phi \) on \( D \) which satisfies \( \phi(z_0) = \frac{f^{(n)}(z_0)}{n!} \neq 0 \) and

\[
 \phi(z) = \frac{f(z)}{(z - z_0)^n}
\]

for all \( z \in D \setminus \{z_0\} \). We may thus extend \( \phi \) to a function holomorphic on \( \Omega \) with the same properties by setting \( \phi(z) = \frac{f(z)}{(z - z_0)^n} \) for \( z \in \Omega \setminus D \). Uniqueness of \( n \) and \( \phi \) is clear.

For the last statement, since \( \phi \) is continuous and \( \phi(z_0) \neq 0 \), there is a neighborhood \( U \) of \( z_0 \) on which \( \phi(z) \neq 0 \) for all \( z \in U \). Then

\[
 f(z) = (z - z_0)^n \phi(z) \neq 0
\]

for all \( z \in U \setminus \{z_0\} \) since both factors are nonzero.

A crucial property of holomorphic functions follows, that the zeroes of a nonconstant holomorphic function are isolated:
X.5.1.2. **Corollary.** Let $\Omega$ be a connected open set in $\mathbb{C}$, and $f$ a holomorphic function on $\Omega$. Suppose there is a sequence $(z_n) \in \Omega$ of distinct points such that $f(z_n) = 0$ for all $n$, and that $z_n \to z_0$ for some $z_0 \in \Omega$. Then $f$ is identically zero on $\Omega$.

**Proof:** We have that $f(z_0) = 0$ by continuity, and thus $f$ is identically zero in a neighborhood of $z_0$ by X.5.1.1.. Let $U$ be the set of all $z \in \Omega$ such that $f$ is identically zero in a neighborhood of $z$. Then $U$ is open and contains $z_0$. If $w \in \bar{U} \cap \Omega$, then there is a sequence $(w_n) \in U$ with $w_n \to w$, and thus $w \in U$ as in the first statement of the proof. So $U$ is relatively closed in $\Omega$. Since $\Omega$ is connected, $U = \Omega$.

X.5.1.3. This result says that the zeroes of a nonconstant holomorphic function on a connected $\Omega$ cannot have an accumulation point in $\Omega$. Note that they can have an accumulation point outside $\Omega$: set $\Omega = \mathbb{C} \setminus \{0\}$ and $f(z) = \sin \frac{1}{z}$. Then $f$ is holomorphic and nonconstant on $\Omega$ and $f$ has a sequence of zeroes (at $\frac{1}{\pi n}$) converging to 0.

Connectedness of $\Omega$ is essential: a holomorphic function on a nonconnected $\Omega$ could be identically zero on a component of $\Omega$ and not on another component.

The statement of X.5.1.2. becomes seemingly much more powerful by the simple trick of considering $f - g$:

X.5.1.4. **Theorem.** [Rigidity Theorem] Let $\Omega$ be a connected open set in $\mathbb{C}$, and $f$ and $g$ holomorphic functions on $\Omega$. Suppose there is a sequence $(z_n) \in \Omega$ of distinct points such that $f(z_n) = g(z_n)$ for all $n$, and that $z_n \to z_0$ for some $z_0 \in \Omega$. Then $f$ and $g$ are identically equal on $\Omega$.

X.5.1.5. This result is called the Rigidity Theorem because it means that if $f$ is holomorphic on $\Omega$, the values of $f$ on any small subregion of $\Omega$ (or even subset of $\Omega$ large enough to have an accumulation point in $\Omega$) completely determine the values of $f$ throughout all of $\Omega$ (if $\Omega$ is connected).

X.5.1.6. If a holomorphic function $f$ has an isolated zero at $z_0$, the number $n$ in X.5.1.1. is called the order of the zero at $z_0$. This language extends the usual notion of order of a zero of a polynomial. Part of the content of X.5.1.1. is that an isolated zero of a holomorphic function has finite order.

**Limits of Sequences of Holomorphic Functions**

Using Montel’s Theorem (), we can obtain a considerable extension of the Rigidity Theorem:

X.5.1.7. **Theorem.** Let $\Omega$ be a connected open set in $\mathbb{C}$, $(f_k)$ a sequence of holomorphic functions on $\Omega$, and $(z_n)$ a sequence of distinct points in $\Omega$ converging to $z_0 \in \Omega$. Suppose the $f_k$ are uniformly bounded on compact subsets of $\Omega$, and that $w_n := \lim_{k \to \infty} f_k(z_n)$ exists for all $n$. Then there is a unique holomorphic function $f$ on $\Omega$ such that $f(z_n) = w_n$ for all $n$, and $f_n \to f$ normally on $\Omega$.

**Proof:** By Montel’s Theorem, there is a subsequence of $(f_k)$ converging u.c. to a holomorphic function $f$ on $\Omega$, and we must have $f(z_n) = w_n$ for all $n$. Similarly, if $(f_{m_k})$ is any subsequence of $(f_k)$, then $(f_{m_k})$ has a subsequence converging u.c. on $\Omega$ to a holomorphic function $g$ with $g(z_n) = w_n$ for all $n$. By the Rigidity
Theorem, \( g = f \) on all of \( \Omega \). Thus, if \( K \) is a compact subset of \( \Omega \) containing all the \( z_n \), every subsequence of \( (f_n) \) has a subsubsequence converging uniformly to \( f \) on \( K \); so \( f_n \to f \) uniformly on \( K \).

This is a powerful result since it shows that to prove that a sequence of holomorphic functions converges normally on \( \Omega \), it suffices to show the sequence is uniformly bounded on compact subsets of \( \Omega \) (a necessary condition for normal convergence) and converges pointwise on a subset of \( \Omega \) with a limit point in \( \Omega \) (e.g. a small subregion of \( \Omega \)). Note that the Rigidity Theorem is the uniqueness part applied to a constant sequence. But the Rigidity Theorem cannot be considered a corollary of this result since it is used in the proof.

### X.5.2. Singularities

#### X.5.2.1. Definition. Let \( f \) be holomorphic on an open set \( \Omega \). Then \( f \) has an isolated singularity at \( z_0 \in \mathbb{C} \setminus \Omega \) if there is a deleted neighborhood of \( z_0 \) contained in \( \Omega \).

In other words, \( f \) has an isolated singularity at \( z_0 \) if it is holomorphic in a neighborhood of \( z_0 \) except at \( z_0 \).

#### X.5.2.2. Examples. (i) The function

\[
    f(z) = \frac{z - 1}{z^2 - 1}
\]

has isolated singularities at 1 and \(-1\).

(ii) The function \( f(z) = e^{1/z} \) has an isolated singularity at 0.

(iii) The function \( f(z) = \tan \frac{z}{n!} \) has isolated singularities at \( \frac{2}{n!} \) for all odd \( n \), and a “singularity” at 0 which is not an isolated singularity.

(iv) The functions \( \log z \) and \( z^c, c \) not an integer, have a “singularity” at 0. But these functions cannot be defined as holomorphic functions on any deleted neighborhood of 0, so no version of them has an isolated singularity at 0. (If \( e \) is a negative integer, \( z^c \) is a well-defined holomorphic function on \( \mathbb{C} \setminus \{0\} \) which has an isolated singularity at 0.)

#### X.5.2.3. We will not consider “singularities” such as the ones at 0 in (iii) or (iv) which are not isolated singularities: we will use the term singularity to mean “isolated singularity.”

The three singularities in (i) and (ii) are examples of the three types of singularity discussed below.

#### X.5.2.4. If \( f \) has a singularity at \( z_0 \), then by X.4.4.11. there is a unique Laurent expansion

\[
    f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}
\]

which converges u.c. in a deleted neighborhood of \( z_0 \).

### Removable Singularities

A singularity such as the one at 1 in X.5.2.2 (i) is “artificial” since the function can be defined there in such a way that the extended function is holomorphic in a whole neighborhood of 1.
**X.5.2.5. Definition.** If \( f \) has a singularity (isolated singularity) at \( z_0 \), the singularity is a **removable singularity** if there is a function \( g \), holomorphic in a neighborhood of \( z_0 \), such that \( f = g \) on a deleted neighborhood of \( z_0 \).

**X.5.2.6. Examples.**

(i) The function \( f(z) = \frac{z-1}{z^2} \) has singularities at 1 and \(-1\). The singularity at 1 is removable, since \( f(z) = g(z) = \frac{1}{z^2} \) in a deleted neighborhood of 1, and \( g \) is holomorphic in a neighborhood of 1. The singularity of \( f \) at \(-1\) is not removable.

(ii) The function
\[
f(z) = \frac{\sin z}{z}
\]
has a singularity at 0. It is not so easy to prove directly that
\[
\lim_{z \to 0} \frac{\sin z}{z}
\]
does not exist (it does, and equals 1). But on \( \mathbb{C} \setminus \{0\} \) \( f \) agrees with the entire function \( g \) defined by the power series
\[
g(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k}
\]
and thus the singularity is removable.

If \( f \) has a removable singularity at \( z_0 \), it is obvious that
\[
\lim_{z \to z_0} f(z)
\]
exists. It is remarkable that this condition is also sufficient for the singularity to be removable; in fact, apparently much weaker conditions are also sufficient:

**X.5.2.7. Theorem.** If \( f \) has a singularity at \( z_0 \), the following are equivalent:

(i) The singularity at \( z_0 \) is removable.

(ii) \( \lim_{z \to z_0} f(z) \) exists.

(iii) \( f \) is bounded in a deleted neighborhood of \( z_0 \).

(iv) \( \lim_{z \to z_0} f(z)(z-z_0) = 0 \).

(v) In the Laurent expansion of \( f \) around \( z_0 \), all \( b_k \) are zero.

**Proof:** (i) \( \Rightarrow \) (ii) \( \Rightarrow \) (iii) \( \Rightarrow \) (iv) are obvious.

(iv) \( \Rightarrow \) (i): Let \( D \) be an open disk centered at \( z_0 \) such that \( f \) is holomorphic on \( D \setminus \{z_0\} \), and let \( \gamma \) be a circle of radius \( r \) centered at \( z_0 \) contained in \( D \setminus \{z_0\} \). Then, for each \( z \) inside \( \gamma \), \( z \neq z_0 \),
\[
f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} d\zeta - \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} d\zeta
\]

950
where \( \gamma_\delta \) is any circle centered at \( z_0 \) of radius \( \delta < |z - z_0| \). Note that, for fixed \( z \), the second integral is independent of \( \delta \), as long as \( 0 < \delta < |z - z_0| \).

Fix \( z \) inside \( \gamma \), and let \( \epsilon > 0 \). If \( \delta > 0 \) is sufficiently small, then \( |f(\zeta)(\zeta - z_0)| < \epsilon \) whenever \( 0 < |\zeta - z_0| \leq \delta \).

We may assume \( \delta \leq \frac{|z - z_0|}{2} \). Then

\[
|f(\zeta)| < \frac{\epsilon}{\delta}
\]

for \( \zeta \in \gamma_\delta \), so

\[
\left| \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta \right| < \frac{2\epsilon}{\delta |z - z_0|} \cdot 2\pi \delta = \frac{4\pi \epsilon}{|z - z_0|}.
\]

Since the integral is independent of \( \delta \), and \( \epsilon > 0 \) is arbitrary, we have

\[
\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta = 0
\]

for \( \zeta \in \gamma \), so

\[
f(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta.
\]

This is true for each \( z \) inside \( \gamma \), \( z \neq z_0 \). But the function

\[
g(z) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(\zeta)}{\zeta - z} \, d\zeta
\]

is holomorphic on the entire disk inside \( \gamma \) by (i).

(v) \( \iff \) (i): Let \( D \) be an open disk centered at \( z_0 \) such that \( f \) is holomorphic on \( D \setminus \{ z_0 \} \), of radius \( r \). If all \( b_k = 0 \), the Laurent series for \( f \) around \( z_0 \) is a power series which converges to \( f \) on \( D \setminus \{ z_0 \} \); thus its radius of convergence is at least \( r \) and it defines a function \( g \) which is holomorphic on \( D \) and agrees with \( f \) on \( D \setminus \{ z_0 \} \). Conversely, if \( g \) is a holomorphic function on an open disk \( D \) centered at \( z_0 \) such that \( f = g \) on \( D \setminus \{ z_0 \} \), the Taylor series for \( g \) around \( z_0 \) converges to \( f \) on \( D \setminus \{ z_0 \} \), so must be the Laurent series for \( f \) around \( z_0 \) by uniqueness of Laurent expansions.

There is a simpler proof that (iii) \( \Rightarrow \) (i) (Exercise X.8.2.12.), and an even simpler proof that (ii) \( \Rightarrow \) (i) (Exercise X.8.2.13.).

\( X.5.2.8. \) The implication (ii) \( \Rightarrow \) (i) means that a continuous function from \( \mathbb{C} \) (or an open subset) to \( \mathbb{C} \) cannot have an isolated point where it is not differentiable (cf. Exercise X.8.2.13.). This is in contrast to the real case, e.g. for the function \( f(x) = |x| \). The theorem also rules out complex analogs of \( g(x) = \sin \frac{1}{x} \) ((iii) \( \Rightarrow \) (i)) and \( h(x) = \frac{1}{\sqrt{|x|}} \) ((iv) \( \Rightarrow \) (i)). (The function \( g(z) = \sin \frac{1}{z} \) is holomorphic with an isolated singularity at 0, but \( g \) is not bounded near 0; in fact the singularity is an essential singularity.)

Poles

\( X.5.2.9. \) Definition. If \( f \) has a singularity (isolated singularity) at \( z_0 \), the singularity is a pole of order \( n \) if \( \frac{1}{f} \) has a removable singularity at \( z_0 \) which is a zero of order \( n \). A pole of order 1 is called a simple pole.

Poles can be simply characterized:
X.5.2.10. Proposition. Suppose $f$ has a singularity at $z_0$. Then $f$ has a pole at $z_0$ if and only if
\[ \lim_{z \to z_0} |f(z)| = +\infty \]
i.e. for any $M > 0$ there is a $\delta > 0$ such that $|f(z)| > M$ whenever $0 < |z - z_0| < \delta$.

Proof: If $f$ has a pole at $z_0$ of order $n$, and $\frac{1}{f} = g$ in a deleted neighborhood of $z_0$ and $g$ has a zero of order $n$ at $z_0$, then there is a deleted neighborhood $U$ of $z_0$ such that $g$ has no zeroes in $U$, and we have
\[ \lim_{z \to z_0} g(z) = 0 \]
and it follows that
\[ \lim_{z \to z_0} |f(z)| = +\infty . \]
Conversely, if
\[ \lim_{z \to z_0} |f(z)| = +\infty \]
there is a deleted neighborhood of $z_0$ on which $f$ is nonzero, so $h = \frac{1}{f}$ is defined in a deleted neighborhood of $z_0$. We have
\[ \lim_{z \to z_0} h(z) = 0 \]
so $h$ has a removable singularity at $z_0$ and the extended function $g$ has an isolated zero at $z_0$.

X.5.2.11. Theorem. If $f$ has a singularity at $z_0$, the following are equivalent:

(i) $f$ has a pole of order $n$ at $z_0$.

(ii) There is a function $\phi$ which is holomorphic in a neighborhood of $z_0$ with $\phi(z_0) \neq 0$, such that
\[ f(z) = \frac{\phi(z)}{(z - z_0)^n} \]
for all $z$ in a deleted neighborhood of $z_0$.

(iii) In the Laurent expansion for $f$ around $z_0$ (X.5.2.4.), $b_n \neq 0$ but $b_k = 0$ for all $k > n$.

Proof: (i) $\Rightarrow$ (ii): There is a function $g$, holomorphic in a neighborhood of $z_0$, such that $\frac{1}{f(z)} = g(z)$ for $z$ in a deleted neighborhood of $z_0$, and there is a function $\psi$, holomorphic in a neighborhood of $z_0$ with $\psi(z_0) \neq 0$, such that $g(z) = (z - z_0)^n \psi(z)$ in a neighborhood of $z_0$. Then $\psi(z) \neq 0$ in a neighborhood $U$ of $z_0$, and for all $z \in U \setminus \{z_0\}$,
\[ f(z) = \frac{1}{(z - z_0)^n} \cdot \frac{1}{\psi(z)} . \]

(ii) $\Rightarrow$ (i): There is a neighborhood $U$ of $z_0$ such that $\phi(z) \neq 0$ for all $z \in U$ and $f(z) = \frac{\phi(z)}{(z - z_0)^n}$ for all $z \in U \setminus \{z_0\}$. Then
\[ g(z) = (z - z_0)^n \frac{1}{\phi(z)} . \]
is holomorphic on \( U \), has a zero of order \( n \) at \( z_0 \), and agrees with \( \frac{1}{f} \) on \( U \setminus \{z_0\} \).

(ii) \( \Rightarrow \) (iii): Suppose the equation holds on \( D \setminus \{z_0\} \), where \( D \) is an open disk centered at \( z_0 \). Let

\[
\sum_{k=0}^{\infty} c_k (z - z_0)^k
\]

be the Taylor expansion of \( \phi \) around \( z_0 \), which converges to \( \phi \) on \( D \). Then \( c_0 = \phi(z_0) \neq 0 \). Then, for \( z \in D \setminus \{z_0\} \),

\[
f(z) = \frac{c_0}{(z - z_0)^n} + \cdots + \frac{c_{n-1}}{z - z_0} + \sum_{k=n}^{\infty} c_k (z - z_0)^{k-n}
\]

where the power series converges u.c. on \( D \). This must be the Laurent expansion for \( f \) around \( z_0 \) by uniqueness of Laurent expansions.

(iii) \( \Rightarrow \) (ii): Suppose

\[
f(z) = \frac{b_n}{(z - z_0)^n} + \cdots + \frac{b_1}{z - z_0} + \sum_{k=0}^{\infty} a_k (z - z_0)^k
\]

is the Laurent expansion of \( f \) around \( z_0 \), with \( b_n \neq 0 \), converging u.c. on \( D \setminus \{z_0\} \), for an open disk \( D \) centered at \( z_0 \). Set \( c_k = b_{n-k} \) for \( 0 \leq k \leq n - 1 \) and \( c_k = a_{k-n} \) for \( k \geq n \). Then the power series

\[
\sum_{k=0}^{\infty} c_k (z - z_0)^k
\]

converges u.c. on \( D \) to a holomorphic function \( \phi \), with the required properties.

\[\Box\]

\textbf{X.5.2.12.} There is thus a symmetry between poles and zeroes: a pole of order \( n \) can be thought of as a zero of order \( -n \), or a zero of order \( n \) can be considered a pole of order \( -n \). Indeed, if \( f \) is holomorphic in a deleted neighborhood of \( z_0 \) and has a pole or removable singularity at \( z_0 \), or is already defined and differentiable also at \( z_0 \), then the Laurent expansion of \( f \) around \( z_0 \) is of the form

\[
f(z) = \sum_{k=-n}^{\infty} a_k (z - z_0)^k
\]

where \( n \in \mathbb{Z} \) may be positive, negative, or 0 and \( a_n \neq 0 \). If \( n > 0 \), \( f \) has a zero of order \( n \) at \( z_0 \) (perhaps after a singularity at \( z_0 \) has been removed). If \( n < 0 \), \( f \) has a pole of order \( -n \) at \( z_0 \). If \( n = 0 \), then \( f \) is holomorphic around \( z_0 \) and \( f(z_0) \neq 0 \), and we may regard \( f \) as having a zero or pole at \( z_0 \) of order 0.

\textbf{X.5.2.13.} Examples. The function \( f(z) = \frac{1}{(z - z_0)^n} \) has a pole of order \( n \) at \( z_0 \), as does the function

\[
g(z) = \frac{e^z}{(z - z_0)^n}.
\]

Since the Taylor series for \( e^z \) around \( z_0 \) is

\[
e^z = \sum_{k=0}^{\infty} \frac{e^{z_0}}{k!} (z - z_0)^k
\]

953
we have that the Laurent series of \( g \) around \( z_0 \) is
\[
e^{z_0} \frac{e^{z_0}}{(z - z_0)^n} + \cdots + \frac{e^{z_0}}{(n - 1)! (z - z_0)^{n-1}} + \sum_{k=n}^{\infty} \frac{e^{z_0}}{k!} (z - z_0)^{k-n}.
\]

**Essential Singularities**

**X.5.2.14.** Definition. If \( f \) has a singularity (isolated singularity) at \( z_0 \), the singularity is an essential singularity if it is not a removable singularity or a pole.

**X.5.2.15.** We should emphasize that the term essential singularity is used only for isolated singularities. Thus the “singularities” at 0 in X.5.2.2. (iii)–(iv) are not essential singularities (these “singularities” are even more “essential” than essential singularities!)

Essential singularities can be characterized from the Laurent expansion by a process of elimination:

**X.5.2.16.** Proposition. If \( f \) has a singularity at \( z_0 \), and
\[
f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}
\]
is its Laurent expansion around \( z_0 \), then the singularity at \( z_0 \) is an essential singularity if and only if infinitely many \( b_k \) are nonzero.

**Proof:** The singularity is removable if and only if all \( b_k \) are zero, and a pole if and only if at least one \( b_k \), but only finitely many, is nonzero.

**X.5.2.17.** Examples. (i) Let \( f(z) = e^{1/z} \). Then \( f \) has a singularity at 0. As \( z \to 0 \) along the positive real axis, \( |f(z)| \to +\infty \), so the singularity at 0 is not removable. And as \( z \to 0 \) along the negative real axis, \( f(z) \to 0 \), so the singularity is not a pole. Thus it is an essential singularity.

This can be seen another way. If \( z \neq 0 \), the inverse power series
\[
\sum_{k=1}^{\infty} \frac{1}{k!} \frac{1}{z^k}
\]
converges to \( e^{1/z} \). Thus this is the Laurent expansion of \( f \) around 0. Since infinitely many negative power coefficients are nonzero, the singularity is essential.

(ii) A closely related example is Cauchy’s example (V.17.2.2.) of a \( C^\infty \) function from \( \mathbb{R} \) to \( \mathbb{R} \) which is not (real) analytic. Let \( f(z) = e^{-1/z^2} \). This function is holomorphic on \( \mathbb{C} \setminus \{0\} \). As \( z \to 0 \) along the real axis, \( f(z) \to 0 \), and if we define \( f(0) = 0 \), it becomes \( C^\infty \) as a function from \( \mathbb{R} \) to \( \mathbb{R} \). But if \( z \to 0 \) along the imaginary axis, \( |f(z)| \to +\infty \). Thus \( f \), as a function from \( \mathbb{C} \setminus \{0\} \) to \( \mathbb{C} \), has an essential singularity at 0.

The next result shows that the behavior of a holomorphic function near an essential singularity is very complicated.
Theorem. [Casorati-Weierstrass] Let \( f \) have an essential singularity at \( z_0 \), and \( U \) a deleted neighborhood of \( z_0 \) contained in the domain of \( f \). Then the image of \( U \) under \( f \) is dense in \( \mathbb{C} \). If \( w \in \mathbb{C} \), then there is a sequence \((z_n)\) in \( U \) with \( z_n \to z_0 \) and \( f(z_n) \to w \).

Proof: The two statements are obviously equivalent (the second follows from the first by taking a decreasing sequence of such neighborhoods), so we prove the first. The proof is very similar to the proof of X.3.3.11. Suppose \( f(U) \) is not dense in \( \mathbb{C} \). Then there is a \( w \in \mathbb{C} \) and an \( \epsilon > 0 \) such that \( |f(z) - w| \geq \epsilon \) for all \( z \in U \). Set

\[
g(z) = \frac{1}{f(z) - w}
\]

for \( z \in U \); then \( g \) is holomorphic on \( U \) and has a singularity at \( z_0 \). We have \( |g(z)| \leq \frac{1}{\epsilon} \) for all \( z \in U \), so by X.5.2.7. this singularity is removable, i.e.

\[
c = \lim_{z \to z_0} g(z)
\]

exists. Then, if \( c \neq 0 \), we have

\[
\lim_{z \to z_0} f(z) = w + \frac{1}{c}
\]

so the singularity of \( f \) at \( z_0 \) is removable, a contradiction. And if \( c = 0 \), we have

\[
\lim_{z \to z_0} |f(z)| = +\infty
\]

so the singularity of \( f \) at \( z_0 \) is a pole, another contradiction.

Actually, a much stronger result is true:

Theorem. [Big Picard Theorem] Let \( f \) have an essential singularity at \( z_0 \), and \( U \) a deleted neighborhood of \( z_0 \) contained in the domain of \( f \). Then, with at most one exception, \( f \) takes every complex value infinitely many times in \( U \).

This result is the best possible: \( e^{1/z} \) never takes the value 0. The proof is beyond the scope of this book; see e.g. [3]. The Big Picard Theorem implies the Little Picard Theorem X.3.3.12.; cf. Exercise X.8.2.14..

Behavior at Infinity

It also makes sense to say that certain functions have singularities at infinity:

Definition. Let \( f \) be holomorphic on a set \( \{ z : |z| > R \} \) for some \( R \). Then \( f \) has a singularity at infinity. We say \( f \) has a removable singularity [resp. pole, essential singularity] at infinity if the function \( g(z) = f \left( \frac{1}{z} \right) \) has a removable singularity [resp. pole, essential singularity] at 0.

We can characterize the types of singularities at infinity:
**X.5.3.2.** **Proposition.** Let $f$ have a singularity at infinity. Then

(i) The singularity at infinity is removable if and only if $\lim_{z \to \infty} f(z)$ exists, i.e. there is a number $c \in \mathbb{C}$ such that, for every $\epsilon > 0$, there is an $M$ such that $|f(z) - c| < \epsilon$ whenever $|z| > M$.

(ii) The singularity at infinity is a pole if and only if $\lim_{z \to \infty} |f(z)| = +\infty$, i.e. for every $N$ there is an $M$ such that $|f(z)| > N$ whenever $|z| > M$.

(iii) Otherwise, the singularity at infinity is an essential singularity.

**X.5.3.3.** If $f$ has a singularity at infinity, i.e. it is holomorphic on an annulus $\Omega = \{ z : |z| > R \}$, then it has a Laurent expansion

$$f(z) = \sum_{k=0}^{\infty} a_k z^k + \sum_{k=1}^{\infty} b_k \frac{1}{z^k}$$

around 0 which converges on $\Omega$. If $g(z) = f \left( \frac{1}{z} \right)$, then the Laurent expansion of $g$ around 0 converging on $\{ z : 0 < |z| < 1/R \}$ is obtained by replacing $z$ by $z^{-1}$ in this expansion, effectively interchanging the $a_k$ and $b_k$. Thus we obtain a classification of singularities at infinity from the Laurent expansion:

**X.5.3.4.** **Proposition.** If $f$ is holomorphic on the annulus $\Omega = \{ z : |z| > R \}$, with Laurent expansion

$$f(z) = \sum_{k=0}^{\infty} a_k z^k + \sum_{k=1}^{\infty} b_k \frac{1}{z^k}$$

around 0 which converges on $\Omega$, then the singularity at infinity is

(i) Removable if (and only if) all $a_k \ (k \geq 1)$ are 0.

(ii) A pole if (and only if) some $a_k \ (k \geq 1)$, but only finitely many, are nonzero.

(iii) Essential if (and only if) infinitely many $a_k$ are nonzero.

As a corollary, we get an important property of entire functions:

**X.5.3.5.** **Corollary.** Let $f$ be an entire function. Then the singularity at infinity is

(i) Removable if $f$ is constant.

(ii) A pole if $f$ is a nonconstant polynomial.

(iii) Essential otherwise.

**Proof:** The Laurent expansion of $f$ on any annulus centered at 0 is just the Taylor series of $f$ around 0. 

The fact that a nonconstant polynomial has a pole at infinity is the essential content of Lemma X.1.4.1.
X.5.3.6. We can also define the order of a pole of \( f \) at infinity to be the order of the pole of \( g(z) = f \left( \frac{1}{z} \right) \) at 0. And we can consider a zero at infinity: \( f \) has a zero at infinity if \( g \) has a removable singularity at 0 which, when removed, is a zero. The order of the zero of \( f \) at infinity is the order of this zero of \( g \). These orders can be read from the Laurent expansion of \( f \) in an unbounded annulus in the obvious way, as the largest \( n \) for which \( a_n \neq 0 \) for a pole, and similarly for the \( b_n \) for a zero.

X.5.3.7. We have used expansions around 0 converging in an unbounded annulus centered at 0 merely for convenience. The same results hold if we use unbounded annuli centered at any \( z_0 \) by using \( g(z) = f \left( \frac{1}{z-z_0} + z_0 \right) \), i.e. \( g(z - z_0) = f \left( \frac{1}{z-z_0} \right) \). This is occasionally useful when \( f \) is naturally expanded around some \( z_0 \neq 0 \).

X.5.4. Meromorphic Functions

X.5.5. The Residue Theorem

X.5.5.1. It has been suggested that Cauchy got the idea for this theorem from his dog, which left a residue at every pole. [Actually Cauchy did not use the term pole, although he did introduce and use résidue.]

X.5.6. Principle of the Argument

Hurwitz’s Theorem

The following theorem has several interesting and useful corollaries, any of which can be called “Hurwitz’s Theorem” (and are in various references).

X.5.6.1. Theorem. [Hurwitz’s Theorem] Let \( \Omega \) be an open set in \( \mathbb{C} \), and \((f_n)\) a sequence of holomorphic functions on \( \Omega \) converging normally on \( \Omega \) to a (holomorphic) function \( f \). Suppose \( f \) has a zero of order \( m \) at \( z_0 \in \Omega \), and let \( \gamma \) be a circle centered at \( z_0 \) bounding a closed disk contained in \( \Omega \) and containing no other zeroes of \( f \). Then, for all sufficiently large \( n \), \( f_n \) has exactly \( m \) zeroes inside \( \gamma \), counted with multiplicity.

Proof: Since \( f_n \to f \) uniformly on \( \Omega \), \( f_n \) has no zeroes on \( \gamma \) for sufficiently large \( n \). Thus, passing to a tail of the sequence \((f_n)\), we may assume no \( f_n \) has zeroes on \( \gamma \). Then \( \frac{f'_n}{f_n} \to \frac{f'}{f} \) uniformly on \( \gamma \), so if \( Z_{f_n} \) is the number of zeroes of \( f_n \) inside \( \gamma \), we have

\[
Z_{f_n} = \frac{1}{2\pi i} \int_{\gamma} \frac{f'_n}{f_n} \to \frac{1}{2\pi i} \int_{\gamma} \frac{f'}{f} = m. 
\]

\[\boxed{\text{Wow!}}\]
Let $\Omega$ be a connected open set in $\mathbb{C}$, and $(f_n)$ a sequence of holomorphic functions on $\Omega$ converging normally on $\Omega$ to a (holomorphic) function $f$ which is not identically 0. If $f(z_0) = 0$, then there is a sequence $(k_n)$ in $\mathbb{N}$ and a sequence $(z_n)$ in $\Omega$ with $z_n \to z_0$ and $f_{k_n}(z_n) = 0$ (informally, every zero of $f$ is a limit of zeroes of the $f_n$).

Proof: Let $(\gamma_n)$ be a sequence of circles centered at $z_0$ of radius decreasing to 0, within a closed disk contained in $\Omega$ containing no other zeroes of $f$. 😊

The restriction that $f$ be not identically 0 is necessary: consider the sequence $f_n(z) = e^{z_n}$.

Corollary. Let $\Omega$ be a connected open set in $\mathbb{C}$, and $(f_n)$ a sequence of holomorphic functions on $\Omega$ converging normally on $\Omega$ to a (holomorphic) function $f$ which is not identically 0. If each $f_n$ has at most $m$ zeroes in $\Omega$ (counted with multiplicity), then $f$ also has at most $m$ zeroes in $\Omega$ (counted with multiplicity).

Proof: Suppose $f$ has zeroes at $z_1, \ldots, z_r$ of multiplicity $m_1, \ldots, m_r$ respectively, with $m_1 + \cdots + m_r > m$. For each $k$, let $\gamma_k$ be a small circle centered at $z_k$ so that the circles do not overlap and the bounded closed disks are contained in $\Omega$ and contain no other zeroes of $f$. Then, for sufficiently large $n$, $f_n$ has $m_k$ zeroes inside $\gamma_k$ for each $k$, and hence more than $m$ zeroes in all, a contradiction. 😊

Under the hypotheses, even if all $f_n$ have exactly $m$ zeroes in $\Omega$, the limit function can have fewer than $m$ if the zeroes of the $f_n$ cluster on the boundary of $\Omega$. As a very simple case, let $\Omega$ be the open right half plane and $f_n(z) = z - \frac{1}{n}$ for each $n$.

The special case $m = 0$ is often stated as Hurwitz’s Theorem:

Corollary. Let $\Omega$ be a connected open set in $\mathbb{C}$, and $(f_n)$ a sequence of holomorphic functions on $\Omega$ converging normally on $\Omega$ to a (holomorphic) function $f$ which is not identically 0. If each $f_n$ is never zero in $\Omega$, then $f$ is also never zero in $\Omega$.

Again, the example $f_n(z) = z_n$ shows that the assumption that $f$ is not identically 0 is necessary.

The following consequence of a special case of Corollary. is especially important:

Corollary. Let $\Omega$ be a connected open set in $\mathbb{C}$, and $(f_n)$ a sequence of holomorphic functions on $\Omega$ converging normally on $\Omega$ to a (holomorphic) function $f$ which is not constant. If each $f_n$ is one-to-one, then $f$ is also one-to-one.

Proof: For any $w \in \mathbb{C}$, the function $g_n = f_n - w$ has at most one zero in $\Omega$, and hence the limit function $f - w$ also has at most one zero in $\Omega$. 😊

This result can be generalized: if each $f_n$ is at most $m$-to-one, and $f$ is not constant, then it is at most $m$-to-one too.
X.6. Conformal Equivalence and the Riemann Mapping Theorem

Throughout this section, we will let $\mathbb{D}$ denote the open unit disk in $\mathbb{C}$.

X.6.1. Conformal Equivalence

X.6.2. Simply Connected Sets in $\mathbb{C}$

It is a remarkable fact, partly topological and partly analytic, that the open sets in $\mathbb{C}$ which are conformally equivalent to $\mathbb{C}$ or $\mathbb{D}$ can be cleanly characterized in a number of ways. We first recall a special case of a definition from topology (5):

**X.6.2.1. Definition.** Let $\Omega$ be a nonempty open set in $\mathbb{C}$. Then $\Omega$ is *simply connected* if it is connected and every path in $\Omega$ is homotopic in $\Omega$ to a constant path.

Some authors do not require a simply connected space to be connected; but this seems a little silly, especially for open subsets of the plane. In fact, a simply connected space should be path-connected.

**X.6.2.2. Theorem.** Let $\Omega$ be a nonempty connected open set in $\mathbb{C}$ ($\mathbb{R}^2$). The following are equivalent:

**Topological conditions:**

(i) $\Omega$ is simply connected.

(ii) For any $p, q \in \Omega$, any two paths from $p$ to $q$ are homotopic in $\Omega$.

(iii) $\Omega$ is contractible.

(iv) $\Omega$ is homeomorphic to $\mathbb{C}$ and $\mathbb{D}$.

**Intermediate condition:**

(v) For every closed contour $\gamma$ in $\Omega$ and $z \in \mathbb{C} \setminus \Omega$, the winding number $w(\gamma, z)$ of $\gamma$ around $z$ is 0.

**Analytic conditions:**

(vi) The contour integral of any holomorphic function on $\Omega$ around any closed curve in $\Omega$ is 0.

(vii) The contour integral of any holomorphic function on $\Omega$ along any contour $\gamma$ in $\Omega$ depends only on the endpoints of $\gamma$.

(viii) Every holomorphic function on $\Omega$ has an antiderivative on $\Omega$.

(ix) $\Omega$ is conformally equivalent to $\mathbb{C}$ or $\mathbb{D}$.

If $\Omega$ is bounded, these are also equivalent to

(x) $\partial \Omega$ is connected.

(xi) $\mathbb{C} \setminus \Omega$ is connected.
X.6.2.3. The implications (iii) ⇒ (i) ⇔ (ii) hold in any topological space (.), and (ix) ⇒ (iv) ⇒ (iii) ⇒ (i) are trivial. The implication (x) ⇒ (xi) is true in great generality (.), but (xi) ⇒ (x) is much more difficult and special (.). Conditions (i)–(iv) and (x) make sense for open sets in \( \mathbb{R}^n \) for any \( n \) (and indeed much more generally; (iv) means \( \Omega \) is homeomorphic to \( \mathbb{R}^n \)), but are not equivalent for \( n \neq 2 \): the interior of a solid torus in \( \mathbb{R}^3 \) satisfies (x) but not (i), and the open region between two concentric spheres in \( \mathbb{R}^3 \) satisfies (i) but not (iii) or (x). And a contractible open set in \( \mathbb{R}^3 \) is not homeomorphic to \( \mathbb{R}^3 \) in general (.). (Note also that (i) ⇒ (x) fails for bounded open sets in \( \mathbb{R} \); (i)–(iv) are equivalent in \( \mathbb{R} \) since a simply connected open set in \( \mathbb{R} \) is just an open interval.) The implications (xi) ⇒ (i) and (iv) ⇒ (xi) do not hold for unbounded open sets in \( \mathbb{R}^2 \), e.g. \( \mathbb{C} \setminus \{0\} \) and an infinite open strip. (If we work in the Riemann sphere, (i), (x), and (xi) are equivalent.)

X.6.2.4. Condition (v) is called an “intermediate condition” since it is essentially topological, but only makes sense in \( \mathbb{R}^2 \) and is closely connected with contour integration. We have (v) ⇔ (vi) ⇔ (vii) ⇔ (viii) by (.) and (.) . If \( \Omega \) is bounded, (x) ⇒ (v) since winding number is constant on connected subsets of \( \mathbb{C} \setminus \Omega \) (.) and is 0 on the unbounded component. Also, (i) ⇒ (vi) is () .

X.6.2.5. Thus, we have the following implications:

\[
\text{(ix)} \Rightarrow \text{(iv)} \Rightarrow \text{(iii)} \Rightarrow \text{(i)} \Leftrightarrow \text{(ii)} \Leftrightarrow \text{(v)} \Leftrightarrow \text{(vi)} \Leftrightarrow \text{(vii)} \Leftrightarrow \text{(viii)}
\]

and, if \( \Omega \) is bounded, (x) ⇒ (v).

X.6.2.6. To finish the proof of the theorem, we will show (viii) ⇒ (ix) and, if \( \Omega \) is bounded, (v) ⇒ (x). The implication (viii) ⇒ (ix) is the most spectacular result, called the Riemann Mapping Theorem; the proof is complicated and is given in X.6.3. In [GK06], an open set satisfying (viii) is called holomorphically simply connected, which seems like good terminology.

X.6.2.7. Even the implication (i) ⇒ (iv), a purely topological result, is highly nontrivial. The simplest proof of this implication seems to be via (ix) using the Riemann Mapping Theorem, a much stronger result! As noted above, in Euclidean spaces this result is special to \( \mathbb{R}^2 \).

X.6.3. The Riemann Mapping Theorem
The Riemann Mapping Theorem is one of the major theorems of Complex Analysis:

X.6.3.1. Theorem. [RIEMANN MAPPING THEOREM] Let \( \Omega \) be a holomorphically simply connected open set in \( \mathbb{C} \), i.e. a nonempty open set satisfying (viii) of X.6.2.2., with \( \Omega \neq \mathbb{C} \). Then \( \Omega \) is conformally equivalent to \( \mathbb{D} \). More precisely, if \( z_0 \in \Omega \), then there is a unique conformal equivalence \( f : \Omega \to \mathbb{D} \) with \( f(z_0) = 0 \) and \( f'(z_0) \) a positive real number.

X.6.3.2. Note that the restriction \( \Omega \neq \mathbb{C} \) is necessary: a conformal bijection from \( \mathbb{C} \) to \( \mathbb{D} \) would be a bounded nonconstant entire function, contradicting Liouville’s Theorem (.) .
X.6.3.3. We first show uniqueness of $f$, assuming existence; we will first relax the last condition and show near uniqueness in that setting. Suppose $f$ and $g$ are conformal equivalences from $\Omega$ to $\mathbb{D}$ with $f(z_0) = g(z_0) = 0$. Set $h = f \circ g^{-1}$; then $h$ is a conformal bijection from $\mathbb{D}$ to $\mathbb{D}$, and $h(0) = 0$. Thus there is an $\alpha \in \mathbb{C}$, $|\alpha| = 1$, with $h(z) = \alpha z$ (note that $g'(z_0) \neq 0$ by (i)). As a consequence, $f(z) = \alpha g(z)$, i.e. $f$ is just $g$ followed by a rotation (and vice versa). By the Chain Rule we have

$$\alpha = h'(0) = f'(g^{-1}(0))(g^{-1})'(0) = f'(z_0)\frac{1}{g'(z_0)}$$

(note that $g'(z_0) \neq 0$ by (i)). If $f'(z_0)$ and $g'(z_0)$ are positive real numbers, then $\alpha$ is a positive real number of absolute value $1$, i.e. $\alpha = 1$, $f = g$.

X.6.3.4. To show existence, we roughly follow the exposition of [GK06]. Let $\mathcal{F}$ be the set of all one-to-one holomorphic functions $f : \Omega \rightarrow \mathbb{D}$ with $f(z_0) = 0$. We will show:

1. $\mathcal{F} \neq \emptyset$.
2. There is an $f \in \mathcal{F}$ for which $|f'(z_0)|$ is maximum for all functions in $\mathcal{F}$.
3. A function $f \in \mathcal{F}$ for which $|f'(z_0)|$ is maximum is a conformal bijection.
4. Such an $f$ can be adjusted to make $f'(z_0)$ a positive real number.

As pointed out in [GK06], this outline can be motivated by the case where $\Omega = \mathbb{D}$ and $z_0 = 0$: a holomorphic map $f$ from $\mathbb{D}$ to $\mathbb{D}$ with $f(0) = 0$ is a conformal bijection if and only if $|f'(0)| = 1$, which is the largest possible value of $|f'(0)|$ for holomorphic functions from $\mathbb{D}$ to $\mathbb{D}$ sending $0$ to $0$ by the Schwarz Lemma.

For both 1 and 3, we need the following lemma:

X.6.3.5. Lemma. Let $\Omega$ be a holomorphically simply connected open set in $\mathbb{C}$, and $f$ holomorphic on $\Omega$. Suppose $f$ is never $0$ on $\Omega$. Then

(i) There is a holomorphic function $g$ on $\Omega$ with $e^g = f$ (i.e. $g$ is a “logarithm” for $f$).

(ii) There is a holomorphic function $h$ on $\Omega$ with $h^2 = f$ (i.e. $h$ is a “square root” for $f$).

Proof: (i): Since $f$ is never $0$ on $\Omega$, $\frac{f'}{f}$ is holomorphic on $\Omega$. Let $\phi$ be an antiderivative for $\frac{f'}{f}$ on $\Omega$. Fix a $z_0 \in \Omega$ and $\alpha$ a value for $\log(f(z_0))$, i.e. so that $e^\alpha = f(z_0)$. Let $g(z) = \phi(z) + \alpha - \phi(z_0))$. Then $g$ is an antiderivative for $\frac{f'}{f}$, and $g(z_0) = \alpha$, so $e^{g(z)} = f(z_0)$. Then, if $\psi = \frac{e^{g}}{f}$, we have, for all $z \in \Omega$,

$$\psi'(z) = \frac{f(z)e^{g(z)}g'(z) - e^{g(z)}f'(z)}{|f(z)|^2} = \frac{f(z)e^{g(z)}f'(z) - e^{g(z)}f'(z)}{|f(z)|^2} = 0$$

so $\psi$ is constant since $\Omega$ is connected. Since $\psi(z_0) = 1$, we have $e^{g(z)} = f(z)$ for all $z \in \Omega$.

(ii): Let $g$ be as in (i), and let $h = e^{g/2}$.  

$\diamondsuit$
Lemma. By the Chain Rule, \[ s \text{Lemma} \]

Later that \( f \) and, since \( ! \)

where \( \phi \)

We have that \( d \)

Montel’s Theorem (\[ X.6.3.7. \])

and note that if \( B \)

is an open mapping, there is an \( h \)

never takes the value 0. By

to get an element of

We now prove 1. If \( X.6.3.6. \)

then \( s \)

\[ w \]

Set \( D \)

Thus \( X.6.3.8. \).

Set \( D \)

be the restriction to \( \Omega \)

onto \( \Omega \) and let \( h \)

is one-to-one, but if it takes the value \( 0 \). By \( \epsilon \)

This completes the proof of the Riemann Mapping Theorem. 

\[ X.6.3.9. \]

Step 4 is easy: If \( f \) is a conformal bijection from \( \Omega \) to \( D \) (e.g. the one constructed in 2), then \( f'(z_0) \neq 0 \). Set \( \alpha = \frac{f'(z_0)}{f(z_0)} \) and \( g(z) = \alpha f(z) \). Then \( g \) is also a conformal bijection with \( g(z_0) = 0 \), and \( g'(z_0) = |f'(z_0)| > 0 \).

This completes the proof of the Riemann Mapping Theorem.
X.7. Analytic Continuation and Riemann Surfaces

In this section we will assume all open sets we work on are connected. We will use the term region to mean a nonempty connected open subset of \( \mathbb{C} \).

X.7.1. Analytic Function Elements

The Rigidity Theorem implies that if \( \Omega \) and \( \bar{\Omega} \) are regions in \( \mathbb{C} \), whose intersection is nonempty, and \( f \) is holomorphic on \( \Omega \), then there is at most one function \( \tilde{f} \) which is holomorphic on \( \bar{\Omega} \) and agrees with \( f \) on the intersection. There may or may not be such an \( \tilde{f} \); but if there is, \( f \) and \( \tilde{f} \) can be reasonably considered to be “pieces” of the same function which is holomorphic on some unspecified domain containing both \( \Omega \) and \( \bar{\Omega} \).

Things can be more complicated, however. The intersection of \( \Omega \) and \( \bar{\Omega} \) is not necessarily connected, and the functions \( f \) and \( \tilde{f} \) can agree on part of the intersection but not all. An example is two branches of \( \log z \) \((\alpha, \beta)\) corresponding to \( \alpha \) and \( \beta \) with \( |\alpha - \beta| < 2\pi \).

A related difficulty which actually provides a solution to the first difficulty is that if \( \Omega_1, \Omega_2, \) and \( \Omega_3 \) are regions, and \( f_j \) is holomorphic on \( \Omega_j \), if \( f_1 = f_2 \) on \( \Omega_1 \cap \Omega_2 \) and \( f_2 = f_3 \) on \( \Omega_2 \cap \Omega_3 \), then \( f_1 \) and \( f_3 \) should be regarded as “pieces” of the same function even if \( \Omega_1 \) and \( \Omega_3 \) are disjoint, or if they overlap but \( f_1 \) and \( f_3 \) do not agree on \( \Omega_1 \cap \Omega_3 \), as can happen, e.g. with suitable branches of the logarithm. The solution is to regard \( f \) on \( \Omega \) and \( \tilde{f} \) on \( \bar{\Omega} \) as “pieces” of the same function if there is a finite chain of intermediate “pieces” \( f_j \) on \( \Omega_j \) which agree on consecutive overlaps. Using this extension of equivalence, we may regard any two branches of \( \log z \), even ones whose angles differ by more than \( 2\pi \), to be “pieces” of the same big function. The domain of this big function is not a subset of \( \mathbb{C} \), but a large abstract set made up of overlapping pieces which can be identified with regions in \( \mathbb{C} \) (technically a complex 1-manifold \( (\cdot) \)). This domain is called the Riemann surface of the function.

We now discuss the details.

X.7.1.1. Definition. An analytic function element is a pair \((\Omega, f)\), where \( \Omega \) is a region (nonempty connected open set) in \( \mathbb{C} \) and \( f \) is holomorphic on \( \Omega \).

Two analytic function elements \((\Omega, f)\) and \((\bar{\Omega}, \tilde{f})\) are equivalent if there is a finite chain \((\Omega_1, f_1), (\Omega_n, f_n) = (\bar{\Omega}, \tilde{f})\) \((1 \leq j \leq n)\) such that \((\Omega_1, f_1) = (\Omega, f)\), \((\Omega_n, f_n) = (\bar{\Omega}, \tilde{f})\), and \(\Omega_j \cap \Omega_{j+1} \neq \emptyset\) and \(f_j = f_{j+1}\) on \(\Omega_j \cap \Omega_{j+1}\) for \(1 \leq j < n\).

A total analytic function is an equivalence class of analytic function elements.

X.7.1.2. It is obvious that equivalence of analytic function elements is indeed an equivalence relation. Every analytic function element determines a unique total analytic function, namely its equivalence class. Equivalent analytic function elements determine the same total analytic function.

X.7.1.3. Proposition. (i) If \((\Omega, f)\) and \((\bar{\Omega}, \tilde{f})\) are analytic function elements, \(\Omega \cap \bar{\Omega} \neq \emptyset\), and \(f = \tilde{f}\) on \(\Omega \cap \bar{\Omega}\), then \((\Omega, f)\) and \((\bar{\Omega}, \tilde{f})\) are equivalent.

(ii) If \((\Omega, f)\) is an analytic function element, \(\Omega\) a subregion of \(\Omega\), and \(\tilde{f} = f|_{\Omega}\), then the analytic function element \((\Omega, \tilde{f})\) is equivalent to \((\Omega, f)\).

(iii) If \(\{(\Omega_k, f_k) : k \in K\}\) are analytic function elements, such that \(f_j = f_k\) on \(\Omega_j \cap \Omega_k\) for all \(j\) and \(k\), and \(\Omega = \cup K \Omega_k\) is connected (e.g. if \(\cap K \Omega_k \neq \emptyset\)), set \(f = f_k\) on \(\Omega_k\) for each \(k\). Then \(f\) is a well-defined holomorphic function on \(\Omega\), and the analytic function element \((\Omega, f)\) is equivalent to each \((\Omega_k, f_k)\).
X.7.2. The Riemann Surface of a Total Analytic Function

X.7.2.1. We can put the domains of the elements of a total analytic function together to get a “surface” which is the natural domain of the function. Informally, we take the “union” of the domains of the elements, with points in overlapping elements identified provided the function agrees around these points. More precisely:

X.7.2.2. Definition. Let \( f = \{ (\Omega_k, f_k) : k \in K \} \) be a total analytic function, i.e. the \((\Omega_k, f_k)\) are equivalent analytic function elements. Let \( D_f \) be the disjoint union of the \( \Omega_k \) (even if \( \Omega_j = \Omega_k \) as subsets of \( \mathbb{C} \), if \( j \neq k \) they are disjoint subsets of \( D_f \)). Put an equivalence relation on \( D_f \) as follows. If \( z_j \in \Omega_j \), \( z_k \in \Omega_k \), set \( z_j \sim z_k \) if \( z_j = z_k \in \Omega_j \cap \Omega_k \subseteq \mathbb{C} \) and \( f_j(z) = f_k(z) \) for all \( z \) in a neighborhood of \( z_j = z_k \). Write \( \hat{\Omega}_k = \{ [z] : z \in \Omega_k \} \). Let \( S_f \) be the set of equivalence classes. \( S_f \) is the Riemann surface of the total analytic function \( f \).

The total analytic function \( f \) gives a well-defined function \( f \) from \( S_f \) to \( \mathbb{C} \) by \( f([z]) = f_k(z) \) for \( z \in \Omega_k \). The function \( f \) is called the (complete) analytic function defined by \( f \). Each \((\Omega_k, f_k)\) is a branch of \( f \) or \( f \).

X.7.2.3. There is a natural injective map \( \iota_k \) from \( \Omega_k \) to \( S_f \) for each \( k \), and the restriction of \( f \) to the image \( \hat{\Omega}_k \) of \( \Omega_k \) is exactly \( f_k \).

We topologize \( S_f \) by taking the \( \hat{\Omega}_k \) as a base for the topology. Then each \( \hat{\Omega}_k \) is an open subset of \( S_f \), and the inclusion of \( \Omega_k \) into \( S_f \) is a homeomorphism onto its image. The function \( f : S_f \to \mathbb{C} \) is continuous.

X.7.2.4. The natural identifications of the subsets \( \hat{\Omega}_k \) of \( S_f \) with the corresponding subsets \( \Omega_k \) of \( \mathbb{C} \) give an atlas defining the structure of a complex analytic 1-manifold on \( S_f \). This atlas has the special property that the transition functions between charts are identity maps on open sets in \( \mathbb{C} \). (This natural atlas is not a maximal atlas.)

The next proposition is an immediate corollary of ().

X.7.2.5. Proposition. Let \( f \) be a total analytic function, and \((\Omega_j, f_j)\) and \((\Omega_k, f_k)\) analytic function elements in \( f \). If \( z_0 \in \Omega_j \cap \Omega_k \), then either \( \iota_j(z) = \iota_k(z) \) for all \( z \) in a neighborhood of \( z_0 \) (if \( \iota_j(z_0) = \iota_k(z_0) \)), or \( f_j(z) \neq f_k(z) \) for all \( z \) in a deleted neighborhood of \( z_0 \) and in particular \( \iota_j(z) \neq \iota_k(z) \) for all \( z \) in a neighborhood of \( z_0 \) (if \( \iota_j(z_0) \neq \iota_k(z_0) \)).

X.7.2.6. Corollary. Let \( f \) be a total analytic function with Riemann surface \( S_f \). Then \( S_f \) is Hausdorff.

X.7.2.7. Proposition. Let \( f \) be a total analytic function with Riemann surface \( S_f \). Then \( S_f \) is path-connected.

Proof: Suppose \((\Omega_1, f_1), \ldots, (\Omega_n, f_n)\) are elements of \( f \), and \( z_j \in \Omega_j \), \( z_k \in \Omega_k \). Let \( (\Omega_1, f_1), \ldots, (\Omega_n, f_n) \) be a chain with \( (\Omega_1, f_1) = (\Omega_j, f_j), (\Omega_n, f_n) = (\Omega_k, f_k) \), and \( \Omega_m \cap \Omega_{m+1} \neq \emptyset \) and \( f_m = f_{m+1} \) on \( \Omega_m \cap \Omega_{m+1} \) for \( 1 \leq m < n \). For each \( m, 1 \leq m \leq n-1 \), choose \( z_m \in \Omega_m \cap \Omega_{m+1} \). We have \( \iota_m(z_m) = \iota_{m+1}(z_m) \) for each

964
There is a path from $z_j$ to $z_1$ in $\Omega_1$, a path from $z_m$ to $z_{m+1}$ in $\Omega_{m+1}$ for $1 \leq m \leq n - 2$, and a path from $z_{n-1}$ to $z_k$ in $\Omega_n$. The corresponding paths in the $\Omega_m$ match up at the endpoints and give a path in $S_f$ from $\iota_j(z_j)$ to $\iota_k(z_k)$.

X.7.2.8. EXAMPLES. (i) The Riemann surface of a total analytic function can sometimes be identified with a subset of $\mathbb{C}$. The simplest case is when $f$ is an entire function. Then $(\mathbb{C}, f)$ is an analytic function element, and the analytic function elements equivalent to it are precisely of the form $(\Omega, f|_{\Omega})$, where $\Omega$ is a region in $\mathbb{C}$. Since the functions in any two elements of the equivalence class agree on the intersection of the regions, every region in $S_f$ is identified with itself as a subset of $\mathbb{C}$. Thus $S_f \cong \mathbb{C}$.

(ii) Similarly, if $f$ is holomorphic on $\mathbb{C}$ except for isolated non-removable singularities, the set $P$ of (non-removable) singularities is a countable subset of $\mathbb{C}$ with no limit point. Then $(\mathbb{C} \setminus P, f)$ is an analytic function element, and as before every equivalent analytic function element is of the form $(\Omega, f|_{\Omega})$, where $\Omega$ is a region in $\mathbb{C} \setminus P$, and $S_f \cong \mathbb{C} \setminus P$. (If $f$ also has removable singularities, then $S_f$ is the domain of $f$ with the removable singularities removed.)

(iii) Here is one of the most important examples, of a total analytic function whose Riemann surface is not a subset of $\mathbb{C}$. If $\Omega$ is $\mathbb{C}$ with a branch cut (ray out from 0) removed (), then there are infinitely many branches of $\log z$ defined on $\Omega$. For each such $f$, $(\Omega, f)$ is an analytic function element, and the copies of $\Omega$ corresponding to different branches are disjoint. If $\bar{\Omega}$ is $\mathbb{C}$ with a different branch cut removed, then $\Omega \cap \bar{\Omega}$ is not connected, and the pieces are identified with parts of different copies of $\Omega$ in $S_f$. The Riemann surface for $\log z$ can be visualized by taking a sequence of copies of $\Omega$, and connecting each sheet along the branch cut with the sheet above, giving a helical surface (Figure X.1, from http://upload.wikimedia.org/wikipedia/commons/4/41/Riemann_surface_log.jpg). Choosing a branch cut just allows the surface to be constructed or visualized in this way; the branch cut has no intrinsic significance as part of the Riemann surface, and any other branch cut would work as well and give the same surface.
Figure X.1: Riemann Surface for $\log z$ (Three levels shown)
(iv) Similarly, the Riemann surface for $\sqrt{z}$ can be constructed by again taking a branch cut out from 0. This time, there are only two branches for each branch cut. The Riemann surface is obtained by gluing each copy of the plane with branch cut removed to the other copy along the branch cut (this cannot be physically accomplished in $\mathbb{R}^3$ since the surfaces would have to pass through each other along the branch cut, but can be done in $\mathbb{R}^4$ or $\mathbb{C}^2$; see Figure X.2, from http://en.wikipedia.org/wiki/File:Riemann_sqrt.jpg).

Figure X.2: Riemann Surface for $\sqrt{z}$

The Riemann surface for $\sqrt{z}$ is similar, with $n$ sheets, each attached to the next one along the branch cut and the last attached to the first.

(v) If $(\Omega_j, f_j)$ and $(\Omega_k, f_k)$ are branches of a total analytic function $f$, and $z \in \Omega_j \cap \Omega_k$, we may have $f_j(z) \neq f_k(z)$ even though $f_j(z) = f_k(z)$: the function with elements $(\Omega, g(z)|g(z) - 2\pi i|)$, where $\Omega$ is a simply connected open set in $\mathbb{C} \setminus \{0\}$ and $g$ is a branch of $\log z$ holomorphic on $\Omega$, has two distinct branches over neighborhoods of 1 both taking the value 0 at 1. (These branches differ at all points of a deleted neighborhood of 1.)

X.7.2.9. If $f$ is a total analytic function with Riemann surface $S_f$, there is a projection map $p_f$ from $S_f$ to $\mathbb{C}$ which sends each $\Omega$ on the surface to the corresponding subset $\Omega$ of $\mathbb{C}$. This projection is continuous and open, its range is an open subset of $\mathbb{C}$, and may be either one-to-one (e.g. in cases X.7.2.8.(i)–(ii)) or many-to-one (e.g. in cases X.7.2.8.(iii)–(iv)). If $p_f$ is one-to-one, it is a homeomorphism onto its image, and $S_f$ can be identified with an open set in $\mathbb{C}$; in this case, $f$ is said to be single-valued; if $p_f$ is not one-to-one, $f$ is said to be multivalued.

In cases X.7.2.8.(iii)–(v), the map $p_f$ is a covering map, but this is not true in general, and the inverse image of different points can even have different cardinality (each point inverse image is always finite or countable since it is relatively discrete in $S_f$). The range of $p_f$ is always a region in $\mathbb{C}$, but it may be quite small.

X.7.2.10. Any analytic function element $(\Omega, f)$ defines a unique total analytic function $f$ and hence a Riemann surface $S_f$, which is the “largest domain” $f$ can be extended to as a holomorphic function. But if $f$
is defined on a small open set \( \Omega \), e.g. on a finite disk by a power series with finite radius of convergence, it can be difficult if not impossible to determine the full Riemann surface of \( f \), or indeed whether \( f \) is single-valued or multivalued.

**X.7.3. Analytic Continuation**

**X.7.3.1.** We now change our point of view slightly, and discuss the problem from X.7.2.10. in more detail. Suppose \( f \) is a holomorphic function on a region \( \Omega \) (i.e. \( (\Omega, f) \) is an analytic function element). How far outside \( \Omega \) can \( f \) be extended as a holomorphic function, and to what degree is the extension unique?

**X.7.3.2.** The first question to discuss is: if \( z \) is a boundary point of \( \Omega \), can a (necessarily unique) holomorphic function \( g \) be defined on an open disk centered at \( z \) which agrees with \( f \) on the part of the disk contained in \( \Omega \)? If not, \( z \) is called a _singular point_ of the boundary (for \( f \)). The set of singular points is a closed subset of the boundary of \( \Omega \). In some cases, it can be shown that there must be singular boundary points, e.g.:

**X.7.3.3.** _Proposition._ Let

\[
\sum_{k=0}^{\infty} a_k (z - z_0)^k
\]

be a power series with radius of convergence \( R \), \( 0 < R < \infty \), and let \( \Omega \) be the open disk of radius \( R \) centered at \( z_0 \). If \( f \) is the holomorphic function on \( \Omega \) defined by the power series, then there is at least one singular point for \( f \) on the boundary of \( \Omega \).

**Proof:** Suppose not. Then for any \( z \) on the boundary circle there is an open disk \( D_z \) of radius \( \epsilon_z \) centered at \( z \) and a holomorphic function \( g_z \) on this disk agreeing with \( f \) on the part of the disk inside \( \Omega \). By the _Rigidity Theorem_, any two \( g_z \) must agree on the intersection of the \( D_z \). Thus there is a well-defined holomorphic function \( g \) on \( U = \Omega \cup \bigcup_z D_z \) which extends \( f \). But \( U \) is an open neighborhood of \( \bar{\Omega} \), which is compact, so \( U \) contains a disk of radius \( R + \epsilon \) centered at \( z_0 \), for some \( \epsilon > 0 \). Then by X.4.4.6. the Taylor series for \( f \) around \( z_0 \) has radius of convergence at least \( R + \epsilon \), a contradiction.

**X.7.3.4.** _Examples._ (i) There may be only one singular boundary point. The power series

\[
\sum_{k=0}^{\infty} z^k
\]

converges to \( f(z) = \frac{1}{1-z} \) on the open unit disk \( \mathbb{D} \). This function can be extended holomorphically to every boundary point of \( \mathbb{D} \) except \( z = 1 \). In fact, it extends to a function holomorphic on \( \mathbb{C} \setminus \{1\} \).

(ii) At the other extreme, all boundary points can be singular points. The power series

\[
\sum_{k=0}^{\infty} z^{2k}
\]
has radius of convergence 1, hence defines a holomorphic function on \( \mathbb{D} \). But if \( \theta \) is a dyadic rational multiple of \( 2\pi \), then it is easily checked that
\[
\lim_{r \to 1^-} |f(re^{i\theta})| = +\infty
\]
so \( e^{i\theta} \) is a singular point of the boundary. Since such points are dense in the unit circle and the set of singular points is closed, every boundary point is a singular point.

(iii) Here is an interesting variation of (ii), a holomorphic function on \( \mathbb{D} \) which can be extended continuously, even smoothly, to the boundary, but for which all boundary points are nonetheless singular. The power series
\[
\sum_{k=1}^{\infty} 2^{-k^2} z^{2^k}
\]
has radius of convergence 1 and converges uniformly on the closed unit disk \( \bar{\mathbb{D}} \), as do the derived series of all orders. Thus the function
\[
f(z) = 1 + 2z + \sum_{k=1}^{\infty} 2^{-k^2} z^{2^k}
\]
is a smooth function (in the vector calculus sense) on \( \bar{\mathbb{D}} \) which is holomorphic on \( \mathbb{D} \). But nonetheless all boundary points are singular by the Hadamard gap theorem. (It turns out that \( f \) is one-to-one on \( \mathbb{D} \), the reason for including the \( 2z \) term. Including the constant term 1 is conventional.)

Examples (ii) and (iii) are lacunary series, power series where the gaps between the indices of consecutive nonzero coefficients become arbitrarily large. There is a discussion of these and related series in [Rem98], where the comment is made that “The gap theorem is in some sense a paradox: power series that, because of their gaps, converge especially fast in the interior of their disk of convergence have singularities almost everywhere on the boundary precisely on account of these gaps.”

We now consider analytic continuation along a path.

**X.7.3.5. Definition.** Let \( \Omega \) be a region in \( \mathbb{C} \), and \( f \) a holomorphic function on \( \Omega \). Let \( \gamma \) a path in \( \mathbb{C} \) from \( z_0 \) to \( z_1 \), with \( z_0 \in \Omega \), parametrized by \( \zeta : [a, b] \to \mathbb{C} \). Then \( f \) can be analytically continued along \( \gamma \) if, for every \( t \in [a, b] \), there is an open disk \( D_t \) centered at \( \zeta(t) \) and a function \( f_t \) holomorphic on \( D_t \) such that, for every \( t_0 \in [a, b] \), \( f_t = f_{t_0} \) on \( D_t \cap D_{t_0} \) for all \( t \) sufficiently close to \( t_0 \), and such that \( f_a = f_{D_0} \).

The essential point is that the path \( \gamma \) begins in \( \Omega \) but may pass or end outside \( \Omega \). In fact this is the only interesting situation: if \( \gamma \) lies inside \( \Omega \), then \( f \) can be analytically continued along \( \gamma \) simply by restriction.

If \( \{(D_t, f_t) : t \in [a, b]\} \) is an analytic continuation of \( f \) along \( \gamma \), then each \( (D_t, f_t) \) is an analytic function element.

**X.7.3.6. Proposition.** If \( \{(D_t, f_t) : t \in [a, b]\} \) is an analytic continuation of \( f \) along \( \gamma \), all analytic function elements \( (D_t, f_t) \) are equivalent.

**Proof:** It is clear that for any \( t_0 \), the function elements \( (D_t, f_t) \) are equivalent to \( (D_{t_0}, f_{t_0}) \) for \( t \) sufficiently close to \( t_0 \). Thus the set of \( t \) in each equivalence class is open in \( [a, b] \). Since the complement of an equivalence class is the union of the other equivalence classes, the set of \( t \) in each equivalence class is also closed in \( [a, b] \). Thus there can be only one equivalence class since \( [a, b] \) is connected.

\[ \_ \]
X.7.3.7. Analytic continuation along a path \( \gamma \) can be alternately considered a finite process: if \( \{(D_t, f_t)\} \) is an analytic continuation, there is a finite partition \( \{a = t_0, t_1, \ldots, t_n = b\} \) of \([a, b]\) such that \( \zeta(t) \in D_{t_k} \cup D_{t_{k+1}} \) for \( t_k \leq t \leq t_{k+1} \), and \( f_{t_k} = f_{t_{k+1}} \) on \( D_{t_k} \cap D_{t_{k+1}} \) for \( 0 \leq k < n \). Conversely, if there is such a partition, disks \( D_k \) centered at \( \zeta(t) \) such that \( \zeta(t) \in D_{t_k} \cup D_{t_{k+1}} \) for \( t_k \leq t \leq t_{k+1} \), and functions \( f_k \) holomorphic on \( D_k \) agreeing on the consecutive overlaps, then \( (D_t, f_t) \) can be defined for each \( t \) by restricting \( f_k \) or \( f_{k+1} \) for \( t_k \leq t \leq t_{k+1} \) to a disk \( D_t \) centered at \( \zeta(t) \) contained in \( D_k \cup D_{k+1} \).

X.7.3.8. Proposition. Let \( \Omega \) be a region in \( \mathbb{C} \), and \( f \) a holomorphic function on \( \Omega \). Let \( \gamma \) a path in \( \mathbb{C} \) from \( z_0 \) to \( z_1 \), with \( z_0 \in \Omega \), parametrized by \( \zeta : [a, b] \to \mathbb{C} \). Then an analytic continuation of \( f \) along \( \gamma \), if it exists, is unique: if \( \{(D_t, f_t)\} \) and \( \{(\tilde{D}_t, g_t)\} \) are analytic continuations of \( f \) along \( \gamma \), then \( f_t = g_t \) on \( D_t \cap \tilde{D}_t \) for all \( t \).

Proof: Let \( S \) be the set of \( t \in [a, b] \) such that \( f_t = g_t \) on \( D_t \cap \tilde{D}_t \). Then \( a \in S \) since both \( f_a \) and \( g_a \) must agree with \( f \) on a neighborhood of \( z_0 \), hence everywhere on \( D_a \cap \tilde{D}_a \) by the Rigidity Theorem. \( S \) is obviously (relatively) open in \([a, b]\), again by the Rigidity Theorem. If \( t_0 \) is a limit point of \( S \), then \( f_{t_0} = g_{t_0} \) on a sequence of points in \( S \) converging to \( t_0 \), and thus \( f_{t_0} = g_{t_0} \) on all of \( D_{t_0} \cap \tilde{D}_{t_0} \) by the Rigidity Theorem, i.e. \( t_0 \in S \). Thus \( S \) is also closed. So \( S = [a, b] \) since \([a, b]\) is connected.

X.7.3.9. Thus, if \( f \) can be analytically continued along \( \gamma \), the value of \( f(\zeta(t)) \) is uniquely determined for each \( t \), and in particular a branch of \( f \) at \( z_1 \) is uniquely determined by the analytic continuation. Note, however, that the curve \( \gamma \) may cross itself, and at crossing or repeated points the value of \( f \) obtained can be different at different times. This is particularly common and important if \( \gamma \) is a closed curve. A good example is \( f(z) = \log z \), \( z_0 = z_1 = 1 \), \( \gamma \) the unit circle parametrized once around counterclockwise. This example also shows that the parametrization (or at least its equivalence class) is essential: if \( \gamma \) is parametrized once around clockwise, or more than once around, different values of \( f \) at \( z_1 \) are obtained.

X.7.3.10. Proposition. Let \( \Omega \) be a region in \( \mathbb{C} \), and \( f \) a holomorphic function on \( \Omega \). If \( \gamma \) is a (parametrized) curve in \( \mathbb{C} \) from \( z_0 \in \Omega \) to \( z_1 \in \mathbb{C} \), then \( f \) can be analytically continued along \( \gamma \) if and only if there is a (necessarily unique) curve \( \hat{\gamma} \) in \( S_f \) from the point \( z_0 \) corresponding to \( z_0 \) on the branch \( (\Omega, f) \) to another point \( \hat{z}_1 \), such that \( p(\hat{\gamma}) = \gamma \).

Proof: If there is such a \( \hat{\gamma} \), \( f \) can clearly be analytically continued along \( \gamma \). Conversely, if \( f \) can be analytically continued along \( \gamma \), there is a partition \( \{a = t_0, t_1, \ldots, t_n = b\} \) of the interval \([a, b]\) such that the disks \( D_{t_k} \) cover \( \gamma \) and \( f_{t_k} = f_{t_{k+1}} \) on \( D_{t_k} \cap D_{t_{k+1}} \) for each \( k \). The pieces of \( \gamma \) for \( t_k \leq t \leq t_{k+1} \) can be successively lifted to \( \tilde{D}_{t_k} \subset S_f \) in a unique way via the local homeomorphisms so that the entire lifted path is continuous.

X.7.3.11. As a result, if \( f \) can be analytically continued along \( \gamma \), there is a “largest” analytic continuation \( (D_t, f_t) \), where \( D_t \) is the open disk of convergence of the Taylor series for \( f_t \) around \( \zeta(t) \). We can use this analytic continuation as the standard one.
X.7.3.12. If $f$ can be analytically continued along $\gamma$, and $\gamma$ is not entirely contained inside $\Omega$, then there is a smallest $t$, say $t = t_1$, for which $\zeta(t_1) \in \partial\Omega$. Then $\zeta(t_1)$ is not a singular point of $\partial\Omega$. There is then a smallest $t_2 > t_1$ such that $\zeta(t_2) \in \partial(\Omega \cup D_{t_1})$ (unless $\gamma \subseteq \Omega \cup D_{t_1}$), and this point is not a singular point. The process can be continued.

Conversely, if $f$ cannot be analytically continued along $\gamma$, then there is a smallest $c \in [a, b]$ such that $f$ cannot be analytically continued along the restriction $\gamma_c$ of $\gamma$ to $[a, c]$. We have $c > a$ since $f$ can be analytically continued along $\gamma_d$ for $d > a$ small enough that $\gamma_d \subseteq \Omega$. Then $\zeta(c)$ can be regarded as a singular point of $f$ along $\gamma$. For example, if $z_0 \in \Omega$, $z_1 \in \partial\Omega$, and $\gamma$ is a path from $z_0$ to $z_1$ with $\zeta(t) \in \Omega$ for $a \leq t \leq b$, then $z_1$ is a singular point of $f$ along $\gamma$ if and only if $z_1$ is a singular point of $\partial\Omega$ in the sense of X.7.3.2. (There is not always such a path for every $z_1 \in \partial\Omega$ if $\partial\Omega$ is not well behaved topologically.)

X.7.3.13. If $f$ is holomorphic on $\Omega$, and $\gamma$ is a path beginning in $\Omega$ ending in a point $z_1$ outside $\Omega$ along which $f$ can be analytically continued, then there is a well-defined value for $f$ at $z_1$ (and for all $z$ in some neighborhood of $z_1$) given by the analytic continuation along $\gamma$. But if $\tilde{\gamma}$ is another path beginning in $\Omega$ and ending at $z_1$, the value of $f$ at $z_1$ defined by analytic continuation along this path need not be the same. A simple example is the principal branch of $\log z$ defined on the right half plane $\Omega$, and the upper and lower unit semicircles from 1 to $-1$. However, homotopic paths give the same result:

X.7.3.14. Theorem. [Monodromy Theorem] Let $f$ be holomorphic on a region $\Omega$, $z_0 \in \Omega$, $z_1 \in \mathbb{C}$. Let $\gamma_0$ and $\gamma_1$ be paths in $\mathbb{C}$ from $z_0$ to $z_1$, which are homotopic by a homotopy $h : [0, 1] \times [a, b]$ (X.3.4.1.). For each $s$, let $\gamma_s$ be the path parametrized by $\zeta_s$, where $\zeta_s(t) = h(s, t)$. Suppose $f$ can be analytically continued along $\gamma_s$ for every $s$. Then the values of $f$ at $z_1$ defined by analytic continuation along $\gamma_0$ and $\gamma_1$ are the same.

X.7.4. The Regular Riemann Surface of a Function

X.7.4.1. If $f$ is a total analytic function with Riemann surface $S_f$, and $f : S_f \rightarrow \mathbb{C}$ is the corresponding complete analytic function, then $f$ has a well-defined derivative at each point of $S_f$ obtained from a local function element representative, and in fact well-defined derivatives of all orders (hence a well-defined Taylor series representation) at each point.

Viewed another way, if $(\Omega_k, f_k)$ is a function element of $f$, then $(\Omega, f')$ is also an analytic function element, and the corresponding total analytic function $f'$ can be regarded as the derivative of the total analytic function $f$.

Although there is a natural map from $S_f$ to $S_{f'}$, which is surjective ($\rightarrow$), the Riemann surface $S_{f'}$ is not the same as the Riemann surface $S_f$ in general; it is only a quotient. (Consider the case of $\log z$; cf. Exercise ($\rightarrow$).) The map from $S_f$ to $S_{f'}$ is a covering map in general ($\rightarrow$).

X.7.4.2. If $f$ is a total analytic function with Riemann surface $S_f$, and $f$ is not constant, then the set of points of $S_f$ where $f'$ is zero is a closed subset $Z_f$ which is discrete in the relative topology ($\rightarrow$). Thus the complement $R_f = S_f \setminus Z_f$ is a dense open subset of $S_f$, called the regular Riemann surface of $f$.

X.7.4.3. If $\zeta \in R_f$, then there is a function element $(\Omega, g)$ of $f$ such that $\Omega$ is a neighborhood of $\zeta$ and $f$ is one-to-one on $\Omega$, i.e. $g$ is one-to-one on $\Omega$ ($\rightarrow$). Then $(g(\Omega), g^{-1})$ is an analytic function element which is a “branch” of $f^{-1}$. It is not hard to see that any two such function elements corresponding to small open sets
in \( R_f \) are equivalent, so the total analytic function defined by them can be called \( f^{-1} \). The correspondence \( \Omega \rightarrow g(\Omega) \) gives a conformal bijection between \( R_f \) and \( R_{f^{-1}} \); thus \( f \) can be regarded as a conformal bijection from \( R_f \) onto \( R_{f^{-1}} \).

**X.7.4.4.** Another way of viewing this is: if \( R_f \) and \( R_{f^{-1}} \) are identified in this way, and \( p \) and \( q \) are the natural projections from \( R_f \) and \( R_{f^{-1}} \) to \( \mathbb{C} \), then \( f \) is completely determined by the two projections \( p \) and \( q \). Thus an analytic function can be thought of as consisting of a Riemann surface and two projections to \( \mathbb{C} \).

**X.7.5. Branch Points**

**X.7.5.1.** The regular Riemann surface \( R_f \) naturally sits as a dense subset of \( S_f \), and the (isolated) points in \( Z_f = S_f \setminus R_f \) do not appear in the Riemann surface \( S_{f^{-1}} \). Conversely, there may be points in \( S_{f^{-1}} \) not corresponding to points of \( S_f \). The points of \( Z_f \) and \( Z_{f^{-1}} \) can all be added to \( R_f \cong R_{f^{-1}} \) to make a larger Riemann surface, and both \( f \) and \( f^{-1} \) can be extended in a meaningful way to these points. Such added points are called branch points.

**X.7.5.2.** Example. Here is a simple example. Let \( f(z) = \sqrt{z} \). Then the Riemann surface \( S_f = R_f \) is a two-sheeted surface which projects to \( \mathbb{C} \setminus \{0\} \) \((\text{X.7.2.8 (iv))}. The inverse function \( f^{-1}(z) = z^2 \) has \( S_{f^{-1}} = \mathbb{C} \) and \( R_{f^{-1}} = \mathbb{C} \setminus \{0\} \). The point 0 can be added to \( R_f \), and \( f \) can be defined to be 0 at this added point, so \( S_f \) can be extended by adding a branch point. There are in this case no branch points to be added to \( R_{f^{-1}} \). The roles of \( f \) and \( f^{-1} \) can of course be interchanged.

**X.7.5.3.** There is an intrinsic way to describe branch points. To motivate this, it is simplest to describe branch points of \( f^{-1} \) corresponding to points of \( Z_f \). Suppose \( \zeta_0 \in Z_f \). Set \( z_0 = p(\zeta_0) \). Then there is a small disk \( \Omega \) centered at \( z_0 \) such that the corresponding function element \((\Omega, g)\) of \( f \) has \( g'(z_0) = 0 \) and no other zeroes of \( g' \). Then \( z_0 \) is a zero of \( g \) of finite order \( n \), and \( g \) is exactly \( n \)-to-one on \( \Omega \setminus \{z_0\} \). The open set \( U \) in \( R_{f^{-1}} \) corresponding to \( \Omega \setminus \{z_0\} \) evenly covers a deleted neighborhood \( V \) of \( g(z_0) \) \( n \) times via \( g \), so there are locally \( n \) corresponding branches of \( f^{-1} \) over \( V \) which fit together like the branches of \( \sqrt{z} \) around 0. (Note that there is no reason there cannot be other branches of \( f^{-1} \) over points of \( V \) corresponding to different branches of \( f \) near \( z_0 \) than the one corresponding to \( \zeta_0 \).)

If \( C \) is a small circle in \( V \) centered at \( g(z_0) \), \( w_1 \) a point of \( C \), and \( \gamma_k \) is \( C \) parametrized \( k \) times around counterclockwise beginning and ending at \( w_1 \), and \( h \) is any of the \( n \) corresponding branches of \( g^{-1} \) in a small neighborhood of \( w_1 \), then \( h \) can be analytically continued along \( \gamma_k \) for any \( k \). For \( 1 \leq k < n \) the terminal branch of \( g^{-1} \) around \( w_1 \) differs from the initial branch chosen, but for \( k = n \) the terminal branch is the same as the initial branch. As the diameter of \( C \) approaches 0, all the values on the branches at \( w_1 \) approach \( z_0 \). So we can “unwind” \( U \), which is homeomorphic to a punctured disk, and add a center point where \( g^{-1} \) takes the value \( z_0 \).

**X.7.5.4.** In general, a branch point of \( f \) corresponds to an open set \( U \) in \( S_f \) homeomorphic to a punctured disk, such that \( p \) gives an even covering \( (\cdot) \) from \( U \) onto a punctured disk \( D \setminus \{z_0\} \) centered at a point \( z_0 \), and such that

\[
w_0 = \lim_{p(\zeta) \rightarrow z_0} f(\zeta)
\]
exists in the sense that for every $\epsilon > 0$ there is a $\delta > 0$ such that $|f(\zeta) - w_0| < \epsilon$ whenever $\zeta \in U$ and $|p(\zeta) - z_0| < \delta$ (it actually suffices that $f$ just be bounded on $U$ because of (1)). Then there is an $n \in \mathbb{N}$ such that $p|_{U}$ is exactly $n$-to-one, and the image of every simple closed curve in $U$ not homotopic to 0 has winding number $\pm n$ around $z_0$. If $n = 1$, there will be a point $\zeta_0$ in the closure of $U$ in $S_f$ such that $p(\zeta_0) = z_0$ and $p$ maps $U \cup \{\zeta_0\}$ homeomorphically onto $D$; and $f(\zeta_0) = w_0$. If $n > 1$, there is no such point on $S_f$; but we can then add a point $\zeta_0$ to $S_f$, and topologize $S_f \cup \{\zeta_0\}$ by taking sets of the form

$$\{\zeta_0\} \cup [U \cap p^{-1} \{\{z : |z - z_0| < \delta\}\}]$$

for $\delta > 0$ as a local base at $\zeta_0$. Extend $f$ to $S_f \cup \{\zeta_0\}$ by setting $f(\zeta_0) = w_0$.

**X.7.5.5.** The extended surface $S_f \cup \{\zeta_0\}$ is a topological 2-manifold with a natural structure as a complex 1-manifold, defined around $\zeta_0$ by taking a “local uniformizing variable.” Consider the map $h(s) = z_0 + s^n$ from a disk $\hat{D}$ centered at $0$ onto $D$. Fix an $s_1 \in \hat{D}$, $s_1 \neq 0$, and let $z_1 = h(s_1)$ be the corresponding point of $D$. Define a holomorphic function $G$ on $\hat{D}$ as follows. Fix a branch $g$ of $f$ on a neighborhood $\Omega$ of $z_1$ corresponding to a point $\zeta_1$ of $U$ lying above it. If $s \in \hat{D} \setminus \{0\}$, let $\gamma$ be a path from $s_1$ to $s$ in $\hat{D} \setminus \{0\}$; then $h(\gamma)$ is a path from $z_1$ to $h(s)$ within $D \setminus \{z_0\}$, and $g$ can be analytically continued along $h(\gamma)$ by (1) and X.7.3.10. Let $G(s)$ be the value of the analytic continuation at the terminal point $h(s)$. Then $G$ is well defined on $D \setminus \{0\}$ (cf. [Ahl78, p. 221-222]). We have

$$\lim_{s \to 0} G(s) = w_0$$

so $G$ has a removable singularity at 0. There is then an obvious induced homeomorphism $\phi$ from $\hat{D}$ to $U \cup \{\zeta_0\}$ sending 0 to $\zeta_0$ and $s_1$ to $\zeta_1$, for which $G = f \circ \phi$, and this homeomorphism can be used to transfer the analytic structure on $\hat{D}$ to $U \cup \{z_0\}$. This analytic structure agrees with the one already defined on $U$ as a subset of $S_f$. It is easily seen that this analytic structure does not depend on the choices made in the construction of $G$. If the branch point arises from a point of $S_{f^{-1}}$ as in X.7.5.3., this procedure gives the same analytic structure as the usual analytic structure from $S_{f^{-1}}$.

**X.7.5.6.** Example. Not all branch points of $f$ arise from $f^{-1}$. Consider $f(z) = z^{2/3}$. The Riemann surface $S_f$ has three sheets over $\mathbb{C} \setminus \{0\}$, and $f$ has a branch point of order 3 at 0. The inverse function is $f^{-1}(z) = z^{3/2}$, whose Riemann surface has two sheets over $\mathbb{C} \setminus \{0\}$ and a branch point of order 2 at 0.

**X.7.6. Points at Infinity and Meromorphic Functions**

Essentially everything done so far in this section works identically if we replace the complex plane $\mathbb{C}$ with the Riemann sphere $\hat{\mathbb{C}}$ and the words analytic and holomorphic with meromorphic.

**X.7.6.1.** Definition. A meromorphic function element is a pair $(\Omega, f)$, where $\Omega$ is a region in $\mathbb{C}$ and $f$ is a meromorphic function on $\Omega$. Two meromorphic function elements $(\Omega, f)$ and $(\Omega, \tilde{f})$ are equivalent if there is a finite chain $(\Omega_j, f_j)$ ($1 \leq j \leq n$) such that $(\Omega_1, f_1) = (\Omega, f)$, $(\Omega_n, f_n) = (\Omega, \tilde{f})$, and $\Omega_j \cap \Omega_{j+1} \neq \emptyset$ and $f_j = f_{j+1}$ on $\Omega_j \cap \Omega_{j+1}$ for $1 \leq j < n$.

A meromorphic function on $\Omega$ can be regarded as an analytic function from $\Omega$ to $\hat{\mathbb{C}}$ in the sense of (1). Any analytic function element is a meromorphic function element.
X.7.6.2. A total meromorphic function can be thought of as an equivalence class of meromorphic function elements (we will later give a slightly different definition, allowing meromorphic function elements including the point at infinity also).

If \( f = \{ (\Omega_k, f_k) \} \) is an equivalence class of meromorphic function elements, we can form a Riemann surface \( S_f \) as in the holomorphic case, where \( z_j \in \Omega_j \) is identified with \( z_k \in \Omega_k \) if \( z_j = z_k \) in \( \mathbb{C} \) and \( f_j = f_k \) on a deleted neighborhood of \( z_j = z_k \). (If we regard \( f_j \) and \( f_k \) as maps to \( \hat{\mathbb{C}} \), they then agree on a neighborhood of \( z_j = z_k \)). There is an induced function from \( S_f \) to \( \hat{\mathbb{C}} \). Thus every meromorphic function element gives a well-defined Riemann surface in this way corresponding to its equivalence class.

X.7.6.3. Note that each analytic function element \( (\Omega, f) \) defines a Riemann surface \( S_f^h \) as in X.7.2.2, when thought of as an analytic function element, and a Riemann surface \( S_f^m \) when thought of as a meromorphic function element. These are not the same in general. \( S_f^b \) is naturally a subset of \( S_f^h \); \( S_f^b \) is exactly the inverse image of \( \mathbb{C} \) in \( S_f^m \) under the induced function from \( S_f^m \) to \( \hat{\mathbb{C}} \). Thus \( S_f^m \) can be regarded as \( S_f^b \) with points added where the extended \( f \) has poles. For example, if \( f(z) = \frac{1}{z} \), then \( S_f^b \cong \mathbb{C} \setminus \{0\} \) and \( S_f^m \cong \mathbb{C} \).

X.7.6.4. We can incorporate the point at infinity more fully into this picture. A meromorphic function element \( (\Omega, f) \), where \( \Omega \) is a region in \( \mathbb{C} \) with bounded complement and \( f \) is a meromorphic function on \( \mathbb{C} \) which has a removable singularity or pole at \( \infty \) in the sense of () (this implies that \( f \) is holomorphic in the complement of a bounded set), can be regarded as a region in \( \hat{\mathbb{C}} \) containing \( \infty \) and an analytic function from this set to \( \hat{\mathbb{C}} \), and will be called a meromorphic function element at infinity.

For each meromorphic function element \( (\Omega, f) \) at infinity, for which \( 0 \notin \Omega \), define a meromorphic function element \( (\Omega, g) \), where

\[
\Omega = \{0\} \cup \left\{ \frac{1}{z} : z \in \Omega \right\} \\
g(z) = f \left( \frac{1}{z} \right) \text{ for } z \in \Omega, \quad g(0) = \lim_{z \to \infty} f(z) \in \hat{\mathbb{C}}.
\]

We incorporate \( (\Omega, g) \) into the construction of \( S_f \) by identifying \( w \in \Omega \setminus \{0\} \) with \( \frac{1}{z} \in \Omega \). This effectively adds a “point at infinity” to \( S_f \), at which \( f \) takes the value \( g(0) \). There may be many such points at infinity added corresponding to different branches of \( f \).

There is then a natural projection map \( p \) from the extended \( S_f \) to \( \hat{\mathbb{C}} \).

X.7.6.5. Branch points can then be defined and added as in the holomorphic case. A branch point corresponds to an open set \( U \) in \( S_f \) for which \( p \) is an even covering from \( U \) onto a deleted open disk \( D \setminus \{z_0\} \) in \( \hat{\mathbb{C}} \), where \( z_0 = \infty \) is allowed in which case \( D \setminus \{\infty\} \) is the complement of a closed disk in \( \mathbb{C} \) centered at \( 0 \), and

\[
w_0 = \lim_{p(\xi) \to z_0} f(\xi)
\]

exists in \( \hat{\mathbb{C}} \), i.e. \( w_0 = \infty \) is allowed.

When all points at infinity and branch points have been added, a Riemann surface \( \tilde{S}_f \) is obtained which is called the total Riemann surface of \( f \). If \( f \) is not constant, the total Riemann surfaces of \( f \) and \( f^{-1} \) are conformally equivalent ().
X.7.7. Germs of Analytic Functions

X.7.7.1. We now go to the opposite extreme from considering the Riemann surface of a total analytic function. If $z_0 \in \mathbb{C}$ is a fixed point of $\mathbb{C}$, and if $f$ is a total analytic function containing an analytic function element $(\Omega, f)$ with $z_0 \in \Omega$, then the restriction of $f$ to any neighborhood of $z_0$ completely determines $f$. Thus every function holomorphic in a neighborhood of $z_0$ determines a unique total analytic function, and two holomorphic functions agreeing on a neighborhood of $z_0$ determine the same total analytic function. (But two different holomorphic functions on small neighborhoods of $z_0$ can give different branches of the same total analytic function.)

X.7.7.2. Definition. Let $z_0 \in \mathbb{C}$. A germ of an analytic function at $z_0$ is an equivalence class of functions holomorphic in neighborhoods of $z_0$, where two such functions are equivalent if they agree on some neighborhood of $z_0$.

Thus the exact domain of a germ of an analytic function at $z_0$ is not specified or relevant.

X.7.7.3. The algebraic operations of addition, multiplication, and scalar multiplication are well defined on germs. If $f$ and $g$ are holomorphic on neighborhoods $U$ and $V$ of $z_0$, then $f + g$ is defined and holomorphic on $U \cap V$. If $\tilde{f}$ is holomorphic near $z_0$ and agrees with $f$ on $\tilde{U} \subseteq U$, and $\tilde{g}$ agrees with $g$ on $\tilde{V} \subseteq V$, then $\tilde{f} + \tilde{g}$ agrees with $f + g$ on $\tilde{U} \cap \tilde{V}$, so addition is well defined. The argument for multiplication and scalar multiplication is almost identical.

X.7.7.4. Thus the set $\mathcal{G}(z_0)$ of germs of analytic functions has a natural structure as a ring, actually an algebra over $\mathbb{C}$.

X.7.7.5. The algebra $\mathcal{G}(z_0)$ can be described in a different way. If $f$ is holomorphic in a neighborhood of $z_0$, then the Taylor series of $f$ around $z_0$ determines $f$ on at least a small disk centered at $z_0$. Conversely, any power series centered at $z_0$ with a positive radius of convergence defines a holomorphic function on a disk centered at $z_0$. Thus there is a one-one correspondence between germs of analytic functions at $z_0$ and power series centered at $z_0$ with a positive radius of convergence (we will call such a series a convergent power series centered at $z_0$). This correspondence preserves addition, multiplication, and scalar multiplication, i.e. is an algebra isomorphism.

Up to isomorphism, the algebra $\mathcal{G}(z_0)$ does not depend on $z_0$; the map $f(z) \to f(z + z_0)$ is an isomorphism of $\mathcal{G}(z_0)$ onto $\mathcal{G}(0)$, which is isomorphic to the algebra of convergent power series centered at 0. So $\mathcal{G}(0)$ can be identified with a subalgebra of the algebra $\mathbb{C}[[X]]$ of formal power series over $\mathbb{C}$.

X.7.7.6. Proposition. The algebra $\mathcal{G}(z_0)$ has no zero divisors: if $[f], [g]$ are nonzero elements of $\mathcal{G}(z_0)$, then $[f][g] \neq 0$. That is, $\mathcal{G}(z_0)$ is an integral domain.

Proof: This can be seen in at least two ways. $\mathcal{G}(z_0)$ is isomorphic to a subring of the integral domain $\mathbb{C}[[X]]$. Alternatively, if $[f]$ and $[g]$ are germs of analytic functions $f$ and $g$ respectively, neither of which are identically zero in a neighborhood of $z_0$, then $f$ and $g$ have at most an isolated zero at $z_0$, i.e. there is a deleted neighborhood $U$ of $z_0$ on which both $f$ and $g$ are never zero. Thus $fg$ is nonzero on $U$, i.e. $[fg] = [f][g] \neq 0$ in $\mathcal{G}(z_0)$.

975
Germs of Meromorphic Functions

We can extend the notion of germ to functions holomorphic on a neighborhood of \( z_0 \) except possibly for a pole at \( z_0 \).

X.7.7.7. Definition. Let \( z_0 \in \mathbb{C} \). A germ of a meromorphic function at \( z_0 \) is an equivalence class of functions holomorphic in deleted neighborhoods of \( z_0 \) with a removable singularity or pole at \( z_0 \), where two such functions are equivalent if they agree on some deleted neighborhood of \( z_0 \).

X.7.7.8. As before, the set \( M(z_0) \) of germs of meromorphic functions at \( z_0 \) is a complex algebra. There is an obvious injective homomorphism from \( G(z_0) \) to \( M(z_0) \), so we may identify \( G(z_0) \) with a subalgebra of \( M(z_0) \). We again have \( M(z_0) \cong M(0) \).

There is a one-one correspondence (algebra isomorphism) between \( M(z_0) \) and the algebra of Laurent series (\( \mathbb{C}((X)) \)) centered at \( z_0 \) which converge on a deleted neighborhood of \( z_0 \) (convergent Laurent series). Thus \( M(z_0) \) is isomorphic to a subalgebra of the algebra \( \mathbb{C}((X)) \) of formal Laurent series with complex coefficients (\( \mathbb{C} \)).

X.7.7.9. Proposition. The algebra \( M(z_0) \) is a field; in fact, it is the quotient field of the ring \( G(z_0) \).

Proof: If \( f \) is holomorphic on a deleted neighborhood of \( z_0 \) and has a pole at \( z_0 \), then \( \frac{1}{g} \) has a removable singularity at \( z_0 \), so agrees on a deleted neighborhood of \( z_0 \) with a function holomorphic in a neighborhood of \( z_0 \). Thus [\( f \)] is invertible in \( M(z_0) \), and its reciprocal is in \( G(z_0) \). And if [\( g \) \( \in G(z_0) \)], [\( g \) not the zero element (i.e. \( g \) not identically zero near \( z_0 \)), then either \( g(z_0) \neq 0 \), in which case [\( g \)] is already invertible in \( G(z_0) \), or \( g \) has an isolated zero at \( z_0 \), in which case \( \frac{1}{g} \) is holomorphic in a deleted neighborhood of \( z_0 \) and has a pole at \( z_0 \), i.e. [\( g \)] is invertible in \( M(z_0) \).

X.7.7.10. Note that this proposition cannot be proved simply by using the isomorphism with a subring of the field \( \mathbb{C}((X)) \), since it is not obvious (although it turns out to be true) that the inverse Laurent series to a convergent Laurent series is also convergent.

X.7.8. General Riemann Surfaces
X.8. Analytic Functions of Several Variables

We now describe the basic theory of complex analytic functions of several variables. Some aspects of the theory are virtually identical to the one-variable case, but there are also significant differences.

This is a large and active subject, and we only cover the most elementary aspects of it; even [BM48] gives a far more comprehensive treatment. The standard general reference for the theory of several complex variables is [GR09]. See [Gun90a]–[Gun90c] for a more modern and detailed study.

X.8.1. Elementary Theory

X.8.1.1. Definition. Let \( U \) be an open set in \( \mathbb{C}^n \). A function \( f : U \to \mathbb{C} \) is analytic if, for every \( c \in U \), \( f \) is represented in a neighborhood of \( c \) by a multivariable power series (VIII.7.1.4.) centered at \( c \). A function \( f = (f_1, \ldots, f_m) : U \to \mathbb{C}^m \) is analytic if each \( f_k \) is analytic.

X.8.1.2. As in the one-variable case, this definition is a local one: a function \( f \) on an open set \( U \) is analytic on \( U \) if and only if it is analytic on a neighborhood of \( c \) for each \( c \in U \). In particular, if \( U = \bigcup_j U_j \) is a union of open sets, then \( f \) is analytic on \( U \) if and only if it is analytic on each \( U_j \).

An immediate corollary of VIII.7.1.13. is:

X.8.1.3. Corollary. Let \( U \) be an open set in \( \mathbb{C}^n \), and \( f \) analytic on \( U \). Then \( f \) is \( C^\infty \) on \( U \), and for any multi-index \( \alpha \), \( \frac{\partial^n f}{\partial z^\alpha} \) is analytic on \( U \).

X.8.1.4. An analytic function \( f \) is “jointly analytic” in its variables. If all but one variable is fixed, it is analytic in the remaining variable in the usual sense of (), since the multivariable power series for \( f \) around any point has a drastic revision obtained by collecting together all terms of like power in the remaining variable, which will converge in a neighborhood of the point. (More generally, if some of the variables are fixed, the function is analytic in the remaining variables.) Thus an analytic function is “separately analytic” in its variables. Conversely, one of the major theorems of the subject is that a separately analytic function is (jointly) analytic:

X.8.1.5. Theorem. [Hartogs’ Theorem] Let \( U \) be an open set in \( \mathbb{C}^n \), and \( f : U \to \mathbb{C} \). Suppose \( f \) is analytic (holomorphic) on \( U \) in each variable separately when the other variables are held fixed, i.e. all first-order partials of \( f \) (in the complex sense) exist everywhere on \( U \). Then \( f \) is analytic on \( U \).

We will prove the theorem only under the additional assumption that \( f \) is continuous (which is of course a consequence of the theorem); for the considerably more difficult proof without this assumption, see [BM48].

X.8.1.6. Theorem. [Osgood’s Lemma] Let \( U \) be an open set in \( \mathbb{C}^n \), and \( f : U \to \mathbb{C} \). Suppose \( f \) is continuous on \( U \), and analytic (holomorphic) on \( U \) in each variable separately when the other variables are held fixed, i.e. all first-order partials of \( f \) (in the complex sense) exist everywhere on \( U \). Then \( f \) is analytic on \( U \).
Proof: The proof is a multidimensional version of the proof of X.4.4.6. Let \( c = (c_1, \ldots, c_n) \in U \), and let \( R = \{(z_1, \ldots, z_n) : |z_k - c_k| \leq r_k\} \) be a closed polydisk around \( c \) contained in \( U \). Fix \((z_1, \ldots, z_n)\) in the interior \( R^\circ \) of \( R \), i.e. \(|z_k - c_k| < r_k\) for all \( k \). We have, by repeated application of the Cauchy Integral Formula

\[
\frac{1}{2\pi i} \int_{\gamma_n} \frac{f(z_1, \ldots, z_{n-1}, \zeta_n) \zeta_n - z_n}{\zeta_n - z_n} \, d\zeta_n
\]

\[
= \frac{1}{2\pi i} \int_{\gamma_{n-1}} \left[ \frac{1}{2\pi i} \int_{\gamma_n} f(z_1, \ldots, z_{n-2}, \zeta_{n-1}, \zeta_n) \zeta_n - z_n \, d\zeta_n \right] \frac{1}{\zeta_{n-1} - z_{n-1}} \, d\zeta_{n-1}
\]

\[
= \cdots = \frac{1}{(2\pi i)^n} \int_{\gamma_1} \cdots \int_{\gamma_n} f(\zeta_1, \ldots, \zeta_n) \prod_{k=1}^n \frac{1}{(\zeta_k - z_k) \cdots (\zeta_n - z_n)} \, d\zeta_1 \cdots d\zeta_n
\]

where \( \gamma_k \) is the circle \( \{z_k : |z_k - c_k| = r_k\} \) parametrized once around counterclockwise.

By parametrizing the integrals, \( f(z_1, \ldots, z_n) \) becomes an iterated integral of a complex-valued continuous function over a rectangle in \( \mathbb{R}^n \). By Fubini’s Theorem, this integral can be converted to a multiple integral (or computed as an iterated integral in any order). We will symbolically denote this multiple integral by

\[
f(z_1, \ldots, z_n) = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)} \, d\zeta.
\]

For each \( k \) set \( b_k = \frac{|z_k - c_k|}{r_k} < 1 \). We have, for \((\zeta_1, \ldots, \zeta_n) \in \partial R\),

\[
\frac{1}{\zeta_k - z_k} = \frac{1}{\zeta_k - c_k} \frac{1 - \frac{z_k - c_k}{\zeta_k - c_k}}{1 - \frac{z_k - c_k}{\zeta_k - c_k}} = \frac{1}{\zeta_k - c_k} \sum_{m=0}^{\infty} \left[ \frac{z_k - c_k}{\zeta_k - c_k} \right]^m
\]

and the terms in the series are dominated in absolute value by the terms of \( \sum_{m=0}^{\infty} b_k^m \), so the series converges absolutely and uniformly on \( \partial R \). Iterating, the iterated sum

\[
\sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \cdots \sum_{m_n=0}^{\infty} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}
\]

converges absolutely and uniformly on \( \partial R \) to \( \frac{1}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)} \). Thus the unordered sum

\[
\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}
\]

converges uniformly on \( \partial R \) to \( \frac{1}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)} \). So we have

\[
f(z_1, \ldots, z_n) = \frac{1}{(2\pi i)^n} \int_{\partial R} f(\zeta_1, \ldots, \zeta_n) \left[ \sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \right] \, d\zeta
\]

\[
= \frac{1}{(2\pi i)^n} \int_{\partial R} \left[ \sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \right] \, d\zeta.
\]

978
Since $f$ is bounded on $R$, by comparison the unordered sum

$$
\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}
$$

converges uniformly on $\partial R$, and hence the summation can be moved past the integral ($\int$). Thus

$$
f(z_1, \ldots, z_n) = \sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \left[ \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} d\zeta \right]
$$

where

$$
a_{m_1, \ldots, m_n} = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} d\zeta
$$

(where the last integral can be interpreted as an iterated contour integral in any order). Thus, since $(z_1, \ldots, z_n)$ is arbitrary in $R^o$, the multivariable power series

$$
\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} a_{m_1, \ldots, m_n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}
$$

converges to $f$ on $R^o$, a neighborhood of $c$. Since $c \in U$ is arbitrary, $f$ is analytic on $U$. 

\[ \Box \]

**X.8.1.7.** Corollary. Let $U$ be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$ a function. Then $f$ is analytic on $U$ if and only if $f$, when regarded as a function $f(x_1, y_1, \ldots, x_n, y_n)$ from $\mathbb{R}^{2n}$ to $\mathbb{R}^2$ with coordinate functions $u$ and $v$, is differentiable in the vector calculus sense on $U$ and the Cauchy-Riemann equations hold for each pair $(x_k, y_k)$, i.e. for all $k$,

$$
\frac{\partial u}{\partial x_k} = \frac{\partial v}{\partial y_k} \quad \text{and} \quad \frac{\partial u}{\partial y_k} = -\frac{\partial v}{\partial x_k}.
$$

**Proof:** If $f$ is analytic, it is differentiable in the vector calculus sense and separately analytic, so by X.1.1.5. the Cauchy-Riemann equations are satisfied. Conversely, if it is differentiable, hence continuous, and the Cauchy-Riemann equations are satisfied, apply X.1.1.5. to conclude that $f$ is separately analytic. 

\[ \Box \]

**X.8.1.8.** Using the Looman-Menchoff Theorem (X.1.1.6.), the assumption that $f$ be differentiable in the vector calculus sense can be relaxed to just assuming $f$ is continuous, and that the first-order partial derivatives of $u$ and $v$ (in the vector calculus sense) all exist everywhere on $U$ and satisfy the Cauchy-Riemann equations. The assumptions can be further relaxed using Hartogs’ Theorem, but they cannot be relaxed to simply assuming the partials exist everywhere and satisfy the Cauchy-Riemann equations (this does not imply analyticity even when $n = 1$; cf. X.8.2.1.).

As a corollary of the proof, we obtain a multidimensional version of the Cauchy Integral Formula and of X.4.4.6.:
X.8.1.9. **Corollary.** Let $U$ be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$ an analytic function. For any $c = (c_1, \ldots, c_n) \in U$, we have:

(i) If $R$ is any closed polydisk in $U$ containing $c$ in its interior, we have

$$f(c) = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1) \cdots (\zeta_n - c_n)} \, d\zeta$$

where the multiple integral can be regarded as an iterated contour integral in any order.

(ii) The multivariable power series

$$\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} a_{m_1, \ldots, m_n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}$$

in standard form representing $f$ in a neighborhood of $c$ is unique and converges u.c. to $f$ on any open polydisk centered at $c$ contained in $U$; we have, for any $(m_1, \ldots, m_n) \in \mathbb{N}_0^n$,

$$\frac{\partial^{m_1 + \cdots + m_n}}{\partial z_1^{m_1} \cdots \partial z_n^{m_n}} f(c) = (m_1!) \cdots (m_n!) a_{m_1, \ldots, m_n} = \frac{(m_1!) \cdots (m_n!)}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \, d\zeta$$

for any closed polydisk $R$ contained in $U$ and containing $c$ in its interior. Thus

$$a_{m_1, \ldots, m_n} = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \, d\zeta$$

where the multiple integral can be regarded as an iterated contour integral in any order.
X.8.2. Exercises

X.8.2.1. (a) (H. Looman) For \( z \in \mathbb{C} \), define \( f(z) = e^{-1/z^4} \) if \( z \neq 0 \), and set \( f(0) = 0 \). Show that the partial derivatives of the real and imaginary parts of \( f \) exist and satisfy the Cauchy-Riemann equations everywhere on \( \mathbb{R}^2 \), but \( f \) is not complex differentiable, or even continuous, at 0.

(b) (D. Menchoff) Set \( f(z) = \frac{z^5}{|z|^4} \) for \( z \neq 0 \), and \( f(0) = 0 \). Show that \( f \) is continuous and has first-order partials everywhere on \( \mathbb{C} \), and the first-order partials at 0 satisfy the Cauchy-Riemann equations, but \( f \) is not complex differentiable at 0.

X.8.2.2. Let \( D \) be the open unit disk in \( \mathbb{C} \).

(a) Show that the power series \( \sum_{k=1}^{\infty} \frac{z^k}{k^2} \) converges uniformly on \( D \).

(b) Show that the series \( \sum_{k=1}^{\infty} \frac{z^{k-1}}{k} \) of term-by-term derivatives converges u.c. but not uniformly on \( D \).

X.8.2.3. (Another proof of the Fundamental Theorem of Algebra, the “standard” one) Let \( f \) be a polynomial with complex coefficients and no root in \( \mathbb{C} \), and let \( g = \frac{1}{f} \). Then \( g \) is an entire function.

(a) Use Lemma X.1.4.1. to show that \( g \) is bounded.

(b) Apply Liouville’s Theorem () to conclude that \( g \), and hence \( f \), is constant.

X.8.2.4. (Yet another proof of the Fundamental Theorem of Algebra [Boa64]) Suppose \( f \) is a polynomial of degree \( n \geq 1 \) with no root in \( \mathbb{C} \).

(a) Set \( p(z) = f(z)\bar{f}(z) \), where \( \bar{f} \) is the polynomial obtained from \( f \) by replacing each coefficient by its complex conjugate. Then \( p \) is a polynomial of degree \( 2n \) with no root in \( \mathbb{C} \), taking strictly positive values on the real axis.

(b) Set \( I = \int_{-\pi}^{\pi} \frac{1}{p(2\cos \theta)} \, d\theta \). Then \( I > 0 \).

(c) Show that \( I = -i \int_{\gamma} \frac{1}{\zeta p(\zeta + \bar{\zeta})^{-1}} \, d\zeta \), where \( \gamma \) is the unit circle.

(d) Show that \( p(z + \zeta^{-1}) = z^{-2n} q(z) \) for a polynomial \( q \) with no roots in \( \mathbb{C} \).

(e) Conclude from Cauchy’s Theorem that \( I = -i \int_{\gamma} \frac{\zeta^{2n-1}}{q(\zeta)} \, d\zeta = 0 \), contradicting (b).

X.8.2.5. Prove the following complex analog of Rolle’s Theorem for polynomials:

**Theorem.** Let \( f(z) \) be a polynomial with complex coefficients. Then the roots of \( f' \) are contained in the (necessarily closed) convex hull \( K \) of the roots of \( f \).

(a) Since \( K \) is an intersection of closed half planes, it suffices to show that if \( H \) is a closed half plane containing all the roots of \( f \), then \( H \) also contains all the roots of \( f' \). Show that it suffices to prove this if \( H \) is the closed upper half plane.

(b) We may assume \( f \) is monic without loss of generality. Use the Fundamental Theorem of Algebra to write

\[
 f(z) = (z - z_1)^{m_1}(z - z_2)^{m_2} \cdots (z - z_n)^{m_n}
\]
where $z_1, \ldots, z_n \in \mathbb{C}$ and $m_1, \ldots, m_n \in \mathbb{N}$.

(c) Use either the Product Rule or logarithmic differentiation to show that

$$\frac{f'(z)}{f(z)} = \frac{m_1}{z - z_1} + \frac{m_2}{z - z_2} + \cdots + \frac{m_n}{z - z_n}$$

for $z \in \mathbb{C} \setminus \{z_1, \ldots, z_n\}$.

(d) Under the assumption that $\text{Im}(z_k) \geq 0$ for all $k$, show that if $\text{Im}(z) < 0$, then $\text{Im} \left( \frac{f'(z)}{f(z)} \right) > 0$ and hence $f'(z) \neq 0$.

(e) Explain why the theorem is a complex analog of Rolle’s Theorem.

**X.8.2.6.** The aim of this problem is to further justify the definition of the complex exponential function (X.1.1.8.(iii)). Suppose $f$ is a holomorphic function on a horizontal strip $\Omega \subset \mathbb{C}$ with $\mathbb{R} \subset \Omega$, which satisfies $f' = f$ on $\Omega$ and $f(0) = 1$. Let $u$ and $v$ be the real and imaginary parts of $f$, i.e. $f(x + iy) = u(x, y) + iv(x, y)$. From the equation $f' = f$ it follows that $u$ and $v$ are $C^\infty$ on $\Omega$.

(a) Using $f' = f$, show that $\frac{\partial u}{\partial x} = u$, and for fixed $y$ use the Uniqueness Theorem for first-order linear differential equations (IX.1.2.7.) to conclude that $u(x, y) = e^x \phi(y)$ for some real-valued $C^\infty$ function $\phi$ defined in a neighborhood of 0 in $\mathbb{R}$. Similarly, $v(x, y) = e^x \psi(y)$ for some $C^\infty$ $\psi$.

(b) Using the Cauchy-Riemann equations, show that $\phi' = -\psi$ and $\psi' = \phi$.

(c) Use $f(0) = 1$ and the Uniqueness Theorem for solutions to systems of first-order linear differential equations (IX.1.3.3.) to conclude that $\phi(y) = \cos y$ and $\psi(y) = \sin y$ for all $y$. Thus $f$ agrees with the complex exponential function on $\Omega$.

(d) Similarly characterize the complex trigonometric and hyperbolic functions.

In fact, by the Rigidity Theorem, any function on such an $\Omega$ which is holomorphic and agrees with the usual exponential function on the real axis must agree with the complex exponential function everywhere on $\Omega$, and similarly for trigonometric and hyperbolic functions.

**X.8.2.7.** Discuss the Riemann surface for the function $\log \log z$.

**X.8.2.8.** Discuss the validity of the following formulas, where all letters denote complex numbers, and interpret them in terms of Riemann surfaces:

(a) $\log(zw) = \log z + \log w$.

(b) $\log(z^c) = c \log z$.

(c) $ze^{z+d} = ze^{z+d}$.

(d) $(ze^d)^c = ze^{cd}$.

(e) $(zw)^c = ze^{cd}$.

**X.8.2.9.** Develop the argument in the last part of the proof of X.3.4.4. to show that if $\gamma_0$ and $\gamma_1$ are piecewise-smooth contours from $p$ to $q$ which are homotopic in an open set $\Omega$, then they are homotopic in $\Omega$ via a piecewise-smooth homotopy. [Choose partitions $\mathcal{P}$ and $\mathcal{Q}$ as in the proof; for $1 < j < m$ let $H(s_j, t)$ be piecewise-linear in $t$ with $H(s_j, t_k) = h(s_j, t_k)$ for each $k$, and then extend $H$ by making it piecewise-linear in $s$ for fixed $t$.]
X.8.2.10. (a) Let $\Omega$ be a connected open set in $\mathbb{C}$ containing an interval $I$ on the real axis. Show that the complex exponential function is the only function which is holomorphic on $\Omega$ and which agrees with the usual real exponential function on $I$. [Use the Rigidity Theorem (X.5.1.4.).] Do the same for the sine and cosine functions.

(b) Show that the power series
$$\sum_{k=0}^{\infty} \frac{z^k}{k!}$$
converges u.c. on all of $\mathbb{C}$ to the function $e^z$. Similarly, show that the series expansions
$$\cos z = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k}$$
$$\sin z = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1}$$
are valid for all $z \in \mathbb{C}$. Give similar series representations for $\cosh z$ and $\sinh z$.

Thus the usual power series representations of the real exponential, trigonometric, and hyperbolic functions are valid also for the complex ones. This problem provides additional justification for why the extensions we gave of the exponential function, etc., to $\mathbb{C}$ are reasonable (and indeed are the only reasonable ones.) In fact, in many references these functions are defined using the power series.

X.8.2.11. [?] Here is an example of an improper real integral which can be computed by a more complicated contour integration. Consider the integral
$$\int_0^\infty \frac{x^{-k}}{1 + x} \, dx$$
(where $k$ is a real constant and $x^{-k}$ is interpreted in the usual real sense, i.e. as a positive real number for $x > 0$).

(a) Show that this improper integral converges if and only if $0 < k < 1$.

We make this restriction in the rest of the problem. Define
$$f(z) = \frac{z^{-k}}{1 + z}$$
where $z^{-k}$ is defined using a branch cut along the positive real axis and taking the argument between 0 and $2\pi$. Thus the argument of $z^{-k}$ lies between $-2\pi k$ and 0.

We also consider the contour $\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$:
where $\gamma_1$ and $\gamma_3$ are line segments along the ray of points with argument $\epsilon$ and $2\pi - \epsilon$ respectively, and $\gamma_2$ and $\gamma_4$ are circle arcs of radius $R$ and $r$ respectively.

(b) Show that if $0 < r < 1$, $R > 1$, and $0 < \epsilon < \pi$ we have
$$\int_\gamma f = 2\pi i \text{Res}(f, -1) = 2\pi i e^{-\pi i k}.$$
(c) Show that
\[ \int_{\gamma_2} f \to 0 \]
as \( R \to +\infty \) and
\[ \int_{\gamma_4} f \to 0 \]
as \( r \to 0 \) uniformly in \( \epsilon \).
(d) Show that, for fixed \( r \) and \( R \), as \( \epsilon \to 0+ \) we have
\[ \int_{\gamma_1} f \to \int_r^R \frac{x^{-k}}{1 + x} \, dx \]
and that
\[ \int_{\gamma_3} f \to -e^{-2\pi ik} \int_r^R \frac{x^{-k}}{1 + x} \, dx \]
[The integrals do not approach the same limit because of the branch cut.]
(e) Conclude that
\[ \int_0^\infty \frac{x^{-k}}{1 + x} \, dx = \frac{2\pi i e^{-\pi i k}}{1 - e^{-2\pi i k}} . \]
(f) Show that
\[ 1 - e^{-2\pi ik} = 2ie^{-\pi i k} \sin \pi k \]
so that
\[ \int_0^\infty \frac{x^{-k}}{1 + x} \, dx = \frac{\pi}{\sin \pi k} . \]

**X.8.2.12.** Suppose \( f \) has a singularity at \( z_0 \), and that \( f \) is bounded in a deleted neighborhood of \( z_0 \). Mimic the proof of (iv) \( \Rightarrow \) (i) in **X.5.2.7.** to show that the singularity is removable, by showing directly that for fixed \( z \) near \( z_0 \),
\[ \lim_{\delta \to 0^+} \left[ \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z} \, d\zeta \right] = 0 \]
by estimating the size of the integral.

**X.8.2.13.** (a) Show that the implication (ii) \( \Rightarrow \) (i) in **X.5.2.7.** is equivalent to the following:
**Theorem.** Let \( D \) be an open disk in \( \mathbb{C} \) centered at \( z_0 \), and \( f : D \to \mathbb{C} \). If \( f \) is continuous on \( D \) and holomorphic on \( D \setminus \{z_0\} \), then \( f \) is holomorphic on \( D \) (i.e. differentiable at \( z_0 \)).

(b) Prove this theorem by showing that the integral of \( f \) around any triangle in \( D \) is zero and applying Morera’s Theorem.

**X.8.2.14.** Using the Big Picard Theorem (**X.5.2.19.**) and **X.5.3.5.**, prove the Little Picard Theorem (**X.3.3.12.**). Do the case of a polynomial separately using the Fundamental Theorem of Algebra.
X.8.2.15. For \( z \in \mathbb{C} \) set \( f(z) = e^{e^z} (= e^{(e^z)}) \). Then \( f \) is an entire function. What is wrong with the following argument? Since \( e^z \) is never 0, \( f \) never takes either the value 0 or the value 1. Thus \( f \) is a counterexample to the Little Picard Theorem (X.3.3.12.).
X.9. Differential Algebra

Differential algebra is, roughly speaking, the algebraic theory of calculus. Versions appear in quite different contexts. A basic application concerns solutions of differential equations, in particular answering questions about whether antiderivatives or other solutions to differential equations have elementary closed formulas. A quite different context is the study of one-parameter automorphism groups of operator algebras, and in forms of amenability.

X.9.1. Derivations

A derivation on a ring is an algebraic operation which abstracts the essential properties of the derivative, specifically the product rule.

X.9.1.1. Definition. Let $R$ be a ring. A derivation on $R$ is a function $\delta : R \to R$ with the properties

$$\delta(x + y) = \delta(x) + \delta(y)$$

$$\delta(xy) = \delta(x)y + x\delta(y)$$

for all $x, y \in R$. If $R$ is an algebra over a field $\mathbb{F}$, we often (but not always) expand the first condition to require that $\delta$ be $\mathbb{F}$-linear, i.e. we add the requirement

$$\delta(\alpha x) = \alpha \delta(x)$$

for all $\alpha \in \mathbb{F}, x \in R$ (cf. X.9.1.5.(iii)).

A differential subring of a differential ring $(R, \delta)$ is a subring $S$ which is closed under $\delta$; the restriction of $\delta$ to $S$ is thus a derivation on $S$.

X.9.1.2. Note that $R$ need not be commutative; if $R$ is not commutative, it is crucial to respect the order of the factors in the product rule terms (since addition is always commutative, the order of the terms does not matter).

Derivations can be defined more generally from a ring $R$ to an $R$-bimodule $M$ with the same properties; but we will primarily stick to the case of derivations from $R$ to $R$, which are general enough for most purposes.

X.9.1.3. Examples. (i) Let $I$ be an interval in $\mathbb{R}$, and $R$ be a ring of real-valued differentiable functions on $I$ which is closed under differentiation, e.g. $C^\infty(I)$ or the set of polynomials, and set $\delta(f) = f'$. Then $\delta$ is a derivation on $R$. This is the prototype example.

$R$ must be closed under differentiation: a ring such as $C^1(I)$ will not work, since the derivative of a $C^1$ function is not in $C^1$ in general. We could more generally define differentiation as a derivation from $C^p(I)$ to the $C^{p-1}(I)$-bimodule $C^{p-1}(I)$ for any $p \geq 1$.

(ii) Similarly, if $U$ is an open set in $\mathbb{R}^2$ and $R = C^\infty(U)$, set $\delta_x(f) = \frac{\partial f}{\partial x}$ and $\delta_y(f) = \frac{\partial f}{\partial y}$. Then $\delta_x$ and $\delta_y$ are derivations on $R$.

(iii): Let $\mathbb{F}$ be a unital commutative ring, usually a field, and let $X, X_1, X_2, \ldots$ be a sequence of indeterminates. Form the polynomial ring

$$R = \mathbb{F}[X, X_1, X_2, \ldots].$$
We will regard $X_n$ as the “$n$’th derivative” of $X$ (we may think of $X$ as $X_0$). Specifically, we define a derivation $\delta$ on $R$ with $\delta(X_k) = X_{k+1}$ for all $k \geq 0$. Linearity and the product rule then dictates the following formula for $\delta$:

$$
\delta \left( \sum_{k,n} \alpha_{k,n} X_k^n \right) = \sum_{k,n} n \alpha_{k,n} X_k^{n-1} X_{k+1}
$$

(the sum is a finite sum). It is easily checked that $\delta$ is a derivation on $R$. The ring $R$ with this derivation is called the ring of differential polynomials in $X$, over $F$, denoted $\mathbb{F}\{X\}$. We may similarly form rings of differential polynomials in more than one variable, e.g. $\mathbb{F}\{X,Y\}$.

(iv) Here is an important class of derivations of interest only on noncommutative rings. Let $R$ be a ring, and $a \in R$. Set

$$
\delta_a(x) = [a,x] = ax - xa
$$

Then $\delta_a$ is a derivation on $R$, called the inner derivation defined by $a$. It is the zero derivation if and only if $a$ is in the center of $R$ (e.g. if $R$ is commutative).

More generally, if $M$ is an $R$-bimodule and $a \in M$, then $\delta_a$ is a derivation from $R$ to $M$. Amenability is characterized by certain types of derivations always being inner (i).

(v) We illustrate the connection between inner derivations and derivatives with the following example, which is a (very) special case of the theory to be developed in (i). Let $\mathcal{H}$ be a (complex) Hilbert space, and let $H$ be a self-adjoint operator in $\mathcal{B}(\mathcal{H})$. For each $t \in \mathbb{R}$, let $U_t = e^{itH}$. Then $U_t$ is a unitary operator on $\mathcal{H}$, $U_{s+t} = U_s U_t$ for all $s,t$, and $t \mapsto U_t$ is norm-continuous. Consider the inner derivation $\delta = \delta_H$ on $\mathcal{B}(\mathcal{H})$, i.e. $\delta(X) = HX - XH$. Then $\delta$ is a derivation on $\mathcal{B}(\mathcal{H})$, and is also a bounded operator on $\mathcal{B}(\mathcal{H})$, i.e. $\delta \in \mathcal{B}(\mathcal{B}(\mathcal{H}))$. Thus we may form the operator

$$
\alpha_t = e^{i\delta} \in \mathcal{B}(\mathcal{B}(\mathcal{H}))
$$

for any $t \in \mathbb{R}$. Then $\alpha_t$ is a *-automorphism of $\mathcal{B}(\mathcal{H})$ for each $t$, $\alpha_{s+t} = \alpha_s \circ \alpha_t$ for all $s,t$, and $t \mapsto \alpha_t$ is norm-continuous ($\{\alpha_t\}$ is a norm-continuous one-parameter group of automorphisms of $\mathcal{B}(\mathcal{H})$). We have

$$
\alpha_t(X) = U_t UX_t^*
$$

$$
H = \lim_{t \to 0} \frac{U_t - I}{t}
$$

$$
\delta(X) = \lim_{t \to 0} \frac{\alpha_t(X) - X}{t}
$$

for all $X \in \mathcal{B}(\mathcal{H})$. Every norm-continuous one-parameter automorphism group of $\mathcal{B}(\mathcal{H})$ is of this form (i), i.e. every bounded *-derivation on $\mathcal{B}(\mathcal{H})$ is inner.

The formal relation between derivations and one-parameter automorphism groups is quite simple (if convergence questions are ignored): if $(\alpha_t)$ is a one-parameter automorphism group and we define

$$
\delta(X) = \lim_{t \to 0} \frac{\alpha_t(X) - X}{t}
$$

(assuming the limit exists for all $X$), then for any $X$ and $Y$ we have

$$
\delta(XY) = \lim_{t \to 0} \frac{\alpha_t(XY) - I}{t} = \lim_{t \to 0} \frac{\alpha_t(X)\alpha_t(Y) - X\alpha_t(Y) + X\alpha_t(Y) - I}{t}
$$
so $\delta$ is a derivation. Conversely, if $\delta$ is a derivation, a formal power-series calculation (assuming the appropriate series converge) shows that $e^\delta$ is an automorphism. Thus derivations are "derivatives" of one-parameter automorphism groups.

One-parameter groups of automorphisms of $B(\mathcal{H})$ which are only point-norm continuous (strongly continuous) are more important in applications, e.g. in the time evolution of a dynamical system, and have a similar description except that the $\mathcal{H}$ is unbounded and only densely defined in general, as is the derivation $\delta$.

Constants

**X.9.1.4.** Let $\delta$ be a derivation on a ring $R$. An element $x \in R$ is a constant (for $\delta$) if $\delta(x) = 0$. Write $R_0$ for the set of constants for $\delta$.

**X.9.1.5.** **PROPOSITION.** Let $\delta$ be a derivation on a ring $R$, and $R_0$ the set of constants for $\delta$. Then

(i) $R_0$ is a subring of $R$.

(ii) If $R$ is a unital ring, then $1 \in R_0$, i.e. $R_0$ is a unital subring of $R$.

(iii) If $K$ is a subring of $R_0$ which is contained in the center of $R$, and $R$ is regarded as an algebra over $K$, then $\delta$ is $K$-linear, i.e.

$$\delta(\alpha x) = \alpha \delta(x)$$

for all $\alpha \in K$, $x \in R$.

**PROOF:** Parts (i) and (iii) are obvious. For (ii), we have

$$\delta(1) = \delta(1 \cdot 1) = \delta(1)1 + 1\delta(1) = 2\delta(1)$$

so $\delta(1) = 0$.

**X.9.1.6.** If $\delta$ is a derivation on a ring $R$, and $\alpha \in R_0$ is in the center of $R$, then $\alpha \delta$ is also a derivation on $R$, where $(\alpha \delta)(x) = \delta(\alpha x)$:

$$\alpha \delta(xy) = \delta(\alpha xy) = \delta(\alpha) \delta(xy) + \alpha \delta(xy) = \alpha \delta(x)y + \alpha x \delta(y) = \alpha \delta(x)y + x \alpha \delta(y).$$

**X.9.1.7.** **EXAMPLES.** (i) In **X.9.1.3.(i)**, the constants are exactly the constant functions.

(ii) In **X.9.1.3.(ii)**, if $U$ is connected, the constants for $\delta_x$ are just the functions whose value depends only on $y$. But if $U$ is not connected, the constants for the derivation are a little more complicated (functions of only $y$ on each component of $U$).

(iii) In **X.9.1.3.(iii)**, the constants are just the elements of $F$.

(iv) In **X.9.1.3.(iv)**, the constants for the inner derivation $\delta_a$ are exactly the elements of $R$ which commute with $a$.

(v) The ring of constants is not necessarily commutative. In the extreme case, if $\delta$ is the zero derivation on any ring $R$, then $R_0 = R$. 988
X.9.2. Differential Fields

In the subject of differential algebra, derivations on fields are usually considered. In this subsection, all rings will be commutative.

X.9.2.1. Definition. A differential field is a pair \((F, \delta)\), where \(F\) is a field of characteristic zero (III.4.3.2.) and \(\delta\) a derivation on \(F\). (In some references the characteristic zero requirement is omitted.)

The restriction to fields is rather natural, due to the following result:

X.9.2.2. Theorem. Let \(R\) be a unital commutative ring, and \(D\) a subset of \(R\) closed under multiplication, \(1 \in D\), which contains neither 0 nor any zero divisors. If \(\delta\) is a derivation on \(R\), then \(\delta\) extends uniquely to a derivation \(\tilde{\delta}\) on the ring \(D^{-1}R\) of quotients (III.4.3.4.). The extension is given by the “quotient rule”

\[
\tilde{\delta}\left(\frac{x}{y}\right) = \frac{y\delta(x) - x\delta(y)}{y^2}
\]

for all \(x \in R, y \in D\). In particular, if \(R\) is an integral domain and \(\delta\) is a derivation on \(R\), then \(\delta\) extends uniquely to the quotient field of \(R\). If \(x \in R\) and \(y \in D\) are constants for \(\delta\), then \(xy\) is a constant for \(\tilde{\delta}\).

Proof: Recall that elements of \(D^{-1}R\) can be symbolically written as \(\frac{x}{y}\) with \(x \in R, y \in D\), with \(\frac{x}{y} = \frac{z}{w}\) if and only if \(xw = yz\). (We should really write \([\frac{x}{y}]\) or \([x, y]\) instead of \(\frac{x}{y}\).)

To motivate the definition of \(\tilde{\delta}\) and prove uniqueness, suppose \(\tilde{\delta}\) is a derivation on \(D^{-1}R\) extending \(\delta\). For each \(y \in D\), we have, by X.9.1.5.(ii),

\[
0 = \tilde{\delta}(1) = \tilde{\delta}\left(\frac{1}{y}\right) = y\tilde{\delta}\left(\frac{1}{y}\right) = y\tilde{\delta}\left(\frac{1}{y}\right) + \frac{1}{y}\delta(y) = \frac{y\delta(x) - x\delta(y)}{y^2}.
\]

and if \(x \in R\), we have

\[
\tilde{\delta}\left(\frac{x}{y}\right) = \tilde{\delta}\left(\frac{1}{y}\right) = \frac{1}{y}\delta(x) + x\tilde{\delta}\left(\frac{1}{y}\right) = \frac{1}{y}\delta(x) - \frac{x}{y^2}\delta(y) = \frac{y\delta(x) - x\delta(y)}{y^2}.
\]

We now show that \(\tilde{\delta}\) is well defined. We first suppose \(x \in R\) and \(y, z \in D\). Then

\[
\tilde{\delta}\left(\frac{xz}{yz}\right) = \frac{yz\delta(xz) - xz\delta(yz)}{y^2z^2} = \frac{yz(x\delta(z) + z\delta(x)) - xz(y\delta(z) + z\delta(y))}{y^2z^2} = \frac{yz^2\delta(x) - xz^2\delta(y)}{y^2z^2} = \frac{y\delta(x) - x\delta(y)}{y^2} = \tilde{\delta}\left(\frac{x}{y}\right).
\]

Now if \(\frac{z}{y} = \frac{z}{w}\), i.e. \(xw = yz\), we have

\[
\tilde{\delta}\left(\frac{x}{y}\right) = \tilde{\delta}\left(\frac{yw^2}{y^2w^2}\right) = \tilde{\delta}\left(\frac{y^2zw}{y^2w^2}\right) = \tilde{\delta}\left(\frac{z}{w}\right).
\]
The embedding of $R$ into $D^{-1}R$ sends $x$ to $\frac{x}{1}$, and

$$\tilde{\delta}\left(\frac{x}{1}\right) = \frac{1 \cdot \delta(x) - x\delta(1)}{1} = \delta(x)$$

since $\delta(1) = 0$. Thus $\tilde{\delta}$ extends $\delta$.

To show that $\tilde{\delta}$ is a derivation, let $x, z \in R$ and $y, w \in D$. Then

$$\tilde{\delta}\left(\frac{x}{y}, \frac{z}{w}\right) = \frac{yvw\delta(xz) - xz\delta(yw)}{y^2w^2} = \frac{yw(x\delta(z) + z\delta(x)) - xz(y\delta(w) + w\delta(y))}{y^2w^2}$$

$$= \frac{yz\delta(x) - xz\delta(y) + xyw\delta(z) - xzw\delta(y)}{y^2w^2}$$

so $\tilde{\delta}$ is a derivation.

The last statement is obvious from the formula for $\tilde{\delta}$.

A corollary of the first part of the proof is:

**X.9.2.3. COROLLARY.** Let $\delta$ be a derivation on a unital ring $R$. If $y \in R$ is invertible, then

$$\delta(y^{-1}) = -y^{-1}\delta(y)y^{-1}.$$  

In particular, if $y$ is a constant for $\delta$, so is $y^{-1}$. So if $(F, \delta)$ is a differential field, the ring of constants for $\delta$ is a subfield of $F$.

Here are the most important examples of differential fields for our purposes:

**X.9.2.4. EXAMPLES.** (i) Let $F$ be a field of characteristic zero (e.g. $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$). The polynomial ring $F[X]$ in one variable over $F$ becomes a differential ring under the derivation

$$\delta \left( \sum_{k=0}^{n} a_k X^k \right) = \sum_{k=1}^{n} k a_k X^{k-1}$$

and the quotient field $F(X)$ of rational functions in one variable over $F$ becomes a differential field with the unique extension from **X.9.2.2.** given by the quotient rule.

There is a minor but technically important point about this example and the succeeding ones which is glossed over in many references on differential fields. The elements of $F(X)$ are called “rational functions,” but they are not exactly functions from $F$ (or a fixed subset) to $F$ since functions which agree except at finitely many points are identified. Thus, for example, $f(X) = X + 1$, $g(X) = \frac{X^2-1}{X-1}$, and $h(X) = \frac{X^2-X-2}{X-2}$ all represent the same element of $F(X)$ even though they represent different functions since their domains are different.

It is customary in differential algebra to play a little fast and loose with domains of “functions” since it turns out that the domains and other functional properties are not really relevant to the algebra, but one should be careful in the process.
(ii) Let $\Omega$ be a connected open set in $\mathbb{C}$. Let $\mathcal{M}(\Omega)$ be the set of meromorphic functions on $\Omega$, complex-valued functions which are differentiable on $\Omega$ except for isolated poles (or removable singularities). We again identify two functions if they agree when all removable singularities have been filled in, i.e. if the functions agree on a nonempty open subset of $\Omega$, or more generally on a subset of $\Omega$ with a limit point in $\Omega$. The reciprocal of a meromorphic function on $\Omega$ which is not identically zero is again a meromorphic function on $\Omega$, so $\mathcal{M}(\Omega)$ is a field. The usual complex differentiation is a derivation on $\mathcal{M}(\Omega)$ making it into a differential field.

$\mathbb{C}(X)$ may be regarded as a differential subfield of $\mathcal{M}(\Omega)$ for any $\Omega$. If $\Omega_1 \subseteq \Omega_2$, then restriction gives a natural embedding of $\mathcal{M}(\Omega_2)$ into $\mathcal{M}(\Omega_1)$, so $\mathcal{M}(\Omega_2)$ can be regarded as a subfield of $\mathcal{M}(\Omega_1)$. (It is a proper subfield if $\Omega_2 \neq \Omega_1$: for example, if $z_0 \in \Omega_2 \setminus \Omega_1$, then $f(z) = e^{1/(z-z_0)}$ is in $\mathcal{M}(\Omega_1)$ but not $\mathcal{M}(\Omega_2)$.) The field of constants in these differential fields is $\mathbb{C}$ (i.e. the usual constant functions).

If $\Omega$ is not connected, then $\mathcal{M}(\Omega)$ has zero divisors and is not a field: a function which is identically zero on one component of $\Omega$ but not on all of $\Omega$ is a zero divisor.

Differential Function Fields

These examples lead to a convenient definition of the principal objects of study in elementary differential algebra:

X.9.2.5. Definition. A differential function field or df is a differential subfield of $\mathcal{M}(\Omega)$ containing $\mathbb{C}(X)$, for some nonempty connected open subset $\Omega$ of $\mathbb{C}$.

As mentioned earlier, the exact set $\Omega$ will turn out to be unimportant, and we will feel free to reduce it whenever necessary. The next convenient result illustrates the technique:

X.9.2.6. Proposition. Let $F$ be a differential function field, and $K$ a finite extension of $F$. Then $K$ is isomorphic to a differential function field, and the derivation on $K$ is uniquely determined by the derivation on $F$.

Proof: Identify $F$ with a subfield of $\mathcal{M}(\Omega)$ for some $\Omega \subseteq \mathbb{C}$. Suppose $K = F(\phi)$ for some $\phi \in K$ ($K$ can be constructed from $F$ by finitely many such steps; in fact, it can be done in one step by the Primitive Element Theorem of algebra). Then $\phi$ is a root of an irreducible polynomial $p(Y)$ with coefficients in $F$. Regard $p$ as a function $p(z, w)$ from $\Omega \times \mathbb{C}$ to $\mathbb{C}$ which is analytic except where $z$ is a pole of a coefficient function of $p$. By the Analytic Implicit Function Theorem (VIII.6.8.4.), there is an open set $\Omega' \subseteq \Omega$ and an analytic function $g$ on $\Omega'$ such that $p(z, g(z)) = 0$ for all $z \in \Omega'$. The embedding of $K$ into $\mathcal{M}(\Omega')$ which is restriction on $F$ and which sends $\phi$ to $g$ represents $K$ as a differential function field.

Algebraic Extensions of Derivations

The uniqueness of the derivation follows from a more general version of this result:

X.9.2.7. Proposition. Let $(F, \delta)$ be a differential field, and $K$ an algebraic extension of $F$. Then $\delta$ extends uniquely to a derivation on $K$. 

991
PROOF: [Ros72] Define two derivations $\delta_0$, $\delta_1$ on $\mathbb{F}[X]$ by

$$
\delta_0 \left( \sum_{k=0}^{n} \alpha_k X^k \right) = \sum_{k=0}^{n} \delta(\alpha_k) X^k
$$

$$
\delta_1 \left( \sum_{k=0}^{n} \alpha_k X^k \right) = \sum_{k=1}^{n} k\alpha_k X^{k-1}
$$

(it is routine to check that $\delta_0$ and $\delta_1$ are derivations). Let $\tilde{\delta}$ be a derivation on $\mathbb{K}$ extending $\delta$. If $p(X) = \sum_{k=0}^{n} \alpha_k X^k$, then for any $z \in \mathbb{K}$,

$$
0 = \tilde{\delta}(p(z)) = \sum_{k=0}^{n} \tilde{\delta}(\alpha_k z^k) = \sum_{k=0}^{n} \tilde{\delta}(\alpha_k) z^k + \sum_{k=1}^{n} \alpha_k z^{k-1} \tilde{\delta}(z) = (\delta_0(p))(z) + (\delta_1(p))(z)\tilde{\delta}(z) .
$$

Let $f$ be the irreducible polynomial of $z$ over $\mathbb{F}$. Taking $p = f$, we have

$$
0 = \tilde{\delta}(0) = \tilde{\delta}(f(z)) = (\delta_0(f))(z) + (\delta_1(f))(z)\tilde{\delta}(z)
$$

$$
(\delta_1(f))(z)\tilde{\delta}(z) = -(\delta_0(f))(z) .
$$

But $\delta_1(f)$ is the “derivative” of $f$ in the usual algebraic sense. Since $\mathbb{F}$ has characteristic zero, $z$ is not a root of $\delta_1(f)$, i.e. $(\delta_1(f))(z) \neq 0$, and hence

$$
\tilde{\delta}(z) = -\frac{(\delta_0(f))(z)}{(\delta_1(f))(z)} .
$$

Thus an extension of $\delta$ to $\mathbb{K}$ (if one exists) is unique, and for any $z \in \mathbb{K}$, we have that $\tilde{\delta}(z) \in \mathbb{F}(z)$.

To show existence, we may assume $\mathbb{K} = \mathbb{F}(z)$ for some $z$. We have $\mathbb{K} = \mathbb{F}(z) \cong \mathbb{F}[X]/(f(X))$, where $f(X)$ is the irreducible polynomial of $z$ over $\mathbb{F}$. We need only show that the derivation

$$
\tilde{\delta} = \delta_0 - \frac{(\delta_0(f))(z)}{(\delta_1(f))(z)} \delta_1
$$

of $\mathbb{F}[X]$ drops to a derivation of $\mathbb{F}[X]/(f(X))$ (X.9.4.3.), since this drop will extend $\delta$. Thus we only need to show that $\tilde{\delta}$ maps the ideal $(f(X))$ into itself, i.e. that $\delta(f)$ is a multiple of $f$, or equivalently that $z$ is a root of $\tilde{\delta}(f)$. But

$$
(\tilde{\delta}(f))(z) = (\delta_0(f))(z) - \frac{(\delta_0(f))(z)}{(\delta_1(f))(z)} (\delta_1(f))(z) = 0 .
$$

Exponentials and Logarithms

X.9.2.8. DEFINITION. Let $(\mathbb{F}, \delta)$ be a differential field, and $x, y \in \mathbb{F}$, $y \neq 0$. Then $y$ is an exponential of $x$, and $x$ is a logarithm of $y$, if $\delta(y) = y\delta(x)$, or equivalently $\delta(x) = \frac{\delta(y)}{y}$. 

992
X.9.2.9. **Proposition.** (i) If $y_1$ is an exponential of $x$, then $y_2$ is an exponential of $x$ if and only if $y_2 = cy_1$ for some nonzero constant $c$.

(ii) If $x_1$ is a logarithm for $y$, then $x_2$ is a logarithm for $y$ if and only if $x_2 = x_1 + c$ for some constant $c$.

**Proof:** (ii) is trivial since $(x_2) = (x_1)$ if and only if $y_2 = y_1 + c$.

(i): The following statements are equivalent for $y_2 \neq 0$:

\[ \frac{\delta(y_2)}{y_2} = \frac{\delta(y_1)}{y_1} = \frac{\delta(y_1)}{y_1} \]

\[ y_2 \delta(y_1) = y_1 \delta(y_2) \]

\[ \frac{y_2 \delta(y_1) - y_1 \delta(y_2)}{y_2^2} = \delta \left( \frac{y_1}{y_2} \right) = 0. \]

X.9.2.10. **Example.** Let $f \in M(\Omega)$ for some connected open set $\Omega \subseteq \mathbb{C}$. An exponential for $f$ is a function $g$ of the form $cf^k$ for some nonzero $c \in \mathbb{C}$. The function $g$ is meromorphic (holomorphic) on any subregion $\Omega' \subseteq \Omega$ containing no poles of $f$, i.e. a subset on which $f$ is holomorphic.

If $f$ is not identically zero, a logarithm for $f$ is a function $h$ of the form $c + \log f$ for some $c \in \mathbb{C}$. The function $h$ can be chosen meromorphic (holomorphic) on any simply connected $\Omega' \subseteq \Omega$ not containing any zeroes or poles of $f$.

Thus, if $F$ is any differential function field, and $f \in F$, there is a df $K$ containing $F$ containing an exponential for $f$; if $f$ is nonzero there is also such a $K$ containing a logarithm for $f$. The smallest such df is well defined up to isomorphism, and is called the df obtained from $F$ by adding an exponential [logarithm] of $F$; it is isomorphic to an evident quotient of the universal differential algebra $F\{X\}$ (by the differential ideal generated by $X_1 - fX$ in the exponential case and $fX_1 - f'$ in the logarithm case).

X.9.2.11. **Proposition.** [*Logarithmic Differentiation Formula*] Let $(F, \delta)$ be a differential field, $a_1, \ldots, a_n$ nonzero elements of $F$, and $r_1, \ldots, r_n \in \mathbb{N}$. Set $a = a_1^{r_1} a_2^{r_2} \cdots a_n^{r_n}$. Then

\[ \frac{\delta(a)}{a} = r_1 \frac{\delta(a_1)}{a_1} + r_2 \frac{\delta(a_2)}{a_2} + \cdots + r_n \frac{\delta(a_n)}{a_n}. \]

**Proof:** This is an easy proof by induction on $r_1 + \cdots + r_n$, using the product rule.
X.9.3. Integration in Finite Terms

The name of this section has been the title of at least two important publications on the subject ([?), [Ros72]). The general problem is: if \( \delta \) is a derivation on a ring \( R \), what is the range of \( \delta \), and in particular is \( \delta \) surjective? If \( x \in R \), can \( \delta \) be extended to a well-behaved larger ring with \( x \) in the range of the extended derivation?

Rather than treating the general problem, we will have the much more modest goal of examining the question for differential function fields. In other words, we will determine conditions for when there is, or is not, a differentiable function \( g \) with an elementary formula whose derivative is a given function \( f \) which has an elementary formula, or equivalently whether

\[
\int f(x) \, dx
\]

has a formula in terms of standard functions.

We must first carefully determine what it means for a function to have an “elementary formula.”

Elementary Formulas

X.9.3.1. We will informally say a function has an elementary formula if it can be written in terms of algebraic operations and exponential, logarithm, trigonometric, and inverse trigonometric functions, which are defined on some common interval. (We could include hyperbolic and inverse hyperbolic functions too, but these functions can be written in terms of exponential and logarithm functions so they come for free).

X.9.3.2. It is convenient and useful to pass to complex functions. Each of the “building block” functions (algebraic, exponential, logarithm, trigonometric, inverse trigonometric) can be regarded as a holomorphic function on large regions of the complex plane, and hence any real-valued function defined on a real interval which has an elementary formula in the above sense will be holomorphic on a large region in \( \mathbb{C} \). In addition, when we expand to complex functions, we get the pleasant fact that trigonometric and inverse trigonometric functions can also be written in terms of exponential and logarithm functions () and thus do not have to be separately considered. We can thus reduce the problem to consideration of only functions which can be written with formulas involving only algebraic operations and exponential and logarithm functions.

X.9.3.3. Definition. Let \((F, \delta)\) be a differential field. A differential field \( K \) containing \( F \) is an elementary extension of \( F \) if there is a finite chain

\[
F = F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n = K
\]

of field extensions where, for each \( k \), \( F_{k+1} \) is either a finite extension of \( F_k \) or is obtained by adding to \( F_k \) an exponential or logarithm of an element of \( F_k \).

An elementary function field is an elementary extension of the field \( \mathbb{C}(X) \) of rational functions over \( \mathbb{C} \). A function \( f \) which is meromorphic on a region in \( \mathbb{C} \) is an elementary function if it lies in some elementary function field.

X.9.3.4. By (), an elementary function field is a differential function field. Any elementary extension of an elementary extension of a differential field \( F \) is an elementary extension of \( F \); in particular, an elementary extension of an elementary function field is an elementary function field.
X.9.3.5. Example. A function with a formula like
\[ f(z) = \frac{\sqrt[3]{e^{\cos(z^2+1)} - \log(\sin^{-1} z)}}{z^5 - iz + 1} \]
is an elementary function: it lies in the field extension

The problem of integration in finite terms thus becomes the following specific problem:

X.9.3.6. Question. Let \( f \) be an elementary function. Is there an elementary function \( g \) with \( g' = f \)?

X.9.3.7. It would be clean to take the field of all elementary functions, with its natural derivative (derivation), and ask what the range of this derivation is. However, it is not clear that this version of the question is well defined. (There is certainly no region in \( \mathbb{C} \) on which all elementary functions are meromorphic.) Thus we have to work locally in elementary function fields.

A more general version of the question is thus:

X.9.3.8. Question. Let \((\mathbb{F}, \delta)\) be a differential field, and \( x \in \mathbb{F} \). Under what conditions is there an elementary extension \((\mathbb{K}, \delta)\) of \( \mathbb{F} \) and an element \( y \in \mathbb{K} \) with \( x = \delta(y) \)?

X.9.3.9. In 1835, Liouville gave a necessary and sufficient condition (X.9.3.11.) for elements of elementary function fields, which was abstracted by Ostrowski [?] in 1946 (cf. [Ros72]):

X.9.3.10. Theorem. Let \((\mathbb{F}, \delta)\) be a differential field, and \( x \in \mathbb{F} \). Then there is an elementary extension \((\mathbb{K}, \delta)\) of \((\mathbb{F}, \delta)\) with the same field of constants as \((\mathbb{F}, \delta)\), and an element \( y \in \mathbb{K} \) with \( x = \delta(y) \), if and only if there are constants \( c_1, \ldots, c_n \) and elements \( u_1, \ldots, u_n, v \) of \( \mathbb{F} \) such that
\[
x = \sum_{k=1}^{n} c_k \frac{\delta(u_k)}{u_k} + \delta(v) .
\]

X.9.3.11. Corollary. Let \((\mathbb{F}, \delta)\) be an elementary function field, and \( f \in \mathbb{F} \). Then there is an elementary function \( g \) with \( f = g' \) if and only if there are constants \( c_1, \ldots, c_n \) and functions \( h_1, \ldots, h_n, \phi \) in \( \mathbb{F} \) such that
\[
f = \sum_{k=1}^{n} c_k \frac{h'_k}{h_k} + \phi' .
\]
X.9.3.12. To obtain the corollary from the theorem, it suffices to note that all differential function fields have the same field of constants, namely the constant functions. One direction of the theorem is very easy: if $x$ is of the given form, one only needs to add logarithms $z_1, \ldots, z_n$ of $u_1, \ldots, u_n$ to $\mathbb{F}$, and then

$$y = c_1 z_1 + \cdots + c_n z_n + v$$

is an antiderivative of $x$. Thus the conclusion of the theorem can be strengthened, since $K$ is obtained by just adding logarithms.

The other direction is a more involved argument requiring some elementary field theory, and is omitted. See e.g. [Ros72]. The requirement that $K$ have the same constants as $F$ is necessary: if $F = \mathbb{R}(X)$, and $f(x) = \frac{1}{x^2 + 1}$, then $f$ is not of the given form in $F$ even though it has an elementary antiderivative in an extension of $\mathbb{C}(X)$ (which is a finite, hence elementary, extension of $F$) by a logarithm. (It does have such an expression in $\mathbb{C}(X)$, to which the theorem applies.)

Applications

We now discuss some simple functions which have no elementary antiderivatives.

X.9.3.13. We first examine integrals of the form

$$\int \frac{1}{\sqrt{p(x)}} \, dx$$

where $p$ is a polynomial with complex coefficients of degree $\geq 3$ and with no multiple roots. Integrals of this type include elliptic integrals, which arise in computing the arc length of an ellipse. Of course, what we want is an elementary function $g$ whose derivative is $f(z) = \frac{1}{\sqrt{p(z)}}$ (say, on some subregion of $\mathbb{C}$; it could be the inverse image under $1/p$ of any simply connected region not containing 0).

X.9.3.14. We now examine integrals of the form

$$\int f(x) e^{g(x)} \, dx$$

where $f$ and $g$ are rational functions. Liouville showed:

X.9.3.15. Theorem. Let $f$ and $g$ be in $\mathbb{C}(X)$, with $f \neq 0$ and $g$ not constant. Then there is an elementary function $h$ such that $h'(z) = f(z)e^{g(z)}$ if and only if there is a rational function $r$ such that

$$r'(X) + g'(X)r(X) = f(X)$$

in $\mathbb{C}(X)$. If there is such an $r$, then $h(z) = r(z)e^{g(z)}$ is an elementary antiderivative of $fe^g$.

X.9.3.16. Note that the condition is considerably stronger than simply asserting that the differential equation

$$y' + g'(t)y = f(t)$$

has a solution, which will always be the case from existence theorems for differential equations (at least if $f$ and $g$ have real coefficients); the condition requires that it have a rational solution. A corollary (equivalent version) of the theorem is that if the differential equation does not have a rational solution, it does not have an elementary solution.
X.9.3.17. For example, consider the integral

\[ \int \frac{e^x}{x} \, dx . \]

Applying the theorem with \( f(X) = \frac{1}{X} \) and \( g(X) = X \), this integral has an elementary formula if and only if the differential equation

\[ r'(X) + r(X) = \frac{1}{X} \]

has a rational solution. Suppose \( r(X) = \frac{p(X)}{q(X)} \) is a solution with \( p(X) \), \( q(X) \) polynomials which are relatively prime. We have

\[ \frac{q(X)p'(X) - p(X)q'(X)}{[q(X)]^2} + \frac{p(X)}{q(X)} = \frac{q(X)p'(X) - p(X)q'(X) + p(X)q(X)}{[q(X)]^2} = \frac{1}{X} \]

in \( \mathbb{C}(X) \), and hence \( X \) divides \([q(X)]^2\) and therefore \( q(X) \). Thus

\[ q(X) = X^n \phi(X) \]

for some \( n \geq 1 \) and some polynomial \( \phi(X) \) not divisible by \( X \). Note that \( X \) does not divide \( p(X) \). We have

\[
\begin{align*}
\frac{X^n \phi(X)p'(X) - p(X)[nX^{n-1} \phi(X) + X^n \phi'(X)]}{X^{2n}[\phi(X)]^2} + \frac{p(X)}{X^n \phi(X)} &= \frac{X^n [\phi(X)p'(X) - \phi'(X)p(X) + \phi(X)p(X)] - nX^{n-1} \phi(X)p(X)}{X^{2n}[\phi(X)]^2} \\
&= \frac{X [\phi(X)p'(X) - \phi'(X)p(X) + \phi(X)p(X)] - n \phi(X)p(X)}{X^{n+1}[\phi(X)]^2} = \frac{1}{X}
\end{align*}
\]

in \( \mathbb{C}(X) \). The denominator is divisible by \( X^{n+1} \), so the numerator must be divisible by \( X^n \), and in particular by \( X \). Thus \( n \phi(X)p(X) \) must be divisible by \( X \). But \( X \) does not divide \( p(X) \), so it must divide \( \phi(X) \), a contradiction. Thus there is no rational solution, and no elementary formula for the antiderivative.

The substitution \( x = \log u \) converts this integral into

\[ \int \frac{1}{\log u} \, du \]

which therefore also does not have an elementary formula. This integral arises in the Prime Number Theorem (\( \)).

X.9.3.18. Now we turn to the best-known example of this type:

\[ \int e^{-x^2} \, dx \]

which is important in probability (\( \)). The arguments for

\[ \int e^{x^2} \, dx \]

997
are essentially identical.

Applying the theorem with \( f(X) = 1 \) and \( g(X) = -X^2 \), we need to show that there is no rational function \( r(X) \) satisfying

\[
r'(X) - 2Xr(X) = 1 .
\]

Clearly there is no polynomial satisfying this equation, since if \( r(X) \) is a polynomial of degree \( d \), the left side is a polynomial of degree \( d + 1 \). Thus we may write \( r(X) = \frac{p(X)}{q(X)} \) in lowest terms with \( q \) not constant. Then \( q \) has a root \( z_0 \in \mathbb{C} \), i.e. \( r \) has a pole of some finite order \( n \geq 1 \) at \( z_0 \), so \( r \) can be written

\[
r(X) = \frac{h(X)}{(X - z_0)^n}
\]

with \( h(X) \in \mathbb{C}(X) \) defined and nonzero at \( z_0 \), and \( h \) and \( h' \) are continuous at \( z_0 \). Set

\[
\phi(X) = r'(X) - 2Xr(X) = \frac{(X - z_0)^n h'(X) - n(X - z_0)^{n-1}h(X)}{(X - z_0)^{2n}} - \frac{2Xh(X)}{(X - z_0)^n}
\]

\[
= \frac{(X - z_0)^n[h'(X) - 2Xh(X)] - n(X - z_0)^{n-1}h(X)}{(X - z_0)^{2n}} = \frac{(X - z_0)[h'(X) - 2Xh(X)] - nh(X)}{(X - z_0)^{n+1}}
\]

and thus, if \( \phi \) is regarded as a meromorphic function on \( \mathbb{C} \), we have

\[
\lim_{z \to z_0} (z - z_0)^{n+1} \phi(z) = \lim_{z \to z_0} \left[ (z - z_0)[h'(z) - 2zh(z)] - nh(z) \right] = nh(z_0) \neq 0
\]

which contradicts that \( \phi(X) = 1 \) in \( \mathbb{C}(X) \).

Thus there is no rational solution to the differential equation, and no elementary formula for the antiderivative.

**X.9.4. Exercises**

**X.9.4.1.** Let \((R, \delta)\) and \((\tilde{R}, \tilde{\delta})\) be differential rings. A ring-homomorphism \( \phi : R \to \tilde{R} \) is a **differential homomorphism**, or **diffomorphism**, if \( \phi \) intertwines \( \delta \) and \( \tilde{\delta} \), i.e. \( \tilde{\delta}(\phi(x)) = \phi(\delta(x)) \) for all \( x \in R \). [The term "diffomorphism" is to "diffeomorphism" exactly as "homomorphism" is to "homeomorphism".]

Show that if \((R, \delta)\) is a commutative differential ring and \( z \in R \), there is a unique diffomorphism from the differential polynomial ring \( R_0\{X\} \) (X.9.1.3.(iii)) to \( R \) sending \( X \) to \( z \). Thus \( R_0\{X\} \) is the "universal" differential algebra with one generator over the ring \( R_0 \) of constants.

**X.9.4.2.** Generalize the construction of the differential polynomial ring as follows. Let \((F, \delta)\) be a commutative differential ring, and let \( X, X_1, X_2, \ldots \) be a sequence of indeterminates. Form the polynomial ring

\[
R = F[X, X_1, X_2, \ldots] .
\]

Define a map \( \tilde{\delta} : R \to R \) by

\[
\tilde{\delta} \left( \sum_{k,n} \alpha_{k,n} X_k^n \right) = \sum_{k,n} \delta(\alpha_{k,n}) X_k^n + \sum_{k,n} n\alpha_{k,n} X_k^{n-1} X_{k+1} .
\]
(i) Show that $\hat{\delta}$ is a derivation on $R$ which extends the derivation $\delta$ on the subring $F$ of “constant polynomials” (which is not the ring of constants for $\hat{\delta}$ in general!) Denote the ring $R$ by $F\{X\}$. Note that if $\delta$ is the zero derivation on $F$, then this definition agrees with the previous definition of $F\{X\}$.

(ii) Show that if $(K, \delta)$ is another commutative differential ring and $\phi : F \to K$ is a diffeomorphism, and $z \in K$, then there is a unique extension $\phi$ of $\phi$ to a diffeomorphism from $F\{X\}$ to $K$ sending $X$ to $z$. Thus $F\{X\}$ is the “universal” differential algebra in one variable over the differential ring $(F, \delta)$.

**X.9.4.3.** Let $(R, \delta)$ be a differential ring, and $J$ a closed two-sided ideal of $R$ (i.e., if $\delta(J) \subseteq J$, show that $\delta$ drops to a derivation $\hat{\delta}$ on the quotient ring $R/J$ satisfying $\hat{\delta}(\pi(x)) = \pi(\delta(x))$ for all $x \in R$, where $\pi : R \to R/J$ is the quotient map (i.e. $\pi$ is a diffeomorphism). An ideal of $R$ invariant under $\delta$ is called a **differential ideal**.

**X.9.4.4.** Let $(K, \delta)$ be a differential field, and $F$ a subfield of $K$ closed under $\delta$. If $y \in K \setminus F$ is an exponential or logarithm of an element of $F$, is $y$ necessarily transcendental over $F$? What if $F$ contains all constants of $K$?

**X.9.4.5.** Define a one-parameter automorphism group of $C^\infty_c(\mathbb{R})$ by

$$[\alpha_t(f)](x) = f(x + t).$$

Show that $\alpha_t$ is an automorphism for each $t$, that $\alpha_{t+s} = \alpha_t \circ \alpha_s$ for all $t, s$. Discuss continuity of this one-parameter automorphism group and how the “derivative” in the sense of X.9.1.3.(v) is the derivation given by ordinary differentiation.
Chapter XI

Topology

Topology is a part of mathematics which provides a general setting for the notions of limits and continuity.

Topology is a huge subject, and goes far beyond the parts which form an essential tool for analysis and
other parts of mathematics. Our treatment of topology will mostly just cover the basic tools of the subject,
with an eye to applications in analysis.

In his classic topology book [Kel75], J. Kelley described the contents as “what every young analyst
should know.” While his judgment was of course subjective, it can be taken as a good guide. The topology
covered in this book largely represents my own judgment of What Every Analyst Should Know, but I admit
some idiosyncracies due to my own personal interests.

The idea behind topology is to describe what it means for points of a set to be “close together.” There
are at least four standard and commonly used equivalent ways to define, specify, or think of a topology on
a set $X$ (usually called a space):

(1) Specifying what the neighborhoods of each point of $X$ are, i.e. specifying which subsets of $X$ are open.
(2) Specifying the closure of each subset of $X$.
(3) Specifying which sequences (or generalized sequences) in $X$ converge, and what their limits are.
(4) Specifying which functions into or out of $X$ are continuous.

A good way of doing (1) is to specify a metric on $X$, and hence open balls in $X$, but there are other ways
to do (1) also.

Approach (1) is normally taken as the definition of a topology. We will in due course describe all four
approaches and how to go from each to the others.

Topology as used by topologists (or as presented in most elementary topology classes) has a rather
different emphasis than topology as used by analysts. Topologists by and large consider spaces which have
only one obvious or natural topology (except sometimes for exotic topologies artificially constructed primarily
to provide counterexamples; there are other exceptions too). When a topologist considers convergence of a
sequence or continuity of a function, there is a natural interpretation, and it is not necessary to go back and
worry about “with respect to what topology?” The topology tends to be taken for granted.
An analyst must quickly overcome this mindset. Spaces arising in analysis, notably function spaces, often have more than one *natural* topology, sometimes a great many, with varying properties. Convergence in one sense does not generally imply convergence in another sense, and some properties (e.g. continuity of functions) are preserved under certain kinds of limits but not others. Thus, for example, an analyst should never say of functions that \( f_n \to f \) without carefully specifying what is meant, because such a statement is totally ambiguous otherwise. An analyst needs to be acutely aware of which topology is being considered in a given situation, and that there is usually a multitude of reasonable candidates.

Topology is a subject which is full of definitions; in fact, the density of definitions in topology may be the highest in all the parts of mathematics. This situation arises because of the great generality of the subject and the enormous variety of topological spaces with varied properties. Most definitions in topology are of little direct interest to analysts. We will for the most part stick to those definitions relevant to analysis to try to avoid getting too bogged down in terminology.
Figure XI.1: Rumored original source of topology in Serbia
XI.1. Topological Spaces

The most important tool used in the study of limits and continuity in Euclidean spaces and metric spaces is the notion of “open set.” The idea behind the definition of a topological space is to abstract the notion of open set directly, without working through intermediate structures like metrics. Thus a topological space will just be an abstract set $X$ together with a specified collection of subsets which we arbitrarily call “open” sets in $X$; the term “open” does not have any intrinsic or geometric meaning in this context. We only require that the “open” sets have some simple union and intersection properties which hold for open sets in Euclidean space, and which are necessary to make a reasonable theory of limits and continuity.

XI.1.1. Definitions

XI.1.1.1. Definition. A topology on a set $X$ is a subset $T$ of $\mathcal{P}(X)$ with the following properties:

(i) $\emptyset, X \in T$.

(ii) $T$ is closed under finite intersections.

(iii) $T$ is closed under arbitrary unions.

The sets in $T$ are called the open sets (or $T$-open sets) in $X$.

A topological space is a pair $(X, T)$, where $X$ is a set and $T$ is a topology on $X$.

There is a technical question whether a topological space is required to be nonempty (the pair $(\emptyset, \mathcal{P}(\emptyset))$ satisfies the definition). We will generally assume topological spaces are nonempty, but explicitly state this when it is important.

XI.1.1.2. A set can have many different topologies. For example, the function spaces studied in analysis typically have several different natural topologies, and it is necessary to carefully distinguish between them. In some settings, there is one standard topology which is obvious from the context, so by abuse of language we often make statements like “let $X$ be a topological space.” However, it should be carefully noted and always kept in mind that the specification of a topology is an essential part of the definition of a topological space.

Topological spaces are the natural abstract setting for the notions of limits and continuity. We typically refer to elements of a topological space as “points.” In this connection, it is convenient to think of an open set containing a point as a “neighborhood” of the point. Slightly more generally:

XI.1.1.3. Definition. Let $(X, T)$ be a topological space, and $x \in X$. A neighborhood (or $T$-neighborhood) of $x$ is a set containing an open set which contains $x$.

Note that by this definition, a neighborhood of a point in $X$ is not necessarily an open set. It is sometimes convenient to work with neighborhoods which are not open. But the most important neighborhoods of points are open neighborhoods; every neighborhood of a point $x$ contains an open neighborhood of $x$, which is simply an open set containing $x$. We will try to always specify “open neighborhood” when open sets are required, even though this sometimes becomes a little pedantic.
Topologies and Metrics.

XI.1.1.4. **Examples.**

(i) If $X$ is any set, then $\mathcal{P}(X)$ is a topology on $X$, called the *discrete topology*.

(ii) If $X$ is any set, then $\{\emptyset, X\}$ is a topology on $X$, called the *indiscrete topology*.

(iii) The most important examples of topologies are those defined by metrics. If $(X, \rho)$ is a metric space, as usual define a subset of $X$ to be open (or $\rho$-open) if it contains an open ball around each of its points. The set $\mathcal{T}_{\rho}$ of $\rho$-open sets is a topology on $X$, called the *topology defined by (or underlying topology of)* the metric space $(X, \rho)$. Equivalent metrics define the same topology.

The discrete topology on a set $X$ comes from a metric (the *discrete metric*) $\rho$, where $\rho(x, y) = 1$ if $x \neq y$.

The usual notions of convergence of sequences and closure of a set in a metric space, or of continuity of functions between metric spaces, depend only on the underlying topology and can all be phrased in terms of the topology. See () and Problem () for continuity, and () for the other notions.

XI.1.1.5. There are, however, many examples of topological spaces which do not come from metric spaces (e.g. the indiscrete topology on a set with more than one element), some of which are important in applications. Even a metrizable topological space (one whose topology comes from a metric) need not have any *natural* metric defining the topology, and it is often unnatural to artificially pick one of the many possible metrics defining a topology which is given in another way. Thus it is useful to work directly with topological spaces rather than metric spaces in many situations.

XI.1.1.6. Although topologies, even interesting topologies, do not come from metrics in general, most interesting topologies share some of the properties of metric spaces: there are “enough” open sets that points have “small” neighborhoods, and in particular distinct points have disjoint neighborhoods. There are varying degrees of similarity to metrizable spaces which are useful, described by the separation axioms developed in ()

XI.1.2. **Bases and Subbases**

XI.1.2.1. If $\{\mathcal{T}_\alpha\}$ is a family of topologies on a set $X$, then $\bigcap_{\alpha} \mathcal{T}_\alpha$ is also a topology on $X$ (Exercise XI.1.6.2.). In particular, if $\mathcal{S}$ is any subset of $\mathcal{P}(X)$, there is a smallest topology containing $\mathcal{S}$, called the *topology generated by* $\mathcal{S}$ (there is always at least one topology containing $\mathcal{S}$, the discrete topology).

**Bases**

It is usually convenient to use a smaller collection of “nice” open sets to describe a topology. For example, in a topology defined by a metric, the open balls generate the topology as in XI.1.2.1. We abstract this situation in the notion of base:
XI.1.2.2. DEFINITION. Let \((X, \mathcal{T})\) be a topological space, and \(p \in X\). A subset \(B\) of \(\mathcal{T}\) is a local base for \(\mathcal{T}\) at \(p\) if, whenever \(U \in \mathcal{T}\) and \(p \in U\), there is a \(V \in B\) with \(p \in V\) and \(V \subseteq U\).

A subset \(B\) of \(\mathcal{T}\) is a base for \(\mathcal{T}\) if it is a local base at every point, i.e. whenever \(U \in \mathcal{T}\) and \(x \in U\), there is a \(V \in B\) with \(x \in V\) and \(V \subseteq U\). In other words, \(U = \bigcup_{V \in B} V\), i.e. every open set is a union of sets in \(B\).

In particular, if \(B\) is a base for \(\mathcal{T}\), then \(\mathcal{T}\) is the topology generated by \(B\).

When a base or local base \(B\) is specified, the sets in \(B\) are usually called basic open sets. Some authors use the term basis instead of base, in analogy with the use of “basis” for a vector space; but this is potentially misleading, since a base for a topology does not necessarily have any type of minimality condition like a basis for a vector space: for example, any collection of open sets containing a base for \(\mathcal{T}\) is also a base for \(\mathcal{T}\).

XI.1.2.3. EXAMPLES. (i) If \((X, \rho)\) is a metric space, and \(p \in X\), the open balls centered at \(p\) form a local base for the topology at \(p\). The set of all open balls in \(X\) forms a base for \(\mathcal{T}_\rho\).

(ii) If \(X\) has the discrete topology, and \(p \in X\), then \(\{\{p\}\}\) is a local base for the topology at \(p\). The collection of all singleton subsets of \(X\) is a base for the discrete topology.

(iii) If \(\mathcal{T}\) is a topology on \(X\), then the collection of all sets in \(\mathcal{T}\), or the collection of all nonempty sets in \(\mathcal{T}\), is a (not very interesting) base for \(\mathcal{T}\).

XI.1.2.4. There is nothing unique about a base for a given topology. For example, if \(\rho_1\) and \(\rho_2\) are equivalent metrics on a set \(X\), the open balls from \(\rho_1\) and \(\rho_2\) may be quite different, but the open sets they generate will be the same. See (i) for an example.

It is useful to know when a collection \(B\) of open subsets of \(X\) forms a base for the topology it generates.

XI.1.2.5. PROPOSITION. Let \(X\) be a set and \(\mathcal{B} \subseteq \mathcal{P}(X)\), and let \(\mathcal{T}\) be the topology generated by \(\mathcal{B}\). Then \(\mathcal{B}\) is a base for \(\mathcal{T}\) if and only if

(i) \(\bigcup_{V \in \mathcal{B}} V = X\) and

(ii) If \(V_1, \ldots, V_n \in \mathcal{B}\) and \(x \in \bigcap_{j=1}^n V_j\), there is a \(V \in \mathcal{B}\) with \(x \in V\) and \(V \subseteq \bigcap_{j=1}^n V_j\).

XI.1.2.6. Note that a base for a topology need not be closed under finite intersections. For example, an intersection of open balls with different centers in a metric space is generally not an open ball. (For XI.1.2.5.(ii), it obviously suffices that \(\mathcal{B}\) be closed under finite intersections, as is frequently the case.)

First and Second Countability

Most interesting topological spaces have uncountably many open sets. But frequently there is a countable base for the topology, or at least a countable local base at each point, and if so the topology is more easily describable and tends to behave more nicely.
XI.1.2.7. **Definition.** Let \((X, \mathcal{T})\) be a topological space.

(i) \((X, \mathcal{T})\) satisfies the *first axiom of countability*, or is *first countable*, if there is a countable local base for \(\mathcal{T}\) at each point.

(ii) \((X, \mathcal{T})\) satisfies the *second axiom of countability*, or is *second countable*, if there is a countable base for \(\mathcal{T}\).

The “first” and “second” are used for historical reasons, but the terminology persists because it is both appropriate and useful.

XI.1.2.8. **Examples.**

(i) The discrete topology on a set \(X\) is always first countable; it is second countable if and only if \(X\) is countable.

(ii) The indiscrete topology on a set \(X\) is always second countable.

(iii) The usual topology on \(\mathbb{R}^n\) is second countable: balls with rational radii around points with rational coordinates form a base for the topology. More generally, the topology of a metric space is always first countable (for any point \(p\), balls of rational radius centered at \(p\) form a local base at \(p\)), and is second countable if and only if the metric space is separable (XI.2.3.15.).

(iv) Let \(X\) be a set. The *finite complement topology* on \(X\) is the topology consisting of \(\emptyset\) and all subsets of \(X\) with finite complement. (It is easily checked that this is a topology; cf. Exercise XI.1.6.4.) If \(X\) is countable, then this topology is second countable: in fact, there are only countably many open sets in all. But if \(X\) is uncountable, then the finite complement topology is not first countable, and in fact no point has a countable neighborhood base. For if \(p \in X\) and \(\{U_n : n \in \mathbb{N}\}\) is a countable collection of open neighborhoods of \(p\), set \(E = \cap_{n=1}^{\infty} U_n\). Then \(E\) has countable complement and is therefore uncountable, and in particular contains a point \(q \neq p\). Then \(X \setminus \{q\}\) is an open neighborhood of \(p\) not containing any \(U_n\), so the \(U_n\) cannot form a local base at \(p\).

Topological spaces which are not first countable are rather unusual, but they do arise naturally in analysis; see (). First countability is important since it is precisely in first countable topological spaces that the topology can be entirely described by convergence of sequences; in spaces which are not first countable it is necessary to work with generalizations of sequences called nets.

XI.1.2.9. **Caution:** For a topological space to be first countable, it is not sufficient that every singleton subset be a countable intersection of open sets (which may be assumed to be decreasing since a finite intersection of open sets is open), i.e. that every singleton subset is a \(G_δ\): there are countable topological spaces which are not first countable (Exercise XI.11.12.8.). (The condition that singletons are \(G_δ\)’s is not necessary either: the indiscrete topology on any set is even second countable. In a first countable \(T_1\) space () , every singleton subset is a \(G_δ\) (XI.7.3.3.). In fact, see XI.7.8.1.)

XI.1.2.10. **Proposition.** Let \((X, \mathcal{T})\) be a second countable topological space, and \(\mathcal{B}\) be a base for \(\mathcal{T}\). Then \(\mathcal{B}\) contains a countable base for \(\mathcal{T}\). In particular, every \(U \in \mathcal{T}\) is a countable union of sets in \(\mathcal{B}\).
XI.1.2.11. We can more generally define the weight \( w(X) \) of a topological space \( X \) to be the minimum cardinality of a base for the topology of \( X \) (this is only well defined in general if the AC is assumed). Thus \( X \) is second countable if and only if \( w(X) \leq \aleph_0 \). Weight is an example of a \textit{cardinal invariant} for topological spaces. The simplest cardinal invariant for \( X \) is \( \text{card}(X) \); another cardinal invariant is \( \text{density} \) \( d(X) \) (XI.2.3.16.). There are many other cardinal invariants, such as:

(i) \textit{character} \( \chi(X) \), minimum cardinal \( \kappa \) such that every point of \( X \) has a local base of cardinality \( \leq \kappa \), e.g. \( X \) is first countable if and only if \( \chi(X) \leq \aleph_0 \).

(ii) \textit{Lindelöf number} \( L(X) \), the smallest cardinal \( \kappa \) such that every open cover of \( X \) has a subcover (XI.1.3.3.) of cardinality \( \leq \kappa \), e.g. \( X \) is Lindelöf (XI.1.3.4.) if and only if \( L(X) \leq \aleph_0 \).

(iii) \textit{cellularity} \( c(X) \), minimum cardinal \( \kappa \) such that every family of pairwise disjoint nonempty open subsets of \( X \) has cardinality \( \leq \kappa \) . \( X \) has the \textit{countable chain condition} if \( c(X) \leq \aleph_0 \).

The invariants \( w(X) \), \( d(X) \), \( L(X) \), and \( c(X) \) coincide for metrizable spaces, but differ in general. A thorough discussion of cardinal invariants can be found in [KV84], and there is also extensive discussion in [Eng89].

Subbases

It is sometimes convenient to work with a less restrictive set of generators:

XI.1.2.12. \textbf{Definition.} Let \((X, \mathcal{T})\) be a topological space, and \( S \subseteq \mathcal{T} \). Then \( S \) is a \textit{subbase} for \( \mathcal{T} \) if the collection of finite intersections of sets in \( S \) forms a base for \( \mathcal{T} \).

XI.1.2.13. \textbf{Example.} In \( \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \), let \( S \) be the collection of all sets of the form \( I \times \mathbb{R} \) or \( \mathbb{R} \times I \), where \( I \) is an open interval in \( \mathbb{R} \). Then \( S \) is a subbase for the usual topology on \( \mathbb{R}^2 \).

More generally, the topology generated by a family of pseudometrics is naturally defined by a subbase. See ( ).

XI.1.2.14. It is sometimes convenient to extend the notion of base to allow sets which are not open. We say that a collection \( \mathcal{B} \) of not necessarily open subsets of \( X \) is a base for the topology \( \mathcal{T} \) if, whenever \( U \in \mathcal{T} \) and \( x \in U \), there is an \( E \in \mathcal{B} \) which is a neighborhood of \( x \) for which \( E \subseteq U \). For example, the set of closed balls of positive radius in a metric space is a base for the topology in this extended sense. We can similarly extend the notion of local base and subbase. Thus, for example, we can in some cases (e.g. in \( \mathbb{R}^n \)) take a base for a topology consisting of compact sets ( ).

If \( \mathcal{B} \) is a base for the topology \( \mathcal{T} \) in this sense, then the interiors (XI.2.3.17.) of the sets in \( \mathcal{B} \) form a base for the topology consisting of open sets. Thus if \( X \) has a countable base in this sense, it is second countable. The same applies to local bases; in particular, if every point of \( X \) has a countable local base in this sense, then \( X \) is first countable.
XI.1.3. Covers and Subcovers

It is often convenient to cover a topological space with “small” subsets, particularly small open sets.

XI.1.3.1. Definition. Let $X$ be a set, and $U = \{U_i : i \in I\}$ a collection of subsets of $X$. If $Y \subseteq X$, then $U$ covers $Y$, or is a cover of $Y$, if $Y \subseteq \bigcup_{i \in I} U_i$. $U$ covers $X$ if $\bigcup_{i \in I} U_i = X$.

If $X$ is a topological space, each $U_i$ is open in $X$, and $U$ covers $X$, then $U$ is an open cover of $X$.

XI.1.3.2. Examples. (i) Any base or subbase for the topology of $X$ is an open cover of $X$.

(ii) If $(X, \rho)$ is a metric space and $\epsilon > 0$ is fixed, then the open balls of radius $\epsilon$ form an open cover of $X$.

XI.1.3.3. Definition. Let $X$ be a set, $Y \subseteq X$, and $U$ a cover of $Y$. A subcover of $U$ is a subcollection of $U$ which also covers $Y$.

Any subcover of an open cover is also an open cover. Although the terminology is imprecise, one must specify the set $Y$ to discuss subcovers; normally one uses “subcover” only for covers of the whole set $X$.

XI.1.3.4. Definition. Let $X$ be a topological space.

(i) $X$ is a Lindelöf space if every open cover of $X$ has a countable subcover.

(ii) $X$ is compact if every open cover of $X$ has a finite subcover.

Compact spaces are discussed in detail in (). Obviously every compact space is Lindelöf; the converse is far from true.

XI.1.3.5. Proposition. Every second countable topological space is Lindelöf.

Proof: Let $\{V_n : n \in \mathbb{N}\}$ be a countable base for the topology of $X$. Let $U$ be an open cover of $X$. Fix a $U_0 \in U$. For each $n \in \mathbb{N}$, if $V_n$ is contained in some $U \in U$, choose $U_n \in U$ with $V_n \subseteq U_n$; if $V_n \not\subseteq U$ for no $U \in U$, set $U_n = U_0$. We claim $\{U_n : n \in \mathbb{N}\}$ is a subcover of $U$ (it is obviously countable). If $p \in X$, then there is a $U \in U$ with $p \in U$ since $U$ is a cover of $X$. Then there is an $n$ such that $p \in V_n$ and $V_n \subseteq U$ since $\{V_n\}$ is a base for the topology. Then $p \in U_n$. Thus $\cup_{n \in \mathbb{N}} U_n = X$. 

Note that this proof requires the Countable AC.

XI.1.3.6. In particular, $\mathbb{R}^n$ is Lindelöf for any $n$ by XI.1.2.8.(iii).

We can sharpen this result:

XI.1.3.7. Definition. A topological space $X$ is locally second countable if every $x \in X$ has an open neighborhood which is second countable in the relative topology.

Since every subspace of a second countable space is second countable, we may eliminate “open” from the definition without changing its meaning.
XI.1.3.8. **Proposition.** Let $X$ be a topological space. Then $X$ is second countable if and only if it is Lindelöf and locally second countable.

**Proof:** If $X$ is second countable, then it is obviously locally second countable, and Lindelöf by XI.1.3.5. Conversely, for each $x \in X$ let $U_x$ be an open neighborhood of $x$ which is second countable. Let $\{U_n : n \in \mathbb{N}\}$ be a countable subcover of $\{U_x : x \in X\}$. For each $n$, let $\mathcal{V}_n = \{V_{n1}, V_{n2}, \ldots\}$ be a countable base for the relative topology on $U_n$. Then $\bigcup_n \mathcal{V}_n$ is a countable base for the topology on $X$. (Note that each $V_{nk}$ is open in $U_n$, and $U_n$ is open in $X$, so $V_{nk}$ is open in $X$.)

This proof uses the Countable AC.

XI.1.4. **Relative Topology on a Subspace**

If $(X, \mathcal{T})$ is a topological space and $Y$ is a subset of $X$, then there is one, and only one, reasonable way to put a topology on $Y$ which is compatible with the topology on $X$:

XI.1.4.1. **Definition.** Let $(X, \mathcal{T})$ and $Y \subseteq X$. The relative topology, or subspace topology, on $Y$ from $\mathcal{T}$ is

$$\mathcal{T}_Y = \{U \cap Y : U \in \mathcal{T}\}$$

i.e. the open sets in $Y$ are the intersections with $Y$ of the open sets in $X$. $Y$ with this topology is called a **subspace** of $X$.

XI.1.4.2. **Proposition.** $\mathcal{T}_Y$ is a topology on $Y$.

Here is the proof as it appears in many standard references:

**Proof:** We have $\emptyset = \emptyset \cap Y$ and $Y = X \cap Y$, so $\emptyset, Y \in \mathcal{T}_Y$. If $V_1, \ldots, V_n \in \mathcal{T}_Y$, then, for each $k$, $V_k = U_k \cap Y$ for some $U_k \in \mathcal{T}$; so

$$V_1 \cap \cdots \cap V_n = (U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y \in \mathcal{T}_Y$$

since $U_1 \cap \cdots \cap U_n \in \mathcal{T}$. If $\{V_i : i \in I\}$ is a collection of sets in $\mathcal{T}_Y$, then, for each $i$, $V_i = U_i \cap Y$ for some $U_i \in \mathcal{T}$; then

$$\bigcup_{i \in I} V_i = \bigcup_{i \in I}(U_i \cap Y) = \left(\bigcup_{i \in I} U_i\right) \cap Y \in \mathcal{T}_Y$$

since $\bigcup_{i \in I} U_i \in \mathcal{T}$.

**XI.1.4.3.** Note that the AC is used in this proof to choose the $U_i$. (Most standard references seem to overlook this point! See XI.1.6.5.) But the result does not depend on the AC, because there is actually a systematic way to choose the $U_i$: for each $i$ let $U_i$ be the union of all $W \in \mathcal{T}$ such that $V = W \cap Y$ (I am indebted to S. Jabuka for this argument). There is also an alternate approach which does not use the AC (XI.1.4.7.).
XI.1.4.4. If $A \subseteq Y$, the phrase “$A$ is open” is ambiguous since also $A \subseteq X$: does it mean that $A$ is open in $Y$ (i.e. $A \in \mathcal{T}_Y$) or that $A$ is open in $X$ (i.e. that $A \in \mathcal{T}$)? To avoid ambiguity, we use the term relatively open for the first: $A \subseteq Y$ is relatively open in $Y$ if $A \in \mathcal{T}_Y$, i.e. $A = U \cap Y$ for some $U \in \mathcal{T}$. If $Y$ is open in $X$, i.e. $Y \in \mathcal{T}$, there is no ambiguity: a subset of $Y$ is relatively open if and only if it is open in $X$ (the converse holds also).

XI.1.4.5. Since $(X \setminus A) \cap Y = Y \setminus (A \cap Y)$ for any $A \subseteq X$, the closed sets in $(Y, \mathcal{T}_Y)$ are precisely the sets of the form $F \cap Y$ for $F$ closed in $X$. These sets are called the relatively closed subsets of $Y$. If $Y$ is closed in $X$, then a subset of $Y$ is relatively closed if and only if it is closed in $X$.

If $A \subseteq Y$, the phrase “closure of $A$” is also ambiguous: we could mean the closure $\bar{A}$ of $A$ in $(Y, \mathcal{T}_Y)$ or the closure $\bar{A}$ in $(X, \mathcal{T})$. But the situation is well behaved:

XI.1.4.6. Proposition. We have $\bar{A} = \bar{A} \cap Y$ for any $A \subseteq Y$.

Proof: Let $y \in \bar{A}$. Then for any $V \in \mathcal{T}_Y$ with $y \in V$ there is a $z \in A \cap V$. If $U \in \mathcal{T}$ with $y \in U$, set $V = U \cap Y$. If $z \in A \cap V$, then $z \in A \cap U$, so $y \in A$. Conversely, if $y \in A \cap Y$ and $V \in \mathcal{T}_Y$ with $y \in V$, then there is a $U \in \mathcal{T}$ with $V = U \cap Y$. Since $y \in \bar{A}$, there is a $z \in A \cap U$. Since $A \subseteq Y$, $z \in Y \cap U = V$. Thus $y \in \bar{A}$.

The set $\bar{A}$ is called the relative closure of $A$ in $Y$.

In light of this result, we can give an alternate construction of the relative topology:

XI.1.4.7. Proposition. Let $(X, \mathcal{T})$ be a topological space, and $Y \subseteq X$. For $A \subseteq Y$, define $\bar{A} = \bar{A} \cap Y$ (\bar{A} denotes the closure of $A$ in $X$). Then $A \mapsto \bar{A}$ is a Kuratowski closure operator (XI.2.4.1.) on $Y$, and the closed sets in the corresponding topology are precisely the sets of the form $F \cap Y$ for $F$ closed in $X$, i.e. is the relative topology on $Y$.

The proof is very simple and is left to the reader. It does not depend on the AC, and hence gives an alternate proof of XI.1.4.2. not using the AC.

The relative topology also has other properties which make it the only reasonable choice:

XI.1.4.8. Proposition. Let $(X, \mathcal{T})$ be a topological space, $Y \subseteq X$, and $\mathcal{T}_Y$ the relative topology on $Y$. Then

(i) If $(y_i)$ is a net in $Y$ and $y \in Y$, then $y_i \to y$ in $(Y, \mathcal{T}_Y)$ if and only if $y_i \to y$ in $(X, \mathcal{T})$.

(ii) If $Z$ is another topological space and $f : X \to Z$ is continuous, then $f|_Y : Y \to Z$ is also continuous.

(iii) The topology $\mathcal{T}_Y$ is the weakest topology on $Y$ making the inclusion map into $X$ continuous.

(iv) If $(Z, S)$ is a topological space and $f : Z \to Y$ is a function, then $f$ is continuous as a function from $(Z, S)$ to $(Y, \mathcal{T}_Y)$ if and only if it is continuous as a function from $(Z, S)$ to $(X, \mathcal{T})$.

XI.1.4.9. Note that (i) does not say that if a net in $Y$ converges in $X$, then it converges in $Y$. This will only be true if $Y$ is closed in $X$. 

1010
XI.1.4.10. Proposition. Let \((X, \mathcal{T})\) be a topological space, and \(Z \subseteq Y \subseteq X\). Then the relative topology on \(Z\) as a subspace of \((Y, \mathcal{T}_Y)\) is the same as the relative topology as a subspace of \((X, \mathcal{T})\), i.e. \((\mathcal{T}_Y)_Z = \mathcal{T}_Z\).

Bases

XI.1.4.11. If \((X, \mathcal{T})\) is a topological space and \((Y, \mathcal{T}_Y)\) is a subspace, and if \(B\) is a base for \(\mathcal{T}\), then it is easily checked that \(B_Y = \{B \cap Y : B \in B\}\) is a base for \(\mathcal{T}_Y\). Similarly, if \(y \in Y\) and \(B\) is a local base for \(\mathcal{T}\) at \(y\), then \(B_Y\) is a local base for \(\mathcal{T}_Y\) at \(y\). We thus obtain:

XI.1.4.12. Proposition. Let \((X, \mathcal{T})\) be a topological space and \((Y, \mathcal{T}_Y)\) a subspace. Then

(i) If \((X, \mathcal{T})\) is first countable, so is \((Y, \mathcal{T}_Y)\).

(ii) If \((X, \mathcal{T})\) is second countable, so is \((Y, \mathcal{T}_Y)\).

XI.1.4.13. It is not true that if \((X, \mathcal{T})\) is separable, then \((Y, \mathcal{T}_Y)\) is necessarily separable. There are many counterexamples, e.g. XI.7.8.10.(h), XI.7.8.14.(d), XI.11.12.4.

XI.1.5. Separated Union

There are many operations making new topological spaces from old ones. One of the simplest is separated union.

XI.1.5.1. Definition. Let \((X, \mathcal{T})\) be a topological space, and \(\{X_i : i \in I\}\) a collection of pairwise disjoint nonempty subsets of \(X\) whose union is \(X\) (i.e. a partition of \(X\)). Then \(X\) is the separated union of the \(X_i\) if a subset \(U\) of \(X\) is open (i.e. \(U \in \mathcal{T}\)) if and only if \(U \setminus X_i\) is relatively open in \(X_i\) for every \(i \in I\).

XI.1.5.2. If \(U \in \mathcal{T}\), then by definition \(U \cap X_i\) is relatively open in \(X_i\) for each \(i\). So the important condition is the converse, which is not automatic [consider a partition of \(X\) into singleton sets; then \(X\) is the separated union if and only if its topology is the discrete topology.] Some authors call a separated union a disjoint union, but the term “separated union” seems clearer and more descriptive.

A subset \(U\) of a topological space \(X\) is called clopen if both \(U\) and \(X \setminus U\) are open in \(X\). (“Clopen” means “both open and closed”; cf. (.)

XI.1.5.3. Proposition. Let \(\{X_i : i \in I\}\) be a partition of a topological space \(X\). Then \(X\) is the separated union of the \(X_i\) if and only if each \(X_i\) is a clopen subset of \(X\).

Proof: Suppose \(X\) is the separated union of the \(X_i\). For each fixed \(j \in I\), \(X_j \cap X_i\) is either \(X_j\) (if \(i = j\)) or \(\emptyset\) (if \(i \neq j\)). Thus \(X_j \cap X_i\) is relatively open in \(X_i\) for each \(i\); hence \(X_j\) is open in \(X\). So for any \(J \subseteq I\), \(\cup_{j \in J} X_j\) is open in \(X\). In particular, for each \(j\), \(X \setminus X_j = \cup_{i \neq j} X_i\) is open in \(X\), hence \(X_j\) is clopen in \(X\). More generally, if \(J \subseteq I\), \(X \setminus \cup_{j \in J} X_j = \cup_{i \in I \setminus J} X_i\) is open in \(X\); so \(\cup_{j \in J} X_j\) is clopen in \(X\).
Conversely, suppose each $X_i$ is clopen (and in particular open) in $X$. Let $U$ be a subset of $X$ such that $U \cap X_i$ is relatively open in $X_i$ for each $i$. Then, since $X_i$ is open in $X$, $U \cap X_i$ is open in $X$ for each $i$, and hence $U = \bigcup_{i \in I} (U \cap X_i)$ is open in $X$.

XI.1.5.4. We can reverse the process. Suppose we are given a family $\{(X_i, T_i) : i \in I\}$ of topological spaces. Form the set-theoretic disjoint union $X = \bigsqcup_{i \in I} X_i$(). Let $\mathcal{T}$ be the collection of all subsets $U$ of $X$ such that $U \cap X_i \in T_i$ for all $i$. It is easily checked that $\mathcal{T}$ is a topology on $X$ and $(X, \mathcal{T})$ is the separated union of the $(X_i, T_i)$.

XI.1.6. Exercises

XI.1.6.1. Give a logical critique of the following statement which can be found in many books:

“A metric space is a topological space.”

As part of your answer, you might consider, if $\rho$ and $\sigma$ are equivalent metrics on a set $X$, whether the spaces $(X, \rho)$ and $(X, \sigma)$ are the “same” or “different.”

XI.1.6.2. Show that if $\{T_\alpha\}$ is a family of topologies on a set $X$, then $\bigcap_\alpha T_\alpha$ is also a topology on $X$.

XI.1.6.3. Prove Proposition XI.1.2.5.

XI.1.6.4. (a) Show that the finite complement topology (XI.1.2.8.(iv)) on a set $X$ is a topology. Show that it is the discrete topology if and only if $X$ is finite.

(b) Let $X$ be a set. The countable complement topology on $X$ consists of $\emptyset$ and all subsets of $X$ with countable complement. Show that this is a topology on $X$, which is the discrete topology if and only if $X$ is countable. If $X$ is uncountable, show that the countable complement topology is not first countable.

XI.1.6.5. Explain why the following argument (which appears or is implicit in many books and does not use the AC) is not adequate to prove that the set $\mathcal{T}_Y$ defined in XI.1.4.1. is closed under arbitrary unions and is thus a topology:

“Let $\{U_i \cap Y : i \in I, U_i \in \mathcal{T}\}$ be an indexed collection of relatively open sets (sets in $\mathcal{T}_Y$). Then

$$\bigcup_{i \in I} (U_i \cap Y) = \left( \bigcup_{i \in I} U_i \right) \cap Y \in \mathcal{T}_Y$$

since $\bigcup_{i \in I} U_i \in \mathcal{T}$.”
XI.2. Closed Sets and Limit Points

XI.2.1. Closed Sets

Complements of open sets in a topological space are called closed sets:

XI.2.1.1. Definition. Let $(X, \mathcal{T})$ be a topological space. A subset $A$ of $X$ is closed in the topology $\mathcal{T}$, or $\mathcal{T}$-closed, if $A^c = X \setminus A \in \mathcal{T}$.

There is an important alternate description of closed sets (XI.2.3.6.).

XI.2.1.2. Examples. (i) A subset of $\mathbb{R}$ is closed in the usual topology if and only if it is closed in the sense of (i).
(ii) If $X$ has the discrete topology, then every subset of $X$ is closed.
(iii) If $X$ has the indiscrete topology, then the only closed subsets of $X$ are $\emptyset$ and $X$.

XI.2.1.3. It is important to note that being closed is not the opposite of being open: on the one hand, a subset of $X$ can be both open and closed (cf. XI.2.1.2.(ii)), and on the other hand a subset of $X$ need not be either open or closed – in fact typically “most” subsets of a topological space are neither open nor closed.

From de Morgan’s laws we immediately obtain the following properties of closed sets:

XI.2.1.4. Proposition. Let $(X, \mathcal{T})$ be a topological space, and $\mathcal{E}$ the collection of $\mathcal{T}$-closed sets. Then

(i) $\emptyset, X \in \mathcal{E}$.
(ii) $\mathcal{E}$ is closed under finite unions.
(iii) $\mathcal{E}$ is closed under arbitrary intersections.

XI.2.1.5. Conversely, if $X$ is a set and $\mathcal{E}$ is a collection of subsets of $X$ satisfying XI.2.1.4.(i)–(iii), then the complements of the sets in $\mathcal{E}$ form a topology $\mathcal{T}$ for which $\mathcal{E}$ is precisely the collection of closed sets. Thus a topology can be equally well defined by specifying the closed sets instead of the open sets.

XI.2.1.6. There is an asymmetry between unions and intersections for both open sets and closed sets. An arbitrary union of closed sets is not closed in general: for example, in $\mathbb{R}$ with its usual topology, points (singleton subsets) are closed, and every subset of $\mathbb{R}$ is a union of singleton sets. Correspondingly, an arbitrary intersection of open sets is not open in general. In fact, this asymmetry is the crucial underlying feature of topology, and it is what gives the subject its power and usefulness.

XI.2.2. Limit Points of a Set

One of the most fundamental notions in topology is limit point. Roughly speaking, a point $x$ in a topological space $X$ is a limit point of a set $A$ if there are points of $A$ “arbitrarily close” to $x$. More specifically:
XI.2.2.1. Definition. Let $(X, T)$ be a topological space, $A \subseteq X$, and $x \in X$. Then $x$ is a limit point of $A$ (in the topology $T$) if every open neighborhood of $x$ contains at least one point of $A$.

Note that if $x \in A$, then $x$ is trivially a limit point of $A$ since every neighborhood of $x$ contains at least one point of $A$, namely $x$ itself. But the importance of the notion of limit point is that limit points of $A$ are not necessarily in $A$. For example, 0 and 1 are limit points of the open interval $(0, 1)$ in $\mathbb{R}$ with its usual topology. Whether a point $x$ is a limit point of a subset $A$ can depend crucially on the topology used (if $x \notin A$); if there is a possibility of confusion we will use the terminology $T$-limit point.

There is a closely related but distinct notion:

XI.2.2.2. Definition. Let $(X, T)$ be a topological space, $A \subseteq X$, and $x \in X$. Then $x$ is an accumulation point of $A$ (in the topology $T$) if every deleted open neighborhood of $x$ contains at least one point of $A$, i.e. if every open neighborhood of $x$ contains at least one point of $A$ other than $x$.

Accumulation points are called cluster points by some authors, but for clarity we will only use the term “cluster point” in reference to sequences or nets (where the meaning is somewhat different, cf. XI.3.1.7.), not subspaces.

XI.2.2.3. Unlike for limit points, a point in $A$ is not necessarily an accumulation point of $A$; a point of $A$ which is not an accumulation point of $A$ is called an isolated point of $A$. A point of $X$ which is not in $A$ is a limit point of $A$ if and only if it is an accumulation point of $A$. There is some nonuniformity of terminology in the literature: some authors take “limit point” to mean what we call “accumulation point.” It will sometimes be convenient to have separate terms for the two notions, and for the purposes of this section limit points are more important. See XI.2.5. for a discussion of accumulation points and isolated points.

XI.2.3. Closure of a Set

XI.2.3.1. If $(X, T)$ is a topological space and $A \subseteq X$, then there is a smallest closed subset of $X$ containing $A$, namely the intersection of all closed subsets of $X$ containing $A$ (there is at least one such set, $X$; recall that any intersection of closed sets is closed).

XI.2.3.2. Definition. Let $(X, T)$ be a topological space and $A \subseteq X$. The smallest $T$-closed subset of $X$ containing $A$ is called the closure of $A$ in $X$ (or $T$-closure of $A$).

The standard notation for the closure of $A$ is $\overline{A}$; the notation $Cl(A)$ is also common. The closure of $A$ depends on $X$ and $T$; when this needs to be made explicit we will use the notation $Cl_X(A)$, $Cl_T(A)$, or $Cl_{(X, T)}(A)$.

Here are some elementary properties of closure:

XI.2.3.3. Proposition. Let $(X, T)$ be a topological space and $A, B \subseteq X$. Then

(i) $A \subseteq \overline{A}$.

(ii) $A$ is closed if and only if $A = \overline{A}$.

(iii) $\overline{A} = \overline{\overline{A}}$.
(iv) If \( A \subseteq B \), then \( \bar{A} \subseteq \bar{B} \).
(v) \( \bar{A \cup B} = \bar{A} \cup \bar{B} \).
(vi) \( \emptyset = \emptyset \) and \( \bar{X} = X \).

Proof: (i) and (ii) are trivial from the definition, and (iii) and (vi) follow immediately from (ii). (iv) is also immediate from the definition [\( B \) is a closed set containing \( A \)]. For (v), note that \( \bar{A} \cup \bar{B} \) is a union of two closed sets, hence closed, and contains \( A \cup B \), so \( \bar{A \cup B} \subseteq \bar{A} \cup \bar{B} \). On the other hand, \( A \subseteq A \cup B \), so \( \bar{A} \subseteq \bar{A \cup B} \) by (iv), and similarly \( \bar{B} \subseteq \bar{A \cup B} \). 

XI.2.3.4. If \( A, B \subseteq X \), it is not generally true that \( \bar{A \cap B} = \bar{A} \cap \bar{B} \). For example, if \( A = (0, 1) \) and \( B = (1, 2) \) in \( \mathbb{R} \) (with its usual topology), then \( \bar{A} \cap \bar{B} = \{ 1 \} \). But \( A \cap B = \emptyset \), so \( \bar{A \cap B} = \emptyset \).

However, we do always have \( \bar{A \cap B} \subseteq \bar{A} \cap \bar{B} \), since the set on the right is a closed set containing \( A \cap B \). More generally, for any collection \( \{ A_i : i \in I \} \) of subsets of \( X \), we have \( \bigcap_{i \in I} \bar{A_i} \subseteq \bigcap_{i \in I} \bar{A_i} \) by the same argument. A similar argument shows that \( \bigcup_{i \in I} \bar{A_i} \subseteq \bigcup_{i \in I} \bar{A_i} \), but we do not have equality in general if \( I \) is infinite.

The definition of closure is precise but often hard to apply directly. There is a more generally useful characterization of closure:

XI.2.3.5. Theorem. Let \((X, T)\) be a topological space and \( A \subseteq X \). Then the closure of \( A \) is exactly the set of all limit points (XI.2.2.1.) of \( A \).

Proof: Let \( U \) be the complement of \( \bar{A} \); then \( U \) is open. Suppose \( x \) is a limit point of \( A \). If \( x \notin \bar{A} \), then \( U \) is an open neighborhood of \( x \) containing no points of \( A \), contradicting that \( x \) is a limit point of \( A \). Thus \( x \in \bar{A} \).

Now suppose \( x \) is not a limit point of \( A \). Then there is an open neighborhood \( V \) of \( x \) containing no point of \( A \). So \( V^c \) is a closed set with \( A \subseteq V^c \), and hence \( \bar{A} \subseteq V^c \), \( x \notin \bar{A} \).

XI.2.3.6. Corollary. Let \((X, T)\) be a topological space. A subset of \( X \) is closed if and only if it contains all its limit points.

Dense Sets and Separability

XI.2.3.7. Definition. Let \((X, T)\) be a topological space, and \( D \subseteq X \). Then \( D \) is dense in \( X \) if \( \bar{D} = X \).

XI.2.3.8. Proposition. Let \( X \) be a topological space, and \( D \subseteq X \). Then \( D \) is dense in \( X \) if and only if every nonempty open subset of \( X \) contains a point of \( D \).

An intersection of two dense subsets is not necessarily dense, or even nonempty: both \( \mathbb{Q} \) and \( \mathbb{J} \) are dense in \( \mathbb{R} \). But if at least one of the sets is open, the intersection is dense:
XI.2.3.9. **Proposition.** Let $X$ be a topological space, and $U$ and $V$ dense subsets of $X$. If $U$ is open, then $U \cap V$ is dense in $X$.

**Proof:** Let $W$ be a nonempty open set in $X$. Then $U \cap W$ is an open set in $X$, which is nonempty since $U$ is dense. Then $W \cap (U \cap V) = (U \cap W) \cap V$ is nonempty since $V$ is dense. \[\Box\]

XI.2.3.10. **Corollary.** Let $X$ be a topological space. The intersection of any finite number of dense open subsets of $X$ is a dense open subset of $X$.

In some, but not all, topological spaces, an intersection of a countable number of dense open sets is always dense (but not necessarily open). See XI.2.11.

Some topological spaces have “small” dense subsets, which can cause better behavior:

XI.2.3.11. **Definition.** A topological space is separable if it has a countable dense subset.

Note that separability in this sense has nothing to do with the separation axioms of XI.71. The coincidence of terminology is regrettable, but well established.

XI.2.3.12. **Examples.** (i) If $X$ has the indiscrete topology, then any nonempty subset of $X$ is dense. Thus $X$ is separable.

(ii) If $X$ has the discrete topology, no proper subset of $X$ is dense in $X$. Thus $X$ is separable if and only if it is countable. In fact, a countable set with any topology is separable.

(iii) If $X$ is an infinite set with the finite complement topology, any infinite subset of $X$ is dense. Thus $X$ is separable if and only if it contains a countably infinite subset (this will always be true if the Countable AC is assumed).

(iv) If $\mathbb{R}$ has the usual topology, $\mathbb{Q}$ is a countable dense subset, so $\mathbb{R}$ is separable. More generally, the points of $\mathbb{R}^n$ with rational coordinates form a countable dense set (for the usual topology), so $\mathbb{R}^n$ is separable.

XI.2.3.13. Separability seems to be related to first or (especially) second countability. There are some connections, as the next two results show. But a topological space can be separable without being first countable (an uncountable set with the finite complement topology; cf. also () and ()), first countable without being separable (an uncountable set with the discrete topology), and both first countable and separable without being second countable ()

XI.2.3.14. **Proposition.** A second countable topological space is separable.

**Proof:** If $\{U_n : n \in \mathbb{N}\}$ is a countable base for the topology consisting of nonempty sets, choose $x_n \in U_n$ for each $n$. Then $\{x_n : n \in \mathbb{N}\}$ is a countable dense set in $X$. \[\Box\]

This proof requires the Countable AC, and the result cannot be proved in ZF.
XI.2.3.15. **Theorem.** A metrizable topological space is second countable if and only if it is separable.

**Proof:** Let \((X, \rho)\) be a metric space. If \(X\) is second countable, it is separable by XI.2.3.14.. Conversely, suppose \(X\) has a countable dense set \(D = \{x_n : n \in \mathbb{N}\}\). Let \(\mathcal{B}\) be the collection of all open balls of the form \(B_r(x_n)\), where \(x_n \in D\) and \(r \in \mathbb{Q}\). Then \(\mathcal{B}\) is a countable collection of open sets in \(X\). We claim \(\mathcal{B}\) is a base for the topology. To show this, we need only show that if \(x \in X\) and \(\epsilon > 0\), there is an \(n\) and \(r\) such that \(x \in B_r(x_n)\) and \(B_r(x_n) \subseteq B_\epsilon(x)\). Choose \(n\) such that \(\rho(x, x_n) < \frac{\epsilon}{3}\) and \(r \in \mathbb{Q}\) such that \(\frac{\epsilon}{3} < r < \frac{2\epsilon}{3}\). Then \(x \in B_r(x_n)\), and \(B_r(x_n) \subseteq B_\epsilon(x)\) by the triangle inequality.

The implication proved in this proof does not require any form of Choice, since only finitely many choices must be made.

XI.2.3.16. More generally, we can define the density \(d(X)\) of a topological space \(X\) to be the minimum cardinality of a dense subset of \(X\) (this is only well defined in general if the AC is assumed). \(X\) is separable if and only if \(d(X) \leq \aleph_0\). A slight variation of the proof of XI.2.3.14. shows that \(d(X) \leq w(X)\) (XI.1.2.11.) for any \(X\). We can have \(d(X) < w(X)\) (a separable space is not necessarily second countable), but if \(X\) is regular we have \(w(X) \leq 2^{d(X)}\) (XI.7.8.6.).

**Interior of a Set**

There is a dual notion to closure:

XI.2.3.17. **Definition.** Let \((X, \mathcal{T})\) be a topological space, and \(A \subseteq X\). The interior of \(A\) (in \(X\), in the topology \(\mathcal{T}\)) is the largest open (\(\mathcal{T}\)-open) subset of \(X\) contained in \(A\).

There is a largest such open set, the union of all open sets contained in \(A\); there is always one such set, the empty set. The interior of \(A\) can very well be empty. The standard notation for the interior of \(A\) is \(A^\circ\); the notation \(\text{Int}(A)\) is also used.

XI.2.3.18. Note that the interior of a subset \(A\) of \(X\) is a relative notion, not intrinsic to \(A\) (if \(A\) is given the relative topology, the interior of \(A\) with respect to \(A\) is all of \(A\)! The interior of \(A\) is also dependent on the topology \(\mathcal{T}\), although the notation \(A^\circ\) does not reflect this; the notation \(\text{Int}(A)\) is thus more flexible, since it can be embellished as \(\text{Int}_X(A)\), \(\text{Int}_\mathcal{T}(A)\), or \(\text{Int}_{(X, \mathcal{T})}(A)\) when necessary.

XI.2.3.19. **Example.** Let \(X = \mathbb{R}\) with its usual topology, \(A = [0, 1]\). Then \(A^\circ = (0, 1)\). If, however, \(X = \mathbb{R}^2\) and \(A\) is the interval \([0, 1]\) on the \(x\)-axis, then \(A^\circ = \emptyset\).

The relation between interior and closure is:

XI.2.3.20. **Proposition.** Let \(X\) be a topological space, and \(A \subseteq X\). Then \((A^\circ)^c = (A^c)^-\), i.e. \(A^\circ = [(A^c)^-]^c\).

From this, de Morgan’s laws, and XI.2.3.3. we obtain the basic properties of interior:
XI.2.3.21. \textbf{Proposition.} Let \((X, T)\) be a topological space and \(A, B \subseteq X\). Then

(i) \(A^o \subseteq A\).

(ii) \(A\) is open if and only if \(A^o = A\).

(iii) \((A^o)^o = A^o\).

(iv) If \(A \subseteq B\), then \(A^o \subseteq B^o\).

(v) \((A \cap B)^o = A^o \cap B^o\).

(vi) \(\emptyset^o = \emptyset\) and \(X^o = X\).

XI.2.3.22. We always have \(A^o \cup B^o \subseteq (A \cup B)^o\), but equality does not hold in general. A dramatic example is to take \(X = \mathbb{R}\), \(A = \mathbb{Q}\), and \(B = \mathbb{J}\); then \(A^o = B^o = \emptyset\) but \((A \cup B)^o = X\).

\textbf{Boundary of a Subset}

XI.2.4. \textbf{Specification of a Topology via Closure}

Instead of defining a topology by specifying the open sets, we can define it by specifying the closure of each set (this will determine the closed sets). This procedure can be axiomatized by a list of properties of the closure operation. This approach to topology is generally attributed to Kuratowski, although he himself in \((\text{Kur66})\) credited the procedure to F. Riesz.

XI.2.4.1. \textbf{Definition.} Let \(X\) be a set. A \textit{Kuratowski closure operation} on \(X\) is an assignment to each subset \(A\) of \(X\) a subset \(\overline{A}\) of \(X\), called the \textit{closure} of \(A\), with the following properties:

(i) \(\overline{\emptyset} = \emptyset\).

(ii) \(A \subseteq \overline{A}\) for all \(A \subseteq X\).

(iii) \(\overline{\overline{A}} = \overline{A}\) for all \(A \subseteq X\).

(iv) \(\overline{A \cup B} = \overline{A} \cup \overline{B}\) for all \(A, B \subseteq X\).

XI.2.4.2. \textbf{Examples.} (i) Let \((X, T)\) be a topological space. For \(A \subseteq X\), let \(\overline{A}\) be the closure of \(A\) in the topological sense (XI.2.3.2). Then \(A \mapsto \overline{A}\) is a Kuratowski closure operation (XI.2.3.3). We will show the converse: if \(A \mapsto \overline{A}\) is a Kuratowski closure operation on \(X\), then there is a unique topology on \(X\) such that the closure operation is closure with respect to the topology (XI.2.4.4).

(ii) If \(X\) is a nonempty set, set \(\overline{A} = X\) for every \(A \subseteq X\). Then this operation satisfies XI.2.4.1.(ii), (iii), and (iv), but not (i).

(iii) If \(X\) is a nonempty set, set \(\overline{A} = \emptyset\) for all \(A \subseteq X\). Then this operation satisfies XI.2.4.1.(i), (iii), and (iv), but not (ii).

(iv) Let \(X = \{1, 2, 3\}\). Set \(\overline{\emptyset} = \emptyset\), \(\overline{\{1\}} = \{1, 2\}\), and \(\overline{A} = X\) for every other \(A\). Then this operation satisfies XI.2.4.1.(i), (ii), and (iv), but not (iii).
(v) Let \(X = \{1, 2, 3\}\). Set \(\emptyset = \emptyset, \{1, 2\} = \{1, 2\}\), and \(A = X\) for every other \(A\). Then this operation satisfies XI.2.4.1.(i), (ii), and (iii), but not (iv).

(ii)–(v) show that the axioms in XI.2.4.1. are independent, and hence all are necessary to specify a topology.

The most powerful of the axioms is (iv):

**XI.2.4.3. Proposition.** Let \(A \mapsto \bar{A}\) be an operation on \(X\) satisfying (iv). Then

(i) If \(A \subseteq B \subseteq X\), then \(\bar{A} \subseteq \bar{B}\).

(ii) If \(\{A_i : i \in I\}\) is a collection of subsets of \(X\), then \(\bigcup_{i \in I} \bar{A}_i \subseteq \bigcup_{i \in I} \bar{A}_i\).

(iii) If \(\{A_i : i \in I\}\) is a collection of subsets of \(X\), then \(\bigcap_{i \in I} \bar{A}_i \subseteq \bigcap_{i \in I} \bar{A}_i\).

**Proof:** (i): If \(A \subseteq B\), then \(\bar{B} = \bar{A} \cup B = \bar{A} \cup \bar{B}\) which implies that \(\bar{A} \subseteq \bar{B}\).

(ii): For each \(j \in I\), \(A_j \subseteq \bigcup_{i \in I} A_i\), so \(\bar{A}_j \subseteq \bigcup_{i \in I} \bar{A}_i\) by (i).

(iii): For each \(j \in I\), \(\bigcap_{i \in I} A_i \subseteq A_j\), so \(\bigcap_{i \in I} \bar{A}_i \subseteq \bar{A}_j\) by (i).

**XI.2.4.4. Theorem.** Let \(A \mapsto \bar{A}\) be a Kuratowski closure operation on a set \(X\). Then there is a unique topology \(T\) on \(X\) such that the closure operation is precisely closure in the topology \(T\).

**Proof:** If there is such a topology, then the closed sets for the topology must be precisely the sets \(A\) for which \(A = \bar{A}\) by axiom (ii), so there is at most one such topology. So we must show that the collection of such sets defines a topology (i.e. the complements of such sets form a topology on \(X\)). Call such sets closed.

By XI.2.1.5., we must show that \(\emptyset\) and \(X\) are closed, and that the collection of closed sets is closed under finite unions and arbitrary intersections.

We have \(\emptyset = \emptyset\) by axiom (i), and \(X \subseteq \bar{X}\) by axiom (ii), hence \(X = \bar{X}\).

Suppose \(A = \bar{A}\) and \(B = \bar{B}\). Then \(\overline{A \cup B} = \overline{A} \cup \overline{B} = \overline{A} \cup \overline{B}\) by axiom (iv), so \(A \cup B\) is closed. By iteration (or induction), the collection of closed sets is closed under finite unions.

Suppose \(A_i = \overline{A}_i\) for each \(i \in I\). Then \(\bigcap_{i \in I} \overline{A}_i \subseteq \bigcap_{i \in I} \bar{A}_i = \bigcap_{i \in I} A_i\) by XI.2.4.3.(iii), and the opposite inclusion follows from axiom (ii). Thus the collection of closed sets is closed under arbitrary intersections.

**XI.2.4.5.** Suppose \(A \mapsto \bar{A}\) is an operation on a set \(X\) satisfying axiom (iv) and also

(v) \(\{p\} = \{p\}\) for every \(p \in X\).

Then the operation satisfies axiom (ii) by XI.2.4.3.(ii) since any subset of \(X\) is a union of singleton subsets. If \(X\) has two distinct points \(p\) and \(q\), then \(\emptyset \subseteq \{p\} \subseteq \{p\}\) by XI.2.4.3.(i), and similarly \(\emptyset \subseteq \{q\}\), so the operation also satisfies axiom (i). Thus if it also satisfies axiom (iii), it is a Kuratowski closure operation. A Kuratowski closure operation satisfies axiom (v) if and only if the corresponding topology is \(T_1()\).
XI.2.5. Isolated Points and Derived Sets

We now look at isolated points and accumulation points (XI.2.2.2.) in more detail. In order to avoid some cardinal-theoretic difficulties, we will assume the AC.

XI.2.5.1. A point \( p \) in a topological space \( X \) is an isolated point of \( X \) if it is not in the closure of \( X \setminus \{p\} \), i.e. if \( \{p\} \) is an open set in \( X \). (In some references, an isolated point is also required to be closed, which is automatic in a \( T_1 \) space; but we will not make this restriction.) A topological space is dense-in-itself if it contains no isolated points. A topological space \( X \) is scattered if every nonempty subset of \( X \) contains an isolated point in the relative topology, i.e. if no (nonempty) subset of \( X \) is dense-in-itself.

Any subset of a scattered space is scattered. A subset of a space which is dense-in-itself is not necessarily dense-in-itself (it could even be scattered); but an open subset of a space which is dense-in-itself must also be dense-in-itself.

XI.2.5.2. If \( A \) is a subset of a topological space \( X \), denote by \( A' \) the set of accumulation points of \( A \) in \( X \). \( A' \) is called the derived set of \( A \) in \( X \). A subset \( A \) of \( X \) is closed if and only if \( A' \subseteq A \). We say \( A \) is a perfect subset of \( X \) if \( A' = A \). \( A \) is perfect in \( X \) if and only if it is closed and dense-in-itself.

Note that the properties of being scattered or dense-in-itself are intrinsic properties of a space (or subspace), but being perfect is a relative property.

The next theorem, the main result of this section, is often stated in references in much less generality than necessary (the proof does not even become simpler for these special cases!) Recall (XI.1.2.11.) that the weight \( w(X) \) of a topological space \( X \) is the smallest cardinality of a base for the topology of \( X \).

XI.2.5.3. Theorem. [General Cantor-Bendixson Theorem] Let \( X \) be a topological space. Then there is a unique decomposition of \( X \) as a disjoint union of a perfect subset \( X^p \) and a (necessarily open) scattered subset \( X^s \). We have \( \text{card}(X^s) \leq \min(\text{card}(X), w(X)) \).

Note that \( X^p \) or \( X^s \) can be empty: \( X^s = \emptyset \) if (and only if) \( X \) is dense-in-itself, and \( X^p = \emptyset \) if (and only if) \( X \) is scattered.

Proof: The proof will actually show more than the stated result; the additional conclusions will be formalized at the end of the proof.

Define a decreasing collection of closed subsets of \( X \) by transfinite induction as follows. Set \( X^{(0)} = X \). For any ordinal \( \alpha \), if \( X^{(\alpha)} \) has been defined, set \( X^{(\alpha+1)} = (X^{(\alpha)})' \). If \( \alpha \) is a limit ordinal, set \( X^{(\alpha)} = \bigcap_{\beta < \alpha} X^{(\beta)} \). We have \( X^{(\alpha)} \subseteq X^{(\beta)} \) if \( \beta < \alpha \). Note that \( X^{(1)} \) is \( X \) with the isolated points removed; since each isolated point is open, \( X^{(1)} \) is closed in \( X \). By the same argument, \( X^{(\alpha+1)} \) is relatively closed in \( X^{(\alpha)} \); thus by transfinite induction \( X^{(\alpha)} \) is a closed subset of \( X \) for all \( \alpha \).

The \( X^{(\alpha)} \) cannot continue to strictly decrease (it will stabilize before \( \text{card}(\alpha) > \text{card}(X) \)); there must be a smallest \( \alpha \) for which \( X^{(\alpha+1)} = X^{(\alpha)} \). If \( X^p = X^{(\alpha)} \) for this \( \alpha \), then \( X^p \) is a perfect subset of \( X \).

Set \( X^s = X \setminus X^p \); then \( X^s \) is open in \( X \). Suppose \( Y \) is a subset of \( X \) which is dense-in-itself. If \( Y \cap X^s \neq \emptyset \), then there is a largest \( \beta \) such that \( Y \subseteq X^{(\beta)} \), and \( \beta < \alpha \). Thus there is a \( y \in Y \cap (X^{(\beta)} \setminus X^{(\beta+1)}) \). But then \( y \) is an isolated point of \( X^{(\beta)} \) and hence an isolated point of \( Y \), a contradiction. Thus every dense-in-itself subset of \( X \) is contained in \( X^p \), and \( X^s \) is scattered.
For uniqueness, suppose $X = Z \cup W$, where $Z$ is perfect and $W$ scattered. We have seen that $Z \subseteq X^p$. If $X^p \cap W \neq \emptyset$, it is a relatively open subset of $X^p$ which contains an isolated point, a contradiction.

For the cardinality assertion, let $\mathcal{B}$ be a base for the topology of $X$ with $\text{card}(\mathcal{B}) = w(X)$. First note that if $p$ is an isolated point of $X$, then $\{p\} \in \mathcal{B}$. By a similar argument, for each $x \in X^{(\alpha)} \setminus X^{(\alpha+1)}$ there is an open set $U_x \in \mathcal{B}$ with $U_x \cap X^{(\alpha)} = \{x\}$. It is clear that $x \mapsto U_x$ is an injective map from $X^s$ to $\mathcal{B}$, so $\text{card}(X^s) \leq w(X)$. The assertion $\text{card}(X^s) \leq \text{card}(X)$ is trivial.

The next corollary, or special cases of it, is usually called the “Cantor-Bendixson Theorem.” Actually Cantor and Bendixson (who worked independently) only stated the result for subsets of $\mathbb{R}^n$ since the formalism of general topology had not yet been developed.

**XI.2.5.4. Corollary. [Cantor-Bendixson Theorem]** Let $X$ be a second countable topological space. Then $X$ can be written uniquely as a disjoint union of a perfect subset $X^p$ and a (necessarily open) scattered subset $X^s$. In addition, $X^s$ is countable.

**XI.2.5.5.** The set $X^{(\alpha)}$ is called the $\alpha$'th derived set of $X$, and the smallest $\alpha$ for which $X^{(\alpha+1)} = X^{(\alpha)} = X^p$ is called the (Cantor-Bendixson) height $ht(X)$ of $X$. The chain of derived subsets and the height are homeomorphism invariants of $X$. The height is an ordinal invariant of $X$, not a cardinal invariant; if $X$ is second countable, then $ht(X)$ is a countable ordinal, but it is not always a finite ordinal or even $\omega$. In fact, every countable ordinal $\alpha$ occurs as the height of a compact subset $X_\alpha$ of $\mathbb{R}$. More generally, every ordinal occurs as the height of a compact Hausdorff space.

**XI.2.6. Exercises**

**XI.2.6.1.** (cf. [Dug78, III.5]) (a) Define a notion of “interior operation” on a set in analogy with the definition of a Kuratowski closure operation, using the properties from XI.2.3.21., and prove the analog of XI.2.4.4. that a topology can be specified by such an interior operation.

(b) Do the same for “boundary operation,” using the properties of ().

(c) Do the same for the “derived set” operation $A \mapsto A'$ (XI.2.5.2.).

**XI.2.6.2.** (a) Show that if $X$ is a $T_1$ topological space and $A \subseteq X$, then $A'$ is closed in $X$.

(b) Show by example that if $X$ is just $T_0$, $A'$ need not be closed in $X$.  

1021
XI.3. Convergence

One of the most important uses of a metric is to define limits of sequences. Limits of sequences can be defined nicely in any topological space, although not all properties of limits of sequences in metric spaces carry over to this case. And it turns out that for (most – cf. XI.3.5.1.) topological spaces which are not first countable, the theory of limits of sequences is not adequate to fully describe convergence and continuity in the space, and one must use a generalized version of sequences known as nets. An alternate equivalent point of view with some technical advantages uses filters in place of nets.

XI.3.1. Limits of Sequences

Limits of sequences in topological spaces are defined just as in metric spaces, replacing open balls by open neighborhoods:

XI.3.1.1. Definition. Let \((X, T)\) be a topological space, and let \((x_n)\) be a sequence in \(X\), and \(x \in X\). Then \((x_n)\) converges to \(x\) in \((X, T)\), denoted \(x_n \to x\) or \(\lim_{n \to \infty} x_n = x\), if for every open neighborhood \(U\) of \(x\) there is an \(N\) such that \(x_n \in U\) for all \(n \geq N\). (We say that \(x_n\) is eventually in \(U\).)

If there is any possible ambiguity about which topology is used on \(X\), we write \(x_n \to_T x\) or \(\lim_{n \to \infty} x_n = x\).

XI.3.1.2. The word “open” can be deleted from this definition without changing its meaning, since every neighborhood of \(x\) contains an open neighborhood. Similarly, the existence of an \(N\) need only be shown for \(U\)’s in a fixed local base for the topology at \(x\).

XI.3.1.3. Unlike in metric spaces, limits of sequences need not be unique: for example, in a space with the indiscrete topology every sequence converges to every point. This undesirable situation does not occur in Hausdorff spaces (XI.7.4.1.).

Subsequences and Tails

Subsequences and tails of sequences are defined just as in \((X, T)\). Convergence of a sequence depends only on the tails of the sequence, and implies convergence of each subsequence, just as in the metric case (the proof is essentially identical):

XI.3.1.4. Proposition. Let \(X\) be a topological space, and \((x_n)\) a sequence in \(X\), \(x \in X\).

(i) \(x_n \to x\) if and only if some tail of \((x_n)\) converges to \(x\).

(ii) If \(x_n \to x\), then every subsequence of \((x_n)\) converges to \(x\). In particular, every tail of \((x_n)\) converges to \(x\).

There is a sort of converse to (ii), which is almost harder to state than to prove:
XI.3.1.5. PROPOSITION. Let $X$ be a topological space, and $(x_n)$ a sequence in $X$, $x \in X$. If $(x_n)$ does not converge to $x$, then there is a subsequence of $(x_n)$, no subsequence of which converges to $x$.

**Proof:** If $(x_n)$ does not converge to $x$, then there is an open neighborhood $U$ of $x$ such that $x_n \notin U$ for infinitely many $n$. These $x_n$ form a subsequence with no terms in $U$, so no subsequence of it can converge to $x$. The result is often stated in the contrapositive:

XI.3.1.6. COROLLARY. Let $X$ be a topological space, and $(x_n)$ a sequence in $X$, $x \in X$. If every subsequence of $(x_n)$ has a subsequence which converges to $x$, then $x_n \not\rightarrow x$.

Cluster Points of Sequences

XI.3.1.7. DEFINITION. Let $X$ be a topological space, and $(x_n)$ a sequence in $X$, $x \in X$. Then $x$ is a cluster point of $(x_n)$ if every open neighborhood $U$ of $x$ contains $x_n$ for infinitely many $n$ (we say $x_n$ is frequently in $U$).

The proof of the next Proposition is essentially identical to the proof of ().

XI.3.1.8. PROPOSITION. Let $X$ be a topological space, and $(x_n)$ a sequence in $X$. For each $n$, let $F_n$ be the closure in $X$ of $\{x_k : k \geq n\}$. Then $\bigcap_n F_n$ is exactly the set of cluster points of $(x_n)$. In particular, the set of cluster points of a sequence is a closed set.

If $x_n \rightarrow x$, or more generally if there is a subsequence of $(x_n)$ which converges to $x$, then $x$ is a cluster point of $(x_n)$. The converse is not true in general (), but is true if the topological space is first countable:

XI.3.1.9. PROPOSITION. Let $X$ be a topological space, $(x_n)$ a sequence in $X$, and $x \in X$. If $x$ is a cluster point of $(x_n)$, and there is a countable local base for the topology at $x$ (in particular, if $X$ is first countable), then there is a subsequence of $(x_n)$ which converges to $x$.

**Proof:** Let $(U_n)$ be a decreasing sequence of open neighborhoods of $x$ forming a local base (). Let $x_{k_1} \in U_1$, and inductively choose $k_{n+1} > k_n$ with $x_{k_{n+1}} \in U_{n+1}$; this is possible since $x_k \in U_{n+1}$ for infinitely many $k$. (For definiteness and to avoid use of the Countable AC, choose $k_{n+1}$ to be the smallest $k > k_n$ for which $x_k \in U_{n+1}$.) Then $(x_{k_n})$ is a subsequence of $(x_n)$, and $x_{k_n} \in U_m$ for $n \geq m$. If $V$ is any neighborhood of $x$, then $U_m \subseteq V$ for some $m$, so $x_{k_n} \in V$ for all $n \geq m$. Thus $x_{k_n} \rightarrow x$.

XI.3.1.10. If $A$ is a subset of a topological space $X$, $x \in X$, and there is a sequence $(x_n)$ in $A$ with $x_n \rightarrow x$, then $x$ is obviously a limit point of $A$, so $x \in \bar{A}$. The converse is not true in general (), but is true if the topological space is first countable:
XI.3.1.11. Proposition. Let $A$ be a subset of a topological space $X$ and $x \in X$. If there is a countable local base for the topology at $x$ (in particular, if $X$ is first countable), then $x \in A$ if and only if there is a sequence $(x_n)$ of points of $A$ which converges to $x$.

Proof: The proof is very similar to the proof of XI.3.1.9. One direction is noted above, and does not require a countable local base. Conversely, suppose $x \in A$. Let $(U_n)$ be a decreasing sequence of open neighborhoods of $x$ forming a local base. Then $A \cap U_n$ is nonempty for every $n$. Choose $x_n \in A \cap U_n$ for each $n$. If $V$ is any neighborhood of $x$, then $U_m \subseteq V$ for some $m$, so $x_n \in V$ for all $n \geq m$. Thus $x_n \to x$.

Note that the Countable AC is needed for this proof.

XI.3.1.12. Thus, if $X$ is a first countable topological space, then (assuming the Countable AC) convergence of sequences tells the whole story about the topology.

Iterated Limits

There is another property of convergent sequences in first countable topological spaces:

XI.3.1.13. Proposition. Let $(X, T)$ be a topological space and $\{x_{k,n} : k, n \in \mathbb{N}\}$ be a doubly indexed set of points in $X$. Suppose $\lim_{n \to \infty} x_{k,n} = x_k$ for each $k$ and $\lim_{k \to \infty} x_k = x$. If $x$ has a countable local base (in particular, if $X$ is first countable), then there are strictly increasing sequences $(k_m), (n_m) \in \mathbb{N}$ such that $\lim_{m \to \infty} x_{k_m,n_m} = x$.

The sequence $(x_{k_m,n_m})$ is a sort of “diagonal” sequence, and this property of limits is sometimes called the diagonal property.

Proof: Let $(U_m)$ be a countable decreasing local neighborhood base at $x$. Then for each $m$ there is a $K_m$ such that $x_k \in U_m$ for all $k \geq K_m$. Then for each $k \geq K_m$ there is an $N_{k,m}$ such that $x_{k,n} \in U_m$ for all $n \geq N_{k,m}$. Inductively choose $k_1 = n_1 = 1$ and $k_m = \max(k_{m-1} + 1, K_m)$ and $n_m = \max(n_{m-1} + 1, N_{k_m,m})$ for $m > 1$. It is routine to check that $x_{k_m,n_m} \to x$.

XI.3.1.14. This result can fail if $x$ does not have a countable local base (). But the net version holds in general ()

XI.3.2. Nets and Limits

In a topological space which is not first countable, sequences are generally inadequate to completely describe the behavior of the topology and convergence. But there is a deceptively simple way to generalize sequences, which in complete analogy does fully describe the topology. These generalized sequences are called nets. On the surface, the results of the preceding subsection work for general topological spaces by simply substituting “net” for “sequence” and “subnet” for “subsequence”; however, below the surface there are some subtleties involved, especially with the notion of a subnet.
XI.3.2.1. Definition. Let $X$ be a topological space (or just a set). A net in $X$ is a function from a directed set $(I, \leq)$ to $X$. The set $I$ is called the index set of the net.

A sequence is just a net whose index set is $\mathbb{N}$. If $\phi : I \to X$ is a net with index set $I$, we usually denote the net by $(x_i)_{i \in I}$ or just by $(x_i)$ if $I$ is understood, where $x_i = \phi(i)$. This notation is a simple extension of sequence notation. However, this notation does not reflect the direction on $I$, which is part of the definition of the net.

XI.3.2.2. Note that the index set of a net does not have to be partially ordered. One can in principle restrict to partially ordered index sets in the theory of nets (cf. XI.3.2.4., XI.3.2.5., II.4.4.11.). However, it is often convenient to use nets whose index sets are not partially ordered.

XI.3.2.3. Definition. Let $X$ be a topological space, $(x_i)$ a net in $X$, and $x \in X$. Then $(x_i)$ converges to $x$, denoted $x_i \to x$ or $\lim_{i \to x} x_i = x$, if for every open neighborhood $U$ of $x$ there is an $i_0 \in I$ such that $x_i \in U$ whenever $i_0 \leq i$. (We say the net $(x_i)$ is eventually in $U$.)

As for sequences, the word "open" can be deleted, and it suffices to let the $U$ range over a local base at $x$.

XI.3.2.4. If the index set of a net is not properly directed (i.e. has a largest element), convergence of the net is pretty trivial: if there is exactly one largest element $i_1$ in $I$, then the net automatically converges to $x_{i_1}$. If there is more than one largest element in $I$, then the net rarely converges (only if every neighborhood of $x_i$ for each largest element $i$ contains $x_j$ for all the other largest $j$; this cannot happen in a $T_0$-space unless $x_i = x_j$ for all largest $i$ and $j$, in which case the net converges to this element). The theory of convergence of nets is only interesting if the index set is properly directed.

The convergence of a net depends on the ordering (direction) of $I$ only up to equivalence, i.e. only on the "cofinal behavior" of the direction:

XI.3.2.5. Proposition. Let $(I, \leq)$ be a directed set, $X$ a topological space, and $(x_i)_{i \in I}$ a net in $X$ with index set $I$. Let $\preceq$ be another direction on $I$ which is equivalent to $\leq$. If $x \in X$, then $x_i \to x$ using the direction $\leq$ on $I$ if and only if $x_i \to x$ using the direction $\preceq$ on $I$.

This is obvious using the definition of equivalence of directions ().

Subnets

The definition of a subnet is not an obvious one. Experience has shown that the following definition is appropriate, and flexible enough to be useful:

XI.3.2.6. Definition. Let $X$ be a topological space, $(I, \leq)$ a directed set, and $\phi : I \to X$ a net in $X$. A subnet of $\phi$ is a net $\phi \circ f$, where $f$ is a directed function from a directed set $(J, \preceq)$ to $(I, \leq)$. The index set of the subnet is $(J, \preceq)$.

Thus the index set of a subnet of a net is not necessarily the same as the index set of the net.
Proof: We give the proof of (ii), leaving the simpler proof of (i) to the reader. Suppose $(x_{f(j)})_{j \in J}$ is a subnet of $(x_i)$, and let $U$ be an open neighborhood of $x$. There is an $i_0$ such that $x_i \in U$ for all $i \geq i_0$. Fix $j_0 \in J$ so that $f(j) \geq i_0$ for all $j \geq j_0$. Then $x_{f(j)} \in U$ for all $j \geq j_0$, so the subnet converges to $x$. \hfill \Diamond

The net version of XI.3.1.5. is the hard-to-state “converse” to (ii):

XI.3.2.11. Proposition. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$.

(i) $x_i \to x$ if and only if some tail of $(x_i)$ converges to $x$.

(ii) If $x_i \to x$, then every subnet of $(x_i)$ converges to $x$. In particular, every tail of $(x_i)$ converges to $x$. 

Proof: We give the proof of (ii), leaving the simpler proof of (i) to the reader. Suppose $(x_{f(j)})_{j \in J}$ is a subnet of $(x_i)$, and let $U$ be an open neighborhood of $x$. There is an $i_0$ such that $x_i \in U$ for all $i \geq i_0$. Fix $j_0 \in J$ so that $f(j) \geq i_0$ for all $j \geq j_0$. Then $x_{f(j)} \in U$ for all $j \geq j_0$, so the subnet converges to $x$. \hfill \Diamond

The proofs of the next two results are almost identical to the sequence versions:

XI.3.2.9. The most obvious subnets of a net $(x_i)$ are formed by letting $J$ be a cofinal subset () of $I$ and $f$ the inclusion of $J$ into $I$. We will call a subnet of this kind a subtail (a nonstandard name; there appears to be no standard one). A tail of a net is a subtail where $J$ is of the form $\{i \in I : i_0 \leq i\}$ for a fixed $i_0 \in I$. A subtail of a sequence is exactly just a subsequence, and a tail of a sequence in this sense is the same as a tail as previously defined.

For purposes of describing topologies and convergence, subtails are inadequate in general, just as in the sequence case. Thus it is necessary to use the more complicated but more general and flexible notion of subnet as we have defined it.

XI.3.10. Proposition. Let $X$ be a topological space and $(x_i)$ a net in $X$. Any subnet of a subnet of $(x_i)$ is a subnet of $(x_i)$. Any subtail of a subtail of $(x_i)$ is a subtail, and any tail of a tail is a tail.
XI.3.2.12. PROPOSITION. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$. If $(x_i)$ does not converge to $x$, then there is a subnet of $(x_i)$ (in fact a subtail), no subnet of which converges to $x$.

PROOF: If $(x_i)$ does not converge to $x$, then there is an open neighborhood $U$ of $x$ such that $J = \{i \in I : x_i \notin U\}$ is cofinal in $I$. This $J$ gives a subtail with no terms in $U$, so no subnet of it can converge to $x$. ⋄

The result is often stated in the contrapositive:

XI.3.2.13. COROLLARY. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$. If every subnet of $(x_i)$ has a subnet which converges to $x$, then $x_i \rightarrow x$.

Weak Subnets
We can extend XI.3.2.12. in an interesting way.

XI.3.2.14. DEFINITION. Let $(x_i)$ be a net in $X$ with index set $I$ defined by $\phi : I \rightarrow X$. A weak subnet of $(x_i)$ is a net defined by $\phi \circ f$, where $f$ is a not necessarily directed function from a directed set $J$ to $I$.

XI.3.2.15. Any subnet is a weak subnet, but there are many weak subnets which are not subnets: for example, $f$ could be a constant function. Even if $f$ is strictly increasing, the weak subnet is not necessarily a subnet since the range of $f$ may not be cofinal in $I$ (this cannot happen for sequences).

The proof of XI.3.2.12. actually shows the following considerably stronger result:

XI.3.2.16. PROPOSITION. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$. If $(x_i)$ does not converge to $x$, then there is a subnet of $(x_i)$ (in fact a subtail), no weak subnet of which converges to $x$.

XI.3.2.17. COROLLARY. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$. If every subnet of $(x_i)$ has a weak subnet which converges to $x$, then $x_i \rightarrow x$.

Cluster Points of Nets

XI.3.2.18. DEFINITION. Let $X$ be a topological space, and $(x_i)$ a net in $X$, $x \in X$. Then $x$ is a cluster point of $(x_i)$ if every open neighborhood $U$ of $x$ contains $x_i$ for a cofinal set of $i$ (we say $x_i$ is frequently in $U$).

The proof of the next Proposition is essentially identical to the proof of (). This result motivates the theory of convergence via filters ( ).
XI.3.2.19. **Proposition.** Let \((x_i)_{i \in I}\) be a net in a topological space \(X\). For each \(i \in I\), set

\[ T_i = \{ x_j : j \geq i \}. \]

Then \(\cap_{i \in I} T_i\) is precisely the set of cluster points of the net \((x_i)\). In particular, the set of cluster points of a net is closed.

The set \(T_i\) is called the *tail set* of the net \((x_i)\) corresponding to \(i\). Thus the convergence of a net depends only weakly on the actual directed set \(I\) but only on the collection of tail sets of the net.

**Proof:** If \(x\) is a cluster point of \((x_i)\), then any neighborhood of \(x\) contains point of \(T_i\) for arbitrarily large \(i\), hence \(x \in \cap_{i \in I} T_i\). Conversely, if \(x \in \cap_{i \in I} T_i\) and \(U\) is an open neighborhood of \(x\), then for each \(i\) we have \(T_i \cap U \neq \emptyset\), and any element of this intersection is of the form \(x_j\) for some \(j \geq i\). Thus \(x\) is a cluster point of \((x_i)\).

If \(x_i \to x\), or more generally if there is a subnet of \((x_i)\) which converges to \(x\), then \(x\) is a cluster point of \((x_i)\). Unlike in the sequence case, the converse is true in general, showing the universal applicability of nets to convergence. The proof is a good example of how nets are used in practice, and beautifully and graphically demonstrates the utility of the subnet definition.

XI.3.2.20. **Theorem.** Let \(X\) be a topological space, \((x_i)\) a net in \(X\), and \(x \in X\). If \(x\) is a cluster point of \((x_i)\), then there is a subnet of \((x_i)\) which converges to \(x\).

**Proof:** Let \(J\) be the set of all pairs \((U, i)\), where \(U\) is an open neighborhood of \(x, i \in I\), and \(x_i \in U\). Put a preorder \(\preceq\) on \(J\) by setting \((U_1, i_1) \preceq (U_2, i_2)\) if \(U_2 \subseteq U_1\) and \(i_1 \leq i_2\).

We show that \((J, \preceq)\) is directed. Suppose \((U_1, i_1)\) and \((U_2, i_2)\) are in \(J\). Set \(V = U_1 \cap U_2\) and let \(i_3\) be an upper bound for \(\{i_1, i_2\}\) in \(I\). Then since \(x\) is a cluster point of \((x_i)\), there is an \(i \geq i_3\) with \(x_i \in V\). Then \((V, i) \in J\) and \((V, i)\) is an upper bound for \(\{(U_1, i_1), (U_2, i_2)\}\).

Define \(f : J \to I\) by \(f(U, i) = i\). We show that \(f\) is directed. Let \(i_0 \in I\). We must find \((U_1, i_1) \in J\) such that \((V, i) \in J\), \((U_1, i_1) \preceq (V, i)\) implies \(i \geq i_0\). Just take \(U_1 = X\) and \(i_1 = i_0\). (Note that \(f\) is in fact order-preserving.)

Finally, we show that the subnet \((x(U, i))\) converges to \(x\). Let \(V\) be an open neighborhood of \(x\), and fix \(i_0\) such that \(x_{i_0} \in V\). Then \((V, i_0) \in J\) and \(x_{(U, i)} \in V\) whenever \((V, i_0) \preceq (U, i)\). Similarly, we get a net version of XI.3.1.11. which is valid in complete generality. The proof shows the utility of considering directed sets which are not partially ordered.

XI.3.2.21. **Theorem.** Let \(A\) be a subset of a topological space \(X\) and \(x \in X\). Then \(x \in \bar{A}\) if and only if there is a net \((x_i)\) of points of \(A\) which converges to \(x\).

**Proof:** If there is a net in \(A\) converging to \(x\), then every neighborhood of \(x\) contains a term of the net, hence a point of \(A\). Thus \(x \in \bar{A}\). Conversely, suppose \(x \in \bar{A}\). Let \(I\) be the set of all pairs \((U, y)\), where \(U\) is an open neighborhood of \(x\) and \(y \in U \cap A\). (We have \(U \cap A \neq \emptyset\) for all \(U\).) Put a preorder \(\leq\) on \(I\) by \((U, y) \leq (V, z)\) if \(V \subseteq U\). (Note that this is rarely a partial order.)

1028
We show that \((I,\leq)\) is directed. If \((U_1,y_1)\) and \((U_2,y_2)\) are in \(I\), set \(V = U_1 \cap U_2\), and let \(z \in V \cap A\). Then \((V,z)\) is an upper bound for \(\{(U_1,y_1),(U_2,y_2)\}\).

Define a net \((x_i)_{i \in I}\) by setting \(x_i(U,y) = y\). Then \(x_i \in A\) for all \(i\). If \(U\) is any open neighborhood of \(x\), let \(y \in U \cap A\), and set \(i_0 = (U,y) \in I\). Then for any \((V,z) \geq (U,y)\) in \(I\), \(x_i(U,z) = z \in V \subseteq U\), i.e. if \(i \geq i_0\), \(x_i \in U\). Thus \(x_i \to x\).

Note that unlike in the sequence case, no form of Choice is needed for this result. This again shows the versatility of the theory of nets and subnets.

**Interleaving Nets**

**XI.3.2.22.** It is technically useful to be able to “interleave” or “intertwine” two nets into a single net. This is easily done for sequences: if \((x_n)\) and \((y_n)\) are sequences in \(X\), simply alternate the terms to form the sequence

\[
(x_1,y_1,x_2,y_2,\ldots,x_n,y_n,\ldots)
\]

which will converge to a limit \(x\) if and only if \(x_n \to x\) and \(y_n \to x\) (cf. (i)).

It is unclear how to do this in the case of nets. There are two technical difficulties:

(i) The two nets need not even have the same index set.

(ii) Even if the index sets are the same, it is not obvious what “alternating” terms means.

The difficulty is in the generality of the indexing for nets, but the generality and flexibility of indexing is also the key to the solution.

**XI.3.2.23.** We first show how to modify two nets to new nets with the same index set. Suppose \((x_i)\) is a net in \(X\) with index set \(I\), and suppose \(J\) is another directed set; we will use \(\prec\) for the direction in both \(I\) and \(J\). Set \(K = I \times J\), and give \(K\) the product ordering: \((i_1,j_1) \prec (i_2,j_2)\) if and only if \(i_1 \prec i_2\) and \(j_1 \prec j_2\) (cf. (i)). For \(k = (i,j)\), set \(x_k = x_i\).

**XI.3.2.24.** **Proposition.** The net \((x_k)\) converges to \(x \in X\) if and only if \(x_i \to x\).

**Proof:** Let \(U\) be a neighborhood of \(x\). If \(x_k \to x\), there is an \(k_0 = (i_0,j_0)\) such that \(x_k \in U\) whenever \(k_0 \prec k\); then \(x_i \in U\) whenever \(i_0 \prec i\). Conversely, if \(x_i \to x\), there is an \(i_0\) such that \(x_i \in U\) whenever \(i_0 \prec i\); if \(j_0\) is any fixed element of \(J\), then \(x_k \in U\) whenever \(k_0 = (i_0,j_0) \prec k\). \(\Box\)

In fact, the modified net \((x_k)\) is a subnet of \((x_i)\) via the projection from \(I \times J\) onto \(I\).

**XI.3.2.25.** If \((x_i)\) and \((y_j)\) are nets in \(X\) with index sets \(I\) and \(J\) respectively, both can be modified to nets with index set \(K = I \times J\); convergence of the nets is unaffected by the reindexing.

**XI.3.2.26.** Now suppose \((x_k)\) and \((y_k)\) are nets with the same index set \(K\). Define a new net \((z_{k,n})\) with index set \(K \times \{1,2\}\) with ordering \((k_1,n_1) \prec (k_2,n_2)\) if and only if \(k_1 \prec k_2\), by taking \(z_{k,1} = x_k\) and \(z_{k,2} = y_k\) for all \(k\) (cf. (i)). The net \((z_{k,n})\) is called the *interleaving* of \((x_k)\) and \((y_k)\). We have:
XI.3.2.27. **Proposition.** The interleaved net \((z_{k,n})\) converges to \(x \in X\) if and only if \(x_k \to x\) and \(y_k \to x\).

The simple proof is almost identical to the proof of XI.3.2.24., and is left to the reader (Exercise ()). Note that \((x_k)\) and \((y_k)\) are subnets of \((z_{k,n})\) via the maps \(k \mapsto (k, 1)\) and \(k \mapsto (k, 2)\) respectively, giving one direction.

**Moore-Smith Convergence**

XI.3.2.28. The theory of convergence of nets is often called the theory of *Moore-Smith convergence*, since it was originated in [?]. The term “net” was coined by *Kelley*.

The Moore of Moore-Smith was E. H. Moore, not R. L. Moore. Both Moores (who were not related) were very important and influential in American mathematics. Both were first-rate mathematicians, but their importance went far beyond their own mathematical work. Both served as presidents of the American Mathematical Society and were instrumental in the development of the AMS.

E. H. Moore was probably more responsible than any other individual for the rise of American mathematics in general, and the University of Chicago in particular, to world-class status. A considerable percentage of current American mathematicians, including the author, are mathematical descendents of his (the Mathematics Genealogy Project [http://www.genealogy.ams.org/] lists at least 18,359 mathematical descendents of E. H. Moore).

R. L. Moore, in addition to being one of the principal founders of the American school of topology, was a gifted and energetic teacher who trained a large percentage of the leading American topologists of the mid-twentieth century. He also developed the “Moore method” of pedagogy (cf. I.5.7.), which is much admired and emulated, although it is not without its detractors. Unfortunately, he was also a virulent racist.

There have been a number of other mathematicians named Moore, including the author’s advisor, C. C. Moore, who is apparently unrelated to either E. H. or R. L.

**XI.3.3. Specification of a Topology via Convergence**

It is often desirable to define a topology by specifying which sequences or nets converge. This is particularly convenient when considering function spaces, where we often want to discuss the “topology of pointwise convergence,” the “topology of uniform convergence,” etc. This can be done rather cleanly, although there are some subtleties.

This approach to topology was developed by Fréchet [], Aleksandrov [] and Urysohn [], and Birkhoff [Bir37]. See [Kur66, §20] and [Kel75] for details.

XI.3.3.1. It is clear that a topology cannot be completely determined by convergence of sequences in general in the non-first-countable case; in fact, two distinct topologies on a set can have the same convergent sequences, such as the usual topology and the discrete topology on \(\beta\mathbb{N}\) (XI.11.9.16.(v); see also () for a natural and important example, and Exercise XI.7.8.12. for a simple one). We have a much better chance of success working with nets, but we will use sequences as much as possible since in applications it is very convenient to use sequences.
XI.3.3.2.  The general setup is: $X$ is a set, and $\mathcal{C}$ a collection of sequences (or nets) in $X$ which we will declare to converge, each with a specified limit. More precisely, $\mathcal{C}$ will be a set of pairs $(s, x)$, where $s$ is a sequence or net in $X$ and $x \in X$; we will declare that the sequence $s$ converges to $x$. We want a topology $\mathcal{T}$ on $X$ such that $s \to x$ in the topology $\mathcal{T}$ for every pair $(s, x)$ in $\mathcal{C}$; ideally, we want $\mathcal{C}$ to be the complete set of convergent sequences (or nets) with respect to $\mathcal{T}$.

We will be primarily interested in the case where a sequence cannot have more than one limit (this was an explicit assumption in [1], [2], [3], [4], [5]), but we will not make this a blanket assumption.

XI.3.3.3.  To make the process work, we make the modest assumption on $\mathcal{C}$ that any subsequence of a sequence in $\mathcal{C}$ is also in $\mathcal{C}$, with the same limit, i.e. that if $(s, x) \in \mathcal{C}$ and $s'$ is a subsequence of $s$, then $(s', x) \in \mathcal{C}$. This subsequence property must be satisfied for $\mathcal{C}$ to have any claim to completeness. (“Sequence” and “subsequence” can be replaced by “net” and “subnet” here.)

XI.3.3.4.  There are two possible approaches to defining $\mathcal{T}$:

(1) We could declare a subset $E$ of $X$ closed if, whenever $(s, x) \in \mathcal{C}$ and $s$ is a sequence in $E$, then $x \in E$. (If $\mathcal{T}$ is a topology on $X$ under which the sequences in $\mathcal{C}$ converge, then any $\mathcal{T}$-closed set has this property.)

(2) We could define a closure operation on $X$ by saying that if $A \subseteq X$, then $\bar{A}$ consists of all $x \in X$ such that there exists $(s, x) \in \mathcal{C}$ with $s$ a sequence in $A$. (If $\mathcal{T}$ is a topology on $X$ under which the sequences in $\mathcal{C}$ converge, then any such $x$ must be a limit point of $A$.) If $A \mapsto \bar{A}$ is a Kuratowski closure operator ( ), it defines a suitable topology $\mathcal{T}$.

Procedure (2) is more constructive. It turns out that procedure (1) works quite nicely, but procedure (2) is more problematic.

First consider procedure (1). Half the answer is almost immediate:

XI.3.3.5.  **Proposition.**  Let $X$ be a set and $\mathcal{C}$ a set of sequences as above, with the subsequence property. Let $\mathcal{E}$ be the collection of closed subsets of $X$ as defined in (1). Then $\mathcal{E}$ is the collection of closed sets in a topology $\mathcal{T}$ which is the strongest topology on $X$ for which $s \to x$ for every $(s, x) \in \mathcal{C}$.

**Proof:** We need to show that $\mathcal{E}$ satisfies the conditions of XI.2.1.4. It is obvious that $\emptyset$ and $X$ are in $\mathcal{E}$, and that $\mathcal{E}$ is closed under arbitrary intersections. So suppose $A$ and $B$ are in $\mathcal{E}$ and $(s, x) \in \mathcal{C}$ with $s$ a sequence in $A \cup B$. Then $s$ either has a subsequence $s'$ in $A$ or a subsequence $s''$ in $B$. In the first case $(s', x) \in \mathcal{C}$, so $x \in A \subseteq A \cup B$ since $A \in \mathcal{E}$, and similarly in the second case $x \in B \subseteq A \cup B$. Thus $A \cup B \in \mathcal{E}$ and $\mathcal{E}$ is closed under finite unions. It is clear from the last observation in (1) that the topology defined by $\mathcal{E}$ is the strongest possible topology on $X$ giving the desired convergence.

XI.3.3.6.  This result and its proof hold if “sequence” is replaced by “net” and “subsequence” by “subnet” throughout (i.e. the subsequence property is replaced by the subnet property, and the definition of “closed” in (1) uses nets).

1031
XI.3.3.7. However, it will rarely be true that $\mathcal{C}$ is the complete set of sequences which converge with respect to $T$. For one thing, for $\mathcal{C}$ to be complete it must have the following property (cf. XI.3.1.6.):

If $s$ is a sequence in $X$, $x \in X$, and every subsequence $s'$ of $s$ has a subsequence $s''$ such that $(s'', x) \in \mathcal{C}$, then $(s, x) \in \mathcal{C}$.

$\mathcal{C}$ must also contain all constant sequences (with limit equal to the constant term).

Even these additional assumptions on $\mathcal{C}$ are not enough in general. Before discussing this further, we will examine procedure (2), beginning with the net version.

XI.3.3.8. So, given a collection $\mathcal{C}$ of nets, let us define a closure operation on $X$ as in (2). If $\mathcal{C}$ is to be the set of convergent nets for a topology, we must have the diagonal property:

It turns out this property along with the previous ones is sufficient to obtain a Kuratowski closure operation:

XI.3.3.9. Proposition. Let $X$ be a set and $\mathcal{C}$ a collection of pairs $(s, x)$ where $s$ is a net in $X$ and $x \in X$. Suppose $\mathcal{C}$ has the following properties:

(i) If $x \in X$ and $s$ is a constant net with all terms $x$, then $(s, x) \in \mathcal{C}$.

(ii) If $(s, x) \in \mathcal{C}$ and $s'$ is a subnet of $s$, then $(s', x) \in \mathcal{C}$.

(iii) Diagonal property

For $A \subseteq X$, let $\bar{A}$ be the set of all $x \in X$ such that there exists a pair $(s, x) \in \mathcal{C}$ with $s$ a net in $A$. Then $A \mapsto \bar{A}$ is a Kuratowski closure operation on $X$. The corresponding topology on $X$ is the strongest topology $T$ on $X$ with the property that $s \rightarrow x$ in $T$ for every $(s, x) \in \mathcal{C}$, and coincides with the topology defined in XI.3.3.5..

Proof: Obviously $\emptyset = \emptyset$, and $X = X$ since $\mathcal{C}$ contains all constant nets. If $A \subseteq X$, then $A \subseteq \bar{A}$ by the same argument, and $\bar{A}$ by (iii). Now suppose $A, B \subseteq X$, and $x \in X$. If $x \in A$ or $x \in B$, then obviously $x \in \bar{A} \cup \bar{B}$. Conversely, if $x \in \bar{A} \cup \bar{B}$, let $s = (x_i)$ be a net in $A \cup B$ with $(s, x) \in S$. Then $s$ is a subnet of $x$ in $A$ or a subnet of $s''$ in $B$. In the first case $(s'', x) \in \mathcal{C}$, so $x \in \bar{A}$, and in the second case $x \in \bar{B}$. In either case, $x \in \bar{A} \cup \bar{B}$. Thus $\bar{A} \cup \bar{B} = \bar{A} \cup \bar{B}$.

XI.3.3.10. From either XI.3.3.5. or XI.3.3.9. we are left with two questions:

1. Is the topology $T$ so defined the unique topology on $X$ for which $s \rightarrow x$ for all $(s, x) \in \mathcal{C}$?

2. Is $\mathcal{C}$ the collection of all convergent sequences or nets with respect to $T$?

The answer to (1) is clearly no in general: we have $s \rightarrow x$ for all $(x, s) \in \mathcal{C}$ for any weaker topology too, e.g. the indiscrete topology.

For question (2) to have a positive answer, it is necessary that $\mathcal{C}$ have the net version of the property of XI.3.3.7., as well as (i)–(iii) of XI.3.3.9.. This suffices:
XI.3.3.11. Theorem. (cf. [Bir37]) Let \( X \) be a set and \( C \) a collection of pairs \((s, x)\) where \( s \) is a net in \( X \) and \( x \in X \). Suppose \( C \) has the following properties:

(i) If \( x \in X \) and \( s \) is a constant net with all terms \( x \), then \((s, x) \in C\).

(ii) If \((s, x) \in C \) and \( s' \) is a subnet of \( s \), then \((s', x) \in C\).

(iii) Diagonal property

(iv) If \( s \) is a net in \( X \), \( x \in X \), and every subnet \( s' \) of \( s \) has a weak subnet \( s'' \) with \((s'', x) \in C \), then \((s, x) \in C\).

Let \( T \) be the topology defined by \( C \) as in XI.3.3.7. or XI.3.3.9. Then \( C \) is precisely the set of nets in \( X \) which converge in \( T \).

Proof: We need only to show that if \( s \) is a net in \( X \) with \( s \to x \), then \((s, x) \in C\). Suppose \((s, x) \notin C\). Then by (iv) there is a subnet \( s' \) of \( s \) such that \((s'', x) \notin C \) for all weak subnets \( s'' \) of \( s' \). Set

\[ E = \{ y \in X | (s'', y) \in C \text{ for some weak subnet } s'' \text{ of } s' \} \, .\]

The set \( E \) is exactly the closure of the set of terms in \( s' \), hence is a closed subset of \( X \). Since all terms of \( s' \) are in \( E \), we cannot have \( s' \to x \) and hence \( s \) does not converge to \( x \).

XI.3.3.12. Hypotheses (iii) and (iv) may be somewhat difficult to check in some cases, but they are indeed necessary conditions for the conclusion (). By a somewhat more complicated argument (cf. [Kel75, p. 74]), it can be shown that hypothesis (iv) can be slightly weakened by replacing “weak subnet” with “subnet.” The proof of this in [Kel75] (implicitly) uses the AC (note that the proof of XI.3.3.11. above does not use any Choice). A class of nets satisfying (i)–(iii) and this version of (iv) is called a convergence class.

XI.3.3.13. It is suggested in [?], p. 19] that hypothesis (iv) is redundant. However, at least some weak version of (iv) is necessary; otherwise, we could take \( C \) to be the class of constant nets. The topology on \( X \) determined by this \( C \) is the discrete topology, but if \( X \) has more than one element \( C \) is not the entire class of convergent nets in \( X \) (the class of eventually constant nets).

The following weak version of (iv) might suffice:

(iv') If \( s \) is a net in \( X \), \( x \in X \), and there is a tail \( s' \) of \( s \) with \((s', x) \in C \), then \((s, x) \in C\).

XI.3.3.14. There is also a set-theoretic problem: the collection of all subnets of a net does not form a set. One might be able to get around this problem by only considering nets whose index set is a subset of some fixed set whose cardinality is large compared to \( |X| \), although construction of diagonal nets may cause difficulty. One really needs a notion of equivalence of nets and that every net on \( X \) is equivalent to a net with an index set of bounded cardinality. This is effectively accomplished using the theory of filter convergence (), which avoids these set-theoretic difficulties.

XI.3.3.15. Unfortunately, there is no simple version of XI.3.3.11. for sequences. The sequence version of (iii) does not hold in general (). If the generated topology is assumed to be first countable, the theorem holds with sequences in place of nets; but there is no simple criterion in terms of convergence of sequences which will characterize when the generated topology is first countable.
XI.3.4. Convergence via Filters

There is another approach to convergence in a topological space which uses filters instead of nets. This approach is somewhat more complicated and less familiar at first, but has some technical advantages once set up. The theories of convergence via nets and filters are entirely equivalent logically, and there is a way of translating directly from one to the other, and it is largely a matter of taste (and, unfortunately, sometimes of national pride, since the theory of net convergence was developed in the United States and the theory of filter convergence in France) which approach is used; but some things are done more easily or efficiently from one of the points of view instead of the other, so it is valuable to be familiar with both approaches.

The theory of convergence via filters was developed by H. Cartan for Bourbaki around 1940. More details can be found in [?], [Dug78].

XI.3.4.1. Recall (XII.1.6.1.) that a filter on X is a nonempty collection $\mathcal{F}$ of subsets of X such that

(i) $\mathcal{F}$ is closed under finite intersections.

(ii) $\mathcal{F}$ has the superset property: if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $B \in \mathcal{F}$.

(iii) $\emptyset \notin \mathcal{F}$.

XI.3.4.2. Definition. A collection of nonempty subsets of a set X is centered if, for every $A, B \in \mathcal{D}$ there is a $C \in \mathcal{D}$ with $C \subseteq A \cap B$. A centered collection of subsets of X is called a filterbase in X.

XI.3.4.3. A collection of nonempty subsets of a set X is centered if and only if it is directed by inverse inclusion, i.e. $A \leq B$ if $B \subseteq A$. A filterbase $\mathcal{D}$ has the finite intersection property (XII.1.6.9.), and hence is contained in a unique smallest filter $\mathcal{F}$ (XII.1.6.10.) called the filter generated by $\mathcal{D}$; $\mathcal{D}$ is called a filterbase for $\mathcal{F}$. In fact, $\mathcal{F}$ consists precisely of all subsets of X which contain a subset in $\mathcal{D}$.

Here are the most important examples of filterbases for topology:

XI.3.4.4. Examples. (i) Let $(X, \mathcal{T})$ be a topological space and $p \in X$. The set of all $\mathcal{T}$-neighborhoods of $p$ is a filter, called the neighborhood filter, or $\mathcal{T}$-neighborhood filter, of $p$ in X, denoted $\mathcal{N}_p$. The collection of open neighborhoods of $p$ is a filterbase for $\mathcal{N}_p$; more generally, any local base for $\mathcal{T}$ at $p$ is a filterbase for $\mathcal{N}_p$.

(ii) Let X be a set, and $(x_i)_{i \in I}$ a net in X. The tail set $\{T_i : i \in I\}$ (XI.3.2.19.) is a filterbase in X.

XI.3.4.5. Example XI.3.4.4.(ii) is actually universal: if $\mathcal{D}$ is a filterbase in X, let $I$ be the set of pairs $(A, x)$, where $A \in \mathcal{D}$ and $x \in A$. Set $(A, x) \leq (B, y)$ if $B \subseteq A$. Then $I$ is a directed set since $\mathcal{D}$ is centered. If $x_{(A, z)} = z$, then $(x_{(A, z)})$ is a net in X with index set $I$, with $T_{(A, z)} = A$ for each $(A, z)$. There is thus a one-one correspondence between filterbases in X and equivalence classes of nets in X, where two nets are regarded as equivalent if their corresponding filterbases are identical (see Exercise XI.3.5.2.; cf. XI.3.3.14.).

Convergence of Filters

The definition of convergence of a filter is very simple:
XI.3.4.6. Definition. Let \((X, \mathcal{T})\) be a topological space. A filter \(\mathcal{F}\) in \(X\) converges to a point \(p \in X\) if \(\mathcal{N}_p \subseteq \mathcal{F}\).

XI.3.4.7. Because of the superset property of filters, a filter \(\mathcal{F}\) converges to a point \(p\) if and only if, whenever \(U\) is a neighborhood of \(p\), there is an \(A \in \mathcal{F}\) with \(A \subseteq U\). Thus we extend the notion of filter convergence to filterbases:

XI.3.4.8. Definition. Let \((X, \mathcal{T})\) be a topological space. A filterbase \(\mathcal{D}\) in \(X\) converges to a point \(p \in X\) if whenever \(U\) is a neighborhood of \(p\), there is an \(A \in \mathcal{D}\) with \(A \subseteq U\).

Thus a filterbase converges to \(p\) if and only if the filter it generates converges to \(p\).

XI.3.4.9. This is consistent with convergence of nets: a net \((x_i)_{i \in I}\) converges to \(p\) if and only if the corresponding filterbase \(\{T_i : i \in I\}\) converges to \(p\).

Closure via Filters

We can give a characterization of the closure of a set using filters. It is cleanest to express it in terms of filterbases, since if \(A\) is a subset of \(X\) and \(\mathcal{D}\) is a filterbase in \(A\), \(\mathcal{D}\) is also a filterbase in \(X\) (this is false for filters).

XI.3.4.10. Proposition. Let \((X, \mathcal{T})\) be a topological space and \(A \subseteq X\). The \(\mathcal{T}\)-closure of \(A\) in \(X\) consists precisely of all points \(x\) with the property that there is a filterbase \(\mathcal{D}\) in \(A\) converging to \(x\). In particular, \(A\) is \(\mathcal{T}\)-closed if and only if, whenever \(\mathcal{D}\) is a filterbase in \(A\) converging to \(x \in X\), then \(x \in A\).

XI.3.4.11. To state the result using filters, we make the following definition. We say a filter \(\mathcal{F}\) in a set \(X\) is eventually in \(A \subseteq X\) if \(A \subseteq \mathcal{F}\) (or, equivalently, some subset of \(A\) is in \(\mathcal{F}\)). We can then state:

XI.3.4.12. Proposition. Let \((X, \mathcal{T})\) be a topological space and \(A \subseteq X\). The \(\mathcal{T}\)-closure of \(A\) in \(X\) consists precisely of all points \(x\) with the property that there is a filter \(\mathcal{F}\) which is eventually in \(A\) and converges to \(x\). In particular, \(A\) is \(\mathcal{T}\)-closed if and only if, whenever \(\mathcal{F}\) is a filter in \(X\) which is eventually in \(A\) and converges to \(x \in X\), then \(x \in A\).

XI.3.4.13. Similarly, we say a filterbase \(\mathcal{D}\) in \(X\) is eventually in \(A \subseteq X\) if there is a \(D \in \mathcal{D}\) with \(D \subseteq A\).

Cluster Points of Filters

XI.3.4.14. Definition. Let \(\mathcal{D}\) be a filterbase in a set \(X\) and \(A \subseteq X\). We say \(\mathcal{D}\) is frequently in \(A\) if \(A \cap D \neq \emptyset\) for all \(D \in \mathcal{D}\). A point \(x\) is a cluster point for a filterbase \(\mathcal{D}\) in \(X\) if \(\mathcal{D}\) is frequently in every neighborhood of \(x\).

This terminology is consistent with the corresponding definition for nets (). There is no modification needed to replace “filterbase” by “filter” (in fact, the statement for filters is a special case since every filter is a filterbase).
XI.3.4.15. If a filter or filterbase converges to \( x \), then \( x \) is a cluster point. We have the following filter version of XI.3.2.19:

XI.3.4.16. Proposition. Let \( (X, \mathcal{T}) \) be a topological space, and let \( D \) be a filterbase in \( X \). A point \( x \in X \) is a cluster point of \( D \) if and only if \( x \in \bigcap_{D \in \mathcal{D}} D \).

Subfilters

What is the right filter or filterbase analog of a subnet? The following definition is reasonable:

XI.3.4.17. Definition. Let \( \mathcal{D} \) and \( \mathcal{D}' \) be filterbases in a set \( X \). Then \( \mathcal{D}' \) is subordinate to \( \mathcal{D} \), written \( \mathcal{D}' \vdash \mathcal{D} \), if for every \( A \in \mathcal{D} \) there is a \( B \in \mathcal{D}' \) with \( B \subseteq A \).

XI.3.4.18. It is easily seen that if \( \mathcal{F} \) and \( \mathcal{F}' \) are the filters generated by \( \mathcal{D} \) and \( \mathcal{D}' \) respectively, then \( \mathcal{D}' \vdash \mathcal{D} \) if and only if \( \mathcal{F} \subseteq \mathcal{F}' \), i.e. the subordinate filter is larger. If \( \mathcal{F} \subseteq \mathcal{F}' \), we say the filter \( \mathcal{F}' \) is finer than \( \mathcal{F} \). Thus it is not appropriate to call \( \mathcal{D}' \) a “subfilterbase” of \( \mathcal{D} \) (which is a mouthful anyway.) However, if \( (x_i)_{i \in I} \) is a net in \( X \) and \( (y_j)_{j \in J} \) is a subnet, then the tail filterbase of \( (y_j) \) is subordinate to the tail filterbase of \( (x_i) \) (the converse is false).

We have:

XI.3.4.19. Proposition. Let \( (X, \mathcal{T}) \) be a topological space, and let \( \mathcal{D}, \mathcal{D}' \) be filterbases in \( X \), with \( \mathcal{D}' \vdash \mathcal{D} \).

(i) If \( \mathcal{D} \) converges to \( x \in X \), then \( \mathcal{D}' \) converges to \( x \).

(ii) If \( x \in X \) is a cluster point of \( \mathcal{D}' \), it is a cluster point of \( \mathcal{D} \).

We next obtain a generalization of XI.3.2.20:

XI.3.4.20. Proposition. Let \( \mathcal{D} \) be a filterbase in a set \( X \), and \( (x_i)_{i \in I} \) a net in \( X \) which is frequently in \( D \) for every \( D \in \mathcal{D} \). Then there is a subnet of \( (x_i) \) which is eventually in every \( D \in \mathcal{D} \).

Proof: Let \( J \) be the set of all pairs \( (D, i) \) where \( D \in \mathcal{D} \), \( i \in I \), and \( x_i \in D \). Set \( (D_1, i_1) \leq (D_2, i_2) \) if \( D_2 \subseteq D_1 \) and \( i_1 \leq i_2 \). Then \( J \) is directed: if \( (D_1, i_1), (D_2, i_2) \in J \), there is a \( D_3 \in \mathcal{D} \), \( D_3 \subseteq D_1 \cap D_2 \), and an \( i_0 \in I \) with \( i_1, i_2 \leq i_0 \), and an \( i_3 \geq i_0 \) such that \( x_{i_3} \in D_3 \) since \( (x_i) \) is frequently in \( D_3 \), so \( (D_3, i_3) \in J \) and \( (D_1, i_1), (D_2, i_2) \leq (D_3, i_3) \). Define \( f : J \to I \) by \( f(D, i) = i \). Then \( f \) is directed: if \( i_0 \in I \), fix \( D_1 \in \mathcal{D} \) and \( i_1 \geq i_0 \) with \( x_{i_1} \in D_1 \); then \( f(D, i) \geq i_0 \) whenever \( (D, i) \geq (D_1, i_1) \). If \( \phi : I \to X \) gives \( \phi(i) = x_i \), then \( \phi \circ f \) defines a subnet \( (x_{f(j)}) \). If \( D_0 \in \mathcal{D} \), then there is an \( i_0 \in I \) with \( (D_0, i_0) \in J \). If \( (D_0, i_0) \leq (D, i) = (D, j) \), then \( x_{f(j)} \in D \subseteq D_0 \), so the subnet \( (x_{f(j)}) \) is eventually in \( D_0 \).

Rephrasing, we get:

1036
XI.3.4.21. COROLLARY. Let $D$ be a filterbase in a set $X$, and $(x_i)_{i \in I}$ a net in $X$. If the tail set filterbase $D'$ of $(x_i)$ satisfies $D \vdash D'$, then there is a subnet of $(x_i)$ whose tail set filterbase $D''$ satisfies $D'' \vdash D$.

We also have the following analog of XI.3.2.12.:

XI.3.4.22. PROPOSITION. Let $(X, T)$ be a topological space, $D$ a filterbase in $X$, and $x \in X$. If $D$ does not converge to $x$, then there is a filterbase $D' \vdash D$ such that no filterbase $D'' \vdash D'$ converges to $x$.

Ultrafilters and Maximal Filterbases

XI.3.4.23. By XI.3.4.18., larger means “finer” for filters. The largest filters are ultrafilters (XII.1.6.1.), which are therefore the “finest” filters. By XII.1.6.15. (which uses AC), every filter is contained in an ultrafilter. By XII.1.6.12., a filter $F$ in a set $X$ is an ultrafilter if and only if for every $A \subseteq X$, either $A \in F$ or $A^c \in F$, i.e. $F$ is either eventually in $A$ or eventually in $A^c$.

XI.3.4.24. A filterbase which generates an ultrafilter is called a maximal filterbase; “ultrafilterbase” would be a better name but too cumbersome (“maximal filterbase” is an abuse of language since a maximal filterbase is not maximal among all filterbases under inclusion unless it is already an ultrafilter; but see Exercise XI.3.5.4.). A filterbase $D$ in $X$ is maximal if and only if for each $A \subseteq X$, there is a $D \in D$ such that either $D \subseteq A$ or $D \subseteq A^c$.

XI.3.4.25. PROPOSITION. Let $(X, T)$ be a topological space, $D$ a maximal filterbase in $X$, and $x \in X$. If $x$ is a cluster point of $D$, then $D$ converges to $x$.

Proof: Let $U$ be a neighborhood of $x$. Then there is a $D \in D$ with either $D \subseteq U$ or $D \subseteq U^c$. But $D \cap U \neq \emptyset$ since $x$ is a cluster point of $D$, so $D \subseteq U^c$ is impossible and $D \subseteq U$.

XI.3.4.26. PROPOSITION. Let $X$ and $Y$ be sets, $f : X \rightarrow Y$ any function, and $D$ a maximal filterbase in $X$. Then $f(D) = \{f(D) : D \in D\}$ is a maximal filterbase in $Y$.

Proof: It is clear that $f(D)$ is a filterbase in $Y$. If $B \subseteq Y$ and $A = f^{-1}(B)$, then $A^c = f^{-1}(B^c)$. Since $D$ is maximal, there is a $D \in D$ such that either $D \subseteq f^{-1}(B)$ or $D \subseteq f^{-1}(B^c)$, hence either $f(D) \subseteq B$ or $f(D) \subseteq B^c$. Thus $f(D)$ is maximal.

Universal Nets

XI.3.4.27. DEFINITION. A net $(x_i)$ in a set $X$ is a universal net if its tail set filterbase is maximal.

The simple verification of the next proposition is left to the reader.

XI.3.4.28. PROPOSITION. A net $(x_i)$ in a set $X$ is universal if and only if, for every subset $A$ of $X$, either $(x_i)$ is eventually in $A$ or $(x_i)$ is eventually in $A^c$. 1037
XI.3.4.29. The existence of universal nets is an immediate consequence of the existence of ultrafilters. But we can do better:

XI.3.4.30. **Theorem.** Let \( X \) be a set, and \( (x_i) \) a net in \( X \). Then \( (x_i) \) has a universal subnet.

**Proof:** Let \( D \) be the tail set filterbase corresponding to \( (x_i) \). Let \( F \) be an ultrafilter containing \( D \) (i.e. containing the filter generated by \( D \)). Since \( D \subseteq F \), if \( F \in F \) we have \( F \cap D \neq \emptyset \) for all \( D \in D \), and it follows that \( (x_i) \) is frequently in \( F \). Since \( F \) is itself a filterbase, by XI.3.4.20., there is a subnet \( (x_{f(j)}) \) of \( (x_i) \) which is eventually in \( F \) for any \( F \in F \). If \( A \subseteq X \), then either \( A \in F \) or \( A^c \in F \), so \( (x_{f(j)}) \) is either eventually in \( A \) or eventually in \( A^c \). Thus \( (x_{f(j)}) \) is a universal net.

This result requires some version of Choice, and is not a theorem of ZF. In fact, it is equivalent to the fact that every filter on a set is contained in an ultrafilter, hence to BPI ().

XI.3.5. **Exercises**

XI.3.5.1. Let \( X \) be a topological space. Make the following definitions:

(i) A subset \( A \) of \( X \) is **sequentially closed** if, whenever \( (x_n) \) is a sequence in \( A \) with \( x_n \to x \in X \), then \( x \in A \).

(ii) If \( A \) is a subset of \( X \), then the **sequential closure** of \( A \) is the set of all \( x \in X \) for which there is a sequence \( (x_n) \) in \( A \) with \( x_n \to x \).

These definitions, from [Fra65], are standard, but somewhat inconsistent: the sequential closure of a set is not sequentially closed in general. Sequential closure can be applied repeatedly, even transfinitely.

(iii) The space \( X \) is **sequential** if every sequentially closed subset of \( X \) is closed.

(iv) The space \( X \) is a **Fréchet-Urysohn space** if the sequential closure of every subset of \( X \) equals its closure (i.e. is closed).

(a) Show that every first countable topological space is Fréchet-Urysohn, and every Fréchet-Urysohn space is sequential. The simplest example of a Fréchet-Urysohn space which is not first countable is the one-point compactification of an uncountable discrete space (so even a compact Hausdorff Fréchet-Urysohn space need not be first countable). See XI.9.5.7. for another example, and [Eng89, 1.6.19] for a sequential space which is not Fréchet-Urysohn.

(b) Show that the space in XI.11.12.8. is not sequential, even though every singleton subset is a \( G_\delta \). (In fact, “most” topological spaces which are not first countable fail to be sequential.)

(c) Show that every quotient of a sequential space is sequential. In particular, every quotient of a first countable space is sequential.

(d) Show that every sequential Hausdorff space is a quotient of a locally compact metrizable space. [Consider the separated union of the space consisting of all convergent sequences in the space and their limits, i.e. a disjoint union of copies of \( \mathbb{N}^1 \).] This is a strong converse to (c).

(e) Show that a topological space \( X \) is sequential if and only if the sequential criterion for continuity holds for functions from \( X \), i.e. if \( Y \) is any topological space and \( f : X \to Y \) is a function, then \( f \) is continuous if and only if, whenever \( x_n \to x \) in \( X \), we have \( f(x_n) \to f(x) \).
(f) If $X$ is a Fréchet-Urysohn space, $(x_n)$ is a sequence in $X$, and $x \in X$, then $x$ is a cluster point of $(x_n)$ if and only if there is a subsequence of $(x_n)$ converging to $X$. Show that this property characterizes Fréchet-Urysohn spaces among sequential spaces (but not among all topological spaces).

(g) Show that a topological space $X$ is Fréchet-Urysohn if and only if every subspace of $X$ is sequential in the relative topology.

For more on these spaces, see [Eng89] (where Fréchet-Urysohn spaces are called Fréchet spaces). Both notions (sequential and Fréchet-Urysohn) are reasonable precise versions of the condition “the topology of $X$ is completely determined by convergence of sequences.”

**XI.3.5.2.** Let $(x_i)_{i \in I}$ and $(y_j)_{j \in J}$ be two nets in a set $X$. Here are various possible notions of equivalence for the nets. Show each is an equivalence relation and compare them.

(a) The nets generate the same filterbase (XI.3.4.5.).

(b) The corresponding filterbases generate the same filter.

(c) Each net is a subnet of the other.

(d) Every subnet of each net has a subnet which is also a subnet of the other net.

[To compare (a) with (c) and (d), consider constant nets with index sets $\mathbb{N}$ and the first uncountable ordinal.]

**XI.3.5.3.** In what sense is XI.3.4.20. a generalization of XI.3.2.20.?

**XI.3.5.4.** Let $X$ be a set, and $\mathcal{D}, \mathcal{D}'$ filterbases in $X$.

(a) Show that we can have $\mathcal{D}' \Vdash \mathcal{D}$ and $\mathcal{D} \Vdash \mathcal{D}'$ even if $\mathcal{D} \neq \mathcal{D}'$. Thus $\Vdash$ is only a preorder and not a partial order on filterbases.

(b) If $\mathcal{D}$ is a maximal filterbase and $\mathcal{D}' \Vdash \mathcal{D}$, then also $\mathcal{D} \Vdash \mathcal{D}'$, and any $\mathcal{D}$ with this property is maximal.
XI.4. Continuous Functions

XI.4.1. Definition and Properties

XI.4.2. Topological Continuity vs. Metric Continuity
XI.5. Comparison of Topologies

XI.5.1. Stronger and Weaker Topologies

XI.5.1.1. Definition. Let $\mathcal{T}$ and $\mathcal{S}$ be topologies on the same set $X$. Then $\mathcal{S}$ is stronger (larger, finer) than $\mathcal{T}$, and $\mathcal{T}$ is weaker (smaller, coarser) than $\mathcal{S}$, if $\mathcal{T} \subseteq \mathcal{S}$, i.e. if every $\mathcal{T}$-open set is also $\mathcal{S}$-open.

The philosophy behind the names reflects the intuitive idea that a topology gives a notion of “closeness” for points, and a stronger or finer topology gives a more discriminating notion of “closeness” than a weaker or coarser one, i.e. makes it easier to distinguish between points using open sets.

A topology is stronger than itself, and also weaker than itself. In fact, a common and useful way to prove that two potentially different topologies are the same is to show that each is stronger than the other. We sometimes use the terms strictly stronger, strictly weaker, etc., when we mean to specify that the topologies are distinct.

XI.5.1.2. It can happen, and often does, that if $\mathcal{T}$ and $\mathcal{S}$ are topologies on a set $X$, then neither topology is stronger than the other. In this case the topologies are said to be incomparable.

XI.5.1.3. Examples. (i) On every set $X$, the discrete topology is the strongest topology: every topology on $X$ is weaker than the discrete topology. Similarly, every topology on $X$ is stronger than the indiscrete topology, so the indiscrete topology on $X$ is the weakest topology on $X$.

(ii) The finite complement topology (XI.1.2.8.(iv)) on a set $X$ is the weakest $T_1$ topology () on $X$. For in any $T_1$ topology on $X$, the singleton subsets are closed, hence finite sets are closed, so complements of finite subsets must be open. In particular, the finite complement topology on $\mathbb{R}$ is weaker than the ordinary topology.

(iii) Let $X = \{a, b\}$, and let $\mathcal{T} = \{\emptyset, \{a\}, X\}$, $\mathcal{S} = \{\emptyset, \{b\}, X\}$. Then $\mathcal{T}$ and $\mathcal{S}$ are incomparable.

(iv) On $C([0, 1])$, the topology of pointwise convergence () is strictly weaker than the topology of uniform convergence (). The topology of mean convergence () is strictly weaker than uniform convergence but incomparable with pointwise convergence. In fact, for every $p$, $0 < p \leq \infty$, there is an $L^p$ topology on $C([0, 1]) ()$, with $p = 1$ giving mean convergence and $p = \infty$ uniform convergence; the $L^p$ topology is weaker than the $L^r$ topology if and only if $p \leq r ()$. So there is an uncountable family of natural topologies on $C([0, 1])$, all distinct and all comparable; only the strongest of these (uniform convergence) is comparable with pointwise convergence. (There are very many other topologies on $C([0, 1])$ too, most of which are incomparable to these.)

(v) On $C_c(\mathbb{R})$, the set of real-valued continuous functions on $\mathbb{R}$ of compact support, i.e. zero outside a finite interval, there is also an $L^p$ topology for $0 < p \leq \infty ()$. In this case no two of the topologies are comparable ()

XI.5.1.4. When the topology is changed, all the ancillary things associated with a topological space potentially change. If the topology on a space $X$ is made stronger, i.e. with more open sets, then

Since closed sets are complements of open sets, there are more closed sets. (Recall that being closed is not the opposite of being open!) Thus it is easier for a subset of $X$ to be closed.
It is harder for a point \( x \) to be a limit point of a subset \( E \), since there are more open neighborhoods of \( x \) which must have nonempty intersection with \( E \). Thus \( E \) has fewer limit points and the closure of \( E \) is smaller.

Similarly, it is harder for a point \( x \) to be the limit of a sequence or net \((x_i)\), since there are more open neighborhoods of \( x \) which must eventually contain the sequence or net. Thus there are fewer convergent sequences and nets. Similarly, a sequence or net has fewer cluster points.

It is harder for a subset \( F \) of \( X \) to be compact, since there are more open covers of \( F \) which must each have finite subcovers.

It is easier for \( X \) to be \( T_0 \), \( T_1 \), or Hausdorff, since there are more open sets to choose from.

However, it is unclear whether regularity or normality becomes easier when the topology becomes stronger, since on the one hand for a given point \( p \) and closed set \( F \) it is easier to find disjoint neighborhoods, but on the other hand there are more pairs \((p,F)\) to consider. Similarly, normality does not obviously become easier. In fact, regularity and normality are not preserved under making the topology stronger or weaker (cf. XI.7.8.13.; for the other direction, consider weakening the discrete topology). In a similar way, one might expect paracompactness, like compactness, to be preserved when weakening a topology, since there are fewer open covers to consider; but on the other hand an open cover has fewer refinements. Paracompactness is in fact not preserved under either strengthening or weakening the topology (XI.7.8.13. and weakening the discrete topology are again counterexamples).

It is harder for a subset of \( X \) to be connected, since there are more potential separations.

If \( Y \) is another space and \( f : X \to Y \), and the topology on \( Y \) is kept fixed and the topology on \( X \) strengthened, it is easier for \( f \) to be continuous since it is easier for \( f^{-1}(V) \) to be open for any given open set \( V \subseteq Y \). However, if the topology on \( X \) is kept fixed and the topology on \( Y \) is strengthened, it is harder for \( f \) to be continuous since the inverse images of more open sets must be considered. If the topologies on both \( X \) and \( Y \) are strengthened, it may be neither easier nor harder for \( f \) to be continuous. In particular, if the topology on \( X \) is strengthened, functions from \( X \) to \( Y \) may change from continuous to discontinuous or vice versa; the set of continuous functions from \( X \) to \( Y \) will normally be different, although neither larger nor smaller (cf. Exercise XI.7.8.11.).

Of course, there are analogous statements if the topology on \( X \) is weakened.

### XI.5.1.5

The following chart summarizes the main features of stronger and weaker topologies:

<table>
<thead>
<tr>
<th>Stronger (larger, finer) topology</th>
<th>Weaker (smaller, coarser) topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>More open sets</td>
<td>Fewer open sets</td>
</tr>
<tr>
<td>More closed sets</td>
<td>Fewer closed sets</td>
</tr>
<tr>
<td>Fewer limit points</td>
<td>More limit points</td>
</tr>
<tr>
<td>Smaller closures</td>
<td>Larger closures</td>
</tr>
<tr>
<td>Fewer compact subsets</td>
<td>More compact subsets</td>
</tr>
<tr>
<td>Fewer connected subsets</td>
<td>More connected subsets</td>
</tr>
<tr>
<td>Fewer convergent sequences (nets)</td>
<td>More convergent sequences (nets)</td>
</tr>
<tr>
<td>More continuous functions out</td>
<td>Fewer continuous functions out</td>
</tr>
<tr>
<td>Fewer continuous functions in</td>
<td>More continuous functions in</td>
</tr>
</tbody>
</table>
The easiest way to remember what happens under strengthening or weakening of the topology is to think of the extreme cases of the discrete or indiscrete topology. (In fact, one good way to analyze many mathematical situations and guess what happens is to pass to the most extreme case.)

XI.5.1.6. Our use of “stronger” and “weaker” in comparing topologies is the standard one, especially in analysis, particularly functional analysis (weak and weak-* topologies), strong and weak operator topologies. However, it should be noted that a few topology references use these terms in the opposite sense.

XI.5.1.7. Another sometimes convenient way to describe stronger and weaker topologies is: if $S$ and $T$ are topologies on a set $X$, then $S$ is stronger than $T$ if and only if the identity map from $(X; S)$ to $(X; T)$ is continuous.

Strongest and Weakest Topologies Defined by a Family of Functions

XI.5.1.8. If $X$ is a set, and $(Y, S)$ is a topological space, and $f : X \to Y$ is a function, then there is a weakest topology on $X$ such that $f$ is continuous. In fact,

$$T = \{f^{-1}(V) : V \in S\}$$

is a topology on $X$, as is easily checked since inverse images respect unions and intersections, and $T$ is obviously the weakest topology on $X$ making $f$ continuous.

XI.5.1.9. More generally, if $X$ is a set, $\{(Y_j, S_j) : j \in J\}$ is a family of topological spaces (not necessarily distinct), and $f_j : X \to Y_j$ is a function for each $j \in J$, then there is a weakest topology on $X$ making all the $f_j$ continuous. The set

$$\{f_j^{-1}(V) : j \in J, V \in S_j\}$$

is not usually a topology on $X$; however, it is a subbase for a topology $T$ which is obviously the weakest topology on $X$ making all the $f_j$ continuous.

XI.5.1.10. The topology $T$ can also be characterized as the intersection of all topologies on $X$ making each $f_j$ continuous (there is always at least one such topology, the discrete topology). However, it is usually better to describe $T$ constructively as in XI.5.1.9.

XI.5.1.11. The topology $T$ can be described in terms of convergence: a net $(x_i)$ in $X$ converges to $x$ in $(X, T)$ if and only if $f_j(x_i) \to f_j(x)$ in $(Y_j, S_j)$ for every $j \in J$. Thus $T$ is often called the weak topology on $X$ defined by the $f_j$.

XI.5.1.12. One of the most important examples of a weak topology defined by a family of functions is the product topology, i.e. the case where $X = \prod j Y_j$ and the $f_j$ are the coordinate projections. The general situation can also be reduced to the case of one target space using the product topology: if $X$ is a set, $\{(Y_j, S_j) : j \in J\}$ is a family of topological spaces (not necessarily distinct), and $f_j : X \to Y_j$ is a function for each $j \in J$, let $Y = \prod j Y_j$ and $f : X \to Y$ the induced map, i.e. $(f(x))_j = f_j(x)$ for all $j$. Give $Y$ the product topology. Then the weak topology on $X$ from the $f_j$ is the weak topology on $X$ from $f$.
XI.5.1.13. If all the $Y_j$ are Hausdorff and the $f_j$ separate points of $X$, then the weak topology on $X$ defined by the $f_j$ is also Hausdorff: if $p, q \in X$, $p \neq q$, choose $j$ with $f_j(p) \neq f_j(q)$, and choose disjoint neighborhoods $U$ and $V$ of $f_j(p)$ and $f_j(q)$ in $Y_j$; then $f_j^{-1}(U)$ and $f_j^{-1}(V)$ are disjoint neighborhoods of $p$ and $q$. The same statement is true if all the $Y_j$ are regular or completely regular ($\star$).

XI.5.1.14. Similarly, if $Y$ is a set, $\{(X_j, T_j) : j \in J\}$ is a set of topological spaces, and $f_j : X_j \to Y$ is a function for each $j$, then there is a strongest topology on $Y$ making all the $f_j$ continuous:

$$S = \{ V \subseteq Y : f_j^{-1}(V) \in T_j \text{ for all } j \} .$$

It is easy to verify that $S$ is a topology on $Y$, and it is obviously the strongest one making all the $f_j$ continuous. However, this topology is often very weak (frequently even the indiscrete topology). The quotient topology (XI.8.1.1.) is an example of this construction, probably the most important example.
XI.6. Products and the Product Topology

XI.6.1. Exercises

XI.6.1.1. Let \( I \) be an uncountable index set, and \( \{ X_i : i \in I \} \) a collection of topological spaces. Suppose for each \( i \) there is a point \( p_i \in X_i \) which has an open neighborhood \( U_i \) which is not all of \( X_i \) (there will be such a \( p_i \) unless \( X_i \) has the indiscrete topology). Let \( X = \prod_{i \in I} X_i \) with the product topology, and \( p = (\cdots p_i \cdots) \in X \). Show that \( p \) does not have a countable neighborhood base, and hence \( X \) is not first countable. [If \( \{ V_n : n \in \mathbb{N} \} \) is a sequence of open neighborhoods of \( p \), then for each \( n \) there is a finite subset \( F_n \) of \( I \) and nonempty open sets \( W_{n_i} \) such that \( V_n \) contains \( \prod_{i \in F_n} W_{n_i} \times \prod_{i \notin F_n} X_i \). If \( j \in I \setminus (\cup_{n=1}^{\infty} F_n) \), then \( U_j \times \prod_{i \neq j} X_i \) does not contain any \( V_n \).]

XI.6.1.2. Let \( \{ X_i : i \in I \} \) be a collection of topological spaces, and let \( X = \prod_{i \in I} X_i \) with the product topology. Suppose each \( X_i \) is separable and \( \text{card}(I) \leq 2^\mathfrak{c} \). Show that \( X \) is separable. [Let \( \{ p_{in} : n \in \mathbb{N} \} \) be a countable dense set in \( X_i \) for each \( i \). Identify \( I \) with a subset of \( \mathbb{R} \). For each finite set \( S = \{(a_1, b_1), \ldots, (a_m, b_m)\} \) of disjoint open intervals in \( \mathbb{R} \) with rational endpoints, and every corresponding tuple \( N = (n_1, \ldots, n_m) \) of natural numbers, let \( p_{(S,N)} \) be the point of \( x \) whose \( i \)'th coordinate is \( p_{in_k} \) if \( i \in (a_k, b_k) \) and \( p_{i1} \) otherwise. Show that the set of all \( p_{(S,N)} \) is a countable dense set in \( X \).]

This result requires the \( \text{AC} \) for index sets of cardinality \( 2^\mathfrak{c} \).

XI.6.1.3. Let \( D \) be a countable dense set in \([0,1]^\mathfrak{c}\). Then \( D \), with the relative topology, is a countable normal topological space in which no point has a countable local base.

XI.6.1.4. Let \( \{ X_i : i \in I \} \) be a collection of topological spaces, and let \( X = \prod_{i \in I} X_i \) with the product topology. Suppose each \( X_i \) has two disjoint nonempty open sets \( U_i \) and \( V_i \) (this will happen if \( X_i \) is Hausdorff and has more than one point), and that \( X \) is separable. Show that \( \text{card}(I) \leq 2^\mathfrak{c} \). [If \( \{ p_n : n \in \mathbb{N} \} \) is dense in \( X \), for each \( i \) let \( S(i) \) be the set of all \( n \) such that the \( i \)'th coordinate of \( p_n \) is in \( U_i \). Show that \( S(i) \) and \( S(j) \) are distinct subsets of \( \mathbb{N} \) for \( i \neq j \).]

XI.6.1.5. Let \( X \) be an arbitrary product of copies of \( \mathbb{N} \) with the finite complement topology. Show that \( X \) is separable in the product topology. [For each \( n \in \mathbb{N} \) let \( p_n \) be the point of \( X \) with all coordinates \( n \).]

XI.6.1.6. \((a)\) \((\text{Sto}48; \text{cf. Eng}89, 2.3.E)\) Show that \( \mathbb{N}^\mathfrak{c} \) is not normal. [For \( j = 1, 2 \) let \( A_j \) be the set of all points for which, for each \( k \neq j \), at most one coordinate is \( k \). Show that \( A_1 \) and \( A_2 \) are closed and disjoint, but do not have disjoint neighborhoods.]

(b) Conclude from XI.11.12.9. that if \( I \) is an index set with \( \text{card}(I) \geq \aleph_1 \), and, for each \( i \in I \), \( X_i \) is a topological space which is not countably compact, then \( \prod_{i \in I} X_i \) is not normal. In particular, if \( X \) is a topological space which is not countably compact, then \( X^\mathfrak{c} \) is not normal.
XI.7. Separation Axioms

The notion of topological space is so general that practically nothing can be said about all topological spaces. All nontrivial results in topology require additional restrictions, usually including separation axioms. Separation axioms roughly insure that there are “enough” open sets in the topology. There are several important separation axioms of varying strength, described in this section. Virtually all topological spaces arising in applications, particularly in analysis, including any metrizable space, satisfy all these separation axioms.

Note that the separation axioms have nothing to do with separability in the sense of XI.2.3.11.

XI.7.1. The $T$-Axioms

The first five separation axioms first appeared together in [AH74], where they were called $T_0–T_4$ ($T$ for Trennungaxiom); the authors attributed each property to someone else (Kolmogorov, Fréchet, Hausdorff, Vietoris, and Tietze respectively). Alternate names for $T_2–T_4$ are usually used today. There are also a $T_{3rac{1}{2}}$ (XI.7.7.3.), and a $T_5$, discussed in XI.10.1.2.; some add $T_{2rac{1}{2}}$ (XI.7.8.17.) and $T_6$ (XI.10.4.10.).

XI.7.1.1. Definition. Let $X$ be a topological space. Then

(o) $X$ is $T_0$ if, whenever $p$ and $q$ are distinct points of $X$, there is an open neighborhood of at least one of the points not containing the other.

(i) $X$ is $T_1$ if each singleton subset of $X$ is closed.

(ii) $X$ is $T_2$ if any two distinct points of $X$ have disjoint open neighborhoods.

(iii) $X$ is $T_3$ if whenever $E$ is a closed set in $X$ and $p$ a point of $X$ not in $E$, then $p$ and $E$ have disjoint open neighborhoods.

(iv) $X$ is $T_4$ if disjoint closed sets in $X$ have disjoint open neighborhoods.

$X$ is Hausdorff if it is $T_2$.

$X$ is regular if it is $T_3$ and Hausdorff.

$X$ is normal if it is $T_4$ and Hausdorff.

Each of these separation axioms is discussed in more detail below.

XI.7.1.2. There is some nonuniformity of use of the terminology in this definition. Some references define “regular” and “normal” to mean what we have called $T_3$ and $T_4$, and define $T_3$ to mean “$T_2$ and regular”, and similarly for $T_4$. We will regard “$T_3$” and “regular” as essentially synonymous (since they differ only in the case of spaces which are not $T_0$ (XI.7.5.2.)), and similarly “$T_4$” and “normal” are essentially equivalent.

Some references use the term separated as a synonym for “Hausdorff.” But this term is too easily confused with “separable” as defined in XI.2.3.11.
XI.7.2.   $T_0$-Spaces

XI.7.2.1.   PROPOSITION. Let $X$ be a topological space, and $p$ and $q$ distinct points of $X$. The following are equivalent:

(i) $\{p\}$ and $\{q\}$ have identical closures.

(ii) $p$ is in the closure of $\{q\}$ and $q$ is in the closure of $\{p\}$.

(iii) Any open set containing $p$ also contains $q$, and vice versa.

$X$ is $T_0$ if and only if none of these occur for any distinct $p$ and $q$.

PROOF: (i) $\Rightarrow$ (iii): if $U$ is an open set containing $p$ but not $q$, then $U^c$ is a closed set containing $q$ and hence containing the closure of $\{q\}$, so the closure of $\{q\}$ is not the same as the closure of $\{p\}$ (in fact, does not contain $p$). The same argument works with $p$ and $q$ interchanged.

(iii) $\Rightarrow$ (ii): If $p$ is not in the closure of $\{q\}$, the complement of the closure of $\{q\}$ is an open set containing $p$ but not $q$. The same argument works if $p$ and $q$ are interchanged.

(ii) $\Rightarrow$ (i) is trivial.

It is obvious that the $T_0$ condition is precisely the negation of (iii).

XI.7.2.2.   COROLLARY. A $T_1$ topological space is $T_0$.

There are $T_0$ spaces which are not $T_1$, e.g. the Sierpiński space.

The $T_0$ property is a local property. Say that a topological space $X$ is locally $T_0$ if every point of $X$ has an open neighborhood which is $T_0$ in the relative topology. Every $T_0$ space is obviously locally $T_0$. The converse is also true:

XI.7.2.3.   PROPOSITION. A locally $T_0$ space is $T_0$.

PROOF: Let $X$ be a locally $T_0$ space, and $p, q$ distinct points of $X$. We must find an open set containing one of $p$ and $q$ and not the other. Let $U$ be an open neighborhood of $p$ which is $T_0$ in the relative topology. If $q \notin U$, then $U$ is the desired set. If $q \in U$, there is a relatively open subset $V$ of $U$ containing one of $p$ and $q$ but not the other. Since $U$ is open in $X$, $V$ is also open in $X$ (.), and is the desired subset.

XI.7.2.4.   Every topological space has a maximal $T_0$ quotient, which captures all the topological information of the space (XI.8.3.10.).
XI.7.2.5. There is really no reason to consider topological spaces which are not at least $T_0$ (especially in light of XI.8.3.10.). Many authors feel the same about restricting to $T_1$ or $T_2$ spaces; indeed, topological spaces which are not Hausdorff are rather bizarre and are rarely encountered. In analysis almost all spaces encountered are even completely regular. However, there are topological spaces arising in applications in algebra and logic, such as spectra of rings (XI.7.8.16., XV.14.4.39.), which are $T_0$ but not necessarily even $T_1$, so such spaces cannot be completely ignored.

But there is one situation in analysis where spaces which are not $T_0$ arise naturally: certain natural topological groups () and topological vector spaces () are not $T_0$. In this case, taking the maximal $T_0$ quotient amounts to dividing out by the closure of the identity (), hence the $T_0$ quotient is also naturally a topological group or topological vector space.

XI.7.3. $T_1$-Spaces

The $T_1$ property has an alternate characterization, which was actually the original definition:

XI.7.3.1. Proposition. A topological space $X$ is $T_1$ if and only if, whenever $p$ and $q$ are distinct points of $X$, each of $p$ and $q$ has an open neighborhood not containing the other.

Proof: If $X$ is $T_1$ and $p$ and $q$ are distinct points of $X$, then $X \setminus \{q\}$ and $X \setminus \{p\}$ are the desired neighborhoods of $p$ and $q$. Conversely, let $p \in X$. For each $q \in X \setminus \{p\}$, there is an open neighborhood $U_q$ of $q$ not containing $p$. The union of the $U_q$ is open, and equals $X \setminus \{p\}$, so $\{p\}$ is closed.

This characterization gives a simpler proof that a $T_1$ space is $T_0$, and also immediately implies that a Hausdorff space is $T_1$. We used the definition we did for $T_1$ since it is the version which is usually verified and used in applications.

XI.7.3.2. Note that the condition in XI.7.3.1. can be rephrased: whenever $p$ and $q$ are distinct points of $X$, then $p$ has an open neighborhood not containing $q$. For if this statement is true for any $p$ and $q$ (in that order), it is also true for $q$ and $p$, i.e. each of $p$ and $q$ has a neighborhood not containing the other.

XI.7.3.3. Corollary. A topological space $X$ is $T_1$ if and only if every singleton subset of $X$ is an intersection of open sets.

XI.7.3.4. Like the $T_0$ property, the $T_1$ property is a local property. Say that a topological space $X$ is locally $T_1$ if every point of $X$ has an open neighborhood which is $T_1$ in the relative topology. Every $T_1$ space is obviously locally $T_1$. The converse is also true:

XI.7.3.5. Proposition. A locally $T_1$ space is $T_1$.

The proof is a slightly simplified version of the argument in the proof of XI.7.2.3., and is left to the reader.
XI.7.4. Hausdorff Spaces

The Hausdorff property is probably the most widely used separation axiom. Here is one of the main reasons. One of the most basic facts about limits in $\mathbb{R}$ or $\mathbb{R}^n$, or in general metric spaces, is that limits are unique (\textit{\textendash}). This property only holds in Hausdorff spaces:

**XI.7.4.1. Proposition.** Let $X$ be a topological space. Then $X$ is Hausdorff if and only if every convergent net in $X$ has a unique limit.

**Proof:** Suppose $X$ is Hausdorff, and $(x_i)$ is a net in $X$ with $x_i \to x$ and $x_i \to y$ with $x \neq y$. Let $U$ and $V$ be disjoint neighborhoods of $x$ and $y$ respectively. Then there is an $i_1$ such that $x_i \in U$ for all $i \geq i_1$, and there is an $i_2$ such that $x_i \in V$ for all $i \geq i_2$. If $i \geq i_1$ and $i \geq i_2$ (there is such an $i$ since the index set is directed), then $x_i \in U \cap V$, a contradiction. Thus the limit of a convergent net is unique.

Conversely, suppose $X$ is not Hausdorff, and let $x$ and $y$ be distinct points which do not have disjoint neighborhoods. Let $I = \{(U, V, z)\}$, where $U$ is a neighborhood of $x$, $V$ is a neighborhood of $y$, and $z \in U \cap V$. Put a preorder on $I$ by setting $(U_1, V_1, z_1) \leq (U_2, V_2, z_2)$ if $U_2 \subseteq U_1$ and $V_2 \subseteq V_1$. Then $I$ is a directed set. Set $x_{(U, V, z)} = z$; this defines a net with index set $I$, and this net converges to both $x$ and $y$. So limits of convergent nets are not unique.

**XI.7.4.2.** Note that the “if” part can be false if “net” is replaced by “sequence” if the space is not first countable (Exercise XI.7.8.12.). In a first countable space, the sequence version is true (\textit{\textendash}).

**XI.7.4.3.** Unlike the $T_0$ and $T_1$ properties, the Hausdorff property is not a local property. Say that a topological space $X$ is \textit{locally Hausdorff} if every point of $X$ has an open neighborhood which is Hausdorff in the relative topology. A Hausdorff space is obviously locally Hausdorff. But a locally Hausdorff space need not be Hausdorff (XI.19.1.3.(v)), although it is $T_1$ (XI.7.3.5.). However, see XI.7.8.4.

XI.7.5. Regular Spaces

The $T_3$ separation axiom has a simple reformulation which reinforces the utility of the axiom:

**XI.7.5.1. Proposition.** Let $X$ be a topological space. Then $X$ is a $T_3$ space if and only if, whenever $p$ is a point of $X$ and $U$ is an open neighborhood of $p$, there is an open neighborhood $V$ of $p$ with $\overline{V} \subseteq U$.

**Proof:** Suppose $X$ is $T_3$, and $U$ is an open neighborhood of $p$. Then $U^c$ is a closed set not containing $p$, so $p$ and $U^c$ have disjoint open neighborhoods $V$ and $W$. Then $\overline{V}$ is contained in the closed set $W^c$, hence in $U$. Conversely, suppose $X$ has the property in the statement, and $E$ is a closed set in $X$ and $p$ is a point of $X$ not in $E$. Then $U = E^c$ is an open neighborhood of $p$, so there is an open neighborhood $V$ of $p$ with $\overline{V} \subseteq U$. Then $V$ and $(\overline{V})^c$ are disjoint open neighborhoods of $p$ and $E$, so $X$ is a $T_3$ space.

In the presence of $T_3$, only the $T_0$ axiom is needed to give the Hausdorff property and thus regularity:
XI.7.5.2. **Proposition.** Let $X$ be a topological space which is $T_3$ and $T_0$. Then $X$ is Hausdorff, and thus regular.

**Proof:** Let $p$ and $q$ be distinct points of $X$. Suppose $p$ is not in the closure $E$ of $\{q\}$. Then $p$ and $E$ have disjoint open neighborhoods, which are disjoint open neighborhoods of $p$ and $q$. The same proof works if $q$ is not in the closure of $\{p\}$. Apply XI.7.2.1.

A set with the indiscrete topology and more than one point is an example of a $T_3$ space which is not $T_0$.

XI.7.6. **Normal Spaces**

XI.7.6.1. In the presence of $T_4$, the $T_1$ condition gives $T_2$, hence normality, as is obvious. However, unlike with regularity, $T_0$ is not enough: the Sierpiński space is $T_0$ and vacuously $T_4$, but not $T_1$. It is also obvious that normality implies regularity, although $T_4$ does not imply $T_3$ in the absence of $T_1$ (the Sierpiński space is again a counterexample).

There is an analog of XI.7.5.1. for normal spaces. The proof is virtually identical.

XI.7.6.2. **Proposition.** Let $X$ be a topological space. Then $X$ is a $T_4$ space if and only if, whenever $E$ is a closed set in $X$ and $U$ is an open neighborhood of $E$, there is an open neighborhood $V$ of $E$ with $V \subseteq U$.

XI.7.6.3. **Proposition.** A closed subspace of a $T_4$ space is $T_4$. Hence a closed subspace of a normal space is normal.

**Proof:** If $A$ is a closed subspace of a $T_4$ space $X$, and $Y, Z$ are (relatively) closed subsets of $A$, then $Y$ and $Z$ are closed in $X$, hence have disjoint neighborhoods $U, V$ in $X$. Then $A \cap U$ and $A \cap V$ are disjoint neighborhoods in $A$.

A general subspace of a normal space is not necessarily normal (XI.11.12.6.). However, an $F_\sigma$ in a normal space is normal (XI.7.8.7.).

An alternate way of phrasing normality is using partitions:

XI.7.6.4. **Definition.** Let $X$ be a topological space, and $Y, Z$ disjoint closed subsets of $X$. A partition between $Y$ and $Z$ in $X$ is a (necessarily closed) subset $P$ of $X$ such that there are open sets $U, V \subseteq X$ with $Y \subseteq U, Z \subseteq V, U \cap V = \emptyset$, and $X \setminus (U \cup V) = P$.

Note that there can be many partitions between disjoint subsets. A partition can be empty: $\emptyset$ is a partition between $Y$ and $Z$ in $X$ if (and only if) there is a clopen set $U$ in $X$ with $Y \subseteq U$ and $Z \subseteq (X \setminus U)$. 1050
XI.7.6.5. **Proposition.** Let $X$ be a topological space. Then $X$ is $T_4$ if and only if every pair of disjoint closed subsets of $X$ has a partition in $X$.

**Proof:** If every pair of disjoint closed subsets of $X$ has a partition in $X$, and $Y, Z$ are disjoint closed subsets of $X$, let $P$ be a partition between $Y$ and $Z$ in $X$; the associated $U$ and $V$ are disjoint neighborhoods of $Y$ and $Z$. Thus $X$ is $T_4$. Conversely, suppose $X$ is $T_4$ and $Y$ and $Z$ are disjoint closed subsets of $X$. Let $U$ and $V$ be disjoint neighborhoods of $Y$ and $Z$ in $X$; then $P = X \setminus (U \cup V)$ is a partition between $Y$ and $Z$ in $X$. \(\Box\)

XI.7.6.6. **Proposition.** A metrizable topological space is normal.

In fact, a metrizable space has a stronger property called *complete normality*. See XI.10.1.1. for a definition and proof. In particular, a metrizable space is regular and Hausdorff.

Every compact Hausdorff space is also normal (XI.11.4.3.).

Here is a useful refinement of the definition of normality:

XI.7.6.7. **Definition.** Let $X$ be a topological space, and $\mathcal{U} = \{U_i : i \in I\}$ a cover of $X$. A *shrinking* of $\mathcal{U}$ is a cover $\mathcal{V} = \{V_i : i \in I\}$ (same index set) with $V_i \subseteq U_i$ for all $i$. An *open [closed] shrinking* is a shrinking in which $\mathcal{V}$ is an open [closed] cover.

XI.7.6.8. **Proposition.** Let $X$ be a topological space. Then $X$ is $T_4$ if and only if every finite open cover of $X$ has a closed shrinking.

**Proof:** If $X$ is not $T_4$, let $Y$ and $Z$ be disjoint closed sets in $X$ which do not have disjoint neighborhoods. Then the open cover $\{X \setminus Y, X \setminus Z\}$ does not have a closed shrinking.

Conversely, suppose $X$ is $T_4$. We show by induction on the number of sets in the cover that every finite open cover of $X$ has a closed shrinking. First suppose $\mathcal{U} = \{U_1, U_2\}$ is an open cover of $X$. Then $X \setminus U_1$ and $X \setminus U_2$ are disjoint closed sets in $X$, and so have disjoint open neighborhoods $V_1, V_2$. Then $\{X \setminus V_1, X \setminus V_2\}$ is a closed shrinking of $\mathcal{U}$.

Now suppose every open cover with $n$ sets has a closed shrinking, and let $\mathcal{U} = \{U_1, \ldots, U_{n+1}\}$ be an open cover of $X$. There is a closed shrinking $\{E_1, \ldots, E_{n-1}, F\}$ for the open cover $\{U_1, \ldots, U_{n-1}, U_n \cup U_{n+1}\}$. Then $F$ is a closed subset of $X$, hence $T_4$, so the open cover $\{F \cap U_n, F \cap U_{n+1}\}$ of $F$ has a closed shrinking $\{E_n, E_{n+1}\}$. Then $\{E_1, \ldots, E_{n+1}\}$ is a closed shrinking of $\mathcal{U}$. \(\Box\)

There is a nice “dual” fact about normal spaces. We first make a definition:

XI.7.6.9. **Definition.** Let $X$ be a topological space, and $A_1, \ldots, A_n$ closed subsets of $X$. A *swelling* of $\{A_1, \ldots, A_n\}$ is a collection $\{U_1, \ldots, U_n\}$ of open sets in $X$ with $A_k \subseteq U_k$ for each $k$ and $U_{k_1} \cap U_{k_2} \cap \cdots \cap U_{k_m} = \emptyset$ for any $k_1, \ldots, k_m$ with $A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m} = \emptyset$. 1051
XI.7.6.10. PROPOSITION. Let \( X \) be a \( T_4 \) topological space, and \( A_1, \ldots, A_n \) closed subsets of \( X \). Then \( \{A_1, \ldots, A_n\} \) has a swelling.

PROOF: Let \( B_1 \) be the union of all sets of the form \( A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m} \) for which
\[
A_1 \cap A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m} = \emptyset .
\]

Since this is a finite union, \( B_1 \) is closed; and \( A_1 \) and \( B_1 \) are disjoint. By normality there are disjoint open neighborhoods \( U_1 \) and \( V_1 \) of \( A_1 \) and \( B_1 \), so that \( U_1 \) is disjoint from \( B_1 \).

Now repeat the process with \( A_2 \), with \( A_1 \) replaced by \( U_1 \): let \( B_2 \) be the union of all finite intersections of \( \{U_1, A_3, \ldots, A_n\} \) which are disjoint from \( A_2 \), and obtain an open neighborhood \( U_2 \) of \( A_2 \) with \( U_2 \) disjoint from \( B_2 \). Repeating the process \( n \) times (at the \( k \)'th step \( A_j \) is replaced by \( U_j \) for all \( j < k \)), we obtain a swelling \( \{U_1, \ldots, U_n\} \).

There are two important theorems about normality with nontrivial proofs. The first is due to ():

XI.7.6.11. THEOREM. A \( T_3 \) Lindelöf space is \( T_4 \). Thus a regular Lindelöf space is normal.

In fact, a stronger result holds: every regular Lindelöf space is paracompact (), and every paracompact space is normal (). But we give a straightforward proof of XI.7.6.11. which is simpler than the proofs of these two stronger results.

PROOF: Let \( X \) be a \( T_3 \) Lindelöf space, and \( E \) and \( F \) disjoint closed sets in \( X \). For each \( p \in E \), there is an open neighborhood \( W_p \) of \( p \) such that \( W_p \cap F = \emptyset \) (XI.7.5.1.). Then \( \{W_p : p \in E\} \cup \{E^c\} \) is an open cover of \( X \), so has a countable subcover \( \{E^c, W_1, W_2, \ldots\} \). Similarly, there is a countable cover \( \{F^c, Z_1, Z_2, \ldots\} \) of \( X \) with \( Z_n \cap E = \emptyset \) for all \( n \). We have \( E \subseteq \bigcup_n W_n \) and \( F \subseteq \bigcup_n Z_n \). Inductively define \( U_1 = W_1 \), \( V_1 = Z_1 \setminus U_1 \), \( U_2 = W_2 \setminus V_1 \), \( V_2 = Z_2 \setminus (U_1 \cup U_2) \), \( U_3 = W_3 \setminus (V_1 \cup V_2) \), etc. Each \( U_n \) and \( V_n \) is open, and \( U_n \cap V_n = \emptyset \) for all \( n \) and \( m \). If \( p \in E \), then \( p \in W_n \) for some \( n \), hence \( p \in U_n \) since \( p \notin V_m \subseteq Z_m \) for any \( m \). Similarly, if \( q \in F \), then \( q \in V_m \) for some \( m \). Thus \( U = \bigcup_n U_n \) and \( V = \bigcup_n V_n \) are disjoint neighborhoods of \( E \) and \( F \).

The second theorem is a key result showing that normal spaces can be completely studied by analytic methods (specifically, real-valued continuous functions):

XI.7.6.12. THEOREM. [URYSOHN’S LEMMA] Let \( X \) be a normal topological space. If \( E_0 \) and \( E_1 \) are disjoint closed sets in \( X \), then there is a continuous function \( f : X \to [0, 1] \) with \( f(x) = 0 \) for all \( x \in E_0 \) and \( f(x) = 1 \) for all \( x \in E_1 \).

This result is a special case of the Tietze Extension Theorem (), but we give a direct proof which is somewhat simpler.

PROOF: We will make repeated use of XI.7.6.2. Let \( U_0 = E_0^c \). Then \( U_0 \) is an open neighborhood of \( E_1 \). Let \( U_{1/2} \) be an open neighborhood of \( E_1 \) with \( U_{1/2} \subseteq U_0 \). Let \( U_{3/4} \) be an open neighborhood of \( E_1 \) with \( U_{3/4} \subseteq U_{1/2} \). \( U_0 \) is an open neighborhood of \( U_{1/2} \); let \( U_{1/4} \) be an open neighborhood of \( U_{1/2} \) with \( U_{1/4} \subseteq U_0 \). Continuing in this way, for each dyadic rational \( t \) with \( 0 \leq t < 1 \) choose an open neighborhood \( U_t \) of \( E_1 \) such that \( U_s \subseteq U_t \) if \( s < t \).
For $x \in X$ define $f(x) = \sup \{t : x \in U_t\}$ if $x \in U_0$ and $f(x) = 0$ if $x \in E_0$. Then $0 \leq f(x) \leq 1$ for all $x \in X$, and $f(x) = 1$ if $x \in E_1$. We show $f$ is continuous. If $(x_i)$ is a net in $X$ with $x_i \to x$, we show that $f(x_i) \to f(x)$. Let $t = f(x)$. If $t > 0$, then for any dyadic rational $s$, $0 \leq s < t$, we have $x \in U_s$; since $U_s$ is open, $x_i \in U_s$ for sufficiently large $i$, i.e. $f(x_i) \geq s$ for sufficiently large $i$. This shows that $f(x_i) \to f(x)$ if $t = 1$. Now suppose $t < 1$. If $s$ is any dyadic rational with $t < s < 1$, choose a dyadic rational $r$ with $t < r < s$. Then $x \notin U_r$. Since $\bar{U}_s \subseteq U_r$, $x$ is in the open set $(\bar{U}_s)^c$, so $x_i \in (\bar{U}_s)^c$ for sufficiently large $i$, i.e. $f(x_i) \leq s$ for sufficiently large $i$. Thus $f(x_i) \to f(x).$  

The Countable AC is used in this proof.

\textbf{XI.7.6.13.} Normality is a nice separation condition to have, but it is somewhat more delicate than the other conditions: for example, unlike all the other separation axioms it does not pass to subspaces or products in general (see XI.11.12.6. and XI.7.8.10. for counterexamples; cf. XI.10.1.2.).

\textbf{XI.7.7. Completely Regular Spaces}

Another separation axiom is commonly added to the hierarchy:

\textbf{XI.7.7.1. Definition.} Let $X$ be a topological space. Then $X$ is \textit{completely regular} if it is $T_0$ and, whenever $E$ is a closed set in $X$ and $p$ a point of $X$ not in $E$, there is a continuous function $f : X \to [0, 1]$ such that $f(p) = 1$ and $f(x) = 0$ for all $x \in E$.

\textbf{XI.7.7.2. Proposition.} A completely regular space is regular (and in particular Hausdorff).

\textbf{Proof:} Suppose $X$ is completely regular. If $E$ is a closed set in $X$ and $p$ a point of $X$ not in $E$, and $f$ is a continuous function from $X$ to $[0, 1]$ as in XI.7.7.1., then $f^{-1}((1/2, 1])$ and $f^{-1}([0, 1/2))$ are disjoint open neighborhoods of $p$ and $E$, so $X$ is $T_3$. Since $X$ is $T_0$, it is regular. 

\textbf{XI.7.7.3.} A normal space is completely regular by Urysohn’s Lemma. Thus the completely regular condition is a separation axiom between $T_3$ and $T_4$, and is sometimes called $T_{3 \frac{1}{2}}$ (with the usual ambiguity about whether the space is required to be Hausdorff). A completely regular space is also sometimes called a \textit{Tikhonov space}.

There are regular spaces which are not completely regular (see [SS95]), and completely regular spaces which are not normal, e.g. XI.19.5.8.

\textbf{XI.7.7.4. Theorem.} Let $X$ be a topological space. Then $X$ is homeomorphic to a subset of a compact Hausdorff space if and only if $X$ is completely regular.

\textbf{Proof:} Since compact Hausdorff spaces are normal (XI.11.4.3.) and hence completely regular, and complete regularity obviously passes to subspaces, the complete regularity condition is necessary. Conversely, let $X$ be completely regular. Let $\mathcal{C}$ be a set of continuous functions from $X$ to $[0, 1]$ such that for every closed set $E$ in $X$ and $p \in X \setminus E$, there is an $f \in \mathcal{C}$ with $f(p) = 1$ and $f = 0$ on $E$, and form the space $Y = [0, 1]^\mathcal{C}$ with the product topology. $Y$ is Hausdorff, and compact by Tikhonov’s theorem. Define a function $\phi : X \to Y$ by
letting ϕ(x) be the point whose f’th coordinate is f(x), for each f ∈ C. ϕ is continuous since its composition with each coordinate projection is continuous. We need only show that it is a homeomorphism, e.g. if (x_i) is a net in X, and x ∈ X with ϕ(x_i) → ϕ(x), then x_i → x. We have ϕ(x_i) → ϕ(x) if and only if f(x_i) → f(x) for every f ∈ C. If x_i ≠ x, there is a neighborhood U of x and a subnet (x_j) such that x_j ∉ U for all j. There is an f ∈ C such that f(x) = 1 and f(y) = 0 for all y ∈ U^c; for this f, f(x_j) = 0 for all j, so f(x_i) ≠ f(x), a contradiction.

XI.7.7.5. A good case can be made that complete regularity is the most natural restriction for topological spaces, and that completely regular spaces form the proper class of spaces to consider, especially for applications in analysis. This is the general thesis of [GJ76], for example. Completely regular spaces are precisely the spaces characterized by each of the following conditions:

(i) Spaces which can be embedded in compact Hausdorff spaces.
(ii) Spaces with a topology arising from a uniform structure.
(iii) Spaces whose topology is defined by a (separating) family of pseudometrics.
(iv) Spaces whose structure can be completely analyzed by study of real-valued continuous functions (i.e. analytic methods).

XI.7.8. Exercises

XI.7.8.1. Let X be a topological space. Show that the following are equivalent:

(i) Every singleton subset of X is an intersection of open sets.
(ii) Every subset of X is an intersection of open sets.
(iii) X is T_1.

XI.7.8.2. Let X = {a, b}, and let T = {∅, {a}, X}.

(a) Show that T is a topology on X.
(b) {b} is closed, but {a} is not closed; {a} is open, and the closure of {a} is X (i.e. {a} is dense in X).
(c) (X, T) satisfies T_0 and T_4, but not T_1, T_2, or T_3.
(d) If x_n = a for all n, then x_n → a and x_n → b.
(e) X is compact () and connected (). In fact, X is path-connected ()

(X, T) is called the Sierpiński space.

XI.7.8.3. (a) Let X be a T_0 space, and U a finite nonempty open subset of X. Show that U contains an open singleton subset of X. [Consider a minimal nonempty open subset of U.]
(b) Let X be a T_1 space, and U a finite nonempty open subset of X. Show that every point of U is an isolated point of X. In particular, a finite T_1 space has the discrete topology.

A T_0 space containing a nonempty finite open set need not have any isolated points, and not every singleton subset of a finite open set is open in general. The Sierpiński space is a counterexample to both.
XI.7.8.4. Let $X$ be a topological space.

(a) If each point of $X$ has a closed neighborhood which is Hausdorff in the relative topology, show that $X$ is Hausdorff.

(b) If each point of $X$ has a closed neighborhood which is $T_3$ in the relative topology, show that $X$ is $T_3$.

(c) (cf. [Eng89, 3.3.1]) If each point of $X$ has a closed neighborhood which is $T_{3\frac{1}{2}}$ in the relative topology, show that $X$ is $T_{3\frac{1}{2}}$.

(d) If each point of $X$ has a closed neighborhood which is [completely] regular in the relative topology, show that $X$ is [completely] regular.

Example XI.19.1.3.(v) shows that the word “closed” cannot be removed from any of the statements. Since there are locally compact Hausdorff spaces which are not normal (XI.11.12.6.), the corresponding statements with “$T_4$” or “normal” do not hold.

XI.7.8.5. (a) Let $X$ be a Hausdorff space. Suppose $X$ has a dense subspace $D$ of cardinality $\kappa$. Show that $\text{card}(X) \leq 2^{2^\kappa}$. Thus $\text{card}(X) \leq 2^{2^{|X|}}$ (XI.2.3.16.). [For each $p \in X$, let $\mathcal{A}_p = \{U \cap D : U$ an open neighborhood of $p\} \in \mathcal{P}(\mathcal{P}(D))\). Show that $\mathcal{A}_p \neq \mathcal{A}_q$ if $p \neq q$. In particular, a separable Hausdorff space has cardinality $\leq 2^{2^{\aleph_0}}$.

There are separable Hausdorff spaces of cardinality $2^{2^{\aleph_0}}$, even compact ones, e.g. $\beta\mathbb{N}$ and $[0,1]^\mathbb{C}$.

(b) Show that if $X$ is any set with the finite complement topology ($\mathcal{O}$), then any infinite subset of $X$ is dense, and hence $X$ is separable if $\text{card}(X) \geq \aleph_0$. Thus the conclusion of (a) fails in general for compact $T_1$ spaces, and there is no bound at all on the cardinality of such spaces.

(c) If $X$ is a separable, first countable Hausdorff space, show that $X$ has cardinality $\leq 2^{\aleph_0}$. [If $D$ is a countable dense set in $X$, for each $x \in X$ choose a sequence in $D$ converging to $x$.] In general, if $X$ is Hausdorff, we have $\text{card}(X) \leq d(X)^{\chi(X)}$ (XI.1.2.11., XI.2.3.16.).

Note that unlike for (a), the Axiom of Choice is needed for (c).

XI.7.8.6. (a) Let $X$ be a regular space. Suppose $X$ has a dense subspace $D$ of cardinality $\kappa$. Show there is a base for the topology of cardinality $\leq 2^\kappa$. Thus $\mathcal{W}(X) \leq 2^{d(X)}$ (XI.1.2.11., XI.2.3.16.). [If $\{U_i : i \in I\}$ is a base, for each $i$ set $V_i = [(U_i \cap D)^-]^{\mathcal{O}}$. Note that $U_i \subseteq V_i \subseteq \bar{U}_i$. Show that $\{V_i : i \in I\}$ is a base of cardinality $\leq 2^\kappa$.]

(b) If $X$ is a $T_0$-space, and there is a base for the topology of cardinality $\kappa$, then $\text{card}(X) \leq 2^\kappa$. Thus $\text{card}(X) \leq 2^{d(X)}$ (XI.1.2.11.). [If $\mathcal{U}$ is the base, associate to each $p$ the collection $\mathcal{U}_p$ of all elements of $\mathcal{U}$ containing $p$.]

(c) Combine (a) and (b) to give an alternate proof of XI.7.8.5.(a) for regular spaces.

XI.7.8.7. (a) Let $X$ be a $T_1$ topological space. Show that $X$ is normal if and only if, whenever $E$ is a closed set in $X$ and $U$ an open set in $X$ containing $E$, there is a sequence $(V_n)$ of open sets in $X$ with $V_n \subseteq U$ for all $n$ and $E \subseteq \bigcup_{n=1}^{\infty} V_n$. If $A$ and $B$ are disjoint closed subsets of $X$, choose such sequences $(V_n)$ for $A$ in $X \setminus B$ and $(W_n)$ for $B$ in $X \setminus A$. Then

$$\bigcup_{n=1}^{\infty} \left( V_n \setminus \left[ \bigcup_{k \leq n} \bar{W}_k \right] \right) \quad \text{and} \quad \bigcup_{n=1}^{\infty} \left( W_n \setminus \left[ \bigcup_{k \leq n} \bar{V}_k \right] \right)$$
are disjoint neighborhoods of $A$ and $B$.]
(b) Use (a) to show that every $F_{\sigma}$ in a normal space is normal.

**XI.7.8.8.** [Pea75] Make the following definitions:

**Definition.** Let $X$ be a topological space.

A subset $Y$ of $X$ is a *generalized $F_{\sigma}$* in $X$ if, whenever $U$ is an open neighborhood of $Y$ in $X$, there is an $F$ which is an $F_{\sigma}$ in $X$ with $Y \subseteq F \subseteq U$.

A subset $Y$ of $X$ is *normally situated* in $X$ if, whenever $U$ is an open neighborhood of $Y$ in $X$, there is a locally finite cover $\{V_i : i \in I\}$ of $Y$ with each $V_i$ an open $F_{\sigma}$ in $X$ contained in $U$.

$X$ is *totally normal* if $X$ is normal and every subset of $X$ is normally situated in $X$.

(a) Show that if $X$ is normal, then every generalized $F_{\sigma}$ in $X$ is normally situated in $X$. [Use (a).]

(b) Show that if $X$ is normal and $Y$ is normally situated in $X$, then $Y$ is normal.

(c) Show that every subspace of a totally normal space is totally normal.

(d) Show that every totally normal space is completely normal (XI.10.1.2.).

(e) Show that the space of ordinals less than or equal to the first uncountable ordinal (which is completely normal by XI.16.1.1.) is not totally normal (cf. XI.11.12.10.).

(f) Show that every perfectly normal space (XI.10.4.10.) is totally normal. Give an example of a totally normal space which is not perfectly normal.

(g) What is the relationship between total normality and (hereditary) paracompactness (XI.11.10.3.)?

**XI.7.8.9.** Show that a separable normal space cannot have a closed discrete subspace of cardinality $2^{\aleph_0}$. [Count the number of real-valued continuous functions and apply Urysohn’s Lemma.] More generally, if $X$ is normal and has a dense set of cardinality $\kappa$, and a closed discrete subspace of cardinality $\lambda$, then $2^\lambda \leq 2^\kappa$; in particular, $X$ cannot have a closed discrete subspace of cardinality $2^\kappa$ (Jones’s Lemma).

**XI.7.8.10.** The Sorgenfrey Line. Put a topology $S$ on $\mathbb{R}$ by taking $\{(a,b) : a, b \in \mathbb{R}, a < b\}$ as a base.

(Many references use base sets of the form $[a,b)$, which gives a different but homeomorphic topology.) $(\mathbb{R}, S)$ is called the Sorgenfrey line, denoted $S$.

(a) Show that $S$ is stronger than the usual topology on $\mathbb{R}$, and thus is Hausdorff.

(b) Show that $S$ is zero-dimensional, hence normal. [Show that $(a,b)$ is clopen for all $a < b$.]

(c) Show that any bounded nondecreasing sequence in $\mathbb{R}$ converges in $S$ to its limit in $\mathbb{R}$, but that no strictly decreasing sequence converges in $S$.

(d) Show that $S$ is first countable and separable.

(e) Show that every subset $A$ of $S$ is separable and Lindelöf. [Show that $\{x \in A : (x - \epsilon, x) \cap A = \emptyset \text{ for some } \epsilon > 0\}$ is countable. Then show that if $U$ is an open cover of $A$, and $V$ is the union of all intervals $(a,b)$ for which $(a,b)$ is contained in some element of $U$, then $A \setminus V$ is countable.] In particular, $S$ is (hereditarily) paracompact.

(f) Show that every closed subset of $S$ is a $G_\delta$. Thus $S$ is perfectly normal (XI.10.4.10.), hence completely normal.
(g) Show that a subset of $S$ is compact if and only if it is closed in the usual topology of $\mathbb{R}$, bounded above, and has no strictly decreasing sequences (i.e. is well ordered in the usual ordering). Conclude that every compact subset of $S$ is countable (cf. ()). In particular, $S$ is not locally compact or $\sigma$-compact.

(h) Show that $S \times S$ is not normal. [The set \{(x, -x) : x \in S\} is a closed discrete subset of $S \times S$ of cardinality $2^{\aleph_0}$. Apply Exercise XI.7.8.9. Alternatively, show using the Baire Category Theorem that \{(x, -x) : x \in \mathbb{Q}\} and \{(x, -x) : x \in \mathbb{I}\} do not have disjoint neighborhoods.] So $S \times S$ is not paracompact or Lindelöf. Note that $S \times S$ is separable, first countable, and completely regular.

(i) Show that $S$ is not second countable. [If it were, it would be metrizable, so $S \times S$ would also be metrizable. There is also a simple direct proof.]

(j) Show that $S$ is a Baire space ().

(k) Give $\mathbb{R} \times \{0, 1\}$ the lexicographic ordering and the order topology. Show that $S$ is homeomorphic to the subspace $\mathbb{R} \times \{0\}$ with the relative topology. Thus $S$ is a suborderable space (XI.1.6.4), and hence completely collectionwise normal (XI.16.1.1). But $S$ is not an orderable space.

The Sorgenfrey line provides a simple counterexample for many seemingly plausible statements, particularly about product spaces. If $T$ is the subset of $S$ consisting of irrational numbers, with the relative topology, it is plausible that $T$ is homeomorphic to $S$ since $S$ is zero-dimensional; but in fact $T$ is not homeomorphic to $S$ [?].

XI.7.8.11. Let $\mathcal{T}$ be the usual topology on $\mathbb{R}$ and $S$ the Sorgenfrey topology (XI.7.8.10.).

(a) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, S)$ to $(\mathbb{R}, T)$?

(b) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, T)$ to $(\mathbb{R}, S)$?

(c) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, S)$ to $(\mathbb{R}, S)$?

XI.7.8.12. Let $X$ be an uncountable set with the countable complement topology (XI.1.6.4.).

(a) Show that $X$ is $T_1$ but not Hausdorff.

(b) Show that a sequence in $X$ converges if and only if it is eventually constant. Thus the convergent sequences in this topology are the same as the convergent sequences in the discrete topology.

(c) Every convergent sequence in $X$ has a unique limit. Contrast with XI.7.4.1..

XI.7.8.13. Smirnov’s Deleted Sequence Topology [SS95, #64]. Let $A = \{\frac{1}{n} : n \in \mathbb{N}\} \subseteq \mathbb{R}$. Let $\mathcal{T}$ be the usual topology on $\mathbb{R}$. Define a topology $\mathcal{S}$ on $\mathbb{R}$ by

$$\mathcal{S} = \{U \setminus B : U \in \mathcal{T}, B \subseteq A\} .$$

(a) Show that $\mathcal{S}$ is a topology on $\mathbb{R}$ which is stronger than $\mathcal{T}$.

(b) Show that $(\mathbb{R}, \mathcal{S})$ is Hausdorff.

(c) Show that $(\mathbb{R}, \mathcal{S})$ is not regular. [$A$ is a closed set, $0 \notin A$, but $0$ and $A$ do not have disjoint neighborhoods.]

(d) Repeat the construction with $A$ replaced by any nonclosed set with dense complement, e.g. $\mathbb{Q}$. 1057
XI.7.8.14. **The Tangent Disk Space** [SS95, Example 82]. Let $X$ be the closed upper half plane in $\mathbb{R}^2$. Let $\mathcal{T}$ be the collection of all unions of subsets of $X$ of the following types:

(i) Any subset of the open upper half plane which is open in the ordinary topology.

(ii) The union of $\{(a,0)\}$ and an open disk of radius $r$ ($r > 0$) centered at $(a, r)$, for some $a \in \mathbb{R}$ (an open disk in the upper half plane tangent to the $x$-axis at $a$).

(a) Show that $\mathcal{T}$ is a topology on $X$ which is stronger than the relative Euclidean topology.

(b) Show that $(X, \mathcal{T})$ is separable and first countable.

(c) Show that $(X, \mathcal{T})$ is completely regular.

(d) Show that $(X, \mathcal{T})$ is not normal. [Show that the relative topology on the $x$-axis is the discrete topology, and apply XI.7.8.9.]

This example is due to Niemytzki [] and is called the Niemytzki tangent disk space or Niemytzki plane. Some references call it the Moore plane, which is not historically accurate and can be confused with the Moore manifold XI.19.5.8., which is not the same (to make matters worse, the Moore manifold has been occasionally called the Niemytzki plane!)

XI.7.8.15. Show that every infinite Hausdorff space contains a sequence of pairwise disjoint nonempty open sets. (There is a simple argument using DC. Some form of Choice is needed; cf. II.9.4.11.)

XI.7.8.16. Let $R$ be a unital ring (III.3.6.3.). Let $\text{Spec}(R)$ be the set of prime ideals (II.6.4.12.) in $R$. Put a topology on $\text{Spec}(R)$ as follows. If $E \subseteq \text{Spec}(R)$, set

$$\ker(E) = \bigcap_{P \in E} P$$

and define

$$\widehat{E} = \{ P \in \text{Spec}(R) : \ker(E) \subseteq P \}.$$

More generally, if $A \subseteq R$, define

$$\text{hull}(A) = \{ P \in \text{Spec}(R) : A \subseteq P \}$$

so that $\widehat{E} = \text{hull}(\ker(E))$.

(a) Show that $E \mapsto \widehat{E}$ is a Kuratowski closure operator (XI.2.4.1.), hence defines a topology on $\text{Spec}(R)$, called the hull-kernel or Jacobson topology on $\text{Spec}(R)$ (if $R$ is commutative, it is also called the Zariski topology).

(b) Show that the hull-kernel topology is always $T_0$.

(c) Show that the hull-kernel topology on $\text{Spec}(\mathbb{Z})$ is not $T_1$.

(d) Show that $\text{Spec}(R)$ is compact in the hull-kernel topology.

(e) Let Maxspec($R$) be the subset of $\text{Spec}(R)$ consisting of the maximal ideals (II.6.4.12.) of $R$, with the hull-kernel topology (which is the relative topology). Show that Maxspec($R$) is closed in $\text{Spec}(R)$, hence also compact, and always $T_1$. Show that Maxspec($\mathbb{Z}$) is not Hausdorff.

(f) Explain why prime ideals must be used, i.e. what goes wrong if we try to take $\text{Spec}(R)$ to be the set of all (proper) ideals in $R$?

$\text{Spec}(R)$, or certain natural subsets like Maxspec($R$), with the hull-kernel topology, is a fundamental object of study in several parts of mathematics, including Operator Algebras () and Algebraic Geometry ().
XI.7.8.17. **Other Separation Axioms.** Several other separation axioms have been considered. Here are two of them.

**Definition.** Let $X$ be a topological space. $X$ is **completely Hausdorff**, or $T_{2^{1/2}}$, if distinct points of $X$ have disjoint closed neighborhoods.

$X$ is **Urysohn** if, whenever $p, q \in X$, $p \neq q$, there is a continuous function $f : X \to [0, 1]$ with $f(p) = 0$ and $f(q) = 1$ (equivalently, there is a continuous $g : X \to \mathbb{R}$ with $g(p) \neq g(q)$).

(a) Show that the two conditions in the definition of an Urysohn space are equivalent.

(b) Show that a regular space is completely Hausdorff (use XI.7.5.1.).

(c) Show that an Urysohn space is completely Hausdorff.

(d) Show that any topology stronger than a completely Hausdorff topology [resp. Urysohn topology] is completely Hausdorff [resp. Urysohn].

A completely Hausdorff space is obviously Hausdorff, and a completely regular space is Urysohn. There are Hausdorff spaces which are not completely Hausdorff (XI.7.8.18.), in fact in which no two points have disjoint closed neighborhoods [SS95, Example 75]. There are Urysohn spaces which are not regular, hence not completely regular (XI.7.8.13.), completely Hausdorff spaces which are neither regular nor Urysohn [SS95, Example 80], and regular spaces which are not Urysohn [SS95, Example 90].

There has been a bizarre reversal in terminology concerning these two axioms. Our definitions are the ones from [SS95], which were apparently universally used until that time (see e.g. [Bou98]). For unknown reasons, the definitions were reversed in [Wil04]: what we call “completely Hausdorff” is called “Urysohn” there, and vice versa. Since the publication of [Wil04], usage of the terms has been inconsistent but has tended to follow [Wil04]. The earlier definitions make more sense to me since “Urysohn functions” separating points or points and closed sets are standard terminology (although using “completely Hausdorff” in WILLARD’S sense would be in analogy with the term “completely regular,” note that the “completely” in “completely normal” means something completely different). My guess is that WILLARD simply made a mistake, which has been perpetuated since; if he had deliberately changed established terminology he probably would have said so and given a reason, and I find it hard to believe he would worry about doing so in such a relatively minor side thread of topology. I have thus reverted to the historical definitions. It would be good if topologists would definitively decide what these definitions should be. There is one additional term sometimes used: spaces which are Urysohn according to our definition are sometimes called **functionally Hausdorff**. (In analogy, completely regular spaces could be called **functionally regular**; “functionally normal” is the same as normal by Urysohn’s Lemma.)

XI.7.8.18. **Double Origin Topology** (ALEKSANDROV and URYSOHN, cf. [SS95, Example 74]). Let $X = \mathbb{R}^2 \cup \{0\}$. Give $X$ a topology $\mathcal{T}$ consisting of all unions of the following types of sets: any subset of $X \setminus \{0, 0'\}$ which is open in $\mathbb{R}^2 \setminus \{0\}$ in the usual topology; the union of $\{0\}$ and an open half-disk in the upper half plane centered at $0$; the union of $\{0'\}$ and an open half-disk in the lower half plane centered at $0$.

(a) Show that $\mathcal{T}$ is a topology on $X$.

(b) Show that $(X, \mathcal{T})$ is Hausdorff.

(c) Show that $0$ and $0'$ do not have disjoint closed neighborhoods. Thus $(X, \mathcal{T})$ is not completely Hausdorff.

(d) Show that $(X, \mathcal{T})$ is path-connected.
XI.8. Quotient Spaces and the Quotient Topology

In this section, we will examine what happens when we take a topological space $X$ and “collapse” various subsets of $X$ to points.

Recall (II.3.3.9.) that there are three equivalent points of view for identifying subsets of a set $X$:

(i) Equivalence relations on $X$.

(ii) Partitions of $X$.

(iii) Functions from $X$ onto another set $Y$.

We suppose the set $X$ has a given topology, and want to topologize the set of equivalence classes, or the set $Y$, in an appropriate way. This topological space will be called the quotient space of $X$ by the relation.

Many constructions in topology amount to forming a quotient space.

XI.8.1. The Quotient Topology

We first take the point of view of having a function $f$ from our set $X$ onto another set $Y$. If $X$ has a specified topology $T$, we want to put a topology $T_f$ on $Y$ which makes $f$ continuous, and which is “universal”, i.e. the strongest possible. There is an obvious way to do this, since for $f$ to be continuous $f^{-1}(V)$ must be in $T$ for every $V \in T_f$ (cf. XI.5.1.14.).

XI.8.1.1. Definition. Let $(X, T)$ be a topological space, and $f : X \to Y$ a surjective function. The quotient topology on $Y$ from $f$ is the topology

$$T_f = \{ V \subseteq Y : f^{-1}(V) \in T \}.$$ 

In other words, a subset of $Y$ is open in the quotient topology if and only if its preimage in $X$ is open. Equivalently, a subset of $Y$ is closed in the quotient topology if and only if its preimage in $X$ is closed.

It is easily checked () that the quotient topology is indeed a topology. The quotient topology on $Y$ has a universal property. If $g$ is a function from $X$ to a set $Z$, and $g(x_1) = g(x_2)$ whenever $x_1, x_2 \in X$ and $f(x_1) = f(x_2)$, then there is an induced function $\tilde{g} : Y \to Z$ with $g = \tilde{g} \circ f$. The next result is obvious from the definition of the quotient topology.

XI.8.1.2. Proposition. Let $X$ be a topological space, and $f$ a surjective function from $X$ to $Y$. Give $Y$ the quotient topology. If $Z$ is a topological space and $g : X \to Z$ factors through $f$, i.e. $g = \tilde{g} \circ f$, then $g$ is continuous if and only if $\tilde{g}$ is continuous.
XI.8.1.3. **Examples.** (i) Let $\pi_1$ be the projection of $\mathbb{R}^2$ onto the $x$-axis, regarded as $\mathbb{R}$, i.e. $\pi_1(x, y) = x$. It is easily seen that the quotient topology on $\mathbb{R}$ is the ordinary topology (inverse images of sets in $\mathbb{R}$ are unions of vertical lines through points of the set). More generally, if $X$ and $Y$ are topological spaces, the projection from $X \times Y$ onto $X$ (or onto $Y$) is a quotient map, i.e. the quotient topology on $X$ is the given topology.

(ii) Let $X = [0, 1]$, $Y$ the unit circle in $\mathbb{R}^2$, and $f(t) = (\cos 2\pi t, \sin 2\pi t)$. Then $f$ is surjective, and one-to-one except at the endpoints. It is easy to check that the quotient topology on $Y$ is the usual topology (relative topology from $\mathbb{R}^2$). Thus we can regard a circle as the quotient of a closed interval with the endpoints identified.

Not all quotient spaces are this nice (cf. XI.8.1.11).

XI.8.1.4. The quotient topology is not the only topology on $Y$ making $f$ continuous in general; any weaker topology (e.g. the indiscrete topology) also will. But by XI.8.1.2. (taking $Z = Y$ and $g = f$), the quotient topology is the strongest topology on $Y$ making $f$ continuous. For example, if $f : X \to X$ is the identity function, the quotient topology on $X$ from $f$ is precisely the original topology on $X$.

**Quotients by Equivalence Relations**

XI.8.1.5. Now switch to the equivalent point of view of an equivalence relation $\sim$ on $X$. The set $Y$ can be identified with the set of equivalence classes, and the function $f : X \to Y$ is the function $f(x) = [x]$. We will write $X/\sim$ for $Y$ from this point of view, and call $f$ the projection, quotient map, or identification map, of $X$ onto $X/\sim$, which is called the quotient space or identification space of $X$ by $\sim$. We give $X/\sim$ the quotient topology.

XI.8.1.6. **Definition.** Let $\sim$ be an equivalence relation on a set $X$.

(i) A subset $A$ of $X$ is saturated if it is closed under $\sim$, i.e. if $x \in A$ and $y \sim x$, then $y \in A$.

(ii) If $A \subseteq X$, the saturation of $A$ is the set of all elements of $X$ equivalent to an element of $A$.

XI.8.1.7. It is obvious that the saturation of a subset of $X$ is saturated, and is the smallest saturated set containing $A$. If $f : X \to X/\sim$ is the projection from $X$ to $X/\sim$, the saturation of $A$ is precisely $f^{-1}(f(A))$. More generally, a subset $A$ of $X$ is saturated if and only if $A = f^{-1}(B)$ for some subset $B$ of $X/\sim$. In fact, $A = f^{-1}(f(A))$ if $A$ is saturated.

A saturated set is just a union of equivalence classes. An arbitrary union, intersection, or complement of saturated sets is saturated.

XI.8.1.8. The quotient map $f : X \to X/\sim$, besides being continuous, has the additional property that if $U$ is a saturated open set in $X$, then $f(U)$ is open in $X/\sim$, or equivalently (since the complement of a saturated open set is a saturated closed set, and vice versa) if $E$ is a saturated closed set in $X$, then $f(E)$ is closed in $X/\sim$. A surjective continuous map from a topological space $X$ to $Y$ with the quotient topology has this property for the equivalence relation defined by $x_1 \sim x_2$ if $f(x_1) = f(x_2)$.

It would be natural to give this type of function a name. Three names have been proposed and used in various references: (1) strongly continuous, first used in [AH74]; (2) bicontinuous, first used in [Kur66];
(3) quasi-compact, first used in [Why50]. None of the names have become generally established, and none is satisfactory since, among other reasons, they are all also sometimes used to mean other things. The term quasi-compact is the most unsatisfactory, since this term is sometimes used for spaces which are compact but not necessarily Hausdorff, and a mapping of the above type has nothing to do with quasi-compactness, and anyway the term quasi-compact mapping is commonly used to mean something else in algebraic geometry (more appropriately related to quasi-compactness). The term strongly continuous is also not appropriate, since it is too easily confused with the common use of “strongly continuous” in functional analysis (), and anyway it does not have the right intuitive connotation: a better term (which we do not seriously advocate) would be barely continuous, since the topology on Y is “just barely” weak enough to make the function continuous. The term bicontinuous is the least objectionable and we will use it, although it is also used to refer to simultaneous continuity with respect to two pairs of topologies (cf. [Kel63]). The term “bicontinuous” is especially appropriate in our context in the case where f is also injective: f is then bicontinuous if and only if both f and f⁻¹ are continuous (i.e. f is a homeomorphism).

It is easily seen that, conversely, if X and Y are topological spaces and f is a surjective continuous map which is bicontinuous in the above sense, then the topology on Y is necessarily the quotient topology. Thus we obtain:

XI.8.9.  PROPOSITION. Let X and Y be topological spaces and f a surjective function from X to Y. Then the topology on Y is the quotient topology from f (i.e. f is a quotient map) if and only if f is bicontinuous.

XI.8.10. This proposition characterizes when a surjective continuous function is a quotient map. Since it is obvious that a surjective continuous function which is either open or closed is bicontinuous, it follows that any such map is a quotient map. But a bicontinuous map need not be either open or closed (XI.8.11).

A quotient space can be badly behaved:

XI.8.11.  EXAMPLE. Let ~ be the equivalence relation on R defined by x ~ y if x − y ∈ Q (). Then each equivalence class is dense in R. Let f : R → R/~ be the projection. If E is a nonempty closed set in R/~, then f⁻¹(E) is a saturated closed subset of R. Since each equivalence class is dense, f⁻¹(E) is dense in R, hence f⁻¹(E) = R, E = R/~. Thus the quotient topology on R/~ is the indiscrete topology.

The same argument works for any X and ~ if all equivalence classes are dense in X.

XI.8.2.  Semicontinuous Decompositions

We want some conditions on ~ which will give that the quotient map f is closed or open. These are easily describable:

XI.8.2.1.  PROPOSITION. Let (X, T) be a topological space, ~ an equivalence relation on X, and f : X → X/~ the quotient map. The following are equivalent:

(i) The saturation of every closed set in X is closed.
(ii) If \( p \) is any point of \( X \), and \( U \) is an open neighborhood of \([p]\), then there is a saturated open set \( V \subseteq X \) with \([p] \subseteq V \subseteq U\).

(iii) If \( A \) is any saturated subset of \( X \), and \( U \) is an open neighborhood of \( A \), then there is a saturated open set \( V \subseteq X \) with \( A \subseteq V \subseteq U\).

(iv) If \( U \) is an open set in \( X \), the union of all saturated subsets of \( U \) is also open.

(v) \( f \) is a closed mapping.

An equivalence relation or partition of \( X \) satisfying these conditions is called an upper semicontinuous (usc) decomposition of \( X \).

**Proof:** (iii) \( \Rightarrow \) (ii) is trivial, and for the converse, suppose \( A \) is saturated and \( U \) is an open neighborhood of \( A \). For each \( p \in A \) let \( V_p \) be a saturated open neighborhood of \([p]\) contained in \( U \). Then \( V = \cup_p V_p \) is a saturated open neighborhood of \( A \) contained in \( U \).

(i) \( \Rightarrow \) (ii): Let \( p \in X \) and \( U \) an open neighborhood of \([p]\). Then \( U^c \) is closed, so the saturation \( B \) of \( U^c \) is closed. If \( p \in B \), then there is a \( q \in X \), \( q \sim p \), with \( q \in U^c \), contradicting that \([p] \subseteq U \). Thus \( V = B^c \) is a saturated open set containing \([p]\) and contained in \( U \).

(iii) \( \Rightarrow \) (i): Let \( E \) be a closed subset of \( X \), and \( B \) its saturation. Then \( U = E^c \) is an open neighborhood of the saturated set \( A = B^c \), so there is a saturated open neighborhood \( V \) of \( A \) contained in \( U \). The saturated closed set \( V^c \) contains \( U^c = E \) and is contained in \( A^c = B \); since \( B \) is the saturation of \( E \), we have \( V^c = B \) and \( B \) is closed.

(i) \( \iff \) (iv): Let \( U \) be open. Then the saturation of \( U^c \) is precisely the complement of the union of all saturated subsets of \( U \).

(i) \( \iff \) (v): If \( E \) is a closed set in \( X \), and \( B \) is the saturation of \( E \), then \( B \) is closed if and only if \( f(B) = f(E) \) is closed.

There is a similar characterization of when a quotient map is open, although there does not appear to be a clean analog of conditions (ii)–(iv).

**XI.8.2.2. Proposition.** Let \((X, T)\) be a topological space, \(\sim\) an equivalence relation on \(X\), and \(f : X \to X/\sim\) the quotient map. Then \(f\) is an open mapping if and only if the saturation of every open set in \(X\) is open.

An equivalence relation or partition of \(X\) satisfying these conditions is called a lower semicontinuous (lsc) decomposition of \(X\).

**XI.8.2.3.** More generally, if \(X\) and \(Y\) are topological spaces and \(2^X\) is given the topology of \((\cdot)\), a similar definition can be made for a function from \(Y\) to \(2^X\) to be upper or lower semicontinuous. There is a (somewhat tenuous) connection with upper and lower semicontinuous functions from \(\mathbb{R}\) to \(\mathbb{R}\) \((\cdot)\), which may have motivated the terminology for decompositions. See [Kur66, p. 174].

A decomposition which is both upper and lower semicontinuous is called continuous. A decomposition is continuous if and only if the quotient map is both open and closed.

Combining these characterizations with **XI.8.1.10.**, we obtain:
XI.8.2.4. Proposition. Let X and Y be topological spaces, and \( f : X \to Y \) a continuous surjection. If \( f \) is closed [open], then \( f \) is a quotient map and the corresponding decomposition of \( X \) isusc [lsc].

XI.8.2.5. Examples. (i) The decomposition of XI.8.1.3.(i) is lower semicontinuous but not upper semicontinuous: if \( E \) is the closed set \( \{ (x, y) : xy = 1 \} \) of \( \mathbb{R}^2 \), the saturation of \( E \) is \( \mathbb{R}^2 \) with the y-axis removed, and \( \pi_1(E) = \mathbb{R} \setminus \{0\} \), which are not closed. The projection \( \pi_1 \) is open ()

(ii) The decomposition of XI.8.1.3.(ii) is upper semicontinuous but not lower semicontinuous. Upper semicontinuity is obvious since only one equivalence class has more than one point. But the saturation and image of the open set \( [0, \frac{1}{2}) \) is not open.

(iii) Let \( \pi_1 \) be projection of \( [0, 1]^2 \) onto the first coordinate. Then the corresponding decomposition of \( [0, 1]^2 \) into vertical line segments is continuous (), ()

(iv) The decomposition of XI.8.1.11. is neither upper nor lower semicontinuous.

XI.8.3. Separation Axioms for Quotient Spaces

Since a primary use of quotient spaces is to construct new topological spaces, it is important to know how properties of \( X \), especially separation properties, pass to \( X/\sim \). Example XI.8.1.11. shows that nothing is preserved in general; however, under reasonable and verifiable conditions separation properties do pass to quotients.

We begin with a simple observation:

XI.8.3.1. Proposition. Let \( \sim \) be an equivalence relation on a topological space \( X \). Then \( X/\sim \) is a \( T_1 \) space if and only if each equivalence class is closed.

XI.8.3.2. Corollary. Let \( \sim \) be an equivalence relation on a \( T_1 \) topological space \( X \), and \( f : X \to X/\sim \) the quotient map. If \( f \) is closed (i.e. if the decomposition is usc), then \( X/\sim \) is \( T_1 \).

Proof: If \( p \in X \), then \( \{p\} \) is closed in \( X \), and hence its saturation \( [p] \) is also closed. 

An important variation is that the equivalence relation is closed:

XI.8.3.3. Definition. Let \( X \) be a topological space. A relation \( R \) on \( X \) is closed if it is a closed subset of \( X \times X \) with the product topology.

More generally, we can define closed relations from one topological space to another.
XI.8.3.4. If is an equivalence relation, the condition that is closed implies that each equivalence class is closed. But the conditions are not equivalent: for example, if is a $T_1$ space which is not Hausdorff, the equality relation on $X$ has closed equivalence classes (singletons) but is not a closed relation (i.e. the diagonal is not closed in $X \times X$, cf. (ii)).

This example shows that for an equivalence relation, being closed is not closely related to being upper or lower semicontinuous: the equality relation on a topological space is always both upper and lower semicontinuous, but only closed if the space is Hausdorff. Conversely, the equivalence relations of XI.8.2.5.(i)-(ii) are closed, but the decompositions fail to be lsc and usc respectively.

XI.8.3.5. Proposition. Let be an equivalence relation on a topological space $X$. If $X/ \sim$ is Hausdorff, then $\sim$ is closed.

Proof: Suppose $\sim$ is not closed. Then there is a net $((x_i, y_i))$ in $X \times X$ converging to an $(x, y) \in X \times X$ with $x_i \sim y_i$ for all $i$ but $x \not{\sim} y$. If $f : X \to X/ \sim$ is the quotient map, then the net $(f(x_i)) = (f(y_i))$ in $X/ \sim$ in $X/ \sim$ converges to both $f(x)$ and $f(y)$, so $X/ \sim$ is not Hausdorff.

The converse of XI.8.3.5. is not true in general. But we have:

XI.8.3.6. Proposition. Let be an equivalence relation on a topological space $X$, and $f : X \to X/ \sim$ the quotient map. If $f$ is open (i.e. the decomposition is lsc) and $\sim$ is closed, then $X/ \sim$ is Hausdorff.

Proof: Let $y_1, y_2 \in X/ \sim$ and $x_1, x_2 \in X$ with $f(x_1) = y_1, f(x_2) = y_2$. Then $x_1 \not{\sim} x_2$, so, since $\sim$ is closed, there are open neighborhoods $V$ and $W$ of $x_1$ and $x_2$ such that $V \times W$ is disjoint from the relation $\sim$. Then $f(V)$ and $f(W)$ are disjoint open neighborhoods of $y_1$ and $y_2$.

XI.8.3.7. The converse of XI.8.3.6. is also false in general: in XI.8.2.5.(ii), the quotient space is Hausdorff but the decomposition is not lsc. We also have:

XI.8.3.8. Proposition. Let be an equivalence relation on a $T_4$ topological space $X$, and $f : X \to X/ \sim$ the quotient map. If $f$ is closed (i.e. if the decomposition is usc), then $X/ \sim$ is $T_4$.

Proof: Let $E_1$ and $E_2$ be disjoint closed subsets of $X/ \sim$. Then $f^{-1}(E_1)$ and $f^{-1}(E_2)$ are disjoint closed subsets of $X$, so they have disjoint open neighborhoods $V$ and $W$. We have that $V^c$ and $W^c$ are closed and $V^c \cup W^c = X$, so $f(V^c)$ and $f(W^c)$ are closed in $X/ \sim$ and $f(V^c) \cup f(W^c) = X/ \sim$. Thus $U_1 = [f(V^c)]^c$ and $U_2 = [f(W^c)]^c$ are disjoint open sets. Since $f^{-1}(E_1) \subseteq V, f(V^c)$ is disjoint from $E_1$ and $E_1 \subseteq U_1$. Similarly, $E_2 \subseteq U_2$.

XI.8.3.9. Corollary. Let be an equivalence relation on a normal topological space $X$, and $f : X \to X/ \sim$ the quotient map. If $f$ is closed (i.e. if the decomposition is usc), then $X/ \sim$ is normal and $\sim$ is closed.

Proof: Combine XI.8.3.8. and XI.8.3.2. to conclude that $X/ \sim$ is normal. Since $X/ \sim$ is Hausdorff, $\sim$ is closed by XI.8.3.5.
The Maximal $T_0$ Quotient

We have so far skipped the $T_0$ condition. There is an important result here:

**XI.8.3.10. Theorem.** Let $X$ be a topological space. Then there is a maximal $T_0$ quotient $\tilde{X}$, a quotient space of $X$, such that:

(i) $\tilde{X}$ is $T_0$.

(ii) If $Y$ is any $T_0$ space and $f : X \to Y$ a continuous function, then there is a unique continuous function $\tilde{f} : \tilde{X} \to Y$ such that $f = \tilde{f} \circ \pi_0$, where $\pi_0 : X \to \tilde{X}$ is the quotient map.

(iii) The quotient map $\pi_0$ is open and closed, i.e. the partition of $X$ is continuous.

(iv) If $A$ is any subset of $X$ and $B = \pi_0^{-1}(\pi_0(A))$ is the saturation of $A$, then

$$\tilde{A} = \tilde{B} = \pi_0^{-1}(\overline{\pi_0(A)}) = \pi_0^{-1}(\overline{\pi_0(A)})$$

and $\overline{\pi_0(A)} = \pi_0(\tilde{A})$. Every closed set and every open set in $X$ is saturated.

The quotient space $\tilde{X}$ is uniquely determined up to bijective equivalence by conditions (i) and (ii). The space $\tilde{X}$ is $X/\sim$, where $\sim$ is the equivalence relation on $X$ defined by $x \sim y$ if $\{x\} = \{y\}$.

**Proof:** We verify that the quotient by the given equivalence relation has properties (i)–(iv). Let $A \subseteq X$, and $B = \pi_0^{-1}(\pi_0(A))$. Let $p \in B$. If $U$ is any open neighborhood of $p$, then there is a $y \in U \cap B$. Then $y \sim x$ for some $x \in A$, so $y \in \{x\}$. Thus any neighborhood of $y$ contains $x$, and in particular $x \in U$. Thus every neighborhood of $p$ contains a point of $A$, so $p \in A$, i.e. $B \subseteq A$, and the opposite inclusion is trivial. Thus $\tilde{A} = \tilde{B}$.

If $A$ is closed, then we have $A = \tilde{A} = \tilde{B}$, so $B \subseteq A$, and the opposite inclusion is trivial, so $A$ is saturated. Since the complement of a saturated set is saturated, every open set is saturated. Thus (iii) is an immediate consequence.

If $A$ is again a general subset of $X$, then $\pi_0(\tilde{A})$ is a closed set containing $\pi_0(A)$, so $\overline{\pi_0(\tilde{A})} \subseteq \pi_0(\tilde{A})$, and the opposite inclusion holds by continuity. Since $\tilde{A}$ is saturated, $\tilde{A} = \pi_0^{-1}(\pi_0(\tilde{A}))$. Thus (iv) is proved.

We then easily obtain (i): if $x, y \in X$ and $\{\pi_0(x)\} = \{\pi_0(y)\}$, we have $\{x\} = \pi_0^{-1}(\{\pi_0(x)\}) = \pi_0^{-1}(\{\pi_0(y)\}) = \{y\}$

so $x \sim y$ and $\pi_0(x) = \pi_0(y)$. Apply XI.7.2.1.

To show (ii), suppose $f$ is a continuous function from $X$ to a $T_0$ space $Y$. It suffices to show that if $x_1, x_2 \in X$ and $x_1 \sim x_2$, then $f(x_1) = f(x_2)$. Suppose $f(x_1) \neq f(x_2)$. Since $Y$ is $T_0$, we may assume without loss of generality that $f(x_1)$ has an open neighborhood $V$ not containing $f(x_2)$. But then $f^{-1}(V)$ is an open neighborhood of $x_1$ not containing $x_2$, so $x_1 \not\sim x_2$.

Finally, we verify uniqueness; in fact, we show slightly more. We use the standard uniqueness argument for universal properties (i). Let $\tilde{X}$ be a $T_0$ space, and $\tilde{f} : X \to \tilde{X}$ a continuous function with the universal property of (ii). Then there is a (unique) $f : X \to \tilde{X}$ such that $\tilde{f} = f \circ \pi_0$, and a (unique) $g : \tilde{X} \to \tilde{X}$ such that $\pi_0 = g \circ \tilde{f}$. But then $(g \circ f) \circ \pi_0 = \pi_0$, so $g \circ f$ is the unique $h : X \to \tilde{X}$ such that $\pi_0 = h \circ \pi_0$. But the identity map $\iota_{\tilde{X}}$ also has this property, so $g \circ f = \iota_{\tilde{X}}$. Similarly, $f \circ g = \iota_{\tilde{X}}$, so $g = f^{-1}$ is a homeomorphism.

This result is due to KOLMOGOROV, and the maximal $T_0$ quotient is sometimes called the Kolmogorov quotient. There is also a maximal Hausdorff quotient (XI.8.5.6.).
Other Properties Preserved in Quotients

We now discuss some properties other than separation properties.

XI.8.3.11. Any property preserved under all surjective continuous maps, e.g. separability and compactness, is of course preserved under arbitrary quotient maps.

For first and second countability, we need the decomposition to be lsc, i.e. the quotient map is open:

XI.8.3.12. Proposition. Let \( f : X \to X/\sim \) be an equivalence relation on a topological space \( X \), and \( f : X \to X/\sim \) the quotient map. If \( f \) is open (i.e. the decomposition is lsc) and \( X \) is first countable \( \) or second countable, then \( X/\sim \) is also first countable \( \) or second countable. \( \square \)

Proof: Let \( p \in X \), and \( \{ U_n : n \in \mathbb{N} \} \) a countable base at \( p \). Then \( \{ f(U_n) \} \) is a collection of open neighborhoods of \( f(p) \). If \( V \) is an open neighborhood of \( f(p) \), then \( f^{-1}(V) \) is an open neighborhood of \( p \), so there is an \( n \) such that \( U_n \subseteq f^{-1}(V) \). Then \( f(U_n) \subseteq V \) and \( \{ f(U_n) : n \in \mathbb{N} \} \) is a countable base at \( f(p) \).

If \( \{ U_n : n \in \mathbb{N} \} \) is a countable base for the topology of \( X \), then by a nearly identical argument \( \{ f(U_n) : n \in \mathbb{N} \} \) is a countable base for the topology of \( X/\sim \).

Compactness and Local Compactness

XI.8.3.13. If \( X \) is compact and \( Y \) is Hausdorff, and \( f : X \to Y \) is continuous, then \( f \) is automatically a closed mapping \( \); so if \( f \) is surjective it is a quotient map and \( Y \) is also compact.

XI.8.4. Quotient Metrics

A construction related to, but not in general identical to, the quotient space construction is the construction of a quotient (pseudo)metric.

XI.8.4.1. Let \( (X, \rho) \) be a metric space and \( \sim \) an equivalence relation on \( X \), and let \( Y = X/\sim \) be the set-theoretic quotient. We want to put a “quotient metric” \( \bar{\rho} \) on \( Y \). As a first try, we could consider setting

\[
\bar{\rho}([x],[y]) = \inf \{ \rho(x',y') : x' \in [x], y' \in [y] \}.
\]

This function is well-defined, nonnegative, and symmetric, but easy examples (e.g. \( \)) show that \( \bar{\rho} \) defined this way usually does not satisfy the triangle inequality (and it is not definite in general, i.e. could only be a pseudometric).

There is a way to modify this definition to give a pseudometric (Exercise XI.8.5.15.), but there is a very slick way to define what is obviously the right candidate for the quotient metric:

XI.8.4.2. Definition. Let \( (X, \rho) \) be a metric space, and \( \sim \) an equivalence relation on \( X \). Define \( \beta : X \times X \to [0, +\infty) \) by

\[
\beta(x,y) = \begin{cases} 
\rho(x,y) & \text{if } x \not\sim y \\
0 & \text{if } x \sim y
\end{cases}
\]

Let \( \rho_\beta \) be the maximum pseudometric of \( \beta \) (VI.2.4.7.), and set \( \bar{\rho}([x],[y]) = \rho_\beta(x,y) \). Then \( \bar{\rho} \) is the quotient pseudometric on \( X/\sim \).

It is not hard to see that \( \bar{\rho} \) is well defined (Exercise XI.8.5.16.), and it is then obviously a pseudometric.
XI.8.4.3. Examples. (i) Let \( X = [0, 1] \) with the usual metric, and \( \sim \) the equivalence relation obtained by identifying 0 and 1 (i.e. 0 \sim 1 \) and the other equivalence classes are singletons). Then \( X/ \sim \) is a circle of circumference 1, and the quotient pseudometric is the arc length metric.

(ii) Let \( X = [0, +\infty) \) with the usual metric \( \rho \), and \( \sim \) the equivalence relation obtained by identifying \( \mathbb{N} \cup \{0\} \) to a point, i.e. the equivalence classes are \( \mathbb{N} \cup \{0\} \) and singletons. \( (X/ \sim, \tilde{\rho}) \) is homeomorphic to the space of XI.18.4.7.

(iii) Let \( X = (0, 1) \) with the usual metric \( \rho \), and \( \sim \) the equivalence relation obtained by identifying \( \frac{1}{n} \) for each \( n \in \mathbb{N} \). Then each equivalence class is closed, but we still have that \( \tilde{\rho} \) is the zero pseudometric.

(iv) Let \( X = \mathbb{R} \) with the usual metric \( \rho \), and \( \sim \) the equivalence relation whose classes are the cosets of \( \mathbb{Q} \). Every equivalence class is dense, so \( \tilde{\rho} \) is the zero pseudometric.

(v) Let \( X = (0, 1) \) with the usual metric \( \rho \). Let \( \sim \) be the equivalence relation obtained by identifying \( x \) with \( x^2 \) for each \( x \in (0, 1) \). Then each equivalence class is closed, but we still have that \( \tilde{\rho} \) is the zero pseudometric.

XI.8.4.4. Examples (iv) and (v) show that \( \tilde{\rho} \) is not always a metric if \( \rho \) is, even if all equivalence classes are closed. Since \( \tilde{\rho}(x, y) \leq \rho(x, y) \) for all \( x, y \in X \), the quotient map from \( (X, \rho) \) to \( (X/ \sim, \tilde{\rho}) \) is continuous, i.e. the \( \tilde{\rho} \) topology on \( X/ \sim \) is weaker than the quotient topology. It is the quotient topology in good cases (e.g. in (i)), but strictly weaker in general (e.g. in (ii), (iii), (v)). The \( \tilde{\rho} \) topology can depend on the exact metric and not just on the topology: examples (ii) and (iii) are homeomorphic respecting the equivalence relation, but the quotient metric topologies are not the same.

XI.8.4.5. But if \( X \) is metrizable and \( X/ \sim \) is also metrizable in the quotient topology (cf. (iv)), there is a metric on \( X \) giving its topology such that the quotient metric gives the quotient topology. For if \( \sigma \) is a metric on \( X \) giving its topology and \( \tau \) a metric on \( X/ \sim \) giving its topology, and \( \pi : X \to X/ \sim \) is the quotient map, then \( \hat{\rho} = \sigma + (\tau \circ \pi) \) is a metric on \( X \) which is equivalent to \( \sigma \), and the \( \hat{\rho} \)-topology on \( X/ \sim \) is stronger than than the \( \tau \)-topology and weaker than the quotient topology, hence they all coincide, i.e. \( \hat{\rho} \) is equivalent to \( \tau \).

XI.8.5. Exercises

XI.8.5.1. Let \( X, Y, Z \) be topological spaces, and \( f : X \to Y \) and \( g : Y \to Z \) surjective continuous functions. Set \( h = g \circ f : X \to Z \).

(a) Show that if \( f \) and \( g \) are quotient maps, then \( h \) is also a quotient map.

(b) Show that if \( h \) is a quotient map, then \( g \) is also a quotient map. Is \( f \) necessarily a quotient map?

XI.8.5.2. (a) Let \( X \) and \( Y \) be topological spaces, and let \( \pi_1 \) be projection of \( X \times Y \) onto \( X \). Show that \( \pi_1 \) is a quotient map and that the corresponding decomposition of \( X \times Y \) is lsc.

(a) Let \( \{X_i : i \in I\} \) be a family of topological spaces, and let \( X = \prod_{i \in I} X_i \). For \( j \in I \), let \( \pi_j \) be projection of \( X \) onto \( X_j \). Show that \( \pi_j \) is a quotient map and that the corresponding decomposition of \( X \) is lsc.
XI.8.5.3. Let \( \sim \) be an equivalence relation on a topological space \( X \), and \( f : X \to X/ \sim \) the quotient map. If \( Z \subseteq X \), let \( \sim_Z \) be the restriction of \( \sim \) to \( Z \). Is \( Z/ \sim_Z \) homeomorphic to \( f(Z) \) in general? What if \( Z \) is open or closed in \( X \)? What if the decomposition is usc or lsc?

XI.8.5.4. Let \( X_1, X_2, Y_1, Y_2 \) be topological spaces, and \( f : X_1 \to Y_1 \) and \( g : X_2 \to Y_2 \) be quotient maps. Define \( h : X_1 \times X_2 \to Y_1 \times Y_2 \) by \( h(x_1, x_2) = (f(x_1), g(x_2)) \). Is \( h \) necessarily a quotient map? What if \( f \) and \( g \) are open or closed? (cf. [Mun75, p. 141-143].)

XI.8.5.5. Show that the maximal \( T_0 \) quotient of a topological space \( X \) can be described alternately as follows.

(a) Define an equivalence relation \( \approx \) on \( X \) by \( x_1 \approx x_2 \) if \( f(x_1) = f(x_2) \) for every continuous function \( f \) from \( X \) to a \( T_0 \) space \( Y_f \) (which can depend on \( f \)). Show that \( X/ \approx \) is \( T_0 \). Conclude that \( \approx = \sim \).

(b) Describe \( \tilde{X} \) explicitly by considering the map \( \pi \) from \( X \) to \( \prod_f Y_f \) by \( (\pi(x))_f = f(x) \) (show that this can be done without set-theoretic difficulties by choosing a suitable representative set of functions). Set \( Y = \pi(X) \). Show that \( Y \cong \tilde{X} \) and that \( \pi_0 \) can be identified with \( \pi \), regarded as a map from \( X \) to \( Y \).

XI.8.5.6. The Maximal Hausdorff Quotient. Let \( X \) be a topological space. Prove the following theorem:

**Theorem.** Let \( X \) be a topological space. Then there is a maximal Hausdorff quotient \( \tilde{X} \), a quotient space of \( X \), such that:

(i) \( \tilde{X} \) is Hausdorff.

(ii) If \( Y \) is any Hausdorff space and \( f : X \to Y \) a continuous function, then there is a unique continuous function \( \tilde{f} : \tilde{X} \to Y \) such that \( f = \tilde{f} \circ \pi_2 \), where \( \pi_2 : X \to \tilde{X} \) is the quotient map.

The quotient space \( \tilde{X} \) is uniquely determined up to bijective equivalence by conditions (i) and (ii).

(a) Define an equivalence relation \( \approx \) on \( X \) by \( x_1 \approx x_2 \) if \( f(x_1) = f(x_2) \) for every continuous function \( f \) from \( X \) to a Hausdorff space \( Y_f \) (which can depend on \( f \)). Show that \( \tilde{X} = X/ \approx \) is Hausdorff. Describe \( \tilde{X} \) as in XI.8.5.5.(b). Give the standard proof of uniqueness.

(b) It is not easy to explicitly describe the equivalence relation \( \approx \). As a first approximation, define \( \sim \) on \( X \) by \( x \sim y \) if \( x \) and \( y \) do not have disjoint neighborhoods. Is \( \sim \) an equivalence relation?

(c) Extend \( \sim \) to an equivalence relation by taking its transitive closure. Is \( X/ \sim \) necessarily Hausdorff? If it is, it must be (up to bijective equivalence) the \( \tilde{X} \) of (a).

(d) If \( X/ \sim \) is not Hausdorff, iterate the construction. Let \( \sim' \) be the relation on \( X/ \sim \) defined in the same way as \( \sim \). Write \( \sim_1 = \sim \) on \( X \), and let \( \sim_2 \) be the equivalence relation on \( X \) corresponding to the relation \( \sim' \) on \( X/ \sim_1 \). Describe \( \sim_2 \) explicitly.

(e) Transfinitely define an increasing collection \( \sim_\alpha \) of equivalence relations on \( X \) in this way. Show that this collection stabilizes at some \( \beta \), and that \( X/ \sim_\beta \) is Hausdorff and is bijectively equivalent to \( \tilde{X} \).

The smallest such \( \beta \) is called the **Hausdorff height** of \( X \) and is an ordinal invariant of \( X \).

The maximal Hausdorff quotient of \( X \) is sometimes called the **Hausdorffization** or **Hausdorffification** of \( X \). See () for details.
XI.8.5.7.  
(a) Is there a maximal $T_1$ quotient of a topological space $X$ along the same lines as the maximal $T_0$ quotient and the maximal Hausdorff quotient?
(b) Show there is no maximal regular, completely regular, normal, or completely normal quotient in general. [There are nonregular topologies on $\mathbb{R}$ which are stronger than the usual topology (e.g. XI.7.8.13.).]

XI.8.5.8.  
There are $2^{\aleph_0}$ pairwise disjoint dense subsets of $\mathbb{R}$ [this can be proved without any form of choice, e.g. by taking rational translates of the von Neumann numbers (III.8.4.13.). See also XI.8.5.9..] Let $\{A_t : t \in \mathbb{R}\}$ be such a collection. We may assume $\bigcup_t A_t = \mathbb{R}$ by expanding one of the sets if necessary. Set

$$X = \{(t, y) \in \mathbb{R}^2 : y \in A_t\} \subseteq \mathbb{R}^2.$$ 

Then projection $\pi_2$ of $X$ onto the second coordinate is a bijection from $X$ onto $\mathbb{R}$. The projection $\pi_1$ is a surjective but not injective function from $X$ to $\mathbb{R}$. Let $\mathcal{E}$ denote the usual topology on $\mathbb{R}$.

(a) Let $\mathcal{T}$ be any topology on $\mathbb{R}$. Let $\mathcal{T}_X$ be the relative topology on $X$ of the product topology $\mathcal{T} \otimes \mathcal{E}$ on $\mathbb{R} \times \mathbb{R}$. Show that $\mathcal{T}_X$ is a Hausdorff topology on $X$. [Distinct points of $X$ have distinct second coordinates.]

(b) If $\mathcal{T}$ is any topology on $\mathbb{R}$, let $\mathcal{H}(\mathcal{T})$ be the topology on $\mathbb{R}$ obtained by transferring $\mathcal{T}_X$ to $\mathbb{R}$ via the bijection $\pi_2$. Then $\mathcal{H}(\mathcal{T})$ is a Hausdorff topology on $\mathbb{R}$ which is stronger than $\mathcal{E}$. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are different topologies on $\mathbb{R}$, then $\mathcal{H}(\mathcal{T}_1)$ and $\mathcal{H}(\mathcal{T}_2)$ are different topologies on $\mathbb{R}$. Thus there are enormously many distinct (Hausdorff) topologies on $\mathbb{R}$ stronger than $\mathcal{E}$.

(c) The surjective map $\pi_1 \circ \pi_2^{-1}$ is continuous and open from $(\mathbb{R}, \mathcal{H}(\mathcal{T}))$ to $(\mathbb{R}, \mathcal{T})$.

(d) Show that $(\mathbb{R}, \mathcal{T})$ has the quotient topology from $(\mathbb{R}, \mathcal{H}(\mathcal{T}))$ under this map.

(e) If $Y$ is any topological space of cardinality $\leq 2^{\aleph_0}$, then $Y$ can be identified as a subset of $\mathbb{R}$ with some relative topology. Thus $Y$ is a continuous open image of a Hausdorff space.

This can be extended to topological spaces of arbitrary cardinality (XI.8.5.9.).

XI.8.5.9.  
(a) Let $\kappa$ be a cardinal. Then there is a compact Hausdorff space $Z$ which has a collection of $\kappa$ pairwise disjoint dense subsets. [Let $Z = \{0, 1\}^\kappa$, and consider the equivalence relation on $Z$ where two points are equivalent if they differ in only finitely many coordinates. Use II.9.7.14.]

(b) If $Y$ is any topological space of cardinality $\leq 2^{\aleph_0}$, then $Y$ can be identified as a subset of $\mathbb{R}$ with some relative topology. Thus $Y$ is a continuous open image of a Hausdorff space.

Repeat problem XI.8.5.8. with this $X$ to show that $Y$ is a continuous open image (quotient) of a Hausdorff space.

XI.8.5.10.  
Let $Y$ be the Sierpiński space. Define $f : [0, 1] \to Y$ by $f(t) = a$ if $t < 1$ and $f(1) = b$. Show that $f$ is a quotient map. However, $f$ is neither open nor closed. Use XI.8.5.8. to write $Y$ as a quotient of a Hausdorff space by an open map; see XI.7.8.13.(d). (It cannot be done by a closed map since the quotient of a $T_1$ space by a closed map is $T_1$.)
XI.8.5.11. Let $X$ and $Y$ be topological spaces, and $f : X \to Y$ a surjective continuous map. Say $f$ is \textit{semiclosed} if the image of the closure of every saturated set in $X$ is closed in $Y$.

(a) Show that $f$ is semiclosed if and only if, for every subset $B$ of $Y$, we have

$$B = f(f^{-1}(B)).$$

(b) Show that every semiclosed map is a quotient map. In fact, $f$ is semiclosed if and only if, for every subset $B$ of $Y$, $f|_{f^{-1}(B)}$ is a quotient map from $f^{-1}(B)$ onto $B$. Thus a semiclosed map is also called a \textit{hereditary quotient map}.

(c) Show that every (surjective) closed map is semiclosed.

(d) Show that all the examples of quotient maps in this section (except XI.8.5.12.) are semiclosed. In particular, a semiclosed map need not be closed.

(e) If $f : X \to Y$ is a quotient map and $Y$ is Hausdorff and first countable, then $f$ is semiclosed.

(f) Not every quotient map is semiclosed (XI.8.5.12.).

(g) Show that if $X$ and $Y$ are first countable and $Y$ is $T_1$, $f : X \to Y$ is semiclosed, $(y_n)$ is a sequence in $Y$, and $y \in Y$, then $y_n \to y$ if and only if for every subsequence of $(y_n)$ there is a subsequence $(y_{n_k})$ for which there is a sequence $(x_{n_k})$ in $X$ and an $x \in X$ with $x_{n_k} \to x$, $f(x_{n_k}) = y_{n_k}$ for all $n$, and $f(x) = y$. In other words, convergent sequences in $Y$ have subsequences which can be lifted to convergent sequences in $X$.

This result is quite useful in identifying and describing quotient topologies. The restriction that $Y$ be $T_1$, or something like it, is necessary (XI.8.5.13.). And even if $X$ and $Y$ are compact metrizable spaces (where the result in (g) is almost trivial), and $f$ is closed, passage to subsequences is necessary in general (XI.8.5.14.).

XI.8.5.12. Let $X = \{a_1, a_2, b_1, b_2, c_1, c_2, x_1, x_2\}$, $Y = \{a, b, c, y\}$, and $f : X \to Y$ with $f(a_j) = a$, $f(b_j) = b$, $f(c_j) = c$, $f(x_j) = y$ for $j = 1, 2$. Give $X$ the topology generated by $\{b_1, x_1\}$ and $\{c_2, x_2\}$.

(a) Show that the quotient topology on $Y$ is the indiscrete topology.

(b) Show that $y$ is in the closure of $\{a\}$ in $Y$, but neither $x_1$ nor $x_2$ is in the closure in $X$ of $f^{-1}(\{a\}) = \{a_1, a_2\}$. Thus $f$ is not semiclosed.

XI.8.5.13. Let $X = \{x_n, z_n : n \in \mathbb{N}\}$, $Y = \{y\} \cup \{y_n : n \in \mathbb{N}\}$, $f : X \to Y$ with $f(x_n) = y_n$ and $f(z_n) = y$ for all $n$. Give $X$ the topology generated by $\{\{x_n\}, \{x_n, z_n\} : n \in \mathbb{N}\}$, and give $Y$ the quotient topology.

(a) Show that $X$ and $Y$ are first countable but not $T_1$, and that the quotient map $f$ is semiclosed.

(b) Show that $y_n \to y$ in $Y$, but no subsequence of $(y_n)$ lifts to a convergent sequence in $X$.

XI.8.5.14. Let $X$ be the compact subset

$$\left\{ \left( \frac{1}{n}, 1 \right) : n \in \mathbb{N} \text{ odd} \right\} \cup \left\{ \left( \frac{1}{n}, 0 \right) : n \in \mathbb{N} \text{ even} \right\} \cup \{(0,0),(0,1)\}$$

of $\mathbb{R}^2$, and $Y = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\} \subseteq \mathbb{R}$. Define $f : X \to Y$ by $f(x, y) = x$.

(a) Show that $f$ is a quotient map which is closed, hence semiclosed.

(b) Show that the convergent sequence $y_n = \frac{1}{n}$ in $Y$ does not lift to a convergent sequence in $X$. 1071
XI.8.5.15. Let $(X, \rho)$ be a metric space, and $\sim$ an equivalence relation on $X$. Show that the quotient metric $\bar{\rho}$ can be explicitly described as

$$\bar{\rho}(a, b) = \inf \sum_{k=1}^{n} \rho(x_k, y_k)$$

where the infimum is over all finite sets \{x_1, \ldots, x_{n+1}, y_1, \ldots, y_n\}, where $x_1 = a$, $x_{n+1} = b$, and $y_k \sim x_{k+1}$ for $1 \leq k \leq n$ ($n$ may vary). The definition does not change if we only require that $x_1 \sim a$, along with the other conditions.

XI.8.5.16. Let $X, \rho, \sim, \beta$ be as in XI.8.4.2. Use the triangle inequality to show that if $\sigma$ is a pseudometric on $X$ with $\sigma \leq \rho_\beta$, then $\sigma$ drops to a well-defined pseudometric on $X/\sim$. In particular, $\bar{\rho}$ is well defined.
XI.9. Identification Spaces

Perhaps the most important method of construction of topological spaces is by identification or “gluing.” There are a few variations, which all fall under the quotient space construction.

XI.9.1. Collapsing a Subset to a Point

XI.9.1.1. Let \( X \) be a topological space, and \( A \) a subset. We can “collapse \( A \) to a point” by taking the equivalence relation ~ for which \( A \) is one equivalence class and all other equivalence classes are single points. Then the space \( X/\sim \) with the quotient topology is usually denoted \( X/A \) and is called the identification space of \( X \) by \( A \), or with \( A \) collapsed to a point. (This notation \( X/A \) is common, but sometimes conflicts with other common notation. For example, if \( X \) is a topological group and \( A \) a subgroup, then \( X/A \) usually denotes the quotient group \( (\cdot) \), which is not the same thing. See also XI.9.3.6.) We may identify \( X/A \) with \((X \setminus A) \cup \{a\}\), where \( a = \{A\} \).

In order to obtain a \( T_1 \) quotient, \( A \) must be closed in \( X \), and the collapsing construction is almost exclusively used in this situation.

More generally, we may take a finite or infinite collection of pairwise disjoint subsets of \( X \) and collapse each one to a point. If \( A \) and \( B \) are disjoint subsets of \( X \) (usually closed), write \( X/(A,B) \) for the space obtained from \( X \) by collapsing \( A \) and \( B \) to (separate) points. The collapsing can be done consecutively, i.e. by forming \((X/A)/B \) or \((X/B)/A \), or simultaneously. Note that \( X/(A,B) \not\cong X/(A \cup B) \).

XI.9.1.2. Proposition. Let \( X \) be a topological space and \( A \) a closed subset of \( X \), and \( f : X \to X/A \) the quotient map. Then \( f \) is closed, i.e. the decomposition is usc, and the restriction of \( f \) to \( X \setminus A \) is a homeomorphism onto the open set \((X/A) \setminus \{a\}\).

Proof: \( X \setminus A \) is open in \( X \), \( W = (X/A) \setminus \{a\} \) is open in \( X/A \), and \( f \) is a bijection of \( X \setminus A \) to \( W \). Identifying \( W \) with \( X \setminus A \), a subset \( V \) of \( W \) is open in \( X/A \) if and only if the corresponding subset of \( X \setminus A \) is open in \( X \), i.e. \( V \) is relatively open in \( W \) if and only if \( f^{-1}(V) \) is relatively open in \( X \setminus A \). Thus \( f|_{X \setminus A} \) is a homeomorphism.

If \( E \) is a subset of \( X \), then \( f^{-1}(f(E)) = E \) if \( E \cap A = \emptyset \), and \( f^{-1}(f(E)) = E \cup A \) if \( E \cap A \neq \emptyset \). In any case, if \( E \) is closed, \( f^{-1}(f(E)) \) is closed, so \( f \) is closed.

All the results can fail if \( A \) is not closed (Exercise XI.9.5.1.).

XI.9.1.3. Conversely, if \( f : X \to Y \) is a quotient map and every point of \( Y \) but one has a singleton preimage, then \( Y = X/f^{-1}\{q\} \), where \( q \) is the exceptional point. If \( \{q\} \) is closed in \( Y \) (e.g. if \( Y \) is \( T_1 \)), then \( f \) is closed and \( f|_{X \setminus f^{-1}(\{q\})} \) is a homeomorphism onto \( Y \setminus \{q\} \).

XI.9.1.4. Examples. (i) Let \( f \) be the map from \([0,1] \) to \( \mathbb{T} \cong S^1 \) from XI.8.1.3.(ii). We have \( S^1 \cong [0,1]/\{0,1\} \), i.e. a circle is a closed interval with endpoints identified.

(ii) More generally, let \( B^n \) be the closed unit ball in \( \mathbb{R}^n \). The boundary is the \((n-1)\)-sphere \( S^{n-1} \). We have that \( B^n/S^{n-1} \cong S^n \), as is easily checked (e.g. via XI.9.1.8.). For another construction of \( S^n \) by collapsing, see XI.9.1.12.(iii).
XI.9.1.5. Proposition. Let $X$ be a Hausdorff topological space, and $A \subseteq X$. Then $X/A$ is Hausdorff if and only if, for each $p \in X \setminus A$, $A$ and $p$ have disjoint neighborhoods. (This entails that $A$ must be a closed subset of $X$.) In particular, if $X$ is regular and $A$ is closed, then $X/A$ is Hausdorff.

XI.9.1.6. An identification space from a regular space need not be regular. For example, if $X$ is regular but not normal, and $A$ and $B$ are disjoint closed sets in $X$ which do not have disjoint neighborhoods, then $X/A$ is not regular. However, we have the following special case of XI.8.3.9:

XI.9.1.7. Proposition. Let $X$ be a normal space and $A$ a closed subset of $X$. Then $X/A$ is normal.

A particularly important special case is:

XI.9.1.8. Proposition. Let $X$ be a compact Hausdorff space, and $A$ a closed subset of $X$. Then $X/A$ is homeomorphic to the one-point compactification (XI.11.8.3.) of $X \setminus A$.

Proof: Combine XI.9.1.3. and XI.11.8.5.

Cones and Suspensions

We can define a cone over any topological space $X$ by forming a cylinder with base $X$ and collapsing the top edge to a point:

XI.9.1.9. Definition. Let $X$ be a topological space. The space $CX = ([0,1] \times X)/(\{1\} \times X)$ is called the cone over $X$.

This construction is primarily used when $X$ is a compact Hausdorff space. In this case, $CX$ is homeomorphic to the one-point compactification of $[0,1) \times X$ (XI.9.1.8).

XI.9.1.10. Examples. (i) The cone over $[0,1]$ is (homeomorphic to) a solid triangle.

(ii) The cone over $S^1$ is an ordinary cone, which is homeomorphic to a closed disk.

(iii) More generally, for any $n$ the cone over $S^n$ is homeomorphic to $B^{n+1}$ (note that $B^{n+1} \setminus \{0\}$ is homeomorphic to $[0,1) \times S^n$).

(iv) The cone $CK$ over the Cantor set $K$ can be identified with the union of all line segments in $\mathbb{R}^2$ between $(1,1)$ and points of $K$, regarded as a subset of the $x$-axis. The point $(\frac{1}{2},1)$ can be replaced by any point of $\mathbb{R}^2$ not on the $x$-axis. This is a good example of a space which is path-connected but not locally path connected. ($CX$ is path-connected for any topological space $X$.)

We can similarly form the “suspension” of $X$ by collapsing each end of the cylinder over $X$ to (separate) points:
**XI.9.1.11. Definition.** Let \( X \) be a topological space. The *suspension* of \( X \) is

\[
SX = ([0,1] \times X)/\{(0) \times X, (1) \times X\}.
\]

This is also primarily used in the case where \( X \) is a compact Hausdorff space, and gives a “two-point compactification” of \((0,1) \times X \cong \mathbb{R} \times X\) in this case. In general, the suspension of \( X \) is formed by collapsing the base of \( CX \) to a point.

The suspension of \( X \) can be thought of as two copies of \( CX \) joined base-to-base (\(\cdot\)).

**XI.9.1.12. Examples.** (i) The suspension of \([0,1]\) can be pictured as a (solid) diamond in \(\mathbb{R}^2\).

(ii) The suspension of \(S^1\) can be pictured as the union of two cones joined along their base circles (\(\cdot\)). This space is homeomorphic to \(S^2\).

(iii) More generally, the suspension of \(S^n\) is homeomorphic to \(S^{n+1}\). This can be seen by first forming \(CS^n \cong B^{n+1}\), and then collapsing the base, which is the boundary \(S^n\), to a point (cf. XI.9.1.4(ii)).

**Wedge Sums**

**XI.9.1.13.** Let \( \{X_i : i \in I\} \) be a collection of topological spaces, and for each \( i \) let \(*_i\) be a designated point of \(X_i\). The *wedge sum* of \( \{(X_i, *_i) : i \in I\} \) is the space \( X \) obtained from the separated union of the \(X_i\) with all the \(*_i\) identified together to one point \(*\). (The homeomorphism class of \( X \) depends in general on the choice of the \(*_i\).) If each \( X_i \) is a circle, the space \( X \) is called a *wedge of circles* or *bouquet of circles*.

If \( I \) is a finite set with \( n \) elements and each \( X_i \) is a circle, the space \( X \) is the bouquet of \( n \) circles, resembling a flower with \( n \) petals. A bouquet of two circles is homeomorphic to a figure 8.

A wedge sum of infinitely many nontrivial spaces is rarely metrizable. For example, a wedge or bouquet of infinitely many circles is not first countable (XI.9.5.7., XI.9.5.8.).

**XI.9.2. Gluing and Attaching**

Gluing and attaching are perhaps the most extensively used constructions in topology.

**XI.9.2.1.** Let \( X \) be a topological space, and \( A \) and \( B \) disjoint subsets of \( X \) (with care, the disjointness can be relaxed). Let \( f : A \to B \) be a homeomorphism, or more generally a quotient map. We take the quotient of \( X \) obtained by identifying \( x \) with \( f(x) \) for all \( x \in A \). More precisely, we define an equivalence relation \( \sim \) on \( X \) by taking

\[
[x] = \{y \in A : f(y) = f(x)\} \cup \{f(x)\}
\]

if \( x \in A \), \( [x] = \{x\} \cup f^{-1}(\{x\}) \) if \( x \in B \), and \( [x] = \{x\} \) if \( x \notin A \cup B \). The space \( X/\sim \) is called the *identification space* of \( X \) with \( A \) identified with \( B \) via \( f \), or with \( A \) glued (or sewed) to \( B \) via \( f \). Note that if \( f \) is a homeomorphism, the equivalence classes just consist of \( \{x, f(x)\} \) for \( x \in A \) and \( \{x\} \) for \( x \notin A \cup B \).

We may more generally do more than one gluing, gluing \( A_1 \) to \( B_1 \) and \( A_2 \) to \( B_2 \), etc. We need that \( A_j \) and \( B_j \) are disjoint for each \( j \), but \( A_j \) and \( A_k \), \( B_j \) and \( B_k \), and \( A_j \) and \( B_k \) need not be disjoint for \( j \neq k \), but the gluings have to be consistent on the overlaps. The gluing can be done consecutively or simultaneously. See e.g. XI.9.2.3(v).
XI.9.2.2. In the gluing construction, to obtain a nice quotient space it is not necessary that \(A\) and \(B\) be closed (XI.9.2.3(ii)); however, the case where they are not closed is rather delicate. The construction is more robust if they are closed.

XI.9.2.3. Examples. (i) Let \(X = [0, 1]\), \(A = \{0\}\), \(B = \{1\}\), \(f\) the obvious homeomorphism. Then the identification space with \(A\) glued to \(B\) is \(S^1\), as in XI.8.1.3.(ii) or XI.9.1.4.

(ii) Let \(X = (0, 1)\), \(A = (0, \frac{1}{3})\), \(B = (\frac{2}{3}, 1)\), \(f(x) = x + \frac{2}{3}\) for \(x \in A\). Then the identification space is homeomorphic to \(S^1\).

The identification space depends crucially on the choice of homeomorphism. If instead we take \(f(x) = 1 - x\) for \(x \in A\), the identification space is not Hausdorff.

(iii) Let \(X = [0, 1]^2\), \(A = [0, 1] \times \{0\}\), \(B = [0, 1] \times \{1\}\). Define \(f : A \to B\) by \(f(t, 0) = (t, 1)\). Then the identification space is a cylinder (Figure XI.2). Informally, we take a rectangle and glue the ends together with no twist.

(iv) Let \(X = [0, 1]^2\), \(A = [0, 1] \times \{0\}\), \(B = [0, 1] \times \{1\}\). Define \(f : A \to B\) by \(f(t, 0) = (1 - t, 1)\). Then the identification space is a Möbius strip (Figure XI.3, from https://upload.wikimedia.org/wikipedia/commons/d/d9/M%C3%B6bius_strip.jpg). Informally, we take a rectangle and glue the ends together with a twist.

Figure XI.2: A cylinder.

Figure XI.3: A Möbius strip.
Figure XI.3: A Möbius strip.
(v) Let $X = [0, 1]^2$, $A_1 = [0, 1] \times \{0\}$, $B_1 = [0, 1] \times \{1\}$, $A_2 = \{0\} \times [0, 1]$, $B_2 = \{1\} \times [0, 1]$. Define $f_1 : A_1 \rightarrow B_1$ by $f(t,0) = (t,1)$ and $f_2 : A_2 \rightarrow B_2$ by $f_2(0,t) = (1,t)$. The identification space is homeomorphic to a 2-torus $T^2 \cong S^1 \times S^1$. It can be made from a rectangle by first gluing the top and bottom edges to form a cylinder, and then gluing the end circles (with no twist); see Figure XI.4, from http://bakingandmath.com/tag/torus/. (See this website, or https://en.wikipedia.org/wiki/Homotopy, for an animation of the topological equivalence of a torus and a coffee cup; the old joke is that a topologist is a mathematician who cannot tell the difference between a doughnut and a coffee cup.)

![Figure XI.4: Construction of a torus.](image)

(vi) Let $X = [0, 1]^2$, $A_1 = [0, 1] \times \{0\}$, $B_1 = [0, 1] \times \{1\}$, $A_2 = \{0\} \times [0, 1]$, $B_2 = \{1\} \times [0, 1]$. Define $f_1 : A_1 \rightarrow B_1$ by $f(t,0) = (t,1)$ and $f_2 : A_2 \rightarrow B_2$ by $f_2(0,t) = (1,1-t)$. The identification space is called a **Klein bottle** (actually in German it is *Kleinsche Fläche*, not *Kleinsche Flasche*; the English name may be a mistranslation). It can be made from a rectangle by first gluing the top and bottom edges to form a cylinder, and then gluing the end circles with an orientation-reversing map; see Figure XI.5 (from https://en.wikipedia.org/wiki/Klein_bottle) or, for an animated version, https://www.youtube.com/watch?v=E8rifKlq5hc; note that the apparent self-intersection is only an artifact of trying to depict the Klein bottle as a subset of $\mathbb{R}^3$ – the picture is really a higher-dimensional version of a knot diagram in $\mathbb{R}^2$. A Klein bottle cannot be topologically embedded in $\mathbb{R}^3$, but it can be embedded in $\mathbb{R}^4$. A Klein bottle can also be constructed from a Möbius strip by suitably identifying points on the boundary circle (Figure XI.6, from http://squareone-learning.com/blog/2014/09/around-and-round-and-never-off-the-surface/), or by gluing two Möbius strips together along their boundary circle. There are many nice pictures and videos of the Klein bottle on the internet, e.g. https://www.youtube.com/watch?v=sRTKSzA0Br4.
Figure XI.5: Construction of a Klein bottle.

Figure XI.6: Construction of a Klein bottle from a Möbius strip.
(vii) Let $X = [0, 1]^2$, $A_1 = [0, 1] 	imes \{0\}$, $B_1 = [0, 1] \times \{1\}$, $A_2 = \{0\} \times [0, 1]$, $B_2 = \{1\} \times [0, 1]$. Define $f_1 : A_1 \to B_1$ by $f(t, 0) = (1 - t, 1)$ and $f_2 : A_2 \to B_2$ by $f_2(0, t) = (1, 1 - t)$. The identification space is called the projective plane. It can be made from a rectangle by first gluing the top and bottom edges with a twist to form a Möbius strip, and then suitably identifying points on the boundary circle (not in the same way as for a Klein bottle). A better way is to take $A$ as the union of the top and left edges, and $B$ the union of the right and bottom edges (these sets are not quite disjoint), and then gluing $A$ to $B$. This amounts to gluing the top semicircle of the closed unit disk in $\mathbb{R}^2$ to the bottom semicircle by a $180^\circ$ rotation, or identifying $x$ with $-x$ for each $x$ of norm 1. See Figure XI.7, from http://www.math.cornell.edu/~mec/Winter2009/Victor/part3.htm.

A projective plane is harder to picture than a Klein bottle, but there are many ways to do it; see Figure XI.8, from http://xahlee.info/math/riemann_sphere__real_projective_plane.html for one way. Unlike this picture which has a pinch point, there is a (nonobvious – HILBERT thought it was impossible!) way to immerse the projective plane in $\mathbb{R}^3$ called Boy’s surface [Boy03]; see https://www.youtube.com/watch?v=9gRx66xXXek for a nice animation. Note again that the apparent self-intersection is only an artifact of trying to depict the projective plane as a subset of $\mathbb{R}^3$. A projective plane cannot be topologically embedded in $\mathbb{R}^3$, but it can be embedded in $\mathbb{R}^4$. There are several other standard ways to construct the projective plane (XI.9.3.7.). The projective plane is denoted $\mathbb{R}P^2$; there are generalizations (real projective spaces) $\mathbb{R}P^n$ for each $n$ (XI.9.3.7.(iii)), as well as complex projective spaces $\mathbb{C}P^n$ (XI.9.3.7.(iv)).

(viii) Every closed surface (closed 2-manifold) (including $S^2$!) can be obtained by starting with a regular $(2n)$-gon for some $n$ and identifying pairs of edges in an appropriate way (XI.9.5.4.).
Attaching

XI.9.2.4. A variation of gluing is to have two spaces $X$ and $Y$, subsets $A$ of $X$ and $B$ of $Y$, and a homeomorphism (or quotient map) $f: A \to B$. This can be put into the previous setting by considering the disjoint union of $X$ and $Y$. The glued identification space is usually denoted $X \cup_f Y$; we say $X$ is attached to $Y$ via $f$.

XI.9.2.5. There are two especially important examples of attaching: CW complexes and surgery.

Handles and Surgery

Surgery is a method of creating new manifolds from known ones. The general setup of surgery in topology is based on the following important fact:

XI.9.2.6. Proposition. Let $M$ and $N$ be $n$-manifolds with boundary, and let $f: \partial M \to \partial N$ be a homeomorphism. Then $M \cup_f N$ is an $n$-manifold without boundary.
XI.9.2.7. The general idea of surgery is to remove from a (usually closed) $n$-manifold $M$ an open submanifold and glue in another $n$-manifold $N$ whose boundary is homeomorphic to the boundary of the removed piece to obtain a new closed $n$-manifold. The simplest interesting example is:

XI.9.2.8. Let $M$ be a closed surface, e.g. $S^2$. Cut two disjoint open disks out of a larger disk in $M$. What is left is a compact 2-manifold with boundary $M_0$ whose boundary consists of a disjoint union of two circles, which may be given a consistent orientation. The cylinder $N = [0, 1] \times S^1$ is also a closed manifold with boundary, whose boundary consists of two circles. Glue $N$ to $M_0$ along the boundary circles, giving them the same orientation. The result is called $M$ with a handle attached. See Figure XI.9, from http://www.uff.br/cdme/pdp/pdp-html/pdp-en.html. It is easy to see that $S^2$ with a handle attached is homeomorphic to $T^2$, and $T^2$ with a handle attached is a two-holed torus. An $n$-holed torus is obtained by adding $n$ handles to $S^1$ (Figure XI.10, from https://upload.wikimedia.org/wikipedia/commons/a/ac/Sphere_with_three_handles.png). It turns out that every closed surface is homeomorphic to either $S^2$ (if orientable) or $\mathbb{R}P^2$ or a Klein bottle (if nonorientable) with some finite number of handles attached. In the orientable case, the number of handles is called the genus and is a topological invariant. Genus can also be defined in the nonorientable case (XI.9.2.12.).

![Figure XI.9: Attaching a handle to a sphere.](image)

If $N$ is attached to $S^2$ with the end circles given opposite orientations, the surface obtained is the Klein bottle, not the torus.

XI.9.2.9. A variation is to take closed surfaces $M$ and $N$, cut out small disks from both of them, and glue the remainders together along the boundary circles (see Figure XI.11, from https://upload.wikimedia.org/wikipedia/commons/thumb/5/52/Connected_sum.svg/1280px-Connected_sum.svg). The resulting surface is called the connected sum of $M$ and $N$, denoted $M \# N$. The connected sum of an $m$-holed torus and an $n$-holed torus is an $(m + n)$-holed torus. Connected sum with a torus is topologically the same thing as adding a handle.
Figure XI.10: Sphere with three handles.

Figure XI.11: Connected sum.
Connected sum can be done for $n$-manifolds by cutting out a small ball $B^n$ from both and gluing the remainders along the boundary $S^{n-1}$'s. In general, the orientation chosen for the identification of the boundaries affects the homeomorphism class of the connected sum ( ).

XI.9.2.10. In higher dimensions, there is more flexibility in surgery. A key observation is that for any $p$ and $q$, the boundary of $S^p \times B^q$ is homeomorphic to $S^p \times S^{q-1}$, as is the boundary of $B^{p+1} \times S^{q-1}$. (We used this for handles with $p = 0$ and $q = 2$). Thus, if $M$ is an $n$-manifold, we can take any $p$ and $q$ with $p + q = n$, cut out an open submanifold whose closure is homeomorphic to $S^p \times B^q$, and glue in a copy of $B^{p+1} \times S^{q-1}$.

If $n = 3$, the only possibility of this type is to add a “three-dimensional handle,” cutting out two 3-balls (a copy of $S^0 \times B^3$) and replacing it by a copy of $B^1 \times S^2 \cong [0,1] \times S^2$. But if $n = 4$, there is another interesting possibility: cutting out a copy of $S^1 \times B^3$ and replacing it with a copy of $B^2 \times S^2$. See ( ).

XI.9.2.11. Actually, even in three dimensions, there is another interesting possibility: the removed manifold with boundary can be glued back in, but with a different homeomorphism on the boundary. (It may be hard to believe this could make a different manifold, but such a self-homeomorphism of the boundary does not generally extend to a self-homeomorphism of the removed manifold.) This can be done also in two dimensions by removing a handle and gluing it back on reversing the orientation of one of the boundary circles, but the three-dimensional case is more interesting since if the removed piece is a solid torus, there are many essentially different ways of gluing it back in ( ), none of which affect orientability. The removed solid torus may be knotted or unknotted, and this type of surgery is called surgery on a knot. More generally, we can do surgery on a link. It is known (Lickorish-Wallace theorem) that every closed orientable 3-manifold can be obtained from $S^3$ by surgery on a link ([Wal60], [Lic62]).

XI.9.2.12. Other types of surgery are possible. For example, the boundary of a Möbius strip is a circle, so if $M$ is a 2-manifold, we may cut a disk out of $M$ and glue in a Möbius strip (called a cross-cap). A 2-sphere with a cross-cap added is a projective plane (Figure XI.8). Adding a cross-cap is the same thing as taking connected sum with a projective plane. Every nonorientable closed surface is obtained by adding a number of cross-caps to a sphere; this number is the genus ( ). Similar things can be done in higher dimensions.

XI.9.3. Group Actions on Spaces

XI.9.3.1. Another important source of quotients arises from group actions. A group of homeomorphisms of a topological space $X$ is a set $G$ of homeomorphisms of $X$ such that

(i) $G$ is closed under composition: if $g, h \in G$, then $g \circ h$ (usually written $gh$) is also in $G$.

(ii) $G$ is closed under inversion: if $g \in G$, then $g^{-1} \in G$.

(recall that compositions and inverses of homeomorphisms are homeomorphisms). It follows that the identity map on $X$ is in $G$.

It turns out to be convenient to change point of view slightly:
XI.9.3.2. **Definition.** Let \( G \) be a group and \( X \) a topological space. An *action* of \( G \) on \( X \) is a homomorphism \( \sigma : G \to \text{Homeo}(X) \) from \( G \) to the homeomorphism group of \( X \).

We usually write \( \sigma_g \) in place of \( \sigma(g) \), and often just write \( g \cdot x \) for \( \sigma_g(x) \). The homomorphism property of \( \sigma \) is that \( \sigma_{gh} = \sigma_g \circ \sigma_h \) (written \( \sigma_g \sigma_h \)), or \( (gh) \cdot x = g \cdot (h \cdot x) \), for \( g, h \in G \), \( x \in X \); we have \( \sigma_{e^{-1}} = \sigma_e^{-1} \) for \( e \in G \), and \( \sigma_e \) is the identity homeomorphism if \( e \) is the identity of \( G \).

XI.9.3.3. The group action picture includes the group of homeomorphisms picture: if \( G \) is a group of homeomorphisms of \( X \), the inclusion map of \( G \) into \( \text{Homeo}(X) \) is an action of \( G \) on \( X \). Conversely, if \( \sigma \) is an injective homeomorphism from \( G \) to \( \text{Homeo}(X) \) (we say \( \sigma \) is a *faithful* action of \( G \) on \( X \)), then \( G \) can be identified with the group \( \sigma(G) \) of homeomorphisms of \( X \). The group action picture is slightly more general since group actions need not be faithful (in the theory of group representations () it turns out to be important to allow nonfaithful representations). The conceptual importance of the group action picture is that the group \( G \) can be considered abstractly and is not *a priori* tied to homeomorphisms of a particular space, and it is useful to let the same abstract group act on many different spaces.

Group actions of groups on sets can be similarly defined with no relation to topology: just consider homomorphisms from a group to the group of bijections (permutations) of the set. This can be regarded as a special case of group actions of groups on topological spaces where the space has the discrete topology.

Actions of topological groups on topological spaces are also considered (). In this section, we will not consider the case of topological groups (or, alternatively, we will give all groups the discrete topology).

**Orbits**

XI.9.3.4. **Definition.** Let \( \sigma \) be an action of a group \( G \) on a topological space \( X \), and \( x \in X \). The *orbit* of \( x \) under \( G \) (or \( \sigma \)) is the set

\[
[x] = \{ g \cdot x : g \in G \} = \{ \sigma_g(x) : g \in G \}.
\]

Of course, \( x \in [x] \) since \( x = e \cdot x \). The crucial fact about orbits is that they form a partition of the space:

XI.9.3.5. **Proposition.** Let \( \sigma \) be an action of a group \( G \) on a topological space \( X \). If \( x, y \in X \), then either \( [x] = [y] \) or \( [x] \) and \( [y] \) are disjoint. Equivalently, the relation \( x \sim y \) if \( [x] = [y] \) is an equivalence relation on \( X \).

**Proof:** Suppose \( z \in [x] \cap [y] \). Then there are \( g, h \in G \) with \( z = g \cdot x \) and \( z = h \cdot y \). Then \( y = h^{-1} \cdot z = (h^{-1} g) \cdot x \), and similarly \( x = (g^{-1} h) \cdot y \), so \( [x] = [y] \).

XI.9.3.6. Thus, if \( \sigma \) is an action of a group \( G \) on a topological space \( X \), we may form the quotient space \( X/\sigma \) of orbits, called the *orbit space* of the action, and give it the quotient topology. The orbit space is also often written \( X/G \), especially if \( G \) is given concretely as a group of homeomorphisms of \( X \), although this notation conflicts with the notation for identification spaces (XI.9.1.1.).

Many interesting topological spaces can be constructed as orbit spaces of group actions:
XI.9.3.7. Examples. (i) Let $X$ be the unit circle in $\mathbb{R}^2$. The set consisting of the identity and reflection across the $x$-axis is a group of homeomorphisms of $X$. Each orbit consists of two points except for $[(1,0)]$ and $[(-1,0)]$, which are singletons. The orbit space is homeomorphic to the closed interval $[-1,1]$, and the quotient map can be identified with projection onto the $x$-axis.

(ii) Let $\mathbb{Z}$ act on $\mathbb{R}$ by translation: $n \cdot x = x + n$. Each orbit contains a unique point in the interval $(0,1]$. But the quotient topology is not the subspace topology on $(0,1]$; for example, we have $\left[\frac{1}{n}\right] \to [0] = [1]$. In fact, the quotient space $\mathbb{R}/\mathbb{Z}$ is homeomorphic to the circle obtained from $[0,1]$ by identifying 0 and 1.

(iii) Let $X$ be the unit sphere $S^{n-1}$ in $\mathbb{R}^n$. The set consisting of the identity and the antipodal map $x \mapsto -x$ is a group of homeomorphisms of $S^{n-1}$. Every orbit has two points. The quotient space consists of $S^{n-1}$ with antipodal points identified, and is called real projective $(n-1)$-space, denoted $\mathbb{R}P^{n-1}$. If $n = 3$, we obtain the (real) projective plane (XI.9.2.3). If $n = 2$, we obtain the quotient of the circle by the antipodal map, which is a smaller circle. Note that this action of $\mathbb{Z}_2$ on the circle is very different from the action in (i).

(iv) We may also define complex projective space. Let $T$ be the multiplicative group of complex numbers of absolute value 1. Regard $S^{2n-1}$ as the unit sphere in $\mathbb{C}^n$. Then $T$ acts on $S^{2n-1}$ by $\lambda \cdot x = \lambda x$. The quotient $S^{2n-1}/T$ is called complex projective $n$-space, denoted $\mathbb{C}P^{n-1}$. It is a $(2n-2)$-dimensional manifold.

Projective spaces can also be described in other ways. $\mathbb{R}P^n$ can be described as the set of lines through the origin in $\mathbb{R}^{n+1}$, the quotient of $\mathbb{R}^{n+1} \setminus \{0\}$ by identifying two points if they lie on the same line through $0$. Each such line contains exactly two points on the unit sphere, which are identified; this leads to the picture of $\mathbb{R}P^n$ as the closed upper hemisphere of $S^n$ with antipodal points on the “equator” identified, the higher-dimension analog of the picture in XI.9.2.3.(vii). Similarly, $\mathbb{C}P^n$ is the set of complex “lines” through the origin (one-dimensional complex subspaces) in $\mathbb{C}^{n+1}$.

$\mathbb{R}P^n$ can also be described as $\mathbb{R}^n$ with certain points at infinity added: identify $x \in \mathbb{R}^n$ with the line through the origin and the point $(x,1) \in \mathbb{R}^{n+1}$ (or, equivalently, the equivalent point of norm 1 on the upper hemisphere of $S^n$). The points at infinity added to make $\mathbb{R}P^n$ can be regarded as “lines at infinity,” which form a copy of $\mathbb{R}P^{n-1}$. There is a similar description for $\mathbb{C}P^n$ as $\mathbb{C}^n$ with “complex lines at infinity” added (a copy of $\mathbb{C}P^{n-1}$). These pictures are the origin of the name “projective space.”

(v) The 3-sphere $S^3$ can be regarded as the unit sphere in $\mathbb{C}^2$:

$$S^3 = \{(z,w) \in \mathbb{C}^2: |z|^2 + |w|^2 = 1\}.$$

Fix $p > 1$ and let $\lambda = e^{2\pi i/p}$. Then $\lambda$ is a complex number of absolute value 1, and $\lambda^p = 1$. Define a faithful action of $\mathbb{Z}_p$ (or a nonfaithful action of $\mathbb{Z}$) on $S^3$ by

$$k \cdot (z,w) = (\lambda^k z, \lambda^k w).$$

The quotient space $S^3/\mathbb{Z}_p$ is called a lens space. If $p = 2$, we obtain $\mathbb{R}P^3$. More generally, if $q$ is an integer relatively prime to $p$ (we may restrict to $0 < q < p$), we can define the action by

$$k \cdot (z,w) = (\lambda^k z, \lambda^{qk} w).$$

The lens spaces obtained, denoted $L(p,q)$, $L(p:q)$, or $L_{q/p}$, depend rather subtly on $p$ and $q$: they can be homotopy equivalent without being homeomorphic, and, although orientable, can fail to have orientation-reversing homeomorphisms. Lens spaces can also be constructed by gluing two solid tori along their boundary (); see [Mun84] for another construction of lens spaces by identifying points on the boundary of $B^3$, or [Thu97] for more constructions and an explanation of why these are called “lens spaces.”
Homogeneous Spaces

XI.9.3.8. A simple but particularly important example of a group action is when the space \( X \) is a topological group and \( G \) is a subgroup of \( X \). A topological group acts on itself by (right) translation \((\cdot)\), and so does any subgroup, i.e. \( g \cdot x = xg^{-1} \) for \( g \in G, x \in X \) (this formula with the inverse is needed to make \((gh) \cdot x = g \cdot (h \cdot x)\) for \( g, h \in G, x \in X \)). The orbit space \( X/G \), which is the space of left cosets \((\cdot)\) of \( G \) in \( X \), with the quotient topology is called a homogeneous space. (There is additional structure here: the group \( X \) acts on \( X/G \) by left translation. If \( G \) is a normal subgroup of \( X \), then \( X/G \) is also a topological group.) The quotient space has reasonable properties, e.g. is Hausdorff, only when \( G \) is a closed subgroup of \( X \) (cf. Exercise XI.9.5.1.). The simplest case is when \( G \) is a closed discrete subgroup.

Example XI.9.3.7.(ii) is a basic example of this construction. A closely related example is to take \( X = \mathbb{R}^n \) and \( G = \mathbb{Z}^n \) the additive subgroup of integer points; the quotient is the \( n \)-torus \( T^n \), which is a product of \( n \) circles. In the case \( n = 2 \) we obtain the ordinary torus (XI.9.2.3.(v)); higher-dimensional tori can be similarly constructed by identifying opposite faces of an \( n \)-cube.

XI.9.3.9. The lens spaces \( L(p, 1) \) (XI.9.3.7.(v)) are another example. The topological group \( SU(2) \) (also isomorphic to \( \text{Sp}(1) \) and to the multiplicative group of unit quaternions) is homeomorphic to \( S^3 \), and we can take the quotient by the cyclic subgroup generated by \( \text{diag}(\lambda, \lambda) \). Other lens spaces do not arise as homogeneous spaces of \( SU(2) \).

\( SU(2) \) has other interesting finite subgroups, perhaps the most interesting of which is the binary icosahedral group, which covers the group of rigid motions of a regular icosahedron centered at the origin (a subgroup of \( SO(3) \), which is a quotient of \( SU(2) \)); the quotient of \( SU(2) \cong S^3 \) by this group is called the Poincaré homology sphere, a closed 3-manifold with the same homology as \( S^3 \) but not homeomorphic to \( S^3 \) (it is not simply connected; its fundamental group is the binary icosahedral group, a group of order 120 with no nontrivial abelian quotients). The Poincaré homology sphere can also be constructed from a regular dodecahedron by identifying opposite faces with an appropriate twist \((\cdot)\). See [Thu97] and [Wee02] for a nice description and for many other examples of 3-manifolds obtained from polyhedra by gluing faces (cf. XI.9.5.6.).

Stabilizers and Freeness

XI.9.4. Local Homeomorphisms and Covering Maps

XI.9.5. Exercises

XI.9.5.1. (a) Let \( X = \mathbb{R} \) and \( A = \mathbb{Q} \), and consider the identification space \( X/A \). Show that all the conclusions of XI.9.1.2. fail, and \( X/A \) is not \( T_1 \). Is \( X/A \ T_0 \)?

(b) Let \( \mathbb{R}/\mathbb{Q} \) be the homogeneous space (XI.9.3.8.) considering \( \mathbb{Q} \) as an additive subgroup of \( \mathbb{R} \). Show that \( \mathbb{R}/\mathbb{Q} \) is the space of XI.8.1.11., and thus has the indiscrete topology.

Note that, although the notation \( \mathbb{R}/\mathbb{Q} \) is used in both (a) and (b), the spaces are quite different.

XI.9.5.2. Figures XI.12 and XI.13 show orientable closed surfaces. By the classification, each is homeomorphic to \( S^2 \) with some number of handles. Find the number of handles (genus), and describe a homeomorphism. (There are not only homeomorphisms, but even isotopies; cf. https://www.youtube.com/watch?v=k8Rxep2Mkp8.)

1087
Figure XI.12: An orientable closed surface.

Figure XI.13: Another orientable closed surface.
XI.9.5.3. (a) Show that $\mathbb{R}P^2 \# \mathbb{R}P^2$ is homeomorphic to the Klein bottle $K$.

(b) Show that $\mathbb{R}P^2 \# K$ is homeomorphic to $\mathbb{R}P^2 \# T^2$, and hence $\mathbb{R}P^2 \# T^2$ is homeomorphic to $\mathbb{R}P^2 \# \mathbb{R}P^2 \# \mathbb{R}P^2$.

(c) Show that $\mathbb{R}P^2$ with $n$ handles attached is homeomorphic to the connected sum of $2n + 1$ copies of $\mathbb{R}P^2$, and that $K$ with $n$ handles attached is homeomorphic to the connected sum of $2n + 2$ copies of $\mathbb{R}P^2$.

XI.9.5.4. Let $P$ be a polygon with $2n$ sides, which we may take to be a regular polygon. A gluing scheme on $P$ consists of a choice of an orientation and a labeling of each edge of $P$, such that each label names exactly two edges. A gluing scheme defines a gluing of $P$ by gluing each pair of edges with the same label according to the designated orientation. For example, the gluing scheme of Figure XI.5 gives a Klein bottle when glued.

(a) Show that the gluing scheme of Figure XI.14 gives a two-holed torus.

(b) Show that the gluing schemes of Figure XI.15 both give $S^2$ when glued.
(c) Show that in any gluing scheme, two consecutive edges with the same label and opposite orientations can be eliminated. [Fold them into the interior of the polygon.]

(d) Show that if a gluing scheme contains a labeled pair of edges with the same orientation (with respect to the boundary circle), then the glued space contains a Möbius strip. [See Figure XI.16.] Thus the glued space is not orientable.

(e) Show that if all pairs of glued edges have opposite orientations, a sphere with some number of handles (perhaps none) is obtained from the gluing.

(f) In the situation of (d), show that the glued space consists of adding a cross-cap to a glued space from a polygon with fewer sides. Conclude by induction that the glued space is a connected sum of projective planes, or, equivalently, a sphere with a number of cross-caps added (cf. Exercise XI.9.5.3.).

(g) Conclude from (e) and (f) that the glued space is always a closed surface which is either a sphere with handles attached if orientable, or a sphere with cross-caps attached (a connected sum of projective planes) if nonorientable. Show that every such surface can be obtained this way (although the polygon and gluing scheme are not unique).

(h) Show that if a connected closed surface can be triangulated, it can be represented as the gluing of a polygon by a gluing scheme. Since every 2-manifold can be triangulated, every connected closed surface (2-manifold) is of the form of (g).

(i) Show that no two of the surfaces of (g) with different numbers of handles or projective planes are homeomorphic. Thus (g) gives a complete classification of closed 2-manifolds.

See [Mas77] (or [Mas91]), or [Blo97] (from which the above pictures are taken), for a nice almost complete discussion of the classification of closed 2-manifolds. See [Tho92] for a different approach, and [FW99] (or [Wee02] for a modern proof due to J. Conway. (Note that these proofs are all incomplete since they use the nontrivial fact that every surface can be triangulated, first proved by T. Radó in 1925 [?], cf. [AS60]. For a short proof of triangulability see [DM68] or [Tho92], and http://arxiv.org/pdf/1312.3518.pdf for a different argument not using the Jordan-Schönflies Theorem. See [Ber03, p. 139] for a simple approach to triangulating a compact surface using Riemannian geometry.) For a complete classification of connected noncompact (triangulable) surfaces, see [Ric63].

XI.9.5.5. (a) Let \(X = [0, 1], n \in \mathbb{N}, A_n = \{\frac{k}{n} : 0 \leq k \leq n\}\). Show that \(X/A_n\) is a bouquet of \(n\) circles. In particular, \(X/A_2\) is a figure 8. Show that \(X/A_n\) is the one-point compactification of the disjoint union of \(n\) open intervals (copies of \(\mathbb{R}\)).

(b) Let \(X = [0, 1], A_\omega = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}\). Show that \(X/A_\omega\) is homeomorphic to the Hawaiian earring (XI.18.4.6.). Show that \(X/A_\omega\) is the one-point compactification of the disjoint union of \(\mathbb{R}_0\) open intervals (copies of \(\mathbb{R}\)).

(c) Let \(K\) be the Cantor set. Show that the quotient \([0, 1]/K\) is also homeomorphic to the Hawaiian earring. [The complement is a countable disjoint union of open intervals.] What about the quotient by identifying other compact subsets of \([0, 1]\)?

XI.9.5.6. [Thu97] Identify the faces of two solid tetrahedra as in Figure XI.16 (there is a unique identification consistent with the picture). Note that the six thick edges are all identified together, as are the six thin edges and all eight vertices.

(a) Show that every point of the identification space \(X\) has a neighborhood homeomorphic to \(\mathbb{R}^3\) except the point \(v\) coming from the vertices. Thus, if \(Y = X \setminus \{v\}\), then \(Y\) is a (noncompact) 3-manifold. If \(Z\) is \(X\)
with a suitable small open neighborhood of \( v \) removed, then \( Z \subseteq Y \) is a compact 3-manifold with boundary homeomorphic to a 2-torus, and there is a deformation retraction of \( Y \) onto \( Z \).

(b) Show that \( Y \) is homeomorphic to the complement in \( S^3 \) of a figure 8 knot \((\cdot)\), and \( Z \) is the complement of a tubular neighborhood of the knot.

(c) [Thu97, p. 129-132] Show similarly that there are two gluing schemes on a pair of octahedra with vertices removed, which respectively give the complement in \( S^3 \) of the Whitehead link (XI.18.5.8.) and the Borromean rings (XI.18.5.4.). Generalize.

**XI.9.5.7.** [Eng89, 1.6.18] Consider the quotient space \( X \) obtained from \( \mathbb{R} \) by collapsing \( Z \) to a point \( z \) (in the notation of XI.9.1.1., this is \( \mathbb{R} / Z \), but this notation is generally reserved for the quotient group \( T \) as in XI.9.3.8., quite a different space.)

(a) Show that \( X \) is perfectly normal (XI.10.4.10.). In particular, \( \{ z \} \) is a \( G_\delta \).

(b) Show that \( z \) does not have a countable local base. Thus \( X \) is not first countable.

(c) Show that \( X \) is a Fréchet-Urysohn space (XI.3.5.1.).

(d) Describe the topological differences, if any, between this space, the Hawaiian earring (XI.18.4.6.), the space of XI.18.4.7., an infinite wedge of circles \((\cdot)\), and the one-point compactification of a countable separated union of copies of \( \mathbb{R} \) (cf. XI.9.5.8.).

**XI.9.5.8.** (a) As a special case of XI.8.5.1.(a), show that collapsing subsets of a space can be done in two or more steps: if \( (A_j) \) and \( (B_k) \) respectively are families of pairwise disjoint subsets of a topological space \( X \) and each \( A_j \) is a union of \( B_k \)'s, the quotient space of \( X \) obtained by collapsing each \( A_j \) to a (separate) point can be obtained by first collapsing the \( B_k \) to (separate) points and then collapsing the set of points corresponding to each \( A_j \) to a point.

(b) As a special case, let \( X \) be a countable separated union of copies of \([0,1]\), indexed by \( \mathbb{Z} \). First identify the endpoints of the \( n \)'th interval to a single point \( *_n \) to obtain a separated union of circles with distinguished
point. Then identify all the $*_{n}$ to a single point. Show that an infinite wedge of circles is obtained.

(c) On the other hand, first identify the right endpoint of the $n$'th interval with the left endpoint of the $(n + 1)$'st interval for each $n$. Show that the quotient space is homeomorphic to $\mathbb{R}$ with the identified points corresponding to $\mathbb{Z}$. Then identify these points together to obtain the space from XI.9.5.7.

(c) Conclude that both the space of XI.9.5.7. and a (countably) infinite wedge of circles are homeomorphic to the quotient of $X$ obtained by identifying all endpoints of all the intervals together to one point, and, in particular, the first two spaces are homeomorphic.
XI.10. Metrizable Spaces

Although the notion of topological space was designed to abstract the concept of a metric space in order to take specific reference to a metric out of the picture, metrizable spaces (topological spaces whose topology comes from a metric) have some nice properties not shared by more general topological spaces, and it is useful to know which topological spaces are metrizable. In this section we give (partial) answers to this question and pursue some of the special properties of metrizable spaces. Completely metrizable spaces (ones whose topology comes from a complete metric) are especially important and also have intrinsic characterizations.

Note that a metrizable space is not the same thing as a metric space: a metrizable space is a topological space for which there exists a metric (which is highly non-unique) giving the topology. A metric space can be regarded as a (metrizable) topological space with a specified metric giving its topology.

XI.10.1. Metrizability

It is a nontrivial problem to determine exactly when a topology comes from a metric, i.e. when a topological space is metrizable. Necessary conditions are that the space be first countable and completely normal (XI.10.1.2.), but these are not sufficient (>). (Note that any subspace of a metrizable space is metrizable).

The first observation is that a metric space \((X, \rho)\) is Hausdorff: if \(x, y \in X \ x \neq y\), and \(\epsilon = \frac{\rho(x,y)}{3}\), then \(B_\epsilon(x)\) and \(B_\epsilon(y)\) are disjoint neighborhoods of \(x\) and \(y\). We also have the following strengthening of \(T_4\):

\[
\text{XI.10.1.1. Proposition. Let } X \text{ be a metrizable space, } A, B \subseteq X \text{ with } \bar{A} \cap B = A \cap \bar{B} = \emptyset \text{ (such } A \text{ and } B \text{ are said to be separated). Then } A \text{ and } B \text{ have disjoint open neighborhoods in } X. \\
\text{Proof: Let } \rho \text{ be a metric on } X \text{ defining the given topology. The conditions on } A \text{ and } B \text{ are equivalent to: } \\
\rho(x, B) > 0 \text{ for all } x \in A \text{ and } \rho(x, A) > 0 \text{ for all } x \in B. \text{ Set } \\
U = \{x \in X : \rho(x, A) < \rho(x, B)\} \\
V = \{x \in X : \rho(x, B) < \rho(x, A)\} \\
\text{and note that } A \subseteq U, B \subseteq V. \text{ An easy application of the triangle inequality shows that } U \text{ and } V \text{ are open, and they are obviously disjoint.}
\]

\[
\text{XI.10.1.2. A } T_1 \text{ topological space satisfying the conclusion of this proposition is called } \text{completely normal. A completely normal space is obviously normal. A topological space } X \text{ is completely normal if and only if every subspace of } X \text{ is normal (XI.10.4.10.), and thus a completely normal space is sometimes called } \text{hereditarily normal. This property is also sometimes called } T_5. \\
\text{The first metrization theorem, and still probably the most useful, was due to Urysohn in (>):}
\]
XI.10.1.3. **Theorem.** [Urysohn Metrization Theorem] Every second countable regular topological space is metrizable.

**Proof:** A second countable regular space is Lindelöf (XI.1.3.5.), hence normal (XI.7.6.11.). Let $X$ be a second countable regular (hence normal) space, and let $U$ be a countable base for the topology of $X$. For each pair $(U, V)$ of sets in $U$ with $V \subseteq U$, let $f_{UV}$ be a continuous function from $X$ to $[0, 1]$ which is 1 on $V$ and 0 on $U^c$; such a function exists by Urysohn’s Lemma (XI.7.6.12.). Let $C$ be the set of all the $f_{UV}$. Then $C$ is countable. If $E$ is a closed set in $X$ and $x \in X \setminus E$, then $E^c$ is an open neighborhood of $x$, so there is a $U \in U$ with $x \in U \subseteq E^c$. There is an open neighborhood $W$ of $x$ with $W \subseteq U$ (XI.7.5.1.), and there is a $V \in U$ with $x \in V \subseteq W$. Then $V \subseteq U$, and $f_{UV}$ satisfies $f_{UV}(x) = 1$ and $f_{UV} = 0$ on $E$. Thus by the proof of XI.7.7.4., $X$ embeds into the cube $[0, 1]^C$. Since $C$ is countable, this cube is metrizable ().

Note that the proof shows slightly more than the statement of the theorem: any second countable regular space embeds in the Hilbert cube ().

XI.10.1.4. There are many other metrization theorems, some of which are quite recent, and there are now various known necessary and sufficient conditions for metrizability. We will state only one of these results here since too many additional definitions are required for the others, and the results are of limited use in analysis. See [SS95] for a survey. There is one version which can be easily stated (cf. [Dug78, p. 207]):

XI.10.1.5. **Theorem.** Let $(X, T)$ be a topological space. Then $X$ is metrizable if and only if it is regular and has a sequence $(U_n)$ of open covers such that if $V$ is any finite open cover of $X$, $U_n$ refines $V$ for some $n$.

**Metrization of Products and Separated Unions**

It is fairly obvious that a finite product of metrizable spaces is metrizable (). It is not true that an arbitrary product of metrizable spaces is metrizable; an uncountable product of nontrivial Hausdorff spaces is not even first countable (XI.6.1.1.). But a *countable* product of metrizable spaces is metrizable. And an arbitrary separated union of metrizable spaces is metrizable.

The technical device giving a simple proof of both facts is:

XI.10.1.6. **Proposition.** Let $\rho$ be a metric on a set $X$. Then there is a uniformly equivalent metric $\tilde{\rho}$ on $X$ with $\tilde{\rho}(x, y) \leq 1$ for all $x, y \in X$. If $\rho$ is complete, then $\tilde{\rho}$ is complete.

**Proof:** There are two standard ways to construct such a metric. One way to do it is to set

$$\tilde{\rho}(x, y) = \frac{\rho(x, y)}{1 + \rho(x, y)} .$$

The triangle inequality needs proof for this metric, which is nontrivial but not difficult (see below). There is, however, a different low-tech way to define $\tilde{\rho}$ which is perhaps less elegant but does the trick more simply: let

$$\tilde{\rho}(x, y) = \min(\rho(x, y), 1) .$$
Proving the triangle inequality for this version of \( \tilde{\rho} \) is just a trivial case-by-case verification. For either definition, we have \( \tilde{\rho}(x, y) \leq \rho(x, y) \) for all \( x, y \in X \) and \( \rho(x, y) \leq 2\tilde{\rho}(x, y) \) if \( \rho(x, y) \leq 1 \), from which it follows easily that \( \rho \) and \( \tilde{\rho} \) are uniformly equivalent. Since uniformly equivalent metrics have the same Cauchy sequences, the completeness statement follows.

For the reader’s edification, we give a proof of the triangle inequality for the metric

\[
\tilde{\rho}(x, y) = \frac{\rho(x, y)}{1 + \rho(x, y)}.
\]

The proof is not difficult, but requires a little trick. Let \( x, y, z \in X \). We may assume \( x \neq z \). Then

\[
\tilde{\rho}(x, z) = \frac{\rho(x, z)}{1 + \rho(x, z)} = \frac{1}{\rho(x, z) + 1} \leq \frac{1}{\rho(x, y) + \rho(y, z) + 1}.
\]

\[
= \frac{\rho(x, y) + \rho(y, z)}{1 + \rho(x, y) + \rho(y, z)} + \frac{\rho(y, z)}{1 + \rho(x, y) + \rho(y, z)}.
\]

\[
\leq \frac{\rho(x, y)}{1 + \rho(x, y)} + \frac{\rho(y, z)}{1 + \rho(y, z)} = \tilde{\rho}(x, y) + \tilde{\rho}(y, z).
\]

\[
\tilde{\rho}(x, z) \leq \tilde{\rho}(x, y) + \tilde{\rho}(y, z) \leq \tilde{\rho}(x, z) + \tilde{\rho}(x, y).
\]

\[
\tilde{\rho}(x, y) = \frac{\rho(x, y)}{1 + \rho(x, y)}.
\]

XI.10.1.7. **Theorem.** A countable product of metrizable spaces is metrizable in the product topology. A countable product of completely metrizable spaces is completely metrizable.

**Proof:** Let \((X_k)\) be a sequence of metrizable spaces, and for each \( k \) let \( \rho_k \) be a metric on \( X_k \) giving the topology on \( X_k \), with the property that \( \rho_k(x, y) \leq 1 \) for all \( x, y \in X_k \), with \( \rho_k \) complete if \( X_k \) is completely metrizable (XI.10.1.6.).

On \( X = \prod_{k \in \mathbb{N}} X_k \), define

\[
\rho(x, y) = \sum_{k=1}^{\infty} 2^{-k} \rho_k(x_k, y_k)
\]

for \( x = (x_1, x_2, \ldots) \) and \( y = (y_1, y_2, \ldots) \) in \( X \). Note that the infinite series converges absolutely for any \( x, y \). It is simple and routine to verify that \( \rho \) is a metric on \( X \). So we need to show that \( \rho \) defines the product topology. It suffices to show that, for fixed \( x = (x_1, x_2, \ldots) \in X \), every \( \rho \)-ball around \( x \) contains an open rectangle around \( x \), and conversely.

Let \( \epsilon > 0 \). Fix \( n \) with \( 2^{-n+1} < \epsilon \). Then

\[
B_{\epsilon/2}^{\rho_1}(x_1) \times B_{\epsilon/2}^{\rho_2}(x_2) \times \cdots \times B_{\epsilon/2}^{\rho_n}(x_n) \times \prod_{k=n+1} B_{\epsilon/2}^{\rho_k}(x_k)
\]

is an open rectangle in \( X \) containing \( x \) and contained in \( B_{\epsilon}^{\rho}(x) \): if \( y = (y_1, y_2, \ldots) \in X \) with \( \rho_k(x_k, y_k) < \frac{\epsilon}{2^n} \) for \( 1 \leq k \leq n \), then

\[
\sum_{k=1}^{\infty} 2^{-k} \rho_k(x_k, y_k) = \sum_{k=1}^{n} 2^{-k} \rho_k(x_k, y_k) + \sum_{k=n+1}^{\infty} 2^{-k} \rho_k(x_k, y_k) < \frac{\epsilon}{2} \sum_{k=1}^{n} 2^{-k} + \sum_{k=n+1}^{\infty} 2^{-k} < \frac{\epsilon}{2} + 2^{-n} = \epsilon.
\]

1095
Conversely, if
\[ V = U_1 \times U_2 \times \cdots \times U_m \times \prod_{k=n+1}^{\infty} X_k \]
is an open rectangle containing \( x \), for \( 1 \leq k \leq m \) there is an \( \epsilon_k \) such that \( B_{\epsilon_k}^m(x_k) \subseteq U_k \). Set
\[ \epsilon = \min(\epsilon_1/2, \epsilon_2/4, \ldots, 2^{-m} \epsilon_m) . \]
If \( y = (y_1, y_2, \ldots) \in X \) with \( \rho(x, y) < \epsilon \), then, for \( 1 \leq k \leq m \)
\[ \rho_k(x_k, y_k) < 2^k \rho(x, y) < 2^k \epsilon \leq \epsilon_k \]
so \( y \in V \), i.e. \( B_{\epsilon}^m(x) \subseteq V \).

If \( (x^{(n)}) = ((x_k^{(n)}) \) is a Cauchy sequence in \( (X, \rho) \), the last inequality shows that \( (x_k^{(n)}) \) is a Cauchy sequence in \( (X_k, \rho_k) \) for each \( k \). If each \( \rho_k \) is complete, then these sequences all converge, and hence \( (x^{(n)}) \) also converges, so \( \rho \) is complete.

This proof requires the Countable AC.

**XI.10.1.8. Proposition.** An arbitrary separated union \( \bigcup \) of metrizable spaces is metrizable. An arbitrary separated union of completely metrizable spaces is completely metrizable.

**Proof:** Let \( \{X_i : i \in I\} \) be a collection of metrizable spaces. Fix a metric \( \rho_i \) on \( X_i \) for each \( i \) giving the topology on \( X_i \), with the property that \( \rho_i(x, y) \leq 1 \) for all \( x, y \in X_i \), with \( \rho_i \) complete if \( X_i \) is completely metrizable (XI.10.1.6.). Let \( X \) be the separated union of the \( X_i \). For \( x, y \in X \), set \( \rho(x, y) = \rho_i(x, y) \) if \( x \) and \( y \) are both in the same \( X_i \), and \( \rho(x, y) = 1 \) otherwise. It is easily checked that \( \rho \) is a metric giving the topology of \( X \), and that \( \rho \) is complete if each \( \rho_i \) is.

This proof uses the AC; the case of \( I \) countable requires the Countable AC.

**Metrizability of Compact Hausdorff Spaces**

**XI.10.1.9. Theorem.** Let \( X \) be a compact Hausdorff space. The following are equivalent:

(i) \( X \) is metrizable.

(ii) \( X \) is second countable.

(iii) \( X \) embeds in the Hilbert cube.

(iv) \( \mathcal{C}_R(X) \) is separable (for the uniform norm).

**Proof:** (i) \( \Rightarrow \) (ii): Let \( d \) be a metric on \( X \). \( (X, d) \) is totally bounded, hence separable, and a separable metric space is second countable.

(ii) \( \Rightarrow \) (iii): Let \( \{U_n : n \in \mathbb{N}\} \) be a countable base for the topology. For each \( n, m \) with \( U_n \subseteq U_m \), let \( f_{n,m} \) be a continuous function from \( X \) to \([0, 1]\) with \( f = 1 \) on \( U_n \) and \( f = 0 \) on \( U_m^c \) (there is such an \( f_{n,m} \) by Urysohn’s
Lemma since $X$ is normal). Set $I = \{(n, m) : \tilde{U}_n \subseteq U_m\} \subset \mathbb{N}^2$ and $Y = [0, 1]^I$ with the product topology. Then $Y$ is homeomorphic to the Hilbert cube (unless $f$ is finite). Define $\phi : X \to Y$ by $\phi(x)_{(n,m)} = f_{n,m}(x)$. Then $\phi$ is continuous by the coordinatewise criterion since each $f_{n,m}$ is continuous. If $x_1, x_2 \in X$, $x_1 \neq x_2$, let $V$ be a neighborhood of $x_1$ not containing $x_2$, $U_m$ a basic neighborhood of $x_1$ contained in $V$, and $U_n$ a basic neighborhood of $x_1$ with $\tilde{U}_n \subseteq U_m$ (there is such an $U_n$ since $X$ is regular). Then $f_{n,m}(x_1) = 1$ and $f_{n,m}(x_2) = 0$, so $\phi(x_1) \neq \phi(x_2)$ and thus $\phi$ is one-to-one. To show $\phi$ is a homeomorphism onto its image, suppose $(x_i)$ is a net in $X$ and $\phi(x_i) \to \phi(x)$ for some $x \in X$. If $x_i \to x$, then there is a neighborhood $V$ of $x$ which does not contain any $x_i$ for a cofinal set of $i$. As in the previous part of the proof, there are $n,m$ such that $x \in U_n \subseteq \tilde{U}_n \subseteq U_m \subseteq V$; then $f_{n,m}(x) = 1$ but $f_{n,m}(x_i) = 0$ for $i$ in a cofinal set, contradicting that $\phi(x_i) \to \phi(x)$. Thus $x_i \to x$.

(iii) $\Rightarrow$ (i) since the Hilbert cube is metrizable.

(iii) $\Rightarrow$ (iv): note first that $C_\mathbb{R}(\mathbb{H})$ is separable: polynomials in the coordinates with rational coefficients are dense by the Stone-Weierstrass Theorem. $C_\mathbb{R}(X)$ is a quotient of $C_\mathbb{R}(\mathbb{H})$ by restriction of functions (this map is surjective by the Tietze Extension Theorem), so $C_\mathbb{R}(X)$ is also separable.

(iv) $\Rightarrow$ (iii): if $C_\mathbb{R}(X)$ is separable, the set $\mathcal{G}$ of continuous functions from $X$ to $[0, 1]$ is also separable in the topology of uniform convergence. Let $\mathcal{F}$ be a countable dense subset of $\mathcal{G}$, and let $Y = [0, 1]^\mathcal{F}$. Then $Y$ is homeomorphic to the Hilbert cube. Define $\phi : X \to Y$ by $\phi(x)_{f} = f(x)$. If $x \in X$ and $V$ is any neighborhood of $x$, there is a $g \in \mathcal{G}$ with $g(x) = 1$ and $g = 0$ on $V'$ since $X$ is completely regular, and $g$ can be uniformly approximated arbitrarily closely by an $f \in \mathcal{F}$. The proof that $\phi$ is a homeomorphism onto its image is then almost identical to the proof of (ii) $\Rightarrow$ (iii).

**XI.10.1.10.** **Corollary.** Let $X$ and $Y$ be compact Hausdorff spaces. If $X$ is metrizable and there is a continuous function from $X$ onto $Y$, then $Y$ is also metrizable.

**Proof:** Let $\phi : X \to Y$ be surjective. Then $f \mapsto f \circ \phi$ embeds $C_\mathbb{R}(Y)$ as a subspace of $C_\mathbb{R}(X)$. Since $C_\mathbb{R}(X)$ is separable and metrizable, so is $C_\mathbb{R}(Y)$.

**XI.10.1.11.** Note that the normed algebra $C(X)$ of complex-valued continuous functions on $X$ is separable if and only if $C_\mathbb{R}(X)$ is separable. Thus, if $X$ is a compact Hausdorff space, $C(X)$ is separable if and only if $X$ is metrizable.

**XI.10.1.12.** Compactness is essential in this result: it fails in general even if $X$ is locally compact and second countable, and $Y$ is perfectly normal. There is a countable perfectly normal space $Y$ which is not first countable (.), hence not metrizable, and $Y$ is a continuous image of $\mathbb{N}$ with the discrete topology. There is even a counterexample if $f$ is a closed map (XI.9.5.7.). The problem seems to be that a single point can be a $G_\delta$ even if the point does not have a countable neighborhood base, which cannot happen in a compact space (XI.11.4.5.).

**XI.10.2. A Universal Space**

We describe a “universal” separable complete metric space:
XI.10.2.1. Definition. Let \( \mathcal{N} = \mathbb{N}^\mathbb{N} \) with the product topology, taking the discrete topology on \( \mathbb{N} \).

\( \mathcal{N} \) consists of all sequences of natural numbers. If \( (a_n) \) is a sequence in \( \mathcal{N} \), say \( a_n = (a_{n,1}, a_{n,2}, \ldots) \), then \( a_n \to b = (b_1, b_2, \ldots) \) if and only if, for each \( k \), there is an \( n_k \) such that \( a_{n,k} = b_k \) for all \( n \geq n_k \).

Recall that set theorists typically write \( \omega \) for \( \mathbb{N} \) (actually \( \mathbb{N} \setminus \{0\} \)), and in set theory \( \mathcal{N} \) is often denoted \( \omega^\omega \).

A notation we will use repeatedly: if \( s = (s_1, s_2, \ldots) \in \mathcal{N} \), write \( s|n = (s_1, \ldots, s_n) \in \mathbb{N}^n \). These are called initial segments of \( s \). Take \( s|0 \) to be the “empty string.”

XI.10.2.2. Proposition. \( \mathcal{N} \) is separable.

Proof: The set of sequences which are eventually 1 is countable and dense in \( \mathcal{N} \).

\[ \mathbb{N} \text{ is a complete metric space under the discrete metric } \delta(a,b) = 1 \text{ if } a \neq b. \]

This is a special case of ( ).

XI.10.2.3. Proposition. The function \( \rho(a,b) = \sum_{k=1}^{\infty} 2^{-k} \delta(a_k, b_k) \) for \( a = (a_k) \) and \( b = (b_k) \) in \( \mathcal{N} \) is a complete metric on \( \mathcal{N} \) giving the product topology.

XI.10.2.4. Proposition. (i) \( \mathcal{N} \) is homeomorphic to \( \mathbb{N} \times \mathcal{N} \), and to \( \{1,\ldots,n\} \times \mathcal{N} \) for any \( n \in \mathbb{N} \). \( \mathcal{N} \) is a disjoint union of a finite or countable sequence of closed subsets, each homeomorphic to \( \mathcal{N} \).

(ii) \( \mathcal{N} \) is homeomorphic to \( \mathbb{N}^n \times \mathcal{N} \) for any \( n \in \mathbb{N} \), and to \( \mathcal{N}^{\mathbb{N}} \).

Proof: (i): The map \( (n, (a_1, a_2, \ldots)) \mapsto (n, a_1, a_2, \ldots) \) is a homeomorphism from \( \mathbb{N} \times \mathcal{N} \) onto \( \mathcal{N} \). If \( n \in \mathbb{N} \),

\[ \{1,\ldots,n\} \times \mathcal{N} \cong \{1,\ldots,n\} \times \mathbb{N} \cong \mathbb{N} \times \mathcal{N} \cong \mathcal{N} \]

taking a one-one correspondence between \( \{1,\ldots,n\} \times \mathbb{N} \) and \( \mathbb{N} \). To obtain an explicit sequence of disjoint closed subsets homeomorphic to \( \mathcal{N} \) partitioning \( \mathcal{N} \), for each \( n \in \mathbb{N} \) let \( \mathcal{N}_n = \{a = (a_k) : a_1 = n\} \).

(ii): If \( n \in \mathbb{N} \), then \( \mathcal{N}^n \cong \mathbb{N}^n \times \{1,\ldots,n\} \), and similarly \( \mathcal{N}^{\mathbb{N}} \cong \mathbb{N}^{\mathbb{N} \times \mathbb{N}} \). Fix one-one correspondences between \( \mathbb{N} \times \{1,\ldots,n\} \) and \( \mathbb{N} \), and between \( \mathbb{N} \times \mathbb{N} \) and \( \mathbb{N} \).

XI.10.2.5. Remark. Note that this proposition does not depend on any form of the Axiom of Choice, since explicit one-one correspondences can be specified (cf. ( ).)

It follows from XI.10.2.2. and XI.10.2.3. that \( \mathcal{N} \) is second countable. An explicit countable base for the topology for \( \mathcal{N} \) can be given:
XI.10.2.6. Proposition. For each \( m \in \mathbb{N} \) and each \((k_1, \ldots, k_m, n_1, \ldots, n_m) \in \mathbb{N}^{2m}\) with \( k_1 < k_2 < \cdots < k_m \), set
\[ U(k_1, \ldots, k_m, n_1, \ldots, n_m) = \{ a = (a_k) : a_{k_j} = n_j \text{ for } 1 \leq j \leq m \}. \]
Then the \( U(k_1, \ldots, k_m, n_1, \ldots, n_m) \) form a base for the topology of \( \mathcal{N} \).

The sets \( U(k_1, \ldots, k_m, n_1, \ldots, n_m) \) can be explicitly arranged in a sequence without using any form of the Axiom of Choice (Exercise (i)).

The space \( \mathcal{N} \) is universal among separable complete metric spaces in the following sense:

XI.10.2.7. Theorem. Let \((X, \sigma)\) be a nonempty separable complete metric space. Then there is a locally contractive (hence continuous) map from \( \mathcal{N} \) onto \( X \).

Proof: By induction on \( n \), choose nonempty closed sets \( F_{s_1, \ldots, s_n} \) for each finite sequence \((s_1, \ldots, s_n) \in \mathbb{N} \) by choosing a sequence \( F_k \) of nonempty closed sets (not necessarily distinct) of diameter \( \leq 1 \) with \( \cup_k F_k = X \), and if sets for sequences of length \( n \) have been chosen, choose nonempty closed sets \( F_{s_1, \ldots, s_n, k} \) for each \((s_1, \ldots, s_n, k) \) of diameter \( \leq 2^{-n} \) such that \( \cup_k F_{s_1, \ldots, s_n, k} = F_{s_1, \ldots, s_n} \) for each \((s_1, \ldots, s_n) \). (Such a collection of subsets is called a Luzin scheme in \( X \).)

The \( F_{s_1, \ldots, s_n} \) can be chosen without any form of choice: fix a dense sequence \((x_k)\) in \( X \), let \( F_k \) be the closed ball of radius \( 2^{-1} \) around \( x_k \), and at the \( n + 1 \)'st stage let \( F_{s_1, \ldots, s_n, k} \) be the intersection of \( F_{s_1, \ldots, s_n} \) with the closed ball of radius \( 2^{-n-1} \) around \( x_k \) if this intersection has nonempty interior, and otherwise the intersection of \( F_{s_1, \ldots, s_n} \) with the closed ball of radius \( 2^{-n-1} \) around \( x_j \) for the first \( j \) for which the intersection has nonempty interior. It is easily checked that these sets have the desired properties.

If \( s = (s_1, s_2, \ldots) \in \mathcal{N} \), there is a unique point in \( \cap_n F_{s_1, \ldots, s_n} \) by completeness; call this point \( f(s) \). Then \( f \) is a map from \( \mathcal{N} \) onto \( X \). If \( s, t \in \mathcal{N} \) with \( \rho(s, t) < 1 \), choose \( n \) so that \( 2^{-n} \leq \rho(s, t) < 2^{-n+1} \); then \( s_j = t_j \) for \( 1 \leq j \leq n \), so \( f(s), f(t) \in F_{s_1, \ldots, s_n} \), \( \sigma(f(s), f(t)) \leq 2^{-n} \), and \( f \) is locally contractive.

Such a map can be explicitly described in some cases:

XI.10.2.8. Examples. (i) Define \( \phi : \mathbb{N} \to \{0, 1\} \) by \( \phi(1) = 1 \) and \( \phi(n) = 0 \) for \( n > 1 \). Applying \( \phi \) coordinatewise, we obtain a map from \( \mathcal{N} \) onto \( \{0, 1\}^\mathbb{N} \) which is a contraction. Since there is a uniform homeomorphism between \( \{0, 1\}^\mathbb{N} \) and the Cantor set \( K \), we obtain a Lipschitz function from \( \mathcal{N} \) onto \( K \). Also, there is a contractive mapping from \( \{0, 1\}^\mathbb{N} \) onto \( [0, 1] \) ( ), hence a contractive mapping from \( \mathcal{N} \) onto \([0, 1] \). Writing \( \mathcal{N} \) as a disjoint union of closed sets \( \mathcal{N}_n \) as in () and taking the corresponding Lipschitz map from \( \mathcal{N}_n \) onto \([n, n + 1] \), we obtain a locally Lipschitz map from \( \mathcal{N} \) onto \([0, +\infty) \). By taking a one-one correspondence between \( \mathbb{N} \) and \( \mathbb{Z} \), we obtain a similar map from \( \mathcal{N} \) onto \( \mathbb{R} \).

(ii) There is a homeomorphism from \( \mathcal{N} \) onto the set \( J \) of irrational numbers given by continued fractions (): the sequence \((n_1, n_2, \ldots)\) corresponds to the number with continued fraction \((n_1, n_2, \ldots)\), i.e.
\[
(n_1, n_2, \ldots) \mapsto n_1 + \frac{1}{n_2 + \frac{1}{n_3 + \cdots}}
\]
(this actually maps to the irrational numbers greater than 1; use a homeomorphism between \( \mathcal{N} \) and \( \mathcal{N} \times \{0, 1\} \) to get a map onto all of \( J \)).
XI.10.2.9. There is an important extension of this result. Recall that if \( X \) is a topological space and \( Y \subseteq X \), a retract of \( X \) onto \( Y \) is a continuous function \( f : X \to Y \) such that \( f(y) = y \) for all \( y \in Y \). \( Y \) is called a retract of \( X \) if there is a retraction from \( X \) onto \( Y \). If \( X \) is Hausdorff, then a retract of \( X \) is necessarily a closed subset of \( X \). It is relatively rare to have a retraction onto a general closed subset of a topological space; however, \( \mathcal{N} \) has the following remarkable property:

XI.10.2.10. Theorem. Let \( Y \) be a nonempty closed subset of \( \mathcal{N} \). Then there is a contractive retraction from \( \mathcal{N} \) onto \( Y \).

The proof is a slight refinement of the proof of XI.10.2.7., and is left to the reader (Exercise ()). Alternatively, an easy proof can be given using the language of trees.

Trees

Trees can be described in various ways. For our purposes, we will take the following definition:

XI.10.2.11. Definition. A tree is a directed graph \( T \) for which:

(i) Every vertex (node) has a rank in \( \mathbb{N} \cup \{0\} \).

(ii) Every edge goes from a node of rank \( n \) to a node of rank \( n + 1 \) for some \( n \).

(iii) There is a unique node of rank 0, called the root.

(iv) For every node \( x \), there is a unique path (necessarily of length equal to the rank of \( x \)) from the root to \( x \). In particular, every node of rank \( n \geq 1 \) has a unique predecessor of rank \( n - 1 \).

A tree \( T \) is pruned if every node has at least one successor.

A subtree of a tree \( T \) is a subgraph which is itself a tree. A subtree is a pruned subtree if it is itself a pruned tree. (We also regard the empty subgraph as a subtree, which is a pruned subtree.)

XI.10.2.12. Example. Take the nodes of rank \( n \) to be \( \{s|n : s \in \mathcal{N}\} \), and take an edge from \( s|n \) to \( s|(n + 1) \) for each \( s \in \mathcal{N} \) and \( n \in \mathbb{N} \cup \{0\} \). This is a pruned tree, called the tree over \( \mathbb{N} \), denoted \( T_{\mathbb{N}} \). The root is the empty string, and \( \mathcal{N} \) can be naturally identified with the infinite paths in \( T_{\mathbb{N}} \).

The proof of the next proposition is a simple exercise which is left to the reader.

XI.10.2.13. Proposition. When \( \mathcal{N} \) is identified with the set of infinite paths in \( T_{\mathbb{N}} \), there is a natural one-one correspondence between the closed subsets of \( \mathcal{N} \) and the pruned subtrees of \( T_{\mathbb{N}} \), where the subset of \( \mathcal{N} \) corresponding to a pruned subtree \( S \) is the set of infinite paths in \( S \).
XI.10.2.14. Proof of Theorem XI.10.2.10.: In light of XI.10.2.13., we need only show that if $S$ is a nonempty pruned subtree of $T_N$, there is a retraction $\phi$ of $T_N$ onto $S$ which preserves rank and edges. Define $\phi$ by induction on rank as follows. Let $\phi(s|0) = s|0$. Suppose $\phi(s|n)$ has been defined for each $s \in \mathcal{N}$. For each node $s|(n + 1)$ (there are countably many distinct nodes of this type), if $s|(n + 1) \in S$, set $\phi(s|(n + 1)) = s|(n + 1)$, and otherwise let $\phi(s|(n + 1))$ be any successor of $\phi(s|n)$ in $S$ (to make it definite and avoid using AC, let it be the successor node in $S$ with smallest possible $(n + 1)'st$ coordinate). It is easily checked that this $\phi$ has the desired properties. □

XI.10.3. Polish Spaces

XI.10.3.1. Definition. A topological space is a Polish space if it is homeomorphic to a separable complete metric space.

Note that a Polish space is a topological space, not a metric space: it does not come with a specified metric, and even if it is given as a metric space, the given metric may not be complete – there need only be an equivalent complete metric (cf. XI.10.3.3.). Any topological space homeomorphic to a Polish space is a Polish space.

XI.10.3.2. Examples. (i) $\mathbb{R}^n$, or more generally any separable Banach space regarded as a topological space. An open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$, hence a Polish space.

(ii) Any (nonempty) closed subset of a Polish space; in particular, any closed subset of $\mathbb{R}^n$.

(iii) Any second countable compact Hausdorff space ().

(iv) The space $\mathcal{N}$ (XI.10.2.1.). More generally, a finite or countable Cartesian product of Polish spaces is Polish with the product topology (XI.10.1.7.).

(v) A countable separated union of Polish spaces is a Polish space (XI.10.1.8.).

A Polish space is second countable, Hausdorff, and normal, and is a Baire space (satisfies the conclusion of the Baire Category Theorem).

XI.10.3.3. Theorem. A subset of a Polish space is a Polish space if and only if it is a $G_\delta$.

Proof: Let $X$ be a Polish space, and fix a complete metric $\rho$ on $X$. Suppose $E$ is a $G_\delta$ in $X$. We could simply write down an explicit formula for a complete metric on $E$ giving the relative topology, but the formula is somewhat complicated and deserves some motivation.

Let $U$ be an open subset of $X$, and for $x \in U$ define $g(x) = \frac{1}{\rho(x,U^c)}$. Then $g$ is a continuous function from $U$ to $\mathbb{R}$, and the graph

$$\Gamma = \{(x,g(x)) : x \in U\} \subseteq X \times \mathbb{R}$$

is a closed subset of $X \times \mathbb{R}$ which is homeomorphic to $U$ via projection onto the first coordinate. $X \times \mathbb{R}$ is a complete metric space under the metric

$$\sigma((x,s),(y,t)) = \rho(x,y) + \min(|s| - |t|, 1) .$$

Hence the restriction of $\sigma$ to $\Gamma$ transfers to a complete metric on $U$, also denoted $\sigma$:

$$\sigma(x,y) = \rho(x,y) + \min(|g(x) - g(y)|, 1) .$$
Now write $E = \cap_n U_n$ with $U_n$ open, and let $g_n$ and $\sigma_n$ be the corresponding function and metric for $U_n$. For $x, y \in E$, set
\[
\tau(x, y) = \sum_{n=1}^{\infty} 2^{-n} \sigma_n(x, y) = \rho(x, y) + \sum_{n=1}^{\infty} 2^{-n} \min(|g_n(x) - g_n(y)|, 1).
\]

Then $\tau$ is a metric on $E$, and it is easily shown (Exercise () that $\tau$ induces the relative topology on $E$. If $(x_k)$ is a $\tau$-Cauchy sequence, then $(x_k)$ is a $\rho$-Cauchy sequence, hence $\rho$-converges to some $x \in X$. Since $(x_k)$ is also a $\sigma_n$-Cauchy sequence for each $n$, it $\sigma_n$-converges to some $x^{(n)} \in U_n$. But the $\sigma_n$-topology on $U_n$ is the relative topology, so $x_k \to x^{(n)}$ in $X$ and hence $x^{(n)} = x$, i.e. $x \in U_n$ for all $n$, so $x \in E$ and $x_k$ $\tau$-converges to $x$; thus $(E, \tau)$ is complete. Since every subset of a separable metric space is separable, $E$ is a Polish space.

For the converse, suppose $E \subset X$ is a complete metric space under a metric $\tau$ giving the relative topology. We may assume that $E$ is dense in $X$, since if $Y$ is the closure of $E$, then $Y$ is a $G_\delta$ in $X$, so every $G_\delta$ in $Y$ is a $G_\delta$ in $X$. For $n \in \mathbb{N}$, let $U_n$ be the set of all $x \in X$ such that there is a neighborhood $V^{(n)}_x$ of $x$ with the $\tau$-diameter of $E \cap V^{(n)}_x$ less than $\frac{1}{n}$. Then $U_n$ is open in $X$. Set $F = \cap_n U_n$; then $F$ is a $G_\delta$ containing $E$. If $x \in F$, choose a sequence $x_n \in E \cap V^{(n)}_x$. Then $(x_n)$ is a $\tau$-Cauchy sequence in $E$, hence converges to an element of $E$ we call $f(x)$. The element $f(x)$ is independent of the choice of the $x_n$ and even the $V^{(n)}_x$ (any neighborhood of $x$ has nonempty intersection with $E$). Then $f : F \to E$ is continuous, and if $x \in E$, we have $f(x) = x$. Thus $E = \{x \in F : f(x) = x\}$, so $E$ is (relatively) closed in $F$. Since $E$ is dense in $F$, $E = F$ is a $G_\delta$ in $X$.

**XI.10.3.4.** Note that separability is not used in the proof (except in the last sentence of the first part), so this result also holds for topological spaces homeomorphic to complete metric spaces, not necessarily separable; such spaces are sometimes called absolute $G_\delta$‘s.

We have the following result of Kuratowski [], which is a sharpening of XI.10.2.7. The proof is rather involved, although fairly elementary, and we omit it; see, for example, [?, 2.4.1].

**XI.10.3.5.** **Theorem.** Let $X$ be an uncountable Polish space which is dense in itself. Then there is an injective continuous function from $\mathcal{N}$ onto $X$ such that the image of every open set is an $F_\sigma$.

Since by the Cantor-Bendixson Theorem (XI.2.5.4.) every Polish space is a disjoint union of a countable open subset and a closed subset which is dense-in-itself, we obtain:

**XI.10.3.6.** **Corollary.** Every uncountable Polish space has cardinality $2^{\aleph_0}$.

**How Many Polish Spaces Are There?**

**XI.10.3.7.** **Theorem.** There are exactly $2^{\aleph_0}$ Polish spaces up to homeomorphism.

**Proof:** Let $\mathcal{P}$ be the set of homeomorphism classes of infinite Polish spaces (there are countably many finite Polish spaces, which are just finite discrete spaces). There are $2^{\aleph_0}$ pairwise nonhomeomorphic compact
subsets of $\mathbb{R}^2$ (XI.18.4.1.), each of which is a Polish space, so $\text{card}(\mathcal{P}) \geq 2^{\aleph_0}$. On the other hand, if $X$ is an infinite Polish space, $\rho$ is a complete metric on $X$, and $D$ is a countable dense set in $X$, then $(X, \rho)$ is the completion of $(D, \rho|_D)$. If $D$ is identified with $\mathbb{N}$, the metric $\rho|_D$ can be transferred to $\mathbb{N}$; thus $X$ is homeomorphic to the completion of $\mathbb{N}$ with respect to some metric. So if $\mathcal{M}$ is the set of all metrics on $\mathbb{N}$, $\text{card}(\mathcal{M}) = 2^{\aleph_0}$ and there is a function from $\mathcal{M}$ onto $\mathcal{P}$. Therefore $\text{card}(\mathcal{P}) \leq 2^{\aleph_0}$.

**XI.10.3.8.** The above proof uses the AC (how?). The full AC is not needed: it suffices to assume that $\mathbb{R}$ can be well ordered, i.e. that $2^{\aleph_0}$ is an aleph. For the best that can be proved in ZF, see Exercise XI.10.4.9.

**XI.10.4. Exercises**

**XI.10.4.1.** Arrange the tuples occurring as indexes in () into an explicit sequence, as follows:

(a) The set of pairs $((n_1, \ldots, n_m), (k_1, \ldots, k_m))$ ($m = 0$ is allowed, giving the “empty pair”), with the $n_j$ strictly increasing in $\mathbb{N}$ and the $k_j \in \mathbb{N}$, is in natural one-one correspondence with the set $\mathcal{C}$ of sequences in $\mathbb{N} \cup \{0\}$ which are eventually 0: the pair $((n_1, \ldots, n_m), (k_1, \ldots, k_m))$ corresponds to the sequence with $k_j$ in the $n_j$th place and zeroes elsewhere.

(b) The set $\mathcal{C}$ is in one-one correspondence with $\mathbb{N}$:

$$(t_1, t_2, \ldots) \longleftrightarrow \prod_j p_j^{t_j}$$

where $p_j$ is the $j$th prime.

**XI.10.4.2.** In the proof of XI.10.3.3., show that $\tau$ gives the relative topology on $E$. [Hint: Suppose $(x_m) \subseteq E$ $\rho$-converges to $x \in E$. Given $\epsilon > 0$, fix $n$ so that $2^{-n} < \epsilon/3$, and find $M$ so that $\rho(x_m, x) < \epsilon/3$ and $|g_k(x_m) - g_k(x)| < \epsilon/3$ for all $m \geq M$ and $1 \leq k \leq n$.]

**XI.10.4.3.** (a) Show that if $K$ is a compact subset of $\mathcal{N}$, then for each $n \in \mathbb{N}$ there is a $k_n \in \mathbb{N}$ such that $s_n \leq k_n$ whenever $s = (s_1, s_2, \ldots) \in K$. [Consider the image of $K$ under the $n$th coordinate map $\pi_n : \mathcal{N} \to \mathbb{N}$.]

(b) Show that if $K \subseteq \mathcal{N}$ is compact, then $\mathcal{N} \setminus K$ is dense in $\mathcal{N}$.

(c) Conclude from the Baire Category Theorem that $\mathcal{N}$ is not $\sigma$-compact.

(d) Prove a similar result about an infinite product of copies of $\mathbb{R}$. Thus no infinite-dimensional Banach space is $\sigma$-compact in the norm topology. (Every infinite-dimensional separable Banach space is homeomorphic as a topological space to a product of a countable number of copies of $\mathbb{R}$ [].)

**XI.10.4.4.** If $(X, \rho)$ is a separable metric space, show that $X$ is homeomorphic to a subspace of the Hilbert cube $[0, 1]^\mathbb{N}$, as follows.

(a) Let $\sigma = \min(\rho, 1)$ be a $[0, 1]$-valued metric equivalent to $\rho$, and let $(y_n)$ be a dense sequence in $X$. Define $f_n : X \to [0, 1]$ by $f_n(x) = \sigma(x, y_n)$. Define $f : X \to [0, 1]^\mathbb{N}$ by $f(x) = (f_1(x), f_2(x), \ldots)$. Show that $f$ is continuous.

(b) Show that $f$ is a homeomorphism from $X$ onto $f(X)$. [If $f(x_n) \to f(x)$, for $\epsilon > 0$ choose $m$ so that $\sigma(x, y_m) < \epsilon/2$. Since $f_n(x_m) = \sigma(x, y_m) \to f_m(x) = \sigma(x, y_m) < \epsilon/2$, there is an $N$ such that $f_m(x_n) < \epsilon/2$ for $n \geq N$, so $\sigma(x_n, x) < \epsilon$ for $n \geq N$.]

1103
XI.10.4.5. Show that every metrizable space is homeomorphic to a closed subspace of a product of completely metrizable spaces. If $X$ is metrizable with metric $\rho$, let $\tilde{X}$ be the completion of $X$ with respect to $\rho$, and consider

$$\prod_{y \in X \setminus \{y\}} (\tilde{X} \setminus \{y\}).$$

XI.10.4.6. [Wil04, 23.2] Give a direct proof of XI.10.1.10. as follows. Let $X$ and $Y$ be compact Hausdorff spaces, and $f : X \to Y$ a surjective continuous function. Suppose $X$ is second countable. Let $\mathcal{B}$ be a countable base for the topology of $X$ which is closed under finite unions. Set

$$D = \{Y \setminus f(X \setminus B) : B \in \mathcal{B}\}.$$ 

Show that $D$ is a countable base for the topology of $Y$.

XI.10.4.7. The Hausdorff Metric. Let $(X, \rho)$ be a metric space, and let $\mathcal{K}(X)$ be the set of nonempty compact subsets of $X$. For $x \in X$ and $F \in \mathcal{K}(X)$, set

$$\rho(x, F) = \inf_{y \in F} \rho(x, y) = \min_{y \in F} \rho(x, y).$$

For $E, F \in \mathcal{K}(X)$, set

$$\rho_1(E, F) = \sup_{x \in E} \rho(x, F) = \max_{x \in E} \rho(x, F)$$

and

$$\rho_h(E, F) = \max(\rho_1(E, F), \rho_1(F, E)).$$

This $\rho_h$ is called the Hausdorff metric (or Hausdorff distance) on $\mathcal{K}(X)$. (It can be defined more generally on the space of closed bounded subsets of $X$, but not all of the same properties hold in this space. Such spaces of subsets are often called hyperspaces.)

(a) Show that $\rho_h$ is a metric on $\mathcal{K}(X)$.

(b) Show that if $\rho$ is complete, then $\rho_h$ is also complete. [It suffices to show that every $\rho_h$-Cauchy sequence has a convergent subsequence. Let $(E_n)$ be a sequence in $\mathcal{K}(X)$ with $\rho_h(E_n, E_{n+1}) < 2^{-n}$. If

$$D = \{x \in X : \exists x_n \in E_n \text{ with } x_n \to x \text{ and } \rho(x_n, x_{n+1}) < 2^{-n}\}$$

then $E = \overline{D}$ is nonempty and compact (totally bounded), and $E_n \to E$.]

(c) Show that if $(X, \rho)$ is separable, then $(\mathcal{K}(X), \rho_h)$ is also separable. [Consider finite subsets of a countable dense set in $X$.]

(d) Show that if $(X, \rho)$ is totally bounded, so is $(\mathcal{K}(X), \rho_h)$. [If $D$ is a $\delta$-cover of $X$, then $\mathcal{P}(D)$ is a $\delta$-cover of $\mathcal{K}(X)$.] Conclude from (b) and (c) that if $(X, \rho)$ is compact, so is $(\mathcal{K}(X), \rho_h)$.

(e) The map $x \mapsto \{x\}$ is an isometry from $(X, \rho)$ onto a closed subset of $(\mathcal{K}(X), \rho_h)$. Deduce the converses of the statements in (b), (c), and (d).

(f) If $E, E_1, E_2, \cdots \in \mathcal{K}(X)$ and $E_n \to E$, then $\cap_{n} E_n \subseteq E \subseteq \cup_{n} E_n$. If $Y$ is a closed subset of $X$, then $\mathcal{K}(Y)$ is a closed subset of $\mathcal{K}(X)$. If $Y$ is compact and $\mathcal{K}(X)_Y = \{E \in \mathcal{K}(X) : Y \subseteq E\}$, then $\mathcal{K}(X)_Y$ is also a closed subset of $\mathcal{K}(X)$. In particular, if $p \in X$, then

$$\mathcal{K}(X)_p = \mathcal{K}(X)_{\{p\}} = \{E \in \mathcal{K}(X) : p \in E\}$$

1104
is a closed subset of \( \mathcal{K}(X) \).

(g) Let \( X = \{ \frac{1}{n} : n \in \mathbb{N} \} \cup \{ 0 \} \subseteq [0,1] \) with the metric \( \rho(x,y) = |x-y| \). Show that \( \mathcal{K}(X)_0 \) is homeomorphic to the Cantor set. [A compact subset of \( X \) containing 0 is of the form \( E_S = \{ \frac{1}{n} : n \in S \} \cup \{ 0 \} \) for a subset \( S \) of \( \mathbb{N} \). Send \( E_S \) to the usual sequence in \( \{0,1\}^\mathbb{N} \) associated to \( S \) as in (i).] The complement of \( \mathcal{K}(X)_0 \) in \( \mathcal{K}(X) \) is countable. Describe the topological space \( \mathcal{K}(X) \).

(h) \( \mathcal{K}([0,1]) \) is topologically infinite-dimensional. [If \( n \in \mathbb{N} \) and \( x_1, \ldots, x_n \) are distinct numbers in \( (0,1) \), then there is a neighborhood \( U \) of \( (x_1, \ldots, x_n) \) in \( \mathbb{R}^n \) such that \( (y_1, \ldots, y_n) \mapsto \{y_1, \ldots, y_n\} \) is a homeomorphism from \( U \) into \( \mathcal{K}([0,1]) \).] In fact, \( \mathcal{K}([0,1]) \) is homeomorphic to the Hilbert cube \([?)\), as is \( \mathcal{K}(X) \) for any compact, connected, locally connected metric space \( X \) [CS78].

(i) Show that \( (\mathcal{K}(X), \rho_h) \) is connected [resp. totally disconnected] if and only if \( (X, \rho) \) is connected [resp. totally disconnected].

### XI.10.4.8. The Hausdorff Metric Continued

[ Mic51 ] Use the notation of Exercise XI.10.4.7. Let \( (X, \rho) \) be a metric space.

(a) Show that if \( E, E_1, E_2, \cdots \in \mathcal{K}(X) \), then \( E_n \to E \) if and only if, for every finite set \( \{U_1, \ldots, U_n\} \) of open subsets of \( X \) with \( E \subseteq \bigcup_k U_k \) and \( E \cap U_k \neq \emptyset \) for all \( k \), we have, for all sufficiently large \( n \), \( E_n \subseteq \bigcup_k U_k \) and \( E_n \cap U_k \neq \emptyset \) for all \( k \). (A topology can be defined on \( \mathcal{K}(X) \) in this way for any topological space \( X \); this topology is called the finite topology or Vietoris topology on \( \mathcal{K}(X) \).)

(b) Let \( (Y, \sigma) \) be another metric space, and \( f : X \to Y \) a continuous function. Define \( f_* : \mathcal{K}(X) \to \mathcal{K}(Y) \) by setting \( f_*(E) = f(E) \) for \( E \in \mathcal{K}(X) \) (recall (i) that \( f(E) \) is compact). Show that \( f_* : (\mathcal{K}(X), \rho_h) \to (\mathcal{K}(Y), \sigma_h) \) is continuous.

(c) Show that if \( \sigma \) is another metric on \( X \) which is equivalent to \( \rho \), then \( \sigma_h \) is equivalent to \( \rho_h \). So the topology on \( \mathcal{K}(X) \) depends only on the topology of \( X \), not on the choice of metric.

(d) Show that if \( E_n \to E \) in \( \mathcal{K}(X) \), then \( F = E \cup \bigcup_{n=1}^\infty E_n \) is compact in \( X \). [If \( \mathcal{U} \) is an open cover of \( F \), let \( \{U_1, \ldots, U_m\} \) be a finite subcover of \( E \). Then \( E_n \subseteq \bigcup_{k=1}^m U_k \) for all but finitely many \( n \).] More generally, if \( \mathcal{E} \) is an open cover of \( \mathcal{K}(X) \), then \( \bigcup_{E \in \mathcal{E}} E \) is an open cover of \( \mathcal{K}(X) \) [Mic51, 2.5.1].

### XI.10.4.9. Let \( \mathcal{P} \) be the set of homeomorphism classes of Polish spaces, and \( \kappa = \text{card}(\mathcal{P}) \). Show that the following weak version of XI.10.3.7. is the best which can be proved in ZF [recall (II.9.7.9.) that it cannot be proved in ZF that \( \aleph_1 \) and \( 2^{\aleph_0} \) are comparable.]

**THEOREM.**

(i) \( \kappa \geq \aleph_1 + 2^{\aleph_0} \).

(ii) \( \kappa \lesssim 2^{\aleph_0} \) (II.9.4.2.).

(iii) \( \kappa < 2^{2^{\aleph_0}} \).

The arguments are the ones from the proof of XI.10.3.7. For (i), note that by XI.18.4.2. there are also \( \aleph_1 \) pairwise nonhomeomorphic compact subsets of \( \mathbb{R} \), none of which are homeomorphic to the ones of XI.18.4.1.. For (iii), use II.9.7.7. and II.5.3.3..
XI.10.4.10. Prove the following theorem:

**Theorem.** Let $X$ be a (necessarily $T_1$) topological space. The following are equivalent:

(i) $X$ is completely normal (XI.10.1.2.).

(ii) Every open subspace of $X$ is normal.

(iii) Every subspace of $X$ is normal.

(iv) Every subspace of $X$ is completely normal.

(a) Show that (i) $\Rightarrow$ (iv).

(b) For (ii) $\Rightarrow$ (i), suppose $X$ is normal, $A, B \subseteq X$ with $A \cap \overline{B} = \overline{A} \cap B = \emptyset$ and $W = X \setminus (\overline{A} \cap \overline{B})$ normal. Show $A$ and $B$ have disjoint neighborhoods in $W$, hence in $X$.

(c) A topological space $X$ is **perfectly normal** if it is normal and every closed set in $X$ is a $G_\delta$ (equivalently, every closed set in $X$ is a zero set $\emptyset$), or that every open set in $X$ is an $F_\sigma$. Use the theorem and XI.7.8.7.(b) to show that every perfectly normal space is completely normal. Show that every metrizable space is perfectly normal. There are completely normal spaces which are not perfectly normal, e.g. the one-point compactification of an uncountable discrete space. A space which is perfectly normal is sometimes called $T_6$. 

1106
XI.11. Compactness

Compactness is one of the most important notions in topology, and especially for its applications in analysis. Compact sets behave roughly like finite sets. Compactness is used in applications to insure boundedness, the existence of limits, and the achievement of extrema, among other things.

Use of the Axiom of Choice is pervasive in the theory of compactness, and we will freely employ it; we will usually point out which results depend on it and which do not. Use of the Axiom of Choice can be almost entirely avoided if one sticks only to second countable topological spaces.

XI.11.1. Basic Definitions

XI.11.1.1. Definition. Let \((X, T)\) be a topological space. An open cover (\(T\)-open cover) is a set \(U = \{U_i : i \in I\}\) of \(T\)-open sets with \(\cup_{i \in I} U_i = X\). The open cover \(U\) is finite if \(I\) is finite. A subset \(V\) of an open cover \(U\) is a subcover if it is itself an open cover. If \(Y \subseteq X\) and \(U = \{U_i : i \in I\} \subseteq T\), then \(U\) is an open cover of \(Y\) in \(X\) if \(Y \subseteq \cup_{i \in I} U_i\). Subcovers of \(Y\) in \(X\) are defined similarly.

XI.11.1.2. Definition. Let \((X, T)\) be a topological space. \((X, T)\) is compact if every open cover of \(X\) has a finite subcover. If \(Y \subseteq X\), then \(Y\) is a compact subspace of \(X\) if every open cover of \(Y\) in \(X\) has a finite subcover in \(X\).

XI.11.1.3. Caution: In some older references, particularly from Eastern Europe, “compact” is used to mean “countably compact” (XI.11.7.1.). A space which is compact according to the modern definition XI.11.1.2. is called “bicom pact” in these references. Also, in some references, especially French ones and [Eng89], the term “compact” is reserved for Hausdorff spaces; a space which is compact in the sense of XI.11.1.2. but not necessarily Hausdorff is called “quasi-compact.” In some references like [Pea75], “bicom pact” means “compact and Hausdorff.” Check the definitions in any reference.

The first observation is that compactness of a subspace depends only on the relative topology of the subspace, not on the way it is embedded:

XI.11.1.4. Proposition. If \(Y\) is a subspace of a topological space \((X, T)\), then \(Y\) is a compact subspace of \(X\) if and only if \((Y, T_Y)\) is compact.

Proof: This is obvious from the fact that every relatively open subset of \(Y\) is of the form \(U \cap Y\), where \(U \in T\).

The next result, an immediate corollary of XI.11.2.2. and the Bolzano-Weierstrass theorem, gives the most important basic examples of compact spaces, and begins to give a picture of the meaning of compactness.

XI.11.1.5. Theorem. A subset of \(\mathbb{R}^n\) is compact if and only if it is closed and bounded.

Here is a far-reaching version (not quite a generalization) of XI.11.1.5.. Because of this, compact sets are often thought of as sets which are “absolutely closed” (although this is only really accurate in Hausdorff spaces). See XI.11.4.1. and XI.11.4.2. for generalizations of the second statement.
XI.11.1.6. Proposition. A closed subspace of a compact space is compact. A compact subspace of a Hausdorff space is closed.

Proof: Let \((X, \mathcal{T})\) be a compact space, and \(Y\) a closed subset of \(X\). Let \(\mathcal{U}\) be an open cover of \(Y\) in \(X\). Then \(\mathcal{V} = \mathcal{U} \cup \{X \setminus Y\}\) is an open cover of \(X\). Let \(\{U_1, \ldots, U_n, X \setminus Y\}\) be a finite subcover of \(\mathcal{V}\); then \(\{U_1, \ldots, U_n\}\) is a finite subcover of the cover \(\mathcal{U}\) of \(Y\).

If \((X, \mathcal{T})\) is a Hausdorff space and \(Y\) is a compact subspace of \(X\), and \(x \in X \setminus Y\), for each \(y \in Y\) let \(U_y\) and \(V_y\) be disjoint open neighborhoods (in \(X\)) of \(x\) and \(y\) respectively. Then \(\{V_y : y \in Y\}\) is an open cover of \(Y\) in \(X\). Let \(\{V_{y_1}, \ldots, V_{y_n}\}\) be a finite subcover. Then \(U = U_{y_1} \cap \cdots \cap U_{y_n}\) is an open neighborhood of \(x\) disjoint from \(Y\). Thus \(X \setminus Y\) is open.

Note that the Axiom of Choice is used in the second half of the proof. The result is true, however, without any Choice assumption (Exercise XI.11.12.2.).

A compact subspace of a general topological space, even a \(T_1\)-space, is not necessarily closed (Exercise XI.11.12.1.).

A continuous image of a closed set is not necessarily closed. But one of the important properties of compactness is that it is preserved under continuous images:

XI.11.1.7. Proposition. Let \((X, \mathcal{T})\) and \((Z, S)\) be topological spaces, and \(f : X \to Z\) a continuous function. If \(Y\) is a compact subset of \(X\), then \(f(Y)\) is a compact subset of \(Z\).

Proof: Let \(\mathcal{V} = \{V_i : i \in I\}\) be an open cover of \(f(Y)\) in \(Z\). For each \(i\), let \(U_i = f^{-1}(V_i)\). Then \(\mathcal{U} = \{U_i : i \in I\}\) is an open cover of \(Y\) in \(X\). If \(\{U_{i_1}, \ldots, U_{i_n}\}\) is a finite subcover of \(Y\), then \(\{V_{i_1}, \ldots, V_{i_n}\}\) is a finite subcover of \(f(Y)\).

An important corollary generalizes the Extreme Value Theorem of calculus:

XI.11.1.8. Corollary. Let \(X\) be a topological space, \(f : X \to \mathbb{R}\) a continuous function, and \(E\) a compact subset of \(X\). Then \(f\) is bounded on \(E\), and attains its maximum and minimum values on \(E\).

Proof: By XI.11.1.7., \(f(E)\) is compact in \(\mathbb{R}\), and therefore closed and bounded by XI.11.1.5..

There is a useful rephrasing of the definition of compactness, using the Finite Intersection Property. This characterization of compactness is strongly reminiscent of the Nested Intervals Property of \(\mathbb{R}\).

XI.11.1.9. Definition. Let \((X, \mathcal{T})\) be a topological space. A (nonempty) collection \(\mathcal{F} = \{F_i : i \in I\}\) of closed subsets of \(X\) has the Finite Intersection Property if, whenever \(F_{i_1}, \ldots, F_{i_n}\) is a finite collection of sets in \(\mathcal{F}\), we have \(F_{i_1} \cap \cdots \cap F_{i_n} \neq \emptyset\).
XI.11.1.10. **Proposition.** Let \((X, \mathcal{T})\) be a topological space. Then \((X, \mathcal{T})\) is compact if and only if,
whenever \(\mathcal{F} = \{F_i : i \in I\}\) is a nonempty collection of closed sets in \(X\) with the Finite Intersection Property,
we have \(\bigcap_{i \in I} F_i \neq \emptyset\).

**Proof:** Suppose \((X, \mathcal{T})\) is compact, and let \(\mathcal{F} = \{F_i : i \in I\}\) be a collection of closed sets in \(X\) with \(\bigcap_{i \in I} F_i = \emptyset\). Let \(U_i = X \setminus F_i\) for each \(i\). Then \(\mathcal{U} = \{U_i : i \in I\}\) is an open cover of \(X\). Let \(\{U_{i_1}, \ldots, U_{i_n}\}\) be a finite subcover. Then \(F_{i_1} \cap \cdots \cap F_{i_n} = \emptyset\), so \(\mathcal{F}\) does not have the Finite Intersection Property.

Conversely, suppose every collection of closed sets in \(X\) with the Finite Intersection Property has a nonempty intersection, and let \(\mathcal{U} = \{U_i : i \in I\}\) be an open cover of \(X\). Set \(F_i = X \setminus U_i\) for each \(i\). Then \(\mathcal{F} = \{F_i : i \in I\}\) is a collection of closed subsets of \(X\) with \(\bigcap_{i \in I} F_i = \emptyset\), so \(\mathcal{F}\) does not have the Finite Intersection Property, i.e. there is a finite set \(\{F_{i_1}, \ldots, F_{i_n}\}\) with \(F_{i_1} \cap \cdots \cap F_{i_n} = \emptyset\). But then \(\{U_{i_1}, \ldots, U_{i_n}\}\) is a finite subcover of \(\mathcal{U}\).

Here is a useful variation:

XI.11.1.11. **Proposition.** Let \(X\) be a topological space, \(\{E_i : i \in I\}\) a collection of closed subsets of \(X\), with at least one compact. If \(U\) is an open subset of \(X\) containing \(\bigcap_{i \in I} E_i\), then there is a finite subset \(F\) of \(I\) such that \(\bigcap_{i \in F} E_i \subseteq U\).

**Proof:** Fix an \(i_0\) with \(E_{i_0}\) compact. For each \(i\), let \(K_i = E_i \cap E_{i_0} \cap U^c\). Then the \(K_i\) are closed subsets of \(E_{i_0}\), and they have empty intersection, so they do not have the Finite Intersection Property, i.e. there are \(i_1, \ldots, i_n\) such that \(K_{i_1} \cap \cdots \cap K_{i_n} = \emptyset\). Then \(E_{i_0} \cap E_{i_1} \cap \cdots \cap E_{i_n} \subseteq U\).

XI.11.1.12. **Corollary.** Let \(X\) be a topological space. Let \((F_n)\) be a decreasing sequence of closed compact subsets of \(X\) ("closed" is redundant if \(X\) is Hausdorff) with \(F = \bigcap_n F_n\). If \(U\) is any open set in \(X\) with \(F \subseteq U\), then \(F_n \subseteq U\) for all sufficiently large \(n\).

Here is an interesting and important special case:

XI.11.1.13. **Corollary.** Let \(X\) be a topological space, \(p \in X\). Let \((V_n)\) be a decreasing sequence of closed compact neighborhoods of \(p\) ("closed" is redundant if \(X\) is Hausdorff) with \(\bigcap_n V_n = \{p\}\). Then the \(V_n\) form a local neighborhood base at \(p\) for the topology.

XI.11.2. Compactness and Convergence

The next theorem gives the main equivalent ways compactness is characterized:
XI.11.2.1. Theorem. Let \((X, \mathcal{T})\) be a topological space. The following are equivalent:

(i) \((X, \mathcal{T})\) is compact.

(ii) Every nonempty collection of closed subsets of \(X\) with the Finite Intersection Property has nonempty intersection.

(iii) Every net in \(X\) has a cluster point.

(iv) Every net in \(X\) has a convergent subnet.

(v) Every filter on \(X\) has a cluster point.

Proof: (i) \(\iff\) (ii) is XI.11.1.10., (iii) \(\iff\) (iv) is (i), and (iii) \(\iff\) (v) is (i).

(ii) \(\Rightarrow\) (iii): Let \((x_i)_{i \in I}\) be a net in \(X\). For each \(i\), let \(F_i\) be the closure of \(\{x_j : j \geq i\}\), and let \(\mathcal{F} = \{F_i : i \in I\}\). If \(i_1, \ldots, i_n \in I\), then since \(I\) is directed there is an \(i_0 \geq i_k\) for \(1 \leq k \leq n\). Thus \(x_{i_0} \in F_{i_1} \cap \cdots \cap F_{i_n}\), and hence \(F_{i_1} \cap \cdots \cap F_{i_n} \neq \emptyset\) and so \(\mathcal{F}\) has the Finite Intersection Property. By (ii), there is an \(x \in \cap_{i \in I} F_i\), and \(x\) is a cluster point of the net \((x_i)\).

(iii) \(\Rightarrow\) (ii): Let \(\mathcal{F}\) be a nonempty collection of closed sets with the Finite Intersection Property. We may assume that \(\mathcal{F}\) is closed under finite intersections. Set

\[
I = \{(x, F) \in X \times \mathcal{F} : x \in F\}.
\]

Put a preorder on \(I\) by setting \((x_1, F_1) \leq (x_2, F_2)\) if \(F_2 \subseteq F_1\). Then \(I\) is directed: for any \((x_1, F_1), (x_2, F_2)\) we have

\[
(x_1, F_1) \leq (x_3, F_1 \cap F_2)
\]

where \(x_3\) is any element of \(F_1 \cap F_2\). Now, for \(i = (x, F) \in I\), define \(x_i = x\). Then \((x_i)\) is a net in \(X\). By (iii) this net has a cluster point \(y\). If \(U\) is any neighborhood of \(y\), and \(F_0 \in \mathcal{F}\), and \(x_0 \in F_0\), then there is an \(i = (x, F)\) with \((x_0, F_0) \leq i\) (i.e. \(F \subseteq F_0\)) and \(x_i = x \in U\). Since \(x \in F \subseteq F_0\), we have that \(U \cap F_0 \neq \emptyset\). Fixing \(F_0\) and letting \(U\) vary, we conclude that \(y\) is a limit point of \(F_0\), hence \(y \in F_0\) since \(F_0\) is closed. This is true for all \(F_0 \in \mathcal{F}\), so \(y \in \cap_{F \in \mathcal{F}} F\) and \(\mathcal{F}\) has nonempty intersection. \(\Box\)

Note that no form of Choice is needed for this result.

XI.11.2.2. Corollary. Let \(X\) be a second countable topological space. Then \(X\) is compact if and only if every sequence in \(X\) has a convergent subsequence (a topological space with this property is called sequentially compact).

This result is false in general for spaces which are just first countable, even for normal spaces: the space \(\omega_1\) of ordinals less than the first uncountable ordinal, with the order topology \(\omega_1\), is a normal Hausdorff space which is first countable and sequentially compact, but not compact. A first countable topological space which is compact is sequentially compact, but \(\beta\mathbb{N}\) (XI.11.9.16.), which is not first countable, is compact but not sequentially compact. See XI.11.7.3.

There is a variation of XI.11.2.1.(v) which is useful. One direction of this result uses the Ultrafilter Property \(\omega_1\), which is equivalent to the Boolean Prime Ideal Theorem \(\omega_1\).
XI.11.2.3. Proposition. (BPI) Let \((X, T)\) be a topological space. Then \((X, T)\) is compact if and only if every ultrafilter on \(X\) converges.

Proof: If \(X\) is compact, then by XI.11.2.1. every ultrafilter on \(X\) has a cluster point, hence converges. Conversely, suppose every ultrafilter on \(X\) converges. Let \(F\) be a filter on \(X\). Then by the Ultrafilter Property there is an ultrafilter \(G\) on \(X\) containing \(F\), which converges by assumption. The limit of \(G\) is a cluster point for \(F\), so \(X\) is compact by XI.11.2.1.

XI.11.3. Compact Metrizable Spaces

There is a very clean characterization of compact metric spaces, using the idea of total boundedness:

XI.11.3.1. Definition. Let \(\rho\) be a metric on a set \(X\). Then \(\rho\) is totally bounded if for every \(\epsilon > 0\), a finite number of open \(\rho\)-balls of radius \(\epsilon\) cover \(X\).

In other words, \(\rho\) is totally bounded if for every \(\epsilon > 0\), there is a finite set of points \(\{x_1, \ldots, x_n\}\) such that every point of \(x\) is within \(\epsilon\) of at least one of the \(x_k\).

A totally bounded metric is bounded: taking \(\epsilon = 1\), if \(\{x_1, \ldots, x_n\}\) are the centers of balls of radius 1 covering \(X\), and \(M\) is the maximum of the distances between the \(x_k\)'s, then any two points of \(X\) are distance less than \(M + 2\) apart. But a bounded metric is not necessarily totally bounded: for example, the discrete metric on an infinite set is bounded but not totally bounded.

XI.11.3.2. Proposition. A totally bounded metric space is separable (and hence second countable by ()).

Proof: For each \(m \in \mathbb{N}\) let \(S_m = \{x_{m,1}, \ldots, x_{m,n_m}\}\) be a finite set of points of \(X\) such that the open balls of radius \(\frac{1}{m}\) centered at the \(x_{m,k}\) cover \(X\). If \(S = \bigcup_m S_m\), then \(S\) is a countable dense subset of \(X\).

XI.11.3.3. Proposition. Let \((X, \rho)\) be a totally bounded metric space. Then every sequence in \(X\) has a Cauchy subsequence.

Proof: For each \(m\), choose a finite set \(\{y_{k,m}\}\) in \(X\) such that every element of \(X\) is within \(\frac{1}{2m}\) of one of the \(y_{k,m}\).

Let \((x_n)\) be a sequence in \(X\). Since there are only finitely many \(y_{k,1}\), there are infinitely many \(x_k\) within \(\frac{1}{2}\) of some single \(y_{k,1}\). Thus there is a subsequence \((x_{n,1})\) of \((x_n)\) all of whose terms lie within \(\frac{1}{2}\) of this \(y_{k,1}\), hence within 1 of each other. Successively let \((x_{n,m+1})\) be a subsequence of \((x_{n,m})\), all of whose terms lie within \(\frac{1}{m+1}\) of some \(y_{k,m+1}\) and hence within \(\frac{1}{m}\) of each other. Then the diagonal sequence \((x_{n,n})\) is a subsequence of \((x_n)\) which is a Cauchy sequence.

The Countable AC is needed in this proof to choose the \(y_{k,m}\), and DC is needed to choose the subsequences.

The main characterization theorem is:
XI.11.3.4. **Theorem.** Let \((X, \rho)\) be a metric space. Then \(X\) is compact (in the metric topology) if and only if \(\rho\) is complete and totally bounded.

**Proof:** Suppose \(X\) is compact. Since \(X\) is first countable, every sequence in \(X\) has a convergent subsequence. In particular, if \((x_n)\) is a Cauchy sequence, it must have a convergent subsequence, hence must converge ( ). Thus \((X, \rho)\) is complete. If \(\rho\) is not totally bounded, there is an \(\epsilon > 0\) such that no finite set of open balls of radius \(\epsilon\) covers \(X\), so a sequence \((x_n)\) can be chosen so that \(\rho(x_n, x_m) \geq \epsilon\) for all \(n\) and \(m\). But such a sequence can have no convergent subsequence, a contradiction. Thus \(\rho\) is totally bounded.

Conversely, suppose \(\rho\) is complete and totally bounded. Suppose \(\mathcal{U} = \{U_i : i \in I\}\) is an open cover of \(X\). Since \(X\) is second countable (XI.11.3.2.), it is Lindelöf ( ), so \(\mathcal{U}\) has a countable subcover \(\{U_1, U_2, \ldots\}\). Suppose this cover does not have a finite subcover. For each \(n\), let \(x_n \in X \setminus \bigcup_{k=1}^{n} U_k\). By XI.11.3.3., \((x_n)\) has a Cauchy subsequence, which must converge to some \(x\) since \((X, \rho)\) is complete. We have \(x \in U_n\) for some \(n\), but \(U_n\) is open and \(x_k \notin U_n\) for all \(k > n\), so no subsequence of \((x_n)\) can converge to \(x\), a contradiction. Thus there is a finite subcover, which is a subcover of \(\mathcal{U}\). Thus \(X\) is compact.

Combining the previous results, we get:

XI.11.3.5. **Corollary.** Let \(X\) be a compact Hausdorff space. Then \(X\) is metrizable if and only if it is second countable.

**Proof:** If \(X\) is metrizable, it is second countable by XI.11.3.4. and XI.11.3.2. Conversely, since a compact Hausdorff space is regular, if it is second countable it is metrizable by the Urysohn Metrization Theorem ( ).

XI.11.3.6. There are compact Hausdorff spaces which are both separable and first countable, but not second countable and hence not metrizable (e.g. XI.11.12.4.).

**Lebesgue Numbers of an Open Cover**

The Lebesgue Property and Lebesgue numbers of an open cover are particularly relevant for compact metric spaces, but can be defined in general:

**XI.11.3.7. Definition.** Let \(\mathcal{U}\) be an open cover of a metric space \((X, \rho)\). \(\mathcal{U}\) has the **Lebesgue Property** if there is a \(\delta > 0\) such that, for every \(x \in X\), there is a \(U \in \mathcal{U}\) with \(B_\delta(x) \subseteq U\). Such a \(\delta\) is a **Lebesgue number** of the open cover \(\mathcal{U}\).

Any positive number smaller than a Lebesgue number for \(\mathcal{U}\) is a Lebesgue number for \(\mathcal{U}\). The supremum of all Lebesgue numbers for \(\mathcal{U}\) can be \(\infty\), and if finite may or may not be a Lebesgue number for \(\mathcal{U}\); thus \(\mathcal{U}\) need not have a largest Lebesgue number. If \(\mathcal{V}\) is a refinement of \(\mathcal{U}\) (in particular, if it is a subcover of \(\mathcal{U}\)), any Lebesgue number of \(\mathcal{V}\) is a Lebesgue number of \(\mathcal{U}\). The Lebesgue Property and Lebesgue numbers depend on the metric and are not topological properties.
XI.11.3.8. Examples. (i) In $\mathbb{R}$ with the usual metric, the open cover $\mathcal{U} = \{-n, n : n \in \mathbb{N}\}$ has the Lebesgue Property, and any positive number is a Lebesgue number for $\mathcal{U}$. 

(ii) In $\mathbb{R}$ with the usual metric, the open cover $\mathcal{U}$ consisting of all open intervals of length less than 2 has the Lebesgue Property, and any positive number less than 1 is a Lebesgue number for $\mathcal{U}$. The supremum of the Lebesgue numbers is 1, but 1 is not a Lebesgue number for $\mathcal{U}$.

(iii) In $(0, 1)$ with the usual metric, the open cover $\{(\frac{1}{n}, 1) : n \in \mathbb{N}\}$ does not have the Lebesgue Property.

(iv) Even a finite open cover need not have the Lebesgue Property. Let $X = (0, 1)$ with the usual metric. Fix two disjoint decreasing sequences $(x_n)$ and $(y_n)$ in $(0, 1)$ converging to 0, and set $U = (0, 1) \setminus \{x_n : n \in \mathbb{N}\}$ and $V = (0, 1) \setminus \{y_n : n \in \mathbb{N}\}$. Then the open cover $\{U, V\}$ does not have the Lebesgue Property.

XI.11.3.9. Proposition. Let $(X, \rho)$ be a metric space of bounded diameter $d$. If $\mathcal{U}$ is an open cover of $X$ by proper subsets, and $\mathcal{U}$ has the Lebesgue property, then any Lebesgue number for $\mathcal{U}$ is $\leq d$. If $\mathcal{U}$ is a finite open cover of $X$ by proper subsets, and $\mathcal{U}$ has the Lebesgue Property, then the supremum of all Lebesgue numbers for $\mathcal{U}$ is a Lebesgue number for $\mathcal{U}$, i.e. $\mathcal{U}$ has a largest Lebesgue number.

Proof: For any $x \in X$ and $U \in \mathcal{U}$, there is a $y \in X$ with $y \notin U$. Since $\rho(x, y) \leq d$, any Lebesgue number of $\mathcal{U}$ must be $\leq d$. If $\mathcal{U} = \{U_1, \ldots, U_n\}$ is a finite open cover by proper subsets, and $\delta$ is the supremum of the Lebesgue numbers of $\mathcal{U}$ ($\delta < \infty$ by the first part of the proof), fix $x \in X$. Then there must be a $k$ such that $B_{\delta - 1/m}(x) \subseteq U_k$ for infinitely many $m$; hence $B_{\delta}(x) \subseteq U_k$. So $\delta$ is a Lebesgue number for $\mathcal{U}$. 

The main result is:

XI.11.3.10. Proposition. Let $(X, \rho)$ be a compact metric space. Then any open cover of $X$ has the Lebesgue Property.

Proof: Let $\mathcal{U}$ be an open cover of $X$. Suppose $\mathcal{U}$ does not have the Lebesgue property. Then for each $n$ there are points $x_n, y_n \in X$ with $\rho(x_n, y_n) < \frac{1}{n}$, but for which there is no $U \in \mathcal{U}$ containing both $x_n$ and $y_n$. Passing to a subsequence if necessary, we may assume $(x_n)$ converges to a point $x \in X$. Then also $y_n \to x$. There is a $U \in \mathcal{U}$ with $x \in U$. Then $x_n$ and $y_n$ are both in $U$ if $n$ is sufficiently large, a contradiction.

See XI.11.12.3. for an alternate proof which gives an explicit Lebesgue number.

Here is a sample application:

XI.11.3.11. Corollary. Let $(X, \rho)$ and $(Y, \sigma)$ be metric spaces, with $X$ compact, and $F : X \to Y$ a continuous function. Then $f$ is uniformly continuous.

Proof: Let $\epsilon > 0$. For each $y \in Y$, let $U_y = f^{-1}(B_{\epsilon/2}(y))$. Then $\{U_y : y \in Y\}$ is an open cover of $X$. If $\delta > 0$ is the Lebesgue number of this open cover, and $x_1, x_2 \in X$ with $\rho(x_1, x_2) < \delta$, there is a $y$ with $B_\delta(x_1) \subseteq U_y$; thus $f(x_1)$ and $f(x_2)$ are both in $B_{\epsilon/2}(y)$ and $\sigma(f(x_1), f(x_2)) < \epsilon$. 

1113
XII.11.4. Separation Properties

The results of this subsection give important senses in which compact sets behave like finite sets (see XII.11.10.17. for another example). The first result is a strengthening of the second half of XII.11.1.6., with essentially the same proof:

XII.11.4.1. Proposition. Let \( X \) be a Hausdorff space, \( E \) a compact subset of \( X \), and \( y \) a point of \( X \setminus E \). Then \( E \) and \( y \) have disjoint open neighborhoods.

Proof: For each \( x \in E \), let \( U_x \) and \( V_x \) be disjoint open neighborhoods of \( x \) and \( y \) in \( X \). Then \( \{U_x : x \in E\} \) is an open cover of \( E \), so has a finite subcover \( \{U_{x_1}, \ldots, U_{x_n}\} \). Then \( U = U_{x_1} \cup \cdots \cup U_{x_n} \) and \( V = V_{x_1} \cap \cdots \cap V_{x_n} \) are disjoint open neighborhoods of \( E \) and \( y \).

There is a nice generalization:

XII.11.4.2. Proposition. Let \( X \) be a Hausdorff space, and \( E \) and \( F \) disjoint compact subsets of \( X \). Then \( E \) and \( F \) have disjoint open neighborhoods.

Proof: For each \( y \in F \), by XII.11.4.1. let \( U_y \) and \( V_y \) be disjoint open neighborhoods of \( E \) and \( y \). Then \( \{V_y : y \in F\} \) is an open cover of \( F \); let \( \{V_{y_1}, \ldots, V_{y_n}\} \) be a finite subcover. Then \( U = U_{y_1} \cap \cdots \cap U_{y_n} \) and \( V_{y_1} \cup \cdots \cup V_{y_n} \) are disjoint open neighborhoods of \( E \) and \( F \).

XII.11.4.3. Corollary. A compact Hausdorff space is normal.

In an almost identical manner (Exercise ()), we can show:

XII.11.4.4. Proposition. Let \( X \) be a regular space, and \( E \) and \( F \) disjoint subsets of \( X \), with \( E \) compact and \( F \) closed. Then \( E \) and \( F \) have disjoint open neighborhoods.

Note the following consequence of XII.11.1.12.:

XII.11.4.5. Proposition. Let \( X \) be a compact Hausdorff space, and \( p \) a point of \( X \). The following are equivalent:

(i) \( \{p\} \) is a \( G_δ \), and whenever \( (F_n) \) is a decreasing sequence of closed neighborhoods of \( p \) with \( \cap_n F_n = \{p\} \), then \( \{F_n : n \in \mathbb{N}\} \) is a local base at \( p \) for the topology of \( X \).

(ii) There is a countable local base at \( p \) for the topology of \( X \).

(iii) \( \{p\} \) is a \( G_δ \).

Proof: (i) \( \Rightarrow \) (ii): Let \( (U_n) \) be a sequence of open sets in \( X \) with \( \cap_n U_n = \{p\} \). Set \( V_0 = X \) and inductively let \( V_n \) be an open neighborhood of \( p \) with \( V_n \subseteq U_n \cap V_{n-1} \) (using regularity of \( X \)). Then \( (V_n) \) is a decreasing sequence of closed neighborhoods of \( p \) with intersection \( \{p\} \), so they form a local base at \( p \).

(ii) \( \Rightarrow \) (iii) is trivial (since \( X \) is Hausdorff, the intersection of any local base at \( p \) is \( \{p\} \)), and (iii) \( \Rightarrow \) (i) is a special case of XII.11.1.12.
XI.11.5. Automatic Continuity and Comparison of Topologies

XI.11.6. Products of Compact Spaces and Tikhonov’s Theorem

Perhaps the most important theorem of general topology is Tikhonov’s Theorem (see XI.11.9.21. for some historical comments):

XI.11.6.1. THEOREM. [Tikhonov’s Theorem] Let \( \{(X_i, T_i) : i \in I\} \) be a collection of topological spaces, \( X = \prod_{i \in I} X_i \), and \( T \) the product topology on \( X \). If each \( (X_i, T_i) \) is compact, then \( (X, T) \) is compact.

This theorem is one of the two theorems of topology, along with the Baire Category Theorem, which are used most extensively in analysis.

Tikhonov’s Theorem can, among other things, be interpreted as giving compelling evidence that the product topology is the “right” topology on the Cartesian product.

XI.11.6.2. Tikhonov’s Theorem depends on the Axiom of Choice; in fact, it is equivalent to the Axiom of Choice (cf. Exercise ()). Important and frequently-used special cases, however, need only weaker versions of AC or no choice at all. We will prove some important versions under weaker choice assumptions.

There are several known proofs of Tikhonov’s Theorem; they are for the most part variations of the arguments we use in the various special cases. For a more complete discussion of these proofs and their history, see [Kel75], [Mun75], . . .

Finite Products

The simplest case is that of finite products, where no version of AC is needed:

XI.11.6.3. THEOREM. Let \( (X_1, T_1), \ldots, (X_n, T_n) \) be a finite number of topological spaces, and \( T \) the product topology on \( X = X_1 \times \cdots \times X_n \). If each \( (X_k, T_k) \) is compact, then \( (X, T) \) is compact.

PROOF: The argument is a slight generalization of the one in (). We prove the theorem in the case \( n = 2 \); this case can be applied successively, or the arguments in the proof repeated finitely many times, to give the general result for finite products. We will also change notation slightly for clarity: we begin with two compact spaces \( (X, T) \) and \( (Y, S) \), and show that \( X \times Y \) is compact in the product topology.

Suppose \( (z_i)_{i \in I} \) is a net in \( X \times Y \), with \( z_i = (x_i, y_i) \). Then \( (x_i) \) is a net in \( X \), and since \( X \) is compact there is a convergent subnet \( (x_{\alpha(j)})_{j \in J} \), where \( \alpha : J \to I \), with limit \( x \in X \). The net \( (y_{\alpha(j)}) \) is a net in \( Y \), and since \( Y \) is compact, so has a convergent subnet \( (y_{\alpha(\beta(k))})_{k \in K} \), where \( \beta : K \to J \), with limit \( y \in Y \). The subnet \( (x_{\alpha(\beta(k))})_{k \in K} \) of \( (x_{\alpha(j)}) \) also converges to \( x \), and hence the subnet \( (z_{\alpha(\beta(k))}) \) of \( (z_i) \) converges to \( z = (x, y) \) by ().

XI.11.6.4. This proof can be modified to show that a countable product of compact spaces is compact, but there is a logical complication and one additional step needed. The additional step is to construct a “diagonal” subnet after a sequence of subnets has been chosen; see Exercise () for details. The logical complication is that an infinite sequence of subnets must be chosen, each depending on the previous one; for this, the Axiom of Dependent Choice is needed. Using another argument (), one can get away with the weaker Countable Axiom of Choice for this result (in fact it is equivalent to Countable AC).
The General Tikhonov Theorem

Next we prove the full Tikhonov Theorem (XI.11.6.1.). This proof is due to P. Chernoff [Che92]; cf. [Fol99, ]. At one point we use an argument similar to the one in the proof of XI.11.6.3.. This proof uses the full Axiom of Choice (in the form of Zorn’s Lemma), as it must.

XI.11.6.5. **Proof:** Let \( (x_\alpha)_{\alpha \in A} \) be a net in \( X = \prod_{i \in I} X_i \). We need to show that \( (x_\alpha) \) has a cluster point. If \( J_1 \subseteq J_2 \subseteq I \), write \( \pi_{J_i} \) for the projection from \( \prod_{i \in J_2} X_i \) onto \( \prod_{i \in J_1} X_i \) (the notation avoids mention of \( J_2 \) to avoid unnecessary complication). Consider the set of all pairs \( (J, y) \), where \( J \) is a nonempty subset of \( I \) and \( y \in \prod_{i \in J} X_i \) is a cluster point for the net \( (\pi_J(x_\alpha))_{\alpha \in A} \). The collection \( S \) of such pairs is nonempty since there is such a \( y \) whenever \( J \) is a singleton by the assumption that each \( X_i \) is compact. Put a partial ordering on \( S \) by setting \( (J_1, y_1) \leq (J_2, y_2) \) if \( J_1 \subseteq J_2 \) and \( y_1 = \pi_{J_1}(y_2) \).

Suppose \( C = \{(J_k, y_k)\} \) is a chain in \( S \), and set \( J_0 = \bigcup_k J_k \). For each \( i \in J_0 \), choose \( k \) with \( i \in J_k \) and let \( z_i = \pi_{\{i\}}(y_k) \). If \( i \in J_k \cap J_{k'} \), then either \( J_k \subseteq J_{k'} \) or vice versa, and if \( J_k \subseteq J_{k'} \) we have \( \pi_{\{i\}}(y_k) = \pi_{\{i\}}(\pi_{J_k}(y_{k'})) = \pi_{\{i\}}(y_{k'}) \), so \( z_i \) is well defined independent of the choice of \( k \). Set \( y = (z_i)_{i \in \prod_{i \in J_0} X_i} \). For each \( i \in J_0 \) we have that \( z_i \) is a cluster point of the net \( (\pi_{\{i\}}(x_\alpha))_{\alpha \in A} \). If \( U \) is a neighborhood of \( y_0 \in \prod_{i \in J_0} X_i \), then there are indices \( i_1, \ldots, i_n \in J_0 \) and neighborhoods \( U_r \) of \( z_{i_r} \) in \( X_{i_r} \) such that

\[
\prod_{r=1}^n U_r \times \prod_{i \in J_0 \setminus \{i_1, \ldots, i_n\}} X_i \subseteq U .
\]

Since \( C \) is totally ordered, there is a \( k \) such that \( i_1, \ldots, i_n \in J_k \); for any \( \alpha_0 \) there is an \( \alpha \geq \alpha_0 \) with

\[
\pi_{J_k}(x_\alpha) \in \prod_{r=1}^n U_r \times \prod_{i \in J_0 \setminus \{i_1, \ldots, i_n\}} X_i
\]

and hence \( \pi_{J_k}(x_\alpha) \in U \). So we have that \( y_0 \) is a cluster point of the net \( (\pi_{J_k}(x_\alpha))_{\alpha \in A} \) by (). Thus \( (J_0, y_0) \) is an upper bound for \( C \).

By Zorn’s Lemma, \( S \) has a maximal element \((J, y)\). We need to show that \( J = I \), for then \( y \) will be the desired cluster point of \((x_\alpha)\). Suppose \( J \neq I \), and let \( i \in I \setminus J \), \( J' = J \cup \{i\} \). There is a subnet \((\pi_{J'}(x_{\alpha})_{\beta \in B})_{\alpha \in A} \) of \((\pi_J(x_{\alpha}))_{\alpha \in A} \) which converges to \( y \) by (). The net \( \pi_{\{i\}}(x_{\alpha(\beta)})_{\beta \in B} \) has a subnet \((\pi_{\{i\}}(x_{\alpha(\beta)})_{\gamma \in C})_{\gamma \in C} \) which converges to some \( z \in X_i \) since \( X_i \) is compact. The subnet \( \pi_{J'}(x_{\alpha(\beta)})_{\gamma \in C} \) of \((\pi_J(x_{\alpha}))_{\beta \in B} \) converges to \( y \), so the subnet \((\pi_{J'}(x_{\alpha(\beta)})_{\gamma \in C})_{\alpha \in A} \) converges to \((y, z)\), and thus \((J', (y, z))\) is an element of \( S \) strictly larger than \((J, y)\), contradicting maximality. Thus \( J = I \) and the theorem is proved. \( \Box \)

Tikhonov’s Theorem for Hausdorff Spaces

Tikhonov’s Theorem for Hausdorff spaces does not require the full Axiom of Choice, only the Ultrafilter Property, which is equivalent to the Boolean Prime Ideal Theorem (II.6.3.11.).

The proof we give was developed by H. Cartan for Bourbaki. Cartan’s proof had an additional step (using the full Axiom of Choice) and gave a proof of the full Tikhonov Theorem, but it simplifies nicely to give an efficient proof in the Hausdorff case. The proof is deceptively quick to write, but it depends on earlier results about convergence and projection of ultrafilters.
XI.11.6.6.  **Theorem.** [Tikhonov’s Theorem for Hausdorff Spaces] Let \( \{ (X_i, \mathcal{T}_i) : i \in I \} \) be a collection of Hausdorff topological spaces, \( X = \prod_{i \in I} X_i \), and \( \mathcal{T} \) the product topology on \( X \). If each \( (X_i, \mathcal{T}_i) \) is compact, then \((X, \mathcal{T})\) is compact.

**Proof:** By XI.11.2.3. (which depends on BPI), it suffices to show that if \( \mathcal{F} \) is an ultrafilter on \( X \), then \( \mathcal{F} \) converges. For each \( i \in I \), the ultrafilter \( \mathcal{F}_i = \pi_i(\mathcal{F}) \) converges to a unique point \( x_i \in X_i \) since \( X_i \) is compact and Hausdorff. Then \( \mathcal{F} \) converges to \( x = (x_i) \in X \) by \( (\cdot) \).

XI.11.6.7.  If the \( X_i \) are not Hausdorff, the same proof works except that the limit point \( x_i \) is not necessarily unique. Thus an application of the Axiom of Choice is necessary to choose an \( x_i \) for each \( i \). (If \( I \) is countable, the Countable AC plus the BPI suffices; cf. XI.11.6.4.)

Tikhonov’s Theorem for the Hilbert Cube

It turns out that it can be proved that the Hilbert cube \([0,1]^N\) is compact without any form of choice. We give an argument here from [Wag93].

XI.11.6.8.  **Theorem.** The Hilbert cube \([0,1]^N\) is compact.

**Proof:** Suppose \( \mathcal{U} \) is an open cover of \([0,1]^N\) with no finite subcover. We will obtain a contradiction by repeatedly subdividing coordinatewise. By passing to a refinement, we may assume that the sets in \( \mathcal{U} \) are rectangular open sets of the form \( U_1 \times \cdots \times U_n \times [0,1] \times [0,1] \times \cdots \), where the \( U_k \) are open sets in \([0,1]\) (this can be done without using any form of choice by simply replacing each \( U \in \mathcal{U} \) with the collection of all rectangular open sets contained in \( U \)).

Divide the space into \([0,1/2] \times [0,1] \times \cdots \) and \([1/2,1] \times [0,1] \times \cdots \). At least one of these sets does not have a finite subcover of \( \mathcal{U} \); choose the first one which does not, and call it \( F_1 \). Further subdivide into eight pieces, ordered lexicographically:

\[
[0,1/4] \times [0,1/2] \times [0,1] \times \cdots, [0,1/4] \times [1/2,1] \times [0,1] \times \cdots, [1/4,1/2] \times [0,1] \times \cdots, etc.
\]

Let \( F_2 \) be the first of these sets in this order for which there is no finite subcover of \( \mathcal{U} \). Continue inductively to generate a decreasing sequence of closed sets \( \{ F_k \} \) which are products of subintervals. For each \( n \) the \( n \)’th coordinate intervals \( I_{k,n} \) of the \( F_k \) have an intersection consisting of a single point \( x_n \), and the point \( x = (x_n) \in [0,1]^N \) is in the intersection of all the \( F_k \) (in fact, this intersection is a singleton). There is a \( U \in \mathcal{U} \) with \( x \in U \). Since \( U \) is a rectangular open set determined by finitely many coordinates, we have that \( F_k \subseteq U \) for some sufficiently large \( k \), contradicting the choice of \( F_k \) as a set with no finite subcover from \( \mathcal{U} \).

XI.11.6.9.  It is shown more generally in [?] by a somewhat similar argument that compactness of a product \( \prod_{i \in I} X_i \) of compact spaces can be proved without any form of choice if the index set can be well ordered and there is a choice function for the family of all, or sufficiently many, closed sets in the \( X_i \). (Of course, if the AC is assumed, this proof then gives the full Tikhonov Theorem.)
XI.11.6.10. If, for each $n$, $X_n$ is a closed subset of the Hilbert cube, then without any form of choice it follows that $X = \prod_{n=1}^{\infty} X_n$ also embeds as a closed subset of the Hilbert cube, and thus from XI.11.6.8. it follows that $X$ and each $X_n$ is compact. But it does not seem to be possible to prove that a countable product of compact metrizable spaces is compact without some form of choice, even though such a space embeds in the Hilbert cube (by making some choices): a specific such embedding must be chosen for each $n$ before one can conclude by the above argument that the product embeds, requiring the Countable AC. If one tries to prove directly that the product space is totally bounded with respect to the product metric (proving completeness is no problem), choices have to be made also using the Countable AC. Thus this form of the theorem seems to require Countable AC; if so, it is logically no easier than the version for countable products of arbitrary compact spaces (XI.11.6.4.). It would also be in contrast with the corresponding results for arbitrary products: no more choice is needed to prove that arbitrary products of compact metrizable spaces (indeed, arbitrary products of compact Hausdorff spaces) are compact than to prove that arbitrary powers of $[0,1]$ are compact (each is equivalent to BPI; cf. ())).

XI.11.7. Variations on Compactness
Countable compactness, sequential compactness, quasicompactness, bicompressness, $\sigma$-compactness, local compactness paracompactness, pseudocompactness, metacompactness, Lindelof spaces, precompactness (relative compactness), …

XI.11.7.1. Definition. A topological space $X$ is countably compact if every countable open cover of $X$ has a finite subcover.

Some useful alternate characterizations of countably compact spaces are given by the next result:

XI.11.7.2. Theorem. Let $X$ be a topological space. The following are equivalent:

(i) $X$ is countably compact.

(ii) If $(F_n)$ is a sequence of closed sets in $X$ with the Finite Intersection Property, then $\cap_{n=1}^{\infty} F_n$ is nonempty.

(iii) Every sequence in $X$ has a cluster point.

Proof: (i) $\iff$ (ii) is proved by taking complements exactly as in the proof of XI.11.1.10. with the word “countable” added in appropriate places.

(iii) $\implies$ (i): Suppose $\{U_1, U_2, \ldots\}$ is a countable open cover of $X$ with no finite subcover. For each $n$, let $x_n \in X \setminus \bigcup_{k=1}^{n} U_k$. If $x \in X$, then $x \in U_m$ for some $m$; but $x_n \notin U_m$ for $n \geq m$, so $x$ is not a cluster point of $(x_n)$. Thus $(x_n)$ is a sequence in $X$ with no cluster point.

(i) $\implies$ (iii): Suppose $(x_n)$ is a sequence in $X$ with no cluster point. For each $n$, let $F_n$ be the closure of $\{x_k : k > n\}$, and let $U_n = F_n^{c}$. Then $U = \{U_n : n \in \mathbb{N}\}$ is an open cover of $X$ since any point of $X$ has a neighborhood containing only finitely many of the $x_n$. Since the $F_n$ are decreasing, the $U_n$ are increasing. If there is a finite subcover of $U$, then $U_n = X$ for some $n$. This is a contradiction since $F_n$ is nonempty. Thus $U$ has no finite subcover.

1118
XI.11.7.3. COROLLARY. A sequentially compact topological space (XI.11.2.2.) is countably compact. A first countable topological space is countably compact if and only if it is sequentially compact.

A countably compact space, even a compact Hausdorff space, which is not first countable need not be sequentially compact (XI.11.9.16.(v)).

XI.11.7.4. PROPOSITION. A countably compact Lindelöf space is compact. A countably compact metrizable space is compact.

PROOF: Let \( X \) be countably compact. If \( X \) is Lindelöf, then any open cover of \( X \) has a countable subcover, which has a finite subcover. If \( X \) is metrizable, fix a metric \( \rho \) on \( X \). Any Cauchy sequence in \( X \) has a cluster point, hence converges by (); thus \( (X, \rho) \) is complete. If \( X \) is not totally bounded, then for some \( \epsilon > 0 \) a sequence \((x_n)\) exists in \( X \) with \( \rho(x_n, x_m) \geq \epsilon \) for \( n \neq m \); this sequence cannot have a cluster point, a contradiction.

XI.11.7.5. EXAMPLE. Let \( \omega_1 \) be the first uncountable ordinal, with the order topology. Every sequence in \( \omega_1 \) has an upper bound, hence is contained in a compact initial segment. Thus \( \omega_1 \) is countably compact. But \( \omega_1 \) is not compact: the “identity net” in \( \omega_1 \) has no cluster point.

As a variation, the “long line” \( (\) is countably compact but not compact.

\(\sigma\)-Compact Spaces

XI.11.7.6. DEFINITION. A topological space \( X \) is \( \sigma\)-compact if \( X \) is a countable union of compact subsets of \( X \). More generally, a subset \( E \) of \( X \) is \( \sigma\)-compact if \( E \) is a union of a countable number of compact subsets of \( X \).

Some authors require a \( \sigma\)-compact space to be locally compact, but we will not.

XI.11.7.7. EXAMPLES. (i) Any compact space is \( \sigma\)-compact.
(ii) \( \mathbb{R}^n \) is \( \sigma\)-compact [the closed ball around 0 of radius \( k \) is compact for any \( k \)].
(iii) A manifold \( (\) is \( \sigma\)-compact if and only if it has only countably many connected components, i.e. satisfies the equivalent conditions (\).
(iv) Any countable topological space is \( \sigma\)-compact. A space with the discrete topology is \( \sigma\)-compact if and only if it is countable.

The next result is an immediate corollary of XI.11.1.7:

XI.11.7.8. PROPOSITION. Let \( X \) and \( Y \) be topological spaces, and \( f : X \rightarrow Y \) a continuous function. If \( E \) is a \( \sigma\)-compact subset of \( X \), then \( f(E) \) is a \( \sigma\)-compact subset of \( Y \).

Recall () that an \( F_{\sigma} \) in a topological space is a countable union of closed subsets.
XI.11.7.9. Proposition. Let $X$ and $Y$ be topological spaces, with $X$ $\sigma$-compact and $Y$ Hausdorff. Then

(i) Every $F_\sigma$ in $X$ is $\sigma$-compact.

(ii) Every $\sigma$-compact subset of $Y$ is an $F_\sigma$.

Proof: (i): Suppose $X = \bigcup_n K_n$, where $K_n$ is compact. If $F$ is a closed subset of $X$, then $F = \bigcup_n (F \cap K_n)$, and by XI.11.1.6. $F \cap K_n$ is compact. Thus every closed set in $X$ is $\sigma$-compact, and hence so is every $F_\sigma$.

(ii): If $E = \bigcup_n E_n \subset Y$ with $E_n$ compact, then $E_n$ is closed by XI.11.1.6., so $E$ is an $F_\sigma$.

XI.11.7.10. Corollary. Let $X$ be a $\sigma$-compact Hausdorff space. Then a subset of $X$ is $\sigma$-compact if and only if it is an $F_\sigma$.

It follows that the set of irrational numbers in $\mathbb{R}$ is not $\sigma$-compact since it is not an $F_\sigma$.

Combining these results, we get an interesting consequence. Although a continuous image of a closed set is not closed in general, we have:

XI.11.7.11. Proposition. Let $X$ and $Y$ be topological spaces, with $X$ $\sigma$-compact and $Y$ Hausdorff, and let $f : X \to Y$ be a continuous function. If $E$ is an $F_\sigma$ in $X$, then $f(E)$ is an $F_\sigma$ in $Y$.

Proof: $E$ is $\sigma$-compact by XI.11.7.9.(i), so $f(E)$ is also $\sigma$-compact by XI.11.7.8., and hence by XI.11.7.9.(ii) we conclude that $f(E)$ is an $F_\sigma$ in $Y$.

In particular, under the hypotheses, the image of any closed subset of $X$ is an $F_\sigma$ in $Y$. The hypotheses hold if $X$ and $Y$ are Euclidean spaces, or manifolds with $X$ separable.

XI.11.7.12. Proposition. A $\sigma$-compact topological space is Lindelöf. Thus a $\sigma$-compact regular space is normal.

Proof: Let $X = \cup_n K_n$, with $K_n$ compact. If $\mathcal{U}$ is an open cover of $X$, for each $n$ let $\mathcal{V}_n$ be a finite subset of $\mathcal{U}$ which covers $K_n$. Then $\mathcal{V} = \cup_n \mathcal{V}_n$ is a countable subset of $\mathcal{U}$ which covers $X$. (This argument uses the Countable AC.) The last statement follows from (i).

XI.11.8. Locally Compact Spaces

Many important spaces such as Euclidean spaces are not compact, but are “locally compact”:

XI.11.8.1. Definition. A Hausdorff space $X$ is locally compact if every point of $X$ has a neighborhood (not necessarily open) which is compact.

We will deal with the non-Hausdorff case in XI.11.8.20.
XI.11.8.2. Examples. (i) Any compact Hausdorff space is locally compact. (The whole space is a compact neighborhood of any point.)

(ii) \( \mathbb{R}^n \) is locally compact since every closed ball is compact. More generally, any open set in \( \mathbb{R}^n \) is locally compact in the relative topology for the same reason.

(iii) Any space with the discrete topology is locally compact.

(iv) Any closed subspace of a locally compact Hausdorff space is locally compact in the relative topology: if \( Y \) is a closed subset of a locally compact Hausdorff space \( X \), \( p \in Y \), and \( K \) is a compact neighborhood of \( p \) in \( X \), then \( K \cap Y \) is a compact neighborhood of \( p \) in \( Y \) in the relative topology. In particular, any closed subset of \( \mathbb{R}^n \) is locally compact in the relative topology.

(v) Any open subset of a compact Hausdorff space is locally compact (in the relative topology): if \( U \) is an open subset of a compact Hausdorff space \( X \), and \( p \in U \), then by XI.11.4.2, \( p \) and \( X \setminus U \) have disjoint neighborhoods \( V \) and \( W \) in \( X \). \( U \) is a compact neighborhood of \( p \) in \( X \) which is contained in \( U \), i.e. a compact neighborhood of \( p \) in \( U \).

Actually, any open subset of a locally compact Hausdorff space is locally compact (XI.11.8.8); in fact, any locally compact Hausdorff space is homeomorphic to an open set in a compact Hausdorff space (XI.11.8.4).

(vi) Spaces such as \( \mathbb{Q}, \mathbb{N} = \mathbb{N}^n \), and \( \mathbb{R}^n \) are not locally compact; in fact, no point in any of these spaces has a compact neighborhood.

The One-Point Compactification

XI.11.8.3. If \( X \) is a locally compact Hausdorff space, a single point can be added to \( X \) to make a compact Hausdorff space, called the one-point compactification of \( X \) and denoted \( X^\dagger \). Specifically, take a point not in \( X \) which we call \( \infty \), and let \( X^\dagger = X \cup \{ \infty \} \). Put a topology on \( X^\dagger \) by taking all open subsets of \( X \) to be open in \( X^\dagger \), along with all subsets consisting of \( \infty \) along with the complement of a compact subset of \( X \) (including the complement of \( \emptyset \), giving the whole space \( X^\dagger \)). It is simple and routine to verify that this is a topology (Exercise XI.11.12.12.), and that \( X \) is an open subset of \( X^\dagger \). See XI.11.12.12. for a generalization of the one-point compactification construction.

XI.11.8.4. Proposition. \( X^\dagger \) is a compact Hausdorff space with this topology.

Proof: It is easy to see that \( X^\dagger \) is compact. If \( \{U_i\} \) is an open cover, there is a set \( U_{i_0} \) containing \( \infty \). The complement of \( U_{i_0} \) is a compact subset of \( X \), so the open cover of \( X^\dagger \setminus U_{i_0} \) by the remaining \( U_i \) has a finite subcover \( \{U_{i_1}, \ldots, U_{i_n}\} \). Then \( \{U_{i_0}, U_{i_1}, \ldots, U_{i_n}\} \) is a finite subcover of \( X^\dagger \).

To see that \( X^\dagger \) is Hausdorff, it suffices to show that if \( p \in X \), then \( p \) and \( \infty \) have disjoint neighborhoods. Since \( X \) is locally compact, there is an open neighborhood \( U \) of \( p \) in \( X \) whose closure in \( X \) is compact (e.g. the interior of a compact neighborhood). Thus \( U \) and \( \{\infty\} \cup (X \setminus U) \) are disjoint neighborhoods of \( p \) and \( \infty \).

The one-point compactification of \( X \) is essentially unique. Thus it can often be realized or recognized in a different form:
XI.11.8.5. PROPOSITION. Let $X$ be a locally compact Hausdorff space, $Y$ a compact Hausdorff space, and $\phi: X \to Y$ a homeomorphism from $X$ onto a (necessarily open) subspace $\phi(X)$ of $Y$ with $Y \setminus \phi(X)$ a single point $q$. Then $\phi$ extends to a homeomorphism $\tilde{\phi}$ from $X^\dagger$ onto $Y$ with $\tilde{\phi}(\infty) = q$.

PROOF: The map $\tilde{\phi}$ is obviously completely determined; it is only necessary to show it is a homeomorphism. If $V$ is an open set in $Y$, and $V$ does not contain $q$, then since $\phi$ is continuous, $\tilde{\phi}^{-1}(V) = \phi^{-1}(V)$ is open in $X$, hence in $X^\dagger$. And if $q \in V$, then $V^c$ is a compact subset of $\phi(X)$, hence $\tilde{\phi}^{-1}(V^c)$ is compact since $\phi^{-1}$ is continuous, so $\tilde{\phi}^{-1}(V) = X^\dagger \setminus \phi^{-1}(V^c)$ is open in $X^\dagger$. Thus $\tilde{\phi}$ is continuous. A continuous bijection from a compact space to a Hausdorff space is a homeomorphism by (i), so $\tilde{\phi}$ is a homeomorphism.

XI.11.8.6. EXAMPLES. (i) Let $Y$ be any compact Hausdorff space, and $q \in Y$. Then $X = Y \setminus \{q\}$ is a locally compact Hausdorff space with $X^\dagger$ homeomorphic to $Y$. By the one-point compactification construction, every locally compact Hausdorff space arises this way.

(ii) An $n$-sphere with a point removed is homeomorphic to $\mathbb{R}^n$, e.g. by stereographic projection (i). Thus $(\mathbb{R}^n)^\dagger$ is homeomorphic to $S^n$. In particular, $\mathbb{R}^\dagger$ is homeomorphic to a circle.

XI.11.8.7. If $X$ is already compact, then $\emptyset$ is the complement of a compact subset of $X$, so $\{\infty\}$ is open in $X^\dagger$, i.e. $\infty$ is an isolated point in $X^\dagger$, and $X^\dagger$ is the separated union of $X$ and a one-point discrete space. Conversely, if $\infty$ is isolated in $X^\dagger$, then $X$ is a closed subset of $X^\dagger$, hence compact. Thus $X$ is dense in $X^\dagger$ if and only if $X$ is noncompact.

XI.11.8.8. It follows that any open subset of a locally compact Hausdorff space is locally compact: if $U$ is an open subset of a locally compact Hausdorff space $X$, then $U$ is an open subset of the compact Hausdorff space $X^\dagger$ and thus is locally compact in the relative topology by XI.11.8.2.(v).

We also get the following corollaries of the one-point compactification construction:

XI.11.8.9. PROPOSITION. A locally compact Hausdorff space is completely regular (i).

PROOF: $X^\dagger$ is normal by XI.11.4.3., hence completely regular, and complete regularity passes to subspaces.

However, a locally compact Hausdorff space need not be normal (XI.11.12.6., XI.19.5.8.).

XI.11.8.10. PROPOSITION. Let $X$ be a locally compact Hausdorff space. If $p \in X$ and $V$ is a neighborhood of $p$, then there is an open neighborhood $U$ of $p$ such that $U$ is compact and contained in $V$.

PROOF: There is a compact neighborhood $K$ of $p$ in $X$; the interior $W$ of $K$ is an open neighborhood of $p$ whose closure is compact. Replacing $V$ by $V \cap W$, we may assume $V^c$ is compact. Since $X$ is regular, $\{p\}$ and $V^c$ have disjoint neighborhoods $U$ and $Z$. Then $\bar{U} \subseteq V$, and $U$ is compact since it is a closed subset of $V$.  

1122
XI.11.8.11. Proposition. Let $X$ be a locally compact Hausdorff space. The following are equivalent:

(i) The point $\infty$ has a countable local base in $X^\dagger$.

(ii) $\{\infty\}$ is a $G_\delta$ in $X^\dagger$.

(iii) $X$ is $\sigma$-compact.

(iv) $X$ is Lindelöf.

Proof: It is obvious from the construction of $X^\dagger$ that (ii) $\iff$ (iii), and it follows from XI.11.4.5. that (i) $\iff$ (ii). (iii) $\implies$ (iv) by XI.11.7.12. For (iv) $\implies$ (iii), each $x \in X$ has an open neighborhood $V_x$ with $\bar{V}_x$ compact. $\{V_x : x \in X\}$ has a countable subcover $\{V_{x_n} : n \in \mathbb{N}\}$, and $X = \cup_n \bar{V}_{x_n}$.

XI.11.8.12. Proposition. Let $X$ be a locally compact, $\sigma$-compact Hausdorff space. Then there is a sequence $(U_n)$ of open sets in $X$ with $\bar{U}_n$ compact and contained in $U_{n+1}$ for each $n$, and $X = \cup_1^\infty U_n$.

Proof: Let $(K_n)$ be a sequence of compact subsets of $X$ with $X = \cup K_n$. Let $U_1$ and $V_1$ be disjoint neighborhoods of $K_1$ and $\{\infty\}$ in $X^\dagger$; then $U_1$ is open and the closure $\bar{U}_1$ of $U_1$ in $X^\dagger$ is compact and contained in $X$. Inductively take $U_{n+1}$ and $V_{n+1}$ to be disjoint open neighborhoods of $U_n \cup K_{n+1}$ and $\{\infty\}$ in $X^\dagger$.

XI.11.8.13. Theorem. Let $X$ be a locally compact Hausdorff space. The following are equivalent:

(i) $X$ is metrizable and $\sigma$-compact.

(ii) There is a countable base for the topology of $X$ consisting of compact sets.

(iii) $X$ is second countable.

(iv) $X$ is a Polish space.

(v) $X^\dagger$ is metrizable.

There are various other ways to phrase these conditions which are equivalent by general considerations, e.g. “$X$ is separable and metrizable” (4). There are metrizable locally compact spaces which do not satisfy the conditions, e.g. an uncountable discrete space.

Proof: (i) $\implies$ (v): Let $(U_n)$ be an increasing sequence of open sets in $X$ as in XI.11.8.12. Then $\bar{U}_n$ is compact and metrizable for each $n$, hence second countable (4). Thus $U_n$ is second countable. There is also a countable local base at $\infty$ in $X^\dagger$ by XI.11.8.11. Putting together a countable base for the topology of $U_n$ for each $n$ and a countable local base at $\infty$, we obtain a countable base for the topology of $X^\dagger$; hence $X^\dagger$ is a second countable compact Hausdorff space, so it is metrizable (4).

(v) $\implies$ (iv): A compact metrizable space is a Polish space; so $X^\dagger$ is Polish. Thus the open subset $X$ of $X^\dagger$ is also a Polish space.
(iv) ⇒ (iii): A separable metric space is second countable (\).

(iii) ⇒ (ii): It suffices to show that if \( \{ U_n \} \) is a countable base, then the collection
\[
\{ \hat{U}_n : n \in \mathbb{N}, \hat{U}_n \text{ compact} \}
\]
is also a base. Let \( p \in X \) and \( V \) a neighborhood of \( p \) in \( X \). There is an open set \( U \) containing \( p \) such that \( \hat{U} \) is compact and contained in \( V \) (XI.11.8.10.). Take \( U_n \) so that \( p \in U_n \subseteq U \). Then \( \hat{U}_n \subseteq V \), and \( \hat{U}_n \) is compact since it is a closed subset of \( \hat{U} \).

(ii) ⇒ (i): \( X \) is regular (XI.11.8.9.), hence metrizable by the Urysohn Metrization Theorem (\( ). (ii) trivially implies that \( X \) is \( \sigma \)-compact. ☑

XI.11.8.14. There is no universally accepted standard notation for the one-point compactification of a space \( X \). I have chosen a notation \( X^\dagger \) that is not widely used, and an explanation is in order. The most common notations are \( X^* \), \( X^\oplus \), \( X_1 \); other notations sometimes used are \( X^\ast \), \( X^\ast_\ast \), \( X^\ast_\ast \), and perhaps others.

There are two reasons I have chosen not to use \( X^\ast \). One is that topologists also often use this notation to mean something else, e.g. the corona space \( \beta X \setminus X \) (\( ). The other reason comes from my background in Operator Algebras, where \( \ast \) is normally used to mean something else (\( ); in fact, it is used to mean too many different things in Operator Algebras!\( ) In addition, there is a natural relationship between adding a point at infinity in topology and adding a unit in algebra: a commutative C*-algebra is an algebra of the form \( C_0(X) \), the continuous functions vanishing at infinity on a locally compact Hausdorff space \( X \) (\( ), and the unitization of this algebra is precisely the continuous functions on the one-point compactification of \( X \). Thus it makes sense to use the same notation for point-one compactification as for adding a unit to a C*-algebra. A common notation in algebra is to write \( A^\dagger \) for the unitization of the algebra \( A \), and thus \( X^\dagger \) would be logical notation for the one-point compactification, giving that \( C_0(X)^\dagger = C(X^\dagger) \). However, a C*-algebra \( A \) also has a positive cone which is usually denoted \( A_\ast \) or sometimes \( A^\dagger \), and the two notations are too easily confused. So in [Bla06] I settled on the notation \( A^\dagger \) for the unitization of a C*-algebra, and thus \( X^\dagger \) for the one-point compactification of a space (so that \( C_0(X)^\dagger = C(X^\dagger) \)), as a variation of \( A^\dagger \) which is easily distinguishable but similar enough to be suggestive, and I have kept that notation here. Some readers may not be happy with my choice, but it is my book . . .

Urysohn’s Lemma for Locally Compact Spaces

Although a locally compact space need not be normal (XI.19.5.8.), there is a version of Urysohn’s Lemma. We first make an observation and a definition:

XI.11.8.15. PROPOSITION. Let \( X \) be a locally compact Hausdorff space, and \( A \) a compact subset of \( X \), and \( V \) a neighborhood of \( A \) in \( X \). Then there is an open neighborhood \( U \) of \( A \) contained in \( V \) with \( \hat{U} \) compact.

PROOF: This is a routine compactness argument from XI.11.8.10. For each \( p \in U \), let \( U_p \) be an open neighborhood of \( p \) contained in \( V \) with \( \hat{U}_p \) compact. The open cover \( \{ U_p : p \in A \} \) of \( A \) has a finite subcover \( \{ U_{p_1}, \ldots, U_{p_n} \} \). Set \( U = U_{p_1} \cup \cdots \cup U_{p_n} \). ☑

1124
XI.11.8.16. **Definition.** Let $X$ be a topological space. A continuous function $f : X \to \mathbb{C}$ has **compact support** if the closure of the support $S_f = \{ x \in X : f(x) \neq 0 \}$ of $f$ is compact. Denote by $C_c(X)$ the set of all continuous functions of compact support from $X$ to $\mathbb{C}$.

XI.11.8.17. It is easily verified that $C_c(X)$ is a subalgebra of $C(X)$. But in general it need not contain any nonzero function. However, if $X$ is locally compact, $C_c(X)$ is large enough to separate points and closed sets:

XI.11.8.18. **Theorem.** [Urysohn’s Lemma, Locally Compact Case] Let $X$ be a locally compact Hausdorff space. If $A$ and $B$ are disjoint closed subsets of $X$ with $A$ compact, then there is a continuous function $f : X \to [0,1]$ of compact support with $f(x) = 1$ for all $x \in A$ and $f(x) = 0$ for all $x \in B$.

**Proof:** The set $V = B^c$ is a neighborhood of $A$. Let $U$ be an open neighborhood of $A$ in $X$ contained in $V$ with $\overline{U}$ compact. Regarding $X$, and hence $A, B, U$, as subsets of $X^\dagger$, let $C = X^\dagger \setminus U$. Then $A$ and $C$ are disjoint closed subsets of the compact, hence normal, space $X^\dagger$. If $g : X \to [0,1]$ with $g(x) = 1$ for all $x \in A$ and $g(x) = 0$ for all $x \in C$, then the restriction of $g$ to $X$ is the desired $f$.

XI.11.8.19. In particular, if $X$ is a locally compact Hausdorff space, $B$ a closed subset of $X$, and $p \in X$, $p \notin B$, then (applying the theorem with $A = \{p\}$) there is a continuous function $f : X \to [0,1]$ of compact support with $f(p) = 1$ and $f(x) = 0$ for all $x \in B$.

**The Non-Hausdorff Case**

XI.11.8.20. There is no natural definition of local compactness for non-Hausdorff spaces. Each of the following conditions, which are equivalent for Hausdorff spaces, is a reasonable potential definition for a topological space $X$:

(i) Every point of $X$ has a neighborhood (not necessarily open) which is compact.

(ii) For any $p \in X$ and any open neighborhood $V$ of $X$, there is a compact neighborhood $U$ (not necessarily open) of $p$ contained in $V$.

(iii) Every point of $X$ has an open neighborhood whose closure is compact.

(iii′) For any $p \in X$ and any open neighborhood $V$ of $X$, there is an open neighborhood $U$ of $X$ contained in $V$ whose closure is compact.

(iv) For any $p \in X$ and any open neighborhood $V$ of $X$, there is an open neighborhood $U$ of $X$ whose closure is compact and contained in $V$.

In general, (iv) $\Rightarrow$ (iii) $\iff$ (iii′) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (ii) $\Rightarrow$ (i). In Hausdorff spaces, (i) $\Rightarrow$ (iv) (XI.11.8.10.), so all the conditions are equivalent. None of the other implications hold in general [SS95, Examples 9, 18, 35].
XI.11.9. Compactifications

It is often very convenient to embed a given topological space into a compact Hausdorff space. The first instance of this is to add \( \pm \infty \) to the real line to obtain the extended real line, which is topologically and order-theoretically the same as the compact space \([0, 1]\). There are some systematic ways to compactify a topological space, which we discuss in this section.

XI.11.9.1. Definition. Let \( X \) be a Hausdorff topological space. A compactification of \( X \) is a compact Hausdorff space \( Y \) and a homeomorphism \( \phi \) of \( X \) onto a dense subset of \( Y \).

Some authors consider compactifications where \( X \) and/or \( Y \) is not Hausdorff (cf. XI.11.9.20.), but we will only consider the Hausdorff case. By (), a (Hausdorff) space has a compactification if and only if it is completely regular.

XI.11.9.2. In some references, a pair \((Y, \phi)\) is called a compactification of \( X \) if \( Y \) is a compact Hausdorff space and \( \phi \) is simply an injective continuous function from \( X \) onto a dense subset of \( Y \). Instances of this are the standard maps of \( \mathbb{R} \) to a figure 8, a figure 6, a torus via an irrational angle wraparound (), or a solenoid (). Another instance would be a map from \( \mathbb{N} \) with the discrete topology onto a dense subset of \([0, 1]\). We will not consider these to be compactifications, but they are what we call compact representations (XI.11.9.9.).

XI.11.9.3. The density of \( X \) in a compactification is a natural condition, but is not restrictive: if \( \phi \) is a homeomorphism of \( X \) with a subspace of a compact Hausdorff space \( Z \), then the closure of \( \phi(X) \) in \( Z \) is a compactification of \( X \).

XI.11.9.4. There are several special instances of compactifications already considered:

(i) If \( (X, \rho) \) is a totally bounded metric space, then the completion of \( (X, \rho) \) is a compactification of \( X \).

(ii) If \( X \) is a totally ordered space with the order topology, then the order-completion of \( X \) with endpoints added if necessary, with the order topology, is a compactification of \( X \) (i).

(iii) If \( X \) is a locally compact, noncompact Hausdorff space, then the one-point compactification \( X^\dagger \) () is a compactification of \( X \). (If \( X \) is already compact, then \( X^\dagger \) is not a compactification of \( X \) since \( X \) is not dense in \( X^\dagger \); in fact, the only compactification of a compact space \( X \) is \( X \) itself. The fact that the one-point compactification is not a compactification in this case is an unfortunate inconsistency in standard terminology.)

(iv) There are more complicated compactifications of \( \mathbb{R} \) than the extended real line and the one-point compactification (circle). For example, there is a homeomorphism of \( \mathbb{R} \) onto the graph of \( y = \sin \left( \frac{1}{x} \right) \) for \( 0 < x < \frac{1}{\pi} \), so the topologist’s sine curve () is a compactification of \( \mathbb{R} \).

(v) If \( X \) is completely regular, and \( \phi \) is a homeomorphism of \( X \) onto a subspace of a cube (), then the closure of \( \phi(X) \) is a compactification of \( X \).

The compactification of (v) is actually universal: every compactification can be done (up to homeomorphism) in this way. For if \( Y \) is any compactification of \( X \), then \( Y \) can be embedded in a cube as in (). The construction of () using the restrictions to \( X \) of the coordinate functions of this embedding yields \( Y \).
XI.11.9.5. Since a compact metrizable space is second countable, a space with a metrizable compactification is second countable. Conversely, by the construction of (I) a second countable completely regular space can be embedded in the Hilbert cube, which is metrizable, and hence has a metrizable compactification.

XI.11.9.6. Although the homeomorphism $\phi$ is often understood, it is logically cleaner to regard it as a part of the compactification. Two compactifications $(Y, \phi)$ and $(Z, \psi)$ of a space $X$ are equivalent if there is a homeomorphism $\omega : Y \to Z$ with $\psi = \omega \circ \phi$. If there is just a continuous function $\omega : Y \to Z$ with $\psi = \omega \circ \phi$, we say $(Y, \phi)$ dominates $(Z, \psi)$. Such an $\omega$ is unique since it is completely determined on the dense subset $\phi(X)$ as $\psi \circ \phi^{-1}$, and it is necessarily surjective since $\omega(Y)$ is a compact subset of $Z$ containing the dense subset $\psi(X)$.

Dominance gives a partial ordering on the equivalence classes of compactifications of $X$, setting $(Y, \phi) \geq (Z, \psi)$ if $(Y, \phi)$ dominates $(Z, \psi)$. If $X$ is locally compact and noncompact, then the one-point compactification of $X$ is the smallest compactification of $X$ under this ordering; conversely, it is not hard to show (cf. (I)) that if $X$ has a minimal compactification, it must be a one-point compactification and $X$ must thus be locally compact. There is always a largest compactification of $X$ if $X$ is completely regular, the Stone-Čech compactification $\beta X$.

XI.11.9.7. More generally, if $X_1$ and $X_2$ are (completely regular) spaces with compactifications $(Y, \phi)$ and $(Z, \psi)$ respectively, and $f : X_1 \to X_2$ is a continuous function, an extension of $f$ to $(Y, Z)$ is a continuous function $\tilde{f} : Y \to Z$ with $\tilde{f} \circ \phi = \psi \circ f$. Such an extension, if it exists, is unique.

XI.11.9.8. A compactification $(Y, \phi)$ of $X$ naturally defines a uniform structure (I) on $X$ by restricting the unique uniform structure on $Y$ (I). (However, not every uniform structure on $X$ arises in this manner in general.) A continuous function $f : X_1 \to X_2$ extends to compactifications $(Y, \phi)$ and $(Z, \psi)$ in the sense of XI.11.9.7. if and only if $f$ is uniformly continuous for these uniform structures.

XI.11.9.9. More generally, we define a compact representation of a (completely regular) space $X$ to be a pair $(Y, \phi)$, where $Y$ is a compact Hausdorff space and $\phi$ is a continuous function from $X$ onto a dense subspace of $Y$. Thus a compactification is a special kind of compact representation, as are the “compactifications” of XI.11.9.2. Another important example of a compact representation is a bounded continuous function from $X$ to $\mathbb{R}$, e.g. a continuous function from $X$ to $[0,1]$. Equivalence, domination, and extension can be defined in this more general context. Any compact representation which dominates a compactification is itself a compactification.

The Stone-Čech Compactification

There is a largest (in the sense of XI.11.9.6.) compactification of any completely regular space $X$, called the Stone-Čech compactification of $X$ and denoted $\beta X$. This compactification is attributed to M. Stone and E. Čech (independently), but see XI.11.9.21.

The Stone-Čech compactification of a space $X$ can be constructed in a number of ways (Stone and Čech used rather different constructions). One construction is via Boolean algebras and/or ultrafilters (cf. XI.11.9.14.). A very slick construction is as the spectrum of the C*-algebra of bounded complex-valued continuous functions on $X$ (XV.14.2.21.), but this approach is probably not the best for beginning students or topologists. We will give two versions of a purely topological construction.

The first version is based on the next result, which is proved by a pullback construction:
XI.11.9.10. Proposition. Let $X$ be a completely regular space, and $\{ (Y_i, \phi_i) : i \in I \}$ a set of compact representations (XI.11.9.9.) of $X$. Then there is a compact representation $(Y, \phi)$, unique up to equivalence, which is the least upper bound of the $(Y_i, \phi_i)$.

Proof: It is obvious that the least upper bound, if it exists, is unique up to equivalence. We show existence. Let $Z = \prod_i Y_i$. Then $Z$ is compact by Tikhonov’s Theorem (the Hausdorff version XI.11.6.6. suffices). Define a map $\phi : X \to Z$ by $(\phi(x))_i = \phi_i(x)$ for $x \in X$. Then $\phi$ is continuous by the universal property of the product topology (). Let $Y'$ be the closure of $\phi(X)$ in $Z$. Then $(Y, \phi)$ obviously dominates $(Y_i, \phi_i)$ via coordinate projections. If $(Y', \psi)$ dominates each $(Y_i, \phi_i)$ via a map $\omega_i : Y' \to Y_i$, define $\omega : Y' \to Y$ by $(\omega(t))_i = \omega_i(t)$ for $t \in Y'$. Then $(Y', \psi)$ dominates $(Y, \phi)$ via $\omega$.

XI.11.9.11. One then gets an apparent proof of the existence of a maximal compactification by simply taking the least upper bound of the “set” of all compact representations of $X$. However, there is a set-theoretic problem: this “set” is not a set. The problem can be circumvented by noting that if $X$ has cardinality $\kappa$, then the cardinality of any compact representation is at most $2^{2^\kappa}$, so we could restrict to only compact representations whose underlying set is a subset of a fixed set of cardinality $2^{2^\kappa}$, and these compact representations form a set. One can also restrict to one compact representation in each equivalence class in the construction. (This appears to require the AC (), but the AC can be avoided by an argument working through XI.11.9.12.; anyway, use of the AC is not much of a problem here since some form of choice in the guise of Tikhonov’s Theorem is necessary to do almost anything with compactifications.)

Call the maximal compact representation $\beta X$; it is a compactification since $X$ has at least one compactification, which is dominated by $\beta X$. It is customary to identify $X$ with its image in $\beta X$.

XI.11.9.12. A more common and concrete construction of $\beta X$ is to let $C$ be the set of all continuous functions from $X$ to $[0, 1]$, and embed $X$ into $[0, 1]^C$ by taking $\phi(x)_f = f(x)$ for $f \in C$ and taking $Y$ to be the closure of $\phi(X)$. Since each such $f$ gives a compact representation of $X$, each of these is dominated by the $\beta X$ constructed previously, hence $(Y, \phi)$ is dominated by $\beta X$. On the other hand, since $\beta X$ is compact, it embeds into a cube (), and the coordinate functions of this embedding, restricted to $X$, are in $C$; it follows that $(Y, \phi)$ dominates $\beta X$, and therefore $(Y, \phi)$ is equivalent to $\beta X$.

We summarize in a theorem:

XI.11.9.13. Theorem. [Stone-Čech Compactification] Let $X$ be a completely regular space. There is a compact Hausdorff space $\beta X$ containing $X$ as a dense subspace, unique up to homeomorphism which is the identity on $X$, with the property that if $f$ is a continuous function from $X$ to any compact Hausdorff space $Y$, then $f$ extends uniquely to a continuous function $\bar{f} : \beta X \to Y$.

XI.11.9.14. We describe the ultrafilter version of the construction of $\beta X$ only for $X$ a set with the discrete topology. For the general version see [GJ76]. If $X$ is a discrete space, $\beta X$ can be identified with the set $\sigma X$ of all ultrafilters on $X$, with the hull-kernel topology (); the closure of a set $S$ of ultrafilters is the set of all ultrafilters containing the filter which is the intersection of the ultrafilters in $S$. It is easy to show that this space is a compact Hausdorff space (), and that the set of fixed ultrafilters, which can be naturally identified with $X$, is dense, so $\sigma X$ is a compactification of $X$. To show that $\sigma X$ dominates $\beta X$,
XI.11.9.15. If $X$ is locally compact, then the Stone-Čech compactification dominates the one-point compactification $X^\dagger$. Since $X$ is open in $X^\dagger$, its inverse image in $\beta X$, the image of $X$ in $\beta X$, is open. Conversely, if $X$ is open in $\beta X$, then $X$ is locally compact by XI.11.8.2.(v).

The set $\beta X \setminus X$ is called the corona or remainder of $X$, especially if $X$ is locally compact; in this case $\beta X \setminus X$ is compact.

XI.11.9.16. The Stone-Čech compactification of a space is rather huge, and often has pathological properties. We list without proof some key properties of some standard spaces:

(i) If $X$ is second countable and noncompact, then $\text{card}(\beta X) = 2^{2^{\omega_0}}$. In particular, $\text{card}(\beta \mathbb{N}) = \text{card}(\beta \mathbb{Q}) = \text{card}(\beta \mathbb{R}) = 2^{2^{\omega_0}}$. If $X$ is any space which is locally compact and $\sigma$-compact, but not compact, then $\text{card}(\beta X \setminus X) \geq 2^{2^{\omega_0}}$.

(ii) Every infinite closed subset of $\beta \mathbb{N} \setminus \mathbb{N}$ contains a copy of $\beta \mathbb{N}$, and in particular has cardinality $2^{2^{\omega_0}}$.

(iii) $\beta \mathbb{N}$ is extremally disconnected: the closure of every open set is open.

(iv) $\beta \mathbb{R}$ is connected (as is $\beta X$ for any connected $X$ (XI.13.1.10.)). $\beta \mathbb{R} \setminus \mathbb{R}$ has two connected components, one at $+\infty$ and the other at $-\infty$ (cf. XI.13.3.6.). But $\beta \mathbb{R} \setminus \mathbb{R}$ is totally path disconnected, and not locally connected at any point. In fact, if $X$ is locally compact and $\sigma$-compact, but not compact, then $\beta X \setminus X$ is an $F$-space: disjoint open $F_{\sigma}$’s have disjoint closures. Any $F$-space is totally path disconnected. In $\beta \mathbb{N} \setminus \mathbb{N}$, $\beta \mathbb{R} \setminus \mathbb{R}$, and many other corona spaces, every nonempty $G_\delta$ has nonempty interior.

(v) Although any sequence in $\beta \mathbb{N}$ has a cluster point (in fact, $2^{2^{\omega_0}}$ cluster points if it has infinite range), the only convergent sequences in $\beta \mathbb{N}$ are ones which are eventually constant. Thus any sequence of distinct points in $\beta \mathbb{N}$ has no convergent subsequence, and $\beta \mathbb{N}$ is compact but not sequentially compact. The same statements are true about $\beta \mathbb{R} \setminus \mathbb{R}$, so $\beta \mathbb{R} \setminus \mathbb{R}$ is also not sequentially compact.

For more detailed structure results about $\beta \mathbb{N}$ and related spaces, see [GJ76] and [vM84].

XI.11.9.17. There are, however, locally compact noncompact spaces for which the one-point compactification and the Stone-Čech compactification coincide. The simplest example is the space $\omega_1$ of ordinals less than the first uncountable ordinal, with the order topology, since every real-valued continuous function on $\omega_1$ is eventually constant; cf. also XI.19.5.6.. These spaces are pseudocompact: every real-valued continuous function is bounded.

XI.11.9.18. Note that either Tikhonov’s Theorem for Hausdorff spaces or the Ultrafilter Property is needed to prove Theorem XI.11.9.13.. In fact, this theorem is equivalent to the BPI (II.6.3.11.).

XI.11.9.19. Theorem. The following theorems are logically equivalent:

(i) Tikhonov’s Theorem for Hausdorff spaces.

(ii) Tikhonov’s Theorem for copies of $[0, 1]$.


**Proof:** (i) ⇒ (ii) and (iv) ⇒ (iii) are trivial. (ii) ⇒ (iv) is the standard construction of the Stone-Čech compactification (XI.11.9.12). So we need only prove (iii) ⇒ (i).

Let \( \{X_i : i \in I\} \) be an indexed collection of compact Hausdorff spaces, and let \( X = \prod X_i \) with the product topology. Let \( Y \) be the set \( X \) with the discrete topology, and \( \beta Y \) its Stone-Čech compactification. For each \( i \), the coordinate map \( \pi_i : X \to X_i \), regarded as a map from \( Y \) to \( X_i \), extends uniquely to a continuous function \( f_i \) from \( Y \) to \( X_i \), by the universal property of \( \beta Y \). The \( f_i \) define a continuous function from \( \beta Y \) to \( X \), which is surjective since its restriction to \( Y \) is the identity map from \( Y \) to \( X \). Thus \( X \) is a continuous image of the compact space \( \beta Y \), hence compact by (i).

Note that (ii) ⇒ (i) is a simple consequence of the fact that every compact Hausdorff space can be embedded in a product of intervals, but we did not need to use this argument.

XI.11.9.20. There is a compactification of any \( T_1 \) space \( X \) called the Wallman compactification, which is similar to the Stone-Čech compactification but not Hausdorff if \( X \) is not normal (it coincides with the Stone-Čech compactification if \( X \) is normal). The Wallman compactification shows that every \( T_1 \) space can be embedded as a dense subset of a compact \( T_1 \) space (for a simpler proof of this, see XI.11.12.12.).

XI.11.9.21. According to [Dud02], Tikhonov was the first to prove existence of the Stone-Čech compactification (in 1930), and Čech elaborated on Tikhonov’s work; Stone’s work appears to have been done independently, but later than Tikhonov (as Gel’fand once observed, Tikhonov obtained his results “previously but independently.”)

On the other hand, according to [Dud02], the general form (XI.11.6.1.) of Tikhonov’s Theorem was first proved by Čech in 1936; Tikhonov only proved that an arbitrary product of copies of \([0, 1]\) is compact.

XI.11.10. Paracompactness and Partitions of Unity

For some purposes, it is enough to be able to make an open cover of a space “locally finite” instead of actually (globally) finite, and it suffices to allow making a given open cover “finer” instead of just working with an actual subcover. It turns out that this is possible in the great majority of topological spaces encountered “naturally”; such spaces are called paracompact.

Paracompact spaces were only defined and named in 1944 by J. Dieudonné.

XI.11.10.1. **Definition.** Let \( \mathcal{U} \) be an open cover of a topological space \((X, T)\). \( \mathcal{U} \) is point finite if every \( x \in X \) is contained in only finitely many \( U \in \mathcal{U} \). \( \mathcal{U} \) is locally finite if each \( x \in X \) has an open neighborhood \( W \) such that \( W \cap U \neq \emptyset \) for only finitely many \( U \in \mathcal{U} \).

A locally finite cover is obviously point finite; the converse is false (Exercise (i).) A finite cover is obviously locally finite.

XI.11.10.2. **Definition.** Let \( \mathcal{U} \) and \( \mathcal{V} \) be open covers of a topological space \((X, T)\). \( \mathcal{V} \) is a refinement of \( \mathcal{U} \) if every \( V \in \mathcal{V} \) is contained in some \( U \in \mathcal{U} \).

Note that \( \mathcal{U} \) itself, or any subcover of \( \mathcal{U} \), is a refinement of \( \mathcal{U} \). If \( X \) is a metric space, \( \epsilon, \delta > 0 \), and \( \mathcal{U} \) and \( \mathcal{V} \) are the collections of all balls of radius \( \epsilon \) and \( \delta \) respectively, then \( \mathcal{V} \) is a refinement of \( \mathcal{U} \) if \( \delta \leq \epsilon \). If \( \mathcal{V} \) is a refinement of \( \mathcal{U} \) and \( \mathcal{W} \) is a refinement of \( \mathcal{V} \), then \( \mathcal{W} \) is a refinement of \( \mathcal{U} \).
It can happen that two distinct open covers are each refinements of the other. For example, if \( \mathcal{U} \) is the set of (bounded) open intervals in \( \mathcal{R} \) of rational length, and \( \mathcal{V} \) the set of open intervals of irrational length, then \( \mathcal{U} \) and \( \mathcal{V} \) are each refinements of the other.

**XI.11.10.3. Definition.** Let \((X, \mathcal{T})\) be a topological space. Then \((X, \mathcal{T})\) is **paracompact** if it is Hausdorff and every open cover of \((X, \mathcal{T})\) has a locally finite refinement.

Note that a paracompact space is by definition Hausdorff. It is clear that any compact Hausdorff space is paracompact. But many noncompact spaces are also paracompact, as the next theorem shows (see also XI.11.10.19.). This result was proved by A. H. Stone [Sto48] (cf. XI.11.12.13.); the slick proof we give is due to M. E. Rudin [Rud69].

**XI.11.10.4. Theorem.** Every metrizable space is paracompact.

**Proof:** Let \( X \) be a metrizable space, and fix a metric \( \rho \) on \( X \). Let \( \mathcal{U} = \{U_i : i \in I\} \) be an open cover of \( X \). We may well-order the index set \( I \) (using AC). Define, inductively on \( n \), a collection of open sets \( V_{im} \) for \( i \in I \) as follows. Suppose \( V_{im} \) has been defined for all \( i \) and for all \( m < n \). Let \( V_{in} \) be the union of all open balls \( B_{2^{-n}}(x) \) such that

1. \( i \) is the smallest index such that \( x \in U_i \).
2. \( x \notin V_{jm} \) for any \( j \) and any \( m < n \).
3. \( B_{3 \cdot 2^{-n}}(x) \subseteq U_i \).

If \( x \in X \), then there is a smallest \( i \) such that \( x \in U_i \) and an \( n \) large enough that (3) holds. Then, by (2), \( x \in V_{im} \) for some \( m \leq n \). Thus \( \mathcal{V} = \{V_{in} : i \in I, n \in \mathbb{N}\} \) is an open cover of \( X \). Obviously \( V_{in} \subseteq U_i \) for each \( i \) and \( n \), so \( \mathcal{V} \) refines \( \mathcal{U} \).

To show that \( \mathcal{V} \) is locally finite, fix \( x \in X \), and let \( i \) be the smallest index for which \( x \in V_{im} \) for some \( n \). Fix the smallest \( n \) for which \( x \in V_{im} \). Choose \( k \) so that \( B_{2^{-k}}(x) \subseteq V_{im} \).

We first claim that if \( m \geq n + k \), then \( B_{2^{-n-k}}(x) \) intersects no \( V_{jm} \) for any \( j \in I \). Fix \( j \). Since \( m > n \), every ball of radius \( 2^{-m} \) used to define \( V_{jm} \) has center \( y \) outside \( V_{in} \), hence \( \rho(x, y) \geq 2^{-k} \) by the choice of \( k \). But \( m \geq k + 1 \) and \( n + k \geq k + 1 \), so \( 2^{-n-k} + 2^{-m} \leq 2^{-k} \), so \( B_{2^{-n-k}}(x) \cap B_{2^{-m}}(y) = \emptyset \).

Finally, we claim that if \( m < n + k \), then \( B_{2^{-n-k}}(x) \) intersects \( V_{jm} \) for at most one \( j \); this will complete the proof that \( B_{2^{-n-k}}(x) \) intersects at most \( n + k \) \( V_{jm} \), and, since \( x \in X \) is arbitrary, that \( \mathcal{V} \) is locally finite. Suppose \( p \in V_{jm} \) and \( q \in V_{jm} \) for \( j_1 < j_2 \); we will show that \( \rho(p, q) > 2^{-n-k+1} \), so both cannot be in \( B_{2^{-n-k}}(x) \). There are points \( y \) and \( z \) such that \( p \in B_{2^{-m}}(y) \subseteq V_{jm} \) and \( q \in B_{2^{-m}}(z) \subseteq V_{jm} \). By (3), \( B_{3 \cdot 2^{-m}}(y) \subseteq U_{j_1} \). But \( z \notin U_{j_1} \) by (2), so \( \rho(y, z) \geq 3 \cdot 2^{-m} \) and \( \rho(p, q) > 2^{-m} \geq 2^{-n-k+1} \).

The AC is used in the proof. It is known that this theorem cannot be proved in ZF+(DC).

Paracompact spaces share nice separation properties with compact spaces:

**XI.11.10.5. Proposition.** Every paracompact space is normal.

Thus there are Hausdorff spaces, even completely regular spaces, which are not paracompact. There are even locally compact completely normal spaces which are not paracompact (Exercise XI.11.12.10.1.)
XI.11.10.6. A subset of a paracompact space is not paracompact in general (e.g. a subset of a compact Hausdorff space need not be normal). Even an open subset of a paracompact space need not be paracompact (Exercise XI.11.12.10.). But:

XI.11.10.7. Proposition. A closed subspace of a paracompact space is paracompact.

The proof is essentially identical to the proof of XI.11.1.6. with “finite subcover” replaced by “locally finite refinement.”

XI.11.10.8. Proposition. Let $X$ be a (Hausdorff) topological space, and $\{X_i : i \in I\}$ a partition of $X$ into clopen subsets. Then $X$ is paracompact if and only if each $X_i$ is paracompact.

Proof: If $X$ is paracompact, then each $X_i$ is paracompact by XI.11.10.7. Conversely, suppose each $X_i$ is paracompact. If $U$ is an open cover of $X$, then, for each $i$, $\mathcal{U}_i = \{U \cap X_i\}$ is an open cover of $X_i$, which has a locally finite refinement $\mathcal{V}_i$. Then $\bigcup_{i \in I} \mathcal{V}_i$ is a locally finite open cover of $X$ which refines $\mathcal{U}$. 

Paracompact spaces are the natural setting for some constructions of algebraic topology.

Partitions of Unity

For analysis purposes, the most important feature of paracompact spaces is the existence of “partitions of unity.” These provide a systematic and controlled way to write a general real- or complex-valued function as a sum of functions of “small” support, and allow “local” phenomena to be stitched together globally.

XI.11.10.9. Definition. Let $(X, T)$ be a topological space. A weak partition of unity on $X$ is a collection of continuous functions $\{f_i : i \in I\}$ from $X$ to $[0, 1]$ such that

(i) For each $x \in X$, $\sum_{i \in I} f_i(x) = 1$.

The collection is a partition of unity if in addition

(ii) Each $x \in X$ has a neighborhood $U$ such that $f_i \neq 0$ on $U$ for only finitely many $i$.

The collection is a strong partition of unity if in addition

(iii) For each $i$, there is an $x \in X$ with $f_i(x) = 1$.

If $\mathcal{U}$ is an open cover of $(X, T)$, and $\{f_i : i \in I\}$ is a partition of unity on $(X, T)$, then $\{f_i\}$ is subordinate to $\mathcal{U}$ if for each $i \in I$ there is a $U \in \mathcal{U}$ such that $Z_{f_i}^c \subseteq U$, where for $f : X \to \mathbb{C}$ continuous $Z_f^c$ is the cozero set of $f$, i.e.

$$Z_f^c = \{x : f(x) \neq 0\}$$

(note that $Z_f^c$ is open for any continuous $f$). The closure of $Z_f^c$ is called the support of $f$.

XI.11.10.10. If $\{f_i : i \in I\}$ is a partition of unity, set $U_i = Z_{f_i}^c = \{x \in X : f_i(x) > 0\}$. Then $\{U_i : i \in I\}$ is a locally finite open cover of $X$. If $\mathcal{U}$ is an open cover of $X$, then the partition of unity is subordinate to $\mathcal{U}$ if and only if this locally finite open cover is a refinement of $\mathcal{U}$. If a partition of unity is subordinate to a refinement of $\mathcal{U}$, it is subordinate to $\mathcal{U}$.
XI.11.10.11. There is some lack of uniformity in this terminology; what we call “weak partitions of unity” are often called “partitions of unity.” Our term “strong partition of unity” is nonstandard, but will be useful; it is a desirable nondegeneracy requirement. It is equivalent to saying that \{U_i\} is a minimal open cover of \(X\), i.e. no proper subcollection covers \(X\).

XI.11.10.12. Example. Let \(X = \mathbb{R}\). ****

XI.11.10.13. Theorem. Let \((X, \mathcal{T})\) be a paracompact space. If \(\mathcal{U}\) is any open cover of \((X, \mathcal{T})\), then there is a strong partition of unity on \(X\) subordinate to \(\mathcal{U}\).

Proof: Passing to a refinement if necessary, we may assume that \(\mathcal{U}\) is locally finite. We first construct a smaller open cover by transfinite induction. If \(\mathcal{U}\) is countable (or has a countable locally finite refinement, e.g. if \(X\) is second countable, or locally compact and \(\alpha\)-compact), the transfinite induction can be simplified to an ordinary induction.

Choose a well ordering \(\{U_\alpha\}\) of \(\mathcal{U}\). We will inductively define an open cover \(\{V_\alpha\}\) such that \(\overline{V_\alpha} \subseteq U_\alpha\) for each \(\alpha\). Suppose that \(V_\beta\) has been defined for all \(\beta < \alpha\) so that \(\{V_\beta : \beta \leq \alpha\} \cup \{U_\beta : \beta > \alpha\}\) is an open cover of \(X\) (this assumption is vacuous if \(\alpha = 0\)). Set

\[
F_\alpha = X \setminus \left[ \left( \bigcup_{\beta < \alpha} V_\beta \right) \cup \left( \bigcup_{\beta > \alpha} U_\beta \right) \right].
\]

Then \(F_\alpha\) is a closed subset of \(U_\alpha\) since \(U_\alpha \cap F_\alpha = X\). Since \(X\) is normal (XI.11.10.5.), there is an open set \(V_\alpha\) with \(F_\alpha \subseteq V_\alpha\) and \(\overline{V_\alpha} \subseteq U_\alpha\). If \(F_\alpha = \emptyset\), choose \(V_\alpha = \emptyset\). We have \(X = (\bigcup_{\beta < \alpha} V_\beta) \cup (\bigcup_{\beta > \alpha} U_\beta) \cup V_\alpha\), so the inductive step is finished.

We claim \(V = \{V_\alpha\}\) is an open cover of \(X\). Suppose not; let \(x \in X \setminus (\bigcup_\alpha V_\alpha)\). Then \(x\) is in only finitely many \(U_\alpha\), so there is an \(\alpha\) such that \(x \notin \bigcup_{\beta \geq \alpha} U_\beta\). For this \(\alpha\), \(x \notin (\bigcup_{\beta < \alpha} V_\beta) \cup (\bigcup_{\beta > \alpha} U_\beta)\). But \(\{V_\beta : \beta < \alpha\} \cup \{U_\beta : \beta \geq \alpha\}\) is an open cover of \(X\), a contradiction. Note that since \(V_\alpha \subseteq U_\alpha\) for each \(\alpha\), \(\{U_\alpha\}\), \(V\) is locally finite. Also note that if \(x \in F_\alpha\), then \(x \notin V_\beta\) for \(\beta < \alpha\) by construction, and \(x \notin V_\beta\) for \(\beta > \alpha\) since \(V_\beta \subseteq U_\beta\) and \(x \notin U_\beta\) by construction. Thus \(x \notin V_\beta\) for any \(\beta \neq \alpha\).

Now repeat the construction using the \(V_\alpha\) in place of the \(U_\alpha\) to inductively define open sets \(W_\alpha\) and

\[
E_\alpha = X \setminus \left[ \left( \bigcup_{\beta < \alpha} W_\beta \right) \cup \left( \bigcup_{\beta > \alpha} V_\beta \right) \right]
\]
such that \(E_\alpha \subseteq W_\alpha\) and \(\overline{W_\alpha} \subseteq V_\alpha\), and \(\{W_\alpha\}\) is an open cover of \(X\). Note that \(F_\alpha \subseteq E_\alpha\) for each \(\alpha\).

Discard any \(\alpha\) for which \(V_\alpha = \emptyset\), i.e. \(F_\alpha = \emptyset\) (since \(F_\alpha \subseteq W_\alpha\), these are also the \(\alpha\) for which \(W_\alpha = \emptyset\)). For each remaining \(\alpha\) let \(g_\alpha\) be a continuous function from \(X\) to \([0,1]\) such that \(g_\alpha \equiv 1\) on \(\overline{W_\alpha}\) and \(g_\alpha \equiv 0\) on \(X \setminus V_\alpha\) (XI.7.6.12.). For \(x \in X\), set \(g(x) = \sum_\alpha g_\alpha(x)\). Since \(V\) is locally finite, only finitely many \(g_\alpha\) are nonzero in a neighborhood of \(x\), so \(g(x)\) is finite, and \(g\) is continuous at \(x\). Since \(g_\alpha(x) = 1\) for at least one \(\alpha\), \(g(x) \geq 1\). For each \(x \in F_\alpha\), \(g_\alpha(x) = g_\alpha(x)/g(x)\); then \(f_\alpha\) is continuous, \(0 \leq f_\alpha(x) \leq 1\) for all \(\alpha\) and \(x\), \(f_\alpha \equiv 0\) on \(X \setminus V_\alpha\), and for each \(x\), only finitely many \(f_\alpha(x)\) are nonzero, and \(\sum_\alpha f_\alpha(x) = 1\). For any \(\alpha\), if \(x \in F_\alpha\), then \(x \notin V_\beta\) for any \(\beta \neq \alpha\), so \(f_\alpha(x) = 1\), and thus the \(f_\alpha\) satisfy (iii). Thus \(\{f_\alpha\}\) is a partition of unity subordinate to \(\mathcal{U}\).

Note that this proof is a little more complex than the standard proof of existence of partitions of unity, because of condition (iii). What the proof really shows is that if \(\mathcal{U}\) is a locally finite cover of a normal space \(X\), then there is a strong partition of unity subordinate to \(\mathcal{U}\). Note that this proof requires the Axiom of Choice.
XI.11.10.14. Conversely, a partition of unity subordinate to an open cover gives a locally finite refinement of the cover, so if every open cover of \( X \) supports a partition of unity subordinate to it, \( X \) must be paracompact. There is a better result:

XI.11.10.15. Theorem. Let \( X \) be a \( T_1 \) space. If for every open cover \( U \) of \( X \) there is a weak partition of unity subordinate to \( U \), then \( X \) is paracompact.

Proof: [Eng89] We first note that \( X \) must be completely regular. If \( E \) is a closed set in \( X \) and \( p \notin E \), choose a partition of unity \( (f_i) \) subordinate to the open cover \( \{ X \setminus E, X \setminus \{p\} \} \). There is an \( i \) such that \( f_i = 0 \) on \( E \) and \( f_i(p) > 0 \), and it gives an Urysohn function for \( (E, p) \).

Now let \( U \) be an open cover of \( X \), and \( \{ f_i : i \in I \} \) a partition of unity subordinate to \( U \). If \( \phi \) is a strictly positive real-valued continuous function on \( X \), then for each \( x_0 \in X \) there is an open neighborhood \( U \) of \( x_0 \) and a finite set \( I_0 \subseteq I \) such that \( f_i(x) < \phi(x) \) for all \( x \in U \) and \( i \notin I_0 \) [Choose \( I_0 \) so that \( 1 - \sum_{i \in I_0} f_i(x_0) < \phi(x_0) \), and \( U = \{ x \in X : 1 - \sum_{i \in I_0} f_i(x) < \phi(x) \} \}.

For \( x \in X \), set \( \phi(x) = \sup_{i \in I} f_i(x) \). Then \( \phi \) is strictly positive, and since the supremum is a finite supremum in a neighborhood of each point, \( \phi \) is continuous. For each \( i \) set \( V_i = \left\{ x \in X : f_i(x) > \frac{1}{2} \phi(x) \right\} \).

Then \( \{ V_i : i \in I \} \) is a locally finite open cover of \( X \) which refines \( U \).

XI.11.10.16. Corollary. Let \( (X, \rho) \) be a metric space. Then for any \( \epsilon > 0 \) there is a strong partition of unity \( \{ f_i : i \in I \} \) such that the diameter of \( Z_{f_i}^\epsilon \) is less than \( \epsilon \) for all \( i \).

Proof: \( X \) is paracompact (XI.11.10.4.), and the balls of radius \( < \epsilon/2 \) form an open cover of \( X \).

Paracompactness and Local Compactness

XI.11.10.17. Proposition. Let \( X \) be a topological space, \( K \) a compact subset of \( X \), and \( \{ U_i : i \in I \} \) a locally finite open cover of \( X \). Then \( K \cap U_i \neq \emptyset \) for only finitely many \( i \).

Proof: For each \( x \in K \), there is an open neighborhood \( V_x \) of \( x \) such that \( V_x \cap U_i \neq \emptyset \) only for \( i \) in a finite subset \( F_x \) of \( I \). Then \( \{ V_x : x \in K \} \) is an open cover of \( K \), so there is a finite subcover \( \{ V_{x_1}, \ldots, V_{x_n} \} \). The set \( U_i \) can have nonempty intersection with \( K \) only if \( i \) is in the finite subset \( F_{x_1} \cup \cdots \cup F_{x_n} \) of \( I \).
XI.11.10.18. **Proposition.** Let $X$ be a (necessarily Hausdorff) topological space which is locally compact, connected, and paracompact. Then $X$ is $\sigma$-compact.

**Proof:** For each $x \in X$, let $U_x$ be an open neighborhood of $x$ with $\overline{U}_x$ compact. Let $\{V_i : i \in I\}$ be a locally finite refinement of $\{U_x : x \in X\}$. Then $V_i$ is compact for each $i$.

Fix $i_1 \in I$ with $V_{i_1}$ nonempty, and set $W_1 = V_{i_1}$, $K_1 = \overline{V}_{i_1}$. There are only finitely many $i$ for which $V_i \cap K_1 \neq \emptyset$; let $W_2$ be their union, and $K_2 = \overline{W}_2$. Then $K_2$ is also compact, and $K_1 \subseteq W_2$. Repeat the process to get $W_n$ and $K_n$ for each $n$. Set $W = \cup_n W_n = \cup_n K_n$. Then $W$ is clearly open in $X$ and $\sigma$-compact. If $i \in I$ and $V_i \cap W \neq \emptyset$, then $V_i \cap K_n \neq \emptyset$ for some $n$, so $V_i \subseteq W_{n+1} \subseteq W$. Thus $X \setminus W$ is the union of the $V_i$ which are disjoint from $W$, hence open; thus $W$ is clopen. Since $X$ is connected and $W \neq \emptyset$, $W = X$. 

Conversely:

XI.11.10.19. **Theorem.** A locally compact, $\sigma$-compact Hausdorff space is paracompact. In fact, any open cover of a locally compact, $\sigma$-compact Hausdorff space has a countable locally finite refinement.

XI.11.11. **Compactly Generated Spaces**

The great majority of topological spaces encountered naturally, if not compact, at least have their topology completely determined by their compact subsets.

**XI.11.11.1. Definition.** Let $X$ be a topological space. $X$ is **compactly generated** if any subset $U$ of $X$ is open (in $X$) if (and only if) $U \cap K$ is relatively open in $K$ for every compact subset $K$ of $X$.

In other words, $X$ has the weak topology generated by the inclusion maps of its compact subsets. A set $U \subseteq X$ such that $U \cap K$ is relatively open in $K$ for every compact $K \subseteq X$ is called a $k$-open set in $X$ (cf. XI.11.12.14.), and a compactly generated space is called a $k$-space in some references like [Dug78].

Any compact space is obviously compactly generated. More generally:

**XI.11.11.2. Proposition.** Every locally compact (Hausdorff) space is compactly generated.

**Proof:** Let $X$ be locally compact, and $U \subseteq X$ $k$-open. Let $x \in U$. Then $x$ has an open neighborhood $V_x$ in $X$ such that $\overline{V}_x$ (closure in $X$) is compact. $U \cap V_x$ is relatively open in $\overline{V}_x$, i.e. there is an open set $W_x$ in $X$ such that $U \cap \overline{V}_x = W_x \cap \overline{V}_x$. Then $V_x \cap W_x$ is an open neighborhood of $x$ in $X$ which is contained in $U$. Since $x \in U$ is arbitrary, $U$ is open in $X$. 

Perhaps more significantly, every first countable space is compactly generated. We need a result of independent interest:
XI.11.11.3. PROPOSITION. Let $X$ be a topological space and $(x_n)$ a sequence in $X$. If $x_n \to x$, then $K = \{x\} \cup \{x_n : n \in \mathbb{N}\}$ is a compact subset of $X$.

Proof: Let $U$ be an open cover of $K$. Then $x \in U$ for some $U \in U$. Since $x_n \to x$, $x_n \in U$ for all but finitely many $n$. Finitely many other sets in $U$ cover the $x_n$ which are not in $U$. 

XI.11.11.4. Caution: This statement is false in general if “sequence” is replaced by “net.”

XI.11.11.5. PROPOSITION. Every first countable topological space is compactly generated.

Proof: Let $X$ be first countable, and $U \subseteq X$ $k$-open. We show that $U^c$ (complement in $X$) is closed. Since $X$ is first countable, it suffices to show that if $(x_n)$ is a sequence in $U^c$ and $x_n \to x$, then $x \in U^c$. Let $K = \{x\} \cup \{x_n : n \in \mathbb{N}\}$; then $K$ is compact, so $U \cap K$ is relatively open, i.e. $U^c \cap K$ is relatively closed. Since $x_n \in U^c \cap K$ for all $n$ and $x_n \to x \in K$, $x \in U^c \subseteq U^c$.

XI.11.11.6. Example. There are topological spaces which are not compactly generated. For example, if $X$ is the space of XI.11.12.8. (b), and $U = \{\alpha\}$, then $U \cap K$ is relatively open for every compact (i.e. finite) subset $K$ of $X$. But $U$ is not open in $X$.

This example is countable, hence $\sigma$-compact, and completely normal.

XI.11.11.7. DEFINITION. Let $X$ and $Y$ be topological spaces, and $f : X \to Y$ a function. Then $f$ is continuous on compact sets if, for every compact $K \subseteq X$, $f|_K$ is a continuous function from $K$ to $Y$.

XI.11.11.8. Any continuous function is continuous on compact sets. The converse is false in general: if $X$ is the space of XI.11.12.8. (b) and $Y$ is any topological space, any function from $X$ to $Y$ is continuous on compact (i.e. finite) sets. However:

XI.11.11.9. PROPOSITION. Let $X$ and $Y$ be topological spaces, with $X$ compactly generated, and $f : X \to Y$ a function. If $f$ is continuous on compact sets, it is continuous.

Proof: Let $V$ be an open set in $Y$, and $U = f^{-1}(V)$. If $K$ is any compact subset of $X$, then $U \cap K = (f|_K)^{-1}(V)$ is relatively open in $K$ since $f|_K$ is continuous. Thus $U$ is open in $X$ since $X$ is compactly generated.

XI.11.12. Exercises

XI.11.12.1. (a) Show that any set with the indiscrete topology () is compact.

(b) Show that any set with the finite complement topology () is compact and $T_1$.

(c) Show that every subset of the examples in (a) and (b) is compact. Thus a compact subset of a compact $T_1$-space need not be closed.

1136
XI.11.12.2. The aim of this exercise is to show without AC that a compact subset of a Hausdorff space is closed. Let $X$ be a Hausdorff space, and $Y \subseteq X$. Suppose $Y$ is not closed, and let $p \in Y \setminus Y$. Fix a net $(y_i)$ in $Y$ converging to $p$ (XI.3.2.21.). Then the net $(y_i)$ cannot have a cluster point in $Y$ (XI.7.4.1.), so $Y$ is not compact (XI.11.2.1.).

XI.11.12.3. Here is an alternate proof of XI.11.3.10. which gives an explicit Lebesgue number. Let $(X, \rho)$ be a compact metric space, and $U$ an open cover of $X$. Let $\{U_1, \ldots, U_n\}$ be a finite subcover of $U$. For $x \in X$, define

$$f(x) = \frac{1}{n} \sum_{k=1}^{n} \rho(x, U_k^c).$$

Then $f$ is a continuous function from $X$ to $(0, \infty)$.

(a) Let $\delta$ be the minimum of $f$ on $X$. Show that $\delta > 0$.

(b) Show that $\delta$ is a Lebesgue number for $U$.

XI.11.12.4. Let $X = F([0,1], [0,1]) \cong [0,1]^\ast$ be the space of all functions from $[0,1]$ to itself, with the product topology (topology of pointwise convergence). Recall (XI.7.4.1.) that $X$ is separable but not first countable. $X$ is compact by Tikhonov’s Theorem (XI.11.3.10.) suffices, and Hausdorff.

Let $Y$ be the subset of $X$ consisting of all nondecreasing functions from $[0,1]$ to $[0,1]$, with the relative topology (also the topology of pointwise convergence). $Y$ is called the Helly space.

(a) Show that $Y$ is a closed subspace of $X$, hence compact and Hausdorff.

(b) Show that the set of nondecreasing step functions (XI.11.12.3) with only rational discontinuities, taking only rational values, is dense in $Y$. Conclude that $Y$ is separable.

(c) Show that $Y$ is first countable. [Use the fact (XI.11.12.3) that a nondecreasing function has only jump discontinuities, and only finitely many jumps of more than $\epsilon$ for any fixed $\epsilon > 0$.]

(d) For each $t \in [0,1]$, let $f_t$ be the function with $f_t(x) = 0$ for $x < t$, $f_t(x) = 1$ for $x > t$, and $f_t(t) = 1/2$. Show that $\{f_t : t \in [0,1]\}$ is an uncountable subset of $Y$ which is discrete in the relative topology. Conclude that $Y$ is not second countable, and hence not metrizable.

(e) Although $F([0,1], [0,1])$ has cardinality $2^{\aleph_0}$, show that $Y$ has cardinality $2^{\aleph_0}$ (cf. XI.7.8.5.).

XI.11.12.5. (a) Let $X$ be the unit square $[0,1] \times [0,1]$. Give $X$ the order topology (XI.11.12.3) from the lexicographic ordering. Show that $X$ is compact, Hausdorff, first countable, but not separable, hence not metrizable.

(b) Do the same as in (a) with the second copy of $[0,1]$ replaced by a three-point set with a total ordering. What if a two-point set is used instead?

XI.11.12.6. The Tikhonov Plank. Let $X$ be the set of ordinals less than or equal to the first uncountable ordinal $\omega_1$, and $Y$ the set of ordinals less than or equal to $\omega$. Then $X$ and $Y$ are compact Hausdorff spaces in the order topology. So $X \times Y$ is a compact Hausdorff space, hence normal. Let $Z = (X \times Y) \setminus (\omega_1, \omega)$ (Z is the “plank” $X \times Y$ with the “upper right-hand corner” removed).

(a) $Z$ is an open set in $X \times Y$, hence locally compact; $Z^f \cong X \times Y$. Show that $X \times Y$ is also $\beta Z$ (cf. XI.11.9.17.). $Z$ is pseudocompact (XI.11.9.17.).

(b) Let $T = \{ (\alpha, \omega) : \alpha < \omega_1 \}$ and $R = \{ (\omega_1, \alpha) : \alpha < \omega \}$ be the “top and right edges” of $Z$. Show that $T$ and $R$ are disjoint closed subsets of $Z$ which do not have disjoint neighborhoods in $Z$. Thus $Z$ is not normal.

1137
XI.11.12.7. Rational Sequence Topology. [SS95, #65] Put a topology on \( \mathbb{R} \) as follows. For each irrational \( x \in \mathbb{R} \), fix a sequence \((q_{x,k})\) of rational numbers converging to \( x \). Let \( \mathcal{B} \) be the following collection of subsets of \( \mathbb{R} \):

(i) \( \{q\} \) for each \( q \in \mathbb{Q} \).
(ii) \( \{x\} \cup \{q_{x,k} : k \geq n\} \) for each \( x \in \mathbb{J} = \mathbb{R} \setminus \mathbb{Q} \), \( n \in \mathbb{N} \).

(a) Show that if \( x, y \in \mathbb{J}, x \neq y \), then \( \{q_{x,k} : k \in \mathbb{N}\} \cap \{q_{y,k} : k \in \mathbb{N}\} \) is finite.
(b) Show that \( \mathcal{B} \) is a base for a topology \( \mathbb{Q} \) on \( \mathbb{R} \).
(c) Show that \( \mathbb{Q} \) is stronger than the ordinary topology on \( \mathbb{R} \). Thus \( (\mathbb{R}, \mathbb{Q}) \) is Hausdorff. Each rational number is an isolated point of \( (\mathbb{R}, \mathbb{Q}) \).
(d) \( (\mathbb{R}, \mathbb{Q}) \) is first countable.
(e) \( \mathbb{Q} \) is a dense open subset of \( (\mathbb{R}, \mathbb{Q}) \); thus \( (\mathbb{R}, \mathbb{Q}) \) is separable.
(f) Each set in \( \mathcal{B} \) is compact in \( (\mathbb{R}, \mathbb{Q}) \) (XI.11.11.3). Thus there is a base for \( \mathbb{Q} \) consisting of compact open sets, so \( (\mathbb{R}, \mathbb{Q}) \) is locally compact and \( ind(\mathbb{R}, \mathbb{Q}) = 0 \) (XI.21.3.2). In particular, \( (\mathbb{R}, \mathbb{Q}) \) is completely regular (XI.11.8.9) and totally disconnected (XI.13.6.1).
(g) \( \mathbb{J} \) is a discrete closed subset of \( (\mathbb{R}, \mathbb{Q}) \) of cardinality \( 2^{\aleph_0} \). Thus by XI.7.8.9., \( (\mathbb{R}, \mathbb{Q}) \) is not normal.
(h) By XI.13.6.5., \( dim(\mathbb{R}, \mathbb{Q}) > 0 \), i.e. \( (\mathbb{R}, \mathbb{Q}) \) is not zero-dimensional (XI.13.6.3.). Also, \( Ind(\mathbb{R}, \mathbb{Q}) > 0 \).

XI.11.12.8. Let \( \alpha \) be a point of \( \beta \mathbb{N} \setminus \mathbb{N} \), and let \( \mathcal{F} \) be the corresponding free ultrafilter on \( \mathbb{N} \). Set \( X = \mathbb{N} \cup \{\alpha\} \), with the relative topology. Then \( X \) is a countable completely regular (hence normal) space.

(a) Show that the open sets in \( X \) consist precisely of all subsets of \( \mathbb{N} \) and all sets of the form \( F \cup \{\alpha\} \), where \( F \in \mathcal{F} \). (The space \( X \) could have been defined without reference to \( \beta \mathbb{N} \) by just taking \( \alpha \) to be a point not in \( \mathbb{N} \) and defining the open sets this way for the free ultrafilter \( \mathcal{F} \).)

(b) Show that \( \alpha \) does not have a countable neighborhood base in \( X \). If \( x_n = \alpha \), then \( \alpha \) is a cluster point of the sequence \( (x_n) \), but no subsequence of \( (x_n) \) converges to \( \alpha \). [Every infinite subset of \( \mathbb{N} \) contains an infinite subset which is not in \( \mathcal{F} \).]

(c) Although every singleton subset of \( X \) is a \( G_\delta \), \( X \) is not first countable.

(d) \( X \) is not compact; in fact, the only compact subsets of \( X \) are the finite subsets. The identity map from \( \mathbb{N} \) to \( X \) extends to a homeomorphism from \( \beta \mathbb{N} \) to \( \beta X \) [GJ76, ].

If \( \gamma \) is another point of \( \beta \mathbb{N} \setminus \mathbb{N} \), corresponding to another free ultrafilter \( \mathcal{G} \) on \( \mathbb{N} \), and \( Y = \mathbb{N} \cup \{\gamma\} \), then \( Y \) is not necessarily homeomorphic to \( X \); in fact, there are \( 2^{2^{\aleph_0}} \) homeomorphism classes of such spaces since there are \( 2^{2^{\aleph_0}} \) points in \( \beta \mathbb{N} \setminus \mathbb{N} \) and only \( 2^{\aleph_0} \) permutations of \( \mathbb{N} \) [GJ76, ].

(e) More generally, consider a topological space \( X \) which is countably infinite and in which every point but one is an isolated point. Identify the isolated points with \( \mathbb{N} \) and call the nonisolated point \( \alpha \). Show that \( X \) is necessarily perfectly normal, and that there is a free filter \( \mathcal{F} \) on \( \mathbb{N} \) such that the open sets in \( X \) consist of all subsets of \( \mathbb{N} \) and all sets of the form \( F \cup \{\alpha\} \) for \( F \in \mathcal{F} \). There is a weakest such topology on \( \mathbb{N} \cup \{\alpha\} \), which is the one-point compactification of \( \mathbb{N} \); the maximal such topologies are the ones described in (i) (i.e. where \( \mathcal{F} \) is an ultrafilter). Every such topology is obtained by taking a compactification \( Y \) of \( \mathbb{N} \) and letting \( \alpha \) be a point of \( Y \setminus \mathbb{N} \). [Consider \( \beta X \).]

See XI.6.1.3. for a more dramatic example of a countable normal space which is not first countable.
XI.11.12.9. Let $X$ be a Hausdorff space. Show that $X$ is countably compact if and only if $X$ does not contain a closed subspace homeomorphic to $\mathbb{N}$ (i.e. a countably infinite closed subset whose relative topology is the discrete topology).

XI.11.12.10. Let $X$ be the space $\omega_1$ of ordinals less than the first uncountable ordinal ($\omega_1$), with the order topology ($\omega_1$). Let $\mathcal{U}$ be the open cover of $X$ consisting of all sets of the form $[0, \alpha)$ for $\alpha \in X$. Show that $\mathcal{U}$ has no locally finite refinement. [If $\mathcal{V}$ is a refinement of $\mathcal{U}$, show that there is a sequence $(V_n)$ of distinct sets in $\mathcal{V}$ and a strictly increasing sequence $(\alpha_n)$ in $X$ with $\alpha_n \in V_n$, and thus that $\mathcal{V}$ cannot be locally finite at $\alpha = \sup \alpha_n$.] Thus $X$ is locally compact and normal, even completely normal (XI.16.1.1.), but not paracompact.

XI.11.12.11. Show that a locally finite open cover of a separable topological space is countable. Conclude that every separable paracompact space is Lindelöf.

XI.11.12.12. Let $X$ be a topological space. Define a one-point compactification of $X$ as follows. Let $X_\infty$ be $X \cup \{\infty\}$, where $\infty$ is something not in $X$. A subset $U$ of $X_\infty$ is open if $U \subseteq X$ and $U$ is open in $X$, or if $\infty \in U$ and $X \setminus U$ is closed and compact in $X$ (regarding $\emptyset$ as a closed compact subset of $X$).

(a) Show that such open sets form a topology on $X_\infty$.
(b) Show that the relative topology on $X$ from $X_\infty$ is the original topology on $X$.
(c) Show that $X_\infty$ is compact.
(d) Show that $X$ is dense in $X_\infty$ if and only if $X$ is not compact.
(e) Show that $\{\infty\}$ is closed in $X_\infty$; hence $X_\infty$ is $T_1$ if and only if $X$ is $T_1$.
(f) Show that $X_\infty$ is Hausdorff if and only if $X$ is Hausdorff and locally compact.

XI.11.12.13. If $X$ is a topological space, $\mathcal{U}$ an open cover of $X$, and $x \in X$, the star of $\mathcal{U}$ at $x$ is

$$\text{st}_x(\mathcal{U}) = \{U \in \mathcal{U} : x \in U\}.$$ 

The collection $\text{st}(\mathcal{U}) = \{\text{st}_x(\mathcal{U}) : x \in X\}$ is an open cover of $X$, and $\mathcal{U}$ is a refinement of this cover. An open cover $\mathcal{V}$ star-refines $\mathcal{U}$ if $\text{st}(\mathcal{V})$ refines $\mathcal{U}$. This implies that $\mathcal{V}$ refines $\mathcal{U}$, but is much stronger. A topological space $X$ is fully normal if $X$ is $T_1$ and every open cover of $X$ has a star-refinement.

(a) Let $(X, \rho)$ be a metric space, and for $\epsilon > 0$ let $\mathcal{U}_\epsilon$ be the open cover of $X$ consisting of all open balls of radius $\epsilon$. Show that $\mathcal{U}_{\epsilon/2}$ is a star-refinement of $\mathcal{U}_\epsilon$ for any $\epsilon > 0$.
(b) Show that a fully normal space is normal. [If $A$ and $B$ are disjoint closed subsets of $X$, consider the open cover $\{X \setminus A, X \setminus B\}$.]
(c) [Sto48] Show that every metrizable space is fully normal.
(d) [Sto48] Show that a topological space is fully normal if and only if it is paracompact.

XI.11.12.14. Let $(X, T)$ be a topological space, and let $T^k$ be the set of $k$-open sets in $X$. Show that $T^k$ is a topology on $X$ which is stronger than $T$; $(X, T)$ is compactly generated if and only if $T^k = T$. Is $(X, T^k)$ always compactly generated? [Show that $T$ and $T^k$ have the same compact sets.]
XI.11.12.15. (D. E. Cohen; cf. [Dug78, XI.9.4]) (a) Show that every compactly generated Hausdorff space is a quotient of a locally compact space. [Consider the separated union of the compact subsets of the space.]

(b) Show that any quotient space of a locally compact space is compactly generated. [If $f : X \to Y$ is a quotient map with $X$ locally compact and $U$ is a $k$-open set in $Y$, show that $f^{-1}(U) \cap V$ is open in $X$ for any open set $V$ in $X$ with compact closure.]

(c) ([Dug78, XI.9.5], but note that a Hausdorff hypothesis is omitted there) Conclude that any quotient space of a compactly generated Hausdorff space is compactly generated.
XI.11.13. Dini’s Theorem

Dini’s Theorem is an assertion that certain sequences of functions which converge pointwise actually converge uniformly. Since uniform convergence has all sorts of nice consequences not shared by pointwise convergence, such a result is very useful.

XI.11.13.1. Definition. Let $X$ be a set, and $(f_n)$ a sequence of real-valued functions on $X$. The sequence $(f_n)$ is nondecreasing if, for each $x \in X$, the sequence $(f_n(x))$ of real numbers is nondecreasing. Similarly, $(f_n)$ is nonincreasing if $(f_n(x))$ is nonincreasing for all $x \in X$. The sequence $(f_n)$ is monotone if it is either a nondecreasing sequence or a nonincreasing sequence.

Note that it is the sequence itself that is nondecreasing (etc.), not the individual functions $f_n$. (Indeed, if $X$ is a general set, it does not make sense to say an individual function from $X$ to $\mathbb{R}$ is nondecreasing.) This notion of nondecreasing (etc.) is sometimes called pointwise nondecreasing (etc.) One can similarly define strictly increasing, strictly decreasing, and strictly monotone sequences of functions.

Here is the general version of Dini’s Theorem:

XI.11.13.2. Theorem. [Dini’s Theorem, General Version] Let $X$ be a compact topological space, and $(f_n)$ a nondecreasing sequence of real-valued continuous functions on $X$ converging pointwise to a function $f$ (not assumed to be continuous). If $g$ is a real-valued continuous function on $X$ satisfying $g(x) < f(x)$ for all $x \in X$, then there is an $n$ such that $g(x) < f_n(x)$ for all $x \in X$. The same statement holds with inequalities reversed if $(f_n)$ is nonincreasing.

Proof: For each $n$, set $K_n = \{x \in X : g(x) \geq f_n(x)\}$. Then $K_n$ is closed since $g$ and $f_n$ are continuous; hence $K_n$ is compact ( ). Since $f_n(x) \leq f_{n+1}(x)$ for all $x \in X$, we have $K_{n+1} \subseteq K_n$ for all $n$. Also, $\bigcap_n K_n = \emptyset$, since if $x \in \bigcap_n K_n$, we have $g(x) \geq f_n(x)$ for all $n$, hence $g(x) \geq f(x)$, contrary to hypothesis. Thus $(K_n)$ is a decreasing sequence of compact subsets of $X$ with empty intersection, so $K_n = \emptyset$ for some $n$ by ( ). The proof of the last statement is analogous. $

As a corollary, we get the commonly stated version of Dini’s Theorem:

XI.11.13.3. Corollary. [Dini’s Theorem, Common Version] Let $X$ be a compact topological space, and $(f_n)$ a monotone sequence of real-valued continuous functions on $X$ converging pointwise to a continuous function $f$. Then $f_n \to f$ uniformly.

Proof: Let $\epsilon > 0$. If $(f_n)$ is nondecreasing, set $g(x) = f(x) - \epsilon$ for all $x$; if $(f_n)$ is nonincreasing, set $g(x) = f(x) + \epsilon$. $

Combining this result with ( ), we obtain:

XI.11.13.4. Corollary. Let $X$ be a compact topological space, and $(f_n)$ a monotone sequence of real-valued continuous functions on $X$ converging pointwise to a function $f$. Then $f_n \to f$ uniformly if and only if $f$ is continuous.
XI.12. Baire Spaces and the Baire Category Theorem

The Baire Category Theorem is one of the two theorems from topology (along with Tikhonov’s Theorem) which are most heavily used in advanced analysis. Some of the conclusions are remarkable, and can seem like “magic” or “too good to be true.”

It will be most convenient to phrase the theorem by defining the class of topological spaces, the Baire spaces, for which the conclusion holds and then showing that complete metric spaces and locally compact Hausdorff spaces are Baire spaces.

XI.12.1. Meager Sets and Baire Spaces

XI.12.1.1. Definition. Let $X$ be a topological space. A subset $A$ of $X$ is nowhere dense in $X$ if the complement of $A$ is dense in $X$.

A subset of $X$ is meager in $X$, or first category in $X$, if it is a countable union of nowhere dense subsets of $X$. A subset is second category in $X$ if it is not meager (first category) in $X$. A subset $A$ is residual in $X$, or comeager or generic in $X$, if $A^c$ is meager in $X$.

The (nonempty) topological space $X$ is a Baire space if no nonempty open subset of $X$ is meager in $X$, i.e. if every residual subset of $X$ is dense in $X$.

The terms “nowhere dense”, “meager”, “first category”, “second category”, “residual”, “generic”, and “comeager” are relative terms, and the ambient topological space $X$ must be mentioned or at least be understood. The term “Baire space”, however, is intrinsic.

The terminology “first category” and “second category” is Baire’s, and is still frequently used, but we prefer the more descriptive terminology “meager” and “residual.” Note that “second category” and “residual” are not the same; in a Baire space, every residual set is second category, but there are many subsets $A$ for which both $A$ and $A^c$ are second category. (In a non-Baire space, even $\emptyset$ can be residual.)

XI.12.1.2. Examples. (i) The complement of a nowhere dense set is dense, but a set with dense complement is not nowhere dense in general. For example, $\mathbb{Q}$ is not nowhere dense in $\mathbb{R}$ (in fact, it is dense). A set $A \subseteq X$ is nowhere dense in $X$ if and only if $A$ is nowhere dense in $X$.

(ii) The complement of a dense open set in a topological space $X$ is nowhere dense in $X$. In particular, the Cantor set is nowhere dense in $\mathbb{R}$.

(iii) $\mathbb{R}$ is not nowhere dense in $\mathbb{R}$ (in fact, no nonempty topological space is nowhere dense in itself). But $\mathbb{R}$ is nowhere dense in $\mathbb{C}$. Thus it is important to specify the ambient space.

(iv) If $X$ is $T_1$ and $x \in X$ is not an isolated point, then $\{x\}$ is nowhere dense in $X$. Thus a countable subset of $X$ containing no isolated points is meager. In particular, every countable subset of $\mathbb{R}$ (e.g. $\mathbb{Q}$) is meager.

(v) Any subset of a nowhere dense set is nowhere dense. Any subset of a meager set is meager. So any subset containing a dense $G_δ$ is residual. The converse is true in a Baire space (XI.12.1.4.).

(vi) Any discrete space is a Baire space, since a nonempty subset of a discrete space cannot be nowhere dense. On the other hand, by (iv), no countable $T_1$ space without isolated points is a Baire space.

(vii) A countable union of meager sets in $X$ is meager in $X$ (cf. II.5.4.8.).
XI.12.1.3.  PROPOSITION. Let $X$ be a topological space. The following are equivalent:

(i) $X$ is a Baire space.

(ii) The intersection of any sequence of dense open subsets of $X$ is dense.

(iii) The intersection of any finite or countable collection of dense $G_{δ}$’s is a dense $G_{δ}$.

PROOF: (i) ⇒ (ii): If $X$ is a Baire space and $(A_n)$ is a sequence of dense open subsets of $X$, then $A_n^c$ is closed and nowhere dense in $X$, so $\bigcup_{n=1}^{\infty}A_n^c$ is meager and $\bigcap_{n=1}^{\infty}A_n$ is residual, hence dense.

(ii) ⇒ (i): Suppose the intersection of any sequence of dense open subsets of $X$ is dense, and let $A$ be a residual subset of $X$. Then $A^c = \bigcup_{n=1}^{\infty}B_n$, where $B_n$ is nowhere dense in $X$. If $A_n$ is the complement of $B_n$, then $A_n$ is a dense open set in $X$, and $\bigcap_{n=1}^{\infty}A_n \subseteq A$. By assumption, $\bigcap_{n=1}^{\infty}A_n$ is dense in $X$, so $A$ is also dense in $X$.

For (iii), the only issue is whether the intersection is dense since in any topological space an intersection of countably many $G_{δ}$’s is a $G_{δ}$.

XI.12.1.4.  COROLLARY. Let $X$ be a Baire space. A subset $A$ of $X$ is residual in $X$ if and only if $A$ contains a dense $G_{δ}$.

PROOF: One direction was observed in XI.12.1.2 (v). Conversely, let $A$ be a residual subset of a Baire space $X$. Then $A^c = \bigcup_{n=1}^{\infty}B_n$, where $B_n$ is nowhere dense in $X$. If $A_n$ is the complement of $B_n$, then $A_n$ is a dense open set in $X$, and $\bigcap_{n=1}^{\infty}A_n$ is a dense $G_{δ}$ in $X$ which is contained in $A$.

XI.12.1.5.  If $X$ is a $T_1$ Baire space and $A$ is a countable dense subset of $X$ containing no isolated points, then $A$ is meager (XI.12.1.2), so cannot be a $G_{δ}$ since if it were $A^c$ would be an $F_{σ}$, hence also meager (Exercise XI.12.5.1). In particular, $\mathbb{Q}$ is not a $G_{δ}$ in $\mathbb{R}$.

XI.12.2.  The Baire Category Theorem

It turns out that many standard topological spaces are Baire spaces. The next theorem is the principal result of this section. It is named after R. Baire, who proved it in 1899 [3] (Baire actually only stated it for $\mathbb{R}^n$, but essentially the same proof gives the general result), and is called the “Baire Category Theorem” because of his category terminology (note that it has nothing to do with category theory []). The case of $\mathbb{R}$ was actually proved earlier (1897) by W. Osgood [Osg97], who used different terminology; Osgood also independently obtained the result for $\mathbb{R}^n$ [Osg00].
XI.12.2.1. **Theorem.** [Baire Category Theorem] (i) Every completely metrizable space (topological space whose topology is given by a complete metric) is a Baire space.

(ii) Every locally compact Hausdorff space is a Baire space.

The first sentence is often stated, “Every complete metric space is a Baire space.” But note that the property of being a Baire space is a purely topological property, not a metric property. Thus, for example, \((0, 1)\) is a Baire space since it is homeomorphic to \(\mathbb{R}\), even though the usual metric on \((0, 1)\) is not complete. See XI.10.3. for more detailed results about completely metrizable spaces.

**Proof:** The proofs of both parts are nearly identical. In both cases, we must show that if \((U_n)\) is a sequence of dense open subsets of \(X\), then \(\bigcap_{n=1}^{\infty} U_n\) is also dense in \(X\), i.e. if \(V\) is a nonempty open set in \(X\), then \(V \cap \bigcap_{n=1}^{\infty} U_n\) is nonempty.

First suppose the topology on \(X\) is given by a complete metric \(\rho\). Since \(U_1\) is dense in \(X\), \(V \cap U_1\) is nonempty. Choose \(x_1 \in V \cap U_1\), and let \(B_1\) be an open ball of radius \(\leq 1\) centered at \(x_1\) whose closure is contained in \(V \cap U_1\) (there exists such a ball since \(V \cap U_1\) is open in \(X\)). Since \(U_2\) is dense in \(X\), there is an \(x_2 \in B_1 \cap U_2\). Let \(B_2\) be an open ball centered at \(x_2\) of radius \(\leq 1/2\) whose closure is contained in \(B_1 \cap U_2\). Inductively define points \(x_n \in B_{n-1} \cap U_n\) and balls \(B_n\) centered at \(x_n\) of radius \(\leq 1/n\) whose closure is contained in \(B_{n-1} \cap U_n\). By construction, \(x_m\) is in the closure of \(B_n\) for all \(m > n\), and thus \(\rho(x_n, x_m) \leq 1/n\) for all \(m > n\). Thus \((x_n)\) is a Cauchy sequence in \((X, \rho)\). Since \(\rho\) is complete, \((x_n)\) converges to some \(x \in X\). We have that \(x\) is in the closure of \(B_n\) for each \(n\); thus \(x \in \bigcap_{n=1}^{\infty} U_n\). And \(x \in B_1 \subseteq V\). Thus \(V \cap \bigcap_{n=1}^{\infty} U_n\) is nonempty.

The proof of (ii) is similar, but easier. Choose \(x_1 \in V \cap U_1\) as before. There is an open neighborhood \(B_1\) of \(x_1\) with compact closure contained in \(V \cap U_1\). Inductively choose points \(x_n\) and open neighborhoods \(B_n\) such that the closure of \(B_n\) is compact and contained in \(B_{n-1} \cap U_n\). Then \((B_n)\) is a decreasing sequence of nonempty compact subsets of \(X\) and hence there is an \(x \in \bigcap_{n=1}^{\infty} B_n\). This \(x\) is in \(U_n\) for all \(n\), and also in \(V\).

XI.12.2.2. Note that the proof of this theorem uses the Axiom of Choice to choose the \(x_n\). In fact, only a sequence of choices must be made, each dependent on the previous ones, so the Axiom of Dependent Choice (DC) suffices. Actually, the Baire Category Theorem is logically equivalent to DC (Exercise II.6.4.7.). But if \(X\) is separable, no form of choice is necessary to prove the Baire Category Theorem (Exercise ())). (It can be proved without any form of choice that a separable metrizable space is second countable (); however, the Countable Axiom of Choice is needed to prove that a second countable topological space, even a second countable metrizable space, is separable, so Countable AC is needed to prove the Baire Category Theorem for second countable complete metrizable spaces.)

XI.12.2.3. There are many Baire spaces, even metrizable ones, which do not satisfy the hypotheses of XI.12.2.1. In particular, a metrizable Baire space need not be completely metrizable (XI.12.5.4.).

XI.12.2.4. One of the important ways the Baire Category Theorem is used in analysis is to prove the existence of objects with a certain (usually pathological) property by showing that the set of objects without the property is a meager subset of some complete metric space. Such proofs are often regarded as “nonconstructive” or “pure existence proofs” which prove existence without exhibiting a specific example. The claim is made in [Oxt71] that this point of view is incorrect (at least in the separable case), and that the proof of the Baire Category Theorem is an explicit construction which in applications gives a
concrete example. This claim is correct, at least in principle: the proof of the theorem does give a recipe for constructing an example. But the actual example is the limit of a Cauchy sequence which may be almost impossible to describe explicitly. Even the terms in the Cauchy sequence can be quite difficult to explicitly describe in many cases. (This point is acknowledged in \([Oxt71]\).)

Here is an interesting sample application of Baire Category. An explicit example of a continuous function which is not differentiable anywhere is given in \(\) . Here is a simpler argument which shows that such functions exist; in fact, in a category sense almost all continuous functions fail to be differentiable anywhere!

**XI.12.2.5.** Theorem. Let \([a, b]\) be a closed bounded interval in \(\mathbb{R}\), and \(C([a, b])\) the real-valued continuous functions on \([a, b]\), with supremum norm. If \(D\) is the set of functions in \(C([a, b])\) which have a (finite) derivative at at least one point of \([a, b]\), then \(D\) is a meager subset of \(C([a, b])\). Hence there is a residual set of functions on \([a, b]\) which are continuous everywhere but differentiable nowhere.

This argument in a sense shows more than the one in \(\) , and in a sense shows less: this argument shows not only that continuous nondifferentiable functions exist, but that such functions are “typical” (despite one’s impression from Calculus) and form a set which is uniformly very dense. However, it is hard to identify a specific example of such a function from the Baire Category argument.

There are some rather astounding variations on this result, such as the following theorem of J. MARCINKIEWICZ (1935).

**XI.12.2.6.** Theorem. Let \([a, b]\) be a closed bounded interval, and \((h_n)\) a sequence of nonzero real numbers converging to zero. Then there is a continuous function \(f : [a, b] \rightarrow \mathbb{R}\) with the following property: if \(\phi : [a, b] \rightarrow \mathbb{R}\) is any measurable function, then there is a subsequence \((h_{k_n})\) of \((h_n)\) such that

\[
\lim_{n \to \infty} \frac{f(x + h_{k_n}) - f(x)}{h_{k_n}} = \phi(x)
\]

for almost all \(x \in [a, b]\). (Thus \(f\) is differentiable at most on a set of measure zero.) In fact, there is a residual set of such \(f\).

See \([Str81, p. 316-317]\) for an outline of the proof.

**XI.12.3. Meager Sets vs. Null Sets**

Meager sets have a close formal resemblance to sets of Lebesgue measure zero. Both notions are reasonable interpretations of what it should mean to be “negligibly small.” The relationship is more than just an analogy; the theme of the book \([Oxt71]\) is the close connections between the concepts.

The two notions of smallness are not the same, however, as the next rather dramatic example shows:
XI.12.3.1. Example. There is a dense $G_δ$ in $\mathbb{R}$ of Lebesgue measure 0. Thus $\mathbb{R}$ can be partitioned into a disjoint union of a meager set and a set of measure 0. This follows immediately from () and the fact that $\mathbb{Q}$ is a dense subset of $\mathbb{R}$ of Lebesgue measure 0, but we give an explicit example.

Let $(r_n)$ be an enumeration of $\mathbb{Q}$. Let $I_{n,k}$ be the open interval of length $\frac{1}{2^n k}$ centered at $r_n$, and for each $k$ let $A_k = \bigcup_{n=1}^{\infty} I_{n,k}$. Then $A_k$ is open, and dense since $\mathbb{Q} \subset A_k$; and

$$\lambda(A_k) \leq \sum_{n=1}^{\infty} \frac{1}{2^n k} = \frac{1}{k}.$$ 

Thus $A = \cap_{k=1}^{\infty} A_k$ is a dense $G_δ$ in $\mathbb{R}$ of Lebesgue measure 0.

An even more dramatic example is possible: there is a dense $G_δ$ in $\mathbb{R}$ which not only has Lebesgue measure 0, but Hausdorff dimension 0, i.e. $s$-dimensional Hausdorff measure 0 for all $s > 0$. Such an example can be readily constructed by a refinement of the argument in XI.12.3.1., but there is a natural and significant example: the Liouville numbers. See ()

XI.12.4. The Baire Property

XI.12.4.1. Definition. Let $X$ be a topological space. If $A, B \subseteq X$, write $A =^* B$ if $A \triangle B$ is a meager subset of $X$ (i.e. “$A$ and $B$ differ by a meager set.”)

A subset $A$ of $X$ has the Baire property if $A =^* U$ for some open set $U$ in $X$, i.e. “$A$ differs from an open set by a meager set.”

XI.12.4.2. It is easy to check that $=^*$ is an equivalence relation on subsets of $X$. Note that the Baire property is a relative property of $A$ in $X$, not an intrinsic property of $A$. For example, every open set and every meager set in $X$ has the Baire property. In particular, there is no direct connection between having the Baire property and being a Baire space.

XI.12.4.3. Proposition. Let $X$ be a topological space. Then the collection of all subsets of $X$ with the Baire property is a $\sigma$-algebra, and is the $\sigma$-algebra generated by the open sets and the meager sets. In particular, every Borel set in $X$ has the Baire property.

Proof: If $U$ is open, then $U =^* \bar{U}$ since $\bar{U} \setminus U$ is nowhere dense, hence meager. Similarly, if $E$ is closed, $E =^* E^c$.

If $A =^* U$ with $U$ open, then $A^c \triangle U^c = A \triangle U$ is meager, so $A^c =^* U^c =^* (U^c)^0$, and $A^c$ has the Baire property. And if $A_n =^* U_n$ for all $n$ with each $U_n$ open, then

$$\left( \bigcup_n A_n \right) \triangle \left( \bigcup_n U_n \right) \subseteq \bigcup_n (A_n \triangle U_n)$$

and a countable union of meager sets is meager, hence $\cup_n A_n =^* \cup_n U_n$, so $\cup_n A_n$ has the Baire property. Thus the collection $\mathcal{D}$ of subsets with the Baire property is a $\sigma$-algebra.

If $\mathcal{A}$ is the $\sigma$-algebra generated by the open sets and the meager sets, then $\mathcal{A} \subseteq \mathcal{D}$. Conversely, if $A \in \mathcal{D}$, then $A \triangle U = \emptyset$ is meager for some open set $U$, and $A = M \triangle U$, so $A \in \mathcal{A}$. Thus $\mathcal{A} = \mathcal{D}$. 

1146
XI.12.4.4. COROLLARY. Let \( X \) be a topological space, and \( A \subseteq X \). The following are equivalent:

(i) \( A \) has the Baire property.

(ii) \( A = E \cup M \) with \( E \) a \( G_\delta \) and \( M \) meager.

(iii) \( A = F \setminus M \) with \( F \) an \( F_\sigma \) and \( M \) meager.

Proof: (ii) \( \Rightarrow \) (i) and (iii) \( \Rightarrow \) (i) by XI.12.4.3. For (i) \( \Rightarrow \) (ii), suppose \( A \triangle U = N \) for a meager set \( N \). Then there is a meager \( F_\sigma F \) with \( N \subseteq F \). Then \( E = U \setminus F \) is a \( G_\delta \), \( E \subseteq A \), and \( A \setminus E \subseteq N \) is meager. For (i) \( \Rightarrow \) (iii), apply (i) \( \Rightarrow \) (ii) for \( A^c \).

Baire Measurable Functions

XI.12.4.5. Definition. Let \( X \) and \( Y \) be topological spaces, and \( f : X \to Y \) a function. Then \( f \) is Baire measurable if \( f^{-1}(U) \) has the Baire property for every open set \( U \) in \( Y \).

XI.12.4.6. It follows from () that if \( f \) is Baire measurable, then \( f^{-1}(B) \) has the Baire property for any Borel set \( B \) in \( Y \). In particular, if \( f \) is Borel measurable, it is Baire measurable. (Of course, if \( f \) is continuous it is Borel measurable, hence Baire measurable.)

The next result is related to Luzin’s Theorem (), at least in spirit. It holds for any topological spaces, but is most interesting if \( X \) is a Baire space:

XI.12.4.7. Theorem. Let \( X \) and \( Y \) be topological spaces, and \( f : X \to Y \) a Baire measurable function. Then there is a residual \( G_\delta E \) in \( X \) such that \( f|_E \) is continuous.

XI.12.4.8. The set \( E \) is necessarily dense in \( X \) if \( X \) is a Baire space. Note that the conclusion is that the restriction of \( f \) to \( E \) is continuous, not that \( f \) is continuous on \( E \). In fact, \( f \) need not be continuous anywhere, even if \( X \) is a Baire space (e.g. \( \chi_Q : \mathbb{R} \to \mathbb{R} \); in this case we may take \( E = \emptyset \)).

In contrast, here is an important result in which actual continuity is obtained on a residual set.

XI.12.4.9. Theorem. Let \( X, Y, \) and \( Z \) be metrizable topological spaces, and \( f : X \times Y \to Z \) a function. Suppose \( f \) is separately continuous, i.e. for each \( x \in X \) the function \( f_x : Y \to Z \) given by \( f_x(y) = f(x, y) \) is continuous, and for each \( y \in Y \) the function \( f^y : X \to Z \) given by \( f^y(x) = f(x, y) \) is continuous. Then there is a residual \( G_\delta E \) in \( X \times Y \), with \( E^y = \{ x \in X : (x, y) \in E \} \) residual in \( X \) for each \( y \in Y \), such that \( f \) is continuous (i.e. jointly continuous) at all points of \( E \).

XI.12.4.10. In particular, if \( X \times Y \) is a metrizable Baire space, e.g. if \( X \) and \( Y \) are completely metrizable, then a separately continuous function from \( X \times Y \) to any metric space \( Z \) is jointly continuous on a dense \( G_\delta \).
XI.12.5. Exercises

XI.12.5.1. Let $X$ be a complete metric space, or a locally compact Hausdorff space, without isolated points. Show that no countable dense subset of $X$ is a $G_{δ}$. (Hence $\mathbb{Q}$ is not a $G_{δ}$ in $\mathbb{R}$.) In particular, $X$ is uncountable. [The complement of a dense $G_{δ}$ is meager.]

XI.12.5.2. (a) Show that a dense $G_{δ}$ in a Baire space is a Baire space in the relative topology.
(b) Use XI.11.9.13., XI.10.3.4., XI.12.2.1.(ii), and (a) to give an alternate proof of XI.12.2.1.(i).

The word “dense” cannot be eliminated in (a) (XI.12.5.4.).

XI.12.5.3. Prove the following strong converse to XI.12.5.2 (a): If a topological space $X$ contains a dense subspace which is a Baire space in the relative topology, then $X$ is a Baire space. [Use XI.2.3.9..]

XI.12.5.4. Let $X = (\mathbb{R} \times [\mathbb{R} \setminus \{0\}]) \cup (\mathbb{Q} \times \{0\})$, with the relative topology from $\mathbb{R}^2$.

(a) Apply XI.12.5.3. to show that $X$ is a Baire space.

(b) $X$ has a closed subspace, a $G_{δ}$, homeomorphic to $\mathbb{Q}$, which is not a Baire space (XI.12.1.2.(vi)). Thus a closed subspace of a Baire space, even a closed $G_{δ}$, is not necessarily a Baire space.

(c) $X$ is metrizable. But a closed subspace of a completely metrizable space is completely metrizable and hence a Baire space, so $X$ is not completely metrizable.

Simpler non-Hausdorff examples of non-Baire closed subspaces of Baire spaces can be given.

XI.12.5.5. [Boa96, p. 59–61] Use the Baire Category Theorem to show that on any closed bounded interval $[a,b]$, there is a continuous function which is not monotone on any subinterval, and in fact a residual set of such functions in $C([a,b])$. [Use subintervals with rational endpoints.] See XI.12.5.8. for a closely related result.

XI.12.5.6. [CSB54] Use the Baire Category Theorem to show that if $f$ is a $C^∞$-function on an interval $[a,b]$, and for every $x \in [a,b]$ there is an $n$ such that $f^{(n)}(x) = 0$, then $f$ is a polynomial, i.e. there is an $n$ such that $f^{(n)}(x) = 0$ for all $x \in [a,b]$.

XI.12.5.7. [Wei76] Let $[a,b]$ be a closed bounded interval, and $D_{0}'([a,b])$ the set of bounded functions on $[a,b]$ which are derivatives of differentiable functions (i.e. bounded functions on $[a,b]$ which have antiderivatives).

(a) Show that $D_{0}'([a,b])$ is a closed subspace of the Banach space of bounded real-valued functions on $[a,b]$ with the supremum norm, hence also a Banach space (cf. V.8.5.7.).

(b) Show that if $f \in D_{0}'([a,b])$, then $Z_f = \{ x \in [a,b] : f(x) = 0 \}$ is a $G_{δ}$ (cf. V.12.1.8., (i)).

(c) Let $D_{0}''([a,b])$ be the set of all $f \in D_{0}'([a,b])$ for which $Z_f$ is dense in $[a,b]$. Use (b) and the Baire Category Theorem to show that $D_{0}''([a,b])$ is a closed subspace of $D_{0}'([a,b])$, and thus also a Banach space.

(d) There is a $g \in D_{0}''([a,b])$ which is not identically zero (XI.22.7.3.). By translating and scaling $g$, show that for any $x \in [a,b]$ and $\epsilon > 0$ there is an $h \in D_{0}''([a,b])$ with $\|h\| < \epsilon$ and $h(x) \neq 0$.

(e) If $I$ is a closed (nondegenerate) subinterval of $[a,b]$, let

$$E^I = \{ f \in D_{0}'([a,b]) : f(x) \geq 0 \text{ for all } x \in I \} .$$
Show that $E_I^+$ is a closed subset of $D'_0([a,b])$. Use (d) to show that $E_I^+$ is nowhere dense in $D'_0([a,b])$. Do the same for $E_I^- = \{ f \in D'_0([a,b]) : f(x) \leq 0 \text{ for all } x \in I \}$. 

(f) Let $E$ be the set of all bounded derivatives of differentiable functions on $[a, b]$ which are monotone on some (nondegenerate) subinterval. Show that $E$ is a meager subset of $D'_0([a,b])$. [Let $(I_n)$ be an enumeration of the closed subintervals of $[a,b]$ with rational endpoints. Show that $E = \bigcup_n (E_{I_n}^+ \cup E_{I_n}^-)$.

XI.12.5.8. This problem is a continuation of XI.12.5.7., using the same notation.

(a) Let $\mathcal{D}_0([a,b])$ be the set of differentiable functions on $[a,b]$ whose derivative is in $D'_0([a,b])$. Show that $\mathcal{D}_0([a,b])$ is a subalgebra of $C_R([a,b])$. (The Baire Category Theorem is needed to prove it is closed under multiplication.)

(b) Show that $\mathcal{D}_0([a,b])$ separates points of $[a,b]$. Thus by the Stone-Weierstrass Theorem (XV.8.2.2.) $\mathcal{D}_0([a,b])$ is uniformly dense in $C_R([a,b])$.

(c) Give a direct proof of the result of (b) without using the Stone-Weierstrass Theorem, using:

**Lemma.** For every $\epsilon > 0$ there is an $f \in \mathcal{D}_0((0,1))$ with the following properties:

(i) $f(0) = 0$, $f(1) = 1$.

(ii) $f'(0) = f'(1) = 0$.

(iii) For every $x \in [0,1]$, $f'(x) < 1 + \epsilon$.

(iv) For every $x \in [0,1]$, $x - \epsilon < f(x) < x + \epsilon$.

(d) To prove the Lemma, let $g$ be an element of $\mathcal{D}_0([a,b])$ for some $[a,b]$ with $1 < \sup g' < 1 + \epsilon$, and let $c, d \in [a,b], c < d$, with $\frac{g(d) - g(c)}{d-c} > 1$. Replace $c$ and $d$ with nearby places where $g' = 0$. Rescale $[c,d]$ to $[0,1]$. Show that (iv) follows from (i), (iii), and the MVT.

(e) By translating and scaling the $f$ of the lemma, show that any piecewise-linear continuous function on any closed bounded interval can be uniformly approximated by functions in $\mathcal{D}_0([a,b])$. 

1149
XI.13. Connectedness

The notion of a connected topological space, or of a connected subset of a topological space, is fundamental in topology and especially in applications in analysis.

The idea is simple. If a topological space can be divided into two disconnected pieces, each piece must be a closed set since no point of one of the sets can be a limit point of the other. Since each set is the complement of the other, the sets must be open as well as closed. Conversely, if a space can be partitioned into two subsets which are each both open and closed, these sets disconnect the space.

A second, more restrictive, notion of connectedness which is perhaps more intuitively appealing is that of path-connectedness, that any two points of the space can be connected by a path. Path-connectedness implies connectedness (once we know that intervals are connected), but not conversely. Both types of connectedness are useful in topology and analysis.

XI.13.1. Connected Spaces

XI.13.1.1. A subset $U$ of a topological space $X$ is clopen if $U$ is both open and closed. ($X \setminus U$ is then also clopen.) In any topological space $X$, $\emptyset$ and $X$ are clopen; in many interesting topological spaces, these are the only clopen subsets.

Finite unions, finite intersections, and set-theoretic differences of clopen sets are clopen.

The term “clopen” is regarded as a language abomination in some circles and is somewhat unfashionable; but the term is useful and there is no good alternative, so we will use it.

XI.13.1.2. DEFINITION. Let $X$ be a topological space. A separation of $X$ is a decomposition $X = U \cup V$ of $X$ into two nonempty disjoint clopen subsets. $X$ is connected if it has no separation, i.e. if the only clopen subsets of $X$ are $\emptyset$ and $X$.

Connectedness is obviously a topological property of a space, preserved under homeomorphisms.

XI.13.1.3. EXAMPLES. (i) Any space with the indiscrete topology is connected.
(ii) A space with the discrete topology and more than one point is not connected.
(iii) The Sierpiński space ($\text{Sierpiński space}$) is connected.
(iv) An infinite set with the finite complement topology ($\text{finite complement topology}$) is connected.
(v) $\mathbb{Q}$ with the relative topology from $\mathbb{R}$ is not connected: for example, $U = \{x \in \mathbb{Q} : x < \sqrt{2}\}$ and $V = \{x \in \mathbb{Q} : x > \sqrt{2}\}$ form a separation of $\mathbb{Q}$.

Examples (i), (iii), and (iv) show that connectedness can behave strangely in non-Hausdorff spaces. The notion is most interesting for spaces with nice separation properties, although even there some of the behavior can be surprising.

By far the most important example of a connected topological space is:
XI.13.1.4. Theorem. $\mathbb{R}$ (with its usual topology) is connected.

Proof: Let $\mathbb{R} = U \cup V$ be a separation, and let $a \in U$, $b \in V$. We may assume $a < b$. Set

$$Y = \{x \in [a, b] : x \in U\}$$

and set $c = \sup(Y)$. We have $c < b$ since $(b - \epsilon, b] \subseteq V$ for some $\epsilon > 0$. Either $c \in U$ or $c \in V$. If $c \in U$, then $[c,c + \epsilon) \subseteq U$ for some $\epsilon > 0$, contradicting that $c = \sup(Y)$. If $c \in V$, then $(c - \epsilon, c] \subseteq V$ for some $\epsilon > 0$, so $c - \epsilon$ would be an upper bound for $Y$, another contradiction. So no separation exists.

Connected Subsets of a Space

XI.13.1.5. If $X$ is a topological space and $E \subseteq X$, we say $E$ is a connected subset of $X$ if $E$ is a connected space in the relative topology. This can be phrased in several equivalent ways:

XI.13.1.6. Proposition. Let $X$ be a topological space and $E \subseteq X$. The following are equivalent:

(i) $E$ is not connected.

(ii) There are open sets $U$ and $V$ in $X$ with $U \cap E$ and $V \cap E$ nonempty, such that $E \subseteq U \cup V$ and $U \cap V \cap E = \emptyset$.

(iii) There are nonempty subsets $A$ and $B$ of $E$ such that $E = A \cup B$ and $\overline{A} \cap B = A \cap \overline{B} = \emptyset$.

The proof is an easy exercise, and is left to the reader. A pair $\{U, V\}$ of open sets as in (ii) is called a separation of $E$ in $X$ (cf. XI.10.1.1.).

XI.13.1.7. Corollary. A subset of $\mathbb{R}$ is connected if and only if it is an interval or a singleton.

An interval is connected by (ii) and essentially the same argument as in the proof of XI.13.1.4. A subset of $\mathbb{R}$ which is not an interval or singleton can be separated as in XI.13.1.3.(v).

XI.13.1.8. An interesting question is whether, if $A$ and $B$ are subsets of a topological space $X$ which are separated as in (iii), then they have disjoint open neighborhoods (essentially, whether the $U$ and $V$ in (ii) can be chosen disjoint). The answer is: not always (Exercise XI.13.9.4.). It is not even always true if $X$ is a compact Hausdorff space [SS95, Example 86]. But it is true if $X$ is metrizable (XI.10.1.1.).

Closure of a Connected Set

XI.13.1.9. Proposition. Let $X$ be a topological space, and $E$ a connected subset of $X$. Then the closure $\bar{E}$ is also connected.

Proof: Suppose $\{U, V\}$ is a separation of $\bar{E}$ as in XI.13.1.6.(ii). Then $U \cap E$ or $V \cap E$ is empty, for otherwise $\{U, V\}$ would be a separation of $E$. If $V \cap E$ is empty, then $E \subseteq U$. Thus $\bar{E} \subseteq U$, since $\bar{E} \cap U = \bar{E} \cap V^c$ is closed. Thus $\bar{E} \cap V = \emptyset$, a contradiction. \(\square\)
XI.13.1.10. Corollary. Let $X$ be a topological space. If $X$ has a dense connected subset, then $X$ is connected. In particular, if $X$ is a topological space, $E$ a connected subset of $X$, and $Y$ a subset of $X$ with $E \subseteq Y \subseteq E$, then $Y$ is connected.

XI.13.2. Continuous Images of Connected Sets

One of the most important properties of connected sets is:

XI.13.2.1. Proposition. Let $X$ and $Y$ be topological spaces and $f : X \to Y$ a continuous function. If $E$ is a connected subset of $X$, then $f(E)$ is a connected subset of $Y$.

Informally, a continuous image of a connected space is connected.

Proof: Suppose $\{U, V\}$ is a separation of $f(E)$ as in XI.13.1.6.(ii). Then $\{f^{-1}(U), f^{-1}(V)\}$ is a separation of $E$.

XI.13.2.2. Corollary. Let $X$ be a topological space, and $f : X \to \mathbb{R}$ a continuous function. If $E$ is a connected subset of $X$, then $f(E)$ is an interval or singleton. In particular, if $f : \mathbb{R} \to \mathbb{R}$ is continuous and $I$ is an interval in $\mathbb{R}$, then $f(I)$ is an interval or singleton.

XI.13.2.3. Corollary. A connected completely regular space with more than one point has cardinality at least $2^{\mathfrak{c}}$.

Proof: Such a space has a nonconstant real-valued continuous function, whose range has cardinality $2^{\mathfrak{c}}$.

XI.13.2.4. More generally, a connected Urysohn space (XI.7.8.17.) has cardinality at least $2^{\aleph_0}$ (same proof). This corollary can fail for spaces which are not completely regular. There are countably infinite connected Hausdorff spaces [SS95, Examples 60, 75], even countable connected completely Hausdorff spaces [SS95, Example 127]. A connected regular space must be uncountable, since a countable topological space is Lindelöf, and a regular Lindelöf space is normal, hence completely regular. There is a connected regular space of cardinality $\aleph_1$ [SS95, Example 92].

The Intermediate Value Theorem

As a corollary, we get one of the most important properties of continuous functions on the real line:

XI.13.2.5. Corollary. [Intermediate Value Theorem] Let $I$ be an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$ a continuous function. If $a, b \in I$ and $d$ is any number between $f(a)$ and $f(b)$, then there is at least one $c \in I$ with $f(c) = d$. (See Figure 1.)
XI.13.3. Unions and Intersections

XI.13.3.1. A union of two or more connected subsets of a topological space $X$ frequently fails to be connected, e.g. the union of two disjoint open sets or of two sets with disjoint closures (e.g. two singletons in a $T_1$ space). However:

XI.13.3.2. **Proposition.** Let $X$ be a topological space, and $\{E_i : i \in I\}$ a collection of connected subsets of $X$. If $\cap_{i \in I} E_i \neq \emptyset$, then $\cup_{i \in I} E_i$ is connected.

**Proof:** Let $E = \cup_{i \in I} E_i$, and let $p \in \cap_{i \in I} E_i$. Suppose $A$ and $B$ are subsets of $E$ with $E = A \cup B$ and $A \cap B = \emptyset$, with $p \in A$. For each $i$ set $A_i = A \cap E_i$ and $B_i = B \cap E_i$. We have $E_i = A_i \cup B_i$ and $A_i \cap B_i = \emptyset$, so by XI.13.1.6.(iii) $A_i$ and $B_i$ cannot both be nonempty. But $p \in A_i$ for all $i$, so $B_i = \emptyset$ for all $i$, and hence $B = \cup_i B_i = \emptyset$. Thus $E$ is connected by XI.13.1.6.(iii).

XI.13.3.3. **Proposition XI.13.3.2.** can be used to “string together” connected subsets: the union of a sequence of connected subsets of a space is connected provided each successive pair has nonempty intersection.

XI.13.3.4. **Corollary.** Every topological vector space is connected. In particular, $\mathbb{R}^n$ is connected for any $n$.

**Proof:** Any topological vector space is the union of all lines through the origin, each of which is homeomorphic to $\mathbb{R}$ and hence connected.

See XI.13.8.3.(ii) for another proof.

Intersections

XI.13.3.5. The intersection of two connected subsets of a topological space is not connected in general. For example, if $X$ is a circle, the closed upper and lower semicircles are connected since they are homeomorphic to intervals (the whole circle is also connected by XI.13.3.2.), but the intersection consists of two points with the discrete topology.

Decreasing intersections are trickier. It turns out that a decreasing sequence of connected sets, even in $\mathbb{R}^2$, does not have connected intersection in general (Exercise XI.13.9.5.). However, we have:

XI.13.3.6. **Theorem.** Let $X$ be a Hausdorff space, and $\{E_i : i \in I\}$ be a collection of compact subsets of $X$. If $\cap_{i \in F} E_i$ is connected for each finite subset $F$ of $I$, then $\cap_{i \in I} E_i$ is connected. In particular, the intersection of a decreasing sequence of connected compact subsets of $X$ is connected.

“Hausdorff” cannot be weakened to $T_1$ (Exercise XI.13.9.6.).

**Proof:** Set $E = \cap_{i \in I} E_i$ and $E_F = \cap_{i \in F} E_i$ for $F \subseteq I$ finite. Then $E = \cap_F E_F$ is compact. Suppose $E$ is not connected. Then there are disjoint nonempty relatively clopen subsets $A, B$ of $E$ with $E = A \cup B$. Since $A$ and $B$ are relatively closed in $E$, they are compact. There are disjoint open neighborhoods $U$ and $V$ of $A$
and $B$ in $X$ (XI.11.4.2.). By XI.11.1.11. there is a finite subset $F$ of $I$ such that $E_F \subseteq U \cup V$. Since $E_F$ is connected, either $E_F \cap U$ or $E_F \cap V$ is empty. But $A \subseteq E_F \cap U$ and $B \subseteq E_F \cap V$, a contradiction. Thus $E$ is connected.

**XI.13.4. Components of a Space**

**XI.13.4.1. Definition.** Let $X$ be a topological space, and $p \in X$. The connected component (or just component) of $p$ in $X$ is the union of all connected subsets of $X$ containing $p$. (There is always at least one such set, $\{p\}$.)

**XI.13.4.2. Proposition.** Let $X$ be a topological space.

(i) The connected component of any point $p$ of $X$ is closed and connected, and is the largest connected subset of $X$ containing $p$.

(ii) If $p$ and $q$ are points of $X$, the connected components of $p$ and $q$ are either identical or disjoint.

**Proof:**

(i): The union $P$ of all connected subsets of $X$ containing $p$ is nonempty and connected by XI.13.3.2., and is obviously the largest connected subset of $X$ containing $p$. By XI.13.1.9., $P$ is a connected set containing $p$, hence $P \subseteq P$, i.e. $P$ is closed.

(ii): If $P$ and $Q$ are the components of $p$ and $q$ in $X$ respectively, and $P \cap Q \neq \emptyset$, then $P \cup Q$ is connected by XI.13.3.2. and contains $p$ and $q$, hence $P \cup Q = P = Q$.

**XI.13.4.3.** So the connected components of a space $X$ partition $X$ into disjoint closed subsets. Connected components can easily contain more than one point: if $X$ is connected, the connected component of any point of $X$ is all of $X$. At the other extreme, connected components can be singletons. This obviously happens for isolated points, but can happen for nonisolated points too: the components of the Cantor set $K$ are connected subsets of $\mathbb{R}$, hence points or intervals; since $K$ contains no intervals, the connected components of $K$ are single points. Intermediate possibilities also occur in many spaces.

The connected component of a point $p$ in a space $X$ depends on $X$; if $Y$ is a subspace of $X$ containing $p$, the connected component of $p$ in $Y$ can be smaller than the connected component in $X$ (it can even be smaller than the intersection of $Y$ with the connected component in $X$, as easy examples show).

**Quasi-Components**

**XI.13.4.4.** If $X$ is a topological space, $p \in X$, and $U$ is a clopen subset of $X$ containing $p$, then any connected subset of $X$ containing $p$ must be contained in $U$ (otherwise $\{U,U^c\}$ would give a separation), and in particular the component $P$ of $p$ in $X$ is contained in $U$. Thus

$$P \subseteq Q = \bigcap \{U \subseteq X \text{ clopen} : p \in U\}.$$

The set $Q$ is sometimes called the quasi-component of $p$ in $X$. It can be strictly larger than the component of $p$ in $X$ (Exercises XI.13.9.7. and XI.13.9.8.). However:
XI.13.4.5. Proposition. Let $X$ be a compact Hausdorff space. Then the quasi-component of any point is just its component.

“Hausdorff” cannot be weakened to $T_1$ in this result (Exercise XI.13.9.7.).

Proof: The argument is similar to the proof of XI.13.3.6. Let $p \in X$, and let $P$ and $Q$ be the component and quasi-component of $p$ in $X$. Since $P \subseteq Q$, it suffices to show that $Q$ is connected. Suppose $Q$ is not connected, and write $Q = A \cup B$ for disjoint nonempty relatively clopen subsets with $p \in A$. Then $A$ and $B$ are compact, so they have disjoint open neighborhoods $U$ and $V$ in $X$. Since $Q$ is the intersection of all clopen subsets of $X$ containing $p$ and each clopen set in $X$ is compact, by XI.11.1.11. there are finitely many clopen subsets $W_1, \ldots, W_n$ of $X$ containing $p$ such that $W = W_1 \cap \cdots \cap W_n$ is contained in $U \cup V$. $W$ is clopen, and $W$ is the disjoint union of $W \cap U$ and $W \cap V$. $W \cap U$ is open, and $W \cap U = W \cap V^c$ is also closed, i.e. $W \cap U$ is clopen and contains $p$, and thus $Q \subseteq W \cup U$, contradicting that $B$ is nonempty. Thus $Q$ is connected.

XI.13.4.6. Quasi-components, like components, are closed, and two quasi-components are either identical or disjoint. Thus the whole space partitions into quasi-components, each of which partitions into one or more components.

XI.13.5. Products of Connected Spaces

XI.13.5.1. Theorem. Let $\{X_i : i \in I\}$ be a collection of topological spaces. Then $\prod_i X_i$ is connected if and only if each $X_i$ is connected.

Proof: Let $X = \prod_i X_i$. If $X$ is connected, since each coordinate projection is continuous it follows from XI.13.2.1. that each $X_i$ is connected. Conversely, suppose each $X_i$ is connected. Let $p = (p_i)$ and $q = (q_i)$ be points of $X$. Let $F = \{i_1, \ldots, i_n\}$ be a finite subset of $I$. The subset $E_1 = \{(x_i) : x_i = p_i \text{ for } i \neq i_1\}$ is homeomorphic to $X_{i_1}$, hence connected, and contains the point $a_1$ with $x_{i_1} = q_{i_1}$ and $x_i = p_i$ for all $i \neq i_1$. Thus $a_1$ is in the connected component of $p$. Similarly, if $a_2$ is the point $(x_i)$ with $x_{i_1} = q_{i_1}$, $x_{i_2} = q_{i_2}$, and $x_i = p_i$ for $i \neq i_1, i_2$, then $a_2$ is in the connected component of $a_1$ and hence in the connected component of $p$. Repeating the argument finitely many times, if $q_F$ is the point $(x_i)$ with $x_i = q_i$ for $i \in F$ and $x_i = p_i$ for $i \notin F$, then $q_F$ is in the connected component of $p$. This is true for any finite $F$. Any open neighborhood of $q$ contains $q_F$ for some finite set $F$, so $q$ is in the closure of the component of $p$. But components are closed, so $q$ is in the component of $p$ in $X$. This is true for any $p$ and $q$, so $X$ consists of only one connected component, i.e. is connected.

XI.13.6. TotallyDisconnected and Zero-Dimensional Spaces

XI.13.6.1. Definition. Let $X$ be a topological space. $X$ is totally disconnected if for every pair $\{p, q\}$ of distinct points of $X$, there is a clopen set $U \subseteq X$ containing $p$ but not $q$. (In other words, the quasi-components of $X$ are single points.)

A totally disconnected topological space is Hausdorff: if $U$ is a clopen set containing $p$ but not $q$, then $U$ and $X \setminus U$ are disjoint open neighborhoods of $p$ and $q$. 

1155
XI.13.6.2. Some references define a space to be totally disconnected if its connected components are single points (what we have called “totally disconnected” is often then called “totally separated”). This definition is equivalent to the one we have given for compact Hausdorff spaces, but not in general; see Theorem XI.13.6.7. Definition XI.13.6.1. appears to be the usual definition used by topologists.

XI.13.6.3. Definition. Let $X$ be a topological space. $X$ is zero-dimensional if $X$ is $T_1$ and every finite open cover of $X$ has a refinement consisting of a partition of $X$ into clopen subsets.

More precisely, this is the specification that the covering dimension (XI.21.3.23.) of $X$ is zero. Some references give a different definition (that there is a base for the topology of $X$ consisting of clopen sets, i.e. that the small inductive dimension of $X$ is zero; cf. XI.21.3.2.). The two definitions are equivalent for $T_1$ Lindelöf spaces (in particular, for compact Hausdorff spaces and second countable metrizable spaces), but not in general; see Theorem XI.13.6.7. The $T_1$ condition cannot be reasonably weakened to $T_0$ (Exercise XI.13.9.1.).

XI.13.6.4. Actually, if $X$ is zero-dimensional, every finite open cover of $X$ has a refinement consisting of a finite partition of $X$ into clopen sets. Even more is true: a finite open cover $\{V_1, \ldots, V_n\}$ has a shrinking consisting of a partition, i.e. there is a partition $\{U_1, \ldots, U_n\}$ of $X$ into clopen sets with $U_j \subseteq V_j$ for each $j$ (XI.21.3.27.).

XI.13.6.5. Proposition. Let $E$ and $F$ be disjoint closed subsets of a zero-dimensional space $X$. Then there is a clopen set $U \subseteq X$ with $E \subseteq U$ and $F \subseteq X \setminus U$. In particular, a zero-dimensional topological space is normal.

Proof: Let $E$ and $F$ be disjoint closed subsets of a zero-dimensional space $X$. The open cover $\{X \setminus E, X \setminus F\}$ has a refinement consisting of a partition of the space into clopen subsets. Let $U$ be the union of all the sets in the partition which contain points of $E$; any such set is disjoint from $F$ since the partition refines $\{X \setminus E, X \setminus F\}$. Then $U$ is an open set containing $E$, and $F \subseteq X \setminus U$. $X \setminus U$ is the union of the remaining sets in the partition, hence is also open, i.e. $U$ is clopen.

There is a converse: if $X$ is $T_1$ and any two disjoint closed sets can be separated by a partition, i.e. there is a clopen set containing one but not the other ($\text{Ind}(X) = 0$ (XI.21.3.9.)), then $X$ is zero-dimensional (XI.21.3.28.).

XI.13.6.6. Proposition. Let $X$ be a $T_0$ space for which there is a base for the topology consisting of clopen sets. Then $X$ is completely regular.

Proof: Let $F$ be a closed set in $X$, and $p \in V = X \setminus F$. Then $V$ is an open neighborhood of $p$, so there is a clopen neighborhood $U$ of $p$ contained in $V$. Define $f : X \to [0, 1]$ by setting $f(x) = 1$ if $x \in U$ and $f(x) = 0$ if $x \in X \setminus U$. Then $f$ is continuous, $f(p) = 1$, and $f$ is identically 0 on $F$.
XI.13.6.7. **Theorem.** Let $X$ be a compact Hausdorff space. The following are equivalent:

(i) Each component of $X$ is a single point.

(ii) $X$ is totally disconnected.

(iii) There is a base for the topology of $X$ consisting of clopen sets ($\text{ind}(X) = 0$).

(iv) $X$ is zero-dimensional ($\text{dim}(X) = 0$).

(v) Any two disjoint closed sets in $X$ can be separated by a partition ($\text{Ind}(X) = 0$).

If $X$ is metrizable, these are also equivalent to

(vi) $X$ is homeomorphic to a subset of the Cantor set.

XI.13.6.8. For general topological spaces, we have (vi) $\Rightarrow$ (iv) $\Rightarrow$ (i), (iv) $\Rightarrow$ (vi) (XI.13.6.5.) and (iv) $\Rightarrow$ (iii); (v) $\Rightarrow$ (iv) for $T_1$ spaces (XI.21.3.28.), and (iii) $\Rightarrow$ (ii) for $T_0$ spaces (a space with the indiscrete topology and more than one point satisfies (iii) but not (ii)). (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) for compact Hausdorff spaces, but not in general: there are $G_\delta$ sets in $\mathbb{R}^2$ satisfying (i) but not (ii), and ones satisfying (ii) but not (iii) ([?, §46], [?], Exercise XI.13.9.3.); thus (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) do not hold even in Polish spaces. See also Exercise XI.13.9.7.. (iii) $\Rightarrow$ (iv) for $T_0$ Lindelöf spaces, but not for general normal spaces [GJ76, 16M], so (iii), (iv), and (v) are equivalent for $T_1$ Lindelöf spaces (which are then automatically normal).

(iv) does not imply (vi) for compact Hausdorff spaces: an uncountable product of two-point spaces is a counterexample (the space from Exercise () is a more interesting counterexample). (iii) $\Rightarrow$ (vi) for arbitrary second countable $T_0$ spaces; thus (iii), (iv), (v), and (vi) are equivalent for second countable $T_1$ spaces (which are then automatically metrizable).

**Proof:** (i) $\Rightarrow$ (ii) is XI.13.4.5., and (ii) $\Rightarrow$ (i) is XI.13.4.4..

(iii) $\Rightarrow$ (ii): We use only that $X$ is $T_0$. Let $p$ and $q$ be points of $X$, and let $V$ be an open set containing $p$ but not $q$. Since the clopen sets form a base for the topology, there is a clopen set $U$ with $p \in U \subseteq V$. Then $U$ separates $p$ and $q$.

(ii) $\Rightarrow$ (iii): Let $p \in X$ and $V$ an open set containing $p$. For each $q \not\in V$, let $U_q$ be a clopen set containing $p$ but not $q$. Then $\{U_q^c : q \in V^c\}$ is an open cover for $V^c$. Since $V^c$ is closed and hence compact, there is a finite subcover $\{U_q^c : q \in V^c\}$. Then $U = U_{q_1} \cap \cdots \cap U_{q_n}$ is a clopen set containing $p$ and contained in $V$. Thus the clopen sets form a base for the topology.

(iv) $\Rightarrow$ (iii): We use only that $X$ is $T_1$. Let $p \in X$ and $V$ an open set containing $p$. Since $\{p\}$ is closed, $\{V, X \setminus p\}$ is an open cover of $X$, so it has a refinement consisting of a partition into clopen sets. The clopen set containing $p$ must be contained in $V$. Thus the clopen sets form a base for the topology.

(iii) $\Rightarrow$ (iv): We will use only that $X$ is $T_0$ and Lindelöf, and satisfies (iii). Then $X$ is completely regular by XI.13.6.6., and in particular $T_1$ (actually $X$ is automatically normal by XI.7.6.11., but we will not need this; after the fact $X$ will be normal by XI.13.6.5.). Let $\mathcal{V}$ be an open cover of $X$ (not necessarily finite). Since the clopen sets form a base for the topology, there is a refinement $\mathcal{W}$ of $\mathcal{V}$ consisting of clopen sets. $\mathcal{W}$ has a countable subcover $\{W_1, W_2, \ldots\}$. Set $U_1 = W_1, U_2 = W_2 \setminus W_1, U_3 = W_3 \setminus (W_1 \cup W_2)$, etc. (Any $U_n$ which are empty can be discarded.) Then $\{U_1, U_2, \ldots\}$ is a partition of $X$ into clopen sets, refining $\mathcal{W}$ and hence refining $\mathcal{V}$.

1157
Actually, any partition of a compact $X$ into clopen sets is necessarily finite since it forms an open cover of $X$ with no proper subcover.

(iii) $\Rightarrow$ (vi): We will use only that $X$ is $T_0$, second countable, and satisfies (iii).
We first show that there is a countable base for the topology of $X$ consisting of clopen sets. Let $\mathcal{V}$ be a countable base for the topology of $X$. For every pair $(V,W)$ in $\mathcal{V}$ with $V \subseteq W$, let $\mathcal{C}(V,W)$ be the set of all clopen sets $U$ with $V \subseteq U \subseteq W$. $\mathcal{C}(V,W)$ might potentially be empty for some $(V,W)$ (after the fact it turns out that it is always nonempty); but if it is nonempty, choose a clopen set $U_{V,W} \in \mathcal{C}(V,W)$. The collection $\mathcal{U}$ of all such $U_{V,W}$ is countable. $\mathcal{U}$ is a base for the topology of $X$ since if $p$ is any point of $X$ and $N$ any open neighborhood of $p$, there is a $W \in \mathcal{V}$ with $p \in W$ and $W \subseteq N$; since the clopen sets form a base, there is a clopen set $U$ with $p \in U$ and $U \subseteq W$; there is then a $V \in \mathcal{V}$ with $p \in V$ and $V \subseteq U$. Thus $\mathcal{C}(V,W) \neq \emptyset$ and $U_{V,W}$ is defined; then $p \in V \subseteq U_{V,W} \subseteq W \subseteq N$.
Now let $\{U_1,U_2,\ldots\}$ be a countable base for the topology of $X$ consisting of clopen sets. Define a function $f : X \rightarrow \mathbb{R}$ by letting $f(x)$ be the number with base 3 expansion
\[ d_1d_2d_3 \cdots \]
where $d_n = 2$ if $x \in U_n$ and $d_n = 0$ if $x \in U_n^c$. $f$ is one-to-one since $X$ is $T_0$. The range of $f$ is contained in the Cantor set $K$. The function $f$ is continuous since if $x_k \rightarrow x$, then for each $n$, if $x \in U_n$ we have $x_k \in U_n$ for all sufficiently large $k$, and similarly if $x \in U_n^c$, so the $n$th decimal term of $x_k$ agrees with the $n$th decimal term of $x$ for all sufficiently large $k$, and hence $f(x_k) \rightarrow f(x)$. Conversely, if $(x_k), x \in X$ and $f(x_k) \rightarrow f(x)$, then for each $n$ we have that the $n$th decimal term of $x_k$ agrees with the $n$th term of $x$ for all sufficiently large $k$, so if $x \in U_n$ we have $x_k \in U_n$ for sufficiently large $k$. Since the $U_n$ form a base for the topology of $X$, we have $x_k \rightarrow x$. Thus $f^{-1} : f(X) \rightarrow X$ is continuous and $f$ is a homeomorphism onto its image.

(vi) $\Rightarrow$ (iii) since the Cantor set satisfies (iii) and (iii) passes to subspaces.

The Countable AC is needed in several places in this proof.

XI.13.6.9. There is an extensive theory of dimension for topological spaces, discussed in (), for which the above theory is the dimension zero part; the theory is based on “higher-order connectedness.”

XI.13.7. Locally Connected Spaces

XI.13.7.1. Definition. Let $X$ be a topological space, $p \in X$. Then $X$ is locally connected at $p$ if every open neighborhood of $p$ contains a connected open neighborhood of $p$, i.e. if there is a local base for the topology at $p$ consisting of connected sets. $X$ is locally connected if it is locally connected at every point, i.e. if there is a base for the topology consisting of connected sets.

XI.13.7.2. Simple examples of locally connected spaces are $\mathbb{R}$ and any discrete space. The latter example shows that a locally connected space need not be connected. Any open set in a locally connected space is locally connected.

The Cantor set is not locally connected at any point. In fact, a nondiscrete totally disconnected space cannot be locally connected.

The topologist’s sine curve () is connected, but not locally connected at points on the $y$-axis. There are more extreme examples: there are compact subsets of Euclidean space, e.g. the pseudo-arc () or a solenoid (), which are connected but not locally connected at any point.
XI.13.7.3. If $X$ is locally connected, then the components of $X$ are open, hence clopen. The converse is not true since there are connected spaces which are not locally connected. Spaces whose components are open are precisely ones in which each point has at least one connected open neighborhood. If $X$ is locally connected, the connected components of any open subset $U$ of $X$ are also open, since $U$ is also locally connected. Conversely, if every component of every open subset of $X$ is open, then $X$ is locally connected.

XI.13.8. Path-Connected Spaces

An intuitively appealing variation of connectedness is path-connectedness. A space is path-connected if any two points can be connected by a continuous path; such a space is necessarily connected. Most familiar topological spaces such as (connected) manifolds and polyhedra are path-connected.

XI.13.8.1. Definition. Let $X$ be a topological space, $p, q \in X$. A path in $X$ from $p$ to $q$ is a continuous function $\gamma : [0, 1] \to X$ with $\gamma(0) = p$, $\gamma(1) = q$.

The space $X$ is path-connected if for any $p, q \in X$, there is a path from $p$ to $q$.

A path-connected space is also called arcwise connected. (Some references make a technical distinction between paths and arcs; cf. XI.13.8.11.)

Since an interval is connected (XI.13.1.7.), we obtain:

XI.13.8.2. Proposition. Let $X$ be a topological space, $p, q \in X$. If there is a path from $p$ to $q$, then $p$ and $q$ are in the same connected component of $X$. In particular, a path-connected space is connected.

Proof: If $\gamma$ is a path from $p$ to $q$, then by XI.13.2.1. the set $\gamma([0, 1])$ is a connected subset of $X$ containing both $p$ and $q$.

XI.13.8.3. Examples. (i) Any connected subset of $\mathbb{R}$ (interval or point) is path-connected.

(ii) Any convex set in $\mathbb{R}^n$ (or, more generally, in any topological vector space) is path-connected: the straight line segment between any two points is a path. In particular, any topological vector space, including $\mathbb{R}^n$, is path-connected.

(iii) The topologist’s sine curve ( ) is connected but not path-connected. There are more dramatic examples of such spaces (XI.13.8.6.).

Define a relation $\sim$ on $X$ by $p \sim q$ if there is a path from $p$ to $q$.

XI.13.8.4. Proposition. The relation $\sim$ is an equivalence relation.

Proof: If $\gamma$ is a path from $p$ to $q$, set $\tilde{\gamma}(t) = \gamma(1 - t)$ for $0 \leq t \leq 1$; then $\tilde{\gamma}$ is a path from $q$ to $p$, so $\sim$ is symmetric. The constant function from $[0, 1]$ to $X$ with value $p$ is a path from $p$ to $p$, so $\sim$ is reflexive. Now suppose $\gamma_1$ is a path from $p$ to $q$ and $\gamma_2$ a path from $q$ to $r$. Define $\gamma : [0, 1] \to X$ by

$$
\gamma(t) = \begin{cases} 
\gamma_1(2t) & \text{if } 0 \leq t \leq \frac{1}{2} \\
\gamma_2(2t - 1) & \text{if } \frac{1}{2} \leq t \leq 1 
\end{cases}
$$
Then $\gamma$ is a path from $p$ to $r$, so $\sim$ is transitive.

**XI.13.8.5.** **Definition.** The equivalence classes of $\sim$ are the **path components of $X$.**

As with components, the space $X$ partitions into a disjoint union of path components. By XI.13.8.2., each path component is contained in a connected component of $X$, i.e. each connected component partitions into one or more path components.

Unlike components, path components need not be closed. For example, the topologist’s sine curve () consists of two path components, one closed and the other dense and open. A solenoid () has uncountably many path components, each dense (and not open).

**XI.13.8.6.** A space $X$ is **totally path disconnected** if its path components are single points, i.e. if every continuous function from $[0,1]$ into $X$ is constant. There are connected compact metrizable spaces which are totally path disconnected, e.g. the pseudo-arc ().

**XI.13.8.7.** **Definition.** A topological space $X$ is **locally path-connected** at $p$ if there is a local base for the topology at $p$ consisting of path-connected sets, i.e. any open neighborhood of $p$ contains a path-connected open neighborhood of $p$. $X$ is **locally path-connected** if it is locally path-connected at each point, i.e. if there is a base for the topology consisting of path-connected sets.

A locally path-connected space is locally connected. There are connected and locally connected compact Hausdorff spaces which are not locally path-connected [?, Examples 46, 48]; however, a connected compact metrizable space which is locally connected is locally path-connected (Exercise XI.13.9.11.; in fact, the whole space is a continuous image of $[0,1]$ by the Hahn-Mazurkiewicz Theorem; cf. [HY88, 3.29–3.30], Exercise XI.13.9.12.). This result is false for noncompact sets (XI.13.9.14.).

**XI.13.8.8.** There are spaces which are path-connected but not locally path-connected: for example, the extended topologist’s sine curve () is not locally path-connected (not even locally connected) at points on the $y$-axis.

**XI.13.8.9.** If $X$ is locally path-connected at $p$, then the path component of $p$ in $X$ is open. If $X$ is locally path-connected, then all its path components are open, hence also closed since the complement of a path component is the union of the other path components. Thus in a locally path-connected space, the path components are the same as the components (and the quasi-components). In particular, if $X$ is connected and locally path-connected, it is path-connected.

**XI.13.8.10.** Since $\mathbb{R}^n$ is locally path-connected (open balls form a base for the topology of path-connected sets), any open set in $\mathbb{R}^n$ is locally path-connected. Hence the connected components of any open set in $\mathbb{R}^n$ are the same as the path components, and are open.

The same is true in any locally convex topological vector space (in particular, any normed vector space). In fact, any topological vector space, locally convex or not, is locally path-connected (there is a base for the topology consisting of balanced open sets, which are path-connected).
Path-Connected vs. Arcwise Connectedness

XI.13.8.11. Definition. A topological space $X$ is arcwise connected if, for every $p, q \in X$, $p \neq q$, there is a homeomorphism $\gamma : [0, 1] \to X$ with $\gamma(0) = p$ and $\gamma(1) = q$.

Arcwise connectedness is more restrictive than path-connectedness, which only requires that there be a continuous $\gamma$. A homeomorphic image of $[0, 1]$ in $X$ is called an arc in $X$. The Sierpiński space is path-connected but not arcwise connected, as it is a set of cardinality $\geq 2^{\aleph_0}$ with the finite complement topology.

The main technical result is that path-connectedness and arcwise connectedness are the same thing for Hausdorff spaces. We give a proof since this result is not discussed in most topology texts, and where it is the proofs given tend to use unnecessarily heavy machinery. Our proof, although somewhat complicated, is quite elementary.

XI.13.8.12. Theorem. Let $X$ be a Hausdorff space, $p, q \in X$, $p \neq q$. If there is a path in $X$ from $p$ to $q$, then there is an arc in $X$ from $p$ to $q$.

Proof: Set $p_0 = p$, $p_1 = q$. Let $\gamma : [0, 1] \to X$ be a path from $p_0$ to $p_1$ (i.e. $\gamma(0) = p_0$, $\gamma(1) = p_1$). We want to find an arc $\hat{\gamma} : [0, 1] \to X$ from $p_0$ to $p_1$. In fact, we will find an arc with $\hat{\gamma}([0, 1]) \subseteq \gamma([0, 1])$.

(i) Replacing $X$ by $\gamma([0, 1])$, we may assume $X$ is compact and metrizable. [The fact that $\gamma([0, 1])$ is metrizable is not essential to the rest of the proof, but makes part (vi) of the argument a little cleaner. The slickest proof I know of this fact is to apply XI.10.110.] Fix a metric $d$ on $X$. Then $\gamma$ is uniformly continuous, i.e. for every $\epsilon > 0$ there is a $\delta(\epsilon) > 0$ such that $d(\gamma(t), \gamma(s)) < \epsilon$ whenever $t, s \in [0, 1]$ and $|t - s| < \delta(\epsilon)$.

(ii) We may assume $\gamma(t) \neq p_0$ and $\gamma(t) \neq p_1$ for all $t \in (0, 1)$: since $X$ is Hausdorff, there is a largest $t \in [0, 1]$ for which $\gamma(t) = p_0$; call it $a_0$. We have $0 \leq a_0 < 1$. Then there is a smallest $t \in [a_0, 1]$ with $\gamma(t) = p_1$; call it $b_0$. We have $a_0 < b_0 \leq 1$. Replace $\gamma$ by $\gamma|_{[a_0, b_0]}$ (rescaled) if necessary.

(iii) Set $\gamma_1 = \gamma$. Let $a_1$ be the smallest $t$ such that $\gamma_1(t) = \gamma_1(s)$ for some $s \in [\frac{1}{2}, 1]$ [there is such a $t$ since $\gamma_1([\frac{1}{2}, 1])$ is compact and hence closed, so $\gamma_1^{-1}(\gamma_1([\frac{1}{2}, 1]))$ is closed]; then $0 < a_1 \leq \frac{1}{2}$. Let $b_1$ be the largest $t$ such that $\gamma_1(b_1) = \gamma_1(a_1)$; then $\frac{1}{2} \leq b_1 < 1$. Set $p_{1/2} = \gamma_1(a_1) = \gamma_1(b_1)$, and let $\gamma_2$ be $\gamma_1|_{[0, a_1]}$ rescaled to $[0, \frac{1}{2}]$ and $\gamma_1|_{[b_1, 1]}$ rescaled to $[\frac{1}{2}, 1]$. We have $\gamma_2([0, 1]) \subseteq \gamma_1([0, 1])$. The sets $A_{11} = \gamma_2([0, \frac{1}{2}])$ and $A_{12} = \gamma_2([\frac{1}{2}, 1])$ satisfy $A_{11} \cap A_{12} = \{p_{1/2}\}$.

(iv) Repeat the construction of (iii) for the restrictions of $\gamma_2$ to $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. We obtain numbers $a_{21} \in (0, \frac{1}{4})$, $b_{21} \in [\frac{1}{4}, \frac{1}{2})$, $p_{1/4} = \gamma_2(a_{21}) = \gamma_2(b_{21})$ and $\gamma_3 : [0, \frac{1}{2}] \to X$ which is rescaled $\gamma_2$ on $[0, a_{21}]$ and $[b_{21}, 1/2]$, and similarly $a_{22} \in (\frac{1}{4}, \frac{1}{2})$, $b_{22} \in [\frac{1}{2}, \frac{3}{4})$, $p_{3/4} = \gamma_2(a_{22}) = \gamma_2(b_{22})$ and $\gamma_3 : [\frac{1}{2}, 1] \to X$ which is rescaled $\gamma_2$ on $[1/2, a_{21}]$ and $[b_{21}, 1]$. Put $\gamma_3$ and $\gamma_2$ together to form $\gamma_3 : [0, 1] \to X$, a path from $p_0$ to $p_1$ with $\gamma_3(1/2) = \gamma_2(1/2) = p_{1/2}$. We have $\gamma_3([0, \frac{1}{2}]) \subseteq \gamma_2([0, \frac{1}{2}])$ and $\gamma_3([\frac{1}{2}, 1]) \subseteq \gamma_2([\frac{1}{2}, 1])$, so $\gamma_3([0, 1]) \subseteq \gamma_2([0, 1])$.

(v) Continuing inductively in this way, a sequence $(\gamma_n)$ of paths from $p_0$ to $p_1$ is constructed, with $\gamma_n(k/2^m) = p_{k/2^m}$ for $m \leq n - 1$. We have $\gamma_n([0, 1]) \subseteq \gamma_n([0, 1])$, and more generally

$$\gamma_{n+1} \left( \left[ \frac{k}{2^m}, \frac{k+1}{2^m} \right] \right) \subseteq \gamma_n \left( \left[ \frac{k}{2^m}, \frac{k+1}{2^m} \right] \right)$$

for any $m \leq n - 1$ and any $k$. If $A_{n+1,k} = \gamma_{n+1} \left( \left[ \frac{k-1}{2^m}, \frac{k}{2^m} \right] \right)$, $A_{n+1,k} \cap A_{n+1,k+1} = \{p_{k/2^m}\}$ and $A_{n+1,k} \cap A_{n+1,k+j} = \emptyset$ for $j > 1$. 

1161
(vi) We claim that \((\gamma_n)\) converges uniformly on \([0, 1]\) to a path \(\tilde{\gamma}\), i.e. that \((\gamma_n)\) is a uniform Cauchy sequence. Let \(\epsilon > 0\), and fix \(N\) with \(2^{-N+1} < \delta(\epsilon)\). For any \(c \in [0, 1]\), there is a \(k\) such that \(\frac{k}{2^{N-1}} \leq c \leq \frac{k+1}{2^{N-1}}\). Then
\[
 Z = \gamma_N \left( \left[ \frac{k}{2^N}, \frac{k+1}{2^N} \right] \right)
\]
is \(\gamma([a, b])\) for some interval \([a, b]\) of length \(\leq 2^{-N+1} < \delta(\epsilon)\), so \(Z\) has diameter less than \(\epsilon\). If \(n, m \geq N\), we have \(\gamma_n(c), \gamma_m(c) \in Z\), so \(d(\gamma_n(c), \gamma_m(c)) < \epsilon\). Note that \(N\) depends only on \(\epsilon\) and not on \(c\).

We have
\[
\tilde{\gamma} \left( \left[ \frac{k}{2^n}, \frac{k+1}{2^n} \right] \right) \subseteq \gamma_n \left( \left[ \frac{k}{2^n}, \frac{k+1}{2^n} \right] \right)
\]
for any \(n, m \leq n - 1\), and \(k\).

(vii) The path \(\tilde{\gamma}\) is an arc. To show this, we need only show that \(\tilde{\gamma}\) is one-to-one (since \(X\) is Hausdorff). So suppose \(\tilde{\gamma}(c) = \tilde{\gamma}(d)\) for some \(c, d, c < d\). Choose \(n, k, j\) with \(j \geq 2\) and
\[
k \leq c < \frac{k+1}{2^n}, \quad \frac{k+j}{2^n} < d \leq \frac{k+j+1}{2^n}.
\]
But then
\[
\tilde{\gamma}(c) \in \gamma_{n+1} \left( \left[ \frac{k}{2^n}, \frac{k+1}{2^n} \right] \right)
\]
and these sets are disjoint, a contradiction.

**X.13.8.13.** COROLLARY. Let \(X\) be a Hausdorff space. If \(X\) is path-connected, then \(X\) is arcwise connected.

**X.13.9.** Exercises

**X.13.9.1.** Let \(X\) be the Sierpiński two-point space (\(\mathbb{S}^2\)). Every open cover of \(X\) has a refinement consisting of clopen sets, namely \(\{X\}\), but there is no base for the topology consisting of clopen sets. \(X\) is \(T_0\), but not \(T_1\) and hence not zero-dimensional by Definition X.13.6.3.; in fact, it is connected, and even path-connected: the function from \([0, 1]\) to \(X\) sending \([0, 1)\) to \(a\) and \(1\) to \(b\) is continuous.

**X.13.9.2.** (Erdős) Let \(X\) be the set of points in \(\ell^2\) (\(\mathbb{E}^2\)) with rational coordinates. Show that \(X\) is totally disconnected, but there is no base for the topology consisting of clopen sets (in fact, no clopen subset is bounded.)

**X.13.9.3.** [?, §27,VI] Let \(K\) be the Cantor set (\(\mathbb{C}\)). Define a function \(f : K \to \mathbb{R}\) as follows. If \(t \in K\), then \(t\) can be uniquely written as a finite or infinite sum
\[
t = \sum_k \frac{2}{3^n_t}
\]
(where \( t = 0 \) is the “empty sum”), i.e. \( n_k \) is the location of the \( k \)'th 2 in the base 3 decimal expansion of \( t \). Then set
\[
f(t) = \sum_{k} \frac{(-1)^{n_k}}{2^k}
\]
(with \( f(0) = 0 \)). Let \( X \) be the graph of \( f \), i.e. \( X = \{(t, f(t)) : t \in K\} \subseteq \mathbb{R}^2 \).
(a) Show that \( f \) is not continuous. Find the points of continuity of \( f \).
(b) Show that \( X \) is totally disconnected.
(c) Show that there is no base for the topology of \( X \) consisting of clopen sets.
(d) Show that there is a point \( p \in \mathbb{R}^2 \) such that \( X \cup \{p\} \) is not totally disconnected, although its connected components are single points.
(e) Show that \( X \) and \( X \cup \{p\} \) are \( G_\delta \) subsets of \( \mathbb{R}^2 \). \([f \) is a pointwise limit of a sequence of continuous functions.]

**XI.13.9.4.**  
Topologize \( X = \{a, b, c\} \) by taking the open sets to be \( \{\emptyset, \{a\}, \{a, b\}, \{a, c\}, X\} \). Then \( X \) is connected. The relative topology on \( E = \{b, c\} \) is the discrete topology and \( E \) is not connected. However, there is no separation \( \{U, V\} \) of \( E \) with \( U \) and \( V \) disjoint.

**XI.13.9.5.**  
Let \( X \) be the subset \( G \cup E \) of the plane, where \( G = \{(x, y) : 0 < x \leq \frac{1}{2}, y = \sin \left( \frac{1}{x} \right) \} \) and \( E = \{(0, 1), (0, -1)\} \). Since \( G \) is homeomorphic to \((0, 1)\), it is connected. \( G \) is dense in \( X \), so \( X \) is connected. Let \( E_n = E \cup \{(x, y) : 0 < x \leq \frac{1}{n^2}, y = \sin \left( \frac{1}{x} \right) \} \). Then \( E_n \) is connected for all \( n \), and \( E = \cap_n E_n \), but \( E \) is not connected. Note that \( E \) is compact, although the \( E_n \) are not compact.

**XI.13.9.6.**  
(a) Give \( \mathbb{N} \) the finite complement topology. Then \( \mathbb{N} \) is compact and \( T_1 \), and every infinite subset of \( \mathbb{N} \) is connected. Set \( E = \{1, 2\} \), and \( E_n = \{1, 2, n + 1, n + 2, \ldots\} \). Then \( E_n \) is compact and connected for all \( n \), but \( E = \cap_n E_n \) is not connected. \((E_n \) is not, however, closed in \( \mathbb{N}\).)
(b) Does the conclusion of **XI.13.3.6.** hold for arbitrary \( X \) if the \( E_i \) are assumed to be *closed* compact subsets?

**XI.13.9.7.**  
Let \( X \) be \( \mathbb{N} \) with two points \( a, b \) added. Topologize \( X \) by taking the open subsets to be all subsets with finite complement and all subsets of \( \mathbb{N} \).
(a) Show that \( X \) is a compact \( T_1 \) space.
(b) Show that the components of \( X \) are single points.
(c) Show that the quasi-component of \( a \) and \( b \) is \( \{a, b\} \), i.e. there is no clopen set containing \( a \) but not \( b \).

**XI.13.9.8.**  
Let \( X \) be the subset of \( \mathbb{R}^2 \) consisting of \( \{(\frac{1}{n}, t) : n \in \mathbb{N}, 0 \leq t \leq 1\} \) along with \((0, 0)\) and \((0, 1)\).
(a) Show that the components of \( X \) are \( \{(0, 0)\}, \{(0, 1)\}, \) and \( \{(\frac{1}{n}, t) : 0 \leq t \leq 1\} \) for fixed \( n \).
(b) Show that \((0, 0)\) and \((0, 1)\) are in the same quasi-component of \( X \).
XI.13.9.9. Let $X$ be the subset of $\mathbb{R}^2$ obtained by beginning with the line segment between $(0, 0)$ and $(1, 0)$ and adding the closed line segments between $\left(\frac{1}{n+1}, \frac{1}{n(n+m)}\right)$ for all $m, n \in \mathbb{N}$. $X$ is called the infinite broom. Is $X$ locally connected at $(0, 0)$? Note that if our usual convention that neighborhoods be open is relaxed, every neighborhood of $(0, 0)$ in $X$ contains a connected neighborhood. (This is false for points $(t, 0)$ with $0 < t < 1$). See [HY88, p. 113]. (Note that this $X$ is quite different from the ones in XI.18.4.1.)

XI.13.9.10. Theorem. Let $Y \subseteq \mathbb{R}^n (n > 1)$ be an absolute retract (e.g. $Y$ is homeomorphic to a closed ball of some dimension; cf. ()). Then $\mathbb{R}^n \setminus Y$ is connected.

Proof: Suppose $\mathbb{R}^n \setminus Y$ is not connected, i.e. it has a nonempty bounded component $U$. Let $p \in U$, and let $X = Y \cup U$. Then $X$ is compact since $\partial U \subseteq Y$. The identity map on $Y$ extends to a continuous function $f$ from $X$ onto $Y$ (i.e. $f$ is a retraction from $X$ onto $Y$).

Let $B$ be a closed ball centered at $p$ large enough that $X$ is contained in its interior. Let $g$ be radial retraction from $B \setminus \{p\}$ to $S = \partial B$. Define $r : B \to S$ by

$$r(x) = \begin{cases} g(f(x)) & \text{if } x \in X \\ g(x) & \text{if } x \in B \setminus X \end{cases}.$$ 

Then $r$ is a retraction from $B$ onto $S$, contradicting the No-Retraction Theorem ().

XI.13.9.11. (a) [HY88, Theorem 3.16] Show that any connected open set in a locally compact, locally connected Hausdorff space is path-connected.

(b) [HY88, Theorem 3.17] Show that every connected open set in a locally connected completely metrizable space is path-connected.

XI.13.9.12. Let $X$ be a compact space, $Y$ a (necessarily compact) Hausdorff space, and $f$ a continuous function from $X$ onto $Y$. If $X$ is locally connected, show that $Y$ is also locally connected. [If $U$ is an open set in $Y$, and $C$ is a component of $U$, show that $f^{-1}(C)$ is a union of components of $f^{-1}(U)$ and hence is open in $X$ (XI.13.7.3.). Use that $X \setminus f^{-1}(C)$ is compact to conclude that $C$ is open in $Y$.]

XI.13.9.13. Let $X$ be a connected locally compact Hausdorff space. Let $p$ be a point of $X$ and $U$ an open neighborhood of $p$ in $X$. Show that there is a nontrivial subcontinuum of $X$ (i.e. a connected compact subset containing more than one point) containing $p$ and contained in $U$. In fact, if $V$ is an open neighborhood of $p$ with compact closure, show that the connected component of $p$ in $V$ contains a point of $\partial V$. [Show that the quasi-component of $p$ in $V$ contains a point of $\partial V$, and apply XI.13.4.5..] This is a weak form of local connectedness.

The statements can fail if $X$ is not locally compact (XI.13.9.5..)

A point in a connected locally compact Hausdorff space need not have a connected compact neighborhood: consider the union of the graph of $y = \frac{1}{x} \sin \frac{1}{x}$ ($x > 0$) and the $y$-axis in $\mathbb{R}^2$. 

1164
XI.13.9.14. (F. Bernstein [Ber07]; cf. [HY88, p. 110]) (a) Show that there are exactly \(2^{\aleph_0}\) nontrivial closed connected subsets of \(\mathbb{R}^2\). Note that each has cardinality \(2^{\aleph_0}\).

(b) (AC) Well-order these nontrivial closed connected subsets \(\{X_\alpha : \alpha < \sigma\}\) according to the first ordinal \(\sigma\) of cardinality \(2^{\aleph_0}\). Inductively choose distinct points \(a_\alpha, b_\alpha\) in \(X_\alpha\), for each \(\alpha < \sigma\) (i.e. \(a_\alpha, b_\alpha, a_\beta, b_\beta\) are all distinct for \(\alpha \neq \beta\)).

(c) Let \(A = \{a_\alpha : \alpha < \sigma\}\) and \(B = \mathbb{R}^2 \setminus A\). Then \(b_\alpha \in B\) for all \(\alpha\).

(d) Show that \(A\) and \(B\) are connected and locally connected.

(e) Show that neither \(A\) nor \(B\) contains a subcontinuum of more than one point. In particular, they are totally path disconnected.

Thus \(\mathbb{R}^2\) is a disjoint union of two connected, locally connected subsets. Compare with XIII.3.9.13.

XI.13.9.15. Let \(X\) be a compact Hausdorff space. Define an equivalence relation \(\sim\) on \(X\) by \(x \sim y\) if \(x\) and \(y\) are in the same connected component of \(X\). Let \(X_c = X/\sim\) with the quotient topology; \(X_c\) is called the component space of \(X\).

(a) Show that the relation \(\sim\) is closed. [Use XI.13.4.5.]

(b) Show that the decomposition is usc (XI.8.2.1.). [If \(E\) is closed in \(X\) and \(x\) is not in the saturation of \(E\), show there is a clopen set in \(X\) containing \(x\) disjoint from \(E\).]

(c) Show that \(X_c\) is a compact Hausdorff space. [Use XI.8.3.9.]

(d) Show that \(X_c\) is totally disconnected.

(e) Is the decomposition necessarily lsc?

(f) Extend the results to the case where \(X\) is locally compact and normal. [Show that \(X_c\) embeds in \((\beta X)_c\).]

(g) To what extent does this construction work if \(X\) is not locally compact?

(h) Does a similar construction work for path components?
XI.14. Direct and Inverse Limits

Direct and, especially, inverse limits are important topological constructions. Actually, they have little to do with topology specifically; they can be defined for systems of sets, or more generally within a category. But they are especially useful in topology.

Caution: The use of the term “limit” in connection with direct and inverse limits has almost nothing to do (at least on the surface) with the usual use of the term “limit” in topology, e.g. “limit of a sequence.”

XI.14.1. Direct Systems and Direct Limits of Sets

XI.14.1.1. Definition. A direct system of sets over \( \mathbb{N} \) is a sequence \( (X_n) \) of sets and, for each \( n \), a function \( \phi_{n;n+1} : X_n \to X_{n+1} \).

Pictorially, we have

\[
X_1 \xrightarrow{\phi_{1;2}} X_2 \xrightarrow{\phi_{2;3}} X_3 \xrightarrow{\phi_{3;4}} \ldots
\]

Both the sets \( X_n \) and the maps \( \phi_{n,n+1} \) must be specified. Usually a direct system of sets over \( \mathbb{N} \) is denoted \( (X_n, \phi_{n,n+1}) \).

If \( n < m \), then there is an induced map \( \phi_{n,m} \) from \( X_n \) to \( X_m \) defined by composition:

\[
\phi_{n,m} = \phi_{m-1,m} \circ \phi_{m-2,m-1} \circ \cdots \circ \phi_{n+1,n+2} \circ \phi_{n,n+1}.
\]

If \( n < m < k \), then \( \phi_{n,k} = \phi_{m,k} \circ \phi_{n,m} \).

We can more generally define a direct system over any directed set \( I \):

XI.14.1.2. Definition. Let \( I \) be a (properly) directed set. A direct system of sets over \( I \) is an indexed collection \( \{X_i : i \in I\} \) of sets and, for each \( i, j \) with \( i < j \), a function \( \phi_{i,j} : X_i \to X_j \) satisfying \( \phi_{i,k} = \phi_{j,k} \circ \phi_{i,j} \) whenever \( i < j < k \).

We usually write \( (X_i, \phi_{i,j}) \) or \( (X_i, \phi_{i,j})_{i \in I} \) for a direct system of sets over \( I \).

If working in a category such as the category of topological spaces, the sets are objects in the category, e.g. topological spaces in the category of topological spaces, and the functions are required to be morphisms in the category, e.g. continuous functions in the category of topological spaces.

XI.14.1.3. We want to associate to a direct system a set \( X \) which “sits at \( +\infty \)”. To see how this works, first consider the case where \( I = \mathbb{N} \) and each \( \phi_{n,n+1} \) is injective. Then \( X_n \) can be identified with the subset \( \phi_{n,n+1}(X_n) \) of \( X_{n+1} \) for each \( n \), and the direct limit \( X \) is then the “union” of the \( X_n \). (It is not truly a union except in the rare case where each \( \phi_{n,n+1} \) is an inclusion map of a subset.)

If the \( \phi_{n,n+1} \) are not all injective, things are a little more complicated: \( X \) is not the “union” of the \( X_n \) since any elements which are identified (i.e. have the same image) in some \( X_m \) must be identified in \( X \). Thus in general the direct limit \( X \) informally consists of all objects which appear at any stage, with objects identified if they are eventually identified at some stage.

Here is the formal property that the direct limit must have:
XI.14.1.4. **Definition.** Let \((X_i, \phi_{i,j})\) be a direct system over a directed set \(I\). A direct limit for the direct system is a pair \((X; \phi_i)\), where \(X\) is a set and, for each \(i\), \(\phi_i : X_i \rightarrow X\) is a function such that \(\phi_i = \phi_j \circ \phi_{i,j}\) for any \(i, j\) with \(i < j\); pictorially, in the case \(I = \mathbb{N}\), we have a commutative diagram:

\[
\begin{array}{c}
X_1 \\
\downarrow_{\phi_{1,2}} \\
X_2 \\
\downarrow_{\phi_{2,3}} \\
X_3 \\
\downarrow_{\phi_{3,4}} \\
\vdots
\end{array}
\rightarrow
\begin{array}{c}
X
\end{array}
\]

with the following universal property:

If \(Y\) is a set and for each \(i\) there is a function \(\psi_i : X_i \rightarrow Y\) such that \(\psi_i = \psi_j \circ \phi_{i,j}\) for any \(i, j\) with \(i < j\), then there is a unique function \(\alpha : X \rightarrow Y\) such that \(\psi_i = \alpha \circ \phi_i\) for all \(i\).

We write \((X, \phi_i) = \lim \left(X_i, \phi_{i,j}\right)\) for a direct limit. A direct limit is often called an inductive limit.

Pictorially, the universal property in the case \(I = \mathbb{N}\) is the following commutative diagram:

\[
\begin{array}{c}
X_1 \\
\downarrow_{\phi_{1,2}} \\
X_2 \\
\downarrow_{\phi_{2,3}} \\
X_3 \\
\downarrow_{\phi_{3,4}} \\
\vdots
\end{array}
\rightarrow
\begin{array}{c}
X
\end{array}
\]

XI.14.1.5. **Theorem.** Let \((X_i, \phi_{i,j})\) be a direct system of sets over a directed set \(I\). Then

(i) There is a direct limit \((X, \phi_i)\).

(ii) If \((X, \phi_i)\) and \((Y, \psi_i)\) are direct limits, there is a unique bijection \(\alpha : X \rightarrow Y\) such that \(\psi_i = \alpha \circ \phi_i\) for all \(i\).

Thus up to natural identification there is a unique direct limit for any direct system.

**Proof:** We begin with (ii). This is a simple standard argument that shows that any object with a universal property is unique up to isomorphism.

Since \((X, \phi_i)\) satisfies the universal property, there is a unique function \(\alpha : X \rightarrow Y\) such that \(\psi_i = \alpha \circ \phi_i\) for all \(i\). Similarly, since \((Y, \psi_i)\) satisfies the universal property, there is a unique function \(\beta : Y \rightarrow X\) such that \(\phi_i = \beta \circ \psi_i\). There is also a unique function \(\gamma : X \rightarrow X\) such that \(\phi_i = \gamma \circ \phi_i\) for all \(i\). Since the identity map \(\iota\) on \(X\) has this property, \(\gamma = \iota\). But for each \(i\) we have

\[
\phi_i = \beta \circ \psi_i = \beta \circ \alpha \circ \phi_i
\]

so \(\beta \circ \alpha\) has the property of \(\gamma\) and thus equals \(\iota\). By a similar argument we have that \(\alpha \circ \beta\) is the identity on \(Y\). Thus \(\alpha\) and \(\beta\) are mutually inverse bijections.

1167
We now show existence. Let $Z$ be the subset of $\bigcup_{i \in I} X_i \times I$ defined by
\[ Z = \{ (x, i) : x \in X_i \} \subseteq \bigcup_{i \in I} X_i \times I \]
($Z$ is sometimes called the disjoint union of the $X_i$). Define a relation $\sim$ on $Z$ by setting $(x, i) \sim (x', j)$ if and only if there is a $k$, $i < k$ and $j < k$, such that $\phi_{i,k}(x) = \phi_{j,k}(x')$. It is easily checked that $\sim$ is an equivalence relation. Let $X$ be the set of equivalence classes. Write $[(x, i)]$ for the equivalence class of $(x, i)$ in $Z$.

Fix $i$. Define a function $\phi_i : X_i \to X$ by $\phi_i(x) = [(x, i)]$ for $x \in X_i$. It is easily checked that $\phi_i = \phi_j \circ \phi_{i,j}$ for $i < j$, since $[(x, i)] = [(\phi_{i,j}(x), j)]$ for $i < j$ and $x \in X_i$.

Now suppose $Y$ is a set and $\psi_i : X_i \to Y$ is a function for each $i$ with $\psi_i = \psi_j \circ \phi_{i,j}$ for $i < j$. Define $\alpha : X \to Y$ by $\alpha([(x, i)]) = \psi_i(x)$ for $x \in X_i$. We need only show that this function is well defined, since then obviously $\psi_i(x) = \alpha([(x, i)]) = \alpha(\phi_i(x))$ for $x \in X_i$. If $[(x, i)] = [(x', j)]$, fix $k$ such that $i < k$, $j < k$, and $\phi_{i,k}(x) = \phi_{j,k}(x')$. Then
\[ \psi_i(x) = \psi_k(\phi_{i,k}(x)) = \psi_k(\phi_{j,k}(x')) = \psi_j(x') \]
so $\alpha$ is well defined.

For uniqueness of $\alpha$, suppose $\beta$ is another map from $X$ to $Y$ with $\psi_i = \beta \circ \phi_i$ for all $i$. Let $[(x, i)] \in X$. We have
\[ \beta([(x, i)]) = \beta(\phi_i(x)) = \psi_i(x) = \alpha(\phi_i(x)) = \alpha([(x, i)]) \]
and so $\beta = \alpha$.

\[ \text{(XI.14.1.6)} \quad \text{Proposition. Let } (X, \phi_i) = \lim_{\to} (X_i, \phi_{i,j}). \text{ Then} \]
\begin{enumerate}
  \item For any $x \in X$, there is an $i$ and $x_i \in X_i$ with $x = \phi_i(x_i)$.
  \item If $x, x' \in X$, $x = \phi_i(x_i)$ and $x' = \phi_j(x'_j)$ for $x_i \in X_i$, $x'_j \in X_j$, then $x = x'$ if and only if there is a $k \in I$, $i < k$, $j < k$, such that $\phi_{i,k}(x_i) = \phi_{j,k}(x'_j)$.
\end{enumerate}

The simple proof is left to the reader.

**Passage to Subsystems**

\[ \text{XI.14.1.7} \quad \text{Suppose } (X_n, \phi_{n,n+1}) \text{ is a direct system over } \mathbb{N}. \text{ Let } (k_n) \text{ be a strictly increasing sequence in } \mathbb{N}, \text{ and let } Y_n = X_{k_n}, \psi_{n,n+1} = \phi_{k_n,k_{n+1}}. \text{ If } (X, \phi_n) = \lim_{\to} (X_n, \phi_{n,n+1}) \text{ and } (Y, \psi_n) = \lim_{\to} (Y_n, \psi_{n,n+1}), \text{ we will show there is a natural bijection between } X \text{ and } Y. \text{ For each } n, \text{ set } \alpha_n = \psi_n \circ \phi_{n,k_n} : X_n \to Y; \text{ there is then a unique } \alpha : X \to Y \text{ such that } \alpha_n = \alpha \circ \phi_n \text{ for all } n. \text{ If } \beta_n = \phi_{k_n} : Y_n = X_{k_n} \to X \text{ and } \beta : Y \to X \text{ is the unique function such that } \beta_n = \beta \circ \psi_n, \text{ we have} \]
\[ \beta \circ \alpha \circ \phi_n = \beta \circ \alpha_n = \beta \circ \psi_n \circ \phi_{n,k_n} = \beta_n \circ \phi_{n,k_n} = \phi_{k_n} \circ \phi_{n,k_n} = \phi_n \]
\[ \alpha \circ \beta \circ \psi_n = \alpha \circ \beta_n = \alpha \circ \phi_{k_n} = \psi_{k_n} \circ \phi_{k_n,k_{k_n}} = \phi_{k_n} \circ \phi_{k_n,k_{k_n}} = \phi_{k_n} = \psi_n \]
so we have $\beta \circ \alpha = \iota_X$ and $\alpha \circ \beta = \iota_Y$, i.e. $\alpha$ and $\beta$ are mutually inverse bijections.

Thus the direct limit over a subsystem is the “same” as the direct limit over the whole system. This is an analog (in a quite different setting) of the fact that the limit of a subsequence is the same as the limit of the whole sequence (if it exists).

More generally, the same result holds for direct limits over arbitrary directed sets (XI.14.1.10.).

1168
Induced Functions

**XI.14.1.8.** Let \((X_i, \phi_{i,j})\) and \((Y_i, \psi_{i,j})\) be direct systems over the same index set \(I\). Suppose there is for each \(i\) a function \(\alpha_i : X_i \to Y_i\) such that \(\alpha_j \circ \phi_{i,j} = \psi_{i,j} \circ \alpha_i\) for all \(i < j\). Let \((X, \phi_i) = \lim (X_i, \phi_{i,j})\), \((Y, \psi_i) = \lim (Y_i, \psi_{i,j})\). The maps \(\psi_i \circ \alpha_i\) are a coherent family of maps from the \(X_i\) to \(Y_i\), so by the universal property there is a unique function \(\alpha : X \to Y\) such that \(\alpha \circ \phi_i = \psi_i \circ \alpha_i\) for all \(i\). The function \(\alpha\) is called the *function induced by the \(\alpha_i\)*. For \(I = \mathbb{N}\) we have a commutative diagram:

![Diagram](image-url)

**XI.14.1.9.** Here is a general procedure including both **XI.14.1.7.** and **XI.14.1.8.**. Let \(I\) and \(J\) be directed sets, and \((X_i, \phi_{i,j})\) and \((Y_j, \psi_{j,j'})\) be inductive systems over \(I\) and \(J\) respectively. (Note that we have slightly changed our standard notation.) Let \(K\) be a cofinal subset of \(I\), and \(f\) a directed function \(\ell\) from \(K\) to \(J\), which for simplicity we will take to be order-preserving (this assumption can be eliminated at some technical cost, in the manner of Exercise II.4.5.3.). Suppose we have a compatible set \((\alpha_k)\) of functions \(\alpha_k : X_k \to Y_{f(k)}\) (i.e. if \(k < k'\), we have \(\alpha_{k'} \circ \phi_{k,k'} = \psi_{f(k),f(k')} \circ \alpha_k\)). For \(i \in I\), define \(\theta_i : X_i \to Y\) by \(\theta_i = \psi_{f(k)} \circ \alpha_k \circ \phi_{i,k}\) for any \(k \in K\), \(i < k\) (by compatibility this is well defined independent of the choice of \(k\)). The \(\theta_i\) are a compatible family of functions from the \(X_i\) to \(Y\), so there is a unique induced map \(\alpha : X \to Y\) such that \(\alpha \circ \phi_i = \theta_i\) for all \(i\). This is the function induced by \(f\) and the \(\alpha_k\).

**XI.14.1.10.** **Proposition.** Let \(I\) and \(J\) be directed sets, and \((X_i, \phi_{i,j})\) and \((Y_j, \psi_{j,j'})\) be direct systems over \(I\) and \(J\) respectively, with direct limits \((X, \phi_i)\) and \((Y, \psi_j)\). Let \(K\) and \(L\) be cofinal subsets of \(I\) and \(J\) respectively, and \(f : K \to J\) and \(g : L \to I\) order-preserving directed maps. Let \(\alpha_k : X_k \to Y_{f(k)}\) and \(\beta_k : Y_k \to X_{g(\ell)}\) be compatible families of maps, with \(\beta_k \circ \psi_{f(k),\ell} \circ \alpha_k = \phi_{k,g(\ell)}\) whenever \(k \in K\), \(\ell \in L\), \(f(k) < \ell\) and \(k < g(\ell)\), and \(\alpha_k \circ \phi_{g(\ell),k} \circ \beta_k = \psi_{\ell,f(k)}\) whenever \(g(\ell) < k\) and \(\ell < f(k)\). Then the induced functions \(\alpha : X \to Y\) and \(\beta : Y \to X\) are mutually inverse bijections.

The proof is very similar to the proof of **XI.14.1.7.**, and is left to the reader. Pictorially, if \(I = J = \mathbb{N}\), we have a commutative diagram:

![Diagram](image-url)
XI.14.1.11. Let \((X_i, \phi_{i,j})\) be a direct system of sets over \(I\), and for each \(i\) suppose \(Y_i \subseteq X_i\) and \(\phi_{i,j}(Y_i) \subseteq Y_j\) whenever \(i < j\). Set \(\psi_{i,j} = \phi_{i,j}^{-1} : Y_i \to Y_j\) for \(i < j\). Then \((Y_i, \psi_{i,j})\) is a direct system of sets over \(I\), and the inclusion maps \(\beta_i : Y_i \to X_i\) induce a function \(\beta : Y \to X\), where \((X, \phi_i) = \lim Y_i, \phi_{i,j}\) and \((Y, \psi_i) = \lim Y_i, \psi_{i,j}\).

XI.14.1.12. **Proposition.** The map \(\beta : Y \to X\) is injective; thus we may write \(\lim Y_i, \psi_{i,j}\) \(\subseteq \lim (X_i, \phi_{i,j})\). (Informally, direct limits of sets commute with taking subsets.)

**Proof:** Suppose \(\beta(y) = \beta(y')\). Let \(y = \psi_i(z)\), \(y' = \psi_j(z')\) for some \(z \in Y_i\), \(z' \in Y_j\). Then there is a \(k\) with \(i < k, j < k\), and \(\phi_{i,k}(z) = \phi_{j,k}(z')\). But then \(\psi_{i,k}(z) = \psi_{j,k}(z')\), so \(y = \psi_i(z) = \psi_j(z') = y'\). \(\square\)

XI.14.2. **Inverse Systems and Inverse Limits of Sets**

Inverse systems and limits work similarly to direct ones, but the arrows are reversed. The details turn out to be subtly different.

XI.14.2.1. **Definition.** An **inverse system of sets** over \(\mathbb{N}\) is a sequence \((X_n)\) of sets and, for each \(n\), a function \(\phi_{n+1,n} : X_{n+1} \to X_n\).

Pictorially, we have

\[
\cdots \xrightarrow{\phi_{4,3}} X_3 \xrightarrow{\phi_{3,2}} X_2 \xrightarrow{\phi_{2,1}} X_1
\]

If \(n < m\), then there is an induced map \(\phi_{m,n}\) from \(X_m\) to \(X_n\) defined by composition:

\[
\phi_{m,n} = \phi_{n+1,n} \circ \phi_{n+2,n+1} \circ \cdots \circ \phi_{m-2,m-1} \circ \phi_{m,m-1}.
\]

If \(n < m < k\), then \(\phi_{k,n} = \phi_{m,n} \circ \phi_{k,m}\).

We can again work over any directed set:

XI.14.2.2. **Definition.** Let \(I\) be a (properly) directed set. An **inverse system of sets over \(I\)** is an indexed collection \(\{X_i : i \in I\}\) of sets and, for each \(i, j\) with \(i < j\), a function \(\phi_{j,i} : X_j \to X_i\) satisfying \(\phi_{j,i} = \phi_{j,i} \circ \phi_{k,j}\) whenever \(i < j < k\).

We usually write \((X_i, \phi_{j,i})\) or \((X_i, \phi_{j,i})_{i \in I}\) for an inverse system of sets over \(I\).

XI.14.2.3. We want to associate to an inverse system a set \(X\) which “sits at \(-\infty\)”; \(X\) will map to each \(X_i\). To see how this works, first consider the case where \(I = \mathbb{N}\) and each \(\phi_{n+1,n}\) is injective (this is not actually a very interesting case for inverse limits). Then \(X_{n+1}\) can be identified with the subset \(\phi_{n+1,n}(X_{n+1})\) of \(X_n\) for each \(n\), and the inverse limit \(X\) is then the “intersection” of the \(X_n\). (It is not truly an intersection except in the rare case where each \(\phi_{n+1,n}\) is an inclusion map of a subset.)

If the connecting maps are surjective but not injective (the most interesting case), the inverse limit will be “larger” than the \(X_i\), since the preimage of any element of \(X_i\) will be at least as large as the preimage under \(\phi_{j,i}\), which increases in size as \(j\) increases.

Here is the universal property the inverse limit must have:
XI.14.2.4. Definition. Let \((X_i, \phi_{j,i})\) be an inverse system over a directed set \(I\). An inverse limit for the inverse system is a pair \((X, \phi_i)\), where \(X\) is a set and, for each \(i\), \(\phi_i : X \to X_i\) is a function such that \(\phi_i = \phi_{j,i} \circ \phi_j\) for any \(i, j\) with \(i < j\); pictorially, in the case \(I = \mathbb{N}\), we have a commutative diagram:

\[
\begin{array}{cccccc}
X & \rightarrow & X_3 & \rightarrow & X_2 & \rightarrow & X_1 \\
\phi_1 & & \phi_{2,3} & & \phi_{3,2} & & \phi_{4,3}
\end{array}
\]

with the following universal property:

If \(Y\) is a set and for each \(i\) there is a function \(\psi_i : Y \to X_i\) such that \(\psi_i = \phi_{j,i} \circ \psi_j\) for any \(i, j\) with \(i < j\), then there is a unique function \(\alpha : Y \to X\) such that \(\psi_i = \phi_i \circ \alpha\) for all \(i\).

We write \((X, \phi_i) = \lim_{\leftarrow} (X_i, \phi_{j,i})\) for an inverse limit. An inverse limit is often called a projective limit.

Pictorially, the universal property in the case \(I = \mathbb{N}\) is the following commutative diagram:

\[
\begin{array}{cccccc}
X & \rightarrow & X_3 & \rightarrow & X_2 & \rightarrow & X_1 \\
\phi_1 & & \phi_{2,3} & & \phi_{3,2} & & \phi_{4,3}
\end{array}
\]

XI.14.2.5. Theorem. Let \((X_i, \phi_{j,i})\) be an inverse system of sets over a directed set \(I\). Then

(i) There is an inverse limit \((X, \phi_i)\).

(ii) If \((X, \phi_i)\) and \((Y, \psi_i)\) are inverse limits, there is a unique bijection \(\alpha : Y \to X\) such that \(\psi_i = \phi_i \circ \alpha\) for all \(i\).

Thus up to natural identification there is a unique inverse limit for any inverse system.

Proof: The proof of (ii) is essentially identical to the proof of XI.14.5.(ii), and is left to the reader.

The argument for (i) is somewhat simpler than for direct limits once infinite Cartesian products have been defined (). Let \(Z\) be the Cartesian product of the \(X_i\), and define

\[X = \{(\cdots x_i \cdots) : x_i = \phi_{j,i}(x_j) \text{ for all } i, j \in I, i < j\} \subseteq \prod_{i \in I} X_i = Z\]

(the elements of \(X\) are called coherent threads in \(Z\)). For each \(i\), let \(\phi_i\) be the restriction to \(X\) of the \(i\)'th coordinate map, i.e.

\[\phi_i(\cdots x_i \cdots) = x_i\ .\]

It is obvious from the coherence that \(\phi_i = \phi_{j,i} \circ \phi_j\) if \(i < j\).
To show the universal property, let \((Y, \psi_i)\) satisfy the coherence condition. Simply define \(\alpha : Y \to X\) by
\[
\alpha(y) = (\cdots \psi_1(y) \cdots).
\]
Then clearly \(\psi_i = \phi_i \circ \alpha\) for all \(i\), and \(\alpha\) is the only possible function from \(Y\) to \(X\) with this property. 

**XI.14.2.6. Proposition.** Let \((X, \phi_i) = \lim_{\longleftarrow}(X_i, \phi_{j,i})\). If \(x, x' \in X\), then \(x = x'\) if and only if \(\phi_i(x) = \phi_i(x')\) for all \(i\).

**XI.14.2.7. Examples.** (i) Let \(X\) be a set and \(i\) the identity map on \(X\). Let \(I\) be a directed set. For \(i \in I\) set \(X_i = X\), and for \(i < j\) set \(\phi_{j,i} = i\). Then \((X_i, \phi_{j,i})\) is an inverse system over \(I\), and \((X, \iota)\) is an inverse limit for \((X_i, \phi_{j,i})\).

(ii) Let \(Y\) be a set, and \((X_n)\) a decreasing sequence of subsets of \(Y\). Let \(\phi_{n+1,n}\) be the inclusion map from \(X_{n+1}\) to \(X_n\). Let \(X = \bigcap_n X_n\) and \(\phi_n\) the inclusion map from \(X\) to \(X_n\). Then \((X_n, \phi_{n+1,n})\) is an inverse system over \(\mathbb{N}\), and \((X, \phi_n)\) is an inverse limit for \((X_n, \phi_{n+1,n})\). Here \(\mathbb{N}\) can be replaced by an arbitrary index set \(I\), if \(X_j \subseteq X_i\) whenever \(i < j\).

(iii) As a special case of (ii), let \(X_n = \{k \in \mathbb{N} : k \geq n\}\). Then \((X_n, \phi_{n+1,n})\) is an inverse system of nonempty sets with \(\lim(X_n, \phi_{n+1,n}) = \emptyset\).

It can even happen that the inverse limit of an inverse system of nonempty sets with surjective connecting maps is empty (Exercise XI.14.5.5.(b)), but not if the index set is \(\mathbb{N}\) (Exercise XI.14.5.5.(a)).

We will see much more interesting examples later.

**Induced Functions and Passage to Subsystems**

Induced functions and passage to subsystems in inverse limits works almost identically to the direct limit case.

**XI.14.2.8.** Suppose \((X_n, \phi_{n+1,n})\) is an inverse system over \(\mathbb{N}\). Let \((k_n)\) be a strictly increasing sequence in \(\mathbb{N}\), and let \(Y_n = X_{k_n}, \psi_{n+1,n} = \phi_{k_{n+1},k_n}\). If \((X, \phi_n) = \lim(X_n, \phi_{n+1,n})\) and \((Y, \psi_n) = \lim(Y_n, \psi_{n+1,n})\), we will show there is a natural bijection between \(X\) and \(Y\). For each \(n\), set \(\alpha_n = \phi_{k_n} : X \to Y_n = X_{k_n}\); there is then a unique \(\alpha : X \to Y\) such that \(\alpha_n = \psi_{n} \circ \alpha\) for all \(n\). If \(\beta_n = \phi_{k_n} \circ \psi_n : Y \to X_n\) and \(\beta : Y \to X\) is the unique function such that \(\beta_n = \phi_n \circ \beta\), we have
\[
\phi_n \circ \beta \circ \alpha = \beta_n \circ \alpha = \phi_{k_n, n} \circ \psi_n \circ \alpha = \phi_{k_n, n} \circ \alpha_n = \phi_{k_n} \circ \phi_{k_n} = \phi_n
\]
\[
\psi_n \circ \alpha \circ \beta = \alpha_n \circ \beta = \phi_k \circ \beta = \beta_k = \phi \circ \psi = \phi \circ \psi = \phi = \psi
\]
so we have \(\beta \circ \alpha = \iota_X\) and \(\alpha \circ \beta = \iota_Y\), i.e. \(\alpha\) and \(\beta\) are mutually inverse bijections.
XI.14.2.9. Let \((X_i, \phi_{j,i})\) and \((Y_i, \psi_{j,i})\) be inverse systems over the same index set \(I\). Suppose there is for each \(i\) a function \(\alpha_i : X_i \rightarrow Y_i\) such that \(\psi_{j,i} \circ \alpha_j = \alpha_i \circ \phi_{j,i}\) for all \(i < j\). Let \((X, \phi_i) = \lim(X_i, \phi_{j,i}),\) \((Y, \psi_i) = \lim(Y_i, \psi_{j,i})\). The maps \(\alpha_i \circ \psi_i\) are a coherent family of maps from \(X\) to \(Y_i\), so by the universal property there is a unique function \(\alpha : X \rightarrow Y\) such that \(\psi_i \circ \alpha = \alpha_i \circ \phi_i\) for all \(i\). The function \(\alpha\) is called the \textit{function induced by the} \(\alpha_i\). For \(I = \mathbb{N}\) we have a commutative diagram:

![Diagram](https://via.placeholder.com/150)

XI.14.2.10. Here is the general procedure including both XI.14.2.8. and XI.14.2.9.. Let \(I\) and \(J\) be directed sets, and \((X_i, \phi_{j,i}')\) and \((Y_j, \psi_{j,j})\) be inverse systems over \(I\) and \(J\) respectively. Let \(K\) be a cofinal subset of \(I\), and \(f\) a directed function \(f\) from \(K\) to \(J\), which for simplicity we will take to be order-preserving (this assumption can be eliminated at some technical cost, in the manner of Exercise II.4.5.3.). Suppose we have a compatible set \((\alpha_k)\) of functions \(\alpha_k : X_k \rightarrow Y_{f(k)}\) (i.e. if \(k < k'\) we have \(\alpha_k \circ \phi_{k',k} = \psi_{f(k'),f(k)} \circ \alpha_{k'}\)). For \(j \in J\), define \(\theta_j : X \rightarrow Y_j\) by \(\theta_j = \psi_{f(k),j} \circ \alpha_k \circ \phi_k\) for any \(k \in K, j < f(k)\) (by compatibility this is well defined independent of the choice of \(k\)). The \(\theta_j\) are a compatible family of functions from \(X\) to the \(Y_j\), so there is a unique induced map \(\alpha : X \rightarrow Y\) such that \(\psi_j \circ \alpha = \theta_j\) for all \(j\). This is the function induced by \(f\) and the \(\alpha_k\).

XI.14.2.11. \textbf{Proposition.} Let \(I\) and \(J\) be directed sets, and \((X_i, \phi_{j,i}')\) and \((Y_j, \psi_{j,j})\) be inverse systems over \(I\) and \(J\) respectively, with inverse limits \((X, \phi_i)\) and \((Y, \psi_j)\). Let \(K\) and \(L\) be cofinal subsets of \(I\) and \(J\) respectively, and \(f : K \rightarrow J\) and \(g : L \rightarrow I\) order-preserving directed maps. Let \(\alpha_k : X_k \rightarrow Y_{f(k)}\) and \(\beta_{\ell} : Y_{\ell} \rightarrow X_{g(\ell)}\) be compatible families of maps, with \(\beta_{\ell} \circ \psi_{f(k),\ell} \circ \alpha_k = \phi_{k,g(\ell)}\) whenever \(k \in K, \ell \in L, \ell < f(k)\) and \(k < g(\ell)\), and \(\alpha_k \circ \phi_{g(\ell),k} \circ \beta_{\ell'} = \psi_{f(k'),f(k)}\) whenever \(k < g(\ell)\) and \(\ell < f(k)\). Then the induced functions \(\alpha : X \rightarrow Y\) and \(\beta : Y \rightarrow X\) are mutually inverse bijections.

The proof is very similar to the proof of XI.14.1.7., and is left to the reader. Pictorially, if \(I = J = \mathbb{N}\), we have a commutative diagram:

![Diagram](https://via.placeholder.com/150)

XI.14.2.12. Like direct limits, inverse limits commute with taking subspaces. Let \((X_i, \phi_{j,i})\) be an inverse system of sets over \(I\), and for each \(i\) suppose \(Y_i \subseteq X_i\) and \(\phi_{j,i}(Y_j) \subseteq Y_i\) whenever \(i < j\). Set
$\psi_{j,i} = \phi_{j,i} y_j : Y_j \to Y_i$ for $i < j$. Then $(Y_i, \psi_{j,i})$ is an inverse system of sets over $I$, and the inclusion maps $\beta_i : Y_i \to X_i$ induce a function $\beta : Y \to X$, where $(X, \phi_i) = \varprojlim (X_i, \phi_{j,i})$ and $(Y, \psi_i) = \varprojlim (Y_i, \psi_{j,i})$. The next proposition is obvious:

**XI.14.2.13. Proposition.** The map $\beta : Y \to X$ is injective; thus we may write $\varprojlim (Y_i, \psi_{j,i}) \subseteq \varprojlim (X_i, \phi_{j,i})$. (Informally, inverse limits of sets commute with taking subsets.)

### XI.14.3. Direct Limits in Topology

If $(X_i, \phi_{i,j})$ is a direct system over a directed set $I$, and each $X_i$ is a topological space and all $\phi_{i,j}$ are continuous, then we want to put a topology on the direct limit so that all the maps in Definition XI.14.1.4. are continuous. So we need each $\phi_i$ continuous, and for the universal property to hold the topology must be the strongest one with this property, i.e. the topology on $X$ is induced by the functions $\phi_i$ in the sense of ()

**XI.14.3.1. Definition.** Let $(X_i, \phi_{i,j})$ be a direct system of topological spaces over a directed set $I$, i.e. each $X_i$ is a topological space and each $\phi_{i,j} : X_i \to X_j$ is continuous for $i < j$. The *inductive limit topology* on $X = \varinjlim (X_i, \phi_{i,j})$ is the topology on $X$ induced by the $\phi_i$, i.e. $U \subseteq X$ is open if and only if $\phi_i^{-1}(U)$ is open in $X_i$ for all $i$.

It is easily checked (cf. ()) that this is a topology on $X$ and that each $\phi_i$ is continuous.

**XI.14.3.2. Proposition.** Let $(X_i, \phi_{i,j})$ be a direct system of topological spaces, with $(X, \phi_i) = \varinjlim (X_i, \phi_{i,j})$. Give $X$ the inductive limit topology. If $Y$ is a topological space and for each $i$ there is a continuous function $\psi_i : X_i \to Y$ such that $\psi_i = \psi_j \circ \phi_{i,j}$ for any $i, j$ with $i < j$, then the unique function $\alpha : X \to Y$ such that $\psi_i = \alpha \circ \phi_i$ for all $i$ is continuous.

**Proof:** Let $V$ be an open set in $Y$, and $U = \alpha^{-1}(V)$. Then $\phi_i^{-1}(U) = \psi_i^{-1}(V)$ is open in $X_i$ for each $i$. Thus $U$ is open in $X$ by definition of the inductive limit topology.

**XI.14.3.3.** There is almost nothing one can say about properties of the inductive limit topology in general; it can be quite pathological (cf. Example XI.14.3.4.(i)). See [], [], ..., for other examples. But the inductive limit topology tends to be fairly strong, as Example XI.14.3.4.(ii) indicates.

**XI.14.3.4. Examples.** (i) Let $X$ be a set, and $\mathcal{T}_1, \mathcal{T}_2, \ldots$ a sequence of topologies on $X$ with $\mathcal{T}_{n+1}$ weaker than $\mathcal{T}_n$ for all $n$. Let $X_n$ be $X$ with the topology $\mathcal{T}_n$ for each $n$, and let $\phi_{n,n+1}$ be the identity map on $X$. Then $\phi_{n,n+1}$ is continuous. The direct limit is again $X$, and the inductive limit topology is $\mathcal{T} = \cap_n \mathcal{T}_n$. If each $\mathcal{T}_n$ is $\mathcal{T}_1$, then $\mathcal{T}$ is also $\mathcal{T}_1$; but if each $\mathcal{T}_n$ is Hausdorff (even metrizable), then $\mathcal{T}$ is not necessarily Hausdorff ()

(ii) Here is a common situation which can be generalized in several ways. Let $(X_n, \rho_n)$ be a metric space for each $n$, and $\phi_{n,n+1}$ an isometry of $X_n$ onto a closed subset of $X_{n+1}$. There is an induced metric $\rho$ on the direct limit $X$, since we can identify $X_n$ with a closed subset of $X$, $X_n \subseteq X_{n+1}$, $X = \cup_n X_n$. It might appear that the inductive limit topology is the topology induced by $\rho$. This is sometimes true, e.g. if $X_n = [0,1]$ with the usual metric, then $X \approx (0,1]$ with the usual topology. But it is false in general. In

\[1174\]
fact, the inductive limit topology is strictly stronger than the $\rho$-topology in general, and is often not even first countable, so sequences are inadequate to describe the topology. An example is obtained by taking $X_n = \mathbb{R}^n$ with the usual metric (or the $p$-metric for $1 \leq p \leq \infty$), identifying $\mathbb{R}^n$ with $\mathbb{R}^n \times \{0\} \subseteq \mathbb{R}^{n+1}$. The direct limit can be described in this case as $c_{\infty}$, the set of sequences of real numbers which are eventually 0. XI.14.3.5. implies that a sequence in $c_{\infty}$ (i.e. a sequence of sequences) converges in the inductive limit topology if and only if there is a fixed $n$ such that all the terms in each sequence vanish after the $n$'th place and the sequence of $k$'th coordinates converges for $1 \leq k \leq n$. The inductive limit topology on $c_{\infty}$ is not first countable: if $\sigma = (\epsilon_n)$ is any sequence of positive numbers, let $U_\sigma$ be the set of all sequences $(x_1, x_2, \ldots)$ in $c_{\infty}$ such that $|x_n| < \epsilon_n$ for all $n$. Then $U_\sigma \cap \mathbb{R}^n$ is open in $\mathbb{R}^n$ for all $n$, hence $U_\sigma$ is open in the inductive limit topology. In fact, the $U_\sigma$ form a local base at 0 for the inductive limit topology. But there is no countable local base at 0, since for any countable set $\{\sigma_k : k \in \mathbb{N}\}$ of such sequences, there is a $\sigma$ whose terms go to 0 faster than those of any $\sigma_k$.

In the situation of XI.14.3.4., we have the following result for sequences (which does not generalize to nets):

XI.14.3.5. Proposition. Let $(X_n, \rho_n)$ be a metric space for each $n$, and $\phi_{n,n+1}$ an isometry of $X_n$ onto a closed subset of $X_{n+1}$. Let $X = \lim(X_n, \phi_{n,n+1})$. Give $X$ the inductive limit topology $T$, and identify $X_n$ with a (closed) subset of $X$ for each $n$. If $(x_k)$ is a sequence in $X$, $x \in X$, then $x_k \to x$ in the $T$-topology if and only if there is an $n$ such that $x_k \in X_n$ for all $k$, $x \in X_n$, and $x_n \to x$ in $X_n$.

XI.14.3.6. Caution: Suppose $(X_i, \phi_{i,j})$ is a direct system of topological spaces, and for each $i$ we have $Y_i \subseteq X_i$, with $\phi_{i,j}(Y_i) \subseteq Y_j$ for $i < j$. Set $\psi_{i,j} = \phi_{i,j}|Y_i : Y_i \to Y_j$ for $i < j$. Give $Y_i$ the induced topology from $X_i$. Then $(Y_i, \psi_{i,j})$ is a direct system of topological spaces. If $\beta_i$ is the inclusion map of $Y_i$ into $X_i$ for each $i$, there is an induced map $\beta : Y \to X$, where $(X, \phi_{i,j}) = \lim(X_i, \phi_{i,j})$ and $(Y, \psi_i) = \lim(Y_i, \psi_{i,j})$. The map $\beta$ is injective by XI.14.1.12. and continuous by XI.14.3.2.

However, $\beta$ is not a homeomorphism onto its image in general. For example (from [Dug78]), let each $X_n$ be $[0, 1]$ with its usual topology, and let $\phi_{n,n+1}$ be the identity map. Let $Y_n = \{0\} \cup \left(\frac{1}{n}, 1\right]$. Then $X$ and $Y$ can both be identified with $[0, 1]$ with $\beta$ the identity map. $X$ has the usual topology on $[0, 1]$, but $Y$ has a strictly stronger topology, since $\{0\}$ is open in $Y$.

Thus direct limits of topological spaces do not commute with taking subspaces in general.

XI.14.4. Inverse Limits in Topology

XI.14.4.1. Inverse limits in topology are somewhat better behaved than direct limits. Suppose $(X_i, \phi_{i,i})$ is an inverse system over a directed set $I$, and each $X_i$ is a topological space and all $\phi_{j,i}$ are continuous. Since the inverse limit $(X, \phi_i)$ is constructed as a subset of the Cartesian product, the obvious topology to use on $X$ is the restriction of the product topology; this coincides with the topology defined by the maps $\phi_i$ (weakest topology for which all $\phi_i$ are continuous; cf. ())). This topology is called the inverse limit topology, but is rarely referred to by name since it is the only reasonable topology on the inverse limit. Another way to describe this topology is: if $(x_\sigma)$ is a net in $X$ and $x \in X$, then $x_\sigma \to x$ in $X$ if and only if $\phi_i(x_\sigma) \to \phi_i(x)$ in $X_i$ for all $i$ ()).
Thus $\varphi$: inverse system of topological spaces. If $i < j$

Suppose its image in the inverse limit case, as follows immediately from the fact that if $\lim_i x_i = x_i$. An inverse limit of a sequence of metrizable spaces is metrizable. These separation properties are preserved under products and taking subspaces. Normality is not, however, preserved in general. An inverse limit of a sequence of metrizable spaces do commute with taking subspaces.

If each $X_i$ is Hausdorff (resp. $T_0$, $T_1$, regular, completely regular), so is the inverse limit, since these separation properties are preserved under products and taking subspaces. Normality is not, however, preserved in general. An inverse limit of a sequence of metrizable spaces is metrizable.

**XI.14.4.3.** Inverse limits behave better than direct limits with respect to taking subspaces. Suppose $(X_i, \phi_{j,i})$ is an inverse system of topological spaces, and for each $i$ we have $Y_i \subseteq X_i$, with $\phi_{j,i}(Y_j) \subseteq Y_i$ for $i < j$. Set $\psi_{j,i} = \phi_{j,i}|_{Y_j} : Y_j \to Y_i$ for $i < j$. Give $Y_i$ the induced topology from $X_i$. Then $(Y_i, \psi_{j,i})$ is an inverse system of topological spaces. If $\beta_i$ is the inclusion map of $Y_i$ into $X_i$ for each $i$, there is an induced map $\beta : Y \to X$, where $(X, \phi_i) = \lim(X_i, \phi_{j,i})$ and $(Y, \psi_i) = \lim(Y_i, \psi_{j,i})$. The map $\beta$ is injective by XI.14.2.13. and continuous by XI.14.4.2. In contrast to the direct limit case (XI.14.3.6.), $\beta$ is a homeomorphism onto its image in the inverse limit case, as follows immediately from the fact that if $\prod_i X_i$ is regarded as a subset of $\prod_i Y_i$, the relative topology on $\prod_i Y_i$ from the product topology on $\prod_i X_i$ is exactly the product topology on $\prod_i Y_i$. Thus inverse limits of topological spaces do commute with taking subspaces.

**XI.14.4.4.** If each $X_i$ is Hausdorff (resp. $T_0$, $T_1$, regular, completely regular), so is the inverse limit, since these separation properties are preserved under products and taking subspaces. Normality is not, however, preserved in general. An inverse limit of a sequence of metrizable spaces is metrizable.

**XI.14.4.5.** Proposition. Let $(X_i, \phi_{j,i})$ be an inverse system of Hausdorff spaces. Then the inverse limit is a closed subset of $\prod_i X_i$.

**Proof:** Let $(x_\sigma)$ be a net in $X = \lim(X_i, \phi_{j,i})$ with $x_\sigma \to x \in \prod_i X_i$. Write $x_\sigma = (\cdots x_\sigma^n \cdots)$, $x = (\cdots x_i \cdots)$. Suppose $i,j \in I$, $i < j$. Then, since coordinate projections are continuous, we have $x_\sigma^j \to x_i$ and $x_\sigma^j \to x_j$. Since $\phi_{j,i}$ is continuous, $x_i^j = \phi_{j,i}(x_\sigma^j) \to \phi_{j,i}(x_j)$. By uniqueness of limits in Hausdorff spaces, $x_i = \phi_{j,i}(x_j)$. Thus $x \in X$.

Combining this result with Tikhonov’s theorem (), we obtain:

**XI.14.4.6.** Corollary. Let $(X_i, \phi_{j,i})$ be an inverse system of compact Hausdorff spaces. Then $\lim(X_i, \phi_{j,i})$ is also a compact Hausdorff space.

**XI.14.4.7.** The conclusion of XI.14.4.5. fails in general if the $X_i$ are not Hausdorff (e.g. if they have the indiscrete topology). But it is nonetheless true that an inverse limit of compact spaces is compact (XI.14.5.6.).

Inverse limits of compact Hausdorff spaces behave somewhat better than general inverse limits. For example:
XI.14.4.8. Proposition. Let \((X_i, \phi_{j,i})\) be an inverse system of compact Hausdorff spaces over a directed set \(I\), and \((X, \phi_i) = \varprojlim(X_i, \phi_{i,j})\). For each \(i\) set \(Y_i = \cap_{j > i} \phi_{j,i}(X_j) \subseteq X_i\). Then for each \(i\) we have that \(\phi_i(X) = Y_i\). In particular, if each \(X_i\) is nonempty, then \(X\) is nonempty.

Proof: (AC) Note that \(Y_i\) is compact, and nonempty if all \(X_i\) are nonempty. Assume all \(X_i\) are nonempty. Fix \(i_0 \in I\), and let \(\mathcal{F}\) be the collection of all finite subsets of \(I\) containing \(i_0\). Let \(x_{i_0} \in Y_{i_0}\). Set \(Z = \prod_{i \in I} X_i\), and for each \(F \in \mathcal{F}\) let \(Z_F\) be the set of all \(\cdots x_i \cdots \) \(\in Z\) such that, whenever \(i, j \in F\), \(i < j\), we have \(x_i = \phi_{j,i}(x_j)\), and \(x_{i_0}\) is the specified element. Then \(Z_F\) is a nonempty compact subset of \(Z\) [choose \(j \in I\) such that \(i < j\) for all \(i \in F\), and let \(x_j \in X_j\) with \(\phi_{j,i_0}(x_j) = x_{i_0}\); set \(x_i = \phi_{j,i}(x_j)\) for each \(i \in F\)]. The collection \(\{Z_F : F \in \mathcal{F}\}\) has the finite intersection property, so the intersection is nonempty, and any point \(x \in \cap F Z_F\) is in the inverse limit \(X\) and satisfies \(\phi_{i_0}(x) = x_{i_0}\). Thus \(\phi_i(X)\) contains \(Y_i\), and the reverse containment is trivial (this containment holds in general inverse limits).

XI.14.4.9. This result fails for general inverse limits (Exercise XI.14.5.6.).

As a companion, it is easy to recognize inverse limits in the compact Hausdorff case:

XI.14.4.10. Proposition. Let \((X_i, \phi_{j,i})\) be an inverse system of compact Hausdorff spaces over a directed set \(I\), and \((X, \phi_i) = \varprojlim(X_i, \phi_{i,j})\). For each \(i\) set \(Y_i = \cap_{j > i} \phi_{j,i}(X_j) \subseteq X_i\). Let \(Y\) be a compact Hausdorff space, and \(\psi\) a compatible family of continuous functions from \(Y\) onto \(Y_i\) (i.e. \(\psi = \phi_{j,i} \circ \psi_j\) for \(i < j\) and \(\psi(Y) = Y_i\) for all \(i\)), which separate points, i.e. if \(y \neq y'\) there is an \(i\) with \(\psi_i(y) \neq \psi_i(y')\). Then \((Y, \psi_i)\) is an inverse limit for \((X_i, \phi_{j,i})\).

Proof: Let \((X, \phi_i) = \varprojlim(X_i, \phi_{j,i})\), and \(\alpha : Y \to X\) the induced map. It suffices to show that \(\alpha\) is a bijection since it is continuous (). Since the \(\psi\) separate points, \(\alpha\) is injective. To show surjectivity, let \(x = (\cdots x_i \cdots) \in X\). Then \(x_i \in Y_i\) for all \(i\). For each \(i\) let \(y_i \in Y\) with \(\psi_i(y_i) = x_i\). Then \((y_i)\) is a net in \(Y\); let \(y\) be a cluster point. Since

\[
\psi_i(y_j) = \phi_{j,i}(\psi_j(y_j)) = \phi_{j,i}(x_j) = x_i
\]

for \(i < j\), we have \(\psi_i(y) = x_i\) for all \(i\) by continuity, so \(\alpha(y) = x\).

XI.14.4.11. Example. For each \(n\) let \(X_n\) be a finite set with \(2^n\) points, with the discrete topology. We may conveniently take \(X_n\) to be the set of all \(n\)-tuples of 0’s and 1’s. Let \(\phi_{n+1,n} : X_{n+1} \to X_n\) be a function for which the inverse image of each point of \(X_n\) consists of exactly two points. For example, we may take

\[
\phi_{n+1,n}((d_1, \ldots, d_n, d_{n+1})) = (d_1, \ldots, d_n)
\]

where \((d_1, \ldots, d_{n+1})\) is an \((n+1)\)’tuple of 0’s and 1’s. The inverse image \((X, \phi_n) = \varprojlim(X_n, \phi_{n+1,n})\) is a compact Hausdorff space which is naturally homeomorphic to the Cantor set: it can be thought of as the set of all sequences of 0’s and 1’s, with the topology of coordinatewise convergence, with

\[
\phi_n((d_1, d_2, \ldots)) = (d_1, \ldots, d_n)
\]

In fact it is easy to conclude from XI.14.4.10. that the Cantor set with these maps is an inverse limit.
XI.14.4.12. **Examples.** Some of the examples from () can be easily written as inverse limits of “nicer” spaces by projection.

(i) Let $X$ be one of the spaces $X_{k}$ of XI.18.4.1. Let $X_{n}$ be the subset of $X$ consisting of the horizontal line and the “whiskers” whose basepoint has $x$-coordinate $\geq \frac{1}{n}$. Define $\phi_{n+1,n} : X_{n+1} \rightarrow X_{n}$ by mapping all the points on the whiskers with basepoint $(\frac{1}{n+1},0)$ to the point $(\frac{1}{n+1},0)$. Then $\lim_{\rightarrow} (X_{n}, \phi_{n+1,n})$ is homeomorphic to $X$, and the map $\phi_{n} : X \rightarrow X_{n}$ just sends all the points of each whisker whose basepoint has $x$-coordinate $< \frac{1}{n}$ to the basepoint of the whisker.

(ii) Similarly, let $X$ be the Hawaiian earring (XI.18.4.6.). Let $X_{n}$ be the subset of $X$ consisting only of the first $n$ circles (with radius $\geq \frac{1}{n}$). Define $\phi_{n+1,n} : X_{n+1} \rightarrow X_{n}$ by mapping all the points on the $(n+1)$'st circle to $(0,0)$. Then $\lim_{\rightarrow} (X_{n}, \phi_{n+1,n})$ is homeomorphic to $X$, and the map $\phi_{n} : X \rightarrow X_{n}$ just sends all the points of each circle of radius $< \frac{1}{n}$ to $(0,0)$. However,

(iii) Let $X$ be the infinite broom (Exercise XI.13.9.9.). For each $n$ let $X_{n}$ be the subset of $X$ consisting only of the line segments from the origin to $((1, \frac{k}{n}))$ for $1 \leq k \leq n$. Define $\phi_{n+1,n} : X_{n+1} \rightarrow X_{n}$ by projecting the $(n+1)$'st line segment upwards onto the $n$'th and leaving the points of $X_{n}$ fixed. Then $X$ is naturally homeomorphic to $\lim_{\rightarrow} (X_{n}, \phi_{n+1,n})$, with $\phi_{n}$ the projection of all segments with right endpoint below $((1, \frac{1}{n}))$ upwards onto the $n$'th segment.

(iv) Let $X$ be the topologist’s sine curve (XI.18.4.3.). Let $X_{n}$ be the subset of $\mathbb{R}^{2}$ obtained by taking the part of $X$ to the right of the vertical line $x = \frac{2}{2n+1}$ and the vertical line segment between $\left(\frac{2}{2n+1},1\right)$ and $\left(\frac{2}{4n+1},1\right)$. Then $X_{n}$ is homeomorphic to $[0,1]$. Define $\phi_{n+1,n} : X_{n+1} \rightarrow X_{n}$ by projecting the points with $x$-coordinates $< \frac{2}{2n+1}$ horizontally to the right onto the line $x = \frac{2}{2n+1}$ and leaving the other points fixed. Then $X$ is naturally homeomorphic to $\lim_{\rightarrow} (X_{n}, \phi_{n+1,n})$, with $\phi_{n}$ the projection of all points with $x$-coordinates $< \frac{2}{4n+1}$ horizontally to the right onto the line $x = \frac{2}{4n+1}$. Thus $X$ is an inverse limit of copies of $[0,1]$. See XI.14.4.18. for more examples of spaces which can be written as inverse limits of copies of $[0,1]$.

XI.14.4.13. **Examples.** For $m \in \mathbb{N}$, write $\mathbb{Z}/m\mathbb{Z}$ for the cyclic group with $m$ elements (we usually denote this group by $\mathbb{Z}_{m}$, but this notation will be reserved for something else here). Let $a = (a_{1}, a_{2}, \ldots)$ be a sequence in $\mathbb{N} \setminus \{1\}$ with $a_{k} | a_{k+1}$ for all $k$. Consider the inverse system

$$
\cdots \xrightarrow{\pi_{4,3}} \mathbb{Z}/a_{3}\mathbb{Z} \xrightarrow{\pi_{3,2}} \mathbb{Z}/a_{2}\mathbb{Z} \xrightarrow{\pi_{2,1}} \mathbb{Z}/a_{1}\mathbb{Z}
$$

where $\pi_{n+1,n}$ is the quotient map from $\mathbb{Z}/a_{n+1}\mathbb{Z}$ to $\mathbb{Z}/a_{n}\mathbb{Z}$. Let $(\mathbb{Z}_{a}, \pi_{n})$ be the inverse limit. Then $\mathbb{Z}_{a}$ is a compact totally disconnected abelian topological group. If $a_{n} = p^{a}$ for a fixed prime $p$, we write $\mathbb{Z}_{p}$ for $\mathbb{Z}_{a}$; $\mathbb{Z}_{p}$ is called the $p$-adic numbers (cf. ()).

There is a natural quotient map $\rho_{n}$ from $\mathbb{Z}$ to $\mathbb{Z}/a_{n}\mathbb{Z}$ for each $n$. The $\rho_{n}$ define a homomorphism $\rho : \mathbb{Z} \rightarrow \mathbb{Z}_{a}$. If $m$ is in the kernel of $\rho$, then $a_{n} | m$ for all $n$, hence $m = 0$ and $\rho$ is injective. But $\rho$ is not surjective: in fact, $\mathbb{Z}_{a}$ has cardinality $2^{\aleph_{0}}$ (it is homeomorphic to the Cantor set by ()). However, $\rho(\mathbb{Z})$ is dense in $\mathbb{Z}_{a}$ (XI.14.4.19.).

The group $\mathbb{Z}_{a}$ is often, but not always, torsion-free (XI.14.5.9.). In particular, $\mathbb{Z}_{p}$ is torsion-free, $\mathbb{Z}_{p}$ and more generally $\mathbb{Z}_{a}$ has a natural structure as a topological ring, but this structure is not easily definable from the inverse limit construction.

There are several results about writing a general space as an inverse limit of “nicer” spaces. For example:
XI.14.4.14. Theorem. Let $X$ be a compact metrizable space. Then there is an inverse system $(X_n, \phi_{n+1,n})$ over $\mathbb{N}$, with each $X_n$ a smooth compact manifold with boundary and $\phi_{n+1,n}$ a smooth map for each $n$, with $X$ homeomorphic to $\lim_{\leftarrow} (X_n, \phi_{n+1,n})$.

Proof: We may identify $X$ with a (compact) subset of the Hilbert cube $(0,1)^n$. Fix a sequence $(\epsilon_n)$ with $0 < \epsilon_{n+1} < \frac{\epsilon_n}{2}$ for all $n$. Let $Y_n$ be the projection of $X$ onto the first $n$ coordinates, and regard $Y_n$ as a subset of $[0,1]^n \subseteq \mathbb{R}^n$. Let $U_n$ be a neighborhood of $Y_n$ in $\mathbb{R}^n$ as in XI.23.10.1. with $\epsilon = \epsilon_n$, and set $X_n = U_n$.

Since the projection from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ is a contraction, it maps $X_{n+1}$ into $X_n$. Let $\phi_{n+1,n}$ be the restriction of the projection map to $X_{n+1}$. Then $(X_n, \phi_{n+1,n})$ is an inverse system of smooth compact manifolds with boundary, with smooth connecting maps.

We show that $X$ is homeomorphic to $\lim (X_n, \phi_{n+1,n})$. For each $n$ let $\psi_n$ be the restriction to $X$ of the projection from the Hilbert cube to $[0,1]^n$; then $\psi_n(X) = Y_n$ by definition of $Y_n$, and the conditions of XI.14.4.10. are satisfied.

If $X$ is a compact subset of $\mathbb{R}^m$, a simplified version of the argument shows that the $X_n$ can be taken to be $m$-dimensional.

XI.14.4.15. There are two drawbacks to this result, however:

(i) The connecting maps are not surjective.

(ii) If $X$ is finite-dimensional, the dimensions of the $X_n$ may be much larger than $\text{dim}(X)$. (If $\text{dim}(X) = d$, $X$ embeds in $\mathbb{R}^{2d+1}$ by the Whitney Embedding Theorem, so the $X_n$ can be chosen to have dimension $\leq 2d + 1$.)

These difficulties cannot be avoided if one wants to write $X$ as an inverse limit of smooth manifolds with boundary. But if we slightly more generally allow the $X_n$ to be polyhedra, we can do better:

XI.14.4.16. Theorem. Let $X$ be a compact metrizable space. Then there is an inverse system $(X_n, \phi_{n+1,n})$ over $\mathbb{N}$, with each $X_n$ a polyhedron and $\phi_{n+1,n}$ a surjective piecewise-linear map for each $n$, with $X$ homeomorphic to $\lim_{\leftarrow} (X_n, \phi_{n+1,n})$. If $\text{dim}(X) = d < \infty$, then the $X_n$ can be chosen so that $\text{dim}(X_n) = d$ for all $n$.

The proof is involved and uses some machinery, and will be given in (). This result was first proved by H. Freudenthal in 1937 [Fre37]. The examples in XI.14.4.11. and XI.14.4.12. are examples of writing compact metrizable spaces as inverse limits of polyhedra.

XI.14.4.17. Examples. (Solenoids.) Let $T$ be the unit circle in the plane, regarded as the compact topological group of complex numbers of absolute value 1. For each $n$, let $X_n = T$ and $\phi_{n+1,n} : X_{n+1} \to X_n$ defined by $\phi_{n+1,n}(z) = z^2$. Then $(X_n, \phi_{n+1,n})$ is an inverse system of compact Hausdorff spaces. Let $(X, \phi_n) = \lim_{\leftarrow} (X_n, \phi_{n+1,n})$. Then $X$ has a natural structure as a compact Hausdorff space which is an abelian topological group (XI.14.5.2.). In fact, it is easily seen using XI.14.4.10. that $X$ is homeomorphic to the solenoid () of type $2^\infty$. Thus a solenoid is homeomorphic to an inverse limit of circles.

The circle $T$ can be regarded as $\mathbb{R}/m\mathbb{Z}$ for any $m \in \mathbb{N}$. Thus we may regard the above $X$ as

$$\lim_{\leftarrow} (\mathbb{R}/2^{n-1}\mathbb{Z}, \pi_{n+1,n})$$
where \( \pi_{n+1,n} : \mathbb{R}/2^n \mathbb{Z} \to \mathbb{R}/2^{n-1} \mathbb{Z} \) is the natural quotient map. We have a commutative diagram

\[
\begin{array}{cccc}
\ldots & \xrightarrow{\pi} & \mathbb{R} & \xrightarrow{\pi} & \mathbb{R} & \xrightarrow{\pi} & \mathbb{R} & \xrightarrow{\pi} & \mathbb{R} \\
\ldots & \xrightarrow{\pi_{n+1,n}} & \mathbb{R}/2^n \mathbb{Z} & \xrightarrow{\pi_{n,n-1}} & \mathbb{R}/2^{n-1} \mathbb{Z} & \xrightarrow{\pi_{n-1,n-2}} & \mathbb{R}/2^{n-2} \mathbb{Z} & \xrightarrow{\pi_{n-2,n-3}} & \mathbb{R}/2^{n-3} \mathbb{Z}
\end{array}
\]

which induces a continuous group homomorphism \( \alpha : \mathbb{R} \to X \), which is injective. The range of \( \alpha \) is dense in \( X \) by XI.14.4.19, but is not all of \( X \): the range consists of all threads \( (x_n) \), where \( x_n \to 1 \) in \( T \) when \( \mathbb{R}/2^{n-1} \mathbb{Z} \) is identified with \( T \). Also, \( \alpha \) is not a homeomorphism onto its image: for example, we have \( \alpha(2^n) \to \alpha(0) \) in \( X \). In fact, the range of \( \alpha \), which is a subgroup of \( X \), is exactly the path component of the identity of \( X \). The other path components of \( X \) are exactly the cosets of \( \alpha(\mathbb{R}) \) in \( X \).

The other solenoids can be constructed by taking \( \phi_{n+1,n}(z) = z^{k_n} \) for a sequence \( (k_n) \) in \( \mathbb{N} \setminus \{1\} \).

**XI.14.4.18. Examples.**
1. **Knaster bucket-handle.** (cf. XI.18.4.28.) The Knaster bucket-handle can be written as an inverse limit of copies of \([0,1]\) by taking for each \( \phi_{n+1,n} \) the piecewise-linear map from \([0,1]\) to \([0,1]\) with \( \phi_{n+1,n}(0) = \phi_{n+1,n}(1) = 0 \) and \( \phi_{n+1,n} \left( \frac{1}{2} \right) = 1 \).

Other bucket-handles can be constructed similarly as inverse limits by taking a sequence \( (k_n) \) of natural numbers \( \geq 2 \) and letting \( \phi_{n+1,n} \) be the sawtooth function taking values 0 and 1 alternately at the numbers \( \frac{m}{2^{k_n+1}} \). The homeomorphism class of the bucket-handle is exactly determined by the equivalence class (XI.18.4.21) of the supernatural number \( \prod n k_n \). There is a qualitative difference between bucket-handles for which infinitely many \( k_n \) are even and ones with only finitely many even \( k_n \) (up to equivalence, ones with all \( k_n \) odd): the latter have two “endpoints” while the former have only one.

2. **Pseudo-Arc.** The pseudo-arc (XI.18.4.29) can also be written as an inverse limit of copies of \([0,1]\).

The maps \( \phi_{n+1,n} \) can also be taken to be piecewise-linear, but zigzag up and down similarly to the definition of a crooked embedding:

See [\( \square \) for details.

**XI.14.4.19. Proposition.** Let \( (X_i, \phi_{j,i}) \) be an inverse system of topological spaces, and \( (X, \phi_i) = \lim (X_i, \phi_{j,i}) \). If \( D \subseteq X \), then \( D \) is dense in \( X \) if and only if \( \phi_i(D) \) is dense in \( \phi_i(X) \subseteq X_i \) for all \( i \).

**Proof:** If \( D \) is dense in \( X \), then \( \phi_i(D) \) is dense in \( \phi_i(X) \) for all \( i \) by (\( \square \)). Conversely, suppose \( \phi_i(D) \) is dense in \( \phi_i(X) \) for each \( i \). Fix \( x \in X \). Let \( J \) be the set of all triples \((i, U, y)\), where \( i \in I \), \( U \) is an open neighborhood of \( \phi_i(x) \) in \( X_i \), and \( y \in D \), \( \phi_i(y) \in U \). Set \((i, U, y) < (i', V, y') \) if \( i < i' \) and \( \phi_{i',i}(V) \subseteq U \). Then \( J \) is a directed set. For \( j = (i, U, y) \in J \) set \( y_j = y \). Then \((y_j)\) is a net in \( D \), and \( y_j \to x \) since \( \phi_i(y_j) \to \phi_i(x) \) for all \( i \). Thus \( D \) is dense in \( X \).

In particular, if \( \phi_i(D) = X_i \) for all \( i \), then \( D \) is dense in \( X \) (this has a simplified direct proof using \( AC \)); but \( D \) can be a proper subset of \( X \) (XI.14.4.17).
XI.14.5. Exercises

XI.14.5.1. Modify the definition of direct and inverse systems and limits by also defining \( \phi_{i,i} \) to be the identity map on \( X_i \) for each \( i \), and replacing \( < \) by \( \leq \) throughout. Show that this modified definition also works over directed sets which are not properly directed. Describe the limit if the index set is not properly directed.

XI.14.5.2. (a) Let \( I \) be a directed set and \( G_i \) a group for each \( i \), with \( \phi_{i,j} : G_i \to G_j \) a homomorphism for \( i < j \). Let \( (G, \phi_i) = \lim \rightarrow (G_i, \phi_{i,j}) \). Show that \( G \) has a unique group structure making each \( \phi_i \) a homomorphism.

(b) If each \( G_i \) is abelian, show that \( G \) is abelian. What about the converse?

(c) Do the same if each \( G_i \) is a ring, vector space, or topological group.

(d) Do the analog of (a)–(c) for inverse systems and limits.

XI.14.5.3. [Dug78] Let \( T \) be the unit circle in the plane, regarded as the compact topological group of complex numbers of absolute value 1. For each \( n \), let \( X_n = T \) and \( \phi_{n,n+1} : X_n \to X_{n+1} \) defined by \( \phi_{n,n+1}(z) = z^n \). Then \( (X_n, \phi_{n,n+1}) \) is a direct system of compact Hausdorff spaces. Let \( (X, \phi_n) = \lim \rightarrow (X_n, \phi_{n,n+1}) \). Then \( X \) has a natural structure as a topological group (XI.14.5.2).

(a) Let \( H \) be the subgroup of \( T \) consisting of all roots of unity whose order is a power of 2. Show that \( X \) can be naturally identified with \( T/H \).

(b) Show that the inductive limit topology on \( X \) is the indiscrete topology.

(c) Repeat the exercise if \( \phi_{n,n+1}(z) = z^{k_n} \) for an arbitrary sequence \( (k_n) \) in \( \mathbb{N} \setminus \{1\} \).

XI.14.5.4. Construct an explicit countable subset \( A \) of \( c_0 \) such that 0 is in the closure of \( A \) in the inductive limit topology, but there is no sequence in \( A \) converging to 0. [For \( n, m \in \mathbb{N} \), consider the sequence with first term \( \frac{1}{n} \), \( n \)th term \( \frac{1}{m} \), and all other terms 0.]

XI.14.5.5. (a) Show (using DC) that if \( (X_n, \phi_{n+1,n}) \) is an inverse system of nonempty sets over \( \mathbb{N} \), and each \( \phi_{n+1,n} \) is surjective, and \( (X, \phi_n) = \lim \leftarrow (X_n, \phi_{n+1,n}) \), then \( X \) is nonempty and each \( \phi_n \) is surjective.

(b) (I. Farah) For each \( \alpha < \omega_1 \) let \( X_\alpha \) be the set of all injective functions from \( \alpha \) to \( \mathbb{N} \) for which the complement of the range is finite. If \( \alpha < \beta \), let \( \phi_{\beta,\alpha} : X_\beta \to X_\alpha \) be defined by \( \phi_{\beta,\alpha}(f) = f|_\alpha \). Then \( (X_\alpha, \phi_{\beta,\alpha}) \) is an inverse system of nonempty sets over the directed (well-ordered) set \( \omega_1 \). Show that each connecting map \( \phi_{\beta,\alpha} \) is surjective, but \( \lim \leftarrow (X_\alpha, \phi_{\beta,\alpha}) = \emptyset \).

XI.14.5.6. (a) If \( (X_i, \phi_{j,i}) \) is an inverse system of nonempty compact spaces, show that the inverse limit can be empty if the \( X_i \) are not Hausdorff. [Give the sets in XI.14.2.7.(iii) or XI.14.5.5.(b) the indiscrete topology or the finite complement topology.]

(b) Prove (using AC) that an inverse limit of compact spaces is compact (regarding \( \emptyset \) as a compact space) by adapting the proof of Tikhonov’s theorem.
XI.14.5.7. Let \((X_n, \phi_{n+1,n})\) be an inverse system over \(\mathbb{N}\) of nonempty finite sets with the discrete topology, and \((X, \phi_n) = \lim\limits_{\mathcal{I}} (X_n, \phi_{n+1,n})\).

(a) If \(\phi_{n+1,n}^{-1}(p)\) has at least two elements for each \(n\) and each \(p \in X_n\), show that \(X\) is homeomorphic to the Cantor set \([use()\]).

(b) Let \(X_n = \{k \in \mathbb{N} : k \leq n\}\), and define \(\phi_{n+1,n}(k) = k\) for \(k \leq n\), \(\phi_{n+1,n}(n + 1) = n\). Show that \(\lim\limits_{\mathcal{I}} (X_n, \phi_{n+1,n})\) is naturally homeomorphic to the one-point compactification of \(\mathbb{N}\) (equivalently, to \(\{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\} \subseteq \mathbb{R}\)).

XI.14.5.8. Let \(G\) be a group, \(I\) a directed set, and \(\{H_i : i \in I\}\) a collection of normal subgroups of \(G\) with \(H_j \subseteq H_i\) for \(i < j\). Set \(G_i = G/H_i\) for all \(i\). Then there is a natural quotient map \(\pi_j, i : G_j = G/H_j \to G/H_i = G_i\) for \(i < j\), and \((G_i, \pi_j, i)\) is an inverse system of groups. Set \((\hat{G}, \pi_i) = \lim\limits_{\mathcal{I}} (G_i, \pi_j, i)\). Then \(\hat{G}\) is a group, and there is a homomorphism \(\alpha : G \to \hat{G}\).

(a) Show that the kernel of \(\alpha\) is exactly \(\cap_i H_i\). In particular, \(\alpha\) is injective if (and only if) \(\cap_i H_i = \{e_G\}\).

(b) Show that \(\alpha\) is not surjective in general (cf. XI.14.4.13., XI.14.4.17.).

(c) If each \(G_i\) has a topology making it into a compact (Hausdorff) topological group, with the \(\pi_j, i\) continuous, then \(\hat{G}\) is a compact (Hausdorff) topological group in the inverse limit topology.

(d) If each \(G_i\) is finite, give \(G_i\) the discrete topology. Then \(\hat{G}\) is a compact totally disconnected (zero-dimensional) topological group. Every compact totally disconnected topological group arises in this manner as an inverse limit of finite groups ()

We may regard \(\hat{G}\) as a “completion” of \(G\) with respect to the subgroups \(H_i\). The examples of XI.14.4.13., XI.14.4.17., and ?? are special cases.

XI.14.5.9. Let \(\mathbb{Z}_a\) be the \(a\)-adic numbers as defined in XI.14.4.13. Show that \(\mathbb{Z}_a\) is torsion-free if and only if the supernatural number \(a_1 \prod_n a_n = \lim_{n \to \infty} a_n\) has all exponents either 0 or \(\infty\). [Use that \(\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}\) if \(m\) and \(n\) are relatively prime () and XI.14.5.11.]

XI.14.5.10. Work out the details of writing the Cantor set, the Hawaiian earring, the infinite broom, the topologist’s sine curve, solenoids, Knaster bucket-handles, and the pseudo-arc as inverse limits as described in XI.14.4.11., XI.14.4.12., XI.14.4.17., and XI.14.4.18..

XI.14.5.11. Let \(I\) be an infinite index set (not ordered or directed). For each \(i \in I\), let \(X_i\) be a set. Let \(\mathcal{F}\) be the set of all finite subsets of \(I\), directed by inclusion. For each \(F \in \mathcal{F}\), define

\[Z_F = \prod_{i \in F} X_i\]

and if \(F \subseteq G \in \mathcal{F}\), let \(\pi_{G,F} : Z_G \to Z_F\) be the projection onto the coordinates in \(F\). Show that

\[\left(\prod_{i \in I} X_i, \pi_F\right)\]

is an inverse limit for \((Z_F, \pi_F)\), where \(\pi_F\) is the natural projection from \(\prod_{i \in I} X_i\) onto \(\prod_{i \in F} X_i\). If each \(X_i\) is a topological space, show that the inverse limit topology is the product topology.
Thus, once finite Cartesian products have been defined, infinite Cartesian products may be regarded as a special case of inverse limits. (However, this is not a logically sound way to define infinite Cartesian products, since infinite Cartesian products are used to construct inverse limits so there is a circularity problem.)

XI.14.5.12. Restricted Direct Products. Let $I$ be an infinite index set (not ordered or directed). For each $i \in I$, let $X_i$ be a locally compact Hausdorff space and $Y_i$ a compact open (hence clopen) subset of $X_i$. Let $\mathcal{F}$ be the set of all finite subsets of $I$, directed by inclusion. For each $F \in \mathcal{F}$, define

$$Z_F = \left\{ (\cdots, x_i, \cdots) \in \prod_{i \in I} X_i : x_i \in Y_i \text{ for all } i \notin F \right\} \subseteq \prod_{i \in I} X_i$$

and give $Z_F$ the restriction of the product topology. Then

$$Z_F \cong \prod_{i \in F} X_i \times \prod_{i \in I \setminus F} Y_i$$

and thus $Z_F$ is locally compact by (a). If $F \subseteq G \in \mathcal{F}$, the inclusion $\phi_{F,G}$ of $Z_F$ into $Z_G$ is a homeomorphism onto a clopen subset of $Z_G$. Let

$$(Z, \phi_F) = \lim\downarrow (Z_F, \phi_{F,G})$$

with the inductive limit topology. $Z$ is called the restricted direct product of the $X_i$ with respect to the $Y_i$, denoted

$$Z = \prod_{i \in I} (X_i, Y_i).$$

(a) As a set, $Z$ can be identified with the subset $\cup_F Z_F$ of $\prod_i X_i$, the set of tuples $(\cdots, x_i, \cdots)$ with $x_i \in Y_i$ for all but finitely many $i$.

(b) Show that $Z$ is locally compact and the image $\phi_F(Z_F)$ of $Z_F$ is a clopen subset (homeomorphic to $Z_F$) for each $F \in \mathcal{F}$.

(c) Show that the inductive limit topology on $Z$ is strictly stronger than the restriction to $Z$ of the product topology if $Y_i \neq X_i$ for infinitely many $i$. [The image $\phi_F(Z_F)$ is not open in $Z$ in the product topology.]

Restricted direct products are often used where the $X_i$ have additional structure such as a topological group (a). Here is one especially important application. The completions of the rational numbers $\mathbb{Q}$ to topological fields are precisely $\mathbb{R}$ and the $p$-adic numbers $\mathbb{Q}_p$ for primes $p$. Each of these is locally compact. The topological field $\mathbb{Q}_p$ contains the ring $\mathbb{Z}_p$ of $p$-adic integers as a compact open subring. Form

$$\mathbb{A}_\mathbb{Q} = \mathbb{R} \times \prod_p (\mathbb{Q}_p, \mathbb{Z}_p)$$

which is naturally a locally compact topological ring, called the ring of adeles of $\mathbb{Q}$. There is a natural embedding of $\mathbb{Q}$ into $\mathbb{A}_\mathbb{Q}$, whose image turns out to be dense. A similar construction can be done for any algebraic number field (finite extension of $\mathbb{Q}$). Rings of adeles are important in number theory; cf. \[.].
XI.14.5.13. (a) If $X$ and $Y$ are sets, use the notation $Y^X$ for the set of all functions from $X$ to $Y$. If $(X_i, \phi_{i,j})$ is a direct system of sets over $I$, and $Y$ is a set, for $i < j$ define $\phi_{i,j}^*: Y^{X_i} \to Y^{X_j}$ by

$$\phi_{i,j}^*(f) = f \circ \phi_{i,j}.$$ 

Then $(Y^{X_i}, \phi_{i,j}^*)$ is an inverse system of sets over $I$. If $(X, \phi)$ is a direct limit for $(X_i, \phi_{i,j})$, show that $(Y^X, \phi^*)$ is an inverse limit for $(Y^{X_i}, \phi_{i,j}^*)$, where $\phi^*(f) = f \circ \phi$. Thus we may write

$$\lim_{\text{inv}} (Y^{X_i}, \phi_{i,j}^*) \cong Y^{\lim_{\text{inv}} (X_i, \phi_{i,j})}.$$ 

(b) If $X$ and $Y$ are topological spaces, use the notation $Y^X$ for the set of all continuous functions from $X$ to $Y$. If $(X_i, \phi_{i,j})$ is a direct system of topological spaces over $I$, and $Y$ is a topological space, for $i < j$ define $\phi_{i,j}^*: Y^{X_j} \to Y^{X_i}$ by

$$\phi_{i,j}^*(f) = f \circ \phi_{i,j}.$$ 

Then $(Y^{X_i}, \phi_{i,j}^*)$ is an inverse system of topological spaces over $I$. If $(X, \phi)$ is a set-theoretic direct limit for $(X_i, \phi_{i,j})$, show that $(Y^X, \phi^*)$ is a set-theoretic inverse limit for $(Y^{X_i}, \phi_{i,j}^*)$, where $\phi^*(f) = f \circ \phi$. If all the spaces of functions are given the topology of pointwise convergence, show that the topology on $Y^X$ is the inverse limit topology.

(c) If $(X_i, \phi_{j,i})$ is an inverse system of sets over $I$, and $Y$ is a set, for $i < j$ define $\phi_{i,j}^*: Y^{X_i} \to Y^{X_j}$ by

$$\phi_{i,j}^*(f) = f \circ \phi_{j,i}.$$ 

Then $(Y^{X_i}, \phi_{i,j}^*)$ is a direct system of sets over $I$. If $(X, \phi)$ is an inverse limit for $(X_i, \phi_{j,i})$, is it true that $(Y^X, \phi^*)$ is a direct limit for $(Y^{X_i}, \phi_{i,j}^*)$, where $\phi^*(f) = f \circ \phi$? What is the relation between the two? What if the sets are topological spaces with modifications as in (b)?
XI.15. Uniform Structures

The theory of uniform structures fits somewhere between the theories of metric spaces and topological spaces. They are more general than metric spaces, but similarly give notions of uniform convergence and uniform continuity which are absent in general topological spaces. There is an underlying topology defined by a uniform structure, but the uniform structure is finer than the underlying topology since in general many different uniform structures can give the same topology, just as with metrics. Not every topology comes from a uniform structure; the ones that do turn out to be precisely the completely regular topologies.

There are two seemingly different standard approaches to the theory of uniform structures which turn out to be equivalent. One is based on metric spaces, the other on topological groups. The second approach seems more general, and it is remarkable that the two are actually the same.

The first approach uses a suitable family of pseudometrics. In the second approach, a uniform structure on a set \( X \) is a set of “neighborhoods” of the diagonal \( \Delta = \{(x, x) : x \in X\} \) in \( X \times X \), satisfying certain natural properties. Such a neighborhood \( W \) gives a uniform notion of closeness in \( X \): \( x \) and \( y \) are “\( W \)-close” if \( (x, y) \in W \). The two prototype examples of such systems are:

(i) If \( (X, \rho) \) is a metric space, then for every \( \epsilon > 0 \) there is a corresponding “neighborhood”

\[
W_{\rho, \epsilon} = \{(x, y) : \rho(x, y) < \epsilon\}
\]

of \( \Delta \) in \( X \times X \).

(ii) If \( X \) is a topological group, and \( U \) is a neighborhood of the identity element in \( X \), then there is a corresponding “neighborhood”

\[
W_U = \{(x, y) : xy^{-1} \in U\}
\]

of \( \Delta \) in \( X \times X \).

If \( X = \mathbb{R} \), these two types of neighborhoods are essentially the same if \( \mathbb{R} \) is regarded as a metric space with the standard metric, or as a topological group under addition; a typical “neighborhood” of \( \Delta \) is

\[
\{(x, y) \in \mathbb{R}^2 : |x - y| < \epsilon\}
\]

for some fixed \( \epsilon > 0 \), which is geometrically an open strip of uniform width around the line \( y = x \).

XI.15.1. Topologies and Uniform Structures Defined by Pseudometrics

A straightforward but powerful generalization of a metric space is a space with a family of pseudometrics. Recall () that a pseudometric on a set \( X \) is a function \( \rho : X \times X \to [0, \infty) \) with all the properties of a metric except that \( \rho(x, y) = 0 \) does not necessarily imply that \( x = y \).

XI.15.1.1. A family of pseudometrics on a set \( X \) is simply a set \( \mathcal{P} = \{\rho_i : i \in I\} \) of pseudometrics on \( X \). The family \( \mathcal{P} \) is called separating if \( \rho_i(x, y) = 0 \) for all \( i \) implies that \( x = y \), i.e. for any distinct \( x, y \in X \) there is an \( i \) such that \( \rho_i(x, y) > 0 \).

Note that a single metric on a set can be regarded as a separating family of pseudometrics with one element.
XI.15.1.2. To each pseudometric \( \rho \) on a set \( X \) we can associate open balls \( B^\rho_\epsilon(x) = \{ y \in X : \rho(x, y) < \epsilon \} \), just as in the metric case. If we have a family \( \mathcal{P} \) of pseudometrics, we may take the set of all open balls with respect to the pseudometrics in \( \mathcal{P} \) as a subbase for a topology, called the **topology induced by \( \mathcal{P} \)**, denoted \( T_\mathcal{P} \). These open balls do not form a base for \( T_\mathcal{P} \) in general if \( \mathcal{P} \) contains more than one pseudometric, but finite intersections do, i.e. sets of the form

\[
B^F_\epsilon(x) = \{ y \in X : \rho_i(x, y) < \epsilon \text{ for all } i \in F \}
\]

where \( F \) is a finite subset of \( I \) and \( \epsilon > 0 \), form a base for \( T_\mathcal{P} \), i.e. a subset \( U \) of \( X \) is open if and only if, for every \( x \in U \), there is a finite subset \( F \) of \( I \) and \( \epsilon > 0 \) such that \( B^F_\epsilon(x) \subseteq U \). (We will also use the analogous notation \( B^F_\epsilon(x) \), for a finite subset \( F \) of \( \mathcal{P} \).)

Exactly as in the metric case (), using the triangle inequality every open ball has this property, hence is an open set.

XI.15.1.3. Convergence in the topology \( T_\mathcal{P} \) is easily described. If \( (x_i) \) is a net in \( X \) and \( x_0 \in X \), then \( x_i \to x \) in \( (X, T_\mathcal{P}) \) if and only if \( \rho(x_i, x) \to 0 \) for every \( \rho \in \mathcal{P} \), i.e. if and only if for every \( \rho \in \mathcal{P} \) and \( \epsilon > 0 \) there is an \( i_0 \) such that \( \rho(x_i, x) < \epsilon \) for all \( i \geq i_0 \).

The next observation generalizes the fact that a metrizable topology is Hausdorff:

XI.15.1.4. **Proposition.** Let \( \mathcal{P} \) be a family of pseudometrics on a set \( X \). Then the following are equivalent:

(i) \( T_\mathcal{P} \) is \( T_0 \).

(ii) \( T_\mathcal{P} \) is Hausdorff.

(iii) \( \mathcal{P} \) is **separating**.

**Proof:** If \( \mathcal{P} \) is separating, for any \( x, y \in X \), \( x \neq y \), let \( \rho \in \mathcal{P} \) with \( \rho(x, y) > 0 \), and set \( \epsilon = \frac{\rho(x, y)}{2} \). Then \( B^\rho_\epsilon(x) \) and \( B^\rho_\epsilon(y) \) are disjoint open neighborhoods of \( x \) and \( y \), so \( T_\mathcal{P} \) is Hausdorff. If \( \mathcal{P} \) is not separating, let \( x, y \in X \), \( x \neq y \), with \( \rho(x, y) = 0 \) for all \( \rho \in \mathcal{P} \). Then any open set containing \( x \) also contains \( y \) and vice versa, so \( T_\mathcal{P} \) is not \( T_0 \).

Uniform Domination and Saturated Families

We need some tools to compare families of pseudometrics, generalizations of the notions of uniform domination and uniform equivalence of metrics ()

XI.15.1.5. **Definition.** Let \( \rho_1, \ldots, \rho_n, \sigma \) be pseudometrics on a set \( X \). Then \( \sigma \) is **uniformly dominated by \( \{\rho_1, \ldots, \rho_n\} \)** if for every \( \epsilon > 0 \) there is a \( \delta > 0 \) such that, whenever \( x, y \in X \) and \( \rho_k(x, y) < \delta \) for \( 1 \leq k \leq n \), then \( \sigma(x, y) < \epsilon \).

If \( \mathcal{P} \) is a family of pseudometrics on \( X \), then \( \sigma \) is **uniformly dominated by \( \mathcal{P} \)** if for every \( \epsilon > 0 \) there is a \( \delta > 0 \) and a finite subset \( \{\rho_1, \ldots, \rho_n\} \) of \( \mathcal{P} \) such that, whenever \( x, y \in X \) and \( \rho_k(x, y) < \delta \) for \( 1 \leq k \leq n \), then \( \sigma(x, y) < \epsilon \).
XI.15.1.6. Proposition. Let \( \mathcal{P} \) be a family of pseudometrics on a set \( X \). Then there is a smallest saturated family \( \mathcal{P} \) of pseudometrics on \( X \) containing \( \mathcal{P} \). \( \mathcal{P} \) consists precisely of all pseudometrics on \( X \) which are uniformly dominated by \( \mathcal{P} \).

If \( \mathcal{Q} \) is another family of pseudometrics on \( X \), then \( \mathcal{Q} \) is uniformly dominated by \( \mathcal{P} \) if and only if \( \mathcal{Q} \subseteq \mathcal{P} \), and \( \mathcal{P} \) and \( \mathcal{Q} \) are uniformly equivalent if and only if \( \mathcal{P} = \mathcal{Q} \). In particular, \( \mathcal{P} \) and \( \mathcal{P} \) are uniformly equivalent.

The proof is left as an exercise (Exercise ())).

XI.15.1.7. Proposition. Let \( \mathcal{P} \) and \( \mathcal{Q} \) be families of pseudometrics on a set \( X \). If \( \mathcal{Q} \) is uniformly dominated by \( \mathcal{P} \), then \( \mathcal{T}_\mathcal{P} \) is stronger than \( \mathcal{T}_\mathcal{Q} \). If \( \mathcal{P} \) and \( \mathcal{Q} \) are uniformly equivalent, then \( \mathcal{T}_\mathcal{P} = \mathcal{T}_\mathcal{Q} \). In particular, \( \mathcal{T}_\mathcal{P} = \mathcal{T}_\mathcal{P} \).

Proof: Let \( \sigma \in \mathcal{Q}, x \in X, \) and \( \epsilon > 0 \). Choose \( \delta > 0 \) and \( F = \{\rho_1, \ldots, \rho_n\} \) as in the definition. Then \( B_\delta(x) \subseteq B_\epsilon(x) \).

XI.15.1.8. The converse of XI.15.1.7. is false in general: families of pseudometrics generating the same topology need not be uniformly equivalent, even if they are singletons (metrics can be equivalent but not uniformly equivalent ())). Two families of pseudometrics generating the same topology are called equivalent.

Equivalently, but not uniformly equivalent families generate the same topology, but different uniform structures; thus uniform structures are more delicate than topologies.

XI.15.1.9. In light of XI.15.1.7., there is no loss of generality in considering only saturated families of pseudometrics. There is some technical advantage in working with saturated families: for example, the open balls form a base for the underlying topology instead of just a subbase. In fact, the open balls from a saturated family frequently give all, or almost all, of the open sets in the topology (they precisely give all the cozero sets, hence all open \( F_\sigma \)'s if the space is normal). And the \( \epsilon \) can be dispensed with in considering open balls from a saturated family \( \mathcal{P} \): since a saturated family of pseudometrics is closed under nonnegative scalar multiples, any ball of the form \( B_\sigma^\rho(x) \) for \( \rho \in \mathcal{P} \) is of the form \( B_1^\sigma(x) \) for some \( \sigma \in \mathcal{P} \) (specifically, \( \sigma = \epsilon \rho \)), so only balls of radius 1 need be considered.

Uniform Continuity and Uniform Convergence
Bounded Pseudometrics

XI.15.1.10. If \( \rho \) is a pseudometric on a set \( X \), then \( \hat{\rho} \), defined by

\[
\hat{\rho}(x, y) = \min\{\rho(x, y), 1\}
\]
is a bounded pseudometric on $X$ which is uniformly equivalent to $\rho$. So if $\mathcal{P}$ is a family of pseudometrics on $X$, define

$$\tilde{\mathcal{P}} = \{ \tilde{\rho} : \rho \in \mathcal{P} \}.$$ 

Then $\tilde{\mathcal{P}}$ is a family of bounded pseudometrics which is uniformly equivalent to $\mathcal{P}$. (Note, however, that $\tilde{\mathcal{P}}$ is never saturated unless $\mathcal{P} = \{0\}$, so if $\mathcal{P}$ is saturated and nonzero, then $\tilde{\mathcal{P}} \neq \mathcal{P}$; in fact, $\tilde{\mathcal{P}}$ is a proper subset of $\mathcal{P}$ in this case.) Thus we may always work with families of pseudometrics taking values in $[0, 1]$.

**Countably Generated Families of Pseudometrics**

**XI.15.1.11.** If $\mathcal{P} = \{ \rho_1, \rho_2, \ldots \}$ is a countable family of pseudometrics on a set $X$, define a pseudometric $\sigma$ by

$$\sigma(x, y) = \sum_{k=1}^{\infty} 2^{-k} \tilde{\rho}_k(x, y)$$

(it is routine and straightforward to verify that $\sigma$ is a pseudometric).

**XI.15.1.12.** Proposition. The pseudometric $\sigma$ is finitely uniformly dominated by $\mathcal{P}$.

**Proof:** Let $\epsilon > 0$. Choose $n$ so that $2^{-n+1} < \epsilon$. Set $\delta = \frac{\epsilon}{2n}$, and $\mathcal{F} = \{ \rho_1, \ldots, \rho_n \}$. Then, if $x, y \in X$ satisfy $\rho_k(x, y) < \delta$ for $1 \leq k \leq n$, we have

$$\sigma(x, y) = \sum_{k=1}^{n} 2^{-k} \tilde{\rho}_k(x, y) + \sum_{k=n+1}^{\infty} 2^{-k} \tilde{\rho}_k(x, y) < n\delta + 2^{-n} < \epsilon.$$ 

\(\Box\)

**XI.15.1.13.** Conversely, it is obvious that each $\rho_k$ is finitely uniformly dominated by $\mathcal{Q} = \{ \sigma \}$. So $\mathcal{Q}$ is a singleton family which is uniformly equivalent to $\mathcal{P}$, so in particular $T_{\mathcal{Q}} = T_{\tilde{\mathcal{P}}}$. Thus a uniform structure generated by a countable family of pseudometrics is equivalent to the uniform structure generated by a singleton.

The pseudometric $\sigma$ is a metric if and only if $\mathcal{P}$ is separating. Thus the uniform structure generated by a countable separating family of pseudometrics is also generated by a single metric. In particular, a topology defined by a countable separating family of pseudometrics is metrizable.

**XI.15.1.14.** Slightly more generally, a family $\mathcal{P}$ of pseudometrics is *countably generated* if there is a countable family $\mathcal{Q}$ of pseudometrics uniformly equivalent to $\mathcal{P}$. A uniform structure generated by a countably generated family of pseudometrics is also generated by one pseudometric (one metric if the family is separating).

**XI.15.1.15.** So the theory of uniform structures generated by families of pseudometrics is technically only a true generalization of the theory of metric spaces in that it allows families of pseudometrics which are not countably generated (as well as families which are not separating). However, as a practical matter the theory has merit even in the countably generated case: a metrizable space often has no natural metric, and a choice of a metric is arbitrary; but it may, and often does, have a natural family of pseudometrics generating the topology (cf. ())).
Pseudometrics and Real-Valued Continuous Functions

There is an intimate connection between pseudometrics and real-valued continuous functions on a set.

**XI.15.1.16.** Let $X$ be a set. If $f$ is a function from $X$ to $\mathbb{R}$, then there is a naturally associated pseudometric $\rho_f$ defined by

$$\rho_f(x, y) = |f(x) - f(y)|.$$  

Conversely, if $\rho$ is a pseudometric on $X$, and $x \in X$, there is an associated nonnegative real-valued function $f_{\rho, x}$ defined by

$$f_{\rho, x}(y) = \rho(x, y).$$

**XI.15.1.17.** If $f : X \to \mathbb{R}$ is a function, then $f$ is (essentially by definition) uniformly continuous with respect to any family of pseudometrics on $X$ containing $\rho_f$ (and the usual metric on $\mathbb{R}$). Conversely, if $\rho$ is a pseudometric on $X$, then it follows easily from the triangle inequality that, for any $x \in X$, $f_{\rho, x}$ is uniformly continuous with respect to any family of pseudometrics on $X$ containing $\rho$.

**XI.15.1.18.** Proposition. Let $\mathcal{P}$ be a family of pseudometrics on a set $X$. Then $\mathcal{T}_{\mathcal{P}}$ is the weakest topology on $X$ making each function $f_{\rho, x}$ continuous, for all $\rho \in \mathcal{P}$ and $x \in X$.

**XI.15.1.19.** Corollary. Let $\mathcal{P}$ be a separating family of pseudometrics on a set $X$. Then the topology $\mathcal{T}_{\mathcal{P}}$ is completely regular.

**Proof:** The proof is similar to the proof of (). Replacing $\mathcal{P}$ by $\hat{\mathcal{P}}$ if necessary, we may assume every $\rho \in \mathcal{P}$ takes values in $[0, 1]$, so $f_{\rho, x}$ takes values in $[0, 1]$ for all $\rho$ and $x$. Let $\mathcal{S}$ be the set of all $f_{\rho, x}$ for $\rho \in \mathcal{P}$, $x \in X$, and set $Z = [0, 1]^\mathcal{S}$. Then $Z$ is a compact Hausdorff space.

Define $\phi : X \to Z$ as follows. For each $y \in X$, let $\phi(y)$ be the point in $Z$ whose $f_{\rho, x}$th coordinate is $f_{\rho, x}(y) = \rho(x, y)$. Then $\phi$ is continuous for $\mathcal{T}_{\mathcal{P}}$ since its composition with each coordinate map is continuous, and $\phi$ is one-to-one since $\mathcal{P}$ is separating. The restriction of the topology on $Z$ to $\phi(X)$ gives a topology on $\phi(X) \cong X$ which is weaker than $\mathcal{T}_{\mathcal{P}}$ for which each $f_{\rho, x}$ is continuous; thus this topology is exactly $\mathcal{T}_{\mathcal{P}}$, i.e. $\phi$ is a homeomorphism onto its image. Thus $(X, \mathcal{T}_{\mathcal{P}})$ is homeomorphic to a subspace of a compact Hausdorff space.

**XI.15.1.20.** Now let $(X, \mathcal{T})$ be a topological space. Set

$$\mathcal{P}_\infty = \{\rho_f : f \text{ is a continuous function from } X \text{ to } [0, 1]\}$$

(one obtains an equivalent set of seminorms if all continuous functions from $X$ to $\mathbb{R}$ are used). Then the topology $\mathcal{T}_{\mathcal{P}_\infty}$ is weaker than $\mathcal{T}$. It may be strictly weaker: if $\mathcal{P}_\infty$ is separating, i.e. if continuous functions from $X$ to $[0, 1]$ separate points of $X$, then $\mathcal{T}_{\mathcal{P}_\infty}$ is completely regular, but $\mathcal{T}$ need not be completely regular even if it is regular [?, Example 91]. $\mathcal{T}_{\mathcal{P}_\infty}$ is the strongest completely regular topology weaker than $\mathcal{T}$ (if $\mathcal{P}_\infty$ is separating); in particular, $\mathcal{T} = \mathcal{T}_{\mathcal{P}_\infty}$ if $\mathcal{T}$ is completely regular. $\mathcal{P}_\infty$ (actually $\mathcal{P}_\infty$) is the “largest” family of seminorms giving the topology $\mathcal{T}_{\mathcal{P}_\infty}$, in the sense that if $\mathcal{Q}$ is a family of seminorms with $\mathcal{T}_{\mathcal{Q}} = \mathcal{T}_{\mathcal{P}_\infty}$, then $\mathcal{Q}$ is uniformly dominated by $\mathcal{P}_\infty$.  

1189
XI.15.1.21. Even if \((X, \mathcal{T})\) is metrizable, the family \(\mathcal{P}_\infty\) is not countably generated in general, and thus is not uniformly equivalent to a singleton family. Every real-valued continuous function on \((X, \mathcal{T}_{\mathcal{P}_\infty})\) is uniformly continuous with respect to \(\mathcal{P}_\infty\).

XI.15.2. General Uniform Structures
XI.16. Order Topology

XI.16.1. Definition and Properties

XI.16.1.1. Theorem. Any subspace of a space with the order topology is completely normal. In particular, any space with the order topology is normal.

Proof: [Cat06] Let $X$ be an ordered space with the order topology, and let $Y$ be a subspace. We will show that $Y$ has the following stronger property: if $\{A_i : i \in I\}$ is a collection of subsets of $Y$ which are separated, i.e. each $A_i$ is disjoint from the closure (in $Y$) of $\cup_{j \neq i} A_j$, then the $A_i$ have disjoint neighborhoods $V_i$ in $Y$ ($Y$ is said to be completely collectionwise normal).

Note that by (), if the $A_i$ are separated in $Y$, then the $A_i$ are separated in $X$, and if $\{U_i\}$ are disjoint neighborhoods in $X$, and $V_i = U_i \cap Y$, then $\{V_i\}$ are disjoint neighborhoods in $Y$. Thus for the proof we may assume $Y = X$.

Suppose $\{A_i\}$ are separated in $X$ (and in particular disjoint), and set $A = \cup_i A_i$. For each convex component $C$ of $X \setminus A$, choose a point $f(C) \in C$. For $p \in A$, $p$ not the largest point of $X$ (if there is one), choose $t_p \in X$ as follows. There is a $u_p > p$ such that $[p, u_p)$ contains no point of $A \setminus A_i$.

Choose $t_p$ to be a point of $A_i \cap (p, u_p)$ if $A_i \cap (p, u_p)$ is nonempty.

Choose $t_p = u_p$ if $(p, u_p) = \emptyset$.

If $(p, u_p) \neq \emptyset$ but $A_i \cap (p, u_p) = \emptyset$, then $(p, u_p)$ is contained in a unique convex component $C$ of $X \setminus A$; in this case $t_p = f(C)$.

Similarly, if $p$ is not the smallest point of $X$, choose $s_p < p$ in the same manner with inequalities reversed. Set $I_p = (s_p, t_p)$. (If $p$ is the largest point of $X$, set $I_p = (s_p, p]$, and if $p$ is the smallest point of $X$, set $I_p = [p, t_p)$.) Then $I_p$ is an open interval in $X$ containing $p$. This defines $I_p$ for each $p \in A$.

For each $i$, set $U_i = \cup_{p \in A_i} I_p$. Then $U_i$ is an open neighborhood of $A_i$. We claim the $U_i$ are disjoint. If $x \in U_i \cap U_j$ for $i \neq j$, there are $p \in A_i$ and $q \in A_j$ such that $x \in I_p \cap I_q$. We may suppose $p < q$. Then $s_q < x < t_p$. Since $q \notin I_p$ and $p \notin I_q$, $p < s_q < x < t_p < q$. Thus $t_p \in I_q$, and hence $t_p \notin A_i$, and $(p, t_p) \neq \emptyset$. Similarly, $s_q \notin A_j$ and $(s_q, q) \neq \emptyset$. By the way $t_p$ and $s_q$ were selected, we have $t_p = s_q = f(C)$ for the same convex component $C$ of $X \setminus A$, contradicting that $s_q < t_p$. Thus the $U_i$ are disjoint. $\emptyset$

Note that the AC is needed in several places in this proof. In fact, the AC (or something only slightly weaker) is necessary for this result [vD85].

XI.16.2. Completion of an Order Topology
XI.16.3. Function Space Notation and the Compact-Open Topology

There are several notations for function spaces used in the literature, and the usage is not always consistent. There is one predominant notation:

XI.16.3.1. Definition. Let $X$ and $Y$ be topological spaces. Denote by $Y^X$ the set of all maps (continuous functions) from $X$ to $Y$.

XI.16.3.2. The notation $C(X,Y)$ is also frequently used for $Y^X$. Some authors use $Y^X$ to denote the set of all functions from $X$ to $Y$, which we will write $F(X,Y)$. Although it is occasionally useful to consider this set, the set $C(X,Y)$ arises far more often. Our use of $Y^X$ to denote $C(X,Y)$, if done with care, includes all standard examples of this exponential notation:

XI.16.3.3. Examples. (i) If $X$ has the discrete topology, then for any $Y$ we have $Y^X = C(X,Y) = F(X,Y)$. If $X$ and $Y$ are just sets, with no topology, we can consider them to be topological spaces with the discrete topology so the notation $Y^X = F(X,Y)$ used in set theory is consistent with the topological notation. Note that the topology on $Y$ has no effect on the definition of $Y^X$ as a set in this case.

(ii) If $Y$ is a topological space (or just a set, regarded as a topological space with the discrete topology), and $n \in \mathbb{N}$, we can regard $n$ as the subset $\{1, \ldots, n\}$ of $\mathbb{N}$ (or, more commonly and properly in set theory, $\{0,1,\ldots,n-1\} \subseteq \mathbb{N}_0$), with the discrete topology. Then $Y^n$ (in the sense of XI.16.3.1.) is naturally the set of $n$-tuples of elements of $Y$, the usual meaning of $Y^n$. Similarly, $Y^\mathbb{N}$ denotes the set of sequences in $Y$ (where $\mathbb{N}$ has its usual discrete topology). This latter set is often, but somewhat less precisely, denoted $Y^\infty$. (iii) More generally, if $\kappa$ is a cardinal, we denote by $Y^\kappa$ the set of all functions from a set of cardinality $\kappa$ to $Y$. To be precise, we regard $\kappa$ as a set of ordinals as in (i), and give it the discrete topology. This set $Y^\kappa$ can be naturally identified with the Cartesian product of $\kappa$ copies of $Y$.

(iv) If $X$ is a set, the power set $\mathcal{P}(X)$ is often denoted $2^X$. If $X$ is a topological space, $2^X$ is often used to denote the set of closed subsets of $X$. Both notations are consistent with XI.16.3.1. if we give $2 = \{0,1\}$ the Sierpiński space topology with $\{0\}$ open and $\{1\}$ closed. We then identify a continuous function $f : X \to 2$ with the subset $f^{-1}(\{1\})$ of $X$.

XI.16.3.4. Both the notations $Y^X$ and $C(X,Y)$ are slight abuses, since the notation does not reflect the topologies of $X$ and $Y$ which are needed to determine $Y^X$ as a set. (The notation $F(X,Y)$ is not ambiguous.) The abuse is the same as in the common statement “Let $X$ be a topological space.”

XI.16.3.5. If $X$ is a topological space, the set $\mathbb{R}^X = C(X,\mathbb{R})$ is usually denoted $C_\mathbb{R}(X)$. Similarly, $\mathbb{C}^X = C(X,\mathbb{C})$ is usually denoted $C_\mathbb{C}(X)$ or just $C(X)$. Here $\mathbb{R}$ and $\mathbb{C}$ are always assumed to have their usual topologies.

The Compact-Open Topology

There are many important notions of convergence for functions, i.e. many topologies on spaces of functions. The most generally useful topology on $Y^X$ is the “topology of uniform convergence on compact sets” (“u.c. convergence”). If $Y$ is metrizable, or more generally if the topology on $Y$ comes from a uniform structure,
this topology can be defined in a straightforward way; the idea is that a net \((f_i)\) converges to \(f\) if and only if, for every compact subset \(K\) of \(X\), for sufficiently large \(i\) \(f_i(x)\) should be close to \(f(x)\) for all \(x \in K\). Even for a general topological space \(Y\), there is a clean way to define such a topology on \(Y^X\), called the compact-open topology.

**XI.16.3.6.** Definition. Let \(X\) and \(Y\) be topological spaces. The **compact-open topology** on \(Y^X\) is the topology generated by the subbase

\[
\{U_{K,O} : K \subseteq X \text{ compact, } O \subseteq Y \text{ open}\}
\]

\[U_{K,O} = \{f \in Y^X : f(K) \subseteq O\}.\]

**XI.16.3.7.** The set of \(U_{K,O}\) does not give a base for a topology in general; one needs to take finite intersections. So a basic open set for the compact-open topology is defined by taking a finite number of compact sets \(K_1, \ldots, K_n\) in \(X\), a corresponding finite number of open sets \(O_1, \ldots, O_n\) in \(Y\), and taking all maps from \(X\) to \(Y\) such that \(f(K_j) \subseteq O_j\) for \(1 \leq j \leq n\).

To obtain a base for the compact-open topology, it suffices to let the \(O\)'s range over a base for the topology of \(Y\).

The compact-open topology can be described in terms of convergence:

**XI.16.3.8.** Proposition. Let \(X\) and \(Y\) be topological spaces. Then a net \((f_i)\) in \(Y^X\) converges to \(f \in Y^X\) in the compact-open topology if and only if, for every compact subset \(K\) of \(X\), and every open set \(O\) in \(Y\) such that \(f(K) \subseteq O\), there is an \(i_0\) such that \(f_i(K) \subseteq O\) for all \(i \geq i_0\).

The proof is immediate from the definition, and is left to the reader.

Since singleton sets are compact, we obtain:

**XI.16.3.9.** Proposition. Let \(X\) and \(Y\) be topological spaces. If a net \((f_i)\) in \(Y^X\) converges to \(f \in Y^X\) in the compact-open topology, then \(f_i \to f\) pointwise on \(X\), i.e. \(f_i(x) \to f(x)\) for all \(x \in X\).

Proof: Let \(x \in X\). If \(O\) is any neighborhood of \(f(x)\) in \(Y\), then \(f(\{x\}) \subseteq O\), so there is an \(i_0\) such that \(f_i(\{x\}) \subseteq O\) for all \(i \geq i_0\), i.e. \(f_i(x) \in O\) for \(i \geq i_0\). \(\Box\)

**XI.16.3.10.** The compact-open topology is thus stronger than the topology of pointwise convergence; it is usually strictly stronger. But if every compact subset of \(X\) is finite, e.g. if \(X\) is discrete, the two topologies coincide. (There are nondiscrete spaces where every compact subset is finite, e.g. XI.11.12.8.; but such a space cannot have any nontrivial convergent sequences. See ()).

1193
XI.16.3.11.  

**Corollary.** If $Y$ is a topological space and $X$ a set with the discrete topology, then $Y^X$ can be identified with the Cartesian product of $\kappa$ copies of $Y$, where $\kappa = \text{card}(X)$, and the compact-open topology is the product topology.

XI.16.3.12.  

If $X$ and $Y$ are topological spaces, any constant function from $X$ to $Y$ is continuous. The subset $C$ of $Y^X$ of constant functions can be identified with $Y$. It is immediate from the definitions that this identification is a homeomorphism if $C$ is given the relative compact-open topology. If $Y$ is Hausdorff, then $C$ is closed in $Y^X$ in the compact-open topology. (The converse does not hold in general, e.g. if $X$ is a one-point space.)

We now examine separation properties.

XI.16.3.13.  

**Proposition.** Let $X$ and $Y$ be topological spaces. Give $Y^X$ the compact-open topology.

(i) $Y^X$ is Hausdorff if and only if $Y$ is Hausdorff.

(ii) $Y^X$ is regular if and only if $Y$ is regular.

**Proof:** $Y$ is homeomorphic to a subset of $Y^X$ (XI.16.3.12.), so if $Y^X$ is Hausdorff or regular, so is $Y$. Conversely, suppose $Y$ is Hausdorff, and $f$ and $g$ are distinct elements of $Y^X$. Then there is an $x \in X$ with $f(x) \neq g(x)$. Let $O_1$ and $O_2$ be disjoint open neighborhoods of $f(x)$ and $g(x)$ in $Y$. Then $U_{\{x\}, O_1}$ and $U_{\{x\}, O_2}$ are disjoint neighborhoods of $f$ and $g$ in $Y^X$.

Now suppose $Y$ is regular. Let $f \in Y^X$ and $V$ a compact-open neighborhood of $f$. Then there are compact sets $K_1, \ldots, K_n$ in $X$ and open sets $O_1, \ldots, O_n$ in $Y$ with $f(K_j) \subseteq O_j$ for $1 \leq j \leq n$ and

$$U_{K_1, O_1} \cap \cdots \cap U_{K_n, O_n} \subseteq V.$$  

Fix $j$. Since $f(K_j)$ is compact, by (i) there is an open set $O'_j$ in $Y$ with

$$f(K_j) \subseteq O'_j \subseteq \overline{O}'_j \subseteq O_j.$$  

Then $W = U_{K_1, O'_1} \cap \cdots \cap U_{K_n, O'_n}$ is a neighborhood of $f$ in $Y^X$. We claim $W \subseteq V$. Suppose $(g_i)$ is a net in $W$ and $g_i \to g \in Y^X$. If $x \in K_j$, then $g_i(x) \in O'_j$ for all $i$, so $g(x) \in O'_j \subseteq O_j$. Thus $g \in V$.  

XI.16.3.14.  

If $Y$ is normal, then $Y^X$ is not necessarily normal: if $S$ is the Sorgenfrey line (XI.7.8.10.), and $2 = \{0,1\}$ has the discrete topology, then $S$ is normal, but $S^2$ is not normal.

XI.16.3.15.  

**Example.** Let $X$ be a topological space. Consider $2^X$, where 2 has the Sierpiński topology. Then $2^X$ can be identified with the set of closed subsets of $X$ as in XI.16.3.3.(iv). What is the corresponding topology on the collection of closed subspaces? Since 2 is not Hausdorff, limits in $2^X$ will not be unique. If $(E_i)$ is a net of closed subspaces of $X$ and $E$ a closed subspace of $X$, then $E_i \to E$ in this topology if and only if, whenever $K$ is a compact subspace of $X$ disjoint from $E$, then there is an $i_0$ such that $K$ is disjoint
from $E_i$ for all $i \geq i_0$. In particular, if $E_i \to E$, and $x \notin E$, then there is an $i_0$ such that $x \notin E_i$ for any $i \geq i_0$. Thus we have

$$
\bigcap_i \left[ \bigcup_{j \geq i} E_j \right] \subseteq E.
$$

This condition is in general not sufficient to guarantee that $E_i \to E$, however. It is sufficient that

$$
E_0 := \bigcap_i \left[ \bigcup_{j \geq i} E_j \right] \subseteq E
$$

for if $E$ is a closed set containing $E_0$, and $K$ is a compact set disjoint from $E$, then $K$ is disjoint from $E_0$, so if $W_i$ is the complement of $\bigcup_{j \geq i} E_j$ the $W_i$ form an increasing open cover of $K$, which has a finite subcover, i.e. $K \subseteq W_{i_0}$ for some $i_0$. Then $K$ is disjoint from $E_i$ for all $i \geq i_0$. Thus $E_i \to E$.

If every point of $X$ has local base of compact neighborhoods, e.g. if $X$ is a locally compact Hausdorff space, $E_i \to E$, and $x \notin E$, then since $E$ is closed $x$ has a compact neighborhood $K$ disjoint from $E$. Then $K$ is disjoint from $E_i$ for all sufficiently large $i$, so $x \notin E_0$, i.e. $E_0 \subseteq E$. So in this case $E_i \to E$ if and only if $E$ is a closed set containing $E_0$.

We now examine the case where $Y$ is metrizable. In this case, the compact-open topology is precisely the topology of u.c. convergence:

**XI.16.3.16. Proposition.** Let $X$ be a topological space and $(Y, \rho)$ a metric space. If $(f_i)$ is a net in $Y^X$, then $f_i \to f \in Y^X$ if and only if, for every compact subset $K$ of $X$ and every $\epsilon > 0$, there is an $i_0$ such that $\rho(f_i(x), f(x)) < \epsilon$ for all $i \geq i_0$ and all $x \in K$.

**Proof:** Suppose the $\epsilon$-condition is satisfied. Let $K$ be a compact subset of $X$ and $O$ an open set in $Y$. Then, since $f(K)$ is compact, there is an $\epsilon > 0$ such that $\rho(f(x), Y \setminus O) \geq \epsilon$ for all $x \in K$. Fix $i_0$ such that $\rho(f_i(x), f(x)) < \epsilon$ for all $i \geq i_0$ and $x \in K$. Then $f_i(K) \subseteq O$ for all $i \geq i_0$. So $f_i \to f$ in the compact-open topology.

Conversely, if $f_i \to f$ in the compact-open topology, let $K$ be a compact subset of $X$ and $\epsilon > 0$. For each $x \in X$ let $V_x$ be the open ball of radius $\frac{\epsilon}{3}$ centered at $f(x)$. Then finitely many of these balls, say $\{V_{x_1}, \ldots, V_{x_n}\}$, cover $f(K)$ since $f(K)$ is compact. For each $j$, $1 \leq j \leq n$, let $K_j = f^{-1}(V_{x_j})$ and $O_j$ the open ball of radius $\frac{\epsilon}{3}$ around $f(x_j)$. For each $j$, $K_j$ is compact and $f(K_j) \subseteq O_j$. There is an $i_0$ such that $f_i(K_j) \subseteq O_j$ for all $j$ and for all $i \geq i_0$. Thus, if $x \in K$, then $x \in K_j$ for some $j$, so for any $i \geq i_0$ we have

$$
\rho(f_i(x), f(x)) \leq \rho(f_i(x), f_i(x_j)) + \rho(f_i(x), f(x_j)) + \rho(f(x), f(x)) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.
$$

**XI.16.3.17.** If $X$ and $Y$ are topological spaces, the compact-open topology can be defined in the same way on $F(X, Y)$ as in XI.16.3.6. The relative topology on the subset $Y^X = C(X, Y)$ is exactly the compact-open topology defined in XI.16.3.6. However, the compact-open topology on $F(X, Y)$ does not reasonably coincide with the topology of uniform convergence on compact sets, and $Y^X$ is not closed in $F(X, Y)$ in general. See Exercise XI.16.4.1.
XI.16.4. Exercises

XI.16.4.1. (a) Let $X = \mathbb{N}^1$, the one-point compactification of $\mathbb{N}$, and let $Y = [0,1]$. Let $f_n : X \to Y$ be the function which is 1 on $\{1, \ldots, n\}$ and 0 otherwise, and let $f$ be the function which is 1 on $\mathbb{N}$ and 0 at $\infty$. Then $f_n \to f$ pointwise, each $f_n$ is continuous, but $f$ is discontinuous. Obviously $(f_n)$ does not converge to $f$ uniformly, hence does not converge uniformly on compact sets since $X$ is compact. Show that $f_n \to f$ in the compact-open topology on $F(X,Y)$.

(b) Let $X = \mathbb{N} \cup \{\alpha\}$ be the space of XI.11.12.8., and let $Y = [0,1]$. The only compact subsets of $X$ are the finite subsets; thus uniform convergence on compact sets is the same as pointwise convergence. If $f_n$ is defined as above, and $f$ is 1 on $\mathbb{N}$ and $f(\alpha) = 0$, then $f_n \to f$ uniformly on compact sets (pointwise), but each $f_n$ is continuous and $f$ is discontinuous. In this case, $f_n \to f$ also in the compact-open topology on $F(X,Y)$.

(c) Let $X = \mathbb{N}^1$ and $Y = \mathbb{R}$. Define $f : X \to Y$ by $f(k) = k$ for $k \in \mathbb{N}$, and $f(\infty) = 0$. Let $f_n(k) = k - \frac{1}{n}$, and $f_n(\infty) = 0$. Then $f_n \to f$ uniformly. Let

$$O = (-1,1) \cup \bigcup_{k \in \mathbb{N}} \left( k - \frac{1}{k}, k + \frac{1}{k} \right).$$

Then $O$ is open in $Y$. We have $f \in U_{X,O}$, but $f_n \notin U_{X,O}$ for any $n$. Thus $(f_n)$ does not converge to $f$ in the compact-open topology.

With some care, this example can be modified to make $Y = [0,1]$.

(d) Show that the proof of Proposition XI.16.3.16. shows the following:

**Proposition.** Let $(f_i)$ be a net of functions from a topological space $X$ to a metric space $(Y, \rho)$, and let $f$ be a function from $X$ to $Y$. If $f$ is continuous on each compact subset of $X$, then $f_i \to f$ in the compact-open topology if and only if $f_i \to f$ uniformly on each compact subset $K$ of $X$.

Note that there is no continuity assumption (or any other assumption) on the $f_i$. Even the limit function need not be continuous on all of $X$. (If $X$ is first countable, then continuity on compact subsets implies continuity on all of $X$ ; however, this can fail in general, as in (b).)

(e) If each $f_i$ is continuous (or just continuous on compact sets), and $f_i \to f$ uniformly on compact sets, then $f$ is continuous on compact sets (), and thus $f_i \to f$ in the compact-open topology.

Note that problem 8b in section 9.1 of [?] is wrong as stated.
XI.17. Cantor Sets

The Cantor set, and its variations, is a complicated but interesting closed subset of the unit interval in $\mathbb{R}$. It is well worth while to develop a good understanding of this set, since it is an important gateway to a number of varied aspects of analysis and topology.

XI.17.1. The Cantor Set

XI.17.2. Other Cantor Sets

The construction of the Cantor set can be obviously generalized: at each step we only need to remove an open interval from the interior of each of the remaining closed intervals at that step; it does not have to be exactly the middle third (the lengths of the remaining intervals must just approach zero). By varying the lengths of the removed intervals, we can obtain versions of the Cantor set with various measure-theoretic properties.

Symmetric Cantor Sets

We first consider the case where all removed intervals at each stage are centered and of the same length, so that all the remaining intervals at each step will have the same length. This restriction is not necessary, but without it notation becomes unwieldy, and the Cantor sets obtained with the restriction, while quite general, retain a pleasant amount of symmetry and homogeneity.

XI.17.2.1. Let $\alpha = (\alpha_0, \alpha_1, \alpha_2, \ldots)$ be a sequence of real numbers such that $\alpha_0 = 1$ and $0 < \alpha_n < \frac{\alpha_{n-1}}{2}$ for $n \geq 1$. Begin with $K_0 = [0, 1]$. Remove an open interval from the middle of $K_0$ to obtain the set $K_1$ consisting of two closed intervals of length $\alpha_1$; the removed interval has length $1 - 2\alpha_1 = \alpha_0 - 2\alpha_1$. Continue inductively: at the $n$'th stage remove open intervals from the middle of each of the closed intervals in $K_{n-1}$ to obtain $2^n$ closed intervals of length $\alpha_n$; the $2^n-1$ removed intervals each have length $\alpha_{n-1} - 2\alpha_n$. Let $K_n$ be the union of the resulting intervals. We have that $K_n$ is a closed subset of $[0, 1]$, hence compact $(\cdot)$, and $K_n+1 \subseteq K_n$ for all $n$.

XI.17.2.2. Definition. The set $K_\alpha = \cap_n K_n$ is the Cantor set of type $\alpha$.

XI.17.2.3. We thus obtain a large variety of Cantor sets. The Cantor set of $(\cdot)$ is the Cantor set of type $\alpha$, where $\alpha_n = 3^{-n}$ for all $n$. This set will simply be denoted $K$ and called “the Cantor set,” i.e. the phrase “the Cantor set” without additional qualification always means the standard one.

All Cantor sets constructed in this way are topologically the same as the Cantor set, and embedded in $[0, 1]$ in the same way:

XI.17.2.4. Proposition. Let $K_\alpha$ be any Cantor set. Then there is a strictly increasing continuous function $\phi : [0, 1] \to [0, 1]$ with $\phi(K) = K_\alpha$. In particular, $K_\alpha$ is homeomorphic to $K$.

We always have $\alpha_n < 2^{-n}$, so $\alpha_n \to 0$ rather rapidly. The faster the $\alpha_n$ converge to 0, the “smaller” the set $K_\alpha$ is. By choosing this rate carefully, one can obtain arbitrary Hausdorff and packing dimensions for $K_\alpha$:
XI.17.2.5. **Theorem.** Let $K_\alpha$ be a Cantor set of type $\alpha = (\alpha_0, \alpha_1, \alpha_2, \ldots)$. Then

(i) The Hausdorff dimension ($\dim$) of $K_\alpha$ is

$$
\dim_H(K_\alpha) = \liminf_{n \to \infty} \frac{n \log 2}{-\log \alpha_n}.
$$

(ii) The packing dimension ($\dim$) of $K_\alpha$ is

$$
\dim_P(K_\alpha) = \limsup_{n \to \infty} \frac{n \log 2}{-\log \alpha_n}.
$$

In particular, if $0 \leq s_H \leq s_P \leq 1$, there is a $K_\alpha$ with $\dim_H(K_\alpha) = s_H$ and $\dim_P(K_\alpha) = s_P$.

XI.17.2.6. **Examples.**

(i) If $0 < s < 1$, set $\alpha_n = 2^{-n/s}$ for all $n$. Then $\dim_H(K_\alpha) = \dim_P(K_\alpha) = s$.

(ii) If $\alpha_n = 2^{-n^2}$ for all $n$, then $\dim_H(K_\alpha) = \dim_P(K_\alpha) = 0$.

XI.17.2.7. If $\dim_H(K_\alpha) = s$, this result does not give information about $\mathcal{H}^s(K_\alpha)$. By a more delicate analysis, we can make this measure arbitrary as long as $0 < s < 1$ (Since $\mathcal{H}^0$ is counting measure, we have $\mathcal{H}^0(K_\alpha) = \infty$ for all $\alpha$, and $\mathcal{H}^1$ is Lebesgue measure on $\mathbb{R}$, so $\mathcal{H}^1(K_\alpha) \leq \mathcal{H}^1([0,1]) = 1$ for all $\alpha$). We can do the same for packing measure. We have:

If $\dim_H(K_\alpha) = 1$, we can make $K_\alpha$ have positive Lebesgue measure:

XI.17.2.8. **Example.** Let $0 < d < 1$. We will construct a $K_\alpha$ such that $\lambda(K_\alpha) = d$. Set

$$
c = \frac{3 - 2d}{1 - d} \in (3, \infty)
$$

and note that $d = \frac{c - 3}{c - 2}$. Choose the $\alpha_n$ so that the removed open intervals at the $n$'th stage have length $c^{-n}$. This can be done inductively by setting $\alpha_0 = 1$ and solving $c^{-n} = \alpha_n - 2\alpha_{n-1}$ for $\alpha_n$. If $b = c^{-1}$, we have successively

$$
\alpha_1 = \frac{1 - b}{2}, \quad \alpha_2 = 1 - b + b^2, \quad \alpha_3 = 2 - 2b + 2b^2 + b^3, \cdots
$$

At the $n$'th step, $2^{n-1}$ intervals of length $c^{-n}$ are removed; hence the complement of $K_\alpha$ in $[0,1]$ has Lebesgue measure

$$
\sum_{n=1}^{\infty} \frac{2^{n-1}}{c^n} = \frac{1}{2} \sum_{n=1}^{\infty} \left[ \frac{2}{c} \right]^n = \frac{1}{2} \cdot \frac{\frac{2}{c}}{1 - \frac{2}{c}} = \frac{1}{c - 2}
$$

since the removed intervals are disjoint, and thus we have

$$
\lambda(K_\alpha) = 1 - \frac{1}{c - 2} = \frac{c - 3}{c - 2} = d.
$$

Denote this Cantor set by $K_d$. We could also take $d = 0$ in this construction ($c = 3$) to get the ordinary Cantor set, although this set does not have Hausdorff dimension 1.

Note that there is no $K_\alpha$ with $\lambda(K_\alpha) = 1$, since for any $\alpha$ we have that $[0,1] \setminus K_\alpha$ is a nonempty open set and thus has strictly positive Lebesgue measure.

More general Hausdorff and packing dimensions

1198
Asymmetric Cantor Sets
For Cantor sets with Lebesgue measure zero, there is a feasible alternate notation which works also in the asymmetric case.

XI.17.2.9. Let \( \lambda = (\lambda_k) \) be a sequence of strictly positive numbers with \( \sum_{k=1}^{\infty} \lambda_k = 1 \). At the first stage, remove a suitable open interval of length \( \lambda_1 \) from the interior of \( I_1 = [0, 1] \). It will turn out that there is a uniquely located such subinterval for which the subsequent process works, not necessarily centered in \([0,1]\). This leaves two closed subintervals \( I_2 \) and \( I_3 \) (numbered from left to right). Remove a suitable open subinterval of length \( \lambda_2 \) from \( I_2 \), and one of length \( \lambda_3 \) from \( I_3 \), leaving four closed intervals \( I_4 \) to \( I_7 \). Continue the process, and let \( C_\lambda \) be the remaining set at the end. Then \( C_\lambda \) is a Cantor set of Lebesgue measure zero (Exercise ()).

XI.17.2.10. To make the process work, the lengths of the remaining intervals at each stage must be sums of suitable subseries of the original series. The length of \( I_2 \) must be \( \sum \lambda_k \), where the \( \lambda_k \) are all numbers of the form \( 2^m + j \), \( 0 \leq j \leq 2^{m-1} - 1 \) for \( m \geq 1 \), and the length of \( I_3 \) must be \( \sum \lambda_k \), where the \( \lambda_k \) are of the form \( 2^m + j \), \( 2^{m-1} \leq j \leq 2^m - 1 \), \( m \geq 1 \), i.e.

\[
\ell(I_2) = \lambda_2 + \lambda_4 + \lambda_5 + \lambda_8 + \lambda_9 + \lambda_{10} + \lambda_{11} + \lambda_{16} + \cdots
\]

\[
\ell(I_3) = \lambda_3 + \lambda_6 + \lambda_7 + \lambda_{12} + \lambda_{13} + \lambda_{14} + \lambda_{15} + \lambda_{24} + \cdots
\]

and these uniquely determine \( I_2 \) and \( I_3 \) (note that \( 1 = \ell(I_1) = \ell(I_2) + \ell(I_3) \)) and hence uniquely determine the interval which must be removed from \( I_1 \). Lengths of subsequent intervals are uniquely determined in the same manner.

XI.17.2.11. The Cantor set \( C_\lambda \) depends on the ordered sequence \( \lambda \). In rare cases, a reordering of the sequence can yield the same Cantor set; two sequences yielding the same Cantor set must be reorderings of each other.

XI.17.2.12. The sets \( K_\alpha \) which have Lebesgue measure 0 are among the \( C_\lambda \). In fact, if \( (\alpha_n) \) is given, take \( \lambda_k = \alpha_{n-1} - 2\alpha_n \) for \( 2^{n-1} \leq k < 2^n \).

More generally, any Cantor set of Lebesgue measure 0 in \([0,1]\) containing 0 and 1 is a \( C_\lambda \). If \( C \) is such a Cantor set, then \( U_1 = [0, 1] \setminus C \) is a countable union of disjoint open intervals. Let \( \lambda_1 \) be the maximum length of the disjoint subintervals. Remove an interval of length \( \lambda_1 \) from \( U \), leaving open sets \( U_2 \) and \( U_3 \) on either side. Let \( \lambda_2 \) be the maximum length of an open subinterval in \( U_2 \), etc. This generates a sequence \( (\lambda_k) \), and it is easily seen that \( C = C_\lambda \) (Exercise ()).

The same construction can be done in any interval \([a,b]\) of \( \mathbb{R} \), starting with a series \( \sum_{k=1}^{\infty} \lambda_k \) whose sum is \( b - a \). Every Cantor subset of \( \mathbb{R} \) of Lebesgue measure 0 is such a set, i.e. a translated and scaled version of a \( C_\lambda \).

The Hausdorff measure and Hausdorff dimension of the \( C_\lambda \) have been studied by a number of authors: [Bor49], [BT54], [CMMS04], [CMPS05], [CHM10].
XI.17.3. Exercises

XI.17.3.1. (i) Show that for every $a \in [0, 1]$, there is a $b \in [0, 1]$ such that $a + b$ and $a - b$ are in $K$. [Write $a$ in its ternary expansion. Let the $n$'th digit of the ternary expansion of $b$ be 1 if the $n$'th digit of the expansion of $a$ is 1, and 0 otherwise.]

(ii) Show that every number in $[0, 2]$ can be written as the sum of two elements of $K$. [If $x \in [0, 2]$, set $a = x/2$ and choose $b$ as in (i).] How many ways are there of doing it?

(iii) Conclude from (ii), (i), and () that the Hausdorff dimension of $K$ is at least $1/2$. (In fact, by XI.17.2.5., the Hausdorff dimension of $K$ is $\frac{\log 2}{\log 3} \approx 0.63$.)
XI.18. The Topology of $\mathbb{R}^n$

XI.18.1. Invariance of Domain and the Jordan Curve Theorem

In this subsection, we describe two remarkable results about Euclidean space. Both of these fall into the category of facts which at first glance are “intuitively obvious,” but which become highly nonobvious when thought about more carefully.

We begin with the original Jordan Curve Theorem in $\mathbb{R}^2$.

XI.18.1.2. Definition. A Jordan curve in $\mathbb{R}^2$ is the image of a one-to-one continuous function from the circle $S^1$ to $\mathbb{R}^2$. A Jordan curve is also called a simple closed curve.

A Jordan curve can also be defined as the image of a continuous function $\gamma$ from a closed bounded interval $[a, b]$ into $\mathbb{R}^2$ which is one-to-one except that $\gamma(a) = \gamma(b)$. There is no assumption of smoothness or differentiability. Such a “parametrization” of the curve is highly nonunique and is not part of the structure of the curve.

A Jordan curve is intuitively a closed loop in the plane which does not cross itself. In fact, since $S^1$ is compact and Hausdorff, a Jordan curve is homeomorphic to $S^1$ by $\gamma$. (Conversely, any subset of $\mathbb{R}^2$ which is homeomorphic to $S^1$ is trivially a Jordan curve.) Geometric circles, ellipses, and polygons are elementary examples of Jordan curves. A “figure 8” is not a Jordan curve. For a more complicated example, see Figure XI.17. For more extreme examples, see Figures XI.18, XIV.25, and XIV.12.1.17.

Figure XI.17: A complicated Jordan curve (in white). Artwork by Robert Bosch
One would intuitively expect that a Jordan curve would divide the plane into two disjoint regions, the “inside” and the “outside” of the curve, just as a geometric circle does. But when more complicated examples are considered, this conclusion is less obvious; for example, is it obvious what is the “inside” and “outside” of the curve in Figure XI.17? (In particular, is the point marked with the yellow x inside or outside? Try to find a path out from this point.) Nonetheless, it does turn out to be true:

\textbf{XI.18.1.3. Theorem.} [Jordan Curve Theorem] Let $C$ be a Jordan curve in $\mathbb{R}^2$. Then $\mathbb{R}^2 \setminus C$ has exactly two connected components, one bounded and one unbounded, whose common boundary is $C$.

This result is due to C. Jordan in 1887. It is questionable whether Jordan’s proof can be regarded as complete; certainly his exposition leaves much to be desired. Conventional wisdom says that Jordan’s proof was flawed and that the first complete, correct proof was given by O. Veblen in 1905; however, this assertion may not stand up under examination [Hal07]. There have been many other proofs given over the years; see [Mae84] for a particularly nice one. We will not give a proof here. If the curve is polygonal or piecewise-smooth, the proof is relatively simple since such a curve locally divides the plane in two parts; however, the general case is difficult since a Jordan curve can be locally pathological such as the example in Figure XI.18 (although it turns out to still locally divide the plane).

Figure XI.18: A locally pathological Jordan curve
This curve, made of two intertwined spirals, is only locally pathological at one point, but small copies of the spirals can be successively added densely, replacing small arcs of the curve, to make a curve locally pathological at every point. (Much more complicated local pathology is possible too!)

There is a more refined version of the Jordan Curve Theorem, first proved by A. Schönflies in 1908(?)[]. This result is commonly called the Jordan-Schönflies theorem.

XI.18.1.4. Theorem. [Jordan-Schönflies Theorem] Let $S^1$ be the unit circle in $\mathbb{R}^2$, and $\gamma : S^1 \to \mathbb{R}^2$ a one-to-one continuous function. Then $\gamma$ extends to a homeomorphism from $\mathbb{R}^2$ onto $\mathbb{R}^2$.

It is easy to see that the Jordan-Schönflies Theorem implies XI.18.1.3. (Exercise ()). The conclusion is stronger: the “inside” (bounded component of the complement) is not only connected, but even homeomorphic to the open unit disk (and hence to $\mathbb{R}^2$ itself), and the “outside” (the unbounded component of the complement) is homeomorphic to $\mathbb{R}^2$ with a point removed. It really says that every subset of $\mathbb{R}^2$ homeomorphic to $S^1$ is topologically embedded in $\mathbb{R}^2$ in the same way.

We omit the proof, which is not easy and would take us too far afield. See [CV98], for example.

The Jordan Curve Theorem in $\mathbb{R}^n$

One natural approach to generalizing the Jordan curve theorem is to consider subsets of $\mathbb{R}^n$ homeomorphic to $S^{n-1}$. Such a subset could be called a Jordan sphere in $\mathbb{R}^n$ (this is nonstandard terminology, and it is questionable whether the name is really appropriate, so we will avoid using it).

XI.18.1.5. One might hope in analogy with the Jordan-Schönflies Theorem that a homeomorphism from the unit sphere $S^{n-1}$ in $\mathbb{R}^n$ to a subset $C$ of $\mathbb{R}^n$ would extend to a homeomorphism from $\mathbb{R}^n$ onto $\mathbb{R}^n$, i.e. that all Jordan spheres in $\mathbb{R}^n$ are topologically embedded in the same way, but the example of the “Alexander horned sphere” in $\mathbb{R}^3$ (XI.18.4.37.) shows that this statement is false in general, and thus the Jordan-Schönflies Theorem is special to $\mathbb{R}^2$. But remarkably, the Jordan Curve Theorem remains true:

XI.18.1.6. Theorem. [Jordan Curve Theorem in $\mathbb{R}^n$] Let $C$ be a subset of $\mathbb{R}^n$ ($n \geq 2$) homeomorphic to $S^{n-1}$. Then $\mathbb{R}^n \setminus C$ has exactly two connected components, one bounded and one unbounded, whose common boundary is $C$.

This theorem is due to Lebesgue and Brouwer. We again omit the proof; see, for example, [].

XI.18.1.7. The statement is not quite true for $n = 1$. The “unit sphere” $S^0$ in $\mathbb{R}^1$ consists of two points $\{\pm 1\}$ whose complement has three connected components. But topologists usually state the theorem by replacing $\mathbb{R}^n$ by its one-point compactification $S^n$ (the statement is easily seen to be equivalent; cf. Exercise ()), and this version also applies when $n = 1$.

There is a related result (which is used in the proof of XI.18.1.6.). Let $B^n$ be the closed unit ball in $\mathbb{R}^n$. Note that $B^n$ is homeomorphic to $[0, 1]^n$.
XI.18.1.8. **Theorem.** Let $B$ be a subset of $\mathbb{R}^n$ ($n \geq 2$) homeomorphic to $B^n$ (i.e. to $[0,1]^n$). Then $\mathbb{R}^n \setminus B$ is connected.

The statement also holds with $\mathbb{R}^n$ replaced by $S^n$, even for $n = 1$.

XI.18.1.9. It has been an active project in topology, not yet completed, to generalize the Jordan-Schönflies Theorem to higher dimensions. Restrictions on the embedding maps must, of course, be made. Perhaps the best result so far is that the statement is true for “locally flat” embeddings (embeddings which extend to a thin spherical shell) ([?], [?]). The result for smooth embeddings is known for $n \neq 4$, but is still open for $n = 4$.

XI.18.1.10. One might also wonder about subsets of $\mathbb{R}^n$ (or $S^n$) homeomorphic to $S^m$ for $m < n - 1$. Such subsets always have connected complements, but the ways they can be embedded (even leaving out “wild” embeddings such as the Alexander Horned sphere, e.g. restricting to smooth or piecewise-linear embeddings) are very complicated. The study of smooth embeddings of $S^1$ into $\mathbb{R}^3$ is a fascinating and active branch of modern topology called knot theory (XI.18.4.32.).

**Brouwer’s Invariance of Domain**

A very important consequence of the Jordan Curve Theorem in $\mathbb{R}^n$ is the theorem that open sets in Euclidean space are invariant under homeomorphisms. This result, called invariance of domain, was proved by L. Brouwer in 1912 [I].

XI.18.1.11. **Theorem.** [Invariance of Domain] Let $U$ be a nonempty open subset of $\mathbb{R}^n$, and $f$ a one-to-one continuous function from $U$ to $\mathbb{R}^m$. Then

(i) $m \geq n$.

(ii) If $m = n$, then $f(V)$ is open in $\mathbb{R}^n$ for every open subset $V$ of $U$ (in particular, $f(U)$ is open in $\mathbb{R}^n$), i.e. $f$ is an open mapping and a homeomorphism onto its range.

Before outlining the proof, we list some of the most important consequences.

XI.18.1.12. **Corollary.** Let $U$ be an open subset of $\mathbb{R}^n$. Then any subset of $\mathbb{R}^n$ which is homeomorphic to $U$ is also open.

The next corollary is especially noteworthy:

XI.18.1.13. **Corollary.** Let $U$ be a nonempty open subset of $\mathbb{R}^n$. Then $U$ is not homeomorphic to any subset of $\mathbb{R}^m$ for any $m < n$. In particular, $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$ for $n \neq m$.

This result also follows from the fact that any nonempty open set in $\mathbb{R}^n$ is topologically $n$-dimensional ().
XI.18.1.14. **Corollary.** Let $U$ be an open subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}^n$ a continuous function which is locally one-to-one. Then $f$ is an open mapping.

XI.18.1.15. We now outline the proof of XI.18.1.11.; actually we give a complete proof assuming XI.18.1.6. and XI.18.1.8.. We first prove (ii), so assume $m = n$. Let $x \in U$, and fix $\epsilon > 0$ with $A = B_\epsilon(x) \subseteq U$. Since $A$ is compact and $f$ is one-to-one on $A$, $B = f(A)$ is homeomorphic to $B^n$, and hence the open set $\mathbb{R}^n \setminus B$ is connected by XI.18.1.8.. If $C = f(\partial A)$, then $C$ is homeomorphic to $\partial A \cong S^{n-1}$, and hence the open set $\mathbb{R}^n \setminus C$ has two components. But $D = f(B_\epsilon(x)) = f(A \setminus \partial A)$ is connected since $B_\epsilon(x)$ is connected (), and $D$ and $\mathbb{R}^n \setminus B$ are disjoint connected sets whose union is $\mathbb{R}^n \setminus C$, so they must be the two components. But the components of an open set are open, so $D$ is open in $\mathbb{R}^n$ and contains $f(x)$. This argument can be repeated for any $x \in U$, so $f(U)$ is open. The same argument works for any open subset $V$ of $U$. Thus the inverse map from $f(U)$ to $U$ is continuous.

To prove (i), Suppose $m < n$. Define $\tilde{f} : U \to \mathbb{R}^n \cong \mathbb{R}^m \times \mathbb{R}^{n-m}$ by $\tilde{f}(x) = (f(x), 0)$. Then $\tilde{f}$ is one-to-one and continuous, but $\tilde{f}(U)$ is not open, a contradiction.

XI.18.1.16. If $U$ is an open set in $\mathbb{R}^n$ and $f$ is a one-to-one continuous function from $U$ to $\mathbb{R}^m$ ($m \geq n$), then the restriction of $f$ to any compact subset $K$ of $U$ is a homeomorphism from $K$ to $f(K)$ by (). But if $m > n$, $f$ does not give a homeomorphism from $U$ to $f(U)$ in general. For example, there is an evident one-to-one continuous map from $\mathbb{R}$ onto a figure 8 in $\mathbb{R}^2$. There are more dramatic examples: there is a one-to-one continuous function from $\mathbb{R}$ onto a dense subset of the 2-torus (or $n$-torus for any $n > 1$) obtained by “wrapping around” at an irrational angle (). See also XI.18.4.16..

XI.18.2. **$\mathbb{R}^n$ is $n$-Dimensional**

Everyone “knows” that $\mathbb{R}^n$ is $n$-dimensional. But what does this mean? Specifically, what does “dimension” mean? And how can $\mathbb{R}^n$ be distinguished from $\mathbb{R}^m$ for $m \neq n$ by its properties?

$\mathbb{R}^n$ is a set with several types of structure: algebraic, geometric, topological, and analytic. Efforts have been made to define and/or describe the “dimension” of $\mathbb{R}^n$ in terms of each of these types of structure, and various theories of dimension have been studied in different contexts; the only thing these various theories have in common is that they all assign the same number $n$ to $\mathbb{R}^n$, justifying the use of the name “dimension” for all of them.

XI.18.2.1. Specifically, we have:

(i) The vector space dimension () of $\mathbb{R}^n$ as a real vector space is $n$.

(ii) The topological dimension () of $\mathbb{R}^n$ as a topological space is $n$.

(iii) The Hausdorff dimension () of $\mathbb{R}^n$ is $n$.

XI.18.2.2. There is one respect in which all Euclidean spaces are the same, however: cardinality. Contrary to CANTOR’s expectations (he first thought the cardinality of $\mathbb{R}^n$ should be $\aleph_n$), the cardinality of $\mathbb{R}^n$ is $2^{\aleph_0}$ for all $n$ (). Thus some additional structure beyond set theory is needed to distinguish $\mathbb{R}^n$ from $\mathbb{R}^m$ for $n \neq m$. 

1205
Figure XI.19: Maple plot
Figure XI.20: Maple plot
Figure XI.21: Maple plot
Figure XI.22: Maple plot
Figure XI.23: Maple plot
Figure XI.24: Maple plot
Figure XI.25: Maple plot
Figure XI.26: Maple plot
Figure XI.27: Maple plot of graph of $f(x) = x^2 \sin \left( \frac{1}{\pi x} \right)$
Figure XI.28: Maple plot of graph of $f(x) = x \sin \left(\frac{x}{2}\right)$
Figure XI.29: Maple plot of $f(x) = \sum_{k=1}^{\infty} 2^{-k} \cos(4^k x)$
Figure XI.30: Mathematica plot
XI.18.3. When is $\mathbb{R}^n$ “Similar” to $\mathbb{R}^m$?

In this section, we ask a rather vague and subjective question: for which $n$ and $m$ are $\mathbb{R}^n$ and $\mathbb{R}^m$ “essentially similar”? The answer depends partly on what type of mathematician is answering it.

XI.18.3.1. Of course, in a certain technical sense, $\mathbb{R}^n$ and $\mathbb{R}^m$ are different for any $n \neq m$ since they are not homeomorphic (or linearly isomorphic); thus they by definition differ in some describable topological (or algebraic) way. But qualitatively all Euclidean spaces, especially ones of large dimension, seem to be “similar.” The question is really: how large does $n$ have to be for Euclidean spaces of dimension $\geq n$ all to behave essentially the same way? Is there really such a dividing line at all?

XI.18.3.2. In calculus or real analysis, the appropriate answer seems to be that all Euclidean spaces of dimension $\geq 2$ are essentially similar, but essentially different from $\mathbb{R}^1$. Calculus on $\mathbb{R}^2$ is quite different in character from calculus on $\mathbb{R}^1$, but calculus on $\mathbb{R}^5$, or $\mathbb{R}^{100}$, is not essentially different from calculus on $\mathbb{R}^2$ except for the trivial difference that notation becomes more complicated in larger dimensions. This statement is not unrestrictedly true: there are some features and results of vector analysis which are special to $\mathbb{R}^2$ and $\mathbb{R}^3$, e.g. ones related to cross product, divergence, and curl (); but even these results are mostly just special cases or interpretations of the general Stokes’ Theorem (), which holds in any dimension.

There are analytic properties unique to three- and four-dimensional space which arise in mathematical physics, e.g. stability of configurations under a gravitational force satisfying an inverse-square law. However, it is unclear whether we should regard these results as deep mathematical reasons why physical space must be three-dimensional, or space-time four-dimensional, or whether the three-dimensionality of the universe (as we currently perceive and understand it) and the compatibility of physical laws with three-dimensionality are just physical facts not entailed by mathematical considerations.

XI.18.3.3. A functional analyst might well give an even less discriminating answer: that all finite-dimensional Hilbert spaces (in particular, all Euclidean spaces) are essentially similar and essentially trivial, and the first important distinction is between finite-dimensional and infinite-dimensional Hilbert spaces.

XI.18.3.4. On the other hand, a topologist would likely say that the Euclidean spaces of dimension $\leq 5$ are all essentially different, while those of dimension $\geq 5$ are essentially similar. They would probably go on to add that $\mathbb{R}^4$ is “more different” than any of the others, for various reasons (). One can identify simple and important topological differences between $\mathbb{R}^2$ and $\mathbb{R}^3$, e.g. XI.18.5.3., Classification of closed manifolds, Banach-Tarski, Poincaré Conjecture, and between $\mathbb{R}^3$ and $\mathbb{R}^n$ for $n \geq 5$ (). Topologists even consider “low-dimensional topology” (of dimension $\leq 4$) to be a separate subdiscipline.

XI.18.3.5. There is a qualitative difference between even- and odd-dimensional spaces, however: for example, even and odd spheres are qualitatively different. Only odd spheres have globally nonvanishing vector fields (), only even spheres have orientation-reversing homeomorphisms without fixed points, etc. The differences can be interpreted as reflecting the difference between the complex $K$-theories of even and odd spheres ()

There are many other differences in the structure of $n$-manifolds for different $n$: for example, every $n$-dimensional PL manifold has a smooth structure (in fact a unique one) if $n \leq 6$, but if $n \geq 7$ there are closed $n$-dimensional PL-manifolds which do not have a smooth structure. And smooth structures, if they exist, need not be unique. The $n$-sphere has a unique smooth structure for $n = 1, 2, 3, 5, 6, 12$ and 61, and
more than one for all other \( n \leq 126 \) except possibly \( n = 4 \) (open), and all other odd \( n \) [\( \text{[]} \)]. The cases of even \( n > 126 \) are partially but not completely known.

It may be a stretch to regard the question as including the structure of manifolds of different dimensions, since the definition of a general manifold is rather subtle; but it is less of a stretch to include the structure of spheres, since the \( n \)-sphere is just the one-point compactification of \( \mathbb{R}^n \), or the unit sphere in \( \mathbb{R}^{n+1} \). One might also legitimately consider the topology of open subsets of \( \mathbb{R}^n \) (cf. XI.18.5.6–XI.18.5.8).

**XI.18.3.6.** There are more esoteric dimensional restrictions arising from topology. Two 1-spheres (circles) can be linked nontrivially in \( \mathbb{R}^n \) if and only if \( n = 3 \). But if \( m > 1 \), the question of whether two disjoint (topological) \( m \)-spheres can be nontrivially linked in \( \mathbb{R}^n \) (\( n > m \)) is much more subtle: for example, two 50-spheres can be nontrivially linked in \( \mathbb{R}^n \) if and only if \( 52 \leq n \leq 95 \) or \( 98 \leq n \leq 101 \) [Rol76].

**XI.18.3.7.** An algebraist might say \( \mathbb{R}^1, \mathbb{R}^2, \mathbb{R}^4, \) and \( \mathbb{R}^8 \) are similar to each other and different from all other \( \mathbb{R}^n \) because of III.6.3.2. In fact, for arbitrarily large \( n \), the algebraic structure of \( \mathbb{R}^n \) is significantly impacted by the number-theoretic properties of \( n \) (https://mathoverflow.net/questions/34186/does-the-truth-of-any-statement-of-real-matrix-algebra-stabilize-in-sufficiently), and much of the algebraic structure can be interpreted geometrically.

**XI.18.3.8.** In geometric measure theory, there is a significant difference between \( \mathbb{R}^n \) for \( n \leq 7 \) and for \( n \geq 8 \): if \( n \leq 7 \), an area-minimizing rectifiable \((n-1)\)-current in \( \mathbb{R}^n \) is a smooth submanifold (\( \text{[]} \)), but this is not necessarily true if \( n \geq 8 \). See [Mor09, 10.7].

**XI.18.3.9.** There is a potentially much higher dividing line from the Borsuk conjecture (or problem): BORSUK asked whether any bounded subset of \( \mathbb{R}^n \) is the union of \( n + 1 \) subsets of strictly smaller diameter. The statement is known to be true for \( n \leq 3 \) and false for \( n \geq 64 \). The smallest \( n \) for which the statement fails is in at least one combinatorial sense an important transition point.

**XI.18.3.10.** There are also interesting and important questions concerning sphere packings. One tries to fill \( \mathbb{R}^n \) with solid equal-size \( n \)-balls as densely as possible. The problem is even tricky to state precisely: we first seek the supremum of the density of all possible packings, and then to find a packing (or “all” packings) realizing the supremum (optimal packings; one should impose additional conditions on optimal packings such as some type of incompressibility, part of the difficulty in phrasing the problem carefully). In \( \mathbb{R}^2 \) a hexagonal lattice is easily shown to give the unique optimal packing. In \( \mathbb{R}^3 \) the cubic face centered lattice (\( \text{[]} \)) is optimal (Kepler’s Conjecture), although the proof is very difficult and was only recently completed [HF11]; it is not the unique optimal packing, although the other optimal packings are easily described variations. Optimal packings are known in dimensions \( \leq 8 \) and 24, and are lattice packings in these dimensions. In dimensions 10, 11, and 13 there are irregular packings which are denser than any known lattice packing, and it may be that in these dimensions, and possibly in all sufficiently large dimensions, optimal packings are necessarily irregular. It is unclear whether there even are any optimal packings (i.e. ones realizing the supremum) in every dimension. See [CS99] for details. Sphere packing problems in large numbers of dimensions have applications in number theory, coding, algebra, and mathematical physics as well as in geometry.

**XI.18.3.11.** As to the general question of whether there exists a “dividing line,” many (but certainly not all) questions or phenomena stabilize for high dimensions and are difficult only due to “accidents of small dimension.” From the right point of view, this philosophy applies even to such problems as the notorious one
of computing higher homotopy groups $\pi_n(S^q)$ of spheres, which is not completely solved and is perhaps not even completely solvable in principle. Viewed in the usual way, by fixing $q$ and letting $n$ vary, the question can seem intractable. But if we instead fix $n$ and let $q$ vary, the problem stabilizes nicely for $q \geq n$ and the complications can be said to be only due to “accidents of small dimension.”

Of course, as seen above, if dividing lines exist at all, they may vary from problem to problem.

XI.18.3.12. Perhaps the best answer to the original question may be that it simply reflects the limitations of our current knowledge. Even topologists as recently as the 1960s might not have said that $\mathbb{R}^4$ is essentially different from $\mathbb{R}^3$; the most profound differences have all been discovered since then. Mathematicians 100 years from now may regard $\mathbb{R}^{18}$ and $\mathbb{R}^{19}$ (or maybe $\mathbb{R}^{17}$ and $\mathbb{R}^{19}$, to pick prime numbers) as being as different as $\mathbb{R}^3$ and $\mathbb{R}^4$ due to some phenomenon not yet imagined.

XI.18.3.13. If you want to start some lively discussions, ask your colleagues or instructors this question! I decided to pose the question on MathOverflow (https://mathoverflow.net/questions/267909/when-are-mathbb-rn-and-mathbb-rm-essentially-similar) and the ensuing discussion has informed and improved my own discussion above.
XI.18.4. Compact Subsets of $\mathbb{R}^n$

In this section, we describe a number of examples of compact subsets of Euclidean space with unusual, even bizarre, properties. We will state some of the properties of the examples without proof; readers with some topological background are encouraged to verify the asserted properties.

Mathematicians tend to have two types of reactions to examples of this sort. Some find them fascinating; others have more a sense of revulsion and want to sweep them into an obscure corner. I myself have the first reaction. But even if you fall into the second camp, it is still very important to be aware of the existence of such examples, since they graphically illustrate two vital mathematical principles:

(i) The number and variety of mathematical objects, even ones of relatively restricted type such as closed bounded subsets of the plane, is often vast, far beyond the limits of one’s initial imagination.

(ii) When proving results about general mathematical objects of a specified type, it is necessary to stick strictly to arguments drawing logical conclusions from explicit axiomatic assumptions, without appeal to intuition or heuristic arguments, especially ones based on pictures or diagrams which may not accurately reflect the full range of possibilities for the objects studied.

Of course, various other examples, such as the examples of continuous and/or differentiable functions with pathological properties which are scattered through this book, also illustrate these principles.

It turns out that subsets similar to some of these occur naturally in mathematics, both pure and applied, such as strange attractors for nonlinear dynamical systems.

Some of the examples of this section are fractals, and their fractal properties will be examined in ( ).

Uncountably Many Mutually Nonhomeomorphic Compact Subsets of $\mathbb{R}^2$

XI.18.4.1. If $k = (k_1, k_2, \ldots)$ is a sequence of natural numbers with $k_1 \geq 2$, construct a compact subset $X_k$ of $\mathbb{R}^2$ as follows. Begin with the closed segment from $(0,0)$ to $(1,0)$. For each $n$, add the closed line segments between $(1/n, 0)$ and $(1/(n+1), 1/(n(n+1)))$ for $1 \leq m \leq k_n$.

For example, Figure XI.31 shows $X_k$ for $k = (2, 3, 1, 1, \ldots )$.

It is clear that $X_k$ is a closed bounded subset of $\mathbb{R}^2$. It can be shown with some straightforward elementary topological arguments (Exercise XI.18.5.1.) that the $X_k$ for distinct sequences $k$ are mutually nonhomeomorphic. Since there are uncountably many $(2^{k_0})$ such $k$, there are uncountably many $(2^{k_0})$ mutually nonhomeomorphic compact subsets of the plane. Since there are only $2^{k_0}$ closed subsets of $\mathbb{R}^2$, there are exactly $2^{k_0}$ homeomorphism classes of closed bounded subsets of the plane.

Each of the $X_k$ is path-connected and topologically one-dimensional. They are even contractible; in fact, they are absolute retracts (dendrites) ( ). Note that these spaces are quite different from the infinite broom of XI.13.9.9.

There is an enormous variety of variations possible on this construction.
Figure XI.31: $X_k$ for $k = (2, 3, 1, 1, \ldots)$.
Uncountably Many Mutually Nonhomeomorphic Compact Subsets of $\mathbb{R}$

XI.18.4.2. Any countable ordinal $\alpha$ can be order-embedded in $\mathbb{R}$ ($\alpha$). Any countable ordinal which is not a limit ordinal can be embedded as a compact subset of $\mathbb{R}$ (with the order topology). There are $\aleph_1$ such ordinals.

However, these are not all distinct topologically. But the ordinals $\omega^\sigma + 1$ are all distinct topologically: the derived series ($\omega$) as a topological space has length $\sigma$. This space is countable if $\sigma$ is countable. Thus there are at least $\aleph_1$ topologically distinct compact countable subsets of $\mathbb{R}$.

It is trickier to explicitly construct $2^{\aleph_0}$ topologically distinct compact subsets of $\mathbb{R}$. Here is a way (cf. [Rei62]). Let $K$ be the Cantor set, and $c_n = 1 - \frac{1}{3^n}$. Then $c_n \in K$, and is the right endpoint of the rightmost open interval $I_n$ removed at the $n$th step of the construction of $K$. Let $m = (m_1, m_2, \ldots)$ be a strictly increasing sequence in $\mathbb{N}$, and for each $n$ attach to $K$ a copy of $\omega^{m_n}$ in $I_n$ which is relatively closed and converges to $c_n$ (see Figure ()). Let $X_m$ be the resulting subset of $\mathbb{R}$. Then $X_m$ is a compact totally disconnected subset of $\mathbb{R}$ for each $m$, and the $X_m$ are topologically distinct for distinct $m$. There are $2^{\aleph_0}$ such $m$, since they are in one-one correspondence with the infinite subsets of $\mathbb{N}$. See Exercise XI.18.5.2.. Since there are only $2^{\aleph_0}$ compact subsets of $\mathbb{R}$, there are exactly $2^{\aleph_0}$ homeomorphism classes.

The Topologist’s Sine Curve

XI.18.4.3. Let $X$ be the graph of $y = \sin \left( \frac{x}{\pi} \right)$ for $0 < x \leq 1$ (the right endpoint is not important), along with the closed segment on the $y$-axis between $(0, -1)$ and $(0, 1)$. Then $X$ is the closure of the graph part, and is a compact subset of $\mathbb{R}^2$. See Figure XI.32.

![The Topologist’s Sine Curve](image)

Figure XI.32: The Topologist’s Sine Curve
Since the graph is connected (even path-connected), $X$ is the closure of a connected set and hence is connected. However, $X$ is not path-connected; it has two path components, the graph and the segment on the $y$-axis.

**XI.18.4.4.** A variation is to “close up” $X$ by adding an arc connecting $(1,0)$ with $(0,-1)$. This is often called the “extended topologist’s sine curve” or “Warsaw circle.” See Figure XI.33. This space is path-connected but not locally path-connected. In some references, the added arc goes from $(1,0)$ around to $(0,0)$ from the left; this space is topologically different but essentially similar.

![Image of the extended topologist’s sine curve or Warsaw circle](image)

**Figure XI.33:** The Extended Topologist’s Sine Curve or Warsaw Circle

**XI.18.4.5.** The topologist’s sine curve can be used to construct a “counterexample” to the Four-Color Theorem (Figure XI.34). Each of the five regions pictured has a boundary with each of the others which at least contains an entire arc, and thus they must all be colored differently. This shows that the Four-Color Theorem must be stated carefully to rule out examples of this sort (cf. [?]).
Figure XI.34: A counterexample to the Four-Color Theorem?
The Hawaiian Earring

XI.18.4.6. Let $X$ be the union of all circles in $\mathbb{R}^2$ with center $(\frac{1}{n}, 0)$ and radius $\frac{1}{n}$, for $n \in \mathbb{N}$ (Figure XI.35).

Figure XI.35: The Hawaiian Earring

The fundamental group of the Hawaiian earring is uncountable: for each subset $S$ of $\mathbb{N}$, there is a loop $\gamma_S$ which goes around the circle of radius $1/n$ if and only if $n \in S$, and no two of these loops are homotopic.

XI.18.4.7. Let $Y$ be the union of all circles in $\mathbb{R}^2$ with center $(n, 0)$ and radius $n$, for $n \in \mathbb{N}$. Then $Y$ is not homeomorphic to $X$, since it is not compact. The fundamental group of $Y$ is a free group on a countable number of generators. In fact, $Y$ is homotopy equivalent (but not homeomorphic) to an infinite wedge of circles (XI.9.1.13., XI.9.5.8.) [collapse a suitable closed neighborhood of $0$ to a point]. See also XI.9.5.5.
Subsets Constructed as Intersections

A number of pathological examples can be made by constructing a decreasing sequence of increasingly complicated compact subsets of Euclidean space and taking the intersection. This construction is already familiar in the definition of the Cantor set ( ).

The Sierpiński Gasket

This subset of $\mathbb{R}^2$ is constructed as an intersection, and is the first of several straightforward higher-dimensional versions of the Cantor set.

Begin with a (solid) triangle in $\mathbb{R}^2$, say an equilateral triangle. Bisecting each side and connecting the midpoints forms four subtriangles. Remove the (open) middle one. Then repeat the process in each of the three remaining triangles and iterate. The intersection is called the Sierpiński gasket (the name is due to B. Mandelbrot; the set is also known as the Sierpiński triangle or the Sierpiński sieve). It is path-connected, compact, and topologically one-dimensional. See Figure XI.36 (from http://en.wikipedia.org/wiki/File:Sierpinski_Triangle.svg).

![Figure XI.36: Six steps in the construction of the Sierpiński Gasket](http://en.wikipedia.org/wiki/File:Sierpinski_Triangle.svg)

The Sierpiński Carpet

XI.18.4.8. As a variation, begin with a (solid) square. Divide it into 9 equal-sized squares, and remove the open middle square. Repeat the process in each of the eight remaining squares, and iterate. See

1227
Figure XI.37 (from http://commons.wikimedia.org/wiki/File:Sierpinski6.png?uselang=fr). This set is called the Sierpiński carpet, or Sierpiński curve.

Like the Sierpiński gasket, the Sierpiński carpet is topologically one-dimensional.

![Sierpiński Carpet](http://commons.wikimedia.org/wiki/File:Sierpinski6.png)

**Figure XI.37**: Six steps in the construction of the Sierpiński Carpet

**XI.18.4.9.** It is not obvious whether the Sierpiński carpet is homeomorphic to the Sierpiński gasket. In fact, they are not homeomorphic. For each point of the Sierpiński gasket has arbitrarily small open neighborhoods whose boundary consists of finitely many points (no more than 4), while small open neighborhoods of points in the Sierpiński carpet always have uncountably many boundary points. See [Kur68, §51] for a discussion.

**XI.18.4.10.** The Sierpiński carpet is universal for one-dimensional compact subsets of the plane: every compact subset of the plane which is topologically one-dimensional is homeomorphic to a subset of the
Sierpiński carpet[]. In particular, the Sierpiński gasket is homeomorphic to a subset of the Sierpiński carpet (try to find one!) A pseudo-arc can also be embedded topologically in the Sierpiński carpet.

**Swiss Cheese**

**XI.18.4.11.** The Sierpiński carpet is (topologically) a special case of a general construction: begin with a closed disk in $\mathbb{R}^2$, and remove a sequence of disjoint open disks to obtain a set with empty interior. Such a subset of $\mathbb{R}^2$ is called a *Swiss cheese*. There is a great topological variety among Swiss cheeses; some are similar topologically to the Sierpiński gasket, and some are homeomorphic to the Sierpiński carpet. The one pictured in [XI.38](http://hans.math.upenn.edu/~pstorm/images/round_Sierpinski_carpet.png) is made in a regular manner and is a fractal; some are much more irregular. As with Cantor sets, Swiss cheeses of positive planar Lebesgue measure can be constructed, as well as ones of zero planar measure; in fact, Swiss cheeses of Hausdorff dimension between 1 and 2 can be constructed, although the exact Hausdorff dimensions are subtle to compute ([Mel66], [Lar66]).

If topological disks are used instead of geometric ones, an even greater variety can be obtained; both the Sierpiński carpet and the Sierpiński gasket are special cases of this construction, as are the Lakes of Wada.

Swiss cheeses are used extensively in the theory of function algebras(); see [FH10] for a good survey, including their topological structure.
Figure XI.38: A “Swiss cheese.” (Small white disks continue throughout the black spots.)
The Menger Sponge

XI.18.4.12. There is a three-dimensional analog of the Sierpiński carpet, due to K. Menger [Men26]. Begin with the unit cube in \( \mathbb{R}^3 \). Dividing each edge in three equal segments leads to a decomposition of the cube into 27 equal-sized subcubes. Remove the central subcube and the central subcubes on each face (i.e. keeping all closed subcubes intersecting an edge of the cube). Seven subcubes are removed in all. Do the same in each of the 20 remaining subcubes and iterate. The intersection is called the Menger sponge. Each outside face of the Menger sponge is a copy of the Sierpiński carpet. See Figure XI.39 (from \( \text{http://en.wikipedia.org/wiki/File:Menger-Schwamm-farbig.png} \)).

![Four steps in the construction of the Menger Sponge](http://en.wikipedia.org/wiki/File:Menger-Schwamm-farbig.png)

Figure XI.39: Four steps in the construction of the Menger Sponge

The Menger sponge is compact, connected, locally connected, and topologically one-dimensional. It is universal for one-dimensional separable metrizable spaces: every separable metrizable space of dimension \( \leq 1 \) is homeomorphic to a subset of the Menger sponge [Men26]. Thus the Menger sponge is sometimes called the Menger Universal Curve (although it is not a “curve” in any normal sense). It is homogeneous and has a topological characterization [And58].

For any \( k, n \in \mathbb{N}, k < n \leq 2k + 1 \), there is a similar subset \( M^n_k \) of \( \mathbb{R}^n \) which is universal for compact subsets of \( \mathbb{R}^n \) of dimension \( \leq k \) [Sta71]; \( M^{n+1}_k \) is universal for all separable metrizable spaces of dimension
$M_3^1$ is the Menger sponge, $M_2^2$ the Sierpiński carpet, and $M_0^1$ the Cantor set. To make $M_2^2$, begin with the unit cube in $\mathbb{R}^3$, subdivide into 27 subcubes, and just remove the center cube; then iterate.

**Sierpiński Pyramids**

At least two versions of the Sierpiński Gasket can be made in $\mathbb{R}^3$. Begin with a pyramid with either a triangular or square base. In the triangular case (say, a regular tetrahedron), bisect each edge and keep the four subpyramids whose vertices are one original vertex and the three midpoints of the adjacent edges; then iterate. The intersection is a compact set called the tetrahedral Sierpiński Pyramid which is topologically one-dimensional. It has the unusual property that its Hausdorff dimension is an integer (2) which is greater than its topological dimension.

A variation is to begin with a pyramid with square base, bisect the edges, and keep the five subpyramids consisting of the top pyramid with four vertices at the midpoints of the slant edges and the four pyramids with top vertex at one of these midpoints and bottom vertices at a corresponding original vertex, two midpoints of base edges, and the midpoint of the base. See Figure XI.40, from [http://en.wikipedia.org/wiki/File:Sierpinski_pyramid.png](http://en.wikipedia.org/wiki/File:Sierpinski_pyramid.png).

![Sierpiński Pyramid](http://en.wikipedia.org/wiki/File:Sierpinski_pyramid.png)

**Figure XI.40:** Seven steps in the construction of the Sierpiński Pyramid with square base and its “inverse”
Solenoids

XI.18.4.13. Let $s = (n_1, n_2, \ldots)$ be a sequence of natural numbers with $n_k \geq 2$ for all $k$. Construct a compact subset of $\mathbb{R}^3$ as follows (in understanding the construction, try taking all $n_k$ to be 2 at first). Let $X_0$ be a solid torus. Inside $X_0$ take a thinner solid torus $X_1$ which “winds around” $n_1$ times. (Think of wrapping a wire around the “hole” in $X_0$ $n_1$ times and connecting the ends of the wire to make a thin solid torus.) See XI.41 for the case $n_1 = 2$ (from http://en.wikipedia.org/wiki/File:Smale-Williams_Solenoid.png).

Now within $X_1$ take an even thinner solid torus $X_2$ which wraps around $X_1$ $n_2$ times, i.e. every cross-

Figure XI.41: First step in the construction of the solenoid of type $(2, 2, \ldots)$

section disk of $X_1$ will cut $X_2$ in $n_2$ cross-sectional disks. Repeat the process countably many times and let $X = \cap_k X_k$. $X$ is called the solenoid of type $s$.

A nice animated construction of the solenoid of type $(2, 2, \ldots)$ can be found at http://en.wikipedia.org/wiki/File:Solenoid.gif.

XI.18.4.14. Solenoids can also be naturally constructed in several other ways. One good way is as inverse limits of circles (XI.14.4.17).

XI.18.4.15. A solenoid is connected but not locally connected: locally, a solenoid looks like $\mathbb{R} \times K$, where $K$ is the Cantor set. In fact, each solenoid can be constructed (up to homeomorphism) from $[0, 1] \times K$ by identifying $\{0\} \times K$ with $\{1\} \times K$ in an appropriate way, e.g. by identifying $K$ with the $p$-adic integers () and identifying $(1, x)$ with $(0, x + 1)$ (this gives the solenoid of type $(p, p, p, \ldots)$).

XI.18.4.16. A solenoid is not path-connected. There are uncountably many path components, each of which is dense and a one-to-one continuous image of $\mathbb{R}$. This is seen most easily by viewing the solenoid as an inverse limit as in (), but can also be deduced from the description in XI.18.4.15. (in this picture, the path components correspond to the cosets of $\mathbb{Z}$ in the $p$-adic integers).
XI.18.4.17. Up to homeomorphism, a solenoid depends only on the sequence $s$. Solenoids corresponding to different sequences are “usually” nonhomeomorphic, but ones corresponding to sequences differing only “slightly” can be homeomorphic.

To give a precise description of the homeomorphism classes, we use the supernatural numbers:

XI.18.4.18. **Definition.** A *supernatural number* is an expression of the form

$$n = 2^{e_2}3^{e_3}5^{e_5} \ldots$$

where each prime appears and each exponent $e_p$ is in $\{0\} \cup \mathbb{N} \cup \{\infty\}$.

An ordinary natural number is a supernatural number in which all $e_p$ are finite and all but finitely many are 0. Sometimes we want to exclude the ordinary natural numbers from the supernatural numbers; we call supernatural numbers which are not ordinary natural numbers *infinite*. When writing a supernatural number, we usually leave out all primes with exponent 0; thus if $n$ is the supernatural number with $e_2 = \infty$ and $e_p = 0$ for $p \neq 2$, we write $n$ as $2^\infty$.

Two supernatural numbers can be multiplied by adding the corresponding exponents. This multiplication extends the ordinary multiplication on $\mathbb{N}$.

We say one supernatural number $n$ *divides* another supernatural number $m$ if $e_p(n) \leq e_p(m)$ for all $p$, i.e. if there is a supernatural number $r$ such that $m = nr$. (Note, however, that $r$ is not necessarily unique, i.e. that there is not a well-defined quotient in general, if the exponents for some $p$ are both infinite.) There is a “largest” supernatural number, the one for which $e_p = \infty$ for all $p$, and all supernatural numbers divide this one.

Note that the supernatural numbers are not the same thing as the nonstandard natural numbers $^*\mathbb{N}$ of ()).

XI.18.4.19. The supernatural numbers are in natural one-one correspondence with the additive subgroups () of $\mathbb{Q}$ containing $\mathbb{Z}$: the group $\mathbb{Z}(n)$ is the set of all rational numbers whose denominator divides $n$. For example, $\mathbb{Z}(2\infty)$ is the set of dyadic rationals ()). The group corresponding to the largest supernatural number is $\mathbb{Q}$ itself; $\mathbb{Z}$ corresponds to the supernatural number 1 (i.e. $e_p = 0$ for all $p$), the smallest supernatural number.

The group $\mathbb{Z}(n)$ is cyclic (has a smallest positive element) if and only if $n$ is an ordinary natural number.

XI.18.4.20. To each sequence $(n_k)$ of natural numbers, there is a naturally associated supernatural number $\prod n_k$, for which, for every $p$, $e_p$ is the sum of the powers of $p$ occurring in the $n_k$’s. This will be an ordinary integer if and only if all but finitely many $n_k$ are 1. Many different sequences can give the same associated supernatural number (for example, reordering the sequence does not change the associated supernatural number, and there are much more drastic changes that work too).

It can be shown without too much difficulty (knowing the first statement of XI.18.4.17.) that if two sequences give the same supernatural number, the corresponding solenoids are homeomorphic. But it is not quite true that solenoids corresponding to different supernatural numbers are nonhomeomorphic. For a correct statement, we need an equivalence relation on supernatural numbers:

XI.18.4.21. **Definition.** Let $n$ and $m$ be supernatural numbers. Then $n$ and $m$ are *equivalent* if there are ordinary natural numbers $n$ and $m$ such that $nn = mm$. 

1234
Supernatural numbers \( n \) and \( m \) are equivalent if and only if only finitely many of their exponents differ, and each by only a finite amount. The supernatural numbers equivalent to any ordinary natural number are precisely all ordinary numbers.

This equivalence is natural because of the following results, the first of which is quite easy:

**XI.18.4.22. Proposition.** Two additive subgroups of \( \mathbb{Q} \) containing \( \mathbb{Z} \) are isomorphic as groups if and only if their associated supernatural numbers are equivalent.

**XI.18.4.23. Theorem.** Two solenoids are homeomorphic if and only if their associated supernatural numbers are equivalent.

**XI.18.4.24.** Thus the equivalence classes of infinite supernatural numbers exactly parametrize the homeomorphism classes of solenoids. The equivalence class of supernatural numbers can be recovered directly from the intrinsic topology of the solenoid \( X \): the first cohomotopy group \( \pi^1(X) \) is isomorphic to the additive subgroup of \( \mathbb{Q} \) associated to the equivalence class of supernatural numbers of \( X \).

**XI.18.4.25.** There is another important connection. Any solenoid \( X \) can be made into a compact abelian topological group. Its Pontrjagin dual group \( \hat{\pi} \) is isomorphic to \( \pi^1(X) \). The solenoids are the only compact connected one-dimensional Hausdorff spaces besides the circle which can be made into abelian topological groups. See XVI.5.2.1.(b).

**XI.18.4.26.** Solenoids are homogeneous. It was shown in [Hag77] that every homogeneous continuum in which every nondegenerate subcontinuum is an arc is a solenoid (or circle). A solenoid is also indecomposable.

**XI.18.4.27.** Solenoids arise naturally as attractors of important nonlinear dynamical systems ().

**Knaster’s Bucket-Handle**

This is a “poor man’s solenoid,” somewhat easier to visualize and with similar topological properties, at least locally. It is perhaps the simplest example of an indecomposable continuum. It can be constructed as an intersection (see Figure XI.43), but has a simpler direct definition:

**XI.18.4.28. Definition.** The *Knaster bucket-handle* is the union of the following closed half-circles in \( \mathbb{R}^2 \):

- All half-circles in the upper half-plane with center \( \left( \frac{1}{2}, 0 \right) \) with endpoints in the Cantor set \( K \).
- All half-circles in the lower half-plane with center \( \left( \frac{5}{3^{2n}}, 0 \right) \) \((n \geq 1)\) passing through a point of \( K \cap \left[ \frac{2}{3^n}, \frac{1}{3^{n-1}} \right] \).
Figure XI.42: The Knaster bucket-handle

See Figure XI.42 (from http://upload.wikimedia.org/wikipedia/commons/thumb/3/30/The_Knaster_%22bucket-handle%22_continuum.svg/2000px-The_Knaster_%22bucket-handle%22_continuum.svg.png).

It can be shown that the Knaster bucket-handle is indecomposable (cf. [? 48, V]): although it is connected, it has uncountably many path components (composants), one an injective continuous image of $[0, +\infty)$ and the rest injective continuous images of $\mathbb{R}$. A proper subcontinuum cannot contain any entire path component, so its intersection with any path component is an arc; it follows that it must be contained in a path component, i.e. the complement of any proper subcontinuum is dense.

It turns out there is no Borel subset containing exactly one point from each path component (cf. [?]).

There are numerous variations of the construction, with slightly different properties. All these can be described as inverse limits of intervals; see XI.14.4.18.

Attractors called “horseshoes” () are types of Knaster bucket-handles.

Figure XI.43: Four steps in the intersection construction of the Knaster bucket-handle
The Pseudo-Arc

XI.18.4.29. The construction of a pseudo-arc is somewhat involved.

We first need the notion of a *crooked* path. Let \( R \) be a (solid) rectangle, subdivided into a strip \( R_1, \ldots, R_m \) of subrectangles, \( a \) a point in the interior of \( R_1 \), and \( b \) a point in the interior of \( R_m \):

\[
\begin{array}{cccc}
& a & & \\
R_1 & R_2 & R_3 & \ldots \\
& b & & \\
& & & R_m
\end{array}
\]

Figure XI.44: Strip Subdivision of a Rectangle

A path \( \gamma \) from \( a \) to \( b \) within \( R \) is *crooked* with respect to \( R_1, \ldots, R_m \) if it has the following property: whenever \( t_1, t_2 \in [0,1], t_1 < t_2, \gamma(t_1) \in R_i, \) and \( \gamma(t_2) \in R_j \) with \( |i-j| > 2 \), there are \( s_1, s_2 \) with \( t_1 < s_1 < s_2 < t_2 \) and

- if \( i < j \), then \( \gamma(s_1) \in R_{j-1} \) and \( \gamma(s_2) \in R_{i+1} \)
- if \( j < i \), then \( \gamma(s_1) \in R_{j+1} \) and \( \gamma(s_2) \in R_{i-1} \).

Instead of a rectangle divided into subrectangles, a chain () of closed balls can be used. A crooked path is defined in the same way.

Thus, for example, if \( m = 4 \), a crooked path from \( a \) to \( b \) must first go to \( R_3 \), then back to \( R_2 \), and then to \( R_4 \). So the simplest crooked path looks like this:

\[
\begin{array}{cccc}
& a & & \\
R_1 & R_2 & R_3 & R_4 \\
& b & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
& & & \\
\end{array}
\]

Figure XI.45: Crooked Path of Length 4

Crooked paths rapidly become more complex as \( m \) increases (if \( \ell(m) \) is the number of horizontal segments in a minimal crooked path of length \( m \), for \( m \geq 4 \) we have the recursive formula

\[
\ell(m) = 2\ell(m-1) + \ell(m-2)
\]

with \( \ell(2) = \ell(3) = 1 \). For example, if \( m = 5 \), a crooked path from \( a \) to \( b \) must first go to \( R_3 \), then back to \( R_2 \), then \( R_4 \), then \( R_2 \), then \( R_4 \), then \( R_3 \), and finally \( R_5 \):

A crooked path from \( a \) to \( b \) in general consists of a crooked path from \( a \) to a point in \( R_{m-1} \), followed by a crooked path from that point to a point in \( R_{m-2} \), followed by a crooked path from that point to \( b \). Thus
Figure XI.46: Crooked Path of Length 5

the picture on [HY88, p. 142] is not correct: an actual crooked path is far more complicated (there would be $14,398,739,476,117,879 \approx 1.4 \times 10^{16}$ horizontal segments for a minimal crooked path of length 45).

To construct a pseudo-arc, fix two points $a$ and $b$ in $\mathbb{R}^2$, and then:

1. Let $X_1$ be a narrow (solid) rectangle of width $< 1$, containing $a$ and $b$ in its interior, and subdivide $X_1$ into a strip of subrectangles $R_{1,1}, \ldots, R_{1,k_1}$, each of diameter $\leq 1$, with $a$ in the interior of $R_{1,1}$ and $b$ in the interior of $R_{1,k_1}$. Alternatively, let $X_1$ be the union of a chain of closed balls from $a$ to $b$.

2. Take a path from $a$ to $b$ in $X_1$ which is simple (does not cross itself) and crooked for the $R_{1,j}$.

3. Fatten up the path within $X_1$ into a thin strip $X_2$ homeomorphic to a (solid) rectangle, and subdivide $X_2$ into a strip of subsets $R_{2,1}, \ldots, R_{2,k_2}$ homeomorphic to subrectangles, of diameter $\leq \frac{1}{2}$, with $a$ in the interior of $R_{2,1}$ and $b$ in the interior of $R_{2,k_2}$. Alternatively, cover the path with a chain of closed balls, and let $X_2$ be the union.

4. Take a simple path from $a$ to $b$ in $X_2$ which is crooked for the $R_{2,j}$, and iterate the process.

Let $X = \cap_{n=1}^{\infty} X_n$. Then $X$ is a compact subset of $\mathbb{R}^2$ containing $a$ and $b$. Since each $X_n$ is homeomorphic to a (solid) rectangle, it is connected; thus $X$ is connected by XI.13.3.6. $X$ is called a pseudo-arc.

A pseudo-arc is compact and connected (a continuum), and topologically one-dimensional. It is an indecomposable continuum: it cannot be written as a union of two proper subcontinua. In this respect it is enormously different from an ordinary arc. In fact, a pseudo-arc is a hereditarily indecomposable continuum: every subcontinuum is also indecomposable. Actually, every subcontinuum of the pseudo-arc with more than one point is homeomorphic to the whole pseudo-arc (an ordinary arc has this property too). And unlike an ordinary arc which has endpoints topologically different from the other points of the arc, a pseudo-arc is homogeneous: for any points $p$ and $q$, there is a homeomorphism of the pseudo-arc to itself sending $p$ to $q$. These properties are proved in [Bin48], where it is also shown that the pseudo-arc is the unique homogeneous hereditarily indecomposable continuum, and also [Bin51a] the unique chainable (hereditarily indecomposable) continuum. In particular, up to homeomorphism the pseudo-arc does not depend on the details of the choice of the $X_n$, so it makes sense topologically to refer to the pseudo-arc. The pseudo-arc can be constructed as an inverse limit of intervals (XI.14.4.18.(ii)).

There is no nonconstant continuous function from an ordinary arc into a pseudo-arc; thus the path components of a pseudo-arc are singletons, i.e. the pseudo-arc is connected but totally path-disconnected.
XI.18.4.30. Pseudo-arcs may seem pathological, but “almost every” continuum in $\mathbb{R}^2$ is a pseudo-arc [Bin51a], in the sense that if the set $C$ of continua in $\mathbb{R}^2$ is given the Hausdorff metric (,) making it into a compact metrizable space, the set of continua which are not pseudo-arcs is meager (). See http://topology.auburn.edu/pm/pseudodraw.pdf for an interesting discussion of the pseudo-arc, including the difficulties inherent in trying to draw a picture of it.

XI.18.4.31. There is also a similar pseudo-circle, described in [Bin51a], whose complement is homeomorphic to the complement of a circle (a bounded component homeomorphic to a disk and an unbounded one). The pseudo-circle is also a hereditarily indecomposable continuum (every proper subcontinuum is a point or a pseudo-arc), and there is also a uniqueness statement up to homeomorphism for pseudo-circles. However, the pseudo-circle is not homogeneous ([?], [?]). This is a little counterintuitive, since an ordinary circle is homogeneous and an ordinary arc is not; the situation is reversed in the pseudo case. The homogeneity/nonhomogeneity of these spaces is a very subtle matter. There is another similar continuum in the plane, called a circle of pseudo-arcs [BJ59], which separates the plane like a circle and is homogeneous (but not indecomposable). There are in fact $2^{\aleph_0}$ topologically distinct hereditarily indecomposable continua in the plane [Bin51a] (there are also hereditarily indecomposable continua of every dimension; cf. [Bin51b], [Kel42]).

There are bizarre plane continua which look nothing like a pseudo-arc: an example is given in [AC59], also constructed as an intersection (or inverse limit), with the property that no two proper subcontinua (containing more than one point) are homeomorphic. Such a continuum can contain neither an arc nor a pseudo-arc (and hence cannot contain any of the other examples of this section), has empty interior and is topologically one-dimensional, and is highly (“totally”) nonhomogeneous. This space is not indecomposable; in fact, it contains no indecomposable subcontinuum of more than one point.

The Lakes of Wada

XI.18.4.32. This subset of $\mathbb{R}^2$, first described by K. Yoneyama [] and credited by him to T. Wada, is constructed as an intersection. The pun name “lakes of Wada” may have been first used in [HY88]. A somewhat similar example was previously described by Brouwer [].

A tropical island $X_0$ named Wada is surrounded by blue ocean. There are two lakes on the island, one with green water and one with brown, with no islands (i.e. the lakes are homeomorphic to open disks). At the first step, canals are dug from the ocean so that every spot of land is within 1 km of blue water. Call the remaining land $X_1$. In the second step, canals are dug from the green lake to within 1/2 km of every spot of land, so that the green lake plus canals is still homeomorphic to an (open) disk. Call the remaining land $X_2$. In the third step, canals are dug from the brown lake to within 1/4 km of every spot of land. In the fourth step, the ocean is used again, and then the green lake, then the brown, repeated cyclically. Call the remaining land $X$ after infinitely many steps, i.e. $X = \cap_{n=1}^{\infty} X_n$.

The construction can (in principle) clearly be continued indefinitely, since each $X_n$ is homeomorphic to a large closed disk with two small open disks removed from its interior (with the boundaries of the three disks pairwise disjoint). At the end, each lake is still homeomorphic to an open disk, and the ocean to the complement in $\mathbb{R}^2$ of a closed disk (if the construction is done on a 2-sphere such as the surface of the earth instead of $\mathbb{R}^2$, the ocean is also an open disk). However, the boundary of each of these three disjoint connected open subsets of $\mathbb{R}^2$ is the set $X$: every point of $X$ is arbitrarily close to blue, green, and brown water. The set $X$ is compact and connected, and has empty interior (it is topologically one-dimensional, and topologically similar to a pseudo-arc). It is unclear to what extent the homeomorphism type of $X$ depends on the details of the construction of the canals.
Instead of two lakes, we can start with $n$ lakes and end up with $n + 1$ disjoint connected open sets in $\mathbb{R}^2$ whose common boundary is the same connected compact set $X$. With care, the construction can even be done with countably many initial lakes. Thus, if $n$ is any natural number or $\aleph_0$, there are $n$ disjoint open subsets of the 2-sphere, each homeomorphic to an open disk, with dense union, which share a common boundary.

Basins of attraction for nonlinear systems frequently resemble lakes of Wada. These basins of attraction are called Wada basins.

The lakes of Wada show again (cf. XI.18.4.5.) that care must be exercised in formulating the Four-Color Theorem.

**Nontrivial Embeddings**

There are classes of examples of subsets of $\mathbb{R}^3$ (or $\mathbb{R}^n$) in which the subset itself is a familiar or nicely behaved one, but is embedded in Euclidean space in an unusual or strange way. There are two general subtypes: “tame” embeddings which are locally nice (e.g. smooth embeddings of manifolds), and “wild” embeddings which are not locally nice.

**Knots**

This is the simplest instance of interesting tame embeddings, consisting of topological circles embedded in $\mathbb{R}^3$ in a nontrivial way.
XI.18.4.33. Definition. A knot is a smooth embedding of $S^1$ into $\mathbb{R}^3$.

Knot theory is a fascinating and active branch of modern topology; see Figure XI.48, from http://www.math.unl.edu/~mbrittenham2/ldt/table9.gif, for some examples of knots (some of the pictures are of links, made up of more than one circle), and Figure XI.49 for a nice depiction of the simplest nontrivial knot, the trefoil knot (in a mathematical knot, the blue tube has no thickness). There are many other knot images and animations on the internet. See e.g. [] for more information on knot theory.

Figure XI.48: Some Knots and Links
Figure XI.49: A Trefoil Knot. Artwork by Jim Belk
Higher-Dimensional Knots

XI.18.4.34. One might also wonder about subsets of $\mathbb{R}^n$ (or $S^n$) homeomorphic to $S^m$ for $1 < m < n - 1$. Such subsets always have connected complements, but the ways they can be embedded (even leaving out “wild” embeddings, e.g. restricting to smooth or piecewise-linear embeddings) are very complicated.

Wild Arcs

XI.18.4.35. An arc in a topological space $X$ is a copy of $[0,1]$ in $X$ (a subspace of $X$ homeomorphic to $[0,1]$). Even an arc in $\mathbb{R}^3$ can be wildly embedded. The simplest and best-known example is due to Fox and Artin [FA48] (Figure XI.50, from http://www.sciencedirect.com/science/article/pii/S0960077901002429, originally from [HY88]):

![Image of the Fox-Artin Wild Arc](http://www.sciencedirect.com/science/article/pii/S0960077901002429)

Figure XI.50: The Fox-Artin Wild Arc

This embedding is wild at the two endpoints. The complement of the arc is not simply connected.

XI.18.4.36. By successively replacing small subarcs of this arc with a small copy of the whole wild arc, a sequence of embeddings of $[0,1]$ into $\mathbb{R}^3$ can be obtained which are wild at many points. If done carefully, this sequence of embeddings can be made to converge uniformly on $[0,1]$ to an embedding (arc) which is wild at every point.

See [Bor47] for a different kind of pathological arc in $\mathbb{R}^3$.

Alexander’s Horned Sphere

XI.18.4.37. The Alexander horned sphere is a subset of $\mathbb{R}^3$ homeomorphic to $S^2$ whose exterior is not homeomorphic to the exterior of the unit sphere in $\mathbb{R}^3$ (it is not simply connected []). The union of the horned sphere and its interior is homeomorphic to $B^3$, and the interior is homeomorphic to the open unit ball. By extending the horns into the ball instead of outside, a homeomorph of $S^2$ in $\mathbb{R}^3$ can be made whose interior is not topologically a ball, and by making horns both inside and outside both the interior and exterior can be made pathological.

There are many pictures and animations of the horned sphere on the internet (Figure XI.52 is from http://conan777.files.wordpress.com/2010/04/hornedsphere3.jpg), although I still think the best picture is the classic one in [HY88] (Figure XI.51, from http://www.sciencedirect.com/science/article/pii/S0960077901002429). For a beautiful animation of the Alexander Horned Sphere, see https://www.youtube.com/watch?v=Pe2mnrLUYFU.
XI.18.4.38. The Alexander Horned Sphere is wild at a Cantor set of points. By successively adding small pairs of horns densely, a sequence of embeddings of $S^2$ into $\mathbb{R}^3$ can be obtained which converges uniformly to an embedding which is wild at every point (cf. XI.18.4.36.). The horns can be added outside, inside, or both inside and outside.

XI.18.4.39. There are other rather different wild embeddings of $S^2$ into $\mathbb{R}^3$. A simple one is to fatten up the Fox-Artin wild arc (Figure XI.50) into a tube tapering to a point at the ends. A variation, also discussed in [FA48], is shown in Figure XI.53, from [GH78] (the knots repeat and approach $p,q,r,s$). See also [Rus73]. Actually, the first wild sphere was obtained by L. ANTOINE (who was blind!) [?], who also constructed Antoine’s necklace (XI.18.4.40.).
Figure XI.52: Construction of an Alexander horned sphere
Figure XI.53: A Few Steps in the Construction of a Wild Sphere

1247
Antoine’s Necklace

XI.18.4.40. This is another subset of $\mathbb{R}^3$ defined by an intersection, a Cantor set wildly embedded in $\mathbb{R}^3$. Begin with a solid torus $X_1$ in $\mathbb{R}^3$. Inside take a chain $X_2$ of smaller interlocked solid tori as in Figure XI.54 (the number of interlocked tori is unimportant as long as it is at least three, although the final embedding of the Cantor set depends on the number of tori at each stage [She68]):

![Figure XI.54: First step in the construction of Antoine's Necklace](image)

Inside each of these smaller solid tori put a similar chain of even smaller interlocked solid tori, and iterate in such a way as to make the diameters of the solid tori at the $n$’th stage $X_n$ go to zero as $n \to \infty$. The intersection $X$ is *Antoine’s Necklace*.

![Figure XI.55: Two steps in the construction of Antoine’s Necklace](image)

Figure XI.56 shows three stages of the construction.

The connected components of $X$ are contained in the connected components of $X_n$ for each $n$; since the diameters of these connected components approach zero as $n \to \infty$, the connected components of $X$ have diameter zero, i.e. are single points. Thus $X$ is totally disconnected. However, it is clear that $X$ can have no isolated points, since every point of $X$ is arbitrarily close to points from different components of $X_n$ for large enough $n$. Thus $X$ is homeomorphic to the Cantor set $K$.

However, the complement of $X$ in $\mathbb{R}^3$ is not simply connected. If $\gamma$ is a loop around the “hole” in $X_1$, then $\gamma$ cannot be shrunk to a point in the complement of $X_n$ for any $n$ (this seems “obvious”, but some
Figure XI.56:  Three steps in the construction of Antoine’s Necklace

nontrivial topology is needed to prove it). If it could be shrunk to a point in the complement of \(X\), the homotopy would trace out a continuous image of a disk, a compact set \(Y\) in \(\mathbb{R}^3\). This set would be disjoint from \(X_n\) for sufficiently large \(n\) by a simple compactness argument (the complements of the \(X_n\) give an open cover of \(Y\)). Thus \(\gamma\) could be shrunk to a point in the complement of some \(X_n\), a contradiction.

So \(X\) is embedded in \(\mathbb{R}^3\) in a very different way than the ordinary Cantor set, regarded as a subset of the \(x\)-axis in \(\mathbb{R}^3\). The construction and its variations are described in [BM88], where the set is described as a necklace that “cannot fall apart” despite having no string.

It also follows from results in [BM88] that any uncountable compact space which can be embedded in a surface in \(\mathbb{R}^3\) (e.g. any uncountable compact subset of \(\mathbb{R}^2\)) can be embedded in \(\mathbb{R}^3\) in uncountably many essentially different ways.

It is known that any subset \(Z\) of \(\mathbb{R}^2\) homeomorphic to the ordinary Cantor set \(K\) (sitting in the \(x\)-axis) is embedded in the same way as \(K\): there is a homeomorphism of \(\mathbb{R}^2\) onto itself carrying \(Z\) onto \(K\) [CV98].

See [Rus92] for some generalizations of Antoine’s necklace and wild embeddings of spheres in \(\mathbb{R}^n\). See also Exercise XI.18.5.4..
**Other Examples**

**XI.18.4.41.** These examples only scratch the surface of the possibilities. Many other examples can be found in books like [SS95], [Bor67], [Edg08], and [Fal14].

When noncompact subsets of Euclidean space are considered, the possibilities become far greater and more bizarre. For example:

**XI.18.4.42.** Proposition. [Kur66, p. 263] Up to homeomorphism, there are $2^{2^\aleph_0}$ distinct subsets of $\mathbb{R}$, or of any separable metric space of cardinality $2^{\aleph_0}$.

**Proof:** If $A$ is a subset of $\mathbb{R}$, then $A$ is separable. There are only $2^{\aleph_0}$ continuous functions from any separable space to $\mathbb{R}$, since such a function is completely determined by its values on a countable dense set. Thus there can only be $2^{\aleph_0}$ subsets of $\mathbb{R}$ homeomorphic to $A$. Since there are $2^{2^\aleph_0}$ subsets of $\mathbb{R}$, there must be $2^{2^\aleph_0}$ homeomorphism classes.

Since there are only $2^{\aleph_0}$ Borel sets in $\mathbb{R}$, “most” of these sets are not even homeomorphic to Borel sets. (Actually, any subset of $\mathbb{R}$ homeomorphic to a Borel set is a Borel set.)

**XI.18.5. Exercises**

**XI.18.5.1.** Let $k$ and $X_k$ be as in XI.18.4.1. Suppose $k'$ is another sequence of natural numbers and $\phi : X_k \to X_{k'}$ is a homeomorphism.

(a) For various points $p$ in $X_k$, consider the number and homeomorphism type of connected components of $X_k \setminus \{p\}$. Observe that this data must be the same for $X_{k'} \setminus \{\phi(p)\}$.

(b) Use (a) to show that necessarily $\phi \left( \left( \frac{1}{2^n}, 0 \right) \right) = \left( \frac{1}{2^n}, 0 \right)$ for all $n$.

(c) Conclude that $k = k'$.

(d) If the sequences $k$ and $k'$ are strictly increasing and distinct, there is a simpler proof that $X_k$ and $X_{k'}$ are not homeomorphic, since for each $n \geq 2$ there is a unique point $p \in X_k$ such that $X_k \setminus \{p\}$ has exactly $k_n + 2$ components. Show that there are $2^{\aleph_0}$ such sequences (they are in one-one correspondence with the infinite subsets of $\mathbb{N}$).

**XI.18.5.2.** This problem is an elaboration of XI.18.4.2. Recall the notation of (). For a compact space $X$, let $X_\infty$ be the stable perfect subset ($K$ in the case $X = X_m$ for some $m$), $S^{(k)}(X)$ the set of isolated points of the $k$’th derived set $X^{(k)}$, and

$$B^{(k)}(X) = \left[ \overline{S^{(k)}(X)} \setminus S^{(k)}(X) \right] \cap X_\infty .$$

A homeomorphism from $X$ to a space $Y$ must send $X_\infty$ to $Y_\infty$ and $B^{(k)}(X)$ to $B^{(k)}(Y)$ for each $k$. Show that $B^{(k)}(X_m) \neq B^{(k+1)}(X_m)$ if and only if $k = m_n$ for some $n$. Conclude that the $X_m$ are mutually nonhomeomorphic for distinct $m$.  

1250
XI.18.5.3. (a) Let $X$ be a compact connected subset of $\mathbb{R}^2$. Show that if $\mathbb{R}^2 \setminus X$ is connected, it is homeomorphic to $\mathbb{R}^2 \setminus \{0\}$. [Use XI.23.10.1., and show the boundary there can be taken to be a single curve.] Contrast with XI.19.5.13..

(b) Show more generally that if $X$ and $Y$ are compact subsets of $\mathbb{R}^2$ such that $U = \mathbb{R}^2 \setminus X$ and $V = \mathbb{R}^2 \setminus Y$ are connected, then $U$ and $V$ are homeomorphic if and only if the component spaces $X_c$ and $Y_c$ (XI.13.9.15.) are homeomorphic. This generalizes both (a) and XI.19.5.13..

(c) Let $X$ be a compact connected subset of $\mathbb{R}^2$. Show that the unbounded component of $\mathbb{R}^2 \setminus X$ is homeomorphic to $\mathbb{R}^2 \setminus \{0\}$, and that each bounded component is homeomorphic to an open disk. [For the unbounded component, apply (a) to the union of $X$ and the bounded components of $\mathbb{R}^2 \setminus X$. For bounded components, invert in a small circle within the component.] Make an analogous statement about complements of compact connected subsets of $S^2$.

These results are special to $\mathbb{R}^2$ (analogous results hold more simply in $\mathbb{R}^1$): several examples from this section show that nothing like this holds in $\mathbb{R}^3$.

XI.18.5.4. Modify the construction of Antoine’s Necklace as follows.

(a) Consider the following linked set of solid tori (picture from [Gar01]; the surfaces should not touch):

![Figure XI.57: Variation of Borromean Rings](figure.png)

The number $n$ of solid tori is not important as long as it is at least 3 (although, as with Antoine’s Necklace, the number probably affects the final embedding); the case $n = 3$ is the usual Borromean Rings (Figure XI.48 $6_2^2$). Show that the entire set of tori is linked nontrivially, that any two of the solid tori are not linked, and that if any one of the tori is removed the remaining ones are unlinked.

(b) Replace each of these solid tori by a set of smaller solid tori contained in it, linked in the same way. Iterate the construction so that the diameters of the solid tori go to zero, and take the intersection $X$.

(c) Show that $X$ is homeomorphic to the Cantor set.
(d) Show that the complement of $X$ is not simply connected; thus $X$ is nontrivially (wildly) embedded in $\mathbb{R}^3$.

(e) If $T$ is one of the solid tori in the construction, then $X \cap T$ is a clopen set in $X$. Show that every clopen set in $X$ is a finite disjoint union of such sets.

(f) Two disjoint compact subsets $A$ and $B$ of $\mathbb{R}^n$ are unlinked (or separated by spheres) if there is a homeomorphism of $\mathbb{R}^n$ sending $A$ and $B$ into disjoint open balls. Show that if $A$ and $B$ are disjoint clopen sets in $X$, then there is a partition of $A$ into clopen sets $C$ and $D$ such that $\{B, C, D\}$ are pairwise unlinked in $\mathbb{R}^3$. Conclude that $X$ is not embedded in $\mathbb{R}^3$ in the same way as Antoine’s Necklace (any version).

(g) What happens if we use other linkings of the embedded solid tori at each stage, and perhaps vary the linkings from stage to stage? We could, for example, at infinitely many stages knot the solid tori, or replace each solid torus with a thinner solid torus wrapping twice around inside it as in the construction of the solenoid (Figure XI.41). See [She68], [Sko86], and [SS13], and the references therein.

**XI.18.5.5. Knotted Solenoids.** Let $S$ be a solenoid (XI.18.4.13).

(a) By construction, $S$ is a subset of a solid torus $T$. If $K$ is a knot, fatten up $K$ into a solid torus $T_K$ and let $\phi$ be a homeomorphism of $T$ onto $T_K$. Let $S_K = \phi(S)$. $S_K$ can be regarded as a knotted version of $S$. Show that if $K_1$ and $K_2$ are knots, show that $S_{K_1}$ and $S_{K_2}$ are embedded in $\mathbb{R}^3$ in the same way if and only if $K_1 \cong K_2$.

(b) If $K$ is a knot, then $K$ can be embedded nontrivially in a solid torus with winding number greater than 1 (write the knot as the closure of a braid; cf. (6)). By fattening up $K$ inside the solid torus, we obtain a knotted solid torus. Another knot can be embedded in this in the same way. Iterating this construction and taking an intersection, an “infinitely knotted” solenoid is obtained. Classify the solenoid obtained via the winding numbers and show that any solenoid can be embedded in $\mathbb{R}^3$ in 280 essentially different ways.

**XI.18.5.6.**

(a) Let $S$ be a solenoid embedded in $\mathbb{R}^3$ in the standard (unknotted) way, with $G = \pi^1(S)$ the corresponding subgroup of $\mathbb{Q}$. Show that the fundamental group of $\mathbb{R}^3 \setminus S$ is $G$. Note that $\mathbb{R}^3 \setminus S$ is a connected open subset of $\mathbb{R}^3$, hence a connected 3-manifold.

(b) Compute the fundamental groups of the complements of the knotted solenoids of XI.18.5.5., which are also connected 3-manifolds.

**XI.18.5.7.** Let $G$ be any countable group. Show that there is a connected open subset of $\mathbb{R}^5$ (hence a connected 5-manifold) with fundamental group $G$. [See http://mathoverflow.net/questions/192230/is-there-a-manifold-with-fundamental-group-mathbbq for an outline of the construction.] Conversely, the fundamental group of any manifold is countable (6).

**XI.18.5.8. The Whitehead Continuum and the Whitehead Manifold.** [6] Here is an interesting construction reminiscent of the construction of solenoids. Let $T_0$ be a solid (unknotted) torus in $\mathbb{R}^3$, and $T_1$ a thinner solid torus embedded in $T_0$ as in Figure XI.58.

For each $k$ let $T_{k+1}$ be a thin solid torus embedded in $T_k$ in the same way. Note that each $T_k$ is unknotted in $\mathbb{R}^3$.

Let $W = \cap_{k=1}^\infty T_k$. Since each $T_k$ is compact and connected, $W$ is also compact and connected. $W$ is called the Whitehead continuum (the exact space depends on the details of the construction).

(a) If $S$ is a meridian circle for $T_0$, show that
(i) $S$ and the longitude axis circle of $T_1$ form a nontrivial link in $\mathbb{R}^3$, called the Whitehead link (XI.59).

(ii) $S$ is not null-homotopic in $T_0 \setminus T_1$.

(iii) $S$ is null-homotopic in $\mathbb{R}^3 \setminus T_1$. [This can be seen directly, but the easiest approach is to note that there is a homeomorphism of $\mathbb{R}^3$ interchanging the circles.]

(b) Discuss other properties of $W$. Is it path-connected? Simply connected?

(c) The complement of $W$ is more interesting. Regard the construction as being done in the one-point compactification $S^3$ of $\mathbb{R}^3$, and let $E = S^3 \setminus W$. The complement of a solid torus in $S^3$ is another solid torus; thus $E$ is an increasing union of solid tori. Show that each is embedded in the next as $T_1$ is embedded in $T_0$. Show that $E$ is a connected, noncompact 3-manifold, and that $E$ is simply connected [use (a)(iii).] Conclude from () that $E$ is contractible (or give a direct argument). $E$ is called the Whitehead manifold. (Unlike for $W$, up to homeomorphism $E$ does not depend on the details of the construction.)

1253
(d) Show that no neighborhood of \( \infty \) in the one-point compactification \( E^\dagger \) of \( E \) is simply connected [\( E^\dagger \) is homeomorphic to the quotient \( S^3/W \) of \( S^3 \) obtained by collapsing \( W \) to a point. Use (a)(ii).] Thus \( E^\dagger \) is not a manifold. Conclude that \( E \) is not homeomorphic to \( \mathbb{R}^3 \).

(e) Show that \( E \times \mathbb{R} \) is homeomorphic to \( \mathbb{R}^4 \), as is \( (\mathbb{R}^3/W) \times \mathbb{R} \). (However, contrary to statements in some references like [CWY10], \( (S^3/W) \times \mathbb{R} = E^\dagger \times \mathbb{R} \) is not homeomorphic to \( \mathbb{R}^4 \) since it has two ends and is not contractible.)

(f) If \( p \) is a point of \( W \), then \( E \subseteq S^3 \setminus \{p\} \cong \mathbb{R}^3 \), and thus \( E \) is homeomorphic to an open subset of \( \mathbb{R}^3 \).

(g) Some more properties of \( E \): \( E \) is the union of two open subsets, each homeomorphic to \( \mathbb{R}^3 \), whose intersection is also homeomorphic to \( \mathbb{R}^3 \) [Gab11]; \( E \) does not have any orientation-reversing diffeomorphism [AW07]; \( E \) does not admit a complete metric with strictly positive scalar curvature [CWY10], unlike \( \mathbb{R}^3 \), which does have such a metric, e.g. by identifying it with

\[ \{(x, y, z, w) : w = x^2 + y^2 + z^2 \} \subseteq \mathbb{R}^4. \]

(h) Let \( n = (n_1, n_2, \ldots) \) be a sequence in \( \mathbb{N} \). Vary the construction of \( W \) and \( E \) by wrapping \( T_k \) around \( T_{k-1} \) \( n_k \) times before linking as in Figure XI.58. Let \( W_n \) be the intersection and \( E_n = S^3 \setminus W_n \). (Thus the Whitehead continuum is \( W_n \) where \( n_k = 1 \) for all \( k \).) Show that \( W_n \) and \( E_n \) have the same properties as before. Show that \( 2^{n_0} \) distinct homeomorphism classes of \( E_n \) are obtained. Thus there are \( 2^{n_0} \) topologically distinct contractible open subsets of \( \mathbb{R}^3 \). See [McM62] for details.

In contrast, every simply connected open subset (in particular, every contractible open subset) of \( \mathbb{R}^2 \) is homeomorphic to \( \mathbb{R}^2 \). In fact, the only simply connected 2-manifolds are \( \mathbb{R}^2 \) and \( S^2 \).

Not every contractible 3-manifold is of the form of (h); cf. [McM62], [TW97]. There does not appear to be any reasonable classification of contractible 3-manifolds. The already known existence of exotic contractible 3-manifolds seems to have been overlooked in [BT82, p. 147].

XI.18.5.9. The House with Two Rooms. Let \( X \) be the subset of \( \mathbb{R}^3 \) pictured in Figure XI.60 (from [Hat02]).

\[ \text{Figure XI.60: The House With Two Rooms} \]

(a) \( X \) is a finite union of planar rectangles, and thus is a two-dimensional (compact) polyhedron in \( \mathbb{R}^3 \).

(b) Let \( Y \) be the rectangular solid in \( \mathbb{R}^3 \) generated by \( X \) (convex hull of \( X \)). Show that there is a deformation retraction of \( Y \) onto \( X \). [Push down the upper hole and expand the indentation to fill the bottom room, and similarly push up the lower hole.] Conclude \( X \) is an absolute retract, and in particular is contractible.
XI.18.5.10. **The Rational Ruler.** Construct a subset $X$ of the closed unit square $[0,1]^2 \subseteq \mathbb{R}^2$ as follows. Begin with the horizontal segment $[0,1]$ on the $x$-axis, and for each $r \in \mathbb{Q} \cap [0,1]$, $r = \frac{p}{q}$ in lowest terms, add the closed vertical segment between $(r,0)$ and $(r,\frac{1}{n})$.

(a) Show that $X$ is a compact subset of $\mathbb{R}^2$.

(b) Construct a retraction of $[0,1]^2$ onto $X$ as follows. Set $X_n = X \cup \left\{ (x,y) : 0 \leq x \leq 1, 0 \leq y \leq \frac{1}{n} \right\}$.

(note that $X = \cap_n X_n$) and construct a uniformly convergent sequence of retractions from $[0,1]^2$ onto $X_n$. Conclude () that $X$ is an absolute retract.

XI.18.5.11. ([Lub53]; cf. [Bor67, p. 150-151]) Let $n \in \mathbb{N}$. Adapt the Lakes of Wada construction to construct in $\mathbb{R}^3$ a connected two-dimensional absolute neighborhood retract $X_n$ such that the complement of $X$ in $\mathbb{R}^3$ has exactly $n$ components, the boundary of each of which is $X$. ($X$ thus has a locally nice structure, a least in a homotopy sense, unlike the examples in $\mathbb{R}^2$ like the Lakes of Wada which necessarily are locally quite pathological [?].)

XI.18.5.12. Show that there is no reasonable analog of the four-color theorem in $\mathbb{R}^3$, even for polyhedra. [Lay $n$ boards with rectangular cross-section side-by-side on the $xy$-plane in $\mathbb{R}^3$ with long dimension parallel to the $x$-axis, and on top of them lay $n$ boards with long dimension parallel to the $y$-axis. Glue the $k$'th horizontal board to the $k$'th vertical board to make a single cross-shaped polyhedron for each $k$. See http://demonstrations.wolfram.com/AnInfinitelyColorableSetOf3DRegions/.] Another more informal argument I heard somewhere: “Imagine a bowl of spaghetti where every noodle touches all the others.”] There are meaningful higher-dimensional analogs of the graph-theoretic versions of the four-color theorem, however.

XI.18.5.13. The Hahn-Mazurkiewicz Theorem says that a Hausdorff topological space is a continuous image of $[0,1]$ if and only if it is compact, metrizable, connected, and locally connected (and thus automatically path-connected and locally path-connected). Thus the Sierpiński gasket, Sierpiński carpet, and Menger sponge are all continuous images of $[0,1]$. Find explicit continuous surjections. [The construction of a space-filling curve () may provide some insight.]

XI.18.5.14. **The Infinite Comb.** Consider the compact subset of $\mathbb{R}^2$ consisting of the line segments from 0 to 1 on the $x$ and $y$ axes and the vertical segments from 0 to 1 with $x$-coordinate $\frac{1}{n}$ for each $n$ (the exact location of the $x$-coordinates of the vertical segments are unimportant as long as they form a sequence of numbers in $(0,1]$ converging to 0). See Figure XI.61.
(a) Show that the infinite comb is contractible.

(b) Show that the infinite comb is not locally connected at points on the positive $y$-axis. Thus it is not strongly contractible to any of these points. It is strongly contractible to any other point.

(c) Figures XI.62 and XI.63 show a number of variations of the infinite comb. Which of these are contractible? Which are strongly contractible, to which points? Which pairs of these are homeomorphic? (The first space in XI.62 is called Zeno’s maze in [Pug15], but I have not found this name in other references.) Show that all have trivial fundamental group, i.e. are simply connected.

Figure XI.61: The Infinite Comb
Figure XI.62: Infinite Comb Variations I
Figure XI.63: Infinite Comb Variations II
XI.18.5.15. **The Infinite Ladder.** A variation of the infinite comb is to add the horizontal segment from 0 to 1 with $y$-coordinate 1 (Figure XI.64). This space is somewhat similar to the Hawaiian Earring, but with a crucial difference: show that the fundamental group is countable.

![The Infinite Ladder](image)
XI.19. Manifolds

There are two types of mathematical objects commonly called manifolds: topological manifolds and differentiable, or smooth, manifolds (there are also PL manifolds, which are something in between). There is also a notion of “manifold with boundary,” both topological and smooth. The term “manifold” is also used in a few other related and even entirely unrelated senses in parts of mathematics. One must be aware of context whenever the term “manifold” is used. Smooth manifolds are special types of topological manifolds with specified extra structure (smooth manifolds come in a number of flavors). Both types of manifolds occur in analysis. In this section, we describe the basics of both theories; each theory is very extensive and goes far beyond what we do here.

XI.19.1. Topological Manifolds

Topological manifolds are generalizations of curves and surfaces. The property of a curve or surface we use is that near each point, a small piece of the curve or surface looks like an open ball in Euclidean space (of dimension 1 or 2 respectively). Our generalization will allow for higher-dimensional objects also, and will be intrinsic, not requiring our space to be embedded in a Euclidean space.

XI.19.1.1. Definition. A topological space $X$ is locally Euclidean if every point $p \in X$ has an open neighborhood which is homeomorphic to an open ball in $\mathbb{R}^n$ for some $n \geq 0$ (which may vary from point to point), where $\mathbb{R}^0$ is a single point. Such a neighborhood is called a coordinate patch and a homeomorphism with an open ball in $\mathbb{R}^n$ is a coordinate chart, and the $n$ is called the local dimension of $X$ at $p$. $X$ is $n$-dimensional if it is has the same local dimension $n$ at every point.

Roughly speaking, a topological manifold will just be a topological space which is locally Euclidean. But it turns out we should make a few additional technical restrictions to rule out pathologies that can occur in bizarre examples.

XI.19.1.2. Some initial remarks about the definition:

(i) The local dimension is well defined by the Invariance of Domain Theorem ().

(ii) All points within a coordinate patch have the same local dimension; thus the local dimension is locally constant (justifying the terminology “local dimension.”) Hence the local dimension is constant on components of $X$ (). In particular, a connected locally Euclidean space has the same local dimension at every point.

(iii) An $n$-dimensional metrizable locally Euclidean topological space has topological dimension $n$ (), so the terminology “$n$-dimensional” is consistent. The topological dimension of a general metrizable locally Euclidean space is the maximum or supremum of the local dimensions at all points. (This can fail in the nonmetrizable case, cf. [Fed95].)

(iv) Since an open ball in $\mathbb{R}^n$ is first countable, even second countable, a locally Euclidean space is first countable. But it need not be second countable (XI.19.1.9., XI.19.5.5.). A locally Euclidean space is locally second countable (XI.1.3.7.).

(v) Since open balls in $\mathbb{R}^n$ are locally compact, a (Hausdorff) locally Euclidean space is locally compact () and hence completely regular ().

(vi) Since an open ball in $\mathbb{R}^n$ is path-connected, a locally Euclidean space is locally path-connected; hence the components and path components coincide and are clopen sets ().
(vii) An open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$. Thus the definition can be equivalently rephrased: $X$ is locally Euclidean if every point $p \in X$ has an open neighborhood which is homeomorphic to $\mathbb{R}^n$ for some $n$. This is the form given in many references. However, it seems preferable to write the definition as we have, since in most examples there are more natural homeomorphisms of point neighborhoods with open balls in $\mathbb{R}^n$ than with $\mathbb{R}^n$ itself. For example, it is immediate from our definition that any open set in $\mathbb{R}^n$ is locally Euclidean (in the relative topology).

(viii) Since an open ball in $\mathbb{R}^n$ is $T_1$, a locally Euclidean topological space is locally $T_1$, hence $T_1$ (XI.7.3.5.). But it need not be Hausdorff (XI.19.1.3.(v)). This is one of the pathologies we want to exclude in the definition of a topological manifold.

**XI.19.1.3. Examples.** (i) Any open set in $\mathbb{R}^n$ is an $n$-dimensional locally Euclidean space with the relative topology.

(ii) A curve in $\mathbb{R}^n$ is a 1-dimensional locally Euclidean space (with the relative topology), provided it has no endpoints and does not cross or loop back onto itself (more precisely, if it is homeomorphic to an open interval or a circle, or a separated disjoint union of such curves). A simple curve (arc) with endpoints is an example of a manifold with boundary (I). “Curves” such as a figure 8, a figure 6, or a “Warsaw circle” (XI.18.4.4.) are not locally Euclidean and are not manifolds with boundary.

(iii) A surface in $\mathbb{R}^3$ is a 2-dimensional locally Euclidean space with the same restrictions as in (ii), which are more complicated to state precisely. Examples include a sphere, torus, a cylinder without its end circles, or a Möbius strip without its edge. A cylinder with end circles included, or a Möbius strip with its edge included, is a manifold with boundary.

(iv) The unit sphere in $\mathbb{R}^{n+1}$ is an $n$-dimensional locally Euclidean space for any $n$. The closed unit ball is not locally Euclidean, but is a manifold with boundary.

(v) A locally Euclidean space need not be Hausdorff. Here is the standard example. Let $X$ be $\mathbb{R} \cup \{0\}$, with the topology consisting of all open subsets of $\mathbb{R}$ and all sets consisting of $0'$ along with an open neighborhood or deleted open neighborhood of 0. Since $\mathbb{R}$ is an open subset of $X$, every point of $\mathbb{R}$ has a locally Euclidean neighborhood; and $0'$ does also, e.g. $\{0\} \cup (\mathbb{R} \setminus \{0\})$. Thus $X$ is locally Euclidean (and 1-dimensional). But 0 and $0'$ do not have disjoint neighborhoods. See Exercise XI.19.5.14. for a more interesting example.

There are far more complicated examples; see Exercises XI.19.5.5. and XI.19.5.8.

**XI.19.1.4. Theorem.** Let $M$ be a topological space which is Hausdorff, locally compact, and locally second countable, in which every component is open (e.g. Hausdorff and locally Euclidean). The following are equivalent:

(i) $M$ is paracompact.

(ii) $M$ is metrizable.

(iii) Every component of $M$ is Lindelöf.

(iv) Every component of $M$ is $\sigma$-compact.

(v) Every component of $M$ is second countable.

(vi) Every component of $M$ is a Polish space.
The following are equivalent, and imply (i)–(vi):

(vii) $M$ is Lindelöf.

(viii) $M$ is $\sigma$-compact.

(ix) $M$ is second countable.

(x) $M$ is a Polish space.

If $M$ satisfies (i)–(vi), then $M$ satisfies (vii)–(x) if and only if $M$ has only countably many components. In particular, if $M$ is connected, then (i)–(x) are equivalent.

**Proof:** Since the components of $M$ are clopen (), obviously (vii)–(x) each imply that $M$ has only countably components; and if $M$ has only countably components, then (vii) $\iff$ (iii), (viii) $\iff$ (iv), (ix) $\iff$ (v), and (x) $\iff$ (vi). So we need only prove that (i)–(vi) are equivalent.

Since $M$ is locally compact, (iii) $\iff$ (iv) by XI.11.8.11., and by XI.11.8.13., (v) and (vi) are equivalent and imply (iv) and that every component of $M$ is metrizable. Since the components of $M$ are clopen, $M$ is metrizable if and only if each component is metrizable (), so (v) $\Rightarrow$ (ii). (ii) $\Rightarrow$ (i) by XI.11.10.4.. $M$ is paracompact if and only if each component of $M$ is paracompact (XI.11.10.8.), and if a component is paracompact it is $\sigma$-compact by XI.11.10.18.. Thus (i) $\Rightarrow$ (iv). And since $M$ is locally second countable, (iii) $\Rightarrow$ (v) by XI.1.3.8..

**XI.19.1.5.** There are connected Hausdorff locally Euclidean topological spaces not satisfying the conditions of Theorem XI.19.1.4.. Such a space can be normal (Exercise XI.19.5.5.) or nonnormal (Exercises XI.19.5.8.–XI.19.5.9.). Separability is also not equivalent to the conditions of the theorem; such examples can be separable (Exercise XI.19.5.9.) or nonseparable (Exercises XI.19.5.5. and XI.19.5.8.). Conditions (vii)–(x) imply separability, and are equivalent to separability in the presence of (i)-(vi).

**XI.19.1.6.** There is no universal consensus on the proper definition of a topological manifold. All of the following are reasonable candidates which have in fact all appeared in various references. They are arranged roughly, but not exactly, in increasingly restrictive order. Throughout, $X$ denotes a topological space.

(i) $X$ is locally Euclidean.

(ii) $X$ is Hausdorff and locally Euclidean.

(iii) $X$ is Hausdorff, locally Euclidean, and satisfies conditions (i)–(vi) of XI.19.1.4..

(iv) $X$ is Hausdorff, locally Euclidean, and satisfies conditions (i)–(x) of XI.19.1.4..

(i)$'$ $X$ is locally Euclidean and has constant local dimension.

(ii)$'$ $X$ is Hausdorff, locally Euclidean, and has constant local dimension.

(iii)$'$ $X$ is Hausdorff, locally Euclidean, satisfies conditions (i)–(vi) of XI.19.1.4., and has constant local dimension.

1262
(iv)' $X$ is Hausdorff, locally Euclidean, satisfies conditions (i)–(x) of XI.19.1.4., and has constant local dimension.

(i)" $X$ is connected and locally Euclidean.

(ii)" $X$ is Hausdorff, connected, and locally Euclidean.

(iii)" $X$ is Hausdorff, connected, locally Euclidean, and satisfies conditions (i)–(vi) of XI.19.1.4..

(iv)" $X$ is Hausdorff, connected, locally Euclidean, and satisfies conditions (i)–(x) of XI.19.1.4..

(Note that (iii)" and (iv)" are identical.)

We will take (iii) as our definition, which is probably the most common choice. Reasonable arguments can be made for all of the versions (although the arguments for (i)–(i)" are weak; manifolds should at least be Hausdorff). Almost all spaces arising in practice which one would want to call manifolds satisfy (iv)" or at least (iv)'. However, almost all properties of such spaces hold equally well for ones satisfying (iii), with identical proofs; practically all interesting things about manifolds happen separately in each component, so it doesn’t really matter how many components there are as long as each individual component is well behaved. Definition (iii) can also be given very cleanly:

XI.19.1.7. Definition. A topological space $X$ is a (topological) manifold if it is metrizable and locally Euclidean.

A space satisfying (iv)" is then just a connected manifold, one satisfying (iv) is a separable manifold, and one satisfying (iii)' is an $n$-dimensional manifold (usually just called an $n$-manifold) for some $n$. It seems rather artificial to allow manifolds with nonconstant local dimension, and they rarely arise in practice; we can restrict to the constant local dimension case by just using the term “$n$-manifold”.

XI.19.1.8. A compact Hausdorff locally Euclidean space automatically satisfies (viii), hence (i)–(x), and in particular is a manifold.

XI.19.1.9. The case $n = 0$ is allowed in the definition of a manifold. A 0-manifold is just a set with the discrete topology.

XI.19.2. Topological Manifolds With Boundary

We would like to extend the notion of manifold to include such examples as curves with endpoints, surfaces with boundary curves, and closed balls in $\mathbb{R}^n$. These will be examples of “manifolds with boundary.”

The basic tool is the notion of a half-space:

XI.19.2.1. Definition. The $n$-dimensional half-space ($n \geq 1$) is the subset

$$\mathbb{H}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_n \geq 0\}$$

with the relative topology from $\mathbb{R}^n$. Its boundary $\partial \mathbb{H}^n$ is the set of $n$-tuples with last coordinate zero. We also define $\mathbb{H}^0$ to be a single point, with $\partial \mathbb{H}^0 = \emptyset$.
XI.19.2.2. Definition. A *topological* manifold with boundary is a metrizable space $X$ in which every point has an open neighborhood homeomorphic to an open set in $\mathbb{H}^n$ for some $n \geq 0$ (which may vary from point to point.) Such an open set is called a *coordinate patch*, and the homeomorphism is called a *coordinate chart*. If the $n$ is constant, $X$ is called an $n$-manifold with boundary. A point of $X$ with a neighborhood homeomorphic to $\mathbb{R}^n$ (or to an open set in $\mathbb{R}^n$) is called an *interior point* of $X$, and a point which maps to a point in $\partial \mathbb{H}^n$ under a coordinate chart is called a *boundary point* of $X$. Denote by $\partial X$ the set of boundary points of $X$, called the *boundary* of $X$.

Note that the use of “interior point” and “boundary point” here is not the same as the general use in topology: such points may not be interior or boundary points respectively in the topological sense. Indeed, since a manifold with boundary is not naturally embedded as a subset of a larger topological space, topological interior and boundary do not really even make sense.

XI.19.2.3. Just as with manifolds, manifolds with boundary are locally compact and locally path-connected, and the components and path components coincide and are clopen subsets. Each component is second countable and $\sigma$-compact, and a Polish space.

XI.19.2.4. Examples. (i) $\mathbb{H}^n$ is an $n$-manifold with boundary. The boundary points are the points of $\partial \mathbb{H}^n$.

(ii) Any (topological) manifold $X$ is a manifold with boundary, with $\partial X = \emptyset$. (So the term “manifold with boundary” does not imply that the boundary is nonempty.)

(iii) Since $\mathbb{H}^1$ is a ray, a closed or half-open interval $I$ in $\mathbb{R}$ is a 1-manifold with boundary $\partial I$ consisting of the endpoint(s). In fact, a complete list (up to homeomorphism) of connected 1-manifolds with boundary is: an open interval, a half-open interval, a closed interval, and a circle.

(iv) A closed disk $D$ in $\mathbb{R}^2$ is a 2-manifold with boundary, with $\partial D$ the boundary circle. More generally, a closed ball in $\mathbb{R}^n$ is an $n$-manifold with boundary.

(v) A Möbius strip $M$ is a 2-manifold with boundary, with $\partial M$ a circle.

(vi) A figure 8 or figure 6 is not a manifold with boundary, since in each case there is one point which does not have any neighborhood homeomorphic to an open set in $\mathbb{H}^1$ (or $\mathbb{H}^n$ for any $n$).

(vii) An open disk with a relatively open subset of the bounding circle added (i.e. a relatively open subset of a closed disk) is a manifold with boundary; more generally, an open subset of any manifold with boundary is a manifold with boundary. But an open disk with a subset of the bounding circle added which is not relatively open is not a manifold with boundary, although this is not entirely obvious.

The next result may seem obvious, but a rigorous proof is rather involved on one point.

XI.19.2.5. Theorem. Let $X$ be an $n$-manifold with boundary. Then

(i) The interior and the boundary of $X$ are disjoint. $\partial X$ is closed in $X$, and the set of interior points is dense and open.

(ii) The interior of $X$ is an $n$-manifold, and $\partial X$ (if nonempty) is an $(n-1)$-manifold.

It is obvious that the set of interior points is dense and open, and that the interior is an $n$-manifold. (Under any coordinate chart, points going to a point of $\mathbb{H}^n \setminus \partial \mathbb{H}^n$ are interior points.) The issue in proving
that the interior and boundary are disjoint is to show that a point of \( \partial \mathbb{H}^n \) does not have an open neighborhood in \( \mathbb{H}^n \) homeomorphic to an open set in \( \mathbb{R}^n \) (or in \( \mathbb{R}^m \) for any \( m \)). This is nontrivial to prove directly, but is an immediate consequence of the Invariance of Domain Theorem. The simplest direct proof, which uses some algebraic topology, is that a point of \( \partial \mathbb{H}^n \) has a deleted neighborhood in \( \mathbb{H}^n \) which is contractible, while no point of \( \mathbb{R}^n \) has a contractible deleted neighborhood. It follows from the disjointness that \( \partial X \) is closed. Showing that \( \partial X \) is an \((n-1)\)-manifold is straightforward, since the points of \( \partial X \) in a coordinate patch around a point of \( \partial X \) form a set homeomorphic to a relatively open set in \( \partial \mathbb{H}^n \), which is homeomorphic to \( \mathbb{R}^{n-1} \).

**XI.19.2.6. Convention:** In geometry/topology, the term “manifold” is often used to mean “manifold with boundary” (with, of course, the possibility that the boundary is empty). In particular, the term “compact manifold” normally means “manifold with boundary which is compact” (i.e. the whole space is compact, not just the boundary). To refer to a manifold in the sense of **XI.19.1.7.** which is compact, the term closed manifold (which has no real logical meaning) is normally used. The term open manifold is also sometimes used to mean a manifold which is noncompact and has empty boundary. The same conventions are used for smooth manifolds.

In some references, a manifold with empty boundary is called unbounded. However, this term seems much too fraught with potential misinterpretation to be appropriate, especially in an analysis book. I also have a very hard time calling a circle or sphere “unbounded.”

In this book, we will strive for minimum ambiguity at the possible cost of excessive pedantry on these terms: we will always use the phrase “manifold with boundary” when a boundary is allowed, and we will avoid using “compact manifold”, writing “manifold with boundary” if a boundary is allowed and “closed manifold” otherwise. (There is still some ambiguity: is a “closed submanifold” of a manifold a submanifold which is a closed subset of the manifold, or a submanifold which is compact? We will be explicit about which is meant when this occurs.) And we will not use the term “unbounded” in the context of manifolds, except when they are explicit subsets of Euclidean space, where the term “unbounded” has its usual meaning for subsets (i.e. not contained in a ball of finite radius). We will avoid the term “open manifold.”

**XI.19.3. New Manifolds From Old**

**XI.19.4. Classification of Topological Manifolds**

**Number of Manifolds**

**XI.19.4.1.** How many topologically distinct topological manifolds (with or without boundary) are there? Since according to our definition any discrete space is a manifold, the question is meaningless; we should restrict to connected manifolds, or at least to manifolds with only countably many components, i.e. separable manifolds.

Any separable manifold with boundary is a Polish space, and there are \( 2^{\aleph_0} \) topologically distinct Polish spaces (\ref{exercise:xi.19.5.10}), so there are at most \( 2^{\aleph_0} \) topologically distinct separable manifolds with boundary. On the other hand, if \( n > 1 \) there are \( 2^{\aleph_0} \) topologically distinct connected open subsets of \( \mathbb{R}^n \) (**XI.19.5.13.**), each of which is a separable connected \( n \)-manifold, so up to homeomorphism there are exactly \( 2^{\aleph_0} \) separable manifolds, and exactly \( 2^{\aleph_0} \) separable manifolds with boundary. The same is true if “separable” is replaced by “connected.”

There are \( 2^{2^{\aleph_0}} \) topologically distinct 2-dimensional connected Hausdorff locally Euclidean spaces, both with and without boundary (Exercise **XI.19.5.10.**).
The question is more interesting and difficult in the compact case (a compact manifold with boundary is automatically separable). In fact:

**XI.19.4.2.** Theorem. There are only countably many homeomorphism classes of compact manifolds with boundary.

For each \( n > 1 \), there are infinitely many topologically distinct connected closed \( n \)-manifolds \([\ldots]\). See [CK70] for a proof of XI.19.4.2.

**XI.19.5. Exercises**

**XI.19.5.1.** (a) Show that an open set in a locally Euclidean space is locally Euclidean in the relative topology. Show that an open set in a topological manifold is a topological manifold in the relative topology.

(b) Show that a finite cartesian product of locally Euclidean spaces is locally Euclidean. Show that a finite cartesian product of topological manifolds is a topological manifold. What about products of topological manifolds with boundary? [Consider \( \mathbb{H}^n \times \mathbb{H}^m \).]

**XI.19.5.2.** Let \( X \) be a locally Euclidean space, and \( U \) and \( V \) open subsets of \( X \). Suppose \( \phi \) is a homeomorphism from \( U \) to \( V \). Let \( X/\phi \) be the topological space obtained from \( X \) by identifying \( U \) and \( V \) via \( \phi \), with the quotient topology.

(a) Show that \( X/\phi \) is locally Euclidean.

(b) Show that the example of XI.19.3.(v) is obtained this way by letting \( X \) be two disjoint copies of \( \mathbb{R} \) and \( U \) and \( V \) the two copies of \( \mathbb{R} \setminus \{0\} \).

(c) Under what conditions will \( X/\phi \) be Hausdorff?

(d) Construct a circle from \( \mathbb{R} \) by this method, and a Möbius strip from \( \mathbb{R}^2 \).

(e) Generalize to a family of open subsets \( \{U_i : i \in I\} \), with specified homeomorphisms \( \phi_i \) of \( U_i \) with a fixed space \( U_0 \).

**XI.19.5.3.** Let \( X \) and \( Y \) be topological manifolds with boundary. Let \( \phi \) be a homeomorphism from \( \partial X \) to \( \partial Y \) (existence of such a \( \phi \) entails that \( X \) and \( Y \) have the same dimension, at least locally), and \( X \cup_\phi Y \) the topological space obtained from the disjoint union of \( X \) and \( Y \) by identifying \( \partial X \) and \( \partial Y \) by \( \phi \). (We say \( X \) and \( Y \) are *glued* along the boundary via \( \phi \)).

(a) Show that \( X \cup_\phi Y \) is a topological manifold.

(b) What manifold results from gluing two disks along the boundary, which is a circle? What if one or both disks are replaced by a Möbius strip, whose boundary is also a circle?

(c) The boundary of a cylinder is two disjoint circles. What if two cylinders are glued together along the boundary circles? Does the resulting manifold depend on the homeomorphisms chosen?

(d) Show that if \( X \) and \( Y \) are smooth manifolds with boundary, and \( \phi \) is a diffeomorphism, then \( X \cup_\phi Y \) is a smooth manifold.

**XI.19.5.4.** Show that \([0,1) \times (0,1), [0,1) \times [0,1), [0,1) \times [0,1)\], and a closed disk with a single boundary point removed are all homeomorphic to \( \mathbb{H}^2 \) and hence are topological 2-manifolds with boundary.
XI.19.5.5. **The Long Line.** Denote by \( \omega_1 \) the ordered set of ordinals less than the first uncountable ordinal \( \omega \). Give \( L_0 := \omega_1 \times [0,1) \) the lexicographic ordering, and the order topology \( \). \( L_0 \) is called the *long ray*.

(a) If \( a, b \in L_0, a < b \), then the interval \( [a, b] \) is homeomorphic to a closed bounded interval in \( \mathbb{R} \). [Use the fact that any countable successor ordinal is order-isomorphic to a closed subset of \([0, 1)\).] Hence the open interval \( (a, b) \) is homeomorphic to \( \mathbb{R} \). Show that \( L_0 \) is order-complete and that it has no countable cofinal subset. Show that \( L_0 \) is countably compact and sequentially compact, and that every real-valued continuous function on \( L_0 \) is eventually constant.

(b) \( L_0 \) is not locally Euclidean since it has an endpoint (smallest point). If this is removed, a 1-dimensional locally Euclidean space \( L_+ \) is obtained which is called the *open long ray* or *half-long line*. \( L_+ \) is normal, even completely normal (and in particular Hausdorff), as is any space with the order topology (XI.16.1.1.), and is connected since it is order-complete. \( L_+ \) has a countable coinitial set but no countable cofinal set.

(c) Show that the open cover \( \{(a, b) : a, b \in L_+, a < b\} \) of \( L_0 \) has no locally finite refinement. Thus \( L_+ \) is not paracompact. Show directly that \( L_+ \) does not satisfy any of the other conditions of XI.19.1.4.. \( L_+ \) is also not separable. \( L_+ \) is not countably compact or sequentially compact.

(d) Let \( -L_0 \) be a copy of \( L_0 \) with ordering reversed, and \( L \) be the union of \( -L_0 \) and \( L_0 \) with the endpoint of \( -L_0 \) identified with the initial point of \( L_0 \), with the order topology (the points of \( -L_0 \) smaller than the points of \( L_0 \)). Then \( L \) is a 1-dimensional locally Euclidean space which is connected and normal, with neither a countable coinitial set nor a countable cofinal set, does not satisfy any of the conditions of XI.19.1.4., and is not separable. \( L \) is called the *long line*. If \( a, b \in L, a < b \), then the subset \( [a, b] \) of \( L \) is homeomorphic to \([0, 1]\), and \((a, b)\) is homeomorphic to \( \mathbb{R} \). Show that \( L \) is countably compact and sequentially compact, and that every real-valued continuous function on \( L \) is eventually constant in each direction.

(e) Show that \( L, L_+, \) and \( L_0 \) are not contractible.

(f) Show that if \( M \) is a connected 1-dimensional Hausdorff locally Euclidean space, then \( M \) is homeomorphic to exactly one of the following: \( \mathbb{R}, L, L_+ \), or a circle \( S^1 \).

XI.19.5.6. Let \( L \) and \( L_0 \) be as in XI.19.5.5.. Show that the Stone-Čech compactification \( \) of \( L_0 \) coincides with the one-point compactification \( \). Show that the Stone-Čech compactification of \( L \) consists of adding two points to \( L \), one at each “end.”

XI.19.5.7. [?, Appendix A] Consider the following Cartesian products of the spaces from XI.19.5.5.(f):

- \( L \times L \), the *big plane*.
- \( L \times L_+ \), the *big half-plane*.
- \( L_+ \times L_+ \), the *big quadrant*.
- \( L \times \mathbb{R} \), the *wide plane* or *long strip*.
- \( L_+ \times \mathbb{R} \), the *half-long strip*.
- \( L \times S^1 \), the *long cylinder*.
- \( L_+ \times S^1 \), the *half-long cylinder*.
(a) Show that each is a connected 2-dimensional Hausdorff locally Euclidean space not satisfying the conditions of XI.19.1.4.

(b) There are four obvious copies of \( L \) in \( L \times L \), the “\( x \)-axis”, the “\( y \)-axis”, and the “diagonals.” Show that the diagonals are embedded differently than the axes. [Show that any copy of \( \omega_1 \) in \( L \times L \) which hits the \( x \)-axis in uncountably many points has only countably many points off the \( x \)-axis, but this is false for the diagonals.] The same is true in \( L_0 \times L_0 \), where \( L_0 \) is the long ray (there is only one diagonal here). Use this to show that that no two of the above spaces are homeomorphic (it is hardest to distinguish between the big half-plane, the big quadrant, and the half-long strip; cf. XI.19.5.4.). [For another approach, use (c).]

(c) Identify the initial circle of \( L_0 \times S^1 \) to a point. Show that a sequentially compact connected 2-dimensional Hausdorff locally Euclidean space is obtained not satisfying the conditions of XI.19.1.4., sometimes called the big disk. Show that the big disk is not homeomorphic to any of the spaces of (a) (it is hardest to distinguish it from the big plane).

(d) Let \( E \) be the “octant” \( \{ (x, y) \in L_0 \times L_0 : y \leq x \} \). \( E \) has two edges, the \( x \)-axis \( A \) and the diagonal \( D \). Let \( \sigma \) be a finite or infinite sequence of \( A \)'s and \( D \)'s. Let \( M_\sigma \) be the space obtained from the finite disjoint union of \( E \)'s, the same number as the number of terms in \( \sigma \), by taking the \( n \)'th term of \( \sigma \) as the first edge of the \( n \)'th copy of \( E \) and the other edge as the second edge, and identifying the first edge of the \( (n+1) \)'st copy of \( E \) with the second edge of the \( n \)'th copy. Then delete the first edge of the first copy of \( E \) and, if \( \sigma \) is finite, the second edge of the last copy. As a variation if \( \sigma \) is finite, identify the second edge of the last copy of \( E \) with the initial edge of the first copy to obtain a space \( M'_{\sigma} \). All these spaces are connected 2-dimensional locally Euclidean Hausdorff spaces, and the \( M'_{\sigma} \) are sequentially compact. Show that all \( M_\sigma \) and \( M'_{\sigma} \) are mutually nonhomeomorphic, except that if \( \sigma \) is finite, \( \sigma \) can be replaced by its complementary sequence obtained by interchanging \( A \)'s and \( D \)'s and reversing order, and in the case of \( M'_{\sigma} \), \( \sigma \) can be cyclically permuted. We have:

The big plane is \( M'_{\sigma} \), where \( \sigma = (A, D, A, D, A, D, A) \) or \( (D, A, D, A, D, A, D, A) \).

The big half-plane is \( M_\sigma \), where \( \sigma = (A, D, A, D) \).

The big quadrant is \( M_\sigma \), where \( \sigma = (A, D) \).

The big disk can be regarded as \( M'_{\sigma} \), where \( \sigma \) is the empty sequence. The “open octant” is the space \( M_{\sigma} \), where \( \sigma = (A) \) or \( (D) \). The \( M'_{\sigma} \) construction works even for infinite sequences, even (countably) transfinitely [Nyi84].

(e) Describe the Stone-Čech compactification of each of the above spaces.

XI.19.5.8. The Prüfer Manifold. This is an example of a 2-dimensional connected locally Euclidean space not satisfying the conditions of XI.19.1.4.

This space was first described by T. Radó in 1925 [?], and credited to his teacher, H. Prüfer. We describe a version of the construction from [?]; see also [?, Appendix A]. There are actually two rather different spaces constructed this way; only the first, described in this problem, is called the “Prüfer manifold,” and the second, described in Exercise XI.19.5.9., is a variation of a space (called the Moore manifold) constructed by R. L. Moore [?]. See [Roy88, p. 208-209] for a description of Moore’s construction. To add to the terminological confusion, all these spaces are examples of a class of spaces called Moore spaces. A variation of the Prüfer manifold is described in [CV98, p. 268-272].

Although these examples are customarily called “manifolds”, they are not manifolds according to the definition in this book since they are not metrizable (in fact, not even normal).
Begin with an uncountable collection \( \{ \mathbb{R}^2_a : a \in \mathbb{R} \} \) of copies of the plane, and an additional copy \( \mathbb{R}^2_+ \) of the open upper half plane. Take the separated union \( U \) of the \( \mathbb{R}^2_a \) and \( \mathbb{R}^2_+ \); then \( U \) is a 2-manifold. Identify the open upper half plane of each \( \mathbb{R}^2_a \) with \( \mathbb{R}^2_+ \) by

\[
(x, y) \in \mathbb{R}^2_a \ (y > 0) \quad \longleftrightarrow \quad (xy + a, y) \in \mathbb{R}^2_+
\]

and let \( P \) be the quotient space. \( P \) is called the Pr"{u}fer manifold.

(a) Show that \( P \) is a connected 2-dimensional locally Euclidean space (cf. Exercise XI.19.5.2.(a)).

(b) Show that \( P \) is Hausdorff (cf. Exercise XI.19.5.2.(c)).

(c) Let \( \mathcal{Z} \) be the union of the copies of the \( x \)-axis in \( \mathbb{R}^2_a \) for all \( a \). Show that in the relative topology \( \mathcal{Z} \) is a separated union of uncountably many copies of \( \mathbb{R} \): the \( x \)-axis of each \( \mathbb{R}^2_a \) is clopen in \( \mathcal{Z} \). Thus \( \mathcal{Z} \) and hence \( P \) is not second countable, so \( P \) does not satisfy the conditions of XI.19.1.4.

(d) Let \( A \) be the subset of \( \mathcal{Z} \) consisting of the points on the \( x \)-axis of \( \mathbb{R}^2_a \) for \( a \) rational, and \( B = \mathcal{Z} \setminus A \) the set of points on the \( x \)-axis of \( \mathbb{R}^2_a \) for \( a \) irrational. Then \( A \) and \( B \) are disjoint closed subsets of \( P \). Use the Baire Category Theorem () to prove that \( A \) and \( B \) do not have disjoint neighborhoods in \( P \). Thus \( P \) is not normal.

(e) Show that \( P \) is not separable. [Consider points in the open lower half-spaces of the \( \mathbb{R}^2_a \).]

(f) Show that \( P \) is contractible (cf. [?, Appendix A]).

**XI.19.5.9.** This problem is a continuation of XI.19.5.8.

(a) Let \( P_0 \) be the union of \( \mathcal{Z} \) and the upper half plane of \( P \) (i.e. the image of \( \mathbb{R}^2_+ \) in \( P \)). Then \( P_0 \) is a “2-dimensional locally Euclidean space with boundary.” \( P_0 \) is connected and separable since \( \mathbb{R}^2_+ \) is dense in \( P_0 \), but \( \partial P_0 = \mathcal{Z} \) consists of a separated union of uncountably many copies of \( \mathbb{R} \).

(b) Let \( Q \) be obtained by gluing two copies of \( P_0 \) along the boundary. Then \( Q \) is a 2-dimensional connected Hausdorff locally Euclidean space which is separable but does not satisfy the conditions of XI.19.1.4.

(c) \( Q \) can alternately be constructed from \( P \) by identifying for each \( a \) the open lower half plane of \( \mathbb{R}^2_a \) with a new copy \( \mathbb{R}^2_0 \) of the open lower half plane by

\[
(x, y) \in \mathbb{R}^2_a \ (y < 0) \quad \longleftrightarrow \quad (xy + a, y) \in \mathbb{R}^2_0
\]

and taking the quotient topology.

(d) The closed subsets \( A \) and \( B \) of \( \mathcal{Z} \) in XI.19.5.8.(d) also do not have disjoint neighborhoods in \( Q \), so \( Q \) is not normal.

(e) \( Q \) is not simply connected: the fundamental group of \( Q \) is quite complicated (uncountable).

(f) Another example can be constructed from \( P_0 \) by identifying, for each \( a \), the positive and negative \( x \)-axes of \( \mathbb{R}^2_a \) by folding over at 0. A separable locally Euclidean space \( M \) is obtained. This is the Moore manifold. The Moore manifold is contractible (), hence simply connected.

**XI.19.5.10.** If \( A \subseteq \mathbb{R} \), modify the construction of \( P \), \( P_0 \), \( Q \), and \( M \) by attaching \( \mathbb{R}^2_a \) if and only if \( a \in A \). Call the resulting spaces \( P_A \), \( P_0^A \), \( Q_A \), \( M_A \).

(a) Show that if \( A \) is countable, \( P_A \) and \( M_A \) are homeomorphic to \( \mathbb{R}^2 \). If \( \text{card}(A) = n \), \( 1 \leq n \leq \aleph_0 \), \( Q_A \) is homeomorphic to a plane with a discrete set of points of cardinality \( n - 1 \) removed (where \( \aleph_0 - 1 = \aleph_0 \)).
(b) [Bla] Show that if $A$ is a subset of $\mathbb{R}$, there at most $2^{\aleph_0}$ subsets $B$ of $\mathbb{R}$ such that the spaces $P_A^0$ and $P_B^0$ are homeomorphic. [Regard all the $P_A^n$ as subsets of $P_0$. A homeomorphism from $P_A^n$ to $P_B^n$ can be regarded as a continuous function from $P_A^n$ to $P_0$. But since $P_A^n$ is separable, and $P_0$ has cardinality $2^{\aleph_0}$, there are only $2^{\aleph_0}$ continuous functions from $P_A^n$ to $P_0$. cf. XI.18.4.42.] So there are $2^{\aleph_0}$ topologically distinct $P_A^0$. Show similarly that there are $2^{\aleph_0}$ topologically distinct $M_A$. The $P_A^0$ and $M_A$ are all separable and contractible. There are also $2^{\aleph_0}$ topologically distinct $P_A$ and $Q_A$. The $M_A$ and $P_A$ constructions can be combined into a unified family.

A family of $2^{\aleph_0}$ mutually nonhomeomorphic simply connected 2-dimensional Hausdorff locally Euclidean spaces, as well as a family of $2^{\aleph_0}$ mutually nonhomeomorphic sequentially compact simply connected 2-dimensional Hausdorff locally Euclidean spaces (the maximum number possible), is constructed in [Nyi84], which is a thorough study of nonmetrizable “manifolds.” However, these examples are not separable or contractible.

XI.19.5.11. ([Kne60]; cf. [Spi79, p. 643]) Construct a three-dimensional version of $Q$ by taking a copy $\mathbb{R}_a^3$ for each $a \in \mathbb{R}$, and identifying $(x, y, z) \in \mathbb{R}_a^3$, $y \neq 0$, with the point in a fixed copy of $\mathbb{R}^4$ with coordinates $(a + yx, y, z + a)$ if $y > 0$ and $(a + yx, y, z - a)$ if $y < 0$.

(a) Show that the quotient $Q^3$ is a connected separable Hausdorff locally Euclidean space with local dimension 3.

(b) Show that the surfaces $z = c$ give a foliation () of $Q^3$ with only one leaf, which is two-dimensional. Thus there is a stronger topology on $Q^3$ under which it is a connected Hausdorff locally Euclidean space of local dimension 2 (not separable).

(c) For an arbitrary subset $A$ of $\mathbb{R}$, similarly construct $Q_A^3$. Show there are $2^{\aleph_0}$ distinct homeomorphism classes. Describe $Q_A^3$ if $A$ is finite or countable. There is still a similar foliation of $Q_A^3$, but it has multiple leaves in general.

XI.19.5.12. [Spi79, Problem A-8] Show that a connected Hausdorff locally Euclidean space which is not a single point has cardinality $2^{\aleph_0}$.

XI.19.5.13. [Spi79, Problem 1-24] (a) Let $X$ and $Y$ be totally disconnected closed subsets of $\mathbb{R}^2$. Show that if $\mathbb{R}^2 \setminus X$ and $\mathbb{R}^2 \setminus Y$ are homeomorphic, then $X$ and $Y$ are homeomorphic. [See XI.18.5.3.]

(b) Conclude from () and () that there are $2^{\aleph_0}$ topologically distinct connected open subsets of $\mathbb{R}^2$.

(c) There are also $2^{\aleph_0}$ topologically distinct connected relatively open subsets of the closed unit disk in $\mathbb{R}^2$, each containing the entire boundary circle.

(d) Show that these spaces remain topologically distinct even if the Cartesian product with $\mathbb{R}^m$ for any $m$ is taken. Thus there are $2^{\aleph_0}$ topologically distinct connected open sets in $\mathbb{R}^n$ for any $n > 1$; in fact, see XI.18.5.6.–XI.18.5.8.(h). (There is only one nonempty connected open set in $\mathbb{R}$ up to homeomorphism, and only countably many open sets in $\mathbb{R}$ in all up to homeomorphism, since every open set is a countable disjoint union of open intervals (.)

XI.19.5.14. [Hat02, 1.3.25] An Interesting Non-Hausdorff Locally Euclidean Space. Let $T$ be the invertible linear transformation on $\mathbb{R}^2$ given by $T(x, y) = (2x, y/2)$. Define an equivalence relation on $E = \mathbb{R}^2 \setminus \{0\}$ by $v \sim w$ if $T^n v = w$ for some $n \in \mathbb{Z}$. Let $M$ be the quotient of $E$ by this equivalence relation, and $\pi: E \to M$ the quotient map.
(a) Show that each equivalence class is closed and discrete in $E$, and that $M$ is $T_1$ and second countable.

(b) Show that $M$ is the disjoint union of four cylinders $C_1, C_2, C_3, C_4$, each homeomorphic to $S^1 \times \mathbb{R}$, where $C_j$ is the image of the open $j$'th quadrant in $E$, and four circles $X_+, X_-, Y_+, Y_-$, where $X_+$ is the image of the positive $x$-axis, $X_-$ the image of the negative $x$-axis, and $Y_\pm$ the images of the positive and negative $y$-axis. $C_1 - C_4$ are open in $M$ and $X_\pm, Y_\pm$ are closed in $M$. The “horizontal” circle in $C_j$ at height $|r|$ is the image of the part of the hyperbola $xy = r$ in the $j$'th quadrant.

(c) The set $\{(x, y) : 1 \leq x < 2\}$ maps one-one onto $C_1 \cup X_+ \cup C_4$ under $\pi$, and the image in $M$ is open and homeomorphic to $S^1 \times \mathbb{R}$. Similarly, $\{(x, y) : -2 \leq x < -1\}$ maps one-one onto $C_2 \cup X_- \cup C_3$, $\{(x, y) : 1 \leq y < 2\}$ maps one-one onto $C_1 \cup Y_+ \cup C_2$ and $\{(x, y) : -2 \leq y < -1\}$ maps one-one onto $C_3 \cup Y_- \cup C_4$. These cylinders are open in $M$. Conclude that $M$ is locally Euclidean.

(d) The closure of $C_1$ is $C_1 \cup X_+ \cup Y_+$. Show that if $p \in X_+$ and $q \in Y_+$, then $p$ and $q$ do not have disjoint neighborhoods. Thus $M$ is not Hausdorff. Similarly, points in $X_\pm$ and $Y_\pm$ do not have disjoint neighborhoods. (However, $X_+$ and $X_-$ have disjoint neighborhoods, as do $Y_+$ and $Y_-$.) Describe the closure of $C_2 - C_4$.

(e) Let $S$ be the image of a vertical line segment $\{(x, y) : x = x_0, 0 < y \leq y_0\}$. Show that the closure of $S$ consists of $S$, one point of $X_+$, and all of $Y_+$. Find a sequence in $S$ which converges to a point in $X_+$ and has every point of $Y_+$ as a cluster point, and a subsequence which converges to any arbitrarily given point of $Y_+$. Do the same for other vertical and horizontal line segments.

(f) Show that $\pi$ is a covering map $()$ from $E$ to $M$.

Let $N$ be the space obtained from $M$ by additionally identifying diametrically opposed points of $E$ ($N$ is the quotient of $E$ by the action of the subgroup $G$ of $SL_2(\mathbb{R})$ generated by $T$ and $-I$, which is isomorphic to $\mathbb{Z} \times \mathbb{Z}_2$). Then $N$ consists of two open cylinders $C_+$ (the image of $C_1$ and $C_3$) and $C_-$ (the image of $C_2$ and $C_4$), and two closed circles $X$ (the image of $X_+$ and $X_-$) and $Y$ (the image of $Y_+$ and $Y_-$. $C_+ \cup X \cup C_-$ and $C_+ \cup Y \cup C_-$ are open subsets of $N$ homeomorphic to $S^1 \times \mathbb{R}$; thus $N$ is a cylinder $S^1 \times \mathbb{R}$ with the circle $S^1 \times \{0\}$ “doubled.” If $C_+$ is identified with $S^1 \times \mathbb{R}_+$ by taking the images of the hyperbolas $xy = r$ as horizontal circles and images of rays from the origin as vertical lines (one only needs rays of slope between $1/2$ and $2$), then points of $X$ are obtained by approaching $0$ along the image of a vertical ray in the right half-plane, i.e. a line spiraling around the cylinder at a certain rate, and points of $Y$ are obtained by approaching along the image of a horizontal ray, i.e. a line spiraling around the cylinder at the same rate but in the opposite direction. The quotient map from $E$ to $N$ is also a covering map, i.e. $G$ acts properly discontinuously on $E$ $()$.

XI.19.5.15. Show that a metrizable topological space $X$ is a manifold with boundary if and only if every point of $X$ has a closed neighborhood homeomorphic to a closed ball in a Euclidean space.
XI.20. The Brouwer Fixed-Point Theorem

The Brouwer Fixed-Point Theorem, in its various equivalent formulations, is arguably the most important fact about the topology of Euclidean space (although the Heine-Borel Theorem () is arguably more fundamental), and indeed one of the most important results in all of topology. It essentially makes possible notions of “higher-order connectedness.” Among other things, the entire subject of algebraic topology is at least implicitly based on this result. A coherent theory of dimension for topological spaces, and many applications in analysis, also depend on this theorem.

XI.20.1. Statements of the Results

XI.20.1.1. We use the following notation. Let $B^n$ be the closed unit ball and $S^{n-1}$ the closed unit sphere in $\mathbb{R}^n$, i.e.

$$B^n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$$
$$S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}.$$ 

In writing these, we implicitly assume $n \in \mathbb{N}$ (although they also make sense if $n = 0$: $B^0$ is a single point and $S^{-1} = \emptyset$). The term “map” will mean “continuous function.”

XI.20.1.2. Theorem. [Brouwer Fixed-Point Theorem] Every continuous function $f : B^n \to B^n$ has a fixed point, i.e. there is a point $x \in B^n$ with $f(x) = x$.

XI.20.1.3. Theorem. [No-Retraction Theorem] There is no retraction () from $B^n$ to $S^{n-1}$.

XI.20.1.4. Theorem. [No-Contraction Theorem] The sphere $S^{n-1}$ is not contractible, i.e. the identity map on $S^{n-1}$ is not homotopic to a constant map.

XI.20.1.5. We will show by elementary arguments that these three statements are equivalent. Thus to prove all the theorems it suffices to prove any one of them. We will also obtain some additional equivalent formulations of the theorem.

XI.20.1.6. If $n = 1$, the statements are quite simple. $B^1 = [-1, 1]$ and $S^0 = \{-1, 1\}$. Thus the Brouwer Fixed-Point Theorem for $n = 1$ is (), essentially the Intermediate Value Theorem of calculus. The No-Retraction Theorem is also the Intermediate Value Theorem, or the fact that a connected space cannot be mapped continuously onto a disconnected space. The No-Contraction Theorem is obvious, since $S^0$ is not even connected (cf. ()).

If $n = 2$, the statements are also fairly elementary to prove; see XI.20.2.. But if $n \geq 3$, all (known) proofs are fairly involved and/or use considerable machinery of some sort. There are many quite varied proofs known, and which one is the “best” is a quite subjective matter.

To prove that XI.20.1.2.–XI.20.1.4. are equivalent, it is convenient to show the negations of the statements are equivalent (although this seems a little perverse, since the ultimate result is that these statements are all false):
XI.20.1.7. **Theorem.** Let $n \in \mathbb{N}$. The following statements are equivalent:

(i) There is a continuous function $f : B^n \to B^n$ with no fixed point.

(ii) There is a retraction $r : B^n \to S^{n-1}$.

(iii) $S^{n-1}$ is contractible.

XI.20.1.8. To prove (ii) $\Rightarrow$ (iii), note that there is a contraction of $B^n$. The simplest one is given by the function $g(t, x) = t x$ for $0 \leq t \leq 1$, $x \in B^n$. If there were a retraction $r : B^n \to S^{n-1}$, the function $h(t, x) = r(tx)$ ($0 \leq t \leq 1$, $x \in S^{n-1}$) would give a contraction of $S^{n-1}$ (to the point $x_0 = r(0)$). [Note that essentially the same argument shows that if $X$ is a contractible space and $r$ is a retraction of $X$ onto a subspace $Y$, then $Y$ is also contractible.]

XI.20.1.9. For the converse (iii) $\Rightarrow$ (ii), suppose $h : [0, 1] \times S^{n-1} \to S^{n-1}$ is a contraction of $S^{n-1}$ with $h(1, x) = x$ for all $x \in S^{n-1}$ and $h(0, x) = x_0$ for all $x \in S^{n-1}$, where $x_0$ is a fixed element of $S^{n-1}$. Define $r : B^n \to S^{n-1}$ by $r(x) = h \left( \|x\|, \frac{x}{\|x\|} \right)$ for $x \neq 0$ and $r(0) = x_0$. It is easily checked that $r$ is continuous (only continuity at 0 is an issue), and is therefore a retraction from $B^n$ to $S^{n-1}$.

Thus XI.20.1.7. (ii) and (iii) are equivalent.

XI.20.1.10. To prove (ii) $\Rightarrow$ (i), let $r$ be a retraction of $B^n$ to $S^{n-1}$. Then the map $f : B^n \to B^n$ defined by $f(x) = -r(x)$ has no fixed point.

XI.20.1.11. The proof of (i) $\Rightarrow$ (ii) is the hardest to write carefully, but the argument is geometrically simple and appealing. Suppose $f : B^n \to B^n$ has no fixed point. For each $x \in B^n$, the line segment from $f(x)$ to $x$ is nontrivial, and can be extended until it hits $S^{n-1}$ in a point we call $r(x)$:

![Figure XI.65: Definition of the retraction](image)

The function thus defined will be a retraction from $B^n$ to $S^{n-1}$.

1273
The only issue is whether \( r \) is continuous. This is geometrically plausible, and can be established carefully as follows. The vector \( r(x) \) is of the form \( f(x) + t(x - f(x)) \) for some \( t \geq 1 \). We can determine \( t \) explicitly using inner products and the quadratic formula:

\[
\begin{align*}
\langle f(x) + t(x - f(x)), f(x) + t(x - f(x)) \rangle &= 1 \\
\langle f(x), f(x) \rangle + 2t\langle x - f(x), f(x) \rangle + t^2\langle x - f(x), x - f(x) \rangle - 1 &= 0 \\
\|x - f(x)\|^2t^2 - 2\langle f(x) - x, f(x) \rangle t + (\|f(x)\|^2 - 1) &= 0 \\
t &= \frac{\langle f(x) - x, f(x) \rangle + \sqrt{(\langle f(x) - x, f(x) \rangle)^2 + \|x - f(x)\|^2(1 - \|f(x)\|^2)}}{\|x - f(x)\|^2}
\end{align*}
\]

which is a continuous function of \( x \) since the denominator is never zero.

This completes the proof of XI.20.1.7.

\[\square\]

**Essential Maps**

**XI.20.1.12.** If \( X \) is a topological space of “dimension” less than \( n \) (we have not yet carefully defined what this means), and \( f \) is a map from \( X \) into \( \mathbb{R}^n \) (or \( B^n \)), the range of \( f \) should be a set of dimension \(< n \) and should thus have dense complement; so it should be possible to perturb \( f \) to a uniformly close function \( g \) such that the range of \( g \) does not include some specified point, e.g. the origin. Actually, the existence of such pathological functions as space-filling curves () casts doubt on the intuition behind this argument, but it turns out rather remarkably that the conclusion is correct.

Another argument leading to the conclusion is that a system of \( n \) equations in fewer than \( n \) variables “usually” does not have a solution. In special cases there might happen to be a solution, but it should be possible to perturb such a system to a “nearby” system with no solution. On the other hand, it is essentially the content of the Inverse Function Theorem () that if we have a “smooth” system of \( n \) equations in \( n \) variables which has a locally unique solution, then any “nearby” smooth system also has a (locally unique) solution.

**XI.20.1.13.** Definition. Let \( X \) be a topological space. A map (continuous function) \( f : X \to B^n \) is **inessential** if for every \( \epsilon > 0 \) there is a map \( g : X \to B^n \) with \( \|f - g\|_{\infty} = \sup\{\|f(x) - g(x)\| : x \in X\} < \epsilon \) and with \( 0 \notin g(X) \). A map \( f \) from \( X \) to \( B^n \) is **essential** if it is not inessential, i.e. every map sufficiently uniformly close to \( f \) has \( 0 \) in the range.

**XI.20.1.14.** This use of the term **essential** is somewhat nonstandard. In topology, it is usually said that a map \( f : X \to Y \) is essential if it is not homotopic to a constant map. Thus any map from a topological space to a contractible space such as \( B^n \) would be inessential. We will use the term in our sense only in relation to maps into \( B^n \).

**XI.20.1.15.** The definition of an essential map makes use of the standard metric on \( B^n \). But the definition would be unchanged if any other metric giving the standard topology were used instead, since any two metrics giving the same compact topology are uniformly equivalent (). Thus the property of being essential or inessential is a purely topological property.

We then obtain another statement equivalent to the Brouwer Fixed-Point Theorem:
XI.20.1.16. \textbf{Theorem.} The identity map from $B^n$ to $B^n$ is essential.


XI.20.1.17. To prove that this statement is equivalent to the No-Retraction Theorem, and thus to the Brouwer Fixed-Point Theorem, suppose that the identity map $f$ on $B^n$ is inessential, and fix $g : B^n \to B^n$ with $\|f-g\| < \frac{1}{2}$ and $g(x) \neq 0$ for all $x \in B^n$. We will gradually modify $g$ on the annulus $\{ x \in B^n : \frac{1}{2} \leq \|x\| \leq 1 \}$ to a function $h : B^n \to B^n$ with $h(x) \neq 0$ for all $x \in B^n$ and $h(x) = x$ for $x \in S^{n-1}$. We will take $h(x)$ to be a point on the line segment between $x$ and $g(x)$, closer to $x$ as $\|x\|$ approaches 1. The precise formula is

$$h(x) = \begin{cases} g(x) & \text{if } \|x\| \leq \frac{1}{2} \\ (2\|x\| - 1)x + (2 - 2\|x\|)g(x) & \text{if } \frac{1}{2} \leq \|x\| \leq 1 \end{cases} .$$

It is clear that $h$ is a continuous function from $B^n$ to $B^n$ and that $h(x) = x$ if $\|x\| = 1$. To show that $h(x) \neq 0$ for all $x$, it suffices to note that if $\|x\| \geq \frac{1}{2}$, then $0$ is not on the line segment between $x$ and $g(x)$ since $\|x - g(x)\| < \frac{1}{2}$.

There is a retraction $r$ from $B^n \setminus \{0\}$ to $S^{n-1}$, e.g. radial projection; the map $r \circ h$ would then be a retraction from $B^n$ to $S^{n-1}$. Thus by the No-Retraction Theorem, $f$ cannot be inessential.

Conversely, suppose there is a retraction $r$ from $B^n$ to $S^{n-1}$. By scaling, for any $\epsilon > 0$ there is a retraction $r_\epsilon$ from $B^n = \{ x \in \mathbb{R}^n : \|x\| \leq \epsilon \}$ onto $S^{n-1} = \{ x \in \mathbb{R}^n : \|x\| = 1 \}$. Define $g_\epsilon : B^n \to B^n$ by

$$g_\epsilon(x) = \begin{cases} r_\epsilon(x) & \text{if } \|x\| \leq \epsilon \\ x & \text{if } \epsilon \leq \|x\| \leq 1 \end{cases} .$$

Then $\|f - g_\epsilon\| \leq 2\epsilon$ and $g_\epsilon(x) \neq 0$ for all $x \in B^n$, so $f$ is inessential. Thus XI.20.1.16. implies the No-Retraction Theorem.

\textbf{Partitions and the Lebesgue Covering Theorem}

XI.20.1.18. We now discuss two more equivalent versions of the Brouwer Fixed-Point Theorem. The statements are slightly more technical, and are phrased in terms of the $n$-cube

$$I^n = [-1,1]^n = \{(x_1, \ldots, x_n): -1 \leq x_k \leq 1\}$$

which is obviously homeomorphic to $B^n$. We will let $A_k$ and $B_k$ $(1 \leq k \leq n)$ be the opposite $k$-faces of $I^n$:

$$A_k = \{(x_1, \ldots, x_n) \in I^n : x_k = -1\}$$

$$B_k = \{(x_1, \ldots, x_n) \in I^n : x_k = 1\}$$

which are obviously disjoint closed subsets of $I^n$. The union of all the $A_k$ and $B_k$ is the boundary $\partial I^n$ of $I^n$, which is homeomorphic to $S^{n-1}$.

Recall the definition of a partition (XI.7.6.4.).

XI.20.1.19. \textbf{Theorem.} [Face Partition Theorem] For $1 \leq k \leq n$, let $P_k$ be a partition between $A_k$ and $B_k$ in $I^n$. Then

$$P_1 \cap \cdots \cap P_n \neq \emptyset .$$
XI.20.1.20. We show that the Face Partition Theorem for \( n \in \mathbb{N} \) is equivalent to the No-Retraction Theorem for the same \( n \), and hence to the Brouwer Fixed-Point Theorem. It is again convenient to prove that the negations of the statements are equivalent.

Suppose there is a retraction \( r \) from \( I^n \) to \( \partial I^n \). Let \((r_1, \ldots, r_n)\) be the coordinate functions of \( r \), and set

\[
P_k = r_k^{-1}(\{0\})
\]

\[
U_k = r_k^{-1}([-1, 0])
\]

\[
V_k = r_k^{-1}((0, 1])
\]

Then \( P_k \) is a partition between \( A_k \) and \( B_k \) for each \( k \), and clearly \( P_1 \cup \cdots \cup P_n = \emptyset \), so the statement of the Face Partition Theorem is false.

Conversely, let \( P_k \) be a partition between \( A_k \) and \( B_k \) for each \( k \), with associated open sets \( U_k \) containing \( A_k \) and \( V_k \) containing \( B_k \), such that \( P_1 \cap \cdots \cap P_n = \emptyset \). For each \( k \), define \( f_k : I^n \to [-1, 1] \) by

\[
f_k(x) = \begin{cases} 
\rho(x, P_k) & \text{if } x \in U_k \\
\frac{\rho(x, P_k) + \rho(x, A_k)}{\rho(x, P_k) + \rho(x, B_k)} & \text{if } x \in P_k \\
\frac{-\rho(x, P_k)}{\rho(x, P_k) + \rho(x, B_k)} & \text{if } x \in V_k 
\end{cases}
\]

where \( \rho \) is the standard metric on \( I^n \). Then \( f_k \) is continuous. If \( f : I^n \to I^n \) is the function with coordinate functions \((f_1, \ldots, f_n)\), then \( f(x) \neq 0 \) for all \( x \in I^n \). Let \( r \) be a retraction of \( I^n \setminus \{0\} \) onto \( \partial I^n \), and let \( g = r \circ f \). It is easy to check that \( g(A_k) \subseteq B_k \) and \( g(B_k) \subseteq A_k \). It follows that \( g \) cannot have a fixed point, so the statement of the Brouwer Fixed-Point Theorem is false.

XI.20.1.21. Theorem. [Lebesgue Covering Theorem] Let \( \mathcal{F} \) be a finite collection of closed sets covering \( I^n \), with the property that no \( F \in \mathcal{F} \) meets both \( A_k \) and \( B_k \) for any \( k \). Then there are distinct \( F_1, \ldots, F_{n+1} \) in \( \mathcal{F} \) with

\[
F_1 \cap \cdots \cap F_{n+1} \neq \emptyset
\]

(i.e. \( \mathcal{F} \) has order \( \geq n \)).

XI.20.1.22. It is easy to show that the Lebesgue Covering Theorem implies the Brouwer Fixed-Point Theorem. Again we show the contrapositive. Suppose \( r \) is a retraction of \( I^n \) onto \( \partial I^n \). Let

\[
\mathcal{F} = \{r^{-1}(A_k), r^{-1}(B_k) : 1 \leq k \leq n\}
\]

which is a finite cover of \( I^n \) by \( 2n \) sets satisfying the hypotheses of the Lebesgue Covering Theorem, but no \( n+1 \) of the sets can have nonempty intersection.

To show the Face Partition Theorem, and hence the Brouwer Fixed-Point Theorem, implies the Lebesgue Covering Theorem is more complicated, and requires a lemma (cf. [Eng95, Lemma 1.2.9]):

1276
**XI.20.1.23. Lemma.** Let $X$ be a metrizable space, or more generally a completely normal space ($\sim$), and $E$ a closed subset of $X$. Let $Y$ and $Z$ be disjoint closed subsets of $X$. Let $\tilde{P}$ be a partition between $Y \cap E$ and $Z \cap E$ in $E$. Then there is a partition $P$ between $Y$ and $Z$ in $X$ with $P \cap E \subseteq \tilde{P}$.

**Proof:** There are relatively open subsets $\tilde{U}$, $\tilde{V}$ of $E$ such that $Y \cap E \subset \tilde{U}$, $Z \cap E \subset \tilde{V}$, $\tilde{U} \cap \tilde{V} = \emptyset$, and $E \setminus (\tilde{U} \cup \tilde{V}) = \tilde{P}$. The closure of $Z \cup \tilde{V}$ (in $X$) is $Z \cup [\tilde{V}]^-$, and $Y \cap [\tilde{V}]^- = (Y \cap E) \cap [\tilde{V}]^- = \emptyset$ since $[\tilde{V}]^- \subset E$ because $E$ is closed, and $Y \cap E \subseteq U$, which is relatively open in $E$. Thus $Y$ is disjoint from the closure of $Z \cup \tilde{V}$. We also have that $\tilde{U}$ is disjoint from both $Z$ and $[\tilde{V}]^-$. So $Y \cup \tilde{U}$ is disjoint from the closure of $Z \cup \tilde{V}$. By a symmetric argument $Z \cup \tilde{V}$ is disjoint from the closure of $Y \cup \tilde{U}$. Thus by (1) there are disjoint open sets $U$ and $V$ in $X$ with $Y \cup U \subseteq U$ and $Z \cup \tilde{V} \subseteq V$. Set $P = X \setminus (U \cup \tilde{V})$. Then $P$ is a partition between $Y$ and $Z$ in $X$, and $P \cap E$ is disjoint from both $\tilde{U}$ and $\tilde{V}$, i.e. $P \cap E \subseteq \tilde{P}$. \hfill $\Diamond$

**XI.20.1.24.** We now show that the Face Partition Theorem for an $n \in \mathbb{N}$ implies the Lebesgue Covering Theorem for the same $n$ (cf. [Eng95, Theorem 1.8.15]). Let $\mathcal{F}$ be a finite collection of closed subsets of $I^n$ satisfying the hypotheses of the Lebesgue Covering Theorem. For $1 \leq k \leq n$ set

$$\mathcal{F}_k = \{ F \in \mathcal{F} : A_k \cap F \neq \emptyset \}$$

$$\mathcal{E}_1 = \mathcal{F}_1, \quad \mathcal{E}_{k+1} = \mathcal{F}_{k+1} \setminus (\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k) \text{ for } 1 \leq k \leq n$$

$$E_k = \cup \{ F : F \in \mathcal{E}_k \} \text{ for } 1 \leq k \leq n + 1.$$ Then the $E_k$ are closed, and we have

$$A_k \subseteq I^n \setminus (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$B_k \subseteq I^n \setminus E_k.$$ Set $\tilde{P}_0 = I^n$, and for $1 \leq k \leq n$ set

$$\tilde{P}_k = E_1 \cap \cdots \cap E_k \cap (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$\tilde{U}_k = \tilde{P}_{k-1} \setminus (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$\tilde{V}_k = \tilde{P}_{k-1} \setminus (E_1 \cap \cdots \cap E_k).$$ Then, for $1 \leq k \leq n$, $\tilde{P}_k$ is closed, $\tilde{P}_k \subseteq \tilde{P}_{k-1}$, $A_k \cap \tilde{P}_{k-1} \subseteq \tilde{U}_k$, $B_k \cap \tilde{P}_{k-1} \subseteq \tilde{V}_k$, $\tilde{U}_k \cap \tilde{V}_k = \emptyset$, and

$$\tilde{P}_{k-1} \setminus (\tilde{U}_k \cup \tilde{V}_k) = \tilde{P}_{k-1} \cap \tilde{P}_k = \tilde{P}_k$$

so $\tilde{P}_k$ is a partition between $A_k \cap \tilde{P}_{k-1}$ and $B_k \cap \tilde{P}_{k-1}$ in $\tilde{P}_{k-1}$. Apply Lemma XI.20.1.23. successively to obtain for each $k$ a partition $P_k$ between $A$ and $B$ in $I^n$ with $P_k \cap \tilde{P}_{k-1} \subseteq \tilde{P}_k$. Thus by the Face Partition Theorem

$$\emptyset \neq P_1 \cap \cdots \cap P_n \subseteq \tilde{P}_n = E_1 \cap \cdots \cap E_{n+1}.$$ If $x \in E_1 \cap \cdots \cap E_{n+1}$, then for $1 \leq k \leq n + 1$ there is an $F_k \in \mathcal{E}_k$ with $x \in F_k \subseteq E_k$; then $F_1 \cap \cdots \cap F_{n+1} \neq \emptyset$, and the $F_k$ are all distinct since each $F \in \mathcal{F}$ is in exactly one $\mathcal{E}_k$.

**XI.20.1.25.** The Face Partition Theorem and the Lebesgue Covering Theorem are versions of the statements $\text{ind}(\mathbb{R}^n) = n$ ($\sim$) and $\text{dim}(\mathbb{R}^n) = n$ ($\sim$) respectively; cf. [Eng95].

1277
XI.20.2. The Case \( n = 2 \)

The Brouwer Fixed-Point Theorem for \( n = 2 \) is not as easy as the case \( n = 1 \) (XI.20.1.6.), but is relatively simple.

XI.20.2.1. The argument uses the map \( p : \mathbb{R} \to S^1 \) defined by

\[
p(t) = (\cos 2\pi t, \sin 2\pi t).
\]

This map is surjective, but far from injective; however, if \( s_0 \in \mathbb{R} \) and \( z_0 = p(s_0) \in S^1 \), the restriction of \( p \) to the open interval \((s_0 - \frac{1}{2}, s_0 + \frac{1}{2})\) is a homeomorphism from \((s_0 - \frac{1}{2}, s_0 + \frac{1}{2})\) onto \( S^1 \setminus \{ -z_0 \} \). We will write \( p^{-1}_{s_0} \) for the inverse map from \( S^1 \setminus \{ -z_0 \} \) onto \((s_0 - \frac{1}{2}, s_0 + \frac{1}{2})\).

XI.20.2.2. Lemma. [Path-Lifting Lemma] Let \( \gamma : [0,1] \to S^1 \) be a path (continuous function), with \( \gamma(0) = z_0 \). If \( s_0 \in \mathbb{R} \) is fixed with \( p(s_0) = z_0 \), then there is a unique path \( \tilde{\gamma} : [0,1] \to \mathbb{R} \) with \( \tilde{\gamma}(0) = s_0 \) and \( \gamma = p \circ \tilde{\gamma} \).

Proof: A partition \( \mathcal{P} = \{0 = t_0, t_1, \ldots, t_n = 1\} \) of \([0,1]\) will be called sufficiently fine if, for each \( k \), \( \|\gamma(t) - \gamma(s)\| < 2 \) for all \( t, s \in [t_{k-1}, t_k] \). There is a sufficiently fine partition since \( \gamma \) is uniformly continuous (\( \cdot \)). Suppose \( \mathcal{P} \) is a sufficiently fine partition. There is a unique path \( \beta : [0,t_1] \to \mathbb{R} \) with \( \beta(0) = s_0 \) and \( \gamma|_{[0,t_1]} = p \circ \beta \), namely \( \beta(t) = p^{-1}_{s_0}(\gamma(t)) \) for \( 0 \leq t \leq t_1 \) [to show uniqueness, note that the range of \( \beta \) must be contained in a connected subset of \( p^{-1}(\gamma([0,t_1])) \subseteq p^{-1}(S^1 \setminus \{ -z_0 \}) \), so \( \beta([0,t_1]) \subseteq (s_0 - \frac{1}{2}, s_0 + \frac{1}{2}) \).] In particular, \( s_1 = \beta(t_1) \) is uniquely determined by \( \gamma \) and the choice of \( s_0 \). Define \( \tilde{\gamma}(t) \) for \( t \in [0,t_1] \) to be \( \beta(t) \).

Now let \( z_1 = p(s_1) = \gamma(t_1) \), and repeat the construction to define \( \tilde{\gamma} \) on \([t_1,t_2]\), and inductively extend to all of \([0,1]\). We then have \( \tilde{\gamma}(0) = s_0 \) and \( \gamma = p \circ \tilde{\gamma} \).

To show uniqueness, suppose \( \tilde{\gamma} \) is another lift of \( \gamma \), i.e. \( \tilde{\gamma}(0) = s_0 \) and \( \gamma = p \circ \tilde{\gamma} \) (e.g. \( \tilde{\gamma} \) could be a lift defined by the same process using a different sufficiently fine partition). By the uniqueness shown above, we must have \( \tilde{\gamma} = \gamma \) on \([0,t_1]\), then on \([t_1,t_2]\), and ultimately on all of \([0,1]\).

XI.20.2.3. The lift \( \tilde{\gamma} \) depends on the choice of \( s_0 \). But if the \( s_0 \) is changed, necessarily by addition of an integer \( n \), then the lift is simply translated by \( n \) (the translation by \( n \) is obviously a lift with the right properties, so must be the unique one).

XI.20.2.4. We will want a similar result for homotopies, i.e. maps from \([0,1]^2\) to \( S^1 \). If \( h \) is a map from \([0,1]^2\) to \( S^1 \), and \( z_0 = h(0) \), and \( s_0 \in \mathbb{R} \) with \( p(s_0) = z_0 \), then by XI.20.2.2. there is a unique lifting \( \tilde{h} : [0,1]^2 \to \mathbb{R} \) with \( \tilde{h}(0) = s_0 \) which is continuous on each line segment in \([0,1]^2\) beginning at \( 0 \), defined for each \( x \in [0,1]^2 \) by lifting the path \( \gamma(t) = h(tx) \). However, it is not obvious that this map is continuous on \([0,1]^2\) (this is roughly the same question as whether the same value for \( \tilde{h}(x) \) is obtained by lifting any other path in \([0,1]^2\) from \( 0 \) to \( x \).) The lifted map \( \tilde{h} \) is continuous in a neighborhood of \( 0 \) since all the lifts are done using \( p^{-1}_{s_0} \), which is continuous; it is possible, but tricky, to expand successively through larger regions to the whole square (cf. \( \cdot \)).

We will take an alternate approach, first due to A. W. Tucker [Tuc46] (cf. [Spa81]).

1278
XI.20.2.5. There is a “multiplication” in \( S^1 \): if \( z = (\cos 2\pi t, \sin 2\pi t) \), write \( z \) symbolically as \( e^{2\pi it} \) (this is actually more than symbolic; cf. (1)). The \( t \) is not unique, but for a given \( z \) any two such \( t \) differ by an integer. If \( w = e^{2\pi is} \), define \( zw = e^{2\pi i(t+s)} \). Although \( t \) and \( s \) are only determined up to adding an integer, and thus \( t+s \) is only determined up to an integer, the point \( e^{2\pi i(t+s)} \) is well defined in \( S^1 \). This multiplication is associative and commutative. The point \( 1 = (1,0) = p(0) \) is an identity for this multiplication, and the point \( z^{-1} = e^{-2\pi it} \) is an inverse for \( z = e^{2\pi it} \). Multiplication is continuous as a function from \( S^1 \times S^1 \) to \( S^1 \), inversion (which geometrically is just reflection across the \( x \)-axis) is also continuous, and \( p \) has the property that \( p(t+s) = p(t)p(s) \) for any \( t, s \in \mathbb{R} \) (\( p \) is a homomorphism).

XI.20.2.6. Now suppose \( h : [0,1]^2 \to S^1 \) is a continuous function, with \( h(0) = z_0 \), and \( s_0 \in \mathbb{R} \) with \( p(s_0) = z_0 \). By uniform continuity there is an \( \epsilon > 0 \) such that \( \|h(x) - h(y)\| < 2 \) whenever \( x, y \in [0,1]^2 \) and \( \|x - y\| < \epsilon \). Fix \( n \in \mathbb{N} \) with \( n > \sqrt{\frac{2}{\epsilon}} \); then \( \|\frac{x}{n}\| < \epsilon \) for all \( x \in [0,1]^2 \). For \( x \in [0,1]^2 \) and \( 1 \leq k \leq n \), define

\[
g_k(x) = h \left( \frac{k}{n} x \right) h \left( \frac{k-1}{n} x \right)^{-1}.
\]

Then \( g_k \) is a continuous function from \( [0,1]^2 \) to \( S^1 \), and

\[
h(x) = z_0 g_1(x) g_2(x) \cdots g_n(x).
\]

The range of \( g_k \) is contained in \( S^1 \setminus \{1\} \): since \( \|\frac{k}{n} x - \frac{k-1}{n} x\| < \epsilon \),

\[
\left\| h \left( \frac{k}{n} x \right) - h \left( \frac{k-1}{n} x \right) \right\| = \left\| h \left( \frac{k}{n} x \right) h \left( \frac{k-1}{n} x \right)^{-1} - 1 \right\| < 2.
\]

So there is a lift \( \tilde{g}_k : [0,1]^2 \to \mathbb{R} \) defined by

\[
\tilde{g}_k(x) = p^{-1}_0(g_k(x))
\]

which is continuous. Set

\[
\tilde{h}(x) = s_0 + \tilde{g}_1(x) + \cdots + \tilde{g}_n(x).
\]

then \( \tilde{h} \) is a finite sum of continuous functions, hence continuous, and is a lift of \( h \). By the uniqueness described in XI.20.2.4., this is the unique continuous lift of \( h \) with \( h(0) = s_0 \). Thus we have proved:

XI.20.2.7. Lemma. [Homotopy Lifting Lemma] Let \( h : [0,1]^2 \to S^1 \) be a continuous function, with \( h(0) = z_0 \), and \( s_0 \in \mathbb{R} \) with \( p(s_0) = z_0 \). Then there is a unique continuous function \( \tilde{h} : [0,1]^2 \to \mathbb{R} \) such that \( \tilde{h}(0) = s_0 \) and \( h = p \circ \tilde{h} \).

Note that the proof actually shows that the same is true for functions from \( [0,1]^n \) to \( S^1 \) for any \( n \).

XI.20.2.8. A loop in \( S^1 \) is a path \( \gamma : [0,1] \to S^1 \) with \( \gamma(1) = \gamma(0) \). A loop in \( S^1 \) is effectively the same thing as a continuous function from \( S^1 \) to \( S^1 \): the restriction \( q \) of \( p \) to \( [0,1] \) is a bijection, and although the inverse map \( q^{-1} \) is not continuous, if \( \gamma \) is a loop in \( S^1 \) the map \( f = \gamma \circ q^{-1} \) is a continuous function from \( S^1 \) to \( S^1 \), and conversely if \( f : S^1 \to S^1 \) is continuous, \( \gamma = f \circ q \) (extended by setting \( \gamma(1) = \gamma(0) \)) is a loop in \( S^1 \).
XI.20.2.9. If \( \gamma \) is a loop in \( S^1 \) with \( \gamma(0) = \gamma(1) = z_0 \), and \( s_0 \in \mathbb{R} \) with \( p(s_0) = z_0 \), and \( \tilde{\gamma} \) is the lift of \( \gamma \) to \( \mathbb{R} \), then it is not generally true that \( \tilde{\gamma}(1) = \tilde{\gamma}(0) \). But \( \tilde{\gamma}(1) - \tilde{\gamma}(0) \) is an integer which does not depend on the choice of \( s_0 \) by XI.20.2.3., called the index or degree of \( \gamma \). Thus each continuous function from \( S^1 \) to \( S^1 \) has a well-defined index.

XI.20.2.10. Proposition. Homotopic maps from \( S^1 \) to \( S^1 \) have the same index.

Proof: Let \( f_0 \) and \( f_1 \) be homotopic maps from \( S^1 \) to \( S^1 \) via maps \( f_s \), and \( \gamma_s \) the loop corresponding to \( f_s \). Then \( h(t, s) = \gamma_s(t) \) is a continuous map from \([0, 1]^2 \) to \( S^1 \). Take a lift \( \tilde{h} \) of \( h \) to a continuous function from \([0, 1]^2 \) to \( \mathbb{R} \) (XI.20.2.7.); then

\[ \phi(s) = \tilde{h}(1, s) - \tilde{h}(0, s) \]

is the index of \( f_s \), and is a continuous function of \( s \). Since it takes only integer values, it must be constant, i.e. the index of \( f_0 \) equals the index of \( f_1 \).

XI.20.2.11. Theorem. The circle \( S^1 \) is not contractible; thus the Brouwer Fixed-Point Theorem holds for \( n = 2 \).

Proof: The identity map on \( S^1 \) has index 1, and a constant map has index 0, so they cannot be homotopic.

XI.20.2.12. The converse of XI.20.2.10. is true (): maps from \( S^1 \) to \( S^1 \) with the same index are homotopic; so we get a parametrization of the homotopy classes of maps from \( S^1 \) to \( S^1 \). It is easy to get a loop with arbitrary index: set \( \gamma_n(t) = p(nt) \). Thus the index map from homotopy classes of maps to \( \mathbb{Z} \) is a bijection. This set of homotopy classes can be made into a group (the fundamental group) \( \pi_1(S^1) \) via concatenation, and \( \pi_1(S^1) \cong \mathbb{Z} \). See () for the general theory of fundamental groups.

XI.20.2.13. The map \( p : \mathbb{R} \to S^1 \) is called the universal covering map, and \( \mathbb{R} \) is the universal cover of \( S^1 \). See () for the general theory of covering maps.

XI.20.3. The General Case

If \( n \geq 3 \), the Brouwer Fixed-Point Theorem for \( n \) is more difficult than for \( n = 2 \) since there is no map from \( \mathbb{R}^{n-1} \) onto \( S^{n-1} \) analogous to the map \( p \) (\( S^{n-1} \) is simply connected ())). So one must proceed differently. One method of proof amounts to again defining a degree (index) which is an integer for every map from \( S^{n-1} \) to \( S^{n-1} \), and proving that homotopic maps have the same degree, and that the identity map has degree 1 and a constant map has degree 0.

The proof must have a combinatorial component of some sort. The standard argument requires working with triangulations (()) and can be phrased in several ways which are all more or less equivalent:

(1) The most direct argument involves a combinatorial result called Sperner’s Lemma. See e.g. (), (), [Spa81, p. 151], . . . for expositions of this argument.
(2) One can develop abstract homology theory up through the Mayer-Vietoris sequence, and use this sequence to give a proof of the degree formula and therefore the No-Retraction Theorem by induction. See e.g. (), (), (), ... for this approach.

(3) The homology argument of (2) can be written out directly in the case of spheres without explicit use of the term “homology.” See [Dug78] or () for this approach.

The Sperner’s Lemma argument actually proves the following more general result (cf. [Spa81, p. 151]):

**XI.20.3.1. Theorem.** Let $X$ be a compact manifold with boundary which can be triangulated (e.g. a smooth compact manifold with boundary). Then there is no retraction of $X$ onto $\partial X$.

As a variation of this result, the following version follows immediately from the Brouwer Fixed-Point Theorem:

**XI.20.3.2. Theorem.** Let $U$ be a bounded open set in $\mathbb{R}^n$. Then there is no retraction from $\bar{U}$ onto $\partial U$.

**Proof:** Suppose $r$ is a retraction of $\bar{U}$ onto $\partial U$. Fix $a \in U$, and $m$ so that $B = B_m(a)$ contains $\bar{U}$. Let $s$ be radial retraction of $B \setminus \{a\}$ onto $\partial B \cong S^{n-1}$. Define, for $x \in B$,

$$f(x) = \begin{cases} s(r(x)) & \text{if } x \in \bar{U} \\ s(x) & \text{if } x \in B \setminus \bar{U} \end{cases}.$$  

It is easy to check that $f$ is continuous, since $\partial (B \setminus \bar{U}) \subseteq \partial U \cup \partial B$. Then $f$ is a retraction of $B$ onto $\partial B$, contradicting the No-Retraction Theorem.

**XI.20.3.3.** This result can fail if $U$ is not bounded, e.g. if $U$ is the open upper half-plane in $\mathbb{R}^2$.

**XI.20.3.4.** We will not give a direct proof of the Brouwer Fixed-Point Theorem for $n \geq 3$ here. We will deduce the theorem from some analysis results, including Sard’s Theorem and the Implicit Function Theorem ().

**XI.20.4. Exercises**

**XI.20.4.1. Fundamental Theorem of Algebra.** Here is a winding-number proof of the Fundamental Theorem of Algebra (X.1.4.2., X.8.2.3., X.8.2.4.); cf. [vM89]. Suppose

$$f(z) = a_0 + a_1 z + \cdots + a_n z^n$$

is a nonconstant polynomial of degree $n$ (i.e. $a_n \neq 0$) with complex coefficients, with no roots in $\mathbb{C}$. Then $a_n \neq 0$.

(a) Set

$$g(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n.$$  

1281
Show that $g(z) = z^n f\left(\frac{1}{z}\right)$ for $z \neq 0$ and $g(0) = a_n$. Thus $g$ also has no roots in $\mathbb{C}$.

(b) Let $r$ be radial retraction of $\mathbb{C} \setminus \{0\}$ onto $S^1$, and define $h : [0,1] \times S^1 \to S^1$ by

$$h(t, z) = r\left(\frac{f(tz)}{g\left(\frac{1}{z}\right)}\right).$$

Show that $h$ is a homotopy between a constant function and $\phi(z) = z^n$, a contradiction since $n > 0$. 

1282
XI.20.5. Proof of the No-Retraction Theorem

In this section, we give two nice analysis proofs of the No-Retraction Theorem, which give a proof of the general Brouwer Fixed-Point Theorem. The first proof was found by Milnor [Mil78] and simplified by Rogers [Rog80]. The second proof is due to Hirsch [Hir94].

We first reduce the problem to the smooth case. For the first proof, we will only need the following result (and not even all of it) for $C^1$, but it is no harder to prove for $C^\infty$, and the second proof needs the $C^\infty$ result. We use the notation $B^n$ for the closed unit ball in $\mathbb{R}^n$, $S^{n-1}$ its boundary, the $(n-1)$-sphere, and $\mathbb{D}^n$ the open unit ball in $\mathbb{R}^n$.

XI.20.5.1. Lemma. If there is a (continuous) retraction of $B^n$ onto $S^{n-1}$, then there is a $C^\infty$ retraction. In fact, there is a $C^\infty$ retraction which is radial projection on a neighborhood of $S^{n-1}$.

Proof: Let $s$ be a retraction. Since $B^n$ is compact, $s$ is uniformly continuous, so there is a $\delta$, $0 < \delta < \frac{1}{4}$, such that $\|s(x) - s(y)\| < \frac{1}{4}$ whenever $x, y \in B^n$ and $\|x - y\| < \delta$. Then if $x \in B^n$ and $\|x\| > 1 - \delta$, we have

$$\|s(x) - x\| \leq \|s(x) - s\left(\frac{x}{\|x\|}\right)\| + \|s\left(\frac{x}{\|x\|}\right) - x\|$$

$$= \left\|s(x) - \frac{x}{\|x\|}\right\| + \left\|\frac{x}{\|x\|} - x\right\| < \frac{1}{4} + \delta < \frac{1}{2}.$$ 

By the General Weierstrass Approximation Theorem (XV.8.3.3.), or () , there is a $C^\infty$ function $f : B^n \to \mathbb{R}^n$ (in fact the coordinate functions can be polynomials) with $\|f - s\|_\infty < \frac{1}{4}$. Then, for $x \in B^n$, $\|x\| > 1 - \delta$, we have

$$\|f(x) - x\| \leq \|f(x) - s(x)\| + \|s(x) - x\| < \frac{1}{4} + \frac{1}{2} = \frac{3}{4}.$$ 

Let $h$ be a $C^\infty$ function from $[0,1]$ to $[0,1]$ which is identically 0 on $[0,1 - \delta]$ and identically 1 on $[1 - \delta,1]$ (). For $x \in B^n$, define

$$g(x) = h(\|x\|) x + (1 - h(\|x\|)) f(x).$$

Then $g$ is $C^\infty$, $g(x) = x$ if $\|x\| \geq 1 - \frac{\delta}{2}$, and $g(x) = f(x)$ if $\|x\| \leq 1 - \delta$.

If $\|x\| \leq 1 - \delta$, we have $\|g(x) - s(x)\| = \|f(x) - s(x)\| < \frac{1}{4}$, and since $\|s(x)\| = 1$, we have $\|g(x)\| > \frac{3}{4}$, and in particular $g(x) \neq 0$. And if $\|x\| > 1 - \delta$, then $g(x)$ lies on the line segment between $x$ and $f(x)$; since

$$\|f(x) - x\| < \frac{3}{4} < 1 - \delta < \|x\|$$

we again have that $g(x) \neq 0$. Thus

$$r(x) = \frac{g(x)}{\|g(x)\|}$$

is a $C^\infty$ retraction of $B^n$ onto $S^{n-1}$ (note that $\phi(x) = \|x\|$ is $C^\infty$ on $\mathbb{R}^n \setminus \{0\}$). We have $r(x) = \frac{x}{\|x\|}$ if $\|x\| \geq 1 - \frac{\delta}{2}$.

Thus the No-Retraction Theorem reduces to:

1283
XI.20.5.2. **Theorem.** Let $n \in \mathbb{N}$. Then there is no $C^\infty$ retraction from $B^n$ to $S^{n-1}$ which is radial projection on a neighborhood of $S^{n-1}$.

The First Proof

XI.20.5.3. **To prove the theorem, suppose $r$ is a $C^1$ retraction.** Set $g(x) = r(x) - x$ and, for $0 \leq t \leq 1$ set

$$f_t(x) = (1-t)x + tr(x) = x + tg(x)$$

for $x \in B^n$. Then $g$ and each $f_t$ is $C^1$, $f_t(x) \in B^n$ for all $t$ and $x \in B^n$, and $f_1(x) = x$ for all $t$ and $x \in S^{n-1}$.

Since $B^n$ is compact, $Dg$ is bounded on $B^n$, and hence there is a constant $K > 0$ such that

$$\|g(x) - g(y)\| \leq K\|x - y\|$$

for all $x, y \in B^n$. If $0 \leq t < \frac{1}{K}$ and $x, y \in B^n$ with $f_t(x) = f_t(y)$, then

$$x - y = t(g(x) - g(y))$$

and since $Kt < 1$, $x = y$. Thus $f_t$ is one-to-one if $0 \leq t < \frac{1}{K}$.

We have $Df_t(x) = I + tDg(x)$ for all $t$ and $x$. Since $Dg$ is bounded on $B^n$, there is a constant $t_0 > 0$, which we may take to be $< \frac{1}{K}$, such that $\text{Det}(Df_t(x)) > 0$ for all $t$, $0 \leq t \leq t_0$, and all $x \in B^n$.

By the Inverse Function Theorem, $f_t$ is an injective open mapping on $\mathbb{D}^n$ for $0 \leq t \leq t_0$. Let $U_t = f_t(\mathbb{D}^n)$; then $U_t$ is an open set in $\mathbb{R}^n$ contained in $B^n$, hence $U_t \subseteq \mathbb{D}^n$.

XI.20.5.4. **Lemma.** Let $f$ be a continuous function from $B^n$ to $B^n$. Suppose $f(S^{n-1}) \subseteq S^{n-1}$ and that $f|_{\mathbb{D}^n}$ is an open mapping (as a map to $\mathbb{R}^n$). Then $f$ maps $\mathbb{D}^n$ onto $\mathbb{D}^n$ and $B^n$ onto $B^n$.

**Proof:** Since $f$ is an open mapping on $\mathbb{D}^n$, we have $f(\mathbb{D}^n) \subseteq \mathbb{D}^n$. If $f(\mathbb{D}^n)$ is a proper subset of $\mathbb{D}^n$, there is a $y \in \mathbb{D}^n$ which is in $f(\mathbb{D}^n) \setminus f(\mathbb{D}^n)$. But $f(B^n)$ is a compact set containing $f(\mathbb{D}^n)$, so $y \in B^n$, and hence $y = f(x)$ for some $x \in B^n \setminus \mathbb{D}^n = S^{n-1}$. But $f(S^{n-1}) \subseteq S^{n-1}$, a contradiction. Thus $f(\mathbb{D}^n) = \mathbb{D}^n$. Then $f(B^n)$ is a closed set containing $\mathbb{D}^n$, so $f(B^n) = B^n$.

XI.20.5.5. Continuing with the proof of the theorem, we conclude from the lemma that $U_t = \mathbb{D}^n$ and $f_t(B^n) = B^n$ for $0 \leq t \leq t_0$.

Now consider the function

$$\phi(t) = \int_{B^n} \text{Det}(Df_t(x)) \, dx = \int_{B^n} \text{Det}(I + tDg(x)) \, dx.$$ 

This $\phi(t)$ is a polynomial (of degree $\leq n$) in $t$. If $0 \leq t \leq t_0$, then $f_1$ is a $C^1$ diffeomorphism from $B^n$ to itself, and thus by the change-of-variables theorem () we have $\phi(t) = \text{Vol}(B^n)$ for $0 \leq t \leq t_0$. Thus $\phi$ is constant on $[0, t_0]$, and therefore constant everywhere, and nonzero.

However, $f_1$ maps $B^n$ into $S^{n-1}$ and thus has rank $\leq n - 1$ at every point, so $\text{Det}(Df_1(x))$ is zero for all $x$, and hence $\phi(1) = 0$, a contradiction.

This completes the proof of Theorem XI.20.5.2. ✇

1284
The Second Proof

The second proof is much shorter (given XI.20.5.1.), but uses Sard’s Theorem, and thus is not as elementary as the first proof.

XI.20.5.6. Let $r$ be a $C^\infty$ retraction from $B^n$ to $S^{n-1}$ which is radial projection on a neighborhood of $S^{n-1}$. By Sard’s Theorem (the version in XI.23.6.3. suffices), there is a regular value $y \in S^{n-1}$ (in fact a dense set of regular values, but we only need one!) Let $X = r^{-1}(\{y\})$. Then $X$ consists of regular points of $r$, hence is a one-dimensional compact manifold with boundary consisting of a radial line segment in from $y$ and possibly some additional points in $D^n$ (cf. VIII.6.7.1.), and $X \cap D^n$ is a manifold without boundary, i.e. $\partial X = \{y\}$. But a compact 1-manifold with boundary is a finite separated union of copies of $[0, 1]$ and $S^1$, and thus has an even number of boundary points, a contradiction.

This completes the proof of Theorem XI.20.5.2..
XI.21. Topological Dimension Theory

It has been a long and occasionally painful process to give a good definition of the “dimension” of a topological space. There are many reasonable definitions of dimension, and it is a remarkable and significant fact that they all give the same numerical value for “reasonable” topological spaces, including separable metrizable spaces. In this section, we describe some of the principal definitions and their properties. Although we give definitions applicable to all topological spaces, we will concentrate on their properties only in the separable metrizable case.

One obvious property any reasonable dimension theory must have is that the dimension of $\mathbb{R}^n$ should be $n$. This is true for all the definitions we will consider, but it is surprisingly hard to prove for most of them.

Some standard references for Dimension Theory are [Men28], [HW41], [Nag65], [Nag70], [AP73], [Pea75], [Eng95]. The first is mainly just of historical value today, but the others are worth reading for content, especially the last. Of course, newer references contain more complete results.

XI.21.1. Essential Dimension

XI.21.1.1. Definition. Let $X$ be a topological space. The essential dimension $ed(X)$ of $X$ is the largest $n$ such that there is an essential map () from $X$ to $B^n$. We say $ed(X) = 1$ if there is an essential map from $X$ to $B^n$ for all $n$, and $ed(X) = 0$ if every map from $X$ to $B^1 = [-1, 1]$ is inessential. Set $ed(\emptyset) = 1$.

The relevance of this definition is the following result, which is essentially equivalent to the Brouwer Fixed-Point Theorem:

XI.21.1.2. Theorem. We have $ed(\mathbb{R}^n) = ed(B^n) = ed(S^n) = n$.

See XI.21.1.5. for a more general result.

The next fact also justifies the form of the definition.

XI.21.1.3. Proposition. Let $X$ be a topological space. If there is an essential map from $X$ to $B^n$, then there is an essential map from $X$ to $B^m$ for any $m < n$.

Proof: Let $f : X \to B^n$ be an essential map, with coordinate functions $(f_1, \ldots, f_n)$. Fix $m < n$, and define $\tilde{f} : X \to B^m$ by

$$\tilde{f}(x) = (f_1(x), \ldots, f_m(x)).$$

Suppose $\tilde{f}$ is inessential. For $\epsilon > 0$, let $\tilde{g} : X \to B^m$ with $\|\tilde{g} - \tilde{f}\|_\infty < \epsilon$ and $0 \notin \tilde{g}(X)$. Let $(g_1, \ldots, g_m)$ be the coordinate functions of $\tilde{g}$. Define $g : X \to \mathbb{R}^n$ by

$$g(x) = (g_1(x), \ldots, g_m(x), f_{m+1}(x), \ldots, f_n(x)).$$

Then $\|g - \tilde{f}\|_\infty < \epsilon$ and $0 \notin g(X)$. The function $g$ does not necessarily map $X$ into $B^n$, but we can retract the part that does not: define $h$ by

$$h(x) = \begin{cases} g(x) & \text{if } \|g(x)\| \leq 1 \\ \frac{g(x)}{\|g(x)\|} & \text{if } \|g(x)\| > 1 \end{cases}.$$
Then $h$ is a map from $X$ to $B^n$, $\|h - f\|_\infty < \epsilon$, and $0 \notin h(X)$. Since $\epsilon > 0$ is arbitrary, this contradicts that $f$ is essential. Thus $f$ is essential.

**XI.21.1.4.** Proposition. Let $X$ be a topological space, and $f : X \rightarrow B^n$ a map. Then $f$ is essential if and only if there is an $\alpha > 0$ such that for every $g : X \rightarrow B^n$ sufficiently uniformly close to $f$, $g(X)$ contains the ball $B^n_\alpha$ of radius $\alpha$ around $0$.

**Proof:** It is obvious that any $f$ with this property is essential. Conversely, suppose $f$ is essential. Fix $0 < \epsilon < 1$ for which there is no $h$ with $\|f - h\|_\infty < \epsilon$ and $0 \notin h(X)$. Let $g : X \rightarrow B^n$ with $\|g - f\| < \frac{\epsilon}{2}$, and suppose there is an $x_0$ with $\|x_0\| < \frac{\epsilon}{2}$ and $x_0 \notin g(X)$. Define $h : B^n \rightarrow B^n$ by

$$h(x) = \begin{cases} x - x_0 & \text{if } \|x - x_0\| \leq 1 \\ \frac{x - x_0}{\|x - x_0\|} & \text{if } \|x - x_0\| > 1 \end{cases}.$$ 

Then $h(x_0) = 0$, and $x_0$ is the only $x$ with $h(x) = 0$; and $\|h - i\|_\infty < \frac{\epsilon}{2}$, where $i$ is the identity map on $B^n$, so $\|h \circ g - f\|_\infty < \epsilon$ and $0 \notin h \circ g(X)$, a contradiction.

We now prove a general result that includes XI.21.1.2..

**XI.21.1.5.** Theorem. Let $X$ be a smooth $n$-manifold with boundary. Then $ed(X) = n$.

**Proof:** $X$ contains a closed subset $Y$ homeomorphic to $B^n$, and there is a retraction $r$ from $X$ onto $Y$ (this requires the Tietze Extension Theorem () in general, but if $X = B^n$, $\mathbb{R}^n$, or $S^n$, such a $Y$ and $r$ can be given explicitly without resort to the Tietze Extension Theorem). Let $f$ be a homeomorphism from $Y$ onto $B^n$. Then $f$ is an essential map by XI.20.1.16., so $f \circ r$ is an essential map from $X$ to $B^n$ [if $g$ approximates $f \circ r$, and $i$ is the embedding of $Y$ into $X$, then $g \circ i$ approximates $f \circ r \circ i = f$] and $ed(X) \geq n$.

On the other hand, suppose $m > n$ and $f$ is a map from $X$ to $B^m$. To show that $f$ is inessential, it suffices to show that the restriction of $f$ to each component of $X$ is inessential, since the components are clopen and $f$ can be modified independently on each component. Thus we may assume $X$ is connected, and in particular second countable. Then $f$ can be arbitrarily closely approximated by a smooth function $g$, and by () the range of $g$ has dense complement. Thus $f$ is inessential by XI.21.1.4. and $ed(X) < m$.

**XI.21.1.6.** The smooth structure was used only to prove $ed(X) \leq n$; the first half of the proof applies to any topological $n$-manifold with boundary (and in fact to any normal topological space $X$ containing a copy of $B_n$) to give $ed(X) \geq n$. Note that the conclusion is purely topological, and the smooth structure is used only to give a simple way of approximating continuous functions by continuous functions whose range is “small.” The result also holds for manifolds which do not have differentiable structures, but requires a different proof in this case (Exercise XI.21.8.3.). An alternate proof that $ed(B^n) \leq n$ can be given using XI.21.4.14. and Exercise XI.21.8.2.

**XI.21.1.7.** Corollary. Let $X$ be a smooth $n$-manifold with boundary, and $Y$ a smooth $m$-manifold with boundary. If $n \neq m$, then $X$ and $Y$ are not homeomorphic. In particular, if $n \neq m$, $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$, or to any open subset of $\mathbb{R}^m$.

1287
Extending Maps to Spheres

Essential dimension can be characterized also by an extension property for maps to spheres:

**XI.21.1.8. Theorem.** Let $X$ be a completely regular space. Then $\text{ed}(X) \leq n$ if and only if, whenever $Y$ is a closed subset of $X$ and $f : Y \to S^n$ is continuous, then $f$ extends to a continuous map $g : X \to S^n$.

**Proof:** Suppose $\text{ed}(X) \leq n$, $Y$ is a closed subspace of $X$, and $f : Y \to S^n$. Regard $f$ as a map from $Y$ to $B^{n+1}$. By the Tietze Extension Theorem (), $f$ extends to a map $h$ from $X$ to $B^{n+1}$. Then $h$ is inessential, so by an argument almost identical to the proof of (), there is a $\phi : B^{n+1} \to B^{n+1}$ which is the identity on a neighborhood of $S^n$ and for which $h = \phi \circ h$ satisfies $0 \notin h(X)$. Then $h|_Y = h|_Y$, so $h$ is also an extension of $f$. If $r$ is the radial retraction of $B^{n+1} \setminus \{0\}$ onto $S^n$, then $g = r \circ h$ is an extension of $f$ to $g : X \to S^n$.

Conversely, suppose $X$ has the extension property, and suppose $f : X \to B^{n+1}$. Fix $\epsilon > 0$, and let $Y = f^{-1}(S^n)$ (where $S^n = \{x \in B^{n+1} : \|x\| = \epsilon \}$). Then $Y$ is a closed subset of $X$, so $f|_Y$ extends to a map $h$ from $X$ to $S^n$. Define $g : X \to B^{n+1}$ by

$$g(x) = \begin{cases} h(x) & \text{if } \|f(x)\| \leq \epsilon \\ f(x) & \text{if } \|f(x)\| \geq \epsilon \end{cases}.$$  

Then $g$ is continuous, $\|g - f\|_{\infty} \leq 2\epsilon$, and $0 \notin g(X)$. Thus $f$ is inessential and $\text{ed}(X) \leq n$. 

A trick gives the following interesting and useful variant:

**XI.21.1.9. Corollary.** Let $X$ be a normal space, $Y$ a closed subspace, and $f : Y \to S^n$ a map. If $\text{ed}(X \setminus Y) \leq n$, then $f$ extends to a map from $X$ to $S^n$.

**Proof:** By (), $f$ extends to a neighborhood $U$ of $Y$ in $X$. Since $X$ is normal, there is a neighborhood $V$ of $Y$ in $X$ with $\bar{V} \subseteq U$. Thus $f$ extends to $\tilde{f} : \bar{V} \to S^n$. Then $Z = (X \setminus Y) \cap \bar{V}$ is a closed subspace of $X \setminus Y$, and $g = \tilde{f}|_Z$ is a map from $Z$ to $S^n$. By **XI.21.1.8.**, $g$ extends to $\tilde{g} : X \setminus Y \to S^n$. Then $\tilde{f}$ and $\tilde{g}$ define a map from $X$ to $S^n$ extending $f$.

**Essential Dimension of Subspaces**

**XI.21.1.10.** If $X$ is a topological space and $Y$ is a subspace of $X$, one would expect that $\text{ed}(Y) \leq \text{ed}(X)$. However, this turns out to be false in general, even if $X$ is a compact Hausdorff space (). The problem is that a map from $Y$ to $B^n$ need not extend to $X$ in general. It is true in general that $\text{ed}(Y) \leq \text{ed}(X)$ if $X$ is separable and metrizable (). And if $Y$ is closed, it is always true, and an immediate consequence of **XI.21.1.8.**:

**XI.21.1.11. Corollary.** Let $X$ be a completely regular space, and $Y$ a closed subspace of $X$. Then $\text{ed}(Y) \leq \text{ed}(X)$.

**Proof:** Suppose $\text{ed}(X) = n$. Let $Z$ be a closed subset of $Y$, and $f : Z \to S^n$ a map. Then $Z$ is closed in $X$, so by **XI.21.1.8.** $f$ extends to a map $g : X \to S^n$. Then $g|_Y$ is an extension of $f$ to $Y$. Thus $\text{ed}(Y) \leq n$ by the other direction of **XI.21.1.8.**.

1288
XI.21.1.12. PROPOSITION. Let $X$ be a completely regular space. Then $ed(X) = ed(\beta X)$, where $\beta X$ is the Stone–Čech compactification () of $X$.

PROOF: Let $f$ be a map from $X$ to $B^n$, and $\bar{f}$ its extension to $\beta X$. If $g : X \to B^n$, then $\|\bar{g} - \bar{f}\|_{\infty} = \|g - f\|_{\infty}$; thus if $f$ is essential, $\bar{f}$ is also essential. On the other hand, if $f$ is inessential, then for every $\epsilon > 0$ there is a $g : X \to B^n$ with $\|g - f\|_{\infty} < \epsilon$ and $g(X)$ bounded away from $0$ (XI.21.1.4); thus $0 \notin \bar{g}(\beta X)$, so $\bar{f}$ is also inessential.

Essential Dimension of Subspaces of $\mathbb{R}^n$

XI.21.1.13. THEOREM. Let $X$ be a subset of $\mathbb{R}^n$. Then $ed(X) \leq n$, and $ed(X) = n$ if and only if $X$ has nonempty interior.

XI.21.2. Invariance of Domain

We can now prove Brouwer’s Invariance of Domain Theorem:

XI.21.2.1. THEOREM. [INVARIANCE OF DOMAIN] Let $U$ be a nonempty open subset of $\mathbb{R}^n$, and $f$ a one-to-one continuous function from $U$ to $\mathbb{R}^m$. Then

(i) $m \geq n$.

(ii) If $m = n$, then $f(V)$ is open in $\mathbb{R}^n$ for every open subset $V$ of $U$ (in particular, $f(U)$ is open in $\mathbb{R}^n$), i.e. $f$ is an open mapping and a homeomorphism onto its range.

This will be a consequence of the following:

XI.21.2.2. THEOREM. Let $X$ be a subset of $\mathbb{R}^n$, and $f$ a one-to-one continuous function from $X$ onto a subset $Y$ of $\mathbb{R}^n$. If $x$ is an interior point of $X$, then $f(x)$ is an interior point of $Y$.

XI.21.2.3. A point $x \in X$ is an interior point of $X$ if and only if it is an interior point of a compact subset of $X$: if $x$ is an interior point of $X$, then $x$ is an interior point of a sufficiently small closed ball around $x$ which is contained in $X$, and the converse is obvious. Thus it suffices to prove XI.21.2.2. for $X$ compact. And in this case, the map $f$ is automatically a homeomorphism.

Thus it suffices to prove that interior points of a compact subset $X$ of $\mathbb{R}^n$ can be characterized topologically, which is the substance of the next result:
**XI.21.2.4. Lemma.** Let $X$ be a compact subset of $\mathbb{R}^n$, and $x \in X$. Then $x \in \partial X$ if and only if, whenever $U$ is a neighborhood of $x$ in $X$, there is an open neighborhood $V$ of $x$ in $X$ contained in $U$ such that every map from $X \setminus V$ to $S^{n-1}$ extends to a map from $X$ to $S^{n-1}$.

**Proof:** Suppose $x \in \partial X$, and $U$ is a neighborhood of $x$ in $X$. Then there is an open ball $W$ around $x$ in $\mathbb{R}^n$ such that $W \cap X \subseteq U$. Set $V = W \cap X$. Let $f$ be a map from $X \setminus V$ to $S^{n-1}$. Then $\partial W$ is homeomorphic to $S^{n-1}$, and $X \setminus V$ is compact and hence closed in $(X \setminus V) \cup \partial W$. The complement $Z = [(X \setminus V) \cup \partial W] \setminus (X \setminus V)$ is an open subset of $\partial W$, hence an $(n-1)$-manifold with a smooth structure, and thus $ed(Z) = n-1$ by XI.21.1.5. (only the “easy” part $ed(Z) \leq n-1$ is needed). So by XI.21.1.9. $f$ extends to a map $\tilde{f}$ from $(X \setminus V) \cup \partial W$ to $S^{n-1}$. Since $x \in \partial X$, there is an $x_0 \in W$ with $x_0 \notin X$, and there is a retraction $r$ from $W \setminus \{x_0\}$ to $\partial W$. Define $g : X \to S^{n-1}$ by

$$g(y) = \begin{cases} \tilde{f}(r(y)) & \text{if } y \in V \\ f(y) & \text{if } y \in X \setminus V \end{cases}.$$  

Then $g : X \to S^{n-1}$ is an extension of $f$ to $X$.

Conversely, suppose $x$ is an interior point of $X$, and $\epsilon$ is small enough that the closed ball $B$ of radius $\epsilon$ around $x$ is contained in $X$. Let $U$ be the open ball of radius $\epsilon$ centered at $x$; then $\partial U = \partial B$ is homeomorphic to $S^{n-1}$, and $X \setminus U$ can be radially projected onto $\partial U$. Call this map $f$. If $V$ is any neighborhood of $x$ contained in $U$, then $f$ can be extended to a map from $X \setminus V$ to $\partial W$ using radial projection from $x$. But $f$ cannot be extended to a map $g$ from $X$ to $\partial U$, since the restriction of $g$ to $B$ would be a retraction of $B$ onto $\partial B$. 

---

**XI.21.3. Other Dimension Functions**

For analysis purposes, essential dimension seems to be the most useful dimension function on spaces. However, it is not commonly used by topologists. There are three standard dimension functions used by topologists: small inductive dimension, large inductive dimension, and covering dimension.

**Small Inductive Dimension**

**XI.21.3.1.** Small inductive dimension $\text{ind}$ is the earliest and probably still the most-used dimension function on spaces. The definition is motivated by the idea that the boundary of a “nice” region of $\mathbb{R}^n$ should have dimension $n-1$. The definition is (surprise!) an inductive one: we successively define what it means for a space to have dimension $\leq n$.

**XI.21.3.2. Definition.** Let $X$ be a topological space. Then

(i) $\text{ind}(X) = -1$ if and only if $X = \emptyset$.

(ii) $\text{ind}(X) \leq n$ if and only if, whenever $x \in X$ and $U$ is an open neighborhood of $x$ in $X$, there is an open neighborhood $V$ of $x$ contained in $U$ with $\text{ind}(\partial V) \leq n - 1$.

(iii) $\text{ind}(X) = n$ if $\text{ind}(X) \leq n$ and it is false that $\text{ind}(X) \leq n - 1$.

(iv) $\text{ind}(X) = \infty$ if for every $n$ it is false that $\text{ind}(X) \leq n$.
XI.21.3.3. Thus, for example, a topological space \( X \) satisfies \( \text{ind}(X) = 0 \) if and only if it is nonempty and, for each \( x \in X \) and each open neighborhood \( U \) of \( x \), there is an open neighborhood \( V \) of \( x \) contained in \( U \) with \( \partial V = \emptyset \). But a subset of \( X \) has empty boundary if and only if it is clopen; so \( \text{ind}(X) = 0 \) if and only if for each \( x \in X \) and each open neighborhood \( U \) of \( x \), there is a clopen neighborhood \( V \) of \( x \) contained in \( U \), i.e. if and only if \( X \) is nonempty and there is a base for the topology of \( X \) consisting of clopen sets. (Thus if \( X \) is \( T_0 \) and \( \text{ind}(X) = 0 \), \( X \) must be completely regular.)

XI.21.3.4. Proposition. Let \( X \) be a regular topological space. Then \( \text{ind}(X) \leq n \) if and only if, whenever \( A \) is a closed set in \( X \) and \( x \in X \setminus A \), there is a partition \( P \) between \( \{x\} \) and \( A \) with \( \text{ind}(P) \leq n-1 \).

Proof: There is an open neighborhood \( U \) of \( x \) with \( \bar{U} \subseteq X \setminus A \). There is an open neighborhood \( V \) of \( x \) contained in \( U \) with \( \text{ind}(\partial V) \leq n-1 \). Then \( P = \partial V \) is a partition between \( \{x\} \) and \( A \).

XI.21.3.5. Proposition. Let \( X \) be a topological \( n \)-manifold with boundary \((n \geq 1)\). Then \( 1 \leq \text{ind}(X) \leq n \). In particular, if \( X \) is a topological 1-manifold with boundary, then \( \text{ind}(X) = 1 \). So \( \text{ind}(\mathbb{R}) = \text{ind}([0,1]) = \text{ind}(S^1) = 1 \).

Proof: By induction on \( n \). If \( n = 1 \), a point of \( X \) has arbitrarily small neighborhoods whose boundary consists of either one or two points, hence of inductive dimension 0. Thus \( \text{ind}(X) \leq 1 \). But \( X \) is locally connected and thus \( \text{ind}(X) = 0 \) is false. And if \( X \) is an \((n+1)\)-manifold with boundary, each point of \( X \) has arbitrarily small neighborhoods whose boundary is homeomorphic to either \( S^n \) or \( B^n \), hence with inductive dimension \( \leq n \); thus \( \text{ind}(X) \leq n + 1 \). Also, \( \text{ind}(X) = 0 \) is false for the same reason as before.

XI.21.3.6. In fact, if \( X \) is a topological \( n \)-manifold with boundary, then \( \text{ind}(X) = n \), but it is much harder to prove equality, even if \( X = B^n \) or \( X = \mathbb{R}^n \). Actually, it is hard to prove that there is any topological space \( X \) with \( \text{ind}(X) > 1 \).

One simple but important property of small inductive dimension is monotonicity:

XI.21.3.7. Proposition. Let \( X \) be a topological space, and \( Y \) a subspace of \( X \). Then \( \text{ind}(Y) \leq \text{ind}(X) \).

Proof: By induction on \( n = \text{ind}(X) \). If \( n = -1 \) there is nothing to prove. Suppose the result is true whenever \( \text{ind}(X) \leq n \). If \( \text{ind}(X) \leq n + 1 \) and \( Y \subseteq X \), and \( y \in Y \), let \( U \) be an open neighborhood of \( y \) in \( Y \). Then there is an open neighborhood \( U \) of \( y \) in \( X \) with \( U = U \cap Y \). There is an open neighborhood \( V \) of \( y \) in \( X \) with \( \partial V \subseteq U \) and \( \text{ind}(\partial V) \leq n \). Set \( V = V \cap Y \); then \( V \) is an open neighborhood of \( y \) in \( Y \) with \( V \subseteq U \). We have \( \partial V \subseteq \partial V \) (\( \subseteq \)), so by the inductive hypothesis \( \text{ind}(\partial V) \leq n \). Thus \( \text{ind}(Y) \leq n + 1 \).

XI.21.3.8. Small inductive dimension is a good notion for separable metrizable spaces, but turns out to have pathological properties outside this class and is regarded by topologists as of little use or interest for more general spaces.
**Large Inductive Dimension**

The small inductive dimension is strictly a local concept. There is a technical variation, called the *large inductive dimension* (Ind), which is more of a global condition.

**XI.21.3.9. Definition.** Let $X$ be a topological space. Then

(i) $\text{Ind}(X) = -1$ if and only if $X = \emptyset$.

(ii) $\text{Ind}(X) \leq n$ if and only if, whenever $Y$ is a closed subset of $X$ and $U$ is an open neighborhood of $Y$ in $X$, there is an open neighborhood $V$ of $Y$ contained in $U$ with $\text{Ind}(\partial V) \leq n - 1$.

(iii) $\text{Ind}(X) = n$ if $\text{Ind}(X) \leq n$ and it is false that $\text{Ind}(X) \leq n - 1$.

(iv) $\text{Ind}(X) = \infty$ if for every $n$ it is false that $\text{Ind}(X) \leq n$.

**XI.21.3.10. Proposition.** Let $X$ be a normal topological space. Then $\text{Ind}(X) \leq n$ if and only if, whenever $A$ and $B$ are disjoint closed sets in $X$, there is a partition $P$ between $A$ and $B$ with $\text{Ind}(P) \leq n - 1$.

The proof is nearly identical to the proof of XI.21.3.4..

**XI.21.3.11. Proposition.** Let $X$ be a $T_1$ topological space. Then $\text{ind}(X) \leq \text{Ind}(X)$.

Proof: It is easily proved by induction that if $\text{Ind}(X) \leq n$, then $\text{ind}(X) \leq n$.

**XI.21.3.12. Theorem.** If $X$ and $Y$ are metrizable, then

$$\text{Ind}(X \times Y) \leq \text{Ind}(X) + \text{Ind}(Y).$$

Equality does not hold in general. For example, the space $X$ of the irrationals satisfies $\text{Ind}(X) = 1$ and $\text{Ind}(X \times X) = 1$ also, since $X \times X$ is homeomorphic to $X$. (This space is separable metrizable, so $\text{ind}(X) = \text{Ind}(X) = \dim(X)$. ) There are compact metrizable spaces for which equality fails []. It is unknown whether equality can fail for ANR’s; for well-behaved ANR’s equality does hold [Bor36].

The inequality holds in somewhat greater generality, but does not even hold for all compact Hausdorff spaces: there are compact Hausdorff spaces $X$ and $Y$ with $\text{Ind}(X) = 1$, $\text{Ind}(Y) = 2$, and $\text{Ind}(X \times Y) = 4$ [Fil72]. Compare with XI.21.3.37..

**Sum and Decomposition Theorems**

We now obtain some moderately complicated but crucial results about dimensions of unions, and decompositions into sets of smaller dimension.

It is clear that if $X$ is a union of two subspaces of dimension $n$, then the dimension of $X$ might be larger than $n$: for example, $\mathbb{R}$ can be decomposed into the rational and irrational numbers, each of which has dimension 0. There is only an inequality in general (). In fact, any sufficiently nice space of dimension $n$ can be written as a union of $n + 1$ subspaces of dimension 0 (). But if the space is nice and the subspaces are closed, the dimension of the union does not grow, even with countably many subspaces:
**XI.21.3.13. Theorem.** [Sum Theorem for ind] Let $X$ be a separable metrizable space. If $X = \bigcup_{j=1}^{\infty}X_j$, with each $X_j$ closed in $X$ and $\text{ind}(X_j) \leq n$ for all $j$, then $\text{ind}(X) = \text{Ind}(X) \leq n$.

The name “Sum Theorem,” which has become standard, comes from the practice in older references on set theory and topology to use sum as a synonym for union. Similarly, product was used to mean intersection.

To prove this theorem, we first prove a zero-dimensional version for $\text{Ind}$ which is valid in much greater generality:

**XI.21.3.14. Theorem.** Let $X$ be a normal space. If $X = \bigcup_{j=1}^{\infty}X_j$, with each $X_j$ closed in $X$ and $\text{Ind}(X_j) = 0$ for all $j$, then $\text{Ind}(X) = 0$.

Before proving XI.21.3.13., we need two lemmas, the first of which includes a useful construction:

**XI.21.3.15. Lemma.** Let $X$ be a topological space and $\mathcal{V}$ a base for the topology of $X$ (consisting of open sets). Set

$$Y = \bigcup_{V \in \mathcal{V}} \partial V$$

and $Z = X \setminus Y$. Then $\text{ind}(Z) \leq 0$.

**Proof:** Let $z \in Z$ and $U$ an open neighborhood of $z$ in $Z$. Then there is an open set $W$ in $X$ with $U = W \cap Z$. There is a $V \in \mathcal{V}$ with $z \in V$ and $V \subseteq U$. Since $\partial V \subseteq Y$, we have that $V \cap Z = V \cap Z$, so $V \cap Z$ is relatively clopen in $Z$ and $z \in V \cap Z \subseteq U$.

The second lemma is a variant of XI.20.1.23. (cf. [Eng95, 1.29]).

**XI.21.3.16. Lemma.** Let $X$ be a metrizable space, or more generally a completely normal space $\langle \rangle$, and $Z$ a subspace of $Z$ with $\text{Ind}(Z) = 0$. If $A$ and $B$ are disjoint closed subsets of $X$, then there are open sets $V_1$, $V_2$ in $X$ with $A \subseteq V_1$, $B \subseteq V_2$, $V_1 \cap V_2 = \emptyset$, and

$$[X \setminus (V_1 \cup V_2)] \cap Z = \emptyset.$$

**Proof:** Let $V_1'$ and $V_2'$ be open neighborhoods of $A$ and $B$ whose closures are disjoint. Let $U_1$ be a relatively clopen subset of $X$ containing $V_1' \cap Z$ (all closures are closures in $X$) and disjoint from $V_2 \cap Z$, and $U_2 = Z \setminus U_1$. Note that $V_1'$ is disjoint from $U_2$ and hence from $U_2$ since $V_1'$ is open; thus $A \cap U_2 = \emptyset$. Similarly, $B \cap U_1 = \emptyset$.

Since $U_1 \cap Z = U_1$ and $U_2 \cap Z = U_2$, the sets $U_1$ and $U_2$ are completely separated in $X$, and hence the sets $A \cup U_1$ and $B \cup U_2$ are also completely separated in $X$. Thus they have disjoint neighborhoods $V_1$ and $V_2$ in $X$. Since $Z \subseteq V_1 \cup V_2$, $[X \setminus (V_1 \cup V_2)] \cap Z = \emptyset$.

We now give the proof of XI.21.3.13.:

**Proof:** By induction on $n$. The case $n = 0$ follows immediately from XI.21.3.14. and the fact that if $Y$ is a $T_0$ Lindelöf space with $\text{ind}(Y) = 0$, then $\text{Ind}(Y) = 0$ (XI.13.6.5. and the proof of XI.13.6.7.((iv) $\Rightarrow$ (iii))). Now suppose $n > 0$ and the statement holds for $n - 1$ (including the statement that $\text{ind}(Y) = \text{Ind}(Y)$ for $Y$).
if \( \text{ind}(Y) \leq n - 1 \). There is a countable base \( \mathcal{V} \) for the topology of \( X \) consisting of open sets \( V \) with \( \text{ind}(\partial V) \leq n - 1 \). Set

\[
Y = \bigcup_{V \in \mathcal{V}} \partial V
\]

and \( Z = X \setminus Y \). Then \( \text{ind}(Z) \leq 0 \) by Lemma XI.21.3.15., so \( \text{Ind}(Z) \leq 0 \) since \( Z \) is Lindelöf. We have that \( \text{ind}(Y) \leq n - 1 \) by the inductive assumption. Let \( A \) and \( B \) be disjoint closed subsets of \( X \). By Lemma XI.21.3.16. there are disjoint neighborhoods \( V_1 \) and \( V_2 \) of \( A \) and \( B \) with \( Z \subseteq V_1 \cup V_2 \). Hence \( P = X \setminus (V_1 \cup V_2) \) is a partition between \( A \) and \( B \). Since \( P \subseteq Y \), \( \text{ind}(P) \leq n - 1 \), so \( \text{Ind}(P) \leq n - 1 \) by the inductive assumption. Thus \( \text{Ind}(X) \leq n \) by XI.21.3.10. So we have for this \( n \) that if each \( X_k \) satisfies \( \text{ind}(X_k) \leq n \), then \( \text{Ind}(X) \leq n \). In particular, if \( \text{ind}(X) = n \), set \( X_k = X \) for all \( k \) to conclude that \( \text{Ind}(X) \leq n \). Since \( \text{ind}(X) \leq \text{Ind}(X) \) (XI.21.3.11.), we have \( \text{Ind}(X) = n \).

We now obtain some corollaries of XI.21.3.13.

**XI.21.3.17.** *Corollary.* Let \( X \) be a separable metrizable space. Then \( \text{ind}(X) = \text{Ind}(X) \).

**XI.21.3.18.** *Corollary.* If \( X \) is a countable regular space, then \( \text{Ind}(X) = \text{ind}(X) = \text{dim}(X) = 0 \).

**Proof:** A countable topological space is Lindelöf, so a countable regular space is normal. Each singleton set \( \{x\} \) has \( \text{Ind} \{x\} = 0 \), thus \( \text{Ind}(X) = 0 \). We then have \( \text{ind}(X) = 0 \), so \( \text{dim}(X) = 0 \) by XI.13.6.7. (iii) \( \Rightarrow \) (iv)).

The regularity assumption is necessary: there exist connected countable Hausdorff spaces ().

**XI.21.3.19.** *Corollary.* [First Decomposition Theorem] Let \( X \) be a separable metrizable space, and \( n \geq 0 \). If \( \text{ind}(X) = n \), then \( X \) can be written as \( Y \cup Z \), where \( \text{ind}(Y) = n - 1 \) and \( \text{ind}(Z) = 0 \). Conversely, if \( X \) can be so written, then \( n - 1 \leq \text{ind}(X) \leq n \).

**Proof:** Since \( \text{ind}(X) = n \), there is an \( x \in X \) and an open neighborhood \( U \) of \( x \) such that no open neighborhood \( V \) of \( x \) contained in \( U \) has \( \text{ind}(\partial V) \leq n - 2 \). There is a countable base \( \mathcal{V} \) for \( X \) with \( \text{ind}(\partial V) \leq n - 1 \) for all \( V \in \mathcal{V} \). Set

\[
Y = \bigcup_{V \in \mathcal{V}} \partial V
\]

and \( Z = X \setminus Y \). Then \( \text{ind}(Z) \leq 0 \) by Lemma XI.21.3.15.. We have \( \text{ind}(Y) \leq n - 1 \) by XI.21.3.13.. But if \( V \in \mathcal{V} \) with \( x \in V \subseteq U \), then \( \text{ind}(\partial V) = n - 1 \), so \( \text{ind}(Y) = n - 1 \) by XI.21.3.7.. For the converse, suppose \( X = Y \cup Z \), where \( \text{ind}(Y) = n - 1 \) and \( \text{ind}(Z) = 0 \). Then \( \text{ind}(X) \geq n - 1 \) by XI.21.3.7.. Let \( x \in X \) and \( U \) an open neighborhood of \( x \). Then by Lemma XI.21.3.16. there are disjoint neighborhoods \( V_1 \) of \( x \) and \( V_2 \) of \( X \setminus U \) such that \( Z \subseteq V_1 \cup V_2 \). Then \( \partial V_1 \subseteq Y \), so \( \text{ind}(\partial V_1) \leq n - 1 \); and \( V_1 \subseteq U \). Thus \( \text{ind}(X) \leq n \).

Applying this result repeatedly, we obtain:

1294
XI.21.3.20. Corollary. [Second Decomposition Theorem] Let $X$ be a separable metrizable space. Then $\text{ind}(X) \leq n$ if and only if $X$ can be written as a union of $n + 1$ subspaces $X_1, \ldots, X_{n+1}$ with $\text{ind}(X_k) \leq 0$ for all $k$.

Covering Dimension
This is perhaps the most generally useful notion of dimension. The idea is due to Lebesgue, but the actual definition was given by Čech.

XI.21.3.21. The idea behind covering dimension is that $\mathbb{R}$ can be covered with small open intervals in such a way that only consecutive intervals overlap, i.e. no three of the intervals have nonempty intersection; however, since it is connected, it cannot be covered by pairwise disjoint bounded open intervals. Similarly, $\mathbb{R}^2$ can be covered by small open sets in such a way that no four of the sets have nonempty intersection (e.g. tile the plane with hexagons, or consider a brick wall, and slightly expand the tiles or bricks to open sets), but it appears that in any covering by bounded open sets at least three of the sets must have a nonempty intersection (although this turns out to be not so easy to prove). It is a little harder, but possible, to cover $\mathbb{R}^3$ with small open sets in such a way that no five of the sets have nonempty intersection (try it!), but in any open cover by bounded sets at least four of the sets must have nonempty intersection.

XI.21.3.22. Definition. Let $X$ be a topological space, and $\mathcal{U} = \{U_i : i \in I\}$ an open cover of $X$. Then $\mathcal{U}$ has order $\leq n$ if, whenever $i_1, \ldots, i_{n+2}$ are distinct elements of $I$, then

$$U_{i_1} \cap \cdots \cap U_{i_{n+2}} = \emptyset$$

(i.e. every point of $X$ is contained in at most $n + 1$ of the sets). $\mathcal{U}$ has order $n$ if it has order $\leq n$ but not order $\leq n - 1$, i.e. there are distinct $i_1, \ldots, i_{n+1}$ in $I$ such that

$$U_{i_1} \cap \cdots \cap U_{i_{n+1}} \neq \emptyset.$$

XI.21.3.23. Definition. Let $X$ be a nonempty topological space. The covering dimension $\text{dim}(X)$ of $X$ is $\leq n$ if every finite open cover of $X$ has a refinement of order $\leq n$. Say $\text{dim}(X) = n$ if $\text{dim}(X) \leq n$ and it is false that $\text{dim}(X) \leq n - 1$. If $\text{dim}(X) \leq n$ is false for all $n$, then $\text{dim}(X) = \infty$. Set $\text{dim}(\emptyset) = -1$.

XI.21.3.24. Most, but not all, authors use $\text{dim}(X)$ to denote covering dimension, as we will; the notation $\text{cov}(X)$ is also sometimes used. But in some references such as [HW41], $\text{dim}(X)$ means something else ($\text{ind}(X)$ in [HW41]); some references on geometric measure theory such as [Edg98] use $\text{dim}(X)$ to denote the Hausdorff dimension of a metric space $X$ (a metric, but not a topological, invariant).

XI.21.3.25. An open cover $\mathcal{V}$ of a topological space $X$ is order zero if and only if $X$ is nonempty and $\mathcal{V}$ is a partition of $X$ into clopen sets. Thus $\text{dim}(X) = 0$ if and only if $X$ is nonempty and every finite open cover of $X$ has a refinement consisting of a partition of $X$ into clopen sets.

1295
XI.21.3.26. Covering dimension can be alternately be defined using open shrinkings (XI.7.6.7.), a special kind of refinement. A refinement $V$ of an open cover $U = \{U_i : i \in I\}$ can be converted to an open shrinking using consolidation. A consolidation of $V$ under $U$ consists of, for each $V \in V$, choosing an $i \in I$ with $V \subseteq U_i$ and calling this $i k(V)$. Then for each $i$ set $W_i = \bigcup \{V \in V : k(V) = i\}$. Then $W_i$ is an open set contained in $U_i$ for each $i$, and $W = \{W_i : i \in I\}$ is an open cover of $X$ since each $V \in V$ is contained in some $W_i$. Thus $W$ is an open shrinking of $U$. If $U$ is a finite open cover, then $W$ is also a finite cover.

A consolidation of $V$ under $U$ is highly nonunique in general, and requires the AC to define.

It is easy to check that if $V$ has order $\leq n$, then any consolidation of $V$ under another open cover $U$ also has order $\leq n$. Thus we obtain:

XI.21.3.27. **Proposition.** Let $X$ be a topological space, and $n \geq 0$. The following are equivalent:

(i) $\dim(X) \leq n$.

(ii) Every finite open cover of $X$ has a finite refinement of order $\leq n$.

(iii) Every finite open cover of $X$ has an open shrinking of order $\leq n$.

The following sharpening is useful:

XI.21.3.28. **Theorem.** Let $X$ be a normal space, and $n \geq 0$. Then $\dim(X) \leq n$ if and only if, whenever $U = \{U_1, \ldots, U_{n+2}\}$ is an open cover of $X$ with $n + 2$ elements, there is an open shrinking of $U$ of order $\leq n$.

**Proof:** If $\dim(X) \leq n$, the condition trivially follows. Conversely, suppose $\dim(X) > n$, and let $V = \{V_1, \ldots, V_m\}$ be a finite open cover of $X$ with no refinement of order $\leq n$ (so $m \geq n + 2$). If $i_1, \ldots, i_k$ are distinct natural numbers $\leq m$, with $V_{i_1} \cap \cdots \cap V_{i_k} \neq \emptyset$, and $W = \{W_1, \ldots, W_m\}$ is an open shrinking of $V$ with $W_{i_1} \cap \cdots \cap W_{i_k} = \emptyset$, replace $V$ by $W$. Doing this process successively for each such $i_1, \ldots, i_k$ (there are only finitely many such sets), we may assume $V$ has the property that whenever $i_1, \ldots, i_k \leq m$ and $V_{i_1} \cup \cdots \cup V_{i_k} \neq \emptyset$, and $W = \{W_1, \ldots, W_m\}$ is an open shrinking of $V$, then $W_{i_1} \cap \cdots \cap W_{i_k} \neq \emptyset$.

The reduction does not change the fact that the order of $V$ is at least $n+1$, so reordering $V$ if necessary we may assume $\cap_{j=1}^{n+2} V_j \neq \emptyset$. Now set $U_j = V_j$ if $1 \leq j \leq n + 1$ and $U_{n+2} = \bigcup_{j=n+2}^m V_j$. If $W = \{W_1, \ldots, W_{n+2}\}$ is an open shrinking of $U$, then

$$\{W_1, \ldots, W_{n+1}, W_{n+2} \cap V_{n+2}, \ldots, W_{n+2} \cap V_m\}$$

is an open shrinking of $V$, and hence

$$\emptyset \neq (\cap_{j=1}^{n+2} W_j) \cap (W_{n+2} \cap V_{n+2}) \subseteq \cap_{j=1}^{n+2} W_j.$$

\(\Box\)

XI.21.3.29. **Proposition.** Let $X$ be a normal space, and $n \geq 0$. If, for every pair $A, B$ of disjoint closed sets of $X$ there is a partition $P$ between $A$ and $B$ with $\dim(P) \leq n-1$, then $\dim(X) \leq n$. 3.1.27

1296
XI.21.3.30. Why finite open covers, and not general open covers? Mainly, the theory just turns out to work well this way. But note that any open cover of finite order is locally finite, and there are spaces, even normal spaces, with open covers which have no locally finite (even no point-finite) refinements. But for paracompact spaces, there is no difference:

XI.21.3.31. Theorem. [Dowker’s Theorem] Let $X$ be a topological space with $\dim(X) \leq n$, and $\mathcal{U}$ a locally finite open cover of $X$. Then $\mathcal{U}$ has a refinement of order $\leq n$.

XI.21.3.32. Corollary. Let $X$ be a paracompact space. The following are equivalent:

(i) $\dim(X) \leq n$.

(ii) Every finite open cover of $X$ has a shrinking of order $\leq n$.

(iii) Every open cover of $X$ has a refinement of order $\leq n$.

(iv) Every open cover of $X$ has an open shrinking of order $\leq n$.

XI.21.3.33. Definition. Let $X$ be a topological space. An open cover $\mathcal{U}$ of $X$ is $n$-decomposable if $\mathcal{U}$ is a disjoint union $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_{n+1}$ of $n + 1$ subcollections, each of which consists of pairwise disjoint sets.

XI.21.3.34. It is obvious that an $n$-decomposable open cover has order $\leq n$. The converse is false: consider a cover of a circle by three arcs. (A cover is 0-decomposable if and only if it has order zero, i.e. consists of a partition into clopen subsets.) However, we do have:

XI.21.3.35. Proposition. Let $X$ be a normal topological space, and $0 \leq n < \infty$. Then every open cover of $X$ of order $n$ has an $n$-decomposable refinement. Every finite open cover of $X$ of order $n$ has a finite $n$-decomposable refinement.

**Proof:** Let $\mathcal{V}$ be an open cover of $X$ of order $n$. Since $\mathcal{V}$ is locally finite, there is a partition of unity $\{f_i\}$ subordinate to $\mathcal{V}$ ( ), which defines a continuous function $\phi$ from $X$ to $|K|$ for a simplicial complex $K$ of dimension $\leq n$ ( ). Let $K'$ be the barycentric subdivision of $K$. Then each vertex $\sigma$ of $K'$ is the barycenter of a unique simplex $\Delta_\sigma$ of $K$; let $d(\sigma) \in \{0, \ldots, n\}$ be the dimension of $\Delta_\sigma$.

Since $|K'|$ can be identified with $|K|$, $\phi$ may be regarded as a map from $X$ to $|K'|$. For each vertex $\sigma$ of $K'$, let $U_\sigma$ be the preimage under $\phi$ of the star of $\sigma$, the points of $|K'|$ whose $\sigma$-coordinate is positive. Then $\mathcal{U} = \{U_\sigma\}$ is an open cover of $X$ which refines $\mathcal{V}$. For each $k$, $0 \leq k \leq n$, let

$$\mathcal{U}_k = \{U_\sigma : d(\sigma) = k\}.$$
Then \( U \) is the disjoint union of the \( U_k \). If \( \sigma, \tau \) are vertices of \( K' \) with \( d(\sigma) = d(\tau) = k \), then \( \sigma \) and \( \tau \) are the barycenters of distinct \( k \)-simplexes in \( K \), so the stars of \( \sigma \) and \( \tau \) are disjoint, and thus \( U_\sigma \) and \( U_\tau \) are disjoint. So \( U \) is \( n \)-decomposable. If \( V \) is finite, \( K' \) (hence \( K' \)) is finite, so \( U \) is also finite.

XI.21.3.36. **Corollary.** Let \( X \) be a normal topological space, and \( n \geq 0 \). Then \( \dim(X) \leq n \) if and only if every finite open cover of \( X \) has an \( n \)-decomposable refinement.

XI.21.3.37. **Theorem.** [Hem46] Let \( X \) and \( Y \) be compact Hausdorff spaces. Then \( \dim(X \times Y) \leq \dim(X) + \dim(Y) \).

Compare with XI.21.3.12.

XI.21.3.38. **Covering dimension is not well behaved for spaces which are not normal.** There is a technical variation which works somewhat better for completely regular spaces (dimension theory is of almost no use or interest for spaces which are not at least completely regular). Recall that a set \( U \) in a topological space \( X \) is a *cozero set* if there is a continuous function \( f : X \to \mathbb{R} \) with \( U = \{ x \in X : f(x) \neq 0 \} \). A cozero set is an open \( F \); if \( X \) is normal, the converse is also true. Some references use the term *functionally open* for cozero sets.

XI.21.3.39. **Definition.** Let \( X \) be a topological space. Then \( \zdim(X) \leq n \) if every finite open cover of \( X \) by cozero sets has a finite refinement by cozero sets which is of order \( \leq n \).

XI.21.3.40. **There is no obvious relation between \( \dim(X) \) and \( \zdim(X) \) for general (even completely regular) spaces \( X \).** But every open cover of a normal space has a shrinking consisting of cozero sets (Exercise XI.21.8.1.); hence if \( X \) is normal, \( \zdim(X) = \dim(X) \). Some authors define the covering dimension of a space \( X \) to be \( \zdim(X) \).

**Partition Dimension**

In analogy with the Lebesgue Covering Theorem (.), we make another definition:

XI.21.3.41. **Definition.** Let \( X \) be a nonempty topological space. The *partition dimension* \( pd(X) \) is the smallest \( n \geq 0 \) such that, whenever \((A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\) are \( n + 1 \) pairs of disjoint closed sets in \( X \), there are partitions \( P_k \) of \((A_k, B_k)\) for \( 1 \leq k \leq n + 1 \) with \( P_1 \cap \cdots \cap P_{n+1} = \emptyset \). If there is no such \( n \), set \( pd(X) = \infty \). Set \( pd(\emptyset) = -1 \).

XI.21.3.42. **If \( X \) is a topological space, then \( pd(X) \leq n \) if and only if, whenever \((A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\) are \( n + 1 \) pairs of disjoint closed sets in \( X \), there are partitions \( P_k \) of \((A_k, B_k)\) for \( 1 \leq k \leq n + 1 \) with \( P_1 \cap \cdots \cap P_{n+1} = \emptyset \). If \( X \) is normal and this condition holds for one \( n \), it holds for all larger \( n \) (and if the condition holds for some \( n \), then \( X \) is normal). The Lebesgue Covering Theorem implies that \( pd(B_n) \geq n \).
Decomposition Dimension

We define one more notion of dimension:

**XI.21.3.43. Definition.** Let \( X \) be a nonempty topological space. The *decomposition dimension* \( \text{dd}(X) \) is the smallest \( n \) such that \( X \) can be written

\[
X = X_1 \cup \cdots \cup X_{n+1}
\]

where \( \text{ind}(X_k) \leq 0 \) for \( 1 \leq k \leq n+1 \). If there is no such \( n \), then \( \text{dd}(X) = \infty \). Set \( \text{dd}(\emptyset) = -1 \).

**XI.21.3.44.** The Second Decomposition Theorem (XI.21.3.20.) says that \( \text{dd}(X) = \text{ind}(X) \) for any separable metrizable \( X \).

**XI.21.4. Comparing the Dimensions**

All the dimensions defined so far agree for separable metric spaces:

**XI.21.4.1. Theorem.** Let \( X \) be a topological space which is separable and metrizable (equivalently, second countable and regular). Then \( \text{ind}(X) = \text{Ind}(X) = \text{dim}(X) = \text{ed}(X) = \text{pd}(X) = \text{dd}(X) \).

This result not only confirms that for “nice” spaces there is an unambiguous notion of topological dimension, but is also useful since some properties of dimension are proved easily for some of the dimension functions and not so easily for others. Proofs of the theorem can be found in [HW41] and [Eng95].

For general spaces, even for metrizable spaces or compact Hausdorff spaces, the dimension functions do not always agree. We state without proof some of the principal relations between them; see e.g. [HW41], [Pea75], [Eng95].

**XI.21.4.2. Theorem.** ([Kat52], [Mor54b]) If \( X \) is a metrizable space (not necessarily separable), then \( \text{Ind}(X) = \text{dim}(X) \).

**XI.21.4.3.** There is a completely metrizable space \( X \) with \( \text{ind}(X) = 0 \) and \( \text{Ind}(X) = \text{dim}(X) = 1 \) [Roy68]. There are first countable compact Hausdorff spaces \( X \) with \( \text{ind}(X) < \text{Ind}(X) \) ([Fil69], [Pas70]). More examples can be found in [Nag70] and [Pea75].

**XI.21.4.4. Theorem.** [?] If \( X \) is a compact Hausdorff space, then \( \text{dim}(X) \leq \text{ind}(X) \).

The result extends to strongly paracompact spaces, but not even to completely metrizable spaces as ROY’S example shows. There is a compact Hausdorff space \( X \) with \( \text{dim}(X) < \text{ind}(X) \) ([Lun49]; cf. [Eng95, 3.1.31]).
**XI.21.4.5.** THEOREM. [Ale47] If $X$ is completely regular, then $ed(X) = zdim(X)$. If $X$ is normal, then $ed(X) = dim(X)$.

The next result is a corollary of XI.21.3.7. and XI.21.4.1:

**XI.21.4.6.** COROLLARY. Let $X$ be a separable metrizable space and $Y$ a subset of $X$. Then $dim(Y) \leq dim(X)$.

**XI.21.4.7.** There is an example of a compact Hausdorff space $X$ and a normal subspace $Y$ such that $dim(X) = Ind(X) = 0$ and $dim(Y) = Ind(Y) = 1$ ([Dow55]; cf. [Eng95, 2.2.11], [GJ76, 16M]).

**XI.21.4.8.** PROPOSITION. Let $X$ be a normal topological space. Then $dim(X) \leq Ind(X)$.

PROOF: This is an easy induction from XI.21.3.29., using XI.21.3.10. and the fact that if $Ind(X) = 0$, then $dim(X) = 0$ (XI.21.3.28.).

For separable metrizable spaces, we obtain a stronger result:

**XI.21.4.9.** PROPOSITION. Let $X$ be a separable metrizable space (or, more generally, a completely normal hereditarily Lindelöf space). Then $dim(X) \leq dd(X) (= ind(X))$.

PROOF: Suppose $dd(X) \leq n$, i.e. $X = X_1 \cup \cdots \cup X_{n+1}$ with $ind(X_k) \leq 0$ for all $k$. Let $\mathcal{U} = \{U_1, \ldots, U_m\}$ be a finite open cover of $X$. For each $k$, we have $dim(X_k) \leq 0$, so the cover $\{U_1 \cap X_k, \ldots, U_m \cap X_k\}$ of $X_k$ has a shrinking $\{F_1, \ldots, F_m\}$ consisting of relatively clopen subsets of $X_k$. Since $X$ is completely normal, there are disjoint open neighborhoods $\{V_{1,k}, \ldots, V_{m,k}\}$ of $F_1, \ldots, F_m$ in $X$. Consider the open cover $W$ of $X$ consisting of all sets of the form $U_j \cap V_{j,k}$. Then $W$ refines $\mathcal{U}$, and has order $\leq n$ since for any $n+2$ distinct sets $U_j \cap V_{j,k}$ in $W$ there must be some $k$ with $k_i = k$ for two distinct $i$. Thus $dim(X) \leq n$.

This result also follows from XI.21.4.8. and the fact that $dd(X) = ind(X) = Ind(X)$ for separable metrizable spaces (XI.21.3.17., XI.21.3.44.), but this requires the Sum Theorem for $ind$; the inequality $dim(X) \leq dd(X)$ for separable metrizable spaces does not require the Sum Theorem.

**XI.21.4.10.** PROPOSITION. Let $X$ be a normal topological space. Then $pd(X) \leq dim(X)$.

PROOF: Suppose $dim(X) \leq n$, and let $\{(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\}$ be a set of $n+1$ pairs of disjoint closed subsets of $X$. Then the open cover

$$\left\{ X \setminus A_1, \ldots, X \setminus A_{n+1}, X \setminus \bigcup_{k=1}^{n+1} B_k \right\}$$

has a shrinking $V = \{V_1, \ldots, V_{n+2}\}$ of order $\leq n$. Let $\{F_1, \ldots, F_{n+2}\}$ be a closed shrinking of $V$. For $1 \leq k \leq n+1$ define

$$\tilde{A}_k = A_k \cup (F_{n+2} \setminus V_k), \quad \tilde{B}_k = B_k \cup F_{n+2}.$$
Since $A_k \cap F_k = \emptyset$ and $B_k \cap F_{n+2} = \emptyset$ for each $k$, we have $\tilde{A}_k \cap \tilde{B}_k = \emptyset$. Also,

$$\bigcup_{k=1}^{n+1} F_k \subseteq \bigcup_{k=1}^{n+1} \tilde{B}_k.$$  

We have that

$$F_{n+2} \cap \left( \bigcap_{k=1}^{n+1} V_k \right) \subseteq V_{n+2} \cap \left( \bigcap_{k=1}^{n+1} V_k \right) = \emptyset$$

since $\mathcal{V}$ has order $\leq n$. Thus $F_{n+2} \subseteq \bigcup_{k=1}^{n+1} \tilde{A}_k$. Since $\bigcup_{k=1}^{n+2} F_k = X$, we have

$$\bigcup_{k=1}^{n+1} (\tilde{A}_k \cup \tilde{B}_k) = X.$$

Let $P_k$ be a partition between $\tilde{A}_k$ and $\tilde{B}_k$ for each $k$. Then $P_k$ is a partition between $A_k$ and $B_k$, and $\bigcap_{k=1}^{n+1} P_k = \emptyset$ since it is disjoint from $\bigcup_{k=1}^{n+1} (\tilde{A}_k \cup \tilde{B}_k) = X$.

**XI.21.4.11.  PROPOSITION.** Let $X$ be a normal space. Then $\dim(X) = pd(X)$.

**PROOF:** Because of XI.21.4.10., we need to show that $\dim(X) \leq pd(X)$. Assume $pd(X) \leq n$, and let $\mathcal{U} = \{U_1, \ldots, U_{n+2}\}$ be an open cover of $X$. There is a closed shrinking $\{A_1, \ldots, A_{n+2}\}$ of $\mathcal{U}$. Set $B_k = X \setminus U_k$.

Then $\{(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\}$ is a family of pairs of disjoint closed sets in $X$, so there are partitions $P_k$ between $A_k$ and $B_k$ for $1 \leq k \leq n+1$ with $\bigcap_{k=1}^{n+1} P_k = \emptyset$, i.e. for each $k \leq n+1$ there are disjoint open sets $V_k, W_k$ containing $A_k$ and $B_k$ respectively with

$$\bigcap_{k=1}^{n+1} X \setminus (V_k \cup W_k) = \emptyset.$$  

Set

$$V_{n+2} = U_{n+2} \cap \left( \bigcup_{k=1}^{n+1} W_k \right).$$

Then $V_k \subseteq U_k$ for all $k$, and

$$\bigcup_{k=1}^{n+2} V_k = \left( \bigcup_{k=1}^{n+1} V_k \right) \cup \left( U_{n+2} \cap \left[ \bigcup_{k=1}^{n+1} W_k \right] \right)$$

$$= \left( \bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \right) \cap \left( \left[ \bigcup_{k=1}^{n+1} V_k \right] \cup \left[ \bigcup_{k=1}^{n+1} W_k \right] \right)$$
$$= \left( \bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \right) \cap X = \bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \supseteq \bigcup_{k=1}^{n+2} A_k = X$$
We need only show \( \mathcal{V} = \{V_1, \ldots, V_{n+2}\} \) is a shrinking of \( \mathcal{U} \). We also have
\[
\bigcap_{k=1}^{n+2} V_k = \left( \bigcap_{k=1}^{n+1} V_k \right) \cap \left( U_{n+2} \cup \bigcup_{k=1}^{n+1} W_k \right) 
\subseteq \left( \bigcap_{k=1}^{n+1} V_k \right) \cap \left( \bigcup_{k=1}^{n+1} W_k \right) = \emptyset
\]
so \( \mathcal{V} \) has order \( \leq n \). By XI.21.3.28., \( \dim(X) \leq n \). \( \Diamond \)

**XI.21.4.12.** Proposition. Let \( X \) be a completely regular space. Then \( pd(X) \leq ed(X) \).

**Proof:** We will use XI.21.1.8.. Suppose \( ed(X) \leq n \), and let \( (A_1, B_1), \ldots, (A_{n+1}, B_{n+1}) \) be \( n+1 \) pairs of disjoint closed sets in \( X \). Let \( Y \) be the union of all the \( A_k \) and \( B_k \). Then \( Y \) is a closed subset of \( X \). For each \( k \), \( 1 \leq k \leq n+1 \), set \( f_k(x) = 0 \) if \( x \in A_k \) and \( f_k(x) = 1 \) if \( x \in B_k \). If \( f = (f_1, \ldots, f_{n+1}) \), then \( f \) is a map from \( Y \) to \( \partial I^{n+1} \approx S^n \). By XI.21.1.8., \( f \) extends to a map \( g \) from \( X \) to \( \partial I^{n+1} \). Let \( g_k (1 \leq k \leq n+1) \) be the \( k \)'th coordinate function of \( g \); then \( g_k \) extends \( f_k \) to \( X \). Set \( P_k = g_k^{-1}(1/2) \). Then \( P_k \) is a partition between \( A_k \) and \( B_k \), and \( P_1 \cap \cdots P_{k+1} = \emptyset \) since \((1/2, \ldots, 1/2) \notin \partial I^{n+1}". Thus \( pd(X) \leq n \). \( \Diamond \)

**XI.21.4.13.** Proposition. Let \( X \) be a normal space. Then \( ed(X) = pd(X) \).

**Proof:** We need only show \( ed(X) \leq pd(X) \). We again use XI.21.1.8.. Suppose \( pd(X) \leq n \), \( Y \) is a closed subset of \( X \), and \( f \) is a map from \( Y \) to \( \partial I^{n+1} \). For \( 1 \leq k \leq n+1 \), let \( A_k \) [resp. \( B_k \)] be the set of \( x \in X \) such that the \( k \)'th coordinate of \( f(x) \) is \( 0 \) [resp. \( 1 \)]. Then there are partitions \( P_k \) between \( A_k \) and \( B_k \) for each \( k \), with \( \bigcap_{k=1}^{n+1} P_k = \emptyset \) since \( pd(X) \leq n \), i.e. for each \( k \leq n+1 \) there are disjoint open sets \( V_k, W_k \) containing \( A_k \) and \( B_k \) respectively with
\[
\bigcap_{k=1}^{n+1} X \setminus (V_k \cup W_k) = \emptyset
\]
Let \( \{U_1, \ldots, U_{n+1}\} \) be a swalling of \( \{P_1, \ldots, P_{n+1}\} \) (XI.7.6.10.), so \( P_k \subseteq U_k \) and \( \bigcap_{k=1}^{n+1} U_k = \emptyset \). Then \( V_k \cup P_k \) and \( W_k \cup P_k \) are closed in \( X \), hence normal, so by Urysohn’s Lemma there is a map \( \phi_k : V_k \cup P_k \to [0,1/2] \) with \( \phi_k(V_k \setminus U_k) = \{0\} \) and \( \phi_k(P_k) = \{1/2\} \), and a map \( \psi_k : W_k \cup P_k \to [1/2,1] \) with \( \psi_k(P_k) = \{1/2\} \) and \( \psi_k(W_k \setminus U_k) = \{1\} \). Then \( \phi_k \) and \( \psi_k \) define a map \( f_k \) from \( X \) to \([0,1]\) since they agree on the closed set \( P_k = (V_k \cup P_k) \cap (W_k \cup P_k) \). We have \( g_k(A_k) = \{0\} \) and \( g_k(B_k) = \{1\} \). The \( g_k \) define a map \( g : X \to I^n \). Since \( g_k^{-1}(\{1/2\}) \subseteq U_k \) for each \( k \) and \( \bigcap_{k=1}^{n+1} U_k = \emptyset \), \( g \) never takes the value \( p = (1/2, \ldots, 1/2) \). If \( r \) is a retraction from \( I^n \setminus \{p\} \) to \( \partial I^n \), then \( r \circ g \) is an extension of \( f \) to \( X \). \( \Diamond \)

Combining XI.21.4.13. with XI.21.4.11., we obtain:

**XI.21.4.14.** Corollary. Let \( X \) be a normal space. Then \( \dim(X) = pd(X) = ed(X) \).
XI.21.4.15. In fact, by an extension of the same methods it can be shown that \( ed(X) = zdim(X) \) for any completely regular \( X \).

The hardest of the inequalities in XI.21.4.1. is that \( ind(X) \leq dim(X) \). We now work on this.

XI.21.4.16. We have \( Ind(X) \leq dim(X) \) (and hence \( Ind(X) = dim(X) \) by XI.21.4.8.) if \( X \) is metrizable.

XI.21.5. More Dimension Theories

Analytic Dimension

XI.21.5.1. Let \( X \) be a completely regular topological space. We will give a characterization due to Katětov of \( zdim(X) \) in terms of the ring (algebra) \( BC_{\mathbb{R}}(X) \) of bounded real-valued continuous functions on \( X \). (This ring is denoted \( C^*(X) \) in many references such as [GJ76], which is a thorough study of topology via these rings. We have adopted different notation since \( C(X) \) is too easily confused with C*-algebra notation.) \( BC_{\mathbb{R}}(X) \) has a norm

\[
\| f \| = \sup_{x \in X} |f(x)|
\]

under which it is a real Banach space. The norm is determined by the algebraic structure since it can be defined using the order structure, and a function is nonnegative if and only if it is a square. Note that every subring of \( BC_{\mathbb{R}}(X) \) containing the constant functions is a subalgebra (all subrings we consider will contain the constant functions).

XI.21.5.2. If \( X \) is compact, or more generally countably compact, then \( BC_{\mathbb{R}}(X) = C_{\mathbb{R}}(X) \), the set of all real-valued continuous functions on \( X \) if \( f : X \to \mathbb{R} \) is continuous, then \( \{ f^{-1}((-n, n)) : n \in \mathbb{N} \} \) is a countable open cover of \( X \) which has a finite subcover. A space \( X \) with \( BC_{\mathbb{R}}(X) = C_{\mathbb{R}}(X) \) is called pseudocompact; there exist pseudocompact spaces which are not countably compact, but every normal pseudocompact space is countably compact [GJ76, 3D]. In general, we have a natural identification \( BC_{\mathbb{R}}(X) \cong C_{\mathbb{R}}(\beta X) \), where \( \beta X \) is the Stone-Čech compactification of \( X \).

XI.21.5.3. The idea behind Katětov’s characterization is that if \( X \) is a compact subset of \( \mathbb{R}^n \), then polynomials in the coordinate functions are dense in \( C_{\mathbb{R}}(X) \), i.e. \( C_{\mathbb{R}}(X) \) has a dense unital subalgebra generated by \( n \) elements (the coordinate functions). This turns out to be the minimum number of generators of a dense unital subalgebra if \( dim(X) = n \), e.g. if \( X \) has nonempty interior.

XI.21.5.4. Simply taking the minimum number of generators of a dense unital subalgebra does not work as a satisfactory notion of dimension for a general \( X \), however. For one thing, \( BC_{\mathbb{R}}(X) \) can easily be nonseparable, and will be if \( \beta X \) is not metrizable, e.g. if \( X \) is not compact; then even no countably generated subalgebra can be dense. And the wrong answer is obtained even for a space as simple as a circle \( S^1 \): the minimum number of generators of a dense unital subalgebra of \( C_{\mathbb{R}}(S^1) \) is two, since no real-valued continuous function on \( S^1 \) separates points.

The key trick is to consider analytic subrings:
XI.21.5.5. Definition. Let $X$ be a completely regular space. An analytic subring of $BC_R(X)$ is a closed unital subalgebra $A$ with the property that if $f^2 \in A$, then $f \in A$. The smallest analytic subring containing a set $F$ of functions is called the analytic subring generated by $F$.

Note that any intersection of analytic subrings is an analytic subring, so there is always a smallest analytic subring containing any given set of functions.

XI.21.5.6. Examples. (i) The example of the unit circle $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ nicely illustrates the idea. Let $f(x, y) = x$. Then the unital subalgebra generated by $f$ is not dense in $C_R(S^1)$. However, if $A$ is the analytic subring of $C_R(S^1)$ generated by $f$, then $A$ contains $g = 1 - f^2$. We have $g(x, y) = 1 - x^2 = y^2$. Thus $A$ also contains $h(x, y) = y$ since $g = h^2$, and it follows from the Stone-Weierstrass Theorem that $A$ is all of $C_R(S^1)$. Thus $C_R(S^1)$ is generated as an analytic subring by one function, i.e. “$S^1$ is one-dimensional.”

(ii) Let $X$ be compact. Consider the analytic subring $A$ generated by the empty set, i.e. the smallest analytic subring of $C_R(X)$. Then $A$ contains the constant function 1 by definition, so it also contains every function taking exactly the values $\pm 1$, and thus the indicator function of every clopen subset of $X$. So all continuous functions with finite range are in $A$. $A$ is just the closure of this set, i.e. $A$ is the closed subalgebra of $C_R(X)$ generated by the indicator functions of clopen sets in $X$. So $A$ is all of $C_R(X)$ if and only if $X$ is totally disconnected, i.e. zero-dimensional.

XI.21.5.7. If $F$ is a subset of $BC_R(X)$, a subset $Y$ of $X$ is a stationary set for $F$ if every $f \in F$ is constant on $Y$. If $Y$ is a stationary set for $F$, so is $\bar{Y}$, so we need consider only closed stationary sets. Of course, each singleton is a stationary set, and the converse is true if and only if $F$ separates points of $X$. The Stone-Weierstrass Theorem can be rephrased:

XI.21.5.8. Theorem. Let $X$ be a compact Hausdorff space, and $F$ a subset of $C_R(X)$. Then the closed unital subalgebra of $C_R(X)$ generated by $F$ consists of all functions in $C_R(X)$ which are constant on each stationary set of $F$.

A variation of this, due to Katětov, is the key result:

XI.21.5.9. Theorem. Let $X$ be a compact Hausdorff space, and $F$ a subset of $C_R(X)$. Then the analytic subring of $C_R(X)$ generated by $F$ consists of all functions in $C_R(X)$ which are constant on each connected stationary set of $F$.

One can restrict attention to maximal connected stationary sets: every connected stationary set is contained in a maximal connected stationary set (XI.13.3.2.), which is automatically closed (XI.13.1.9.).

XI.21.5.10. We can provisionally define the analytic dimension of $X$ to be the smallest $n$ such that $BC_R(X)$ is generated as an analytic subring by $n$ elements. This is a good definition if $BC_R(X)$ is separable, which occurs if and only if $X$ is compact and metrizable. In general, we make the following definition:
XI.21.5.11. Definition. Let $X$ be a completely regular space. The *analytic dimension* $ad(X)$ is the smallest $n$, $0 \leq n \leq \aleph_0$, such that every countably generated (i.e. separable) closed subalgebra of $BC_\mathbb{R}(X)$ is contained in an analytic subring generated by $n$ elements.

The analytic dimension $ad(X)$ is the smallest $n$ such that every finite subset of $BC_\mathbb{R}(X)$ is contained in an analytic subring generated by $n$ elements. If $X$ is compact and metrizable, then $ad(X)$ is the smallest $n$ such that $BC_\mathbb{R}(X)$ is generated as an analytic subring by $n$ elements. Note that the analytic subring generated by even finitely many elements need not be separable: if $X$ is the product of uncountably many two-point spaces, then $X$ is compact and zero-dimensional, and the analytic subring of $C_\mathbb{R}(X)$ generated by $\emptyset$ is all of $C_\mathbb{R}(X)$, which is nonseparable since $X$ is not metrizable.

The main theorem of Katětov is:

XI.21.5.12. Theorem. Let $X$ be a completely regular space. Then $ad(X) = zdim(X)$. Thus, if $X$ is normal, $ad(X) = dim(X)$.

See [GJ76] for a complete discussion and proof, which is fairly involved. Since $ad(X)$ is transparently equal to $ad(\beta X)$ for any $X$, we obtain:

XI.21.5.13. Corollary. Let $X$ be a completely regular space. Then $zdim(X) = zdim(\beta X) = dim(\beta X)$. If $X$ is normal, then $dim(X) = dim(\beta X)$.

The corollary can be proved directly without introducing analytic dimension (cf. XI.21.1.12., XI.21.4.5.; [GJ76, 16.11]).

Affine Dimension
Homological Dimension

XI.21.6. Embedding in Euclidean Space

In this subsection, we prove two important theorems about embedding of topological spaces in Euclidean space.

If $X$ is a topological space of dimension $n$, we want to embed $X$ in $\mathbb{R}^p$ for some $p$ which is a function of $n$. Of course, for $X$ to embed in any Euclidean space it is necessary that $X$ be separable and metrizable, so we make an assumption that all spaces in this section are separable and metrizable. Thus all standard dimension functions agree on such an $X$, so “dimension $n$” is unambiguous.

XI.21.6.1. Theorem. Let $X$ be a subset of $\mathbb{R}^n$. Then $X$ has dimension $n$ if and only if the interior of $X$ is nonempty, i.e. if and only if $X$ contains a nonempty open set in $\mathbb{R}^n$.

The proof is deferred to XI.21.6.14..

XI.21.6.2. If a (separable metrizable) space $X$ has dimension $n$, it cannot be embedded in $\mathbb{R}^n$ in general (consider a circle). It cannot even generally be embedded in $\mathbb{R}^{n+1}$: a graph cannot always be embedded in $\mathbb{R}^2$, and the projective plane and the Klein bottle are closed 2-manifolds which cannot be embedded in $\mathbb{R}^3$. The $n$-skeleton of a $2n + 2$-simplex cannot be embedded in $\mathbb{R}^{2n}$[]. The best which can be done in general is the following:
XI.21.6.3. Theorem. Let $X$ be a separable metrizable space with $\dim(X) \leq n$. Then $X$ embeds in $\mathbb{I}^{2n+1}$. In fact, every continuous function from $X$ to $\mathbb{I}^{2n+1}$ is a uniform limit of embeddings.

To prove XI.21.6.3. (and in fact to prove a slightly better result), we introduce notation for some useful subsets of $\mathbb{R}^n$:

XI.21.6.4. Definition. Let $0 \leq n \leq m$. Then

- $\mathcal{P}_n^m$ is the set of points in $\mathbb{R}^m$ with precisely $n$ rational coordinates.
- $\mathcal{M}_n^m$ is the set of points in $\mathbb{R}^m$ with at most $n$ rational coordinates.
- $\mathcal{L}_n^m$ is the set of points in $\mathbb{R}^m$ with at least $n$ rational coordinates.
- $\mathcal{N}_n = \mathcal{M}_n^{2n+1}$ is the Nöbeling universal space of dimension $n$.

Note that $\mathcal{M}_n^m \cup \mathcal{L}_n^m = \mathbb{R}^m$ and $\mathcal{M}_n^m \cap \mathcal{L}_n^m = \mathcal{P}_n^m$ for any $n$ and $m$.

XI.21.6.5. Proposition. For any $0 \leq n \leq m$, $\dim(\mathcal{P}_n^m) = 0$, $\dim(\mathcal{M}_n^m) = n$, $\dim(\mathcal{L}_n^m) = m - n$, $\dim(\mathcal{N}_n) = n$.

Proof: $\mathcal{P}_n^m$ is a countable union of (relatively) closed subsets where $n$ specified coordinates are specified fixed rational numbers and the other $m - n$ coordinates are irrational. Such a set is homeomorphic to $\mathbb{I}^{m-n}$, which has dimension 0. Thus $\dim(\mathcal{P}_n^m) = 0$ by the Sum Theorem (the zero-dimensional version XI.21.3.14. suffices).

The observations

$$\mathcal{M}_n^m = \mathcal{P}_0^m \cup \mathcal{P}_1^m \cup \cdots \cup \mathcal{P}_n^m$$

$$\mathcal{L}_n^m = \mathcal{P}_n^m \cup \mathcal{P}_{n+1}^m \cup \cdots \cup \mathcal{P}_m^m$$

and the Sum Theorem imply that $\dim(\mathcal{M}_n^m) \leq n$ and $\dim(\mathcal{L}_n^m) \leq m - n$. But $\mathcal{M}_n^m$ contains a subspace homeomorphic to $\mathbb{R}^n$ (say the set of all points with the first $m - n$ coordinates $\pi$), and $\mathcal{L}_n^m$ similarly contains a copy of $\mathbb{R}^{m-n}$, so $\dim(\mathcal{M}_n^m) \geq n$ and $\dim(\mathcal{L}_n^m) \geq m - n$ by XI.21.3.7..

The theorem we will prove is:

XI.21.6.6. Theorem. Let $X$ be a separable metrizable space of dimension $\leq n$. Then there is an embedding $f : X \to \mathcal{N}_n$ such that the closure of $f(X)$ in $\mathcal{N}_n$ is compact. In fact, any bounded continuous function from $X$ to $\mathbb{R}^{2n+1}$ can be uniformly approximated arbitrarily closely by such a function.

This result justifies the name “Nöbeling universal space.” Theorem XI.21.6.3. is an immediate corollary. Another important corollary is:
XI.21.6.7. **Corollary.** Let $X$ be a separable metrizable space of dimension $n$. Then $X$ has a metrizable compactification of dimension $n$. In fact, there is a compact subset $K_n$ of $\mathbb{R}^{2n+1}$ with $\dim(K_n) = n$ such that every separable metrizable space of dimension $\leq n$ embeds in $K_n$.

**Proof:** There is an embedding $f: \mathfrak{M}_n \to \mathfrak{M}_n$ such that $K_n = f(\mathfrak{M}_n)$ is a compact subset of $\mathfrak{M}_n$.

Actually, the Menger Universal Space $M_n^{2n+1}$ is such a space (XI.18.4.12.).

XI.21.6.8. The strategy of the proof of XI.21.6. will be to show that the set $E$ of embeddings of $X$ into $\mathfrak{M}_n$ such that the closure of $f(X)$ in $\mathfrak{M}_n$ is compact is a countable intersection of dense open sets in $BC(X, \mathbb{R}^{2n+1})$. The theorem will then follow from the Baire Category Theorem (XI.21.6.6).

We fix a totally bounded metric $\rho$ on $X$. There is such a metric since $X$ embeds in the Hilbert cube (XI.21.6.6). Actually, the Menger Universal Space $M_n^{2n+1}$ is such a space (XI.18.4.12.).

Let $\mathcal{F}$ be the set of all tuples of the form

$$\mathcal{F} = \{ (i_1, \ldots, i_{n+1}; q_1, \ldots, q_{n+1}) : 1 \leq i_1 < i_2 < \cdots < i_{n+1} \leq 2n+1, \ q_k \in \mathbb{Q} \}.$$ 

Then $\mathcal{F}$ is countable. For each $F \in \mathcal{F}$, define

$$L_F = \{ (x_1, \ldots, x_{2n+1}) \in \mathbb{R}^{2n+1} : x_{ik} = q_k \text{ for } 1 \leq k \leq n+1 \}.$$ 

Then $L_F$ is a hyperplane in $\mathbb{R}^{2n+1}$ of dimension $n$, $L_F \cap \mathfrak{M}_n = \emptyset$, and

$$\bigcup_{F \in \mathcal{F}} L_F = \mathfrak{M}_{n+1} = \mathbb{R}^{2n+1} \setminus \mathfrak{M}_n.$$ 

For each $F \in \mathcal{F}$ and $m \in \mathbb{N}$, let $W_{F,m}$ be the set of all $f \in BC(X, \mathbb{R}^{2n+1})$ such that $f(X)$ is disjoint from $L_F$ and $f$ is an $\mathcal{U}_m$-map, i.e. for every $y \in f(X)$ there is an open neighborhood $V$ of $y$ such that $f^{-1}(V) \subseteq U$ for some $U \in \mathcal{U}_m$. We will show that $W_{F,m}$ is a dense open set in $BC(X, \mathbb{R}^{2n+1})$. This will prove the theorem since

$$E = \bigcap_{F,m} W_{F,m}.$$ 

XI.21.6.9. To show that $W_{F,m}$ is open in $BC(X, \mathbb{R}^{2n+1})$, fix $f \in W_{F,m}$. Since $f(X)$ is compact and disjoint from $L_F$, the distance from $f(X)$ to $L_F$ is a positive number $\eta$. For each $y \in f(X)$, there is an open neighborhood $V_y$ in $\mathbb{R}^{2n+1}$ such that $f^{-1}(V_y)$ is contained in some $U \in \mathcal{U}_m$. Let $\gamma$ be the Lebesgue number of the cover $V = \{ V_y : y \in f(X) \}$ of $f(X)$, i.e. if $B$ is a ball of radius $\leq \gamma$ in $\mathbb{R}^n$ intersecting $f(X)$, then $B \subseteq V_y$ for some $y \in f(X)$, and thus $f^{-1}(B)$ is contained in some $U \in \mathcal{U}_m$.

Set $\delta = \min(\gamma/3, \eta/2)$. If $g \in BC(X, \mathbb{R}^{2n+1})$ with $\|f-g\| < \delta$, then for any $x \in X$ and $a \in L_F$ we have

$$\|g(x) - a\| \geq \|f(x) - a\| - \|f(x) - g(x)\| > \frac{\eta}{2}$$

so $g(X) \cap L_F = \emptyset$. 1307
Let \( y \in \overline{g(X)} \). There is a \( z \in X \) with \( \|y - g(z)\| < \delta \). If \( B \) is the open ball of radius \( \delta \) around \( y \), and \( x \in g^{-1}(B) \), then
\[
\|g(x) - f(z)\| \leq \|g(x) - g(z)\| + \|g(z) - f(z)\| < 3\delta \leq \gamma 
\]
since \( g(x) \) and \( g(z) \) are both in \( B \) and \( \|f - g\| < \delta \). Thus \( g(x) \) is in the open ball \( C \) around \( f(z) \) of radius \( \gamma \), i.e. \( g^{-1}(B) \subseteq f^{-1}(C) \), which is contained in some \( U \in \mathcal{U}_m \). Thus \( g \in W_{F,m} \). So \( W_{F,m} \) is open.

Showing that \( W_{F,m} \) is dense in \( BC(X, \mathbb{R}^{2n+1}) \) takes more work. We need to approximate using vectors in general position.

**XI.21.6.10.** **Proposition.** If \( S \) is a finite union of affine subspaces of \( \mathbb{R}^n \) of dimension \( \leq n - 1 \), then the complement of \( S \) is dense (and open) in \( \mathbb{R}^n \).

**Proof:** The complement of an affine subspace of \( \mathbb{R}^n \) of dimension \( \leq n - 1 \) is dense and open, and in any topological space a finite intersection of dense open sets is dense and open.

In fact, by the Baire Category Theorem, a countable union of affine subspaces of \( \mathbb{R}^n \) of dimension \( \leq n - 1 \) has dense (but usually not open) complement. This also follows from the Sum Theorem XI.21.3.13.

**XI.21.6.11.** **Corollary.** If \( G = \{x_1, \ldots, x_p\} \) is a subset of \( \mathbb{R}^n \) in general position, \( y \in \mathbb{R}^n \), and \( \epsilon > 0 \), then there is an \( x_{p+1} \in \mathbb{R}^n \) with \( \|y - x_{p+1}\| < \epsilon \) and \( \{x_1, \ldots, x_p, x_{p+1}\} \) in general position.

**Proof:** If \( \{x_{i_1}, \ldots, x_{i_r}\} \) is a subset of \( G \) with \( r \leq n \), and \( x \in \mathbb{R}^n \), then \( \{x_{i_1}, \ldots, x_{i_r}, x\} \) span an \( r \)-dimensional affine subspace if and only if \( x \) is not in the \((r-1)\)-dimensional affine subspace spanned by \( \{x_{i_1}, \ldots, x_{i_r}\} \). So we need only choose \( x_{p+1} \) to not be in any of these finitely many affine subspaces.

Applying this result repeatedly, we obtain:

**XI.21.6.12.** **Corollary.** If \( \{x_1, \ldots, x_p\} \) is a subset of \( \mathbb{R}^n \) in general position, \( y_1, \ldots, y_q \in \mathbb{R}^n \), and \( \epsilon > 0 \), then there are \( x_{p+1}, \ldots, x_{p+q} \in \mathbb{R}^n \) such that \( \|x_{p+k} - y_k\| < \epsilon \) for \( 1 \leq k \leq q \) and \( \{x_1, \ldots, x_p, x_{p+1}, \ldots\} \) is in general position.

In fact, if \( \{x_1, \ldots, x_p\} \) is in general position, \( (y_k) \) is any sequence in \( \mathbb{R}^n \), and \( (\epsilon_k) \) is any sequence of positive numbers, there is a sequence \( (x_{p+k}) \) in \( \mathbb{R}^n \) such that \( \|x_{p+k} - y_k\| < \epsilon_k \) for every \( k \) and the infinite set \( \{x_1, \ldots, x_p, x_{p+1}, \ldots\} \) is in general position.

**XI.21.6.13.** Now fix \( F \) and \( m \), and let \( f \in BC(X, \mathbb{R}^{2n+1}) \) and \( \epsilon > 0 \). Let \( \{x_0, \ldots, x_n\} \) be a set of points in general position spanning \( L_F \) as an affine subspace. Cover \( f(X) \) by finitely many open balls of diameter \( \leq \frac{\epsilon}{2} \); by increasing \( m \) if necessary, we may assume \( \mathcal{U}_m \) refines the cover of \( X \) consisting of inverse images of these balls, i.e. that the diameter of \( f(U) \) is less than \( \frac{\epsilon}{2} \) for all \( U \in \mathcal{U}_m \). Let
\[
\mathcal{V} = \{V_1, \ldots, V_q\}
\]
be a shrinking of \( \mathcal{U}_m \) of order \( \leq n \). For each \( k, 1 \leq k \leq q \), fix \( z_k \in V_k \). Choose \( x_{n+1}, \ldots, x_{n+q} \in \mathbb{R}^{2n+1} \) such that \( \|x_{n+k} - f(z_k)\| < \frac{\epsilon}{2} \) for each \( k \) and \( \{x_0, \ldots, x_{n+q}\} \) is in general position.
Let \( \{ h_k : 1 \leq k \leq q \} \) be a partition of unity on \( X \) subordinate to \( V \), e.g. the standard partition of unity from \( \rho (\cdot) \). For \( x \in X \), set

\[
g(x) = \sum_{k=1}^{q} h_k(x) x_{n+k}.
\]

We claim \( g \in W_{F,m} \) and \( \| f - g \| \leq \epsilon \). If \( x \in X \), then for each \( k \) for which \( h_k(x) > 0 \), we have \( x \in V_k \), so \( \| f(x) - f(z_k) \| < \frac{\epsilon}{2} \) and thus \( \| f(x) - x_{n+k} \| < \epsilon \). We thus have

\[
\| f(x) - g(x) \| = \left\| \sum_{k=1}^{q} h_k(x) f(x) - \sum_{k=1}^{q} h_k(x) x_{n+k} \right\|
\]

\[
= \left\| \sum_{k=1}^{q} h_k(x) [f(x) - x_{n+k}] \right\| \leq \sum_{k=1}^{q} h_k(x) \| f(x) - x_{n+k} \| \leq \epsilon h_k(x) = \epsilon
\]

so \( \| f - g \| < \epsilon \).

To show \( g \in W_{F,m} \), we first show \( g(X) \cap L_F = \emptyset \). If \( x \in X \), then since \( V \) has order \( \leq n \), \( g(x) \) is a convex combination of at most \( n + 1 \) of the \( x_{n+k} \), say \( x_{n+i_0}, \ldots, x_{n+i_n} \). Since

\[
\{ x_0, \ldots, x_n, x_{n+i_0}, \ldots, x_{n+i_n} \}
\]

is a set of at most \( 2n + 2 \) vectors in \( \mathbb{R}^{2n+1} \) which is in general position, it is affinely independent; thus the affine subspaces generated by \( \{ x_0, \ldots, x_n \} \) (which is \( L_F \)) and \( \{ x_{n+i_0}, \ldots, x_{i_n} \} \) do not intersect (\( \cdot \)). In particular, the convex hull of \( \{ x_{n+i_0}, \ldots, x_{i_n} \} \) (which is a simplex) does not intersect \( L_F \). There are finitely many such simplexes as \( x \) ranges over \( X \); the union of these simplexes is a compact subset (polyhedron) \( P \) of \( \mathbb{R}^{2n+1} \) containing \( g(X) \) which does not meet \( L_F \). Thus \( g(X) \) does not meet \( L_F \).

If \( y \in P \), let \( \phi_k(y) \) be the barycentric coordinate of \( y \) with respect to \( x_{n+k} \). Then \( \phi_k \) is a continuous function from \( P \) to \( [0,1] \), and \( h_k = \phi_k \circ g \). If \( y \in g(X) \), then since the \( \phi_k(y) \) sum to 1, there is a \( k \) with \( \phi_k(y) > 0 \). Fix such a \( k \), and let \( Z = \{ z \in P : \phi_k(z) > 0 \} \). Then \( Z \) is a neighborhood of \( y \), and \( g^{-1}(Z) \subseteq V_k \).

Thus \( g \in W_{F,m} \).

This completes the proof of Theorem XI.21.6.6.

**XI.21.6.14.** We now turn to the proof of Theorem XI.21.6.1. If \( X \) contains an open subset of \( \mathbb{R}^{n} \), it obviously has dimension \( n \). The proof of the converse involves a somewhat delicate construction, and uses the following functions. If \( a \in \mathbb{R}^{n} \) and \( \epsilon > 0 \), let \( h_{a,\epsilon} \) be the function from \( \mathbb{R}^{n} \setminus \{ a \} \) to \( \mathbb{R}^{n} \) which moves each point of \( \mathbb{R}^{n} \setminus \{ a \} \) radially away from \( a \) a distance \( \epsilon \). This function has the following properties for all \( x, y \in \mathbb{R}^{n} \setminus \{ a \} \), as is easily verified:

(i) \( h_{a,\epsilon} \) is a homeomorphism from \( \mathbb{R}^{n} \setminus \{ a \} \) onto \( \mathbb{R}^{n} \setminus B_\epsilon(a) \).

(ii) \( \| h_{a,\epsilon}(x) - x \| = \epsilon \).

(iii) \( \| x - y \| \leq \| h_{a,\epsilon}(x) - h_{a,\epsilon}(y) \| \leq \| x - y \| + 2\epsilon \).

We will actually prove the following stronger result:
XI.21.6.15. Theorem. Let \( X \) be a subset of \( \mathbb{R}^n \) with empty interior. Then \( X \) is homeomorphic to a subset \( Y \) of \( \mathbb{R}^n \) with empty interior. In particular, \( Y \) is nowhere dense in \( \mathbb{R}^n \) and \( \dim(Y) = \dim(X) \leq n - 1 \).

Proof: By XI.21.2.2., any subset of \( \mathbb{R}^n \) homeomorphic to \( X \) has empty interior. The complement of \( M^n_{n-1} \) in \( \mathbb{R}^n \) is the set of points with all rational coordinates, hence is countable. Let \( \{a_k : k \in \mathbb{N}\} \) be an enumeration of this complement. Set \( Y_1 = X \), and inductively define a sequence \( Y_k \) of subsets of \( \mathbb{R}^n \) and homeomorphisms \( h_k : Y_k \to Y_{k+1} \) as follows. \( Y_k \) has empty interior, so there is a \( b_k \in \mathbb{R}^n \setminus Y_k \) with

\[ \|b_k - a_k\| < \frac{3^{-k}}{4}. \]

Define \( h_k \) to be the restriction of \( h_{b_k,3^{-k}} \) to \( Y_k \), and \( Y_{k+1} = h_k(Y_k) \). We have

\[ \|h_k(y) - a_k\| \geq \|h_k(y) - b_k\| - \|b_k - a_k\| > 3^{-k} - \frac{3^{-k}}{4} = \frac{3^{-k+1}}{4} \]

for all \( y \in Y_k \).

For each \( k \) set \( f_k = h_k \circ \cdots \circ h_1 \). Then \( f_k \) is a homeomorphism from \( X \) onto \( Y_{k+1} \). By XI.21.14.(ii) and the triangle inequality we have, for \( j < k \) and \( x \in X \),

\[ \|f_k(x) - f_j(x)\| \leq \sum_{m=j+1}^{k} 3^{-m} < \sum_{m=j+1}^{\infty} 3^{-m} = \frac{3^{-j}}{2}. \]

Thus the \( f_k \) form a uniform Cauchy sequence and converge uniformly to a continuous function \( f : X \to \mathbb{R}^n \). Set \( Y = f(X) \). For \( x, z \in X \), we have \( \|f(x) - f(z)\| \geq \|x - z\| \) by XI.21.14.(iii), so \( f \) is injective and \( f^{-1}|_Y \) is a contraction, hence continuous; thus \( f \) is a homeomorphism from \( X \) onto \( Y \).

For any \( k \) we have, for any \( x \in X \),

\[ \|f(x) - f_k(x)\| \leq \frac{3^{-k}}{2} \]

and hence

\[ \|f(x) - a_k\| \geq \|f_k(x) - a_k\| - \|f(x) - f_k(x)\| > \frac{3^{-k+1}}{4} - \frac{3^{-k}}{2} = \frac{3^{-k}}{4}. \]

Thus \( a_k \) is not in \( \hat{Y} \), i.e. \( \hat{Y} \subseteq M^n_{n-1} \).

XI.21.7. Infinite-Dimensional Spaces

XI.21.7.1. A topological space can be infinite-dimensional in any of the above senses. There are in fact examples of (normal) spaces which are infinite-dimensional in some senses but not others. For separable metrizable spaces the notion of infinite-dimensionality is unambiguous.

XI.21.7.2. There are degrees of infinite-dimensionality, however. One might expect, for example, that \([0,1]^\kappa\) should be “\(\kappa\)-dimensional.” A more important distinction, however, at least for metrizable spaces, is between “countable-dimensional” and “uncountable-dimensional” spaces:
XI.21.7.3. Definition. Let $X$ be a topological space. Then $X$ is countable-dimensional if it is a countable union of finite-dimensional subspaces. $X$ is countably infinite-dimensional if it is countable-dimensional but not finite-dimensional.

This definition can be made for any of the dimension functions; we will be imprecise as to the meaning of “finite-dimensional” but note this is well-defined for separable metrizable spaces.

XI.21.7.4. By (), a normal space $X$ is countable-dimensional (in the sense of $\text{Ind}$) if and only if it can be written as a countable union of subspaces $X_n$ with $\text{Ind}(X_n) = 0$ for all $n$. $X$ is countably infinite-dimensional if and only if it can be written as a countable union of zero-dimensional subspaces, but not as a finite union of zero-dimensional subspaces.

XI.21.7.5. It is easy to give examples of countably infinite-dimensional separable metrizable spaces: the disjoint union $\sqcup_{n} [0,1]^n$ works. For a compact metrizable example, take the one-point compactification of the disjoint union, or the infinite join if a connected compact metrizable example is desired. It is only slightly harder to string together the cubes to get a compact AR which is countably infinite-dimensional.

Another distinction comes from considering a strengthening of the partition dimension concept:

XI.21.7.6. Definition. Let $X$ be a normal topological space. Then $X$ is strongly infinite-dimensional if there is a sequence of pairs $(A_k, B_k)$ of disjoint closed sets in $X$ such that, if $P_k$ is any partition between $A_k$ and $B_k$ for each $k$, then $\bigcap_{k=1}^{\infty} P_k \neq \emptyset$.

$X$ is weakly infinite-dimensional if $\text{pd}(X) = \infty$ but $X$ is not strongly infinite-dimensional, i.e. for every sequence of pairs $(A_k, B_k)$ of disjoint closed sets in $X$, there are partitions $P_k$ such that $\bigcap_{k=1}^{\infty} P_k = \emptyset$, but for each $n$ there are pairs $(A_1, B_1), \ldots, (A_n, B_n)$ such that, if $P_k$ is any partition between $A_k$ and $B_k$ for each $k$, then $\bigcap_{k=1}^{n} P_k \neq \emptyset$.

XI.21.7.7. Proposition. Every countably infinite-dimensional completely normal space is weakly infinite-dimensional.

Proof: This is a simple adaptation of (). Write $X = \bigcup_{n=1}^{\infty} X_n$, where $X_n$ is zero-dimensional. If $\{ (A_n, B_n) : n \in \mathbb{N} \}$ is a sequence of pairs of disjoint closed sets in $X$, then by XI.21.3.16 there is for each $n$ a partition $P_n$ between $A_n$ and $B_n$ with $P_n \cap X_n = \emptyset$. Then $\bigcap_{n=1}^{\infty} P_n = \emptyset$.

On the contrary, we have:

XI.21.7.8. Theorem. The Hilbert cube $\mathbb{H}$ is strongly infinite-dimensional.

Proof: (cf. [HW41, p. 49]) In $\mathbb{H} = [0,1]^{\mathbb{N}}$, let $A_k$ be the set of all points with $k$’th coordinate 0, and $B_k$ the set of points with $k$’th coordinate 1. Then $A_k$ and $B_k$ are disjoint closed sets in $X$. Suppose $P_k$ is a partition between $A_k$ and $B_k$ for each $k$. For each $n$ let $f_n : [0,1]^n \to \mathbb{H}$ be defined by

$$f_n(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, 0, \ldots) .$$
Fix $n$. If $\tilde{A}_k = f_n^{-1}(A_k)$, $\tilde{B}_k = f_n^{-1}(B_k)$, $\tilde{P}_k = f_n^{-1}(P_k)$ for $1 \leq k \leq n$, then $\tilde{A}_k$ and $\tilde{B}_k$ are just the opposite $k$-faces of $[0,1]^n$, and $\tilde{P}_k$ is a partition in $[0,1]^n$ between $\tilde{A}_k$ and $\tilde{B}_k$, so by the Face Partition Theorem (XI.20.1.19.), $\cap_{k=1}^n \tilde{P}_k \neq \emptyset$; thus $\cap_{k=1}^n \tilde{P}_k \neq \emptyset$ in $X$. This is true for all $n$, so the $P_k$ have the Finite Intersection Property, and thus $\cap_{k=1}^\infty P_k \neq \emptyset$ since the $P_k$ are compact.

**XI.21.7.9.** Corollary. The Hilbert cube is not countable-dimensional.

Since the Hilbert cube is separable and metrizable, it is unambiguous what dimension function we use.

**XI.21.7.10.** It is thus highly questionable whether we should say that $\dim ([0,1]^\aleph_0) = \aleph_0$.

There is a theory of transfinite dimension for (some) countable-dimensional spaces (Exercise XI.21.8.7.).

**XI.21.7.11.** There are many pathologies in infinite-dimensional spaces. For example, if $X$ is $n$-dimensional for finite $n$, then $X$ has closed subspaces of dimension $k$ for all $0 \leq k \leq n$ (cf. the proof of XI.21.3.19.). But there are examples of an infinite-dimensional compact metrizable space $X$ such that every subspace of $X$ is either zero-dimensional or infinite-dimensional (such a space is called hereditarily infinite-dimensional). Using the CH, an $X$ can be constructed with the property that every uncountable subspace is infinite-dimensional (the CH is necessary; existence of such a space cannot be proved in ZFC). Note that every (regular) space has zero-dimensional subspaces, e.g. any countable subspace (XI.21.3.18.).

Examples of the following types of spaces (and many more) can be found in [vM89]:

- A weakly infinite-dimensional compact space which is not countable-dimensional.
- A strongly infinite-dimensional compact space which is hereditarily infinite-dimensional.
- A strongly infinite-dimensional totally disconnected space (XI.21.8.6.; there can be no such compact example by XI.13.6.7.).

All these examples are separable and metrizable.

**XI.21.7.12.** Until quite recently, the structure theory of infinite-dimensional spaces lagged far behind that of finite-dimensional spaces. But there has been a lot of recent progress. See [vM01] and [Eng95] for more details.

**XI.21.8.** Exercises

**XI.21.8.1.** Let $X$ be a normal space.

(a) Use XI.7.6.8. and Urysohn’s Lemma to show that any open cover $\mathcal{U}$ of $X$ has an open shrinking consisting of cozero sets.

(b) Show that $zdim(X) = dim(X)$.
XI.21.8.2. \([vM89]\) This problem gives a direct proof that \(pd(B^n) \leq n\) (cf. XI.21.1.5., XI.21.4.14.): we actually show that if \(X\) is any compact subset of \(\mathbb{R}^n\), then \(pd(X) \leq n\).

(a) Show that if \(D\) is a dense subset of \(\mathbb{R}\), and

\[
D^{(n)} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_k \in D \text{ for at least one } k\}
\]

and if \(A\) and \(B\) are disjoint compact subsets of \(\mathbb{R}^n\), then there is a partition between \(A\) and \(B\) in \(\mathbb{R}^n\) contained in \(D^{(n)}\). [Cover \(A\) with finitely many open rectangular solids which are products of intervals with endpoints in \(D\), whose closures are disjoint from \(B\).]

(b) Show that if \((A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\) are \(n + 1\) pairs of disjoint closed sets in \(X\), then there are partitions \(P_k\) between \(A_k\) and \(B_k\) in \(X\) with \(\bigcap_{k=1}^{n+1} P_k = \emptyset\). [Let \(D_1, \ldots, D_{n+1}\) be \(n + 1\) disjoint dense subsets of \(\mathbb{R}\), and choose \(P_k \subseteq D_k^{(n)}\).]

XI.21.8.3. Show that every topological \(n\)-manifold \(X\) with boundary has \(dim(X) = n\). [Write each component of \(X\) as a countable union of subsets homeomorphic to \(B_n\), and apply \(??\).]

XI.21.8.4. (Local Dimension.) If \(X\) is a topological space and \(x \in X\), define \(ind_x(X)\) inductively: 

\[ind_x(X) \leq n\] if and only if, whenever \(U\) is an open neighborhood of \(x\) in \(X\), there is an open neighborhood \(V\) of \(x\) contained in \(U\) with \(ind(\partial V) \leq n - 1\). Let \(dim_x(X)\) be the minimum of \(dim(U)\) for neighborhoods \(U\) of \(x\).

(a) Show that \(\sup_{x \in X} ind_x(X)\) and, if \(X\) is separable and metrizable, \(\dim(X) = \sup_{x \in X} \dim_x(X)\). [For the second, use the Sum Theorem for \(dim\).]

(b) Show that \(ind_x(X)\) and \(dim_x(X)\) can differ, even if \(X\) is a compact metrizable space. [Consider the one-point compactification of a disjoint union of a sequence of compact metrizable spaces of positive dimension.]

(c) \([?]\) Let \(X\) be a separable metrizable space of dimension \(n\). The \(n\)-dimensional kernel of \(X\) is

\[DK(X) = \{x \in X : ind_x(X) = n\}.
\]

Show that \(DK(X)\) is an \(F_n\) and \(ind(DK(X)) \geq n - 1\). [Write the complement as a union of a set of dimension \(\leq n - 2\) and a set of dimension \(0\). Use the Sum Theorem and the argument of XI.21.3.15.\] Of course, \(ind(DK(X)) \leq n\) by XI.21.3.7.. For “most” \(X\), \(ind(DK(X)) = n\).

(d) \(X\) is \(\text{weakly } n\text{-dimensional} if \(\text{ind}(DK(X)) = n - 1\). Show that the space of \((X)\) is weakly 1-dimensional.

For every \(n\) there exist \(n\)-dimensional spaces. A compact metrizable space cannot be weakly \(n\)-dimensional; in fact, a weakly \(n\)-dimensional space cannot contain a compact subset of dimension \(n\) [HW41, p. 95].

XI.21.8.5. Let \(X\) be a topological space and \(Y\) a (nonempty) subspace of \(X\).

(a) Show that \(\text{ind}(Y) = 0\) if and only if there is a base \(\mathcal{B}\) for the topology of \(X\) such that \(Y \cap \partial U = \emptyset\) for all \(U \in \mathcal{B}\).

(b) Apply XI.1.2.10. to show that if \(X\) is second countable and \(\text{ind}(Y) = 0\), there is a countable base \(\mathcal{B}\) for the topology of \(X\) such that \(Y \cap \partial U = \emptyset\) for all \(U \in \mathcal{B}\).

(c) Show that if \(X\) is second countable and \(\text{ind}(Y) = 0\), then there is a \(G_\delta\) subset \(Y_\delta\) of \(X\) with \(Y \subseteq Y_\delta\) and \(\text{ind}(Y_\delta) = 0\). [Consider \(X \setminus \cup_{U \in \mathcal{B}} \partial U\), where \(\mathcal{B}\) is a countable base as in (b); cf. XI.21.3.15.]

(d) Using the Second Decomposition Theorem XI.21.3.20., prove the Extension Theorem: if \(X\) is second countable, \(Y \subseteq X\), and \(\text{ind}(Y) = n\), then there is a \(G_\delta\) subset \(Y_\delta\) of \(X\) with \(Y \subseteq Y_\delta\) and \(\text{ind}(Y_\delta) = n\).
XI.21.8.6. ([?]; cf. [Eng95, 1.4.F]) Let $K$ be the Cantor set, and $n \in \mathbb{N} \cup \{\infty\}$. Set $X = [0,1]^n$. Let $G$ be the set of $G_\delta$’s in $K \times X$. Then $\text{card}(G) = 2^{8^n}$. Let $\phi$ be a bijection from $K$ to $G$. Fix a point $x_0 \in X$, and define a function $f : K \to X$ by choosing $f(t)$ to be a point such that $(t,f(t)) \notin \phi(t)$ if there is such a point and $f(t) = x_0$ otherwise. Let $Y = \{(t,f(t)) : t \in K\} \subseteq K \times X$ be the graph of $f$; give $Y$ the subspace topology.

(a) Show that projection onto the first coordinate is an injective continuous function from $Y$ onto $K$. Thus $Y$ is totally disconnected. (So $Y$ may be thought of as $K$ with a topology stronger than the usual topology.)

(b) If $G$ is a $G_\delta$ in $K \times X$ containing $Y$, then $G = \phi(t)$ for some $t \in K$, and hence $G$ contains $\{t\} \times X$. Conclude that $\text{ind}(G) = n$.

(c) Apply the Extension Theorem (XI.21.8.5.(d)) to conclude that $\text{ind}(Y) = n$.

Note that $Y$ is separable and metrizable, so all dimension functions coincide on $Y$. This argument uses the AC. $K$ and $X$ can be replaced in the construction by any separable metrizable spaces of cardinality $2^{8^n}$; thus any separable metrizable space of cardinality $2^{8^n}$ is an injective continuous image of a separable metrizable space of arbitrarily larger dimension, i.e. has a stronger separable metrizable topology of arbitrarily larger dimension. In particular, for any $n \in \mathbb{N} \cup \{\infty\}$ there is a separable metrizable topology $T_n$ on $\mathbb{R}$, or $[0,1]$, stronger than the usual topology, with $\dim(\mathbb{R}, T_n) = n$. [Do the construction with $K$ and $X$ replaced by $\mathbb{R}$ and $\mathbb{R}^n$ respectively. To see that the $Y$ has dimension exactly $n$, note that $Y$ cannot contain any nonempty open set in $\mathbb{R} \times \mathbb{R}^n \cong \mathbb{R}^{n+1}$ by Invariance of Domain, and apply XI.21.6.1.]

There is also such a topology on $\mathbb{R}$ for $n = 0$, obtained by expanding the usual topology by making each rational number an isolated point.

XI.21.8.7. [ ] Extend the definitions of $\text{ind}$ and $\text{Ind}$ transfinitely: if $\alpha$ is an ordinal, then $\text{ind}(X) \leq \alpha$ if and only if, whenever $x \in X$ and $U$ is an open neighborhood of $x$ in $X$, there is an open neighborhood $V$ of $x$ contained in $U$ with $\text{ind}(\partial V) < \alpha$, and similarly for $\text{Ind}$.

(a) If $X$ is a $T_1$ space, and $\text{Ind}(X)$ is defined, then $\text{ind}(X)$ is defined and $\text{ind}(X) \leq \text{Ind}(X)$.

(b) If $X$ is a separable metrizable space, and $\text{ind}(X)$ is defined, then $X$ is countable-dimensional. Thus $\text{ind}(X)$ is not defined even for every compact metrizable space; in particular, $\text{ind}(\mathbb{R})$ is not defined. If $X$ is compact, metrizable, and countable-dimensional, then $\text{ind}(X)$ and $\text{Ind}(X)$ are defined.

(c) If $X$ is a separable metrizable space, and $\text{ind}(X)$ is defined, then $\text{ind}(X)$ is a countable ordinal.

(d) For every countable ordinal $\alpha$, there are compact metrizable spaces $X_\alpha$ and $Y_\alpha$ with $\text{ind}(X_\alpha) = \alpha$ and $\text{Ind}(Y_\alpha) = \alpha$. There are compact metrizable spaces $X$ with $\omega < \text{ind}(X) < \text{Ind}(X)$.

(e) If $\text{ind}_x(X)$ is finite for every $x \in X$, then $\text{ind}(X) \leq \omega$. The converse is false, even for compact metrizable $X$. Define a transfinite version of local dimension $\text{ind}_x$, and show that $\text{ind}(X) = \sup_{x \in X} \text{ind}_x(X)$ for any $X$.  

1314
XI.22. Homotopy and the Fundamental Group

XI.22.1. Homotopy

Homotopy is the careful mathematical formulation of the notion of “continuous deformation,” one of the most important ideas in topology and the basis for the entire subject of algebraic topology. There are several slight variations which we will discuss. If \( f \) and \( g \) are continuous functions between topological spaces \( X \) and \( Y \), we want to describe how \( f \) can be “continuously deformed” to \( g \); the intuitive idea is that there should be a “continuous path” \((f_t)\) (say, for \( 0 \leq t \leq 1 \)) of continuous functions with \( f_0 = f \), \( f_1 = g \). Experience shows that the following definition is the appropriate one:

**XI.22.1.1. Definition.** Let \( X \) and \( Y \) be topological spaces, and \( f_0, f_1 : X \to Y \) continuous functions. A homotopy from \( f_0 \) to \( f_1 \) is a continuous function \( H : X \times [0, 1] \to Y \) with \( H(x, 0) = f_0(x) \) and \( H(x, 1) = f_1(x) \) for all \( x \in X \). Two continuous functions \( f \) and \( g \) from \( X \) to \( Y \) are homotopic, written \( f \simeq g \), if there exists a homotopy from \( f \) to \( g \).

**XI.22.1.2.** We can then define, for fixed \( t \in [0, 1] \), a continuous function \( f_t : X \to Y \) by \( f_t(x) = H(x, t) \) for \( x \in X \). The functions \( f_t \) “vary continuously” from \( f_0 \) to \( f_1 \) as \( t \) varies from 0 to 1. The \( f_t \) are “jointly continuous” in \( x \) and \( t \). We often think of the homotopy \( H \) as being the family \((f_t)\) of deformations of \( f_0 \) to \( f_1 \).

**XI.22.1.3.** The choice of \([0, 1]\) for the parametrizing interval is arbitrary, as is the specification that \( H \) goes from \( X \times [0, 1] \) to \( Y \) instead of from \([0, 1] \times X \) to \( Y \) (the opposite convention is also common in references). To add to the potential notational confusion, in many applications the space \( X \) will also be \([0, 1]\).

**XI.22.1.4.** In considering homotopies, the target space \( Y \) must be carefully specified or understood. If \( Y \) is a subspace of a larger space \( Z \), and \( f, g : X \to Y \), then \( f \) and \( g \) can well be homotopic as maps from \( X \) to \( Z \) but not as maps from \( X \) to \( Y \), i.e. a homotopy might necessarily pass through points of \( Z \) outside \( Y \). Thus we should really write \( f \simeq g \) in \( Y \) (although we usually don’t).

**XI.22.1.5.** If \( f \simeq g \) via a homotopy \( H \), then for any \( x \in X \) the function \( \gamma_x : [0, 1] \to Y \) defined by \( \gamma_x(t) = H(x, t) \) is a continuous path in \( Y \) from \( f(x) \) to \( g(x) \). In particular, if \( f \simeq g \), then for every \( x \in X \), \( f(x) \) and \( g(x) \) must be in the same path component of \( Y \).

**XI.22.1.6.** Examples. (i) Two constant functions from \( X \) to \( Y \) with values \( y_0 \) and \( y_1 \) are homotopic if and only if \( y_0 \) and \( y_1 \) are in the same path component of \( Y \). Necessity comes from XI.22.1.5. Conversely, if \( \gamma : [0, 1] \to Y \) is a path from \( y_0 \) to \( y_1 \), set \( H(x, t) = \gamma(t) \); then \( H \) is a homotopy (via constant functions). In particular, if \( Y \) is path-connected, any two constant functions from any \( X \) to \( Y \) are homotopic.

(ii) A very special but important case is when the space \( Y \) is a convex subset of \( \mathbb{R}^n \) or, more generally, of a topological vector space. If \( X \) is any topological space, then any two continuous functions \( f \) and \( g \) from \( X \) to \( Y \) are homotopic by a homotopy of the form

\[
H(x, t) = tg(x) + (1 - t)f(x).
\]
Such a homotopy is called a \textit{linear homotopy}, and is defined only when \( Y \) has additional structure from a topological vector space.

The first fundamental property of homotopy is:

\textbf{XI.22.1.7. PROPOSITION.} Let \( X \) and \( Y \) be topological spaces. Then homotopy is an equivalence relation on the set of maps (continuous functions) from \( X \) to \( Y \).

\textbf{Proof:} Reflexivity and symmetry are very simple, but transitivity, although it is relatively simple also, involves an important notion of concatenation for homotopies.

If \( f : X \rightarrow Y \), define \( H(x,t) = f(x) \) for all \( x, t \). Then \( H \) is a homotopy from \( f \) to \( f \), so homotopy is reflexive.

If \( H \) is a homotopy from \( f_0 \) to \( f_1 \), define \( \tilde{H} : X \times [0,1] \rightarrow Y \) by \( \tilde{H}(x,t) = H(x,1-t) \). Then \( \tilde{H} \) is a homotopy from \( f_1 \) to \( f_0 \). Thus homotopy is symmetric.

Let \( H_1 \) be a homotopy from \( f_0 \) to \( f_1 \), and \( H_2 \) a homotopy from \( f_1 \) to \( f_2 \). Define \( H : X \times [0,1] \rightarrow Y \) by

\[
H(x,t) = \begin{cases} 
H_1(x,2t) & \text{if } 0 \leq t \leq \frac{1}{2} \\
H_2(x,2t-1) & \text{if } \frac{1}{2} < t \leq 1
\end{cases}
\]

Then \( H \) is continuous since \( H_1(x,1) = H_2(x,0) = f_1(x) \) for all \( x \), and is a homotopy from \( f_0 \) to \( f_2 \).

Homotopy respects compositions:

\textbf{XI.22.1.8. PROPOSITION.} Let \( X, Y, Z \) be topological spaces, and let \( f, f_0, f_1 : X \rightarrow Y \) and \( g, g_0, g_1 : Y \rightarrow Z \) with \( f_0 \simeq f_1 \) and \( g_0 \simeq g_1 \). Then \( g_0 \circ f_0 \simeq g_1 \circ f_1 \). In particular, \( g \circ f_0 \simeq g \circ f_1 \) and \( g_0 \circ f \simeq g_1 \circ f \).

\textbf{Proof:} We need only prove the last statement, since then \( g_0 \circ f_0 \simeq g_0 \circ f_1 \) and \( g_0 \circ f_1 \simeq g_1 \circ f_1 \), so \( g_0 \circ f_0 \simeq g_1 \circ f_1 \) by \textbf{XI.22.1.7.}.

If \( H \) is a homotopy from \( f_0 \) to \( f_1 \), define \( \tilde{H}(x,t) = g(H(x,t)) \); then \( \tilde{H} \) is a homotopy from \( g \circ f_0 \) to \( g \circ f_1 \).

If \( H \) is a homotopy from \( g_0 \) to \( g_1 \), define \( \tilde{H}(x,t) = H(f(x),t) \); then \( \tilde{H} \) is a homotopy from \( g_0 \circ f \) to \( g_1 \circ f \).

\textbf{Contractible Spaces}

\textbf{XI.22.1.9. DEFINITION.} If \( X \) and \( Y \) are topological spaces, and \( f : X \rightarrow Y \) is a map, then \( f \) is \textit{nullhomotopic} if it is homotopic to a constant function.

A topological space \( X \) is \textit{contractible} if the identity map on \( X \) is nullhomotopic, i.e. homotopic to a constant function from \( X \) to \( X \).

\textbf{XI.22.1.10.} If \( X \) is contractible and \( H \) is a homotopy from the identity map \( \iota_X \) to a constant function with value \( p \), then the functions \( f_t \) “contract” the space down to the point \( p \), and we say \( X \) is \textit{contractible} to \( p \). But a contractible space is path-connected since every point of \( X \) can be connected to \( p \) by a path, so if \( X \) is contractible to \( p \) it is also contractible to \( q \) for any other point \( q \).
XI.22.1.11. **Examples.** (i) If $X$ is a convex subset of $\mathbb{R}^n$, or more generally of a topological vector space, then $X$ is contractible to any point via a linear homotopy (XI.22.1.6.(ii)). In particular, $[0,1]$ is contractible, as is any ball in $\mathbb{R}^n$.

(ii) Let $X$ be the infinite comb $(\cdot)$. Then $X$ is contractible by first moving all points down to the $x$-axis and then moving them to $0$.

(iii) It is a fundamental theorem of topology, equivalent to the Brouwer fixed-point theorem, that $S^n$ is not contractible for any $n$ $(\cdot)$. Note that $S^n$ is path-connected for $n > 0$.

The next result is a special case of XI.22.1.8.

XI.22.1.12. **Proposition.** Let $X$ and $Y$ be topological spaces. If either $X$ or $Y$ is contractible, then every map from $X$ to $Y$ is nullhomotopic. If $Y$ is path-connected, any two maps from $X$ to $Y$ are homotopic.

**Homotopy Relative to a Subspace**

An important variation is homotopy relative to a subspace:

XI.22.1.13. **Definition.** Let $X$ and $Y$ be topological spaces, and $A \subseteq X$, $B \subseteq Y$. Let $f_0, f_1$ be maps from $X$ to $Y$ for which $f_0(A), f_1(A) \subseteq B$. Then a *homotopy from $f_0$ to $f_1$ relative to $(A,B)$* is a continuous function $H : X \times [0,1] \to Y$ such that $H(x,0) = f_0(x)$ and $H(x,1) = f_1(x)$ for all $x \in X$ and $H(x,t) \in B$ for all $x \in A$ and all $t$. If there is such a homotopy, write $f_0 \simeq f_1 \text{ rel } (A,B)$.

XI.22.1.14. A homotopy relative to a subspace is a homotopy. But the notion is much more restrictive, since the image of the set $A$ must remain within $B$ under the entire homotopy. In many applications, $A$ and $B$ are sets with only one or two points, and the restriction that $f$ take $A$ into $B$ amounts to specifying the value of $f$ at the point(s) of $A$.

XI.22.1.15. An important example of relative homotopy is when $X = [0,1]$ and $A = \{0,1\}$. A map $\gamma$ from $[0,1]$ to $Y$ is a path in $Y$. If $\gamma_0$ and $\gamma_1$ are paths in $Y$ with the same initial point $p$ and final point $q$, then $\gamma_0$ and $\gamma_1$ are always homotopic since they are both homotopic to the constant map with value $p$; but if $B = \{p,q\}$ a homotopy rel $(A,B)$ is one in which all intermediate paths also go from $p$ to $q$ (as long as $p$ and $q$ have disjoint neighborhoods in $Y$), so $\gamma_0$ and $\gamma_1$ may not be homotopic rel $(A,B)$. For example, if $Y$ is the unit circle in $\mathbb{R}^2$ and $\gamma_0$ and $\gamma_1$ are paths in $Y$ from $p = (-1,0)$ to $q = (1,0)$ via the upper and lower semicircles respectively, then $\gamma_0$ and $\gamma_1$ are not homotopic rel $(\{0,1\},\{p,q\})$ (this seems intuitively obvious, but requires a nontrivial proof, and is equivalent to the Brouwer fixed-point theorem in the plane). In this case we often notationally suppress the $B$ (if $\gamma_0$ and $\gamma_1$ have the same endpoints) and simply write $\gamma_0 \simeq \gamma_1$ rel $\{0,1\}$. Even the $A$ is often understood in this context, but a reader must carefully understand whether it is assumed (as it most commonly is) that homotopies of paths fix endpoints.

XI.22.1.16. An especially important special case of XI.22.1.15. comes when $p = q$, i.e. $B$ is a singleton $p$. We thus consider *loops with basepoint $p$*, maps $\gamma : [0,1] \to Y$ with $\gamma(0) = \gamma(1) = p$. Two such loops are *homotopic rel $p$* if they are homotopic rel $(\{0,1\},\{p\})$, i.e. homotopic via loops $\gamma_t$ with basepoint $p$ for all $t$. These will be discussed in detail in (\cdot).
XI.22.1.17. Here is another interesting special case. If $X$ is a topological space and $p \in X$, then $X$ is contractible to $p$ relative to $p$ (rel $p$) if the identity map is homotopic to the constant map with value $p$ relative to $(\{p\}, \{p\})$, i.e. if the space can be contracted to $p$ in a way leaving $p$ fixed. A convex set $X$ in a topological vector space is contractible to $p$ rel $p$ for any $p \in X$, but if $X$ is the infinite comb and $p$ is a point on the $y$-axis other than $0$, then $X$ is not contractible to $p$ rel $p$ since all points must move through $0$ in any contraction to $p$ (XI.22.5.1.).

A slight variation of the proofs of XI.22.1.7. and XI.22.1.8. (left to the reader) show:

XI.22.1.18. **Proposition.** Let $X, Y, Z$ be topological spaces, and $A \subseteq X$, $B \subseteq Y$, $C \subseteq Z$. Then

(i) Homotopy rel $(A, B)$ is an equivalence relation on the set of maps from $X$ to $Y$.

(ii) If $f_0, f_1 : X \to Y$, $f_0 \simeq f_1$ rel $(A, B)$, and $g_0, g_1 : Y \to Z$, $g_0 \simeq g_1$ rel $(B, C)$, then $g_1 \circ f_1 \simeq g_0 \circ f_0$ rel $(A, C)$.

**Homotopy Equivalence**

XI.22.1.19. **Definition.** Let $X$ and $Y$ be topological spaces. Then $X$ and $Y$ are homotopy equivalent if there are maps $f : X \to Y$ and $g : Y \to X$ with $g \circ f \simeq \iota_X$ and $f \circ g \simeq \iota_Y$. $X$ homotopically dominates $Y$ if there are maps $f : X \to Y$ and $g : Y \to X$ with $f \circ g \simeq \iota_Y$.

XI.22.1.20. It is an easy consequence of XI.22.1.8. that homotopy dominance is transitive and homotopy equivalence is an equivalence relation. Homotopy dominance is a sort of “homotopy retraction” ()

XI.22.1.21. **Examples.** (i) Every topological space homotopically dominates any contractible space. A topological space is homotopy equivalent to a contractible space if and only if it is contractible. Any two contractible spaces are homotopy equivalent.

(ii) A circle in $\mathbb{R}^2$ is not homotopy equivalent to $\mathbb{R}^2$ since $\mathbb{R}^2$ is contractible. But a circle is homotopy equivalent to $\mathbb{R}^2$ with a point removed.

(iii) The greek letter $\theta$ and a figure 8 are homotopy equivalent.

(iv) A cylinder and a Möbius strip are homotopy equivalent since each is homotopy equivalent to a circle.

(v) If $X$ and $Y$ are homotopy equivalent, they each homotopy dominate the other. But the converse is false in general:

XI.22.1.22. Homotopy equivalence is much weaker than homeomorphism in general: for example, every contractible space is homotopy equivalent to a one-point space. (Homeomorphism can be concluded from homotopy equivalence in some special cases.) Most methods and results of algebraic topology do not distinguish between homotopy equivalent spaces, i.e. algebraic topology describes and studies topological spaces up to homotopy equivalence.
XI.22.1.23. We can also define relative homotopy equivalence. If \( A \subseteq X \) and \( B \subseteq Y \), a homotopy equivalence between \( X \) and \( Y \) rel \((A, B)\) is a pair of maps \( f : X \to Y \) and \( g : Y \to X \) with \( f(A) \subseteq B \), \( g(B) \subseteq A \), and \( g \circ f \simeq \iota_X \) rel \((A, A)\) and \( f \circ g \simeq \iota_Y \) rel \((B, B)\). Relative homotopy dominance can be defined similarly.

Smooth Homotopy

XI.22.1.24. In situations where there is a notion of differentiability, one also wants to consider smooth homotopies. For example, if \( X \) and \( Y \) are \( C^r \)-manifolds and \( f, g \) are \( C^r \) maps from \( X \) to \( Y \), we say \( f \) and \( g \) are \( C^r \)-homotopic if there is a homotopy \( H \) which is a \( C^r \) function from \( X \times [0, 1] \) to \( Y \). The relation between ordinary homotopies and smooth homotopies in this context is somewhat subtle, but in good cases an ordinary homotopy can be “smoothed” to a smooth one.

Isotopy

In some applications in topology, such as knot theory, it is important not to allow the image to “pass through itself” during the deformation. This is accomplished by the notion of isotopy.

XI.22.1.25. Definition. Let \( X \) and \( Y \) be topological spaces, and \( f_0, f_1 \) topological embeddings of \( X \) into \( Y \) (homeomorphisms onto their images). An isotopy from \( f_0 \) to \( f_1 \) is a homotopy \( H = (f_t) \) from \( f_0 \) to \( f_1 \) in which each \( f_t \) (\( 0 \leq t \leq 1 \)) is a topological embedding.

Isotopic maps are obviously homotopic, but the converse is false:

XI.22.1.26. Examples. (i) Let \( X = \{0, 1\}, Y = \{0, 1\} \) (or \( Y = \mathbb{R} \)). Let \( f_0(t) = t \), and \( f_1(t) = 1 - t \). Then \( f_0 \simeq f_1 \) (any two maps from \( X \) to \( Y \) are homotopic). But \( f_0 \) and \( f_1 \) are not isotopic: if \( (f_t) \) is a homotopy from \( f_0 \) to \( f_1 \), there is a \( t \) such that \( f_t(0) = f_t(1) \), so \( f_t \) is not injective [set \( \phi(t) = f_t(1) - f_t(0) \) and apply the Intermediate Value theorem].

(ii) Any two knots are homotopic as maps from \( S^1 \) to \( \mathbb{R}^3 \). But a nontrivial knot like the trefoil is not isotopic to an unknot. In fact, two knots are isotopic if and only if they are equivalent as knots.

(iii) The Fox-Artin wild arc is isotopic to an ordinary arc as maps from \( [0, 1] \) into \( \mathbb{R}^3 \). But the more complicated embedding of \( [0, 1] \) into \( \mathbb{R}^3 \) outlined at the end of (i) which is wild at every point is not isotopic to an ordinary arc.

Ambient Isotopy

XI.22.2. The Fundamental Group

XI.22.3. Covering Spaces

XI.22.4. Other Homotopy Groups

XI.22.5. Exercises

XI.22.5.1. Let \( X \) be the infinite comb, and \( p \) a point of \( X \) on the \( y \)-axis with positive \( y \)-coordinate. Then \( X \) is contractible to \( p \). Show that \( X \) is not contractible to \( p \) rel \( p \).
(a) If \( q \) is a point of \( X \) with positive \( x \)-coordinate, then any path in \( X \) from \( q \) to \( p \) must pass through \( 0 \).

(b) Let \( (p_n) \) be a sequence in \( X \) of points with positive \( x \)-coordinate with \( p_n \to p \). If \( H \) is a contraction of \( X \) to \( p \), then for each \( n \) there must be a \( t_n \) with \( H(p_n, t_n) = 0 \).

(c) If \( t_0 \) is a limit point of the sequence \( (t_n) \), then \( H(p, t_0) = 0 \). Thus we cannot have that \( H \) is a contraction of \( X \) to \( p \) rel \( p \).
XI.22.6. Fubini’s Differentiation Theorem

Although even uniformly convergent infinite series of functions cannot be differentiated term-by-term in general, there are a few types of series for which term-by-term differentiation is valid. The most notable example is convergent power series. Another example is convergent series of nondecreasing functions, the topic of this subsection.

The main theorem is due to Fubini. The name “Fubini’s Theorem” is, however, normally reserved for another of Fubini’s results on interchanging the order of multiple integration. This result is sometimes called “Fubini’s Differentiation Theorem,” a name we will use. This theorem can be stated for sequences (as can any result about infinite series, since the theories of sequences and series are formally equivalent), but the hypotheses are clumsy to state in the sequence setting (cf. Exercise XI.22.7.1.).

XI.22.6.1. Theorem. [Fubini’s Differentiation Theorem] Let \( \sum_{k=1}^{\infty} f_k(x) \) be an infinite series of nondecreasing functions on an interval \([a,b]\), converging pointwise to a (nondecreasing) function \( f \). Then

(i) The series converges uniformly to \( f \) on \([a,b]\).

(ii) The series \( \sum_{k=1}^{\infty} f'_k(x) \) converges pointwise a.e. on \([a,b]\) to \( f'(x) \).

In particular, \( f'_k(x) \to 0 \) for almost all \( x \in [a,b] \).

(Conclusion (ii) is the main part of the theorem; (i) is an easy observation.)

Proof: Begin with (i). For any \( x \in [a,b] \), we have

\[
f(x) = [f(x) - f(a)] + f(a) = \sum_{k=1}^{\infty} [f_k(x) - f_k(a)] + \sum_{k=1}^{\infty} f_k(a)
\]

with both series convergent. If \( \epsilon > 0 \), choose \( N \) such that for all \( n \geq N \),

\[
\left| \sum_{k=n+1}^{\infty} [f_k(b) - f_k(a)] \right| < \frac{\epsilon}{2} \quad \text{and} \quad \left| \sum_{k=n+1}^{\infty} f_k(a) \right| < \frac{\epsilon}{2}.
\]

Then, for any \( n \geq N \) and any \( x \in [a,b] \), we have

\[
\left| f(x) - \sum_{k=1}^{n} f_k(x) \right| = \left| \sum_{k=n+1}^{\infty} [f_k(x) - f_k(a)] + \sum_{k=n+1}^{\infty} f_k(a) \right|
\]

\[
\leq \left| \sum_{k=n+1}^{\infty} [f_k(x) - f_k(a)] \right| + \left| \sum_{k=n+1}^{\infty} f_k(a) \right| \leq \left| \sum_{k=n+1}^{\infty} [f_k(b) - f_k(a)] \right| + \sum_{k=n+1}^{\infty} f_k(a) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon
\]

(the middle inequality on the last line holds because \( f_k \) is nondecreasing) and thus the series converges uniformly to \( f \) on \([a,b]\).
We now turn to (ii). First assume all the \( f_k \) are nonnegative. There is a subset \( A \) of \([a, b]\) whose complement has measure 0 and such that \( f \) and all the \( f_k \) are differentiable at all \( x \in A \). Let \( s_n(x) = \sum_{k=1}^{n} f_k(x) \) be the \( n \)'th partial sum and \( r_n(x) = f(x) - s_n(x) = \sum_{k=n+1}^{\infty} f_k(x) \) the remainder. Then \( s_n \) and \( r_n \) are nonnegative, nondecreasing, and differentiable on \( A \), with \( s'_n(x) \geq 0 \) and \( r'_n(x) \geq 0 \) for all \( x \in A \). Since \( s'_n(x) \leq s'_n(x) + r'_n(x) = f'(x) \) for all \( x \in A \), we have that \( \sum_{k=1}^{\infty} f'_k(x) \) converges for all \( x \in A \) since it is a nonnegative series with bounded partial sums, and

\[
\sum_{k=1}^{\infty} f'_k(x) \leq f'(x)
\]

for all \( x \in A \).

To show that equality holds a.e. on \([a, b]\), it suffices to find a subsequence \((s'_{n_j})\) which converges to \( f' \) a.e. on \([a, b]\). Since \( s_n(b) \to f(b) \), for each \( j \) there is an \( n_j > n_j-1 \) such that \( r_{n_j}(b) = f(b) - s_{n_j}(b) < 2^{-j} \) and hence, for each \( x \in [a, b] \), we have

\[
0 \leq \sum_{j=1}^{\infty} r_{n_j}(x) = \sum_{j=1}^{\infty} (f(b) - s_{n_j}(b)) < \infty
\]

and thus the series \( \sum_{j=1}^{\infty} r_{n_j}(x) \) satisfies the hypotheses in the preceding paragraph, so there is a subset \( B \) of \([a, b]\) such that \([a, b] \setminus B \) has measure 0 and

\[
0 \leq \sum_{j=1}^{\infty} r'_{n_j}(x) = \sum_{j=1}^{\infty} [f'(x) - s'_{n_j}(x)] < \infty
\]

for all \( x \in B \). Thus, for every \( x \in B \), we have \( \lim_{j \to \infty} [f'(x) - s'_{n_j}(x)] = 0 \), so

\[
\lim_{j \to \infty} s'_{n_j}(x) = f'(x)
\]

for almost all \( x \in [a, b] \).

Now drop the assumption that the \( f_k \) are nonnegative. (The only place this assumption was used was to obtain that the \( r_n \) were nonnegative.) Set \( g(x) = f(x) - f(a) \) and \( g_k(x) = f_k(x) - f_k(a) \) for each \( k \) and each \( x \in [a, b] \). Then the \( g_k \) are nonnegative and nondecreasing, \( g' = f' \), \( g'_k = f'_k \) for all \( k \), and \( g(x) = \sum_{k=1}^{\infty} g_k(x) \) for all \( k \). Thus the previous argument applies to \( g \) and the \( g_k \), and shows that

\[
f'(x) = g'(x) = \sum_{k=1}^{\infty} g'_k(x) = \sum_{k=1}^{\infty} f'_k(x) \text{ a.e.}
\]
XI.22.7. Exercises

XI.22.7.1. Give a correct restatement of the Fubini Differentiation Theorem for sequences. Show that the following is not a correct statement:

“Let \((f_n)\) be a nondecreasing sequence of nondecreasing functions on an interval \([a, b]\) converging uniformly on \([a, b]\) to a (nondecreasing) function \(f\). Then \(f'_n \to f'\) pointwise a.e. on \([a, b]\).”

Consider a strictly increasing sequence \((f_n)\) of nondecreasing step functions converging uniformly to \(f(x) = x\) on \([0, 1]\). Modify this example to make the \(f_n\) continuous by connecting the steps by line segments of large slope. In these examples, there is no \(x \in [0, 1]\) for which \(f_n'(x) \to f'(x)\).

XI.22.7.2. Let \(g\) be the Cantor function \((\)\). Regard \(g\) as a function on \(\mathbb{R}\) by setting \(g(x) = 0\) for \(x < 0\) and \(g(x) = 1\) for \(x > 1\). Then \(g\) is nonnegative, continuous, nondecreasing, \(g(0) = 0\), \(g(1) = 1\), and \(g' = 0\) a.e. Let \([a_k, b_k] : k \in \mathbb{N}\) be an enumeration of the closed bounded intervals with rational endpoints \(a_k < b_k\).

(a) For each \(k\) let \(f_k\) be a scaled copy of \(g\) transferred to \([a_k, b_k]\):

\[
f_k(x) = 2^{-k} f\left(\frac{x - a_k}{b_k - a_k}\right).\]

Show that \(f_k\) is nonnegative, continuous, nondecreasing, \(f_k(x) = 0\) for \(x \leq a_k\), \(f_k(x) = 2^{-k}\) for \(x \geq b_k\), and \(f'_k = 0\) a.e.

(b) Set \(f(x) = \sum_{k=1}^{\infty} f_k(x)\). Show that the series converges uniformly on all of \(\mathbb{R}\), and that \(f\) is continuous and strictly increasing on \(\mathbb{R}\).

(c) Apply XI.22.6.1. to conclude that \(f' = 0\) a.e.

XI.22.7.3. Move to section on series of functions

(a) Let \(\{a_k : k \in \mathbb{N}\}\) be an enumeration of \(\mathbb{Q} \cap [0, 1]\). Show that the infinite series \(\sum_{k=1}^{\infty} \frac{3}{k!} \sqrt{x - a_k}\) converges uniformly on \([0, 1]\) to a function \(g\) which is continuous and strictly increasing.

(b) Let \([a, b]\) be the range of \(g\), and \(f = g^{-1} : [a, b] \to [0, 1]\). Show that \(f\) is continuous, strictly increasing, and that \(f' = 0\) on the dense set \(g(\mathbb{Q})\).

(c) Show that \(g\) is differentiable at every irrational number \(x\) in \([0, 1]\) and that \(g'(x) > 1\). Thus \(f\) is differentiable at every point of \([a, b]\) and \(0 \leq f'(x) < 1\) for all \(x \in [a, b]\).

This example is due to D. Pompeiu (cf. [vRS82, 13.3] or [Str81, p. 215-216]). For this \(f\), the set \(\{x \in [a, b] : f'(x) = 0\}\), while dense in \([a, b]\), is countable. There are more dramatic examples, e.g. there is a strictly increasing continuous function \(f\) on \([a, b]\) such that \(f' = 0\) a.e. (Exercise XI.22.7.2.).

XI.22.7.4. Let \(K \subseteq [0, 1]\) be a Cantor set of positive measure, and let \(g\) be a nonnegative real-valued continuous function on \([0, 1]\) which is zero precisely on \(K\), e.g. \(g(x)\) is the distance from \(x\) to \(K\). Set

\[
f(x) = \int_0^x g(t) \, dt.\]

Then \(f\) is strictly increasing on \([0, 1]\), differentiable everywhere with \(f' = g\) (\(), i.e. \(C^1\), and \(f' = 0\) on a set of positive measure.

1323
XI.22.7.5. Does there exist a strictly increasing differentiable function \( f \) on an interval \([a, b]\) such that the set
\[
S = \{ x \in [a, b] : f'(x) = 0 \}
\]
satisfies that \( S \cap [c, d] \) has positive measure for all \( a \leq c < d \leq b \)? (Compare XI.22.7.3., XI.22.7.4.)

XI.22.7.6. Let \( f_k(x) = \frac{1}{k} \tan^{-1} kx - \frac{1}{k+1} \tan^{-1}(k+1)x \).
(a) Show that \( f_k \) is strictly increasing on \([-1, 1]\).
(b) Show that \( \sum_{k=1}^{\infty} f_k(x) \) converges (uniformly) on \([-1, 1]\) to a differentiable function \( f \).
(c) Although \( f \) and each \( f_k \) are differentiable everywhere, it is not true that \( f'(x) = \sum_{k=1}^{\infty} f'_k(x) \) everywhere (we do not have \( f'(0) = \sum_{k=1}^{\infty} f'_k(0) \)).
XI.23. Sard’s Theorem

Sard’s Theorem in its various forms is one of the most remarkable properties of Euclidean space with its differentiable structure, and is perhaps the fundamental result of the subject of differential topology.

Consider a smooth map $f$ from $\mathbb{R}^n$ to $\mathbb{R}^m$, or, more generally, from a smooth $n$-manifold $N$ to a smooth $m$-manifold $M$. Then, for any $x_0 \in \mathbb{R}^n$, $Df(x_0)$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. The point $x_0$ is singular if $Df(x_0)$ has rank strictly smaller than $m$; otherwise, $x_0$ is regular. The image in $M$ of the set of singular points of $f$ is called the set of singular values of $f$; a $y_0$ in the range of $f$ which is not a singular value is called a regular value of $f$. (The image of a regular point can be a singular value if it is also the image of a singular point.) The names reflect the fact that $f$ is well-behaved in a neighborhood of a regular point but can be quite pathological near a singular point (cf. ()). Points of $M$ not in the range of $f$ are also (rather perversely) usually called regular values of $f$. Every critical value of $f$ is a singular value; the converse is not necessarily true if $m > 1$.

Although the set of singular points of $f$ can be a large subset of $\mathbb{R}^n$ ($f$ could even be a constant function!), Sard’s Theorem states that the set of singular values of $f$ is a “negligibly small” subset of $\mathbb{R}^m$ (or $M$), in a certain precise sense (measure zero and meager). Thus, in many situations, it can be concluded that “almost all” values of $f$ are regular values; even the existence of one regular value for $f$ can sometimes have important consequences (cf. ()).

Sard’s Theorem is intuitively quite plausible, and relatively easy to prove, if $n \leq m$, and only quite mild differentiability assumptions on $f$ are needed. In the case $n < m$, where every point of $N$ is a singular point (since the rank of $Df(x_0)$ is always $\leq n$ for any $x_0$), it simply says that the range of $f$ is a “small” subset of $M$ (intuitively it has dimension $\leq n$); although the existence of “space-filling curves” (cf.) can shake one’s confidence about this, Sard’s Theorem in the case $n < m$ says that there can be no smooth space-filling curve or lower-dimensional surface. And in the case $n = m = 1$, the statement says that for a continuously differentiable function, “most” horizontal lines are not tangent to the graph.

In the case $n > m$, or $n = m > 1$, the result is also intuitively plausible: if $x_0$ is a singular point, then near $x_0$ the function $f$ is approximately a linear function whose range is a proper subspace of $\mathbb{R}^m$, and such a subspace is negligibly small in a measure or category sense, and thus a small neighborhood of $x_0$ should have an image of “much smaller” size in $\mathbb{R}^m$. However, this intuitive idea turns out not to be directly translatable into a proof in this case, and is even false in general unless some delicate smoothness conditions are assumed (the underlying intuitive reason for the smoothness conditions is closely related to $V.10.6.11.$). The fact that the conclusion holds in the case $n > m$ is rather subtle and quite remarkable, and has all sorts of important consequences, e.g. it easily implies the Brouwer Fixed-Point Theorem, which is the most fundamental fact about the topology of Euclidean space, as well as underlying transversality, the fundamental technical tool of differential topology.

Sard’s Theorem is often called the Morse-Sard Theorem since the case $m = 1$ was proved by A. P. Morse in 1939 [Mor39]; the general theorem was proved by A. Sard in 1942 [Sar42]. Sard, a student of Marston Morse, obtained some of the results independently of A. P. Morse in his Ph.D. dissertation of 1936. Note that A. P. Morse was not the same person as Marston Morse (they were apparently unrelated). Sard states in [Sar42] that Marston Morse previously proved the theorem of A. P. Morse in the cases $n \leq 6$, in part jointly with Sard (unpublished). A closely related result was obtained in [Bro35]. There have been various generalizations and refinements, cf. [Fed69], [Yom83], [Bat93].
XI.23.1. Sets of Measure Zero

The notion of a set of measure zero is really a concept from measure theory (). But little if any actual measure theory is needed to develop the definition and basic properties needed in this section.

XI.23.1.1. Definition. A subset $A \subseteq \mathbb{R}^n$ has measure zero in $\mathbb{R}^n$ if, for every $\epsilon > 0$, there is a sequence $(B_k)$ of open balls in $\mathbb{R}^n$ such that $A \subseteq \bigcup_k B_k$ and

$$
\sum_{k=1}^{\infty} r_k^n < \epsilon
$$

where $r_k$ is the radius of $B_k$. We allow empty balls, i.e. open balls of radius 0; thus the collection $\{B_k\}$ may be finite or countably infinite.

Note that up to a constant factor, the sum $\sum_{k=1}^{\infty} r_k^n$ is just the sum of the $n$-dimensional volumes $\sum_{k=1}^{\infty} V_n(B_k)$ of the $B_k$. By adjusting $\epsilon$, the sum of the volumes can be used instead in the definition.

XI.23.1.2. We should properly say “$A$ has $n$-dimensional Lebesgue measure zero” instead of “$A$ has measure zero in $\mathbb{R}^n$.” Note that the “$in \mathbb{R}^n$” is a crucial part of the terminology: it does not make sense to simply say that a subset of Euclidean space has measure zero without specifying which Euclidean space it is with respect to. For example, the interval $[0, 1]$ does not have measure zero in $\mathbb{R}^1$, but if it is regarded as a segment on the $x$-axis it does have measure zero in $\mathbb{R}^2$. The term null set is often used for a set of measure zero.

XI.23.1.3. A ball of radius $r$ in $\mathbb{R}^n$ is contained in a cube of side length $2r$ with sides parallel to the coordinate axes; conversely, a cube of side length $2s$ is contained in the ball of radius $s / \sqrt{n}$ around its center. So the definition can be equivalently rephrased replacing “open ball” with “open cube” and “radius” with “side length.” The cubes can be restricted to ones with sides parallel to the axes, or not.

An open ball intersecting $A$ is contained in an open ball of twice the radius centered at a point of $A$. Thus all the balls in the definition may be taken with centers in $A$. The same consideration can be made with cubes.

There is much more flexibility. If $B$ is a ball, there is a ball with rational radius with the same center whose radius is less than twice the radius of $B$; thus we may restrict to using open balls with rational radius. Similarly, open cubes with rational edges can be used. In a similar manner, we may restrict to using only open balls or cubes whose centers have rational coordinates (and rational radii or sides).

Another especially useful variation is to use open rectangular solids. For if $R$ is an open rectangular solid, $R$ can be covered by a finite number of open cubes whose total volume is less than $2V_n(R)$. Thus we obtain:

XI.23.1.4. Proposition. Let $A \subseteq \mathbb{R}^n$. Then $A$ has measure zero in $\mathbb{R}^n$ if and only if, for every $\epsilon > 0$, there is a sequence $(R_k)$ of open rectangular solids in $\mathbb{R}^n$ with $A \subseteq \bigcup_k R_k$ and

$$
\sum_{k=1}^{\infty} V_n(R_k) < \epsilon
$$

As before, some of the $R_k$ may be empty, i.e. there may be either a finite or countably infinite number of $R_k$. 
XI.23.1.5. **Proposition.** (i) Any subset of a subset of \( \mathbb{R}^n \) of measure zero in \( \mathbb{R}^n \) has measure zero in \( \mathbb{R}^n \).

(ii) A countable union of subsets of \( \mathbb{R}^n \) of measure zero in \( \mathbb{R}^n \) has measure zero in \( \mathbb{R}^n \).

**Proof:** Part (i) is obvious. For (ii), let \( (A_j) \) be a sequence of subsets of \( \mathbb{R}^n \) of measure zero in \( \mathbb{R}^n \), and \( A = \bigcup_j A_j \). Let \( \epsilon > 0 \). For each \( j \) let \( (B_{jk}) \) be a sequence of open balls in \( \mathbb{R}^n \) covering \( A_j \) for which the sum of the volumes is less than \( 2^{-j} \epsilon \). Then the \( B_{jk} \) cover \( A \) and the sum of the volumes is less than \( \epsilon \).

XI.23.1.6. **Corollary.** Let \( A \subseteq \mathbb{R}^n \). Then \( A \) has measure zero in \( \mathbb{R}^n \) if and only if every \( x_0 \in A \) has a neighborhood \( U \) such that \( A \setminus U \) has measure zero in \( \mathbb{R}^n \).

**Proof:** This is really the Lindelöf property of \( \mathbb{R}^n \). If \( x_0 \in A \) and \( U \) is any neighborhood of \( x_0 \), then there is a ball \( V \) of rational radius, whose center has rational coordinates, with \( x_0 \in V \subseteq U \). If \( A \cap U \) has measure zero in \( \mathbb{R}^n \), so does \( A \cap V \). There are only countably many distinct such \( V \), so \( A \) is a countable union of sets of measure zero in \( \mathbb{R}^n \). The converse is trivial.

XI.23.1.7. **Examples.** (i) Let \( S \) be the \((n - 1)\)-dimensional subspace
\[
\{(x_1, \ldots, x_{n-1}, 0) : x_1, \ldots, x_{n-1} \in \mathbb{R}\}
\]
of \( \mathbb{R}^n \). Then \( S \) is a countable union of bounded subsets. Each bounded subset can be contained in a single open rectangular solid whose thickness is so small that its volume is arbitrarily small. Thus \( S \) has measure zero in \( \mathbb{R}^n \).

(ii) No (nonempty) open rectangle in \( \mathbb{R}^n \) can have measure zero in \( \mathbb{R}^n \). This fact, while intuitively quite plausible, is somewhat tricky to prove; see (). Since any nonempty open set in \( \mathbb{R}^n \) contains an open rectangle, no nonempty open set can have measure zero, i.e. every set of measure zero in \( \mathbb{R}^n \) has empty interior.

XI.23.1.8. **Proposition.** Let \( U \) be a subset of \( \mathbb{R}^n \), \( f : U \rightarrow \mathbb{R}^n \) a Lipschitz function (). If \( A \subseteq U \) has measure zero in \( \mathbb{R}^n \), so does \( f(A) \). In particular, isometries and affine functions preserve sets of measure zero.

**Proof:** Suppose \( A \subseteq U \) has measure zero in \( \mathbb{R}^n \). Let \( M \) be the Lipschitz constant of \( f \). Let \( \epsilon > 0 \), and let \( (B_k) \) be a sequence of open balls with centers \( x_k \in A \) covering \( A \) with total volume less than \( \frac{\epsilon}{M} \). If \( \rho_k \) is the radius of \( B_k \), then \( f(B_k \cap U) \) is contained in the open ball of radius \( M \rho_k \) around \( f(x_k) \); thus \( f(A) \) is contained in the union of these open balls, which have total volume \( < \epsilon \).

XI.23.1.9. **Corollary.** Let \( U \) and \( V \) be open subsets of \( \mathbb{R}^n \), and \( f \) a \((C^1)\) diffeomorphism from \( U \) to \( V \). If \( A \subseteq U \), then \( A \) has measure zero in \( \mathbb{R}^n \) if and only if \( f(A) \) does.

**Proof:** Every point of \( U \) has a compact neighborhood \( V \) contained in \( U \), on which \( f \) is Lipschitz (); thus if \( A \cap V \) has measure zero in \( \mathbb{R}^n \), so does \( f(A \cap V) \).
XI.23.1.10. Sets of measure zero are not preserved under homeomorphisms in general (). Thus having measure zero is a metric concept but not a topological concept.

The next result is a special case of Tonelli’s Theorem (), needed for the proof of Sard’s Theorem. This proof appears in [Ste83, p. 51] and is attributed to H. FURSTENBERG. An \( F_{\sigma} \) is a countable union of closed sets. Since every closed set in \( \mathbb{R}^n \) is a countable union of compact sets, a subset of \( \mathbb{R}^n \) is an \( F_{\sigma} \) if and only if it is \( \sigma \)-compact. Of course, every closed set in \( \mathbb{R}^n \) is an \( F_{\sigma} \). See ().

XI.23.1.11. Theorem. [Baby Tonelli Theorem] Let \( K \) be an \( F_{\sigma} \) in \( \mathbb{R}^n \). Suppose, for each \( t \in \mathbb{R} \), we have that
\[
K_t := \{(x_1, \ldots, x_{n-1}) : (x_1, \ldots, x_{n-1}, t) \in K\} \subseteq \mathbb{R}^{n-1}
\]
has measure zero in \( \mathbb{R}^{n-1} \). Then \( K \) has measure zero in \( \mathbb{R}^n \).

Informally, if each horizontal cross-section of \( K \) has \((n-1)\)-dimensional measure zero, then \( K \) has \( n \)-dimensional measure zero.

We need two lemmas for the proof. The first is a variation of () (note that this result is not quite correctly stated in [Ste83]):

XI.23.1.12. Lemma. Let \([a, b]\) be a closed bounded interval in \( \mathbb{R} \) \( c > 0 \), and \( \mathcal{U} \) a cover of \([a, b]\) by open intervals of length \( \leq c \). Then there is a finite subcover \( \{I_1, \ldots, I_n\} \) of \( \mathcal{U} \) for which
\[
\sum_{k=1}^n \ell(I_k) \leq 2(b - a + c).
\]

Proof: Since \([a, b]\) is compact there is a finite subcover. Any finite subcover contains a minimal subcover \( \mathcal{V} \) (i.e. \( \mathcal{V} \) is a cover but no proper subset of \( \mathcal{V} \) is a cover). Number the intervals in \( \mathcal{V} \) in increasing order of left endpoint (no two can have the same left endpoint by minimality), i.e. \( \mathcal{V} = \{I_1, \ldots, I_n\} \) where \( I_k = (a_k, b_k) \) with \( a_k < a_{k+1} \) for \( 1 \leq k < n \). These endpoints necessarily satisfy
\[
a_1 < a \leq a_2 < b_1 \leq a_3 < b_2 \leq \cdots \leq a_n < b_{n-1} \leq b < b_n
\]
(if one of these inequalities is violated, the collection of intervals is not minimal). We then have
\[
\sum_{k=1}^n \ell(I_k) = \sum_{k=1}^n (b_k - a_k) = (a - a_1) + (b - a) + (b_n - b) + \sum_{k=1}^{n-1} (b_k - a_{k+1}).
\]
The last sum is the total length of the overlaps of the intervals; since the overlaps are disjoint and contained in \([a, b]\), this total length is \( \leq b - a \). Also, \( a - a_1 \leq c \) and \( b_n - b \leq c \). Thus we have
\[
\sum_{k=1}^n \ell(I_k) \leq c + (b - a) + c + (b - a) = 2(b - a + c).
\]

\(\triangleright\)

1328
**XI.23.1.13. Lemma.** Let $K$ be a compact subset of $\mathbb{R}^n \cong \mathbb{R}^{n-1} \times \mathbb{R}$, and $t \in \mathbb{R}$. Set

$$K_t := \{(x_1, \ldots, x_{n-1}) : (x_1, \ldots, x_{n-1}, t) \in K\} \subseteq \mathbb{R}^{n-1}.$$ 

If $U$ is an open set in $\mathbb{R}^{n-1}$ containing $K_t$, then there is a $\delta > 0$ such that

$$K \cap \left[\mathbb{R}^{n-1} \times (t - \delta, t + \delta)\right] \subseteq U \times (t - \delta, t + \delta).$$ 

**Proof:** Suppose there is no such $\delta$. Then there is a sequence $(x_k, t_k)$ of points of $K$ with $t_k \to t$ and $x_k \not\in U$ for all $k$. Passing to a subsequence, we may assume $(x_k, t_k) \to (x, t) \in K$ for some $x$ by compactness of $K$. But then $x \in K_t$, and $x_k \to x$ in $\mathbb{R}^{n-1}$, contradicting that $U$ is an open neighborhood of $x$. \(\Box\)

**XI.23.1.14.** We now prove Theorem XI.23.1.11. It suffices to assume that $K$ is compact. Then there is an interval $[a, b]$ such that $K \subseteq \mathbb{R}^{n-1} \times [a, b]$. Fix $\epsilon > 0$. For each $t \in [a, b]$ there is a countable (even finite) collection $\mathcal{R}_t$ of open rectangular solids in $\mathbb{R}^{n-1}$ covering $K_t$ whose total $(n-1)$-dimensional volume is

$$v_t \leq \frac{\epsilon}{2(b - a + 1)}.$$ 

Let $U_t$ be the union of the rectangular solids in $\mathcal{R}_t$. By XI.23.1.13, there is a $\delta_t > 0$ such that

$$K \cap \left[\mathbb{R}^{n-1} \times (t - \delta_t, t + \delta_t)\right] \subseteq U_t \times (t - \delta_t, t + \delta_t).$$ 

We may assume $\delta_t \leq 1/2$ for all $t$. The intervals $\{(t - \delta_t, t + \delta_t) : t \in [a, b]\}$ form an open cover of $[a, b]$, so by Lemma XI.23.1.12, there is a subcover

$$\{(t_1 - \delta_{t_1}, t_1 + \delta_{t_1}), \ldots, (t_m - \delta_{t_m}, t_m + \delta_{t_m})\}$$

whose total length $\sum_{k=1}^m 2\delta_{t_k}$ is $\leq 2(b - a + 1)$. If $\mathcal{R}$ is the collection of open rectangular solids of the form

$$R \times (t_k - \delta_{t_k}, t_k + \delta_{t_k})$$

for $R \in \mathcal{R}_{t_k}, 1 \leq k \leq m$, then the collection $\mathcal{R}$ covers $K$ and the total $n$-dimensional volume is

$$\sum_{k=1}^m 2\delta_{t_k} v_{t_k} \leq \frac{\epsilon}{2(b - a + 1)} \sum_{k=1}^m 2\delta_{t_k} \leq \epsilon.$$ 

Since $\epsilon > 0$ is arbitrary, $K$ has measure zero in $\mathbb{R}^n$. This completes the proof of Theorem XI.23.1.11. \(\Box\)

**XI.23.1.15.** This theorem extends (XIV.6.1.12) from $F_\alpha$'s to a much larger class of subsets of $\mathbb{R}^n$ (Lebesgue measurable subsets), but not to arbitrary subsets. See XIV.6.3.1..

**Sets of Measure Zero in a Manifold**

We can extend the notion of subset of measure zero to smooth manifolds:
XI.23.1.16.  **Definition.** Let $M$ be a $C^r$ $n$-manifold ($r \geq 1$). A subset $A$ of $M$ has measure zero in $M$ if, for every $x \in A$, there is a coordinate chart $(U, \phi)$ for $M$ with $x \in U$ such that $\phi(A \cap U)$ has measure zero in $\mathbb{R}^n$.

XI.23.1.17.  It follows easily from XI.23.1.9. that if $A \subseteq M$ has measure zero in $M$, and $(U, \phi)$ is any coordinate chart for $M$, then $\phi(A \cap U)$ has measure zero in $\mathbb{R}^n$. This could have been taken as the definition of a subset of measure zero, but the condition of XI.23.1.16. is easier to check.

If $M$ is not second countable, what we have technically defined is a locally null subset of $M$. (This technicality may be safely ignored as irrelevant in our applications.) There is no analogous reasonable definition of a subset of measure zero in a topological manifold ($\ast$).

The next result is a straightforward corollary of XI.23.1.8.:

XI.23.1.18.  **Proposition.** Let $N$ and $M$ be $C^r$ $n$-manifolds ($r \geq 1$), with $N$ separable, and $f : N \to M$ a $C^r$ function. If $A$ is a subset of $N$ which has measure zero in $N$, then $f(A)$ has measure zero in $M$.

XI.23.1.19.  This result holds also if $N$ is $n$-dimensional and $M$ is $m$-dimensional with $n < m$ ($\ast$), but not necessarily if $m < n$.

**XI.23.2. Statement of Sard’s Theorem**

XI.23.2.1.  **Theorem.** [Sard’s Theorem] Let $N$ be a separable $C^r$ $n$-manifold, and $M$ a $C^r$ $m$-manifold, with $r \geq \max(n - m + 1, 1)$, and let $f : N \to M$ be a $C^r$ function. Then the set of singular values of $f$ is a subset of $M$ of measure zero in $M$.

XI.23.2.2.  We will prove special cases of the theorem separately before turning to the most general version. The special cases are written separately because the proof simplifies in these cases and sometimes a more general result holds. The cases we discuss are:

(i) $n < m$.

(ii) $n = m = 1$.

(iii) $n = m > 1$.

(iv) $n > m$, $r$ large.

(v) $n > m$, sharp $r$ bound.

Our proof is adapted from [Ste83] and [Hir94].

XI.23.2.3.  The first reduction in all cases is to note that $N$ can be covered by countably many coordinate charts, each of which is contained in the preimage of a coordinate chart in $M$, and since the union of countably many null sets is a null set and null sets are preserved under diffeomorphisms, it suffices to prove the theorem for $N = \mathbb{R}^n$ and $M = \mathbb{R}^m$. In fact, since $N$ is $\sigma$-compact, it suffices to prove the theorem for the set of singular values in the image of an arbitrary compact subset of $\mathbb{R}^n$. 1330
XI.23.2.4. The restriction that \( N \) be separable, i.e. that it has only countably many components, is necessary in general: if \( N \) has \( 2^{\aleph_0} \) components, a function from \( N \) to \( M \) can be surjective even if it is constant on each component of \( N \); such a function has identically zero derivative, and hence every point of \( M \) is a singular value.

XI.23.3. The Case \( n < m \)

XI.23.4. The Case \( n = m = 1 \)

XI.23.4.1. The simplest case is \( n = m = 1 \). Suppose \( f \) is a \( C^1 \) function from \( \mathbb{R} \) (or an open interval in \( \mathbb{R} \)) to \( \mathbb{R} \). The singular points of \( f \) are just the critical numbers (the \( x_0 \) for which \( f'(x_0) = 0 \)). It suffices to show that the image of the critical numbers in any closed bounded subinterval of the domain has measure zero in \( \mathbb{R} \).

So suppose \([a,b]\) is such an interval. Since \( f' \) is continuous on \([a,b]\), it is uniformly continuous there \((\ast)\). Fix \( \epsilon > 0 \). Then there is a \( \delta > 0 \) such that \(|f'(x) - f'(y)| < \epsilon \) whenever \( x, y \in [a,b], |x - y| < \delta \). Partition \([a,b]\) into subintervals \( I_k = [x_{k-1}, x_k] \) with \( \ell(I_k) = x_k - x_{k-1} < \delta \) for all \( k \). For every \( I_k \) containing a singular point of \( f \), we have \(|f'(x)| < \epsilon \) for all \( x \in I_k \). Thus by the Mean Value Theorem we have \(|f(x) - f(y)| < \epsilon \ell(I_k)\) for all \( x, y \in I_k \), for any such \( k \). Thus \( f(I_k) \) is contained in an interval \( J_k \) of length \( < \epsilon \ell(I_k) \) for any such \( k \) (actually \( f(I_k) \) is itself such an interval). Thus the set of singular values of \( f \) in \([a,b]\) is contained in a finite number of intervals of total length

\[
\sum_{k=1}^{n} \ell(J_k) < \epsilon \sum_{k=1}^{n} \ell(I_k) = \epsilon (b - a)
\]

and since \( \epsilon > 0 \) is arbitrary, this set of singular values has measure zero (in fact zero content) in \( \mathbb{R} \).

This completes the proof in the case \( n = m = 1 \). \( \diamond \)

The \( C^1 \) assumption on \( f \) can be weakened and even eliminated in this case. The following result is due to LUZIN (XIV.17.1.14.); we give an alternate proof due to S. IVANOV (http://mathoverflow.net/questions/113991/counterexample-to-sard-theorem-for-a-not-c1-map/114000#114000):

XI.23.4.2. Theorem. Let \( f \) be a function from \( \mathbb{R} \) (or an interval in \( \mathbb{R} \)) to \( \mathbb{R} \). Set

\[
\mathcal{C} = \{ x \in \mathbb{R} : f'(x) \text{ exists and equals 0} \}
\]

and let \( S = f(\mathcal{C}) \). Then \( S \) has measure zero in \( \mathbb{R} \).

Proof: We will use the Vitali Covering Theorem \((\ast)\). It suffices to show that the part of \( \mathcal{C} \) in \((a,b)\) has image of zero measure, for any bounded open interval \((a,b)\) in the domain of \( f \), so we may assume the domain of \( f \) is \((a,b)\) and hence that \( \mathcal{C} \) has finite outer measure. Fix \( \epsilon > 0 \). For every \( x_0 \in \mathcal{C} \), there is a \( \delta_0 \) such that \((x_0 - \delta_0, x_0 + \delta_0) \subseteq (a,b)\) and

\[
\left| \frac{f(x) - f(x_0)}{x - x_0} \right| < \frac{\epsilon}{5}
\]

for all \( x \) with \( |x - x_0| < \delta_0 \). Set

\[
\mathcal{I}_{x_0} = \left\{ \left( x - \delta, x + \frac{\delta}{5} \right) : 0 < \delta \leq \delta_0 \right\}
\]

1331
and let $I = \bigcup_{x_0 \in \mathbb{C}} I_{x_0}$. Then $\mathcal{I}$ is a Vitali covering of $\mathbb{C}$.

For each $I = (x_0 - \frac{\delta}{5}, x_0 + \frac{\delta}{5}) \in \mathcal{I}$, let $I'$ be the interval $(x_0 - \delta, x_0 + \delta)$ with the same midpoint as $I$ of five times the length. By (i) there is a sequence $(I_n)$ of disjoint sets in $\mathcal{I}$ such that the $I_n'$ cover $\mathbb{C}$. If $x \in I_n'$, and $x_n$ is the midpoint of $I_n$, then

$$|f(x) - f(x_n)| < \frac{\epsilon}{5} |x - x_n| < \frac{\epsilon}{5} \ell(I_n') = \frac{\epsilon}{5} \ell(I_n)$$

and thus, if $x, y \in I_n'$ we have

$$|f(x) - f(y)| \leq |f(x) - f(x_n)| + |f(x_n) - f(y)| < \epsilon \ell(I_n)$$

and so $f(I_n')$ is contained in an interval $J_n$ of length $\epsilon \ell(I_n)$. The $J_n$ cover $\mathcal{S}$.

Since the $I_n$ are disjoint, we have $\sum_n \ell(I_n) \leq b - a$, so

$$\sum_{n=1}^{\infty} \ell(J_n) = \epsilon \sum_{n=1}^{\infty} \ell(I_n) \leq \epsilon (b - a)$$

and since $\epsilon > 0$ is arbitrary, we have that $\mathcal{S}$ has measure zero in $\mathbb{R}$. 

XI.23.5. The Case $n = m > 1$

XI.23.6. The Case $n > m$, $r$ Large

We now show that the theorem holds for general $n$ and $m$ if $r$ is sufficiently large. Most importantly, this result applies for all $n$ and $m$ if $f$ is $C^\infty$. The estimate on the minimum $r$ from this argument is not sharp in general (it is quadratic in $n - m$, not linear). This proof is adapted from [Hir94].

XI.23.6.1. For $n, m \in \mathbb{N}$, define $s(n, m) \in \mathbb{N}$ as follows:

$$s(n, m) = \begin{cases} 1 & \text{if } n \leq m \text{ and } m + 1 \text{ divides } (n - m + 1)(n - m + 2) \\ \frac{1}{2} & \text{if } n > m \end{cases}.$$ 

The following properties needed for the proof of the theorem are easily verified:

XI.23.6.2. Proposition. The numbers $s(n, m)$ have the following properties:

(i) $s(n, m) > \frac{n}{m}$ if $n > m$.

(ii) $s(n, m) - k \geq s(n - 1, m)$ if $1 \leq k \leq \frac{n}{m}$.

(iii) $s(n, m) = s(n - 1, m - 1)$ if $m > 1$.

XI.23.6.3. Theorem. Let $n, m \in \mathbb{N}$. Then if $r \geq s(n, m)$, $N$ is a separable $C^r$ manifold, $M$ a $C^r$ $m$-manifold, and $f : N \to M$ is a $C^r$ function, then the set of singular values of $f$ is a subset of $M$ of measure zero in $M$. 

1332
XI.23.6.4. The proof will be by double induction: induction on \( m \), and for fixed \( m \) induction on \( n \). The cases \( n \leq m \) have already been proved (), so fix \( n \) and \( m \) and assume that \( n > m \) and that the theorem is proved for \((n - 1, m)\), and for \((n - 1, m - 1)\) if \( m > 1 \). If \( \Sigma \) is the set of singular points of \( N \), it suffices to show that every \( x_0 \in \Sigma \) has a neighborhood \( U \) such that \( f(\Sigma \cap U) \) has measure 0 in \( M \). Write

\[
\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3
\]

where

\( \Sigma_1 \) is the set of \( x_0 \in \Sigma \) for which all partial derivatives of \( f \) of orders \( \leq \frac{n}{m} \) vanish at \( x_0 \).

\( \Sigma_2 \) is the set of \( x_0 \in \Sigma \) such that the first-order partials of \( f \) vanish at \( x_0 \) but some partial derivative of \( f \) of order \( k \), \( 2 \leq k \leq \frac{n}{m} \), is nonzero at \( x_0 \).

\( \Sigma_3 \) is the set of \( x_0 \in \Sigma \) such that some first-order partial of \( f \) is nonzero at \( x_0 \).

XI.23.6.5. First suppose \( x \in \Sigma_1 \).

XI.23.6.6. Now suppose \( x_0 \in \Sigma_2 \), \( y_0 = f(x_0) \). For each multiindex \( \alpha \) of order \( k \), \( 2 \leq k < \frac{n}{m} \), set

\[
g_{\alpha,i} = \frac{\partial^n f_i}{\partial x_\alpha}
\]

where \( f_i \) is the \( i \)th coordinate function of \( f \). Then \( g_{\alpha,i} \) is \( C^q \), where

\[
q = r - k \geq s(n, m) - \frac{n}{m} \geq s(n - 1, m)
\]

For each multiindex \( \alpha \) of order \( k \), \( 2 \leq k < \frac{n}{m} \), and each \( i \) and \( j \), let \( X_{\alpha,i,j} \) be the set of points \( x \in \Sigma_2 \) such that \( g_{\alpha,i}(x) = 0 \) and \( \frac{\partial g_{\alpha,i}}{\partial x_j}(x) \neq 0 \). Then \( x_0 \in X_{\alpha,i,j} \) for some \( \alpha, i, j \).

Fix \( \alpha, i, j \) with \( x_0 \in X_{\alpha,i,j} \). Then, since \( g_{\alpha,i}(x_0) = 0 \) and \( \frac{\partial g_{\alpha,i}}{\partial x_j}(x_0) \neq 0 \), \( x_0 \) is a regular point of \( g_{\alpha,i} \). Thus there is a neighborhood \( U \) of \( x_0 \) such that 0 is a regular value of \( h_{\alpha,i} = g_{\alpha,i}|U \) and therefore \( Z_{\alpha,i} = h_{\alpha,i}^{-1}(\{0\}) \) is a \( C^q \)-submanifold of \( N \) of dimension \( n - 1 \) containing \( x_0 \). Since \( q \geq s(n - 1, m) \) and \( f|Z_{\alpha,i} \) is \( C^q \), the inductive hypothesis applies. Any point of \( \Sigma_2 \) which is in \( Z_{\alpha,i} \) is a singular point of \( f|Z_{\alpha,i} \); since its differential is zero, so we conclude that the measure of \( f(\Sigma_2 \cap Z_{\alpha,i}) \) is zero. Since \( X_{\alpha,i,j} \cap U \subseteq Z_{\alpha,i} \), we have that the measure of \( X_{\alpha,i,j} \cap U \) is zero. There is a neighborhood of \( x_0 \) in \( \Sigma_2 \) contained in a finite union of such sets for various \( \alpha, i, j \), so we conclude that the measure of \( f(\Sigma_2) \) is zero.

XI.23.6.7. If \( m = 1, \Sigma_3 = \emptyset \) so the proof is finished. If \( m > 1 \), suppose \( x_0 \in \Sigma_3 \), \( y_0 = f(x_0) \). By (), there are neighborhoods \( U \) of \( x_0 \) and \( V \) of \( y_0 \), and diffeomorphisms \( g : \mathbb{R}^{n-1} \times \mathbb{R} \to U \) and \( h : V \to \mathbb{R}^{m-1} \times \mathbb{R} \) with \( g(0, 0) = x_0 \), \( h(y_0) = (0, 0) \), such that \( \phi = h \circ f \circ g \) is of the form

\[
\phi(z, t) = (\psi(z, t), t)
\]

for some \( C^r \)-function \( \psi : \mathbb{R}^{n-1} \times \mathbb{R} \to \mathbb{R}^{m-1} \). The singular values of \( f \) in \( V \) are precisely the images under \( h^{-1} \) of the singular values of \( \phi \), so it suffices to show that the set \( S \) of singular values of \( \phi \) has measure 0. Fix \( t \), and let \( \omega_t : \mathbb{R}^{n-1} \to \mathbb{R}^{m-1} \) be defined by \( \omega_t(z) = \psi(z, t) \). Then \( \omega \) is \( C^r \), so by the inductive hypothesis
the set \( S_t \) of singular values of \( \omega_t \) has measure 0 in \( \mathbb{R}^{m-1} \). We have that \( z \) is a singular point of \( \omega_t \) if and only if \((z,t)\) is a singular point of \( \phi \), so

\[
S = \{(z,t) : t \in \mathbb{R}, z \in S_t \}.
\]

Thus by the Baby Tonelli Theorem (XI.23.1.11) \( S \) has measure zero in \( \mathbb{R}^m \times \mathbb{R} \).

This completes the proof of Theorem XI.23.6.3.

XI.23.7. The Case \( n > m \), Sharp \( r \) Bound

We now turn to the final and most delicate part of the theorem, establishing the sharp bound on \( r \). It is questionable whether it is worth proving this version, since the softer result XI.23.6.3. suffices for almost all applications. But the conceptual reason behind the sharp bound on \( r \) is important and instructive.

The reason for the sharp bound \( r \geq n - m + 1 \) is a higher-dimensional version of the following one-dimensional fact (cf. ()): XI.23.7.1. Proposition. Let \( f : \mathbb{R} \to \mathbb{R} \) be a \( C^r \) function (\( r \geq 0 \)). If \( f \) has a nonisolated zero at \( x_0 \), then \( f(x_0) = f'(x_0) = \cdots = f^{(r)}(x_0) = 0 \). It follows that \( f \) vanishes at \( x_0 \) at least to order \( r^+ \), i.e.

\[
\lim_{x \to x_0} \frac{f(x)}{(x - x_0)^r} = 0.
\]

Proof: If \( r = 0 \) there is nothing to prove. If \( r \geq 1 \), let \((x_n)\) be a sequence of zeroes of \( f \) converging to \( x_0 \). By Rolle’s Theorem, for each \( n \) there is a \( y_n \) between \( x_0 \) and \( x_n \) with \( f'(y_n) = 0 \). We have \( y_n \to x_0 \) by the Squeeze Theorem, so \( f'(x_0) = 0 \) since \( f' \) is continuous. If \( r > 1 \), repeat the argument using \( f' \) in place of \( f \) and the \( y_n \) in place of the \( x_n \) to conclude that \( f''(x_0) = 0 \), and continue inductively. The last statement follows from V.10.1.3.

XI.23.7.2. This result is sharp. If \( f(x) = \frac{|x|}{(\log |x|)^r} \sin(\log |x|) \) for \( x \neq 0 \) and \( f(0) = 0 \), then \( f \) is \( C^1 \) and has a nonisolated zero at 0, but vanishes only to order \( 1^+ \) at 0. There are similar counterexamples for larger \( r \).

XI.23.8. The Category Version

The following result is a simple corollary of the measure-theoretic version of Sard’s Theorem (cf. [Bro35]).

XI.23.8.1. Theorem. Let \( N \) be a separable \( C^r \) \( n \)-manifold, \( M \) a \( C^r \) \( m \)-manifold, and \( f : N \to M \) a \( C^r \) function, where \( r \geq \max(n - m + 1, 1) \). Then the set of singular values of \( f \) is a meager set in \( M \).

Proof: First note that the set of singular points of \( f \) is closed in \( N \): it is the set of \( x_0 \in N \) such that \( \text{Det}(Df(x_0)Df(x_0)^t) = 0 \). Thus, although the set of singular values of \( f \) need not be closed in \( M \), it is an \( F_\alpha \) (XI.11.7.11.), and has measure zero by XI.23.2.1.; thus it has empty interior (\( \) ), i.e. its complement in \( M \) is a dense \( G_\delta \).
**XI.23.8.2.** Corollary. Let $N$ be a separable $n$-manifold, $M$ an $m$-manifold, and $f : N \to M$ a $C^1$ function. If $n < m$, then the range of $f$ is a meager subset of $M$.

**XI.23.8.3.** These results apply to the case where $N$ and $M$ are Euclidean spaces. They imply that manifolds of different dimensions cannot be diffeomorphic; in particular, $\mathbb{R}^n$ and $\mathbb{R}^m$ are not diffeomorphic if $n \neq m$ (actually more is true: they cannot even be homeomorphic, cf. (i)).

**XI.23.9.** Counterexamples

The differentiability restrictions in Sard’s theorem are sharp (as long as only $C^r$ functions are considered for $r \in \mathbb{N}$; some relaxation is possible if more delicate degrees of differentiability are allowed, cf. [Bat93]). Even before any of the publications of A. P. Morse or Sard, H. Whitney [Whi35] described counterexamples to stronger statements.

**XI.23.9.1.** Whitney described a (highly nondifferentiable) parametrized curve $\gamma : [0, 1] \to \mathbb{R}^2$, with $\gamma$ injective, such that

$$\lim_{t \to t_0} \frac{|t - t_0|}{\|\gamma(t) - \gamma(t_0)\|} = 0$$

for every $t_0 \in [0, 1]$. Then the function $\gamma^{-1} : C = \gamma([0, 1]) \to [0, 1]$ has “zero derivative” at each point $a$ of $C$ in the sense that

$$\lim_{x \to a} \frac{\|\gamma^{-1}(x) - \gamma^{-1}(a)\|}{\|x - a\|} = 0$$

where the limit is taken within $C$. By Whitney’s Extension Theorem (i), $\gamma^{-1}$ extends to a $C^1$ function $f$ from $\mathbb{R}^2$ to $\mathbb{R}$; every point of $C$ is a critical point of $f$, so every number in $[0, 1]$ is a critical value.

As described in [Hir94], “This example leads to the following paradox: the graph of $f$ is a surface $S \subset \mathbb{R}^3$ on which there is an arc $A$, at every point of which the surface has a horizontal tangent plane, yet $A$ is not at a constant height. To make this more vivid, imagine that $S$ is a hill and $A$ a path on the hill. The Hill is level at every point of the path, yet the path goes up and down.” In fact, the path can be made to go monotonically up from bottom to top. We should point out that such a path would not be a very good route to follow to ascend the hill, since it has infinite arc length; in fact, its projection into the $xy$-plane has positive Lebesgue measure (the curve has Hausdorff dimension 2 and positive 2-dimensional Hausdorff measure). There cannot be a smooth path with this property because of the Chain Rule.

Whitney briefly described a higher-dimensional version, giving a $C^{n-1}$ function $f_n$ ($n \geq 2$) from $\mathbb{R}^n$ to $\mathbb{R}$ for which the set of critical values contains an interval. If $f_n$ is such a function, then for any $m$ the function $g_{n,m} : \mathbb{R}^{n+m-1} \to \mathbb{R}^m$ given by

$$g_{n,m}(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m-1}) = (f_n(x_1, \ldots, x_n), x_{n+1}, \ldots, x_{n+m-1})$$

is a $C^{n-1}$ function whose singular values contain an entire rectangle in $\mathbb{R}^m$.

**XI.23.9.2.** Here is a simpler example from [Gri85] for the case $n = 2$, $m = 1$.

We need a $C^1$ function $f$ from $\mathbb{R}$ to $\mathbb{R}$ for which every point of the Cantor set $K$ is a singular value. Here is one construction. Fix $\gamma$, $2 < \gamma < 3$, and let $\alpha = (\alpha_1, \alpha_2, \ldots)$ with $\alpha_n = \gamma^{-n}$. Construct the Cantor set $K_\alpha$ as in XI.17.2.1., i.e. subintervals of length $\gamma^{-n}$ are removed at each step. Let $(I_k)$ be a list of the removed
intervals. Let $\phi$ be a mesa function () on $[0,1]$ and $\beta = \int_0^1 \phi(t) \, dt$; then $0 < \beta < 1$. If the length of $I_k$ is $\gamma^{-n}$, let $h_k$ be a translated and scaled version of $\phi$ on $I_k$ whose integral is $3^{-n}$, i.e. scaled horizontally by $\gamma^{-n}$ and vertically by $\frac{1}{\beta} \left( \frac{2}{3} \right)^n$. Since

$$\lim_{n \to \infty} \left( \frac{2}{3} \right)^n = 0$$

we have that $\sum_{k=1}^{\infty} h_k$ converges uniformly to a continuous function $h$. Then $h(t) \geq 0$ for all $t \in \mathbb{R}$ and $h(t) = 0$ for all $t \in K_\alpha$, and for $t < 0$, $t > 1$.

Set $g(x) = \int_{-\infty}^x h(t) \, dt$. Then $g$ is a nondecreasing $C^1$ function on $\mathbb{R}$, $g(x) = 0$ for $x \leq 0$ and $g(x) = 1$ for $x \geq 1$, and $g(x) \in K$ for $x \in K_\alpha$. For each such $x$, $g'(x) = h(x) = 0$. Thus every point of $K$ is a critical value of $g$ (the converse holds too).

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = g(x) + g(y)$. Then $f$ is $C^1$, and if $x$ and $y$ are in $K_\alpha$, then $Df(x,y) = 0$. If $z \in [0,2]$, then there are $c,d \in K$ such that $z = c + d$ (Exercise XI.17.3.1). There are $x,y \in K_\alpha$ with $g(x) = c$ and $g(y) = d$; so $f(x,y) = z$ and $z$ is a critical value, hence singular value, of $f$. Thus the set of singular values of $f$ does not have measure zero in $\mathbb{R}$. (Actually $[0,2]$ is the entire range of $f$, and consists of singular values.)

Note that $h$ is not differentiable everywhere, hence $g$, and therefore $f$, are not $C^2$; of course a $C^2$ function with this property cannot exist by Sard’s Theorem.

### XI.23.10. Applications

#### XI.23.10.1. Theorem. Let $X$ be a compact subset of $\mathbb{R}^n$. Then for any $\epsilon > 0$ there is an open neighborhood $U$ of $X$ in $\mathbb{R}^n$ with $\partial U$ a smooth submanifold of $\mathbb{R}^n$ (i.e. $\partial U$ is a compact smooth submanifold with boundary), such that $\rho(y,X) < \epsilon$ for all $y \in U$, $U$ contains all $y \in \mathbb{R}^n$ with $\rho(y,X) < \frac{\epsilon}{4}$, and, for every $y \in \partial U$, $\frac{\epsilon}{4} \leq \rho(y,X) < \epsilon$.

**Proof:** Let $V = \{ y \in \mathbb{R}^n : \rho(y,X) < \epsilon \}$. Then $V$ is a neighborhood of $X$ in $\mathbb{R}^n$ with compact closure. Define $g : \mathbb{R}^n \to \mathbb{R}$ by $g(y) = \rho(y,X)$; then $g$ is continuous. Let $f$ be a $C^\infty$ function from $\mathbb{R}^n$ to $\mathbb{R}$ (e.g. a polynomial in the coordinates) uniformly approximating $g$ within $\frac{\epsilon}{4}$ on $\bar{V}$; there is such an $f$ by the Stone-Weierstrass Theorem (). There is a regular value $\alpha$ for $f$ with $\frac{\epsilon}{2} < \alpha < \frac{4\epsilon}{3}$ by Sard’s Theorem. Set

$$U = \{ y \in V : f(y) < \alpha \} .$$

Then $U$ is open. Since for $y \in \bar{V}$ we have $f(y) - \frac{\epsilon}{4} < \rho(y,X) < f(y) + \frac{\epsilon}{4}$, we have that $\bar{U} \subseteq V$, so

$$\partial U = \{ y \in V : f(y) = \alpha \}$$

is a smooth closed submanifold of $\mathbb{R}^n$, and the desired inequalities follow.

#### XI.23.10.2. In fact the $\frac{\epsilon}{4}$ can be replaced by any $\eta < \epsilon$ by a suitable modification of the proof.

#### XI.23.10.3. In the case $n = 2$, if $X$ is a compact subset of the plane there is for any $\epsilon > 0$ an $\epsilon$-neighborhood $U$ of $X$ in $\mathbb{R}^2$ whose boundary consists of finitely many disjoint smooth closed curves. (Such a neighborhood is not necessarily homeomorphic to a disjoint union of closed disks, however, e.g. if $X$ is a circle.) See XI.23.11.3. for a constructive proof not using Sard’s Theorem.
XI.23.11. Exercises

XI.23.11.1. Carefully justify all the assertions in XI.23.1.3., i.e. show that adding each of the various restrictions or modifications to the definition does not change the class of null sets.

XI.23.11.2. Show that if $C$ is any Cantor set in $\mathbb{R}$ of Lebesgue measure 0, there is a $C^1$ function $f : \mathbb{R} \to \mathbb{R}$ whose set of critical values is $C$. [Write $C = C_\lambda$ as in XI.17.2.9., and mimic the argument of XI.23.9.2., using IV.2.16.5.]

XI.23.11.3. Give a constructive proof of the case $n = 2$ of XI.23.10.1., as follows.
(a) Show that the diameter of a regular hexagon of side $s$ is $2s$. [Write the hexagon as a union of six equilateral triangles.]
(b) If $X$ is a compact subset of $\mathbb{R}^2$ and $V$ is an open neighborhood of $X$, there is a bounded open neighborhood $W$ of $X$ contained in $V$. Let $\epsilon$ be the distance between $X$ and $\partial W$. Show that $\epsilon > 0$.
(c) Cover the plane with a hexagonal honeycomb with sides of length $s < \frac{\epsilon}{4}$, and let $A$ be the union of all the closed hexagons which are contained in $W$. Show that $X$ is contained in the interior $U$ of $A$.
(d) Show that $\partial A = \partial U$ consists of finitely many disjoint simple closed polygonal curves.
(e) If desired, the corners of these boundary curves can be rounded off to give smooth boundary curves.
(f) Try to do the same argument using squares (try staggering the squares so no four edges meet at a point), and try generalizing to $\mathbb{R}^n$. 

1337
XI.24. Piecewise-Linear Topology

The idea behind piecewise-linear (PL) topology is to reduce the study of the topology of “reasonable” subsets of Euclidean space, and eventually all “reasonable” (e.g. compact metrizable) topological spaces, to the affine structure of $\mathbb{R}^n$. This is clearly possible in principle since the affine structure on $\mathbb{R}^n$ determines its topological structure, and it turns out in practice to be a slick and clever way to express many topological notions in essentially combinatorial language.

PL topology has somewhat passed out of fashion. However, the basic setup (which is all we will do here, what topologists would consider to be “preliminaries”) consisting of polyhedra and simplicial complexes and their maps, is material every mathematician should be familiar with.

The standard reference for PL topology is [RS72], on which much of our exposition is based.

XI.24.1. Polyhedra, Cells, and Simplexes

The term “polyhedron” traditionally refers to a three-dimensional figure made up of a finite number of vertices, edges, and faces; it is the three-dimensional analog of a polygon. We will use the term in a somewhat different sense, applying to objects of arbitrary dimension; thus (most) polyhedra in the traditional sense will be polyhedra in our sense, as will be polygons in the plane, closed intervals, points, and analogous higher-dimensional figures.

There is an ambiguity in the usual use of terms like “polyhedron,” “polygon,” “triangle,” “square,” etc.: do we include the “interior” or just the “surface”? Our definition of polyhedron will include both cases, although when a specific polyhedron is considered we must carefully specify which we mean. For example, a solid triangle (homeomorphic to a disk) and a triangle not including the interior (homeomorphic to a circle) will both be polyhedra; the first is a 2-dimensional polyhedron and the second a 1-dimensional polyhedron.

Cells

XI.24.1.1. Definition. A cell in $\mathbb{R}^n$ is a compact convex set with only finitely many extreme points. An extreme point of a cell is called a vertex of the cell. (The plural of vertex is vertices.) The dimension of a cell is the dimension of the affine subspace it generates.

XI.24.1.2. Since a compact convex subset of $\mathbb{R}^n$ is the convex hull of its extreme points ($\{\}$), a cell is precisely the convex hull of a finite set $S$ of points in $\mathbb{R}^n$. (Note, however, that not all the points of $S$ are necessarily vertices of the cell defined by $S$, although each vertex of the cell is in $S$.)

XI.24.1.3. Example. A rectangle (product of intervals) is a cell. A rectangle whose sides all have the same length is called a cube.

It is convenient in PL-topology to use the $\rho_{\infty}$ metric on $\mathbb{R}^n$. The closed balls in this metric are cubes, hence cells. A closed ball in the usual $\rho_2$ metric, or the $\rho_p$ metric for $1 < p < \infty$, is not a cell except in $\mathbb{R}^1$, since it has infinitely many extreme points.

XI.24.1.4. If a cell $C$ in $\mathbb{R}^m$ has $n + 1$ vertices $v_0, v_1, \ldots, v_n$, then the affine subspace it generates is parallel to the subspace spanned by $\{v_1 - v_0, \ldots, v_n - v_0\}$, hence the dimension of $C$ is the dimension of this subspace, which is $\leq n$. It can be strictly less than $n$ (e.g. for a convex polygon in $\mathbb{R}^2$).
XI.24.1.5. Recall () that a face of a convex set \( C \) is a subset \( F \) such that, whenever an interior point of a line segment in \( C \) is in \( F \), the endpoints of the segment are in \( F \). We have easily from () (or with an elementary proof):

XI.24.1.6. Proposition. Let \( C \) be a cell in \( \mathbb{R}^n \), and \( F \) a face of \( C \). Then \( F \) is also a cell, and the vertices of \( F \) are a subset of the vertices of \( C \). The dimension of \( F \) is strictly smaller than the dimension of \( C \) if \( F \) is a proper face of \( C \).

XI.24.1.7. Not every set of vertices of a cell \( C \) determines a face: for example, if \( C \) is a (solid) square in \( \mathbb{R}^2 \), the faces are just the vertices, the edges, and the whole square. Two opposite vertices do not determine a face.

XI.24.1.8. Proposition. Let \( C \) be a cell in \( \mathbb{R}^n \) and \( y \in C \). Let \( F_y \) be the smallest face of \( C \) containing \( y \). Then \( F_y \) is either \( \{y\} \) or the union \( F \) of all line segments in \( C \) containing \( y \) in the interior.

Proof: If \( F \) is empty, then \( y \) is an extreme point of \( C \) and \( F_y = \{y\} \). Suppose \( F \neq \emptyset \). By definition, any face containing \( y \) contains \( F \), so we need only show that \( F \) is a face of \( C \). To show \( F \) is convex, suppose \( a \) and \( b \) are in \( F \), and that \( c \) is a convex combination of \( a \) and \( b \). There are line segments \( \overline{ad} \) and \( \overline{be} \) in \( C \) containing \( y \) in their interior. Let \( Q \) be the quadrilateral in the plane generated by \( \overline{bd} \) and \( \overline{be} \) with vertices \( a, b, d, \) and \( e \); then \( Q \subseteq C \) by convexity. Let \( L \) be the line through \( c \) and \( y \). Then \( L \cap C \) contains \( L \cap Q \), which is a line segment with \( y \) in its interior. Thus \( c \in F \) and \( F \) is convex.

Now suppose \( a, b \in C \), and \( c \) is a (nontrivial) linear combination of \( a \) and \( b \) which is in \( F \). Then there is a line segment \( \overline{cd} \) in \( C \) containing \( y \) in its interior. Let \( T \) be the triangle (in the plane generated by \( \overline{ab} \) and \( \overline{cd} \) with vertices \( a, b, \) and \( d \); then \( T \subseteq C \) by convexity. If \( L \) is the line through \( a \) and \( y \), then \( L \cap C \) contains \( L \cap T \), which is a line segment containing \( y \) in its interior. Thus \( a \in F \). Similarly \( b \in F \). So \( F \) is a face of \( C \).

XI.24.1.9. If \( C \) is a cell in \( \mathbb{R}^n \), let \( \hat{C} \) be the set of all points of \( C \) not contained in any proper face of \( C \), and \( \hat{C} \) the union of the proper faces of \( C \). Then \( C \) and \( \hat{C} \) are called the interior and boundary of \( C \) respectively (they are the topological interior and topological boundary of \( C \) in the affine subspace it generates with the relative topology). The interior \( C \) is nonempty: if \( v_1, \ldots, v_m \) are the vertices of \( C \), then \( \hat{C} \) contains all points of the form \( \sum_{k=1}^m \lambda_k v_k \) where all \( \lambda_k > 0 \) (and \( \sum_{k=1}^m \lambda_k = 1 \)). In fact, \( C \) is precisely this set. \( \hat{C} \) is a relatively open subset of \( C \), and \( \hat{C} = C \setminus \hat{C} \) is compact.

XI.24.1.10. Proposition. Let \( C \) be a cell in \( \mathbb{R}^n \), and \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \) an affine function. Then \( f(C) \) is a cell in \( \mathbb{R}^m \).

Proof: If \( v_1, \ldots, v_m \) are the vertices of \( C \), then \( f(C) \) is exactly the convex hull of \( \{f(v_1), \ldots, f(v_m)\} \), hence a cell. (But \( f(v_k) \) is not necessarily a vertex of \( f(C) \) for every \( k \)).

1339
XI.24.1.11. PROPOSITION. Let $C$ be a convex subset of $\mathbb{R}^n$. Then $C$ is a cell if and only if $C$ is a finite intersection of (closed) half-spaces.

One direction of this is proved in XI.24.1.31., and the converse in Exercise XI.24.5.1.

Simplexes

XI.24.1.12. DEFINITION. An $n$-simplex is an $n$-dimensional cell with exactly $n + 1$ extreme points. A simplex is an $n$-simplex for some $n$.

The plural of simplex should properly be simplices, but (unlike with vertex) the anglicized plural simplexes has come into general use.

XI.24.1.13. PROPOSITION. Let $C$ be a cell in $\mathbb{R}^m$ with vertices $v_0, \ldots, v_n$. Then the following are equivalent:

(i) $C$ is an $n$-simplex.

(ii) The set $\{v_1 - v_0, \ldots, v_n - v_0\}$ is linearly independent in $\mathbb{R}^m$.

(iii) Every $a \in C$ can be uniquely written as a convex combination $a = \sum_{k=0}^{n} \lambda_k v_k$

with all $\lambda_k \geq 0$ and $\sum_{k=0}^{n} \lambda_k = 1$.

PROOF: (i) $\iff$ (ii): This follows immediately from XI.24.1.4., since $C$ is $n$-dimensional if and only if $\{v_1 - v_0, \ldots, v_n - v_0\}$ spans an $n$-dimensional subset of $\mathbb{R}^m$.

(ii) $\Rightarrow$ (iii): Since $C$ is a cell, every element of $C$ can be written as a convex combination of the $v_k$. So if (iii) is false, there are $(\lambda_0, \ldots, \lambda_n) \neq (\mu_0, \ldots, \mu_n)$ satisfying the conditions of (iii) with

$$0 = \sum_{k=0}^{n} \lambda_k v_k = \sum_{k=0}^{n} \mu_k v_k$$

$$0 = \sum_{k=0}^{n} (\lambda_k - \mu_k) v_k = \sum_{k=1}^{n} (\lambda_k - \mu_k) (v_k - v_0) + \sum_{k=0}^{n} (\lambda_k - \mu_k) v_0 = \sum_{k=1}^{n} (\lambda_k - \mu_k) (v_k - v_0)$$

since $\sum_{k=0}^{n} (\lambda_k - \mu_k) = 0$ because $\sum_{k=0}^{n} \lambda_k = \sum_{k=0}^{n} \mu_k = 1$. So $\{v_1 - v_0, \ldots, v_n - v_0\}$ is not linearly independent.

(iii) $\Rightarrow$ (ii):

XI.24.1.14. If $C$ is a simplex, the numbers $\lambda_k$ for which $a = \sum_{k=0}^{n} \lambda_k v_k$ are called the barycentric coordinates of $a$ in $C$. These are indexed by the vertices of $C$, i.e. $\lambda_k$ is the barycentric coordinate of $a$ with respect to $v_k$ (in $C$). If $C$ is not a simplex, then barycentric coordinates are not well defined since some points can be written in more than one way as convex combinations of the vertices.
XI.24.1.15. In $\mathbb{R}^{n+1}$, let $\Sigma_n$ be the convex hull of the set of standard basis vectors $\{e_0, \ldots, e_n\}$ (it is convenient here to index the coordinates from 0 to $n$ rather than from 1 to $n+1$). Then

$$\Sigma_n = \left\{ (\lambda_0, \ldots, \lambda_n) \in \mathbb{R}^{n+1} : \lambda_k \geq 0, \sum_{k=0}^{n} \lambda_k = 1 \right\}$$

is called the standard $n$-simplex. If $S$ is any $n$-simplex in $\mathbb{R}^m$ with vertices $v_0, \ldots, v_n$, then there is a unique affine map $\phi : \mathbb{R}^{n+1} \to \mathbb{R}^m$ with $\phi(e_k) = v_k$ for each $k$, which maps $\Sigma_n$ homeomorphically onto $S$. Thus every $n$-simplex “looks the same” and is affinely homeomorphic to the standard $n$-simplex.

We sometimes regard the simplex in $\mathbb{R}^n$ with vertices $0, e_1, \ldots, e_n$, i.e.

$$\left\{ (\lambda_1, \ldots, \lambda_n) : \lambda_k \geq 0, \sum_{k=1}^{n} \lambda_k \leq 1 \right\}$$

as the standard $n$-simplex. There is an obvious affine homeomorphism between this simplex and $\Sigma_n$.

A 0-simplex is a single point. A 1-simplex is a (closed) line segment. A 2-simplex is a (solid) triangle. A 3-simplex is a (solid) tetrahedron.

XI.24.1.16. If $S$ is an $n$-simplex with vertices $v_0, \ldots, v_n$, and $w_0, \ldots, w_n$ are any $n+1$ points in $\mathbb{R}^m$, then there is a unique affine function $f$ from $S$ onto the convex hull of $\{w_0, \ldots, w_n\}$ with $f(v_k) = w_k$ for all $k$. The obvious formula is

$$f \left( \sum_{k=0}^{n} \lambda_k v_k \right) = \sum_{k=0}^{n} \lambda_k w_k$$

for $\lambda_k \geq 0$, $\sum_{k=0}^{n} \lambda_k = 1$.

XI.24.1.17. Proposition. Every simplex is a finite intersection of (closed) half-spaces.

Polyhedra

XI.24.1.18. Definition. A polyhedron is a subset of some $\mathbb{R}^n$ which is a finite union of cells.

This is not the way the definition of a polyhedron is usually expressed in PL topology, but is equivalent (in the compact case). The PL definition (XI.24.1.34.) is hard to motivate geometrically, so we have chosen this alternative definition.

XI.24.1.19. Definition. Let $B$ be a (usually compact) subset of $\mathbb{R}^n$, and $a$ a point of $\mathbb{R}^n$ not in $B$. The join $aB$ of $a$ and $B$ is the union of all line segments in $\mathbb{R}^n$ between $a$ and a point of $B$. The join is a cone with vertex $a$ and base $B$ if every $x \in aB \setminus \{a\}$ can be uniquely written as $\lambda a(1 - \lambda)b$ for some $b \in B$ and $0 \leq \lambda < 1$.

XI.24.1.20. It is easily seen that $aB$ is a cone if and only if every ray beginning at $a$ intersects $B$ in at most one point. A cone $aB$ is compact if $B$ is compact.
XI.24.1.21. Definition. Let $X$ be a subset of $\mathbb{R}^n$, $a \in X$, and $\epsilon > 0$. Set

$$N_\epsilon(a, X) = \{ x \in X : \rho_\infty(a, x) \leq \epsilon \}$$

$$\hat{N}_\epsilon(a, X) = \{ x \in X : \rho_\infty(a, x) = \epsilon \}$$

$$\check{N}_\epsilon(a, X) = \{ x \in X : \rho_\infty(a, x) < \epsilon \} .$$

Then $\hat{N}_\epsilon(a, X)$ is an open neighborhood of $a$ in $X$, and $N_\epsilon(a, X)$ is a closed neighborhood of $a$ in $X$. If $N_\epsilon(a, X)$ is a cone with vertex $a$ and base $\hat{N}_\epsilon(a, X)$, $\check{N}_\epsilon(a, X)$ is called the $\epsilon$-star of $a$ in $X$; if not, the $\epsilon$-star of $a$ in $X$ is not defined. $X$ is locally conical if for every $a \in X$ there is an $\epsilon$-star of $a$ in $X$ for some $\epsilon > 0$.

The set $N_\epsilon(a, X)$ is the intersection of $X$ with a cube of side $2\epsilon$ centered at $a$, and $\check{N}_\epsilon(a, X)$ is the intersection of $X$ with the boundary of this cube.

If $a$ has an $\epsilon$-star in $X$, it also has an $\epsilon'$-star for any $\epsilon' < \epsilon$.

Roughly speaking, a set $X$ in $\mathbb{R}^n$ is locally conical if every point $a$ of $X$ has a neighborhood $U$ in $\mathbb{R}^n$ such that $X \cap U$ is a union of ray segments with endpoint $a$ with lengths bounded away from zero.

XI.24.1.22. Examples. (i) Any open set in $\mathbb{R}^n$ is locally conical: if $U$ is open and $a \in U$, then for sufficiently small $\epsilon > 0$ $N_\epsilon(a, U)$ is an entire closed cube centered at $a$.

(ii) Any affine subspace or half-space in $\mathbb{R}^n$ is locally conical: in fact, any point in such a subset has an $\epsilon$-star for every $\epsilon > 0$.

(iii) A polygon in $\mathbb{R}^2$, with or without interior, is locally conical.

(iv) A circle is not locally conical: in fact, no point has an $\epsilon$-star for any $\epsilon > 0$. Similarly, a closed disk is not locally conical.

The important fact justifying our nonstandard definition of polyhedron is:

XI.24.1.23. Theorem. Let $X$ be a compact subset of $\mathbb{R}^m$. The following are equivalent:

(i) $X$ is a polyhedron.

(ii) $X$ is a finite union of simplexes.

(iii) $X$ is locally conical.

The theorem will be proved using some lemmas.

XI.24.1.24. Lemma. Let $X_1, \ldots, X_n$ be subsets of $\mathbb{R}^m$, $X = \bigcap_{k=1}^n X_k$, and $a \in X$. If $N_{\epsilon_k}(a, X_k)$ is an $\epsilon_k$-star of $a$ in $X_k$ for each $k$, and $\epsilon = \min(\epsilon_1, \ldots, \epsilon_n)$, then $N_\epsilon(a, X) = \bigcap_{k=1}^n N_\epsilon(a, X_k)$ is an $\epsilon$-star of $a$ in $X$.

Proof: Obvious from the fact that $N_\epsilon(a, X_k)$ is an $\epsilon$-star of $a$ in $X_k$ for all $k$. ☐
Let $a \in X$. By renumbering, suppose $a \in \cap_{k=1}^r X_k$ and $x \notin \cup_{k=r+1}^n X_k$. Let $N_{\epsilon_k}(a, X_k)$ be an $\epsilon_k$-star of $a$ in $X_k$ for $1 \leq k \leq r$. Choose $\epsilon > 0$ so that $\epsilon \leq \epsilon_k$ for $1 \leq k \leq r$ and $\epsilon < \rho_{\infty}(a, \cup_{k=r+1}^n X_k)$. Then it is easily checked that $N_\epsilon(a, X)$ is an $\epsilon$-star of $a$ in $X$.

**Lemma.** Every simplex is locally conical.

Proof: A half-space is locally conical (XI.24.1.22.(ii)); thus any simplex is locally conical by XI.24.1.17. and XI.24.1.25.

**Lemma.** Let $C$ be a cell in $\mathbb{R}^n$, and $a \in C$. Then $C$ is the union of cones of the form $aF$, where $F$ is a face of $C$ not containing $a$.

Proof: Let $x \in C$, $x \neq a$. If $R$ is the ray starting at $a$ through $x$, then $R \cap C$ is a segment from $a$ to a point $y$. The face $F_y$ (XI.24.1.8.) cannot contain $a$ since there is no line segment in $C$ through $a$ and $y$ containing $y$ in its interior (and $y \neq a$). Thus $x$ is in the union of the sets $aF$. Such a set is a cone, since if a ray from $a$ intersects a face $F$ in points $z$ and $z'$ with $z$ between $a$ and $z'$, by definition of a face $F$ must contain $a$.

**Proof of XI.24.1.23:** We have (ii) $\Rightarrow$ (iii) by XI.24.1.27. and XI.24.1.26., and (ii) $\Rightarrow$ (i) is trivial.

(i) $\Rightarrow$ (ii): By XI.24.1.26., it suffices to show that every cell is a finite union of simplexes. We prove this by induction on the dimension of the cell. If $C$ is a cell of dimension $\leq 1$, then $C$ is a point or line segment, hence a simplex. Now suppose all cells of dimension $< n$ are unions of simplexes, and let $C$ be a cell of dimension $n$. If $a \notin C$, then by XI.24.1.28. $C$ is a (finite) union of cones of the form $aF$, where $F$ is a face of $C$ not containing $a$. Each such $F$ is a cell of dimension $< n$, hence a finite union of simplexes $S_1, \ldots, S_r$. Then $aF$ is the union of $aS_1, \ldots, aS_r$, each of which is a simplex by (i).

It remains to show (iii) $\Rightarrow$ (ii). This will be shown by induction on the dimension of the Euclidean space containing the set. A compact locally conical subset of $\mathbb{R}^1$ is a finite union of points and closed line segments, each of which is a simplex. Now suppose all compact locally conical subsets of Euclidean spaces of dimension less than $n$ are finite unions of simplexes. Let $X$ be a compact locally conical subset of $\mathbb{R}^n$. To show that $X$ is a finite union of simplexes, it suffices to show that for every $a \in X$, an $\epsilon$-star of $a$ in $X$ is a finite union of simplexes for some $\epsilon = \epsilon_a$. For then the sets $N_{\epsilon_a}(a, X)$ form an open cover of $X$ which has a finite subcover. So suppose $N_\epsilon(a, X)$ is an $\epsilon$-star. Let $F$ be a proper face of the closed cube $\{x \in \mathbb{R}^n : \rho(a, x) \leq \epsilon\}$. Then $F$ is compact and locally conical, and thus $X \cap F$ is compact and locally conical by XI.24.1.25. Since $X \cap F$ is contained in a Euclidean space of dimension less than $n$, it is a finite union of simplexes. Then $N_\epsilon(a, X)$,
which is a finite union of the $X \cap F$, is a finite union of simplexes $S_1, \ldots, S_r$. But then $N_\epsilon(a, X)$ is the union of $aS_1, \ldots, aS_r$, each of which is a simplex by ()

This completes the proof of XI.24.1.23.

Here is a consequence which is not at all obvious directly from the definition:

**XI.24.1.29. Corollary.** A finite intersection of polyhedra is a polyhedron.

**Proof:** A finite intersection of polyhedra is compact, and locally conical by XI.24.1.25.

**XI.24.1.30. Proposition.** Let $X$ be a polyhedron in $\mathbb{R}^n$. Then the convex hull of $X$ is a cell. In particular, a convex polyhedron is a cell.

**Proof:** $X$ is a finite union of cells $C_1, \ldots, C_r$. Let $S$ be the set of all extreme points of $C_1, \ldots, C_r$. Then $S$ is a finite set and the convex hull of $S$ is the convex hull of $X$.

We then obtain one direction of XI.24.1.11.:

**XI.24.1.31. Corollary.** Let $X$ be a compact set in $\mathbb{R}^n$ which is a finite intersection of half-spaces. Then $X$ is a cell.

**Proof:** By XI.24.1.25., $X$ is locally conical, hence a polyhedron, and $X$ is also convex.

**XI.24.1.32.** Let $X$ be a subset of $\mathbb{R}^n$. If $a \in X$, a star of $a$ in $X$ is a closed neighborhood of $a$ in $X$ which is a cone of the form $aL$, where $L$ is a compact subset of $X$ not containing $a$ ($L = \emptyset$ is allowed, in which case $aL = \{a\}$). An $\epsilon$-star of $a$ in $X$ is a star with $L = N_\epsilon(a, X)$ provided this set is compact (e.g. if $X$ is compact).

**XI.24.1.33. Proposition.** Let $X \subseteq \mathbb{R}^n$. Then every point of $X$ has a star if and only if $X$ is locally compact and locally conical.

**Proof:** If $X$ is locally compact and locally conical, and $a \in X$, then $\mathcal{N}_\epsilon(a, X)$ is compact for sufficiently small $\epsilon$, so $a$ has a star. Conversely, if $aL$ is a star of $a$, it is a compact neighborhood of $a$, and $N_\epsilon(a, X)$ is an $\epsilon$-star for any $\epsilon < \rho_\infty(a, L)$, so if every point of $X$ has a star, then $X$ is locally compact and locally conical.

**XI.24.1.34.** In PL topology, a polyhedron is defined to simply be a subset of $\mathbb{R}^n$ in which every point has a star, i.e. a locally compact locally conical subset. While it is useful in PL topology to consider locally conical sets (polyhedra) which are not compact, we will not need to do so, so to avoid having to constantly write “compact polyhedron" we will require polyhedra to be compact.
XI.24.2. Complexes

XI.24.2.1. Definition. A cell complex in $\mathbb{R}^n$ is a finite collection $\mathcal{K}$ of cells in $\mathbb{R}^n$ with the following properties:

(i) If $C \in \mathcal{K}$ and $F$ is a face in $C$, then $F \in \mathcal{K}$.

(ii) If $C_1, C_2 \in \mathcal{K}$ then $C_1 \cap C_2$ is a face of both $C_1$ and $C_2$ (we make the convention that $\emptyset$ is a face of every cell, so $C_1 \cap C_2 = \emptyset$ is allowed).

The dimension of $\mathcal{K}$ is maximum of the dimensions of the cells of $\mathcal{K}$.

The vertices of $\mathcal{K}$ are the 0-cells of $\mathcal{K}$ and the edges of $\mathcal{K}$ are the 1-cells.

If $\mathcal{K}$ is a cell complex in $\mathbb{R}^n$, the union of the cells in $\mathcal{K}$ is a polyhedron in $\mathbb{R}^n$ denoted $|\mathcal{K}|$.

A cell complex $\mathcal{K}$ is a collection of cells that fit together nicely. The cell complex $\mathcal{K}$ can be thought of as the polyhedron $|\mathcal{K}|$ with a specified nice subdivision of $|\mathcal{K}|$ into cells. The vertices of $\mathcal{K}$ are the extreme points of the cells in $\mathcal{K}$.

Infinite cell complexes are often used in topology, but for simplicity we will make the convention that cell complexes are finite.

XI.24.2.2. Definition. A simplicial complex in $\mathbb{R}^n$ is a cell complex in which every cell is a simplex.

A triangulation of a polyhedron $X$ is a simplicial complex $\mathcal{K}$ with $|\mathcal{K}| = X$.

A simplicial complex in $\mathbb{R}^n$ is often called a geometric simplicial complex to distinguish it from an abstract simplicial complex (XI.24.2.12.), which is a combinatorial object.

We will work almost exclusively with simplicial complexes.

XI.24.2.3. Examples. (i) Any cell $C$ can be regarded as a cell complex $\mathcal{C}$ consisting of all faces of $C$. Then $\mathcal{C} = |\mathcal{C}|$. In particular, any simplex $S$ can be regarded as a simplicial complex $\mathcal{S}$ consisting of the faces of $S$. $\mathcal{S}$ is a triangulation of $S$.

(ii) Let $X$ be a finite union of line segments in $\mathbb{R}^n$, such that any two of the line segments are either disjoint or have only an endpoint in common. Then $X$ is a polyhedron which is the underlying space of a simplicial complex $\mathcal{K}$ consisting of the line segments and their endpoints (plus $\emptyset$); $\mathcal{K}$ is a triangulation of $X$. In particular, any (simple) polygon in $\mathbb{R}^2$, not including interior, is the underlying space of a simplicial complex. (A simple closed polygon including interior is also the underlying space of a simplicial complex, although this is somewhat harder to show (i).)

Subcomplexes

XI.24.2.4. Definition. Let $\mathcal{K}$ be a cell complex. A subcomplex of $\mathcal{K}$ is a subset of $\mathcal{K}$ which is itself a cell complex.

A subset $\mathcal{K}'$ of a complex $\mathcal{K}$ is a subcomplex if and only if, whenever $C \in \mathcal{K}'$, then all faces of $C$ are also in $\mathcal{K}'$. A subcomplex of a simplicial complex is a simplicial complex.

There are important special subcomplexes of any complex:
XI.24.2.5. Definition. Let $K$ be a cell complex, and $d \in \mathbb{N} \cup \{0\}$. The collection $K_d$ of all cells in $K$ of dimension $\leq d$ is a subcomplex of $K$ called the $d$-skeleton of $K$.

Thus the 0-skeleton consists of the vertices, the 1-skeleton the vertices and edges, etc.

Subdivisions

XI.24.2.6. Definition. Let $K$ and $K'$ be cell complexes. $K'$ is a subdivision of $K$ if $|K'| = |K|$ and every cell in $K'$ is contained in a cell in $K$.

It is easily seen that $K'$ is a subdivision of $K$ if and only if each cell in $K$ is a union of cells in $K'$.

Simplicial Maps

XI.24.2.7. Definition. (i) Let $K$ be a cell complex in $\mathbb{R}^n$. A function $f : |K| \to \mathbb{R}^m$ is simplicial (with respect to $K$), or a simplicial map, if the restriction of $f$ to each cell of $K$ is affine.

(ii) Let $K$ and $K'$ be cell complexes. A function $f : |K| \to |K'|$ is simplicial (with respect to $K$ and $K'$) if the image of each cell in $K$ is contained in a cell of $K'$ and the restriction of $f$ to each cell of $K$ is affine.

A simplicial isomorphism between cell complexes $K$ and $K'$ is a bijection $f$ between $|K|$ and $|K'|$ such that $f$ and $f^{-1}$ are simplicial.

XI.24.2.8. A simplicial map is continuous by (), since each cell is closed in $|K|$ and the restriction to each cell is continuous. A composition of simplicial maps is simplicial.

XI.24.2.9. Examples. (i) If $K'$ is a subcomplex of $K$, the inclusion map from $|K'|$ to $|K|$ is simplicial for $K'$ and $K$.

(ii) If $K'$ is a subdivision of $K$, the identity map on $|K'| = |K|$ is simplicial for $K'$ and $K$. But it is not simplicial for $K$ and $K'$ unless $K' = K$.

(iii) Let $P$ be a (simple closed) polygon in $\mathbb{R}^2$, not including interior, with $n$ sides. Then $P$ can be regarded as a simplicial complex $\mathcal{P}$ with $n$ 0-simplexes (the vertices) and $n$ 1-simplexes (the sides). If $Q$ is another polygon with $n$ sides, regarded as a simplicial complex $\mathcal{Q}$, then $\mathcal{P}$ and $\mathcal{Q}$ are simplicially isomorphic.

But if $P$ and $Q$ are convex polygons with $n$ sides and the interiors are included, regarded as cell complexes $\mathcal{P}$, $\mathcal{Q}$ with one 2-cell and $n$ 1-cells and 0-cells as above, then $\mathcal{P}$ and $\mathcal{Q}$ are not generally simplicially isomorphic, and the conditions for simplicial isomorphism are restrictive: for example, if $P$ is a square, then $\mathcal{P}$ and $\mathcal{Q}$ are simplicially isomorphic if and only if the quadrilateral $Q$ is a parallelogram.

For simplicial complexes, conditions for simplicial isomorphism are much less stringent (XI.24.2.10).

XI.24.2.10. Proposition. Let $K$ and $K'$ be simplicial complexes, and let $\phi$ be a one-one correspondence between the simplexes in $K$ and the simplexes in $K'$ with the property that if $S \in K$ and $F$ is a face of $S$, then $\phi(F)$ is a face of $\phi(S)$. Then there is a unique simplicial isomorphism $f : |K| \to |K'|$ such that $f(S) = \phi(S)$ for every $S \in K$.

Proof: By counting numbers of faces, it follows that $\phi$ must give a bijection between the $d$-simplexes in $K$ and the $d$-simplexes of $K'$ for each $d$. Since the 0-simplexes are just the vertices, $\phi$ gives a one-one
correspondence between the vertices of $\mathcal{K}$ and the vertices of $\mathcal{K}'$. If $v_0, \ldots, v_n$ are the vertices of $\mathcal{K}$, set $w_k = \phi(v_k)$ for each $k$. Then if $S$ is a simplex in $\mathcal{K}$ with vertices $v_{i_1}, \ldots, v_{i_r}$, then $\phi(S)$ must be a simplex in $\mathcal{K}'$ with vertices $w_{i_1}, \ldots, w_{i_r}$, and we may define $f : S \to \phi(S)$ by

$$f \left( \sum_{k=1}^{r} \lambda_k v_{i_k} \right) = \sum_{k=1}^{r} \lambda_k w_{i_k}.$$  

It is easily checked that if $x \in |\mathcal{K}|$, this definition of $f(x)$ is independent of which of the simplexes of $\mathcal{K}$ containing $x$ is used, and thus we obtain a well-defined function $f : |\mathcal{K}| \to |\mathcal{K}'|$ which obviously has the right properties.

\textbf{Barycentric Coordinates}

\textbf{XI.24.2.11.} The construction in the proof of XI.24.2.10. can be thought of in another way. If $\mathcal{K}$ is a simplicial complex and $v_0, \ldots, v_n$ are the vertices of $\mathcal{K}$, and $x \in |\mathcal{K}|$, then for any simplex $S$ of $\mathcal{K}$ containing $x$ with vertices $v_{i_1}, \ldots, v_{i_r}$, $x$ has barycentric coordinates $(\lambda_{i_k})$ with respect to $v_{i_k}$. These barycentric coordinates are defined for any vertex of any simplex of $\mathcal{K}$ containing $x$, and do not depend on which simplex containing $x$ and that vertex is used. If $v_j$ is a vertex of $\mathcal{K}$ and $x$ is not contained in any simplex of $\mathcal{K}$ with $v_j$ a vertex, set the barycentric coordinate $\lambda_j(x)$ equal to 0. Thus the barycentric coordinates $\lambda_0(x), \ldots, \lambda_n(x)$ of $x$ are well defined for every vertex, and satisfy $\lambda_k(x) \geq 0$ for all $k$ and $\sum_{k=0}^{n} \lambda_k(x) = 1$. We may symbolically write

$$x = \sum_{k=0}^{n} \lambda_k(x)v_k$$

interpreted as follows: if the set of vertices $v_k$ for which $\lambda_k(x) > 0$ is $\{v_{i_1}, \ldots, v_{i_r}\}$, then the simplex $S$ with vertices $v_{i_1}, \ldots, v_{i_r}$ is in $\mathcal{K}$ and $x$ is the point of $S$ with barycentric coordinates $\lambda_{i_1}(x), \ldots, \lambda_{i_r}(x)$ with respect to $v_{i_1}, \ldots, v_{i_r}$.

\textbf{Abstract Simplicial Complexes}

The existence of barycentric coordinates (XI.24.2.11.) in any simplicial complex allows the definition of simplicial complexes as combinatorial objects:

\textbf{XI.24.2.12.} \textbf{Definition.} An \textit{abstract simplicial complex} is a collection $\mathcal{K}$ of subsets of a finite set $\{v_0, \ldots, v_n\}$ such that $\{v_k\} \in \mathcal{K}$ for all $k$ and if $S \in \mathcal{K}$, then all subsets of $S$ are also in $\mathcal{K}$. The $v_k$ are called the \textit{vertices} of $\mathcal{K}$.

If $\mathcal{K}$ and $\mathcal{K}'$ are abstract simplicial complexes with sets of vertices $V$ and $V'$ respectively, a \textit{simplicial isomorphism} between $\mathcal{K}$ and $\mathcal{K}'$ is a bijection $\phi$ from $V$ to $V'$ such that $F \subseteq V$ is in $\mathcal{K}$ if and only if $\phi(F) \in \mathcal{K}'$.

A \textit{geometric realization} of an abstract simplicial complex $\mathcal{K}$ is a (geometric) simplicial complex $\mathcal{K}'$ and a bijection $\phi$ from the set $\{v_0, \ldots, v_n\}$ of vertices of $\mathcal{K}$ to the vertices of $\mathcal{K}'$, say $w_k = \phi(v_k)$, such that the simplex with vertices $\{w_{i_1}, \ldots, w_{i_r}\}$ is in $\mathcal{K}'$ if and only if $\{v_{i_1}, \ldots, v_{i_r}\} \in \mathcal{K}$.

The next result is obvious from XI.24.2.10.:
XI.24.2.13. Proposition. Let $\mathcal{K}$ be an abstract simplicial complex. Then there is a unique simplicial isomorphism between any two geometric realizations of $\mathcal{K}$ which extends the identification of the vertices by the realizations. More generally, if $\mathcal{K}$ and $\mathcal{K}'$ are abstract simplicial complexes and $\phi$ is a simplicial isomorphism from $\mathcal{K}$ to $\mathcal{K}'$, then there is a unique simplicial isomorphism from any geometric realization of $\mathcal{K}$ to any geometric realization of $\mathcal{K}'$ extending the identification of the vertices via $\phi$.

XI.24.2.14. Every abstract simplicial complex has a geometric realization. In fact, if an ordering $v_0, v_1, \ldots, v_n$ on the vertices of an abstract simplicial complex $\mathcal{K}$ is fixed, there is a corresponding standard realization of $\mathcal{K}$ as a subcomplex $\mathcal{K}'$ of the standard $n$-simplex $\Sigma_n$: just let $\mathcal{K}'$ be the collection of all faces of $\Sigma_n$ with vertices $\{e_i, \ldots, e_r\}$ for sets $\{v_i, \ldots, v_r\}$ in $\mathcal{K}$.

Thus there is a natural one-one correspondence between simplicial isomorphism classes of abstract simplicial complexes and simplicial isomorphism classes of geometric simplicial complexes.

XI.24.3. Triangulations

Recall (XI.24.2.2) that a triangulation of a polyhedron $P$ is a simplicial complex $\mathcal{K}$ with $|\mathcal{K}| = P$.

XI.24.3.1. Theorem. Every polyhedron has a triangulation.

Triangulations of Topological Spaces

XI.24.3.2. Definition. Let $X$ be a topological space (usually a compact Hausdorff space). A weak triangulation of $X$ is a simplicial complex $\mathcal{K}$ and a continuous function $f : X \to |\mathcal{K}|$ such that every vertex of $\mathcal{K}$ is in the range of $f$. A weak triangulation $f$ of $X$ is a triangulation of $X$ if $f$ is a homeomorphism from $X$ onto $|\mathcal{K}|$.

This terminology is consistent: if $P$ is a polyhedron, then a triangulation of $P$ in the simplicial sense (XI.24.2.2) is a triangulation in the topological sense (XI.24.3.2) when $P$ is regarded as a topological space. The other direction is not well defined, since a polyhedron is more than a topological space: it also has a “local affine structure” (XI.24.4.1) which is not part of the topological structure. See () for further comments.

Note that a topological space with a triangulation must be homeomorphic to a polyhedron, hence a compact metrizable space. But spaces which are not compact and/or metrizable can have weak triangulations: for example, any constant function from a topological space $X$ to $\mathbb{R}^n$ is a weak triangulation.

Triangulations and Partitions of Unity

Weak triangulations of a topological space $X$ are effectively the same thing as finite partitions of unity on $X$.  

1348
XI.24.3.3. Let $X$ be a topological space and \{w_0, \ldots, w_n\} a finite partition of unity on $X$. Define a continuous function $f$ from $X$ to the standard $n$-simplex $\Sigma_n$ by

$$f(x) = \sum_{k=0}^{n} f_k(x) e_k.$$ 

If $\Sigma_n$ is regarded as a simplicial complex, then $f$ is a weak triangulation of $X$. (The requirement that each $f_k$ take the value 1 is precisely the same as the requirement that the range of $f$ contain all vertices of $\Sigma_n$.) Note that the range of $f$ may well be contained in a proper subcomplex of $\Sigma_n$ containing all the vertices.

XI.24.3.4. Conversely, let $f$ be a weak triangulation from $X$ to a simplicial complex $K$, and let $\{w_0, \ldots, w_n\}$ be the vertices of $K$. If $a \in |K|$, let $\lambda_k(a)$ be the barycentric coordinate of $a$ with respect to $w_k$ in $K$. For each $k$ set $f_k = \lambda_k \circ f$, i.e. for $x \in X$ set $f_k(x) = \lambda_k(f(x))$. Then the $f_k$ give a finite partition of unity on $X$.

XI.24.3.5. Suppose $f : X \to |K|$ is a weak triangulation, and let $v_0, \ldots, v_n$ be the vertices of $K$ and $f_0, \ldots, f_n$ the corresponding partition of unity. For each $k$ let $V_k$ be the set of $a \in |K|$ with $\lambda_k(a) > 0$; $V_k$ is the set of points $a$ of $|K|$ such that every simplex in $K$ containing $a$ has $v_k$ as a vertex. Then $V_k$ is open in $|K|$, and if $U_k = f^{-1}(V_k) = \{x \in X : f_k(x) > 0\}$, then $U = \{U_0, \ldots, U_n\}$ is an open cover of $X$, and $\{f_0, \ldots, f_n\}$ is a partition of unity subordinate to $U$.

XI.24.3.6. So if $X$ is a compact Hausdorff space, the theory of partitions of unity assures the existence of “fine” weak triangulations. In fact, we have:

XI.24.3.7. Theorem. Let $X$ be a compact Hausdorff space and $W$ an open cover of $X$. Then there is a surjective weak triangulation $f : X \to |K|$ such that the cover $U$ of XI.24.3.5. is a refinement of $W$.

Proof: By passing to a subcover, we may assume $W$ is finite. Let $V$ be an open star-refinement of $W$, which we may also assume finite, and $\{f_0, \ldots, f_n\}$ a partition of unity subordinate to $V$. The corresponding weak triangulation $f : X \to |K| = \Sigma_n$ works except that $f$ may not be surjective.

We modify $f$ in stages and reduce $K'$ to a subcomplex, beginning with the largest simplex. To do the modification on a simplex $S$ which is not a face of another simplex in $K'$, suppose the interior of $S$ is not entirely contained in the range of $f$ (if it is, no modification for this $S$ is necessary). If $a$ is an interior point of $S$ not contained in the range, there is a retraction $r$ (e.g. radial projection from $a$) from $S \setminus \{a\}$ onto its boundary. Replace $K'$ by $K' \setminus \{S\}$ and $f$ by $r \circ f$. Then the inverse image of the star of any vertex in the reduced complex under $r \circ f$ is contained in the inverse image of the star of that vertex in $K'$ under $f$, so is contained in one of the sets of $W$. The modified $r \circ f$ agrees with $f$ on any $x \in X$ for which $f(x)$ is not in the interior of $S$. Repeat the reduction through all simplexes in $K'$ which are not faces of remaining simplexes after the previous reduction. Since all vertices of $K'$ are in the range of $f$ at every stage, after a finite number of steps a subcomplex $K$ and modified $f$ will be surjective. 

\[\Box\]
XI.24.3.8. **Corollary.** Let $X$ be a compact metrizable space, $\rho$ a fixed metric on $X$, and $\epsilon > 0$. Then there is a surjective weak triangulation $f : X \to |K|$ of $X$, such that for every vertex $v_k$ of $|K|$, the $\rho$-diameter of $f^{-1}(V_k)$ is less than $\epsilon$.

**Proof:** Let $W$ be the open cover of $X$ by balls of $\rho$-diameter $< \epsilon$. 

The Nerve of an Open Cover

XI.24.4. **Piecewise-Linear Maps**

**XI.24.4.1.** If $X$ is a polyhedron and $a \in X$, then for any $b \in X$ sufficiently close to $a$ (within a star of $a$ in $X$) there is an unambiguous notion of the line segment from $a$ to $b$ in $X$, and of $\lambda a + (1 - \lambda)b$ for $0 \leq \lambda \leq 1$, i.e. $X$ has a “local affine structure.” (Existence of such a structure does not characterize polyhedra, however; for example, any convex set also has such a local affine structure.)

The local affine structure is “continuous”:

**XI.24.4.2.** **Proposition.** Let $X$ be a subset of $\mathbb{R}^n$, $a \in X$, and $aL$ a star of $a$ in $X$ (with $L$ compact). If $(\lambda_n a + (1 - \lambda_n)b_n)$ is a sequence in $aL$ (with $0 \leq \lambda_n \leq 1$ and $b_n \in L$ for all $n$), then $\lambda_n a + (1 - \lambda_n)b_n \to \lambda a + (1 - \lambda)b \in \mathbb{R}^n$ with $b \in L$ if and only if $\lambda_n \to \lambda$ and, if $\lambda \neq 1$, $b_n \to b$.

**Proof:** If $\lambda_n \to \lambda$ and $b_n \to b$, then $\lambda_n a + (1 - \lambda_n)b_n \to \lambda a + (1 - \lambda)b$, and $b \in L$ since $L$ is closed in $\mathbb{R}^n$. Conversely, suppose $\lambda_n a + (1 - \lambda_n)b_n \to \lambda a + (1 - \lambda)b$ with $b \in L$. Since $[0, 1]$ and $L$ are compact, any subsequence of $(\lambda_n a + (1 - \lambda_n)b_n)$ has a subsubsequence $(\lambda_{k_n} a + (1 - \lambda_{k_n})b_{k_n})$ such that $(\lambda_{k_n})$ and $(b_{k_n})$ both converge, say to $\mu$ and $c$ respectively, with $0 \leq \mu \leq 1$ and $c \in L$. Then $\lambda_{k_n} a + (1 - \lambda_{k_n})b_{k_n} \to \mu a + (1 - \mu)c$. By uniqueness of limits in $\mathbb{R}^n$, $\mu a + (1 - \mu)c = \lambda a + (1 - \lambda)b$, and by uniqueness of representation of points of $aL$, we must have $\mu = \lambda$ and, if $\lambda \neq 1$, $c = b$. Thus every subsequence of $(\lambda_n)$ has a subsequence converging to $\lambda$, and if $\lambda \neq 1$ every subsequence of $(b_n)$ has a subsubsequence converging to $b$, so $\lambda_n \to \lambda$ and, if $\lambda \neq 1$, $b_n \to b$.

We can use the local affine structure to define piecewise-linear maps between polyhedra:

**XI.24.4.3.** **Definition.** Let $X$ and $Y$ be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. A function $f : X \to Y$ is piecewise-linear (PL) if for every $a \in X$ there are $\delta > 0$, $\epsilon > 0$ such that

(i) $a$ has a $\delta$-star in $X$ and $f(a)$ has an $\epsilon$-star in $Y$

(ii) $f(N_\delta(a, X)) \subseteq N_\epsilon(f(a), Y)$

(iii) $f(\lambda a + (1 - \lambda)b) = \lambda f(a) + (1 - \lambda)f(b)$ for all $b \in N_\delta(a, X), 0 \leq \lambda \leq 1$.

Such a function might be more properly called something like *locally affine*, but the name “piecewise-linear” is standard, and is appropriate since $X$ can indeed be broken up into pieces on which $f$ is linear (affine) (XI.24.4.6).

The next fact may seem obvious, but requires proof:

1350
XI.24.4. \textbf{Proposition.} Let $X$ and $Y$ be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a PL function. Then $f$ is continuous.

\textbf{Proof:} Let $(c_n)$ be a sequence in $X$ with $c_n \to a$. Let $\delta > 0$ and $\epsilon > 0$ be as in XI.24.4.3. for $a$. Passing to a tail, we may assume $c_n \in N_{\delta}(a, X)$ for all $n$. Write $c_n = \lambda_n a + (1 - \lambda_n) b_n$ for $0 \leq \lambda_n \leq 1$ and $b_n \in N_{\delta}(a, X)$. Since $c_n \to a$, we have $\lambda_n \to 1$ by XI.24.4.2. We have

$$f(c_n) = \lambda_n f(a) + (1 - \lambda_n) f(b_n) \to f(a) + 0 = f(a)$$

since $\lambda_n \to 1$ and $(f(b_n))$ is a bounded sequence. \hfill \(<)

XI.24.4.5. \textbf{Proposition.} Let $X$ and $Y$ be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a function. Then $f$ is PL if and only if the graph $\Gamma(f)$ is a polyhedron in $\mathbb{R}^{n+m}$.

\textbf{Proof:} It is easy to verify that $\Gamma(f)$ is compact and locally conical if and only if $f$ is PL. \hfill \(<)

XI.24.4.6. \textbf{Corollary.} Let $X$ and $Y$ be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a PL function. Then there is a triangulation of $X$ such that $f$ is simplicial for the triangulation. Conversely, such a map is PL.

\textbf{Proof:} If $f$ is PL, triangulate the polyhedron $\Gamma(f)$. The projection map from $\Gamma(f)$ to $X$ is affine, and is a homeomorphism, so the projections of the simplexes in the triangulation of $\Gamma(f)$ are simplexes in $X$ and give a triangulation of $X$ as desired. The converse is obvious. \hfill \(<)

This result justifies the name “piecewise-linear.”

\textbf{XI.24.5. Exercises}

XI.24.5.1. (a) Show that every affine subspace of $\mathbb{R}^n$ is a finite intersection of (closed) half-spaces.

(b) Let $C$ be a cell in $\mathbb{R}^n$ with vertices $v_1, \ldots, v_m$. Suppose $C$ generates $\mathbb{R}^n$ as an affine subspace. For $1 \leq k \leq m$, let $C_k$ be the cone (in the vector space sense) generated by $C$ with vertex $v_k$, i.e. the union of all rays from $v_k$ passing through other points of $C$. Show that $C = C_1 \cap \cdots \cap C_m$. [Use an argument like the one in the proof of XI.24.1.28.]

(c) Show that each $C_k$ is a finite intersection of half-spaces.

(d) Conclude that every cell is a finite intersection of half-spaces. Use this to give a direct proof of (i) $\Rightarrow$ (iii) of XI.24.1.23.

XI.24.5.2. Let $T$ be an open triangle in $\mathbb{R}^2$ with one vertex added. Show that $T$ is locally conical, but not a polyhedron in the PL sense (XI.24.1.34.) since it is not locally compact.
XI.25. The Metric Space of a Measure Space

If \((X, \mathcal{A}, \mu)\) is a measure space, there is a natural pseudometric on \(\mathcal{A}\) for which the metric quotient is complete.

Recall that if \(A\) and \(B\) are sets, the symmetric difference of \(A\) and \(B\) is

\[A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B).\]

\(A\triangle B\) consists of elements which are in either \(A\) or \(B\), but not both. The size of \(A\triangle B\) is a measure of how much the sets \(A\) and \(B\) differ.

If \(\mathcal{A}\) is an algebra of subsets of a set \(X\), and \(A, B \in \mathcal{A}\), then \((A \triangle B) \in \mathcal{A}\).

Here are some important simple properties of symmetric difference:

**Proposition.** Let \(A, B, C\) be sets. Then

1. \((A \triangle B) \triangle C = A \triangle (B \triangle C)\).
2. \(A \triangle C \subseteq (A \triangle B) \cup (B \triangle C)\).
3. \(A \triangle B = B \triangle A\).
4. \(A \triangle \emptyset = \emptyset\).
5. \(A \triangle \emptyset = A\).

**Definition.** Let \((X, \mathcal{A}, \mu)\) be a measure space. If \(A, B \in \mathcal{A}\), set

\[\rho_\mu(A, B) = \min(\mu(A \triangle B), 1)\]

**Proposition.** If \((X, \mathcal{A}, \mu)\) is a measure space, then \(\rho_\mu\) is a pseudometric on \(\mathcal{A}\).

**Proof:** It is obvious from XI.25(iii) that \(\rho_\mu(A, B) = \rho_\mu(B, A) \geq 0\) for any \(A, B \in \mathcal{A}\), and the triangle inequality follows from XI.25(ii) and monotonicity of \(\mu\).

There is inconsistency in the exact way \(\rho_\mu\) is defined in the literature, but the differences are not important since all standard definitions give uniformly equivalent metrics. The “natural” definition is just \(\rho_\mu(A, B) = \mu(A \triangle B)\), but this can take the value \(\infty\), which is not allowed for a pseudometric.

As usual, we can take a quotient of \(\mathcal{A}\) on which \(\rho_\mu\) defines a metric. The equivalence class of a set \(A \in \mathcal{A}\) is

\[[A] = \{B \in \mathcal{A} : \mu(A \triangle B) = 0\}\]

The set of equivalence classes is often written \(\mathcal{A}/\mathcal{N}\), where \(\mathcal{N}\) is the \(\sigma\)-ideal of \(\mu\)-null sets. We will denote the induced metric on \(\mathcal{A}/\mathcal{N}\) also by \(\rho_\mu\).

If \((A_n)\) is a sequence of sets, recall the definition of \(\limsup_{n \to \infty} A_n\) and \(\liminf_{n \to \infty} A_n\) from (i).
Theorem. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\mathcal{N}\) the set of \(\mu\)-null sets. Let \((A_n)\) be a Cauchy sequence in \(\mathcal{A}\) for \(\rho_\mu\). Then

\[
\left(\limsup_{n \to \infty} A_n\right) \setminus \left(\liminf_{n \to \infty} A_n\right) \in \mathcal{N}
\]

and \([\limsup A_n] = [\liminf A_n] = \lim_{n \to \infty} [A_n]\) in \(\mathcal{A}/\mathcal{N}\). Thus \((\mathcal{A}/\mathcal{N}, \rho_\mu)\) is a complete metric space.

Note that this result has nothing to do with completeness of \(\mu\). In fact, if \((X, \bar{\mathcal{A}}, \bar{\mu})\) is the completion of \((X, \mathcal{A}, \mu)\), and \(\bar{\mathcal{N}}\) is the set of \(\bar{\mu}\)-null sets, there is a natural bijection between \(\bar{\mathcal{A}}/\bar{\mathcal{N}}\) and \(\mathcal{A}/\mathcal{N}\), and \(\rho_\mu\) agrees with \(\rho_\bar{\mu}\) on this space, so the metric space of \(\bar{\mu}\) is the “same” as for \(\mu\).

The usual set operations are continuous with respect to the topology on \(\mathcal{A}\):

**Theorem.** Let \((X, \mathcal{A}, \mu)\) be a measure space and \((\mathcal{A}/\mathcal{N}, \rho_\mu)\) the corresponding metric space. Then

(i) The functions \(([A], [B]) \mapsto [A \cup B], ([A], [B]) \mapsto [A \cap B], ([A], [B]) \mapsto [A \setminus B],\) and \(([A], [B]) \mapsto [A \Delta B]\)

are well defined and jointly continuous, i.e. continuous functions from \((\mathcal{A}/\mathcal{N}) \times (\mathcal{A}/\mathcal{N})\) to \(\mathcal{A}/\mathcal{N}\).

(ii) The function \([A] \mapsto [X \setminus A]\) is a well-defined continuous function from \(\mathcal{A}/\mathcal{N}\) to \(\mathcal{A}/\mathcal{N}\).

(iii) If \(\mathcal{B}\) is an algebra of sets containing \(\mathcal{N}\), the closure of \(\mathcal{B}/\mathcal{N}\) in \(\mathcal{A}/\mathcal{N}\) is \(\sigma(\mathcal{B})/\mathcal{N}\), where \(\sigma(\mathcal{B})\) is the \(\sigma\)-algebra generated by \(\mathcal{B}\).
Chapter XII

Measurable Spaces and Subset Structures

In this chapter, we will set up the structures on which measures can be naturally defined. While these structures are of interest in their own right, the reader should not lose sight of the primary goal of this work: developing a good notion of measure and integration. In this context, this chapter should be regarded as a set of elaborate and technical preliminary work necessary to obtain a good theory of measure. Strictly speaking, not all of the material of this chapter will be necessary for basic measure theory, but it makes sense to give a reasonably complete study of these topics all at once.

The set structures discussed in this chapter should be compared to topologies, another very important kind of set structure discussed in Chapter (). Although the various kinds of set structures have properties making them useful in different contexts, all are of the same general nature and have many features in common.

XII.1. Algebras and σ-Algebras of Sets
XII.1.1. Rings, Semirings, and Semialgebras

In this subsection, we will discuss various types of collections of subsets of a set $X$ which are important to a greater or lesser extent in measure theory. Throughout the section, $X$ will be a fixed set which we will regard as our “universe.”

XII.1.1.1. We first recall some standard operations on subsets of $X$. If $A, B \subseteq X$, then

- The union $A \cup B$ is $\{x \in X : x \in A \text{ or } x \in B\}$.
- The intersection $A \cap B$ is $\{x \in X : x \in A \text{ and } x \in B\}$.
- The complement of $A$ is $A^c = \{x \in X : x \notin A\}$.
- The difference $A \setminus B$ is $A \cap B^c = \{x \in A : x \notin B\}$.
- The symmetric difference $A \triangle B$ is $\{x \in X : x \in A \text{ or } x \in B \text{ but not both}\} = (A \setminus B) \cup (B \setminus A)$.

More generally, if $\{A_i : i \in I\}$ is an indexed collection of subsets of $X$, then

- $\bigcup_{i \in I} A_i = \{x \in X : x \in A_i \text{ for some } i \in I\}$
- $\bigcap_{i \in I} A_i = \{x \in X : x \in A_i \text{ for all } i \in I\}$

XII.1.1.2. Definition. A nonempty collection $S$ of subsets of $X$ is a semiring if

(i) $S$ is closed under finite intersections.
(ii) If $A, B \in S$, then $A \setminus B$ is a finite disjoint union of sets in $S$.

A semiring $S$ for which $X \in S$ is a semialgebra. A semiring $S$ with the property that $X$ is a countable union of sets in $S$ is of countable type. A semialgebra is a semiring of countable type.

Taking $B = A$ in (ii), it follows that $\emptyset$ is in any semiring.

The definition of a semiring may look strange at first, but many important collections of subsets which naturally arise in analysis turn out to be semirings, and on the other hand semirings have enough structure that important constructions can be done with them.

XII.1.1.3. Examples. (i) Here is perhaps the most important example. Let $I$ be the set of intervals in $\mathbb{R}$, either open, closed, or half-open; include in $I$ the “degenerate” intervals $\emptyset$ and singleton sets $\{x\}$ (intervals of length 0). $I$ is closed under finite intersections (in fact under arbitrary intersections), and the complement of one interval in another is either an interval or the disjoint union of two intervals, so $I$ is a semiring. In fact, $I$ is a semialgebra.

(ii) As a variation of (i), let $I_I$ be the set of bounded intervals in $\mathbb{R}$, including degenerate intervals. By the same argument, $I_I$ is a semiring. It is not a semialgebra, but it is a semiring of countable type.

(iii) Similarly, if $\delta > 0$, let $I_\delta$ be the set of intervals in $\mathbb{R}$ of length $\leq \delta$, again including degenerate intervals. By the same argument, using the Archimedean property of $\mathbb{R}$, $I_\delta$ is a semiring of countable type.

(iv) If $C$ is any semiring (e.g. an algebra or $\sigma$-algebra) on $X$, and $\{A_i : i \in I\}$ is a nonempty collection of sets in $C$, then $S = \{S \subseteq C : S \subseteq A_i \text{ for some } i \in I\}$ is a semiring. More generally, if $P$ is a property of sets inherited by all subsets, then the set of $C \subseteq C$ with property $P$ form a semiring.

1355
XII.1.4. The intersection of two semirings is not necessarily a semiring, and hence if \( \mathcal{C} \) is a collection of subsets of \( X \), there is not necessarily a smallest semiring containing \( \mathcal{C} \). But there is a standard construction of a semiring containing \( \mathcal{C} \), which is particularly important in probability:

XII.1.5. Proposition. Let \( \mathcal{C} \) be any nonempty collection of subsets of \( X \), and let

\[
\mathcal{D} = \{ A \setminus B : A, B \in \mathcal{C} \}.
\]

Let \( \mathcal{S} \) be the set of finite intersections of sets in \( \mathcal{C} \cup \mathcal{D} \). Then \( \mathcal{S} \) is a semiring containing \( \mathcal{C} \).

Note that \( \emptyset \in \mathcal{D} \) since it is \( A \setminus A \) for any \( A \in \mathcal{C} \); if \( \emptyset \in \mathcal{C} \), then \( \mathcal{C} \subseteq \mathcal{D} \) since \( A = A \setminus \emptyset \) for any \( A \).

Proof: (Outline) This is completely routine, but tedious to write out carefully; see XII.1.5.1. The proof consists of repeated applications of the following facts for subsets \( A, B, C \) of \( X \):

\[
(A \setminus B) \setminus C = (A \setminus C) \cap (B \setminus C).
\]

\[
(A \setminus B) \setminus C = (A \setminus C) \cap (A \setminus B).
\]

\( A \setminus (B \setminus C) \) is the disjoint union of \( A \setminus B \) and \( A \setminus B \cap C \).

\( A \setminus (B \cap C) \) is the disjoint union of \( A \setminus B \) and \( (A \setminus B) \setminus C \).

The details of the proof are left to the reader, who should at least verify the above facts.

A similar proof shows the following:

XII.1.6. Proposition. Let \( \mathcal{C} \) be a nonempty collection of subsets of \( X \) with the property that \( A \in \mathcal{C} \) implies \( A^c \in \mathcal{C} \). Let \( \mathcal{S} \) be the collection of finite intersections of sets in \( \mathcal{C} \). Then \( \mathcal{S} \) is a semialgebra.

Collections with more structure behave more nicely and are of central importance in measure theory:

XII.1.7. Definition. A ring of subsets of a set \( X \) is a nonempty collection of subsets of \( X \) which is closed under finite unions and relative complements.

XII.1.8. Proposition. A ring is closed under symmetric difference and finite intersections; in particular, a ring is a semiring.

Proof: If \( A, B \subseteq X \), then \( A \Delta B = (A \setminus B) \cup (B \setminus A) \) and \( A \cap B = (A \cup B) \setminus (A \Delta B) \).

1356
XII.1.1.9. The converse also holds: a nonempty collection of subsets of \( X \) closed under finite intersections and symmetric differences is a ring, because of the relations

\[
A \cup B = (A \oplus B) \Delta (A \cap B), \quad A \setminus B = A \oplus (A \cap B)
\]

for subsets of \( X \). This explains the use of the term “ring”: the operations \( A + B = A \oplus B \) and \( AB = A \cap B \) make \( \mathcal{P}(X) \) into a ring in the algebraic sense of III.3.6.3. (in fact, a special kind of ring called a Boolean algebra); a ring of subsets is precisely a subring of \( \mathcal{P}(X) \) with these operations. This use of ring (of subsets) is apparently due to Halmos [Hal50].

In many references, especially older ones, \( A \cap B \) is written \( AB \), and in many of these \( A \setminus B \) is written \( A - B \). But \( \mathcal{P}(X) \) is not a ring under these algebraic operations. Also, for sets of numbers or subsets of a topological group, these notations are commonly used to denote something quite different (). Thus we will not use the notation \( AB \) or \( A + B \) for set-theoretic operations.

XII.1.1.10. Any intersection of rings is a ring (the intersection is nonempty since \( \emptyset \) is in every ring). Thus if \( \mathcal{C} \) is any collection of subsets of \( X \), there is a smallest ring of subsets containing \( \mathcal{C} \), called the ring generated by \( \mathcal{C} \) (there is at least one ring containing \( \mathcal{C} \), namely \( \mathcal{P}(X) \)). The ring generated by \( \emptyset \) is \( \{ \emptyset \} \).

XII.1.1.11. Proposition. Let \( \mathcal{S} \) be a semiring. The ring generated by \( \mathcal{S} \) is precisely the set of finite disjoint unions of sets in \( \mathcal{S} \).

Proof: Let \( \mathcal{A} \) be the set of finite disjoint unions of sets in \( \mathcal{S} \). We show that \( \mathcal{A} \) is closed under finite intersections, then finite unions and relative complements.

For finite intersections, let \( A \) be the disjoint union of \( A_1, \ldots, A_n \in \mathcal{S} \) and \( B \) the disjoint union of \( B_1, \ldots, B_m \in \mathcal{S} \). Then \( A \cap B \) is the disjoint union of

\[
\{ A_j \cap B_k : 1 \leq j \leq n, 1 \leq k \leq m \}.
\]

For finite unions, it suffices to show that if \( A_1, \ldots, A_n, B \in \mathcal{S} \) with \( A_1, \ldots, A_n \) pairwise disjoint, then \( A_1 \cup \cdots \cup A_n \cup B \) can be written as a finite disjoint union of sets in \( \mathcal{S} \). For each \( j, B \setminus A_j \) can be written as a finite disjoint union of sets in \( \mathcal{S} \). Thus by the previous argument

\[
B \setminus (A_1 \cup \cdots \cup A_n) = (B \setminus A_1) \cap \cdots \cap (B \setminus A_n)
\]

can be written as a finite disjoint union of sets in \( \mathcal{S} \); the union of these sets and the \( A_j \) is \( A_1 \cup \cdots \cup A_n \cup B \).

Now let \( A, B \in \mathcal{A} \), with \( A \) the disjoint union of \( A_1, \ldots, A_n \in \mathcal{S} \) and \( B \) the disjoint union of \( B_1, \ldots, B_m \in \mathcal{S} \). We have

\[
A \setminus B = \bigcup_{j=1}^{n} \bigcap_{k=1}^{m} (A_j \setminus B_k)
\]

and each \( A_j \setminus B_k \) is a finite disjoint union of sets in \( \mathcal{S} \), hence in \( \mathcal{A} \). So \( A \setminus B \) is in \( \mathcal{A} \) by the previous arguments. \( \blacksquare \)

XII.1.1.12. Corollary. A semiring which is closed under finite disjoint unions is a ring.
XII.1.1.13. \textbf{Corollary.} Let $S$ be a semiring, $A, A_1, \ldots, A_n$ sets in $S$ with $A_j \subseteq A$ for all $j$ and the $A_j$ pairwise disjoint. Then there are finitely many sets $B_1, \ldots, B_m \in S$ such that $\{A_1, \ldots, A_n, B_1, \ldots, B_m\}$ are pairwise disjoint and
\[
A = \left( \bigcup_{j=1}^{n} A_j \right) \cup \left( \bigcup_{k=1}^{m} B_k \right).
\]

\textbf{Proof:} Let $\mathcal{A}$ be the ring generated by $S$. If $D = \bigcup_{j=1}^{n} A_j$, then $D \in \mathcal{A}$, so $(A \setminus D) \in \mathcal{A}$, i.e. $A \setminus D$ can be written as a finite disjoint union of sets in $S$.

XII.1.1.14. \textbf{Proposition.} Let $S$ be a semiring of subsets of a set $X$, and let $A, A_1, A_2, \ldots$ be sets in $S$. Suppose $A \subseteq \bigcup_{n=1}^{\infty} A_n$. Then there is a sequence $(B_k)$ of pairwise disjoint sets in $S$ with each $B_k$ contained in a unique $A_n$, only finitely many $B_k$ in each $A_n$, such that $A = \bigcup_{k=1}^{m_1+1} B_k$.

\textbf{Proof:} Set $B_1 = A \cap A_1 \in S$. Write $A \setminus A_1 = C_{11} \cup \cdots \cup C_{1m_1}$ with the $C_{1j}$ pairwise disjoint sets in $S$. Set $B_2 = A_2 \cap C_{11}, \ldots, B_{m_1+1} = A_2 \cap C_{1m_1}$. Then the $B_k$ (for $1 \leq k \leq m_1 + 1$) are in $S$ and are pairwise disjoint, $B_1 \subseteq A_1, B_k \subseteq A_2 \setminus A_1$ for $2 \leq k \leq m_1 + 1$, and
\[
\bigcup_{k=1}^{m_1+1} B_k = A \cap (A_1 \cup A_2).
\]

Write $A \setminus A_2 = C_{21} \cup \cdots \cup C_{2m_2}$ with the $C_{2j}$ pairwise disjoint sets in $S$. The next set of $B_k$’s (for $m_1 + 2 \leq k \leq m_1 + m_1 m_2 + 1$) are of the form $A_3 \cap C_{i1} \cap C_{2j}$ for $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$. The new $B_k$ are pairwise disjoint and contained in $A_3 \setminus (A_1 \cup A_2)$, and
\[
\bigcup_{k=1}^{m_1+m_1 m_2+1} B_k = A \cap (A_1 \cup A_2 \cup A_3).
\]

The process can be continued inductively, although the notation becomes increasingly complicated. At the end we have
\[
\bigcup_{k=1}^{\infty} B_k = A \cap \left( \bigcup_{n=1}^{\infty} A_n \right) = A.
\]

XII.1.2. \textbf{Algebras and $\sigma$-Algebras}

For measure theory purposes, the most important set structures are algebras and especially $\sigma$-algebras.

XII.1.3. \textbf{$\pi$-Systems, $\lambda$-Systems, and Monotone Classes}

We will occasionally have use for some collections of sets with other combinations of closure properties, and it will be convenient to have names for them.
XII.1.3.1. **Definition.** Let \( X \) be a set.

A nonempty collection \( Q \) of subsets of \( X \) is a \( \pi \)-system if it is closed under finite intersections.

A collection \( L \) of subsets of \( X \) is a \( \lambda \)-system if it contains \( X \) and is closed under complements and countable disjoint unions.

A nonempty collection \( \mathcal{N} \) of subsets of \( X \) is a monotone class if it is closed under countable monotone unions and intersections: if \((A_n)\) is an increasing sequence in \( \mathcal{N} \), then \( \bigcup_{n=1}^{\infty} A_n \in \mathcal{N} \), and if \((B_n)\) is a decreasing sequence in \( \mathcal{N} \), then \( \bigcap_{n=1}^{\infty} B_n \in \mathcal{N} \).

XII.1.3.2. **Examples.**

(i) A semiring is a \( \pi \)-system. A \( \sigma \)-algebra is both a \( \lambda \)-system and a monotone class. A \( \pi \)-system contains \( \emptyset \), and is thus also closed under finite disjoint unions.

(ii) A \( \pi \)-system need not be a semiring. For example, a topology on \( X \) is a \( \pi \)-system but usually not a semiring.

(iii) A \( \lambda \)-system need not be closed under even finite unions. For example, let \( X \) be a finite set with an even number of elements (> 2), and \( L \) the collection of all subsets with an even number of elements. Then \( L \) is a \( \lambda \)-system, but is not closed under finite unions. \( L \) is also a monotone class.

(iv) A monotone class need not be closed under either finite unions or finite intersections, or even finite disjoint unions. In fact, any finite collection of subsets of a set \( X \) is a monotone class.

(v) Any intersection of \( \pi \)-systems (on a fixed set \( X \)) is a \( \pi \)-system, and similarly for \( \lambda \)-systems and monotone classes. So if \( S \) is any collection of subsets of a set \( X \), there is a smallest \( \pi \)-system containing \( S \), denoted \( \pi(S) \); similarly there is a smallest \( \lambda \)-system \( \lambda(S) \) and a smallest monotone class \( \nu(S) \) containing \( S \). \( \pi(S) \) just consists of all finite intersections of sets in \( S \), but \( \lambda(S) \) and \( \nu(S) \) do not have simple descriptions in general.

XII.1.3.3. **Proposition.**

(i) A \( \lambda \)-system of subsets of a set \( X \) is closed under relative monotone complements: if \( A, B \in \mathcal{L} \) and \( A \subseteq B \), then \( B \setminus A \in \mathcal{L} \).

(ii) Let \( \mathcal{L} \) be a collection of subsets of a set \( X \) containing \( X \) and closed under relative monotone complements and finite and countable disjoint unions. Then \( \mathcal{L} \) is a \( \lambda \)-system.

**Proof:**

(i): \( B \setminus A = [A \cup B^c]^c \), and \( A \cap B^c = \emptyset \).

(ii): If \( A \in \mathcal{L} \), then \( A^c = X \setminus A \in \mathcal{L} \) since \( X \in \mathcal{L} \).

XII.1.3.4. **Proposition.** A collection \( A \) of subsets of a set \( X \) which is both a \( \pi \)-system and a \( \lambda \)-system is a \( \sigma \)-algebra.

**Proof:** \( A \) is closed under complements and finite intersections, hence an algebra. Thus \( A \) is closed under relative complements. If \( (A_n) \) is a sequence of sets in \( A \), let \( B_1 = A_1 \) and \( B_n = A_n \setminus A_{n-1} \) for \( n > 1 \), then \( (B_n) \) is a pairwise disjoint sequence of sets in \( A \), so \( \bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n \in A \), so \( A \) is closed under countable unions.

The following variation of this result will be important:
XII.1.3.5.  **Theorem.**  [Dynkin’s $(\pi-\lambda)$-Theorem] Let $X$ be a set, $Q$ a $\pi$-system in $X$, and $L$ a $\lambda$-system in $X$. If $Q \subseteq L$, then $\sigma(Q) \subseteq L$.

**Proof:** By XII.1.3.4., it suffices to show that $\lambda(Q)$ is a $\pi$-system, since $\lambda(Q)$ is a $\lambda$-system contained in $L$ by minimality.

If $A \subseteq X$, let

$$L_A = \{ B \subseteq X : A \cap B \in \lambda(Q) \}.$$ 

If $A \in \lambda(Q)$, then we show that $L_A$ is a $\lambda$-system. $X$ is trivially in $L_A$. If $B, C \in L_A, C \subseteq B$, then $A \cap B$ and $A \cap C$ are in $\lambda(Q)$, so $A \cap (B \setminus C) = (A \cap C) \setminus (A \cap B) \in \lambda(Q)$ by XII.1.3.3. (i), since $\lambda(Q)$ is a $\lambda$-system. Thus $B \setminus C \in L_A$ and $L_A$ is closed under relative monotone complements. If $(B_n)$ is a pairwise disjoint sequence in $L_A$, then $(A \cap B_n)$ is a pairwise disjoint sequence in $\lambda(Q)$, hence $A \cap \bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} (A \cap B_n) \in \lambda(Q)$, i.e. $\bigcup_{n=1}^{\infty} B_n \in L_A$, and $L_A$ is closed under countable disjoint unions. So $L_A$ is a $\lambda$-system if $A \in \lambda(Q)$, and in particular if $A \in Q$.

Now suppose $A \in Q$. If $B \in Q$, then $A \cap B \in Q$ since $Q$ is a $\pi$-system, and thus $B \in L_A$. So if $A \in Q$, then $Q \subseteq L_A$, and thus $\lambda(Q) \subseteq L_A$ by minimality of $\lambda(Q)$.

Note that if $A, B \subseteq X$, then $B \in L_A$ if and only if $A \in L_B$. If $A \in Q$ and $B \in \lambda(Q)$, then $B \in L_A$ since $\lambda(Q) \subseteq L_A$, and thus $A \in L_B$. Thus, for $B \in \lambda(Q), Q \subseteq L_B$, and hence $\lambda(Q) \subseteq L_B$ by minimality of $\lambda(Q)$. If $B, C \in \lambda(Q)$, we have $C \in L_B$, i.e. $B \cap C \in \lambda(Q)$, so $\lambda(Q)$ is a $\pi$-system.  

We have similar results for monotone classes, due to Halmos, which will be useful:

XII.1.3.6.  **Proposition.**  An algebra of subsets of a set $X$ which is also a monotone class is a $\sigma$-algebra.

**Proof:** Let $A$ be an algebra of subsets of $X$, and suppose $A$ is also a monotone class. Let $(A_n)$ be a sequence of sets in $A$. For each $n$, set $B_n = \bigcap_{k=1}^{n} A_k$. Then $(B_n)$ is a decreasing sequence of sets in $A$ since $A$ is an algebra, and hence $\bigcap_{n=1}^{\infty} B_n = \bigcap_{n=1}^{\infty} A_n \in A$ since $A$ is a monotone class.  

XII.1.3.7.  **Theorem.**  [Monotone Class Theorem] Let $X$ be a set, $A$ an algebra of subsets of a set $X$, and $\mathcal{N}$ a monotone class of subsets of $X$. If $A \subseteq \mathcal{N}$, then $\sigma(A) \subseteq \mathcal{N}$.

XII.1.4.  **Measurable Spaces and Measurable Functions**

XII.1.4.1.  **Definition.**  A measurable space $(X, A)$ is a set $X$ with a specified $\sigma$-algebra $A$ of subsets.

Measurable spaces are the natural setting for measure theory, which will be developed in (). Note the analogy with the definition of a topological space. Measurable spaces are often called Borel spaces, especially (but not necessarily) if they come from a topological space as in ().

As usual with mathematical theories, it is important not only to have sets with a structure, but to have a class of functions which preserve the structure to serve as the morphisms of the theory. Experience shows that the appropriate morphisms between measurable spaces are the measurable functions:

1360
XII.1.4.2. **Definition.** Let \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) be measurable spaces, and \(f : X \to Y\) a function. Then \(f\) is \((\mathcal{A}, \mathcal{B})\)-measurable (or just measurable if the \(\mathcal{A}\) and \(\mathcal{B}\) are understood) if \(f^{-1}(B) \in \mathcal{A}\) for every \(B \in \mathcal{B}\).

Note the close analogy with the definition of a continuous function between topological spaces \((\text{topo})\) (although there is no useful notion of “measurability at a point.”) When measurable spaces are called Borel spaces, measurable functions between them are often called Borel functions, although this terminology can sometimes be slightly ambiguous.

XII.1.4.3. **Proposition.** Let \((X, \mathcal{A}), (Y, \mathcal{B})\), and \((Z, \mathcal{C})\) be measurable spaces, and \(f : X \to Y\) and \(g : Y \to Z\) functions. If \(f\) is \((\mathcal{A}, \mathcal{B})\)-measurable and \(g\) is \((\mathcal{B}, \mathcal{C})\)-measurable, then \(g \circ f\) is \((\mathcal{A}, \mathcal{C})\)-measurable.

The proof is virtually identical to the proof that a composition of continuous functions is continuous.

XII.1.4.4. Thus the measurable spaces and measurable functions form a category. An isomorphism in this category, i.e. a bijection \(f\) such that both \(f\) and \(f^{-1}\) are measurable, is often called a Borel isomorphism.

XII.1.4.5. Just as with continuity, the larger \(\mathcal{A}\) is, or the smaller \(\mathcal{B}\) is, the easier it is for a function to be measurable. Given \(\mathcal{A}\), there is a largest \(\mathcal{B}\) making \(f\) measurable, and similarly, given \(\mathcal{B}\), there is a smallest corresponding \(\mathcal{A}\). These are given by the following construction:

XII.1.4.6. **Definition.** Let \(X\) and \(Y\) be sets, and \(f : X \to Y\) a function. If \(\mathcal{A}\) is a \(\sigma\)-algebra on \(X\), set
\[
\mathcal{f}^{-}(\mathcal{A}) = \{ B \subseteq Y : f^{-1}(B) \in \mathcal{A} \}.
\]
If \(\mathcal{B}\) is a \(\sigma\)-algebra on \(Y\), set
\[
\mathcal{f}^{+}(\mathcal{B}) = \{ f^{-1}(B) : B \in \mathcal{B} \}.
\]

XII.1.4.7. **Proposition.** Let \(X\) and \(Y\) are measurable spaces and \(f : X \to Y\) a function.

(i) If \(\mathcal{A}\) is a \(\sigma\)-algebra on \(X\), then \(\mathcal{f}^{-}(\mathcal{A})\) is a \(\sigma\)-algebra on \(Y\).

(ii) If \(\mathcal{B}\) is a \(\sigma\)-algebra on \(Y\), then \(\mathcal{f}^{+}(\mathcal{B})\) is a \(\sigma\)-algebra on \(X\).

**Proof:** Both parts follow easily from the fact that \(f^{-1}(\emptyset) = \emptyset\), \(f^{-1}(Y) = X\), \(f^{-1}(B^c) = [f^{-1}(B)]^c\), and \(f^{-1}(\bigcup_{n=1}^{\infty} B_n) = \bigcup_{n=1}^{\infty} f^{-1}(B_n)\) for any \(B, B_n \subseteq Y\). \(\square\)

XII.1.4.8. In fact, if \((X, \mathcal{A})\) is a measurable space and \(Y\) is a set, and \(f : X \to Y\) a function, then \(\mathcal{f}^{-}(\mathcal{A})\) is the largest \(\sigma\)-algebra on \(Y\) making \(f\) measurable. Similarly, if \(f : X \to (Y, \mathcal{B})\), then \(\mathcal{f}^{+}(\mathcal{B})\) is the smallest \(\sigma\)-algebra on \(X\) making \(f\) measurable.

Note, however, that \(\mathcal{f}^{-}(\mathcal{A})\) and \(\mathcal{f}^{+}(\mathcal{B})\) can be quite degenerate: for example, if \(f\) is a constant function, then \(\mathcal{f}^{-}(\mathcal{A}) = \mathcal{P}(Y)\) and \(\mathcal{f}^{+}(\mathcal{B}) = \{\emptyset, X\}\). There are examples where \(\mathcal{f}^{-}(\mathcal{A}) = \{\emptyset, Y\}\) even for nondegenerate \(f\) and \(\mathcal{A}\); \(\mathcal{f}^{+}(\mathcal{B})\) is less degenerate in general.

Although it would appear that \(\mathcal{f}^{-}\) and \(\mathcal{f}^{+}\) are “inverse” to each other, this is not the case in general: we always have that \(\mathcal{f}^{+}(\mathcal{f}^{-}(\mathcal{A})) \subseteq \mathcal{A}\) and \(\mathcal{f}^{-}(\mathcal{f}^{+}(\mathcal{B})) \supseteq \mathcal{B}\), but equality does not hold in general.\(\square\)

The next theorem is one of the most useful facts about mappings of \(\sigma\)-algebras.
XII.1.4.9. Theorem. Let $X$ and $Y$ be sets, $f : X \to Y$ a function, and $\mathcal{B}$ a $\sigma$-algebra on $Y$. If $\mathcal{C} \subseteq \mathcal{B}$ is a subset which generates $\mathcal{B}$ as a $\sigma$-algebra, then $f^{-1}(\mathcal{C}) = \{ f^{-1}(C) : C \in \mathcal{C} \}$ generates $f^{-1}(\mathcal{B})$ as a $\sigma$-algebra.

Proof: Let $\mathcal{A}$ be the $\sigma$-algebra on $X$ generated by $f^{-1}(\mathcal{C})$. Then $\mathcal{A} \subseteq f^{-1}(\mathcal{B})$ since by XII.1.4.7, $f^{-1}(\mathcal{B})$ is a $\sigma$-algebra containing $f^{-1}(\mathcal{C})$. Let

$$B_0 = f^{-1}(A) \cap B = \{ B \in \mathcal{B} : f^{-1}(B) \in \mathcal{A} \}.$$  

By XII.1.4.7, $B_0$ is the intersection of two $\sigma$-algebras, hence is a $\sigma$-algebra; and $B_0 \subseteq \mathcal{B}$. Also, $\mathcal{C} \subseteq B_0$. But $\mathcal{C}$ generates $\mathcal{B}$ as a $\sigma$-algebra, so $B_0 = \mathcal{B}$, and hence $f^{-1}(B) \in \mathcal{A}$ for any $B \in \mathcal{B}$, i.e. $f^{-1}(\mathcal{B}) \subseteq \mathcal{A}$, so $\mathcal{A} = f^{-1}(\mathcal{B})$.

XII.1.4.10. Corollary. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $f : X \to Y$ a function, and $\mathcal{C}$ a subset of $\mathcal{B}$ which generates $\mathcal{B}$ as a $\sigma$-algebra. Then $f$ is measurable if and only if $f^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$.

Proof: It is obvious that if $f$ is measurable, then $f^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$. The converse follows immediately from XII.1.4.9., since by assumption the $\sigma$-algebra on $X$ generated by $f^{-1}(\mathcal{C})$, which by the theorem is $f^{-1}(\mathcal{B})$, is contained in $\mathcal{A}$.

The next simple observation is extraordinarily useful in defining measurable functions piecewise:

XII.1.4.11. Proposition. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let $(E_n)$ be a sequence in $\mathcal{A}$ with $\bigcup_{n=1}^{\infty} E_n = X$. (Since $\emptyset \in \mathcal{A}$, $(E_n)$ could also be a finite sequence.) Let $f : X \to Y$ be a function, and set $f_n = f|_{E_n} : E_n \to Y$. If each $f_n$ is measurable, then $f$ is measurable.

Proof: Let $B \in \mathcal{B}$. We need to show that $f^{-1}(B) \in \mathcal{A}$. By assumption, $f_n^{-1}(B) = \{ x \in E_n : f_n(x) = f(x) \in B \}$ is in $\mathcal{A}$. But $f^{-1}(B) = \bigcup_{n=1}^{\infty} f_n^{-1}(B)$.

A slight rephrasing is usually used in piecewise definition of functions:

XII.1.4.12. Corollary. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let $(E_n)$ be a sequence in $\mathcal{A}$ which partitions $X$, i.e. the $E_n$ are pairwise disjoint and $\bigcup_{n=1}^{\infty} E_n = X$. (Since $\emptyset \in \mathcal{A}$, $(E_n)$ could also be a finite sequence.) For each $n$, let $f_n : E_n \to Y$ be a function, and define $f : X \to Y$ by $f(x) = f_n(x)$ if $x \in E_n$. If each $f_n$ is measurable, then $f$ is measurable.

XII.1.4.13. We can also rephrase the last two results in another way. If $(X, \mathcal{A})$ is a measurable space, and $E \in \mathcal{A}$, the restriction of $\mathcal{A}$ to $E$ is

$$\mathcal{A}_E = \{ A \cap E : A \in \mathcal{A} \} = \{ B : B \in \mathcal{A}, B \subseteq E \}.$$  

which can be easily verified to be a $\sigma$-algebra of subsets of $E$. The measurable space $(E, \mathcal{A}_E)$ is called a measurable subspace, or Borel subspace, of $(X, \mathcal{A})$.  

1362
XII.1.4.14. **Definition.** (i) If \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) are measurable spaces, the *sum* or *disjoint union* is the measurable space \((Z, \mathcal{C})\), where \(Z\) is the disjoint union \(X \sqcup Y \) and
\[
\mathcal{C} = \{ A \sqcup B : A \in \mathcal{A}, B \in \mathcal{B} \}
\]
(it is easy to check that \(\mathcal{C}\) is a \(\sigma\)-algebra of subsets of \(X \sqcup Y\)). The sum is denoted \((X, \mathcal{A}) \oplus (Y, \mathcal{B})\).

(ii) Let \((X_n, \mathcal{A}_n)\) be a finite or countable set of measurable spaces. The *sum* or *disjoint union*, denoted \(\bigoplus_n (X_n, \mathcal{A}_n)\), is the measurable space \((Z, \mathcal{C})\), where \(Z = \sqcup_n X_n\) and
\[
\mathcal{C} = \{ \sqcup_n A_n : A_n \in \mathcal{A}_n \text{ for all } n \}
\]
(it is again easy to check that \(\mathcal{C}\) is a \(\sigma\)-algebra of subsets of \(Z\) (Exercise ()).)

Then XII.1.4.11. can be rephrased:

XII.1.5. **Proposition.** Let \((X, \mathcal{A})\) be a measurable space, and let \((E_n)\) be a sequence in \(\mathcal{A}\) which partitions \(X\), i.e. the \(E_n\) are pairwise disjoint and \(\sqcup_{n=1}^\infty E_n = X\). (Since \(\emptyset \in \mathcal{A}\), \((E_n)\) could also be a finite sequence.) Let \(f_n\) be the inclusion map of \(E_n\) into \(X\), and \(f : \sqcup_n E_n \to X\) the induced map. Then \(f\) is a Borel isomorphism from \(\bigoplus_n (E_n, \mathcal{A}_{E_n})\) onto \((X, \mathcal{A})\). In particular, if \(E \in \mathcal{A}\), then \((X, \mathcal{A})\) is Borel isomorphic to \((E, \mathcal{A}_E) \oplus (X \setminus E, \mathcal{A}_{X \setminus E})\).

XII.1.5. **Exercises**

XII.1.5.1. This exercise should convince the reader of (1) the truth of the statement of XII.1.1.5. and (2) the author’s wisdom in not attempting to write out a full proof. Adopt the situation and notation of XII.1.1.5..

(a) Verify the four identities in the proof outline.

(b) Let \(A_i, B_i, C_i, D_i\) \((i = 1, 2)\) be sets in \(\mathcal{C}\). Show that
\[
[(A_1 \setminus B_1) \cap (A_2 \setminus B_2)] \setminus [(C_1 \setminus D_1) \cap (C_2 \setminus D_2)]
\]
is a finite disjoint union of finite intersections of sets in \(\mathcal{C} \cup \mathcal{D}\), using the four identities. [The first step is to use the fourth identity to rewrite the set as
\[
[(A_1 \setminus B_1) \cap (A_2 \setminus B_2)] \setminus (C_1 \setminus D_1) \cup \{(A_1 \setminus B_1) \cap (A_2 \setminus B_2) \cap (C_1 \setminus D_1)\}\setminus (C_2 \setminus D_2)
\]
where \(\sqcup\) denotes disjoint union.]

XII.1.5.2. Let \(\mathcal{E}\) be a collection of subsets of a set \(X\). Set \(\tilde{\mathcal{E}} = \{ X \setminus E : E \in \mathcal{E}\}\).

(a) Show that if \(\mathcal{E}\) is closed under finite unions and finite intersections, and \(\mathcal{E} \cap \tilde{\mathcal{E}}\) is nonempty, then \(\mathcal{E} \cap \tilde{\mathcal{E}}\) is an algebra of subsets of \(X\) and is the largest algebra contained in \(\mathcal{E}\).

(b) Show that if \(\mathcal{E}\) is closed under countable unions and countable intersections, and \(\mathcal{E} \cap \tilde{\mathcal{E}}\) is nonempty, then \(\mathcal{E} \cap \tilde{\mathcal{E}}\) is a \(\sigma\)-algebra of subsets of \(X\) and is the largest \(\sigma\)-algebra contained in \(\mathcal{E}\).

(c) Even if \(\mathcal{E}\) is infinite and closed under countable unions and countable intersections, \(\mathcal{E} \cap \tilde{\mathcal{E}}\) can be empty. [Let \(\mathcal{E}\) be the collection of countable subsets of an uncountable set.]
XII.1.5.3. Prove the Schröder-Bernstein Theorem for measurable spaces: if \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) are measurable spaces, \((X, \mathcal{A})\) is Borel isomorphic to a measurable subspace of \((Y, \mathcal{B})\), and \((Y, \mathcal{B})\) is Borel isomorphic to a measurable subspace of \((X, \mathcal{A})\), then \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) are Borel isomorphic. [Mimic the proof of the Schröder-Bernstein Theorem of set theory (II.5.6.11.).]
XII.1.6. Filters and Ultrafilters

Filters are another important type of subspace structure. Filters are useful in a wide variety of mathematical applications, especially in topology.

XII.1.6.1. Definition. Let $X$ be a set. A filter on $X$ is a nonempty collection $\mathcal{F}$ of subsets of $X$ such that

(i) $\mathcal{F}$ is closed under finite intersections.

(ii) $\mathcal{F}$ has the superset property: if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $B \in \mathcal{F}$.

(iii) $\emptyset \notin \mathcal{F}$.

A filter on $X$ which is not properly contained in any other filter on $X$ is called an ultrafilter on $X$.

XII.1.6.2. It follows from the superset property that a filter $\mathcal{F}$ is closed under arbitrary unions, and, since $\mathcal{F}$ is nonempty, that $X \in \mathcal{F}$. Thus a filter is somewhat similar to a topology. But property (iii) gives a crucial difference between a filter and a topology (a topology also does not have the superset property in general).

In fact, if $\mathcal{F}$ is a filter on $X$ and $A \subseteq X$, we cannot have both $A$ and $A^c$ in $\mathcal{F}$, since then we would have $A \cap A^c = \emptyset \notin \mathcal{F}$. It can happen, though, that neither $A$ nor $A^c$ is in $\mathcal{F}$. (This cannot happen in an ultrafilter, and in fact gives a characterization of ultrafilters (XII.1.6.12.).)

XII.1.6.3. A useful point of view, which is helpful in determining abstract properties of filters and appropriate for some, but not all, applications, is to regard a filter (especially an ultrafilter) as giving a notion of “size” for subsets of $X$: sets in the filter are regarded as “large” subsets and those not in the filter as “small” subsets.

XII.1.6.4. As with many other subset structures, the intersection of any collection $\{\mathcal{F}_i : i \in I\}$ of filters on $X$ is a filter on $X$ (note that the intersection is a nonempty collection of subsets since $X \in \mathcal{F}_i$ for all $i$). However, the union of two filters is not a filter in general (in fact, see XII.1.6.8.); but the union of a chain (under inclusion) of filters is a filter (Exercise (i)).

XII.1.6.5. If $\mathcal{F}$ is a filter on $X$, the set $\mathcal{I} = \{A^c : A \in \mathcal{F}\}$ is closed under finite unions and subsets, and does not contain the whole set $X$. Such a collection is called an order ideal. Order ideals and hence filters can be defined in general partially ordered sets (i).

XII.1.6.6. Examples. (i) Let $A$ be a fixed nonempty subset of $X$, and let

$$\mathcal{F}_A = \{B \subseteq X : A \subseteq B\}.$$ 

Then $\mathcal{F}_A$ is a filter on $X$, which is an ultrafilter if and only if $A$ is a singleton (note that $\mathcal{F}_A \subseteq \mathcal{F}_B$ if and only if $A \subseteq B$). If $x \in X$, write $\mathcal{F}_x$ to mean $\mathcal{F}_{\{x\}}$. Thus $\mathcal{F}_x = \{B \subseteq X : x \in B\}$.

(ii) Let $X$ be an infinite set, and $\mathcal{F}$ the collection of all subsets of $X$ with finite complement. Then $\mathcal{F}$ is a filter on $X$. This filter is never an ultrafilter (cf. XII.1.6.12.). More generally, if $\kappa$ is an infinite cardinal less
than or equal to $\text{card}(X)$, the set $\mathcal{F}$ of subsets $A$ of $X$ for which $\text{card}(A^c) < \kappa$ is a filter on $X$; in particular, if $X$ is uncountable, the set of cocountable subsets of $X$ is a filter on $X$.

(iii) Let $(X, \mathcal{T})$ be a topological space, and $x \in X$. Let $\mathcal{O}_x$ be the collection of all $\mathcal{T}$-neighborhoods () of $x$. Then $\mathcal{O}_x$ is a filter on $X$ which is contained in $\mathcal{F}_x$; they are not the same in general since $\{x\} \in \mathcal{O}_x$ if and only if $x$ is an isolated point in $X$ (i.e. $\{x\} \in \mathcal{T}$). Thus $\mathcal{O}_x$ is usually not an ultrafilter on $X$. More generally, $x$ can be replaced by any nonempty subset of $X$. (The notation $\mathcal{O}_x$ should really include the topology $\mathcal{T}$ since this filter depends on $\mathcal{T}$ as well as $x$.)

The intersection of all the sets in a filter may be empty (XII.1.6.6.(ii)) or nonempty (XII.1.6.6.(i)). We make the following definition:

XII.1.6.7. Definition. Let $\mathcal{F}$ be a filter on a set $X$. $\mathcal{F}$ is a fixed filter with fixed set $A$ if the intersection of all the sets in $\mathcal{F}$ is the nonempty set $A$. If the intersection of all sets in $\mathcal{F}$ is empty, $\mathcal{F}$ is a free filter on $X$.

An ultrafilter on a set $X$ is either of the form $\mathcal{F}_x$ (XII.1.6.6.(i)) for some $x \in X$, or is free. The ultrafilter $\mathcal{F}_x$ is often called the principal ultrafilter at the point $x$.

XII.1.6.8. If $\mathcal{F}$ and $\mathcal{G}$ are filters on $X$, then there may not be any filter on $X$ at all containing $\mathcal{F} \cup \mathcal{G}$; this will fail if there is a subset $A$ of $X$ with $A \in \mathcal{F}$ and $A^c \in \mathcal{G}$ (XII.1.6.2.), e.g. if $\mathcal{F}$ and $\mathcal{G}$ are $\mathcal{F}_x$ and $\mathcal{F}_y$ as in XII.1.6.6.(i) for distinct $x$ and $y$. If there is such a set $A$, $\mathcal{F}$ and $\mathcal{G}$ are said to be separated (). Thus an arbitrary collection of nonempty subsets of $X$ need not be contained in a filter on $X$. We will characterize when a collection of subsets is contained in a filter.

XII.1.6.9. Definition. A collection $\mathcal{S}$ of subsets of a set $X$ has the finite intersection property if the intersection of any finite number of sets in $\mathcal{S}$ is nonempty.

Any subset of a filter on a set $X$ has the finite intersection property. The converse is also true:

XII.1.6.10. Proposition. Let $\mathcal{S}$ be a collection of subsets of a set $X$, with the finite intersection property. Then there is a filter on $X$ containing $\mathcal{S}$.

Proof: Let $\mathcal{F}$ be the collection of all subsets of $X$ which contain a finite intersection of sets in $\mathcal{S}$. It is easy to verify that $\mathcal{F}$ is a filter containing $\mathcal{S}$, and is the smallest such filter. ☐

We now aim towards the question of whether every filter is contained in an ultrafilter, and along the way we will obtain an important explicit characterization of ultrafilters. We first obtain a fairly simple, but important, one-step extension theorem.

XII.1.6.11. Theorem. [Filter Extension Theorem] Let $\mathcal{F}$ be a filter on a set $X$, and $A \subseteq X$. If $A^c \notin \mathcal{F}$, then there is a filter $\mathcal{G}$ on $X$ containing $\mathcal{F}$ and $A$, i.e. with $\mathcal{F} \subseteq \mathcal{G}$ and $A \in \mathcal{G}$.

Note that the condition $A^c \notin \mathcal{F}$ is necessary for the existence of $\mathcal{G}$, since otherwise both $A$ and $A^c$ would be in $\mathcal{G}$.
Proof: Let
\[ G = \{ C \subseteq X : C \supseteq (B \cap A) \text{ for some } B \in \mathcal{F} \} . \]
If \( B \in \mathcal{F} \), then \( B \supseteq (B \cap A) \), so \( B \in \mathcal{G} \), and thus \( \mathcal{F} \subseteq \mathcal{G} \). We also have \( A \in \mathcal{G} \) since \( A \supseteq (X \cap A) \).

We show that \( \mathcal{G} \) is a filter on \( X \). \( \mathcal{G} \) obviously has the superset property, and is nonempty since \( \mathcal{F} \subseteq \mathcal{G} \).

If \( C_1, C_2 \in \mathcal{G} \), with \( C_1 \supseteq (B_1 \cap A) \) and \( C_2 \supseteq (B_2 \cap A) \) for \( B_1, B_2 \in \mathcal{F} \), then
\[
C_1 \cap C_2 \supseteq (B_1 \cap A) \cap (B_2 \cap A) = (B_1 \cap B_2) \cap A
\]
and \( (B_1 \cap B_2) \in \mathcal{F} \), so \( (C_1 \cap C_2) \in \mathcal{G} \). Hence \( \mathcal{G} \) is closed under finite intersections. Finally, we show \( \emptyset \notin \mathcal{G} \).

If \( \emptyset \in \mathcal{G} \), then \( \emptyset \supseteq (B \cap A) \) for some \( B \in \mathcal{F} \); for this \( B \), we have \( B \cap A = \emptyset \), so \( B \subseteq A^c \) and \( A^c \in \mathcal{F} \) by the superset property, contrary to hypothesis. Thus \( \mathcal{G} \) is the desired filter.

As an immediate consequence, we get a characterization of ultrafilters:

XII.1.6.12. Theorem. [Ultrafilter Characterization Theorem] Let \( \mathcal{F} \) be a filter on a set \( X \). Then \( \mathcal{F} \) is an ultrafilter if and only if, for every subset \( A \) of \( \mathcal{F} \), exactly one of \( A \) and \( A^c \) is in \( \mathcal{F} \).

Proof: We cannot have both \( A \) and \( A^c \) in \( \mathcal{F} \) (XII.1.6.2.). If neither \( A \) nor \( A^c \) is in \( \mathcal{F} \) for some subset \( A \) of \( X \), then the filter \( \mathcal{G} \) of XII.1.6.11. properly contains \( \mathcal{F} \), so \( \mathcal{F} \) is not an ultrafilter. Conversely, if \( \mathcal{F} \) is not an ultrafilter, and \( \mathcal{G} \) is a filter on \( X \) properly containing \( \mathcal{F} \), let \( A \in \mathcal{G} \setminus \mathcal{F} \). Then \( A^c \notin \mathcal{G} \) by XII.1.6.2., so neither \( A \) nor \( A^c \) is in \( \mathcal{F} \).

Here is another useful observation which is a simple corollary:

XII.1.6.13. Corollary. Let \( \mathcal{F} \) be a filter on a set \( X \). Then \( \mathcal{F} \) is an ultrafilter if and only if, whenever \( A, B \subseteq X \) and \( A \cup B \in \mathcal{F} \), we have either \( A \in \mathcal{F} \) or \( B \in \mathcal{F} \).

Proof: A filter with this property is an ultrafilter by setting \( B = A^c \). Conversely, suppose \( \mathcal{F} \) is an ultrafilter and \( A \cup B \in \mathcal{F} \). If \( A \notin \mathcal{F} \) and \( B \notin \mathcal{F} \), we have \( A^c \in \mathcal{F} \) and \( B^c \in \mathcal{F} \), so
\[
A^c \cap B^c = (A \cup B)^c \in \mathcal{F}
\]
which contradicts XII.1.6.2..
XII.1.6.15. Theorem. (AC) Every filter on a set $X$ is contained in an ultrafilter on $X$.

The proof of this theorem is a standard Zorn’s Lemma argument, and is left to the reader (Exercise ()).

XII.1.6.16. This theorem cannot be proved without some form of the Axiom of Choice. But the full AC is not needed: the statement of this theorem is equivalent to the Boolean Prime Ideal Theorem (II.6.3.11.), which is strictly weaker than AC.

We cannot even prove without some form of AC that there is a free ultrafilter on $\mathbb{N}$ (this statement actually follows from the Measure Extension Theorem (), which is strictly weaker than the Boolean Prime Ideal Theorem). With AC (or BPI), this is an immediate application of XII.1.6.15. to the filter of cofinite subsets of $\mathbb{N}$ (XII.1.6.6.(ii)).

It can be shown even without AC that many sets have a large number of pairwise separated filters. To give the best statement possible without AC, we first discuss two key property of at least some cardinals:

XII.1.6.17. If $\kappa$ is a cardinal, and $X$ is a set of cardinality $\kappa$, then we denote by $\kappa^2$ the cardinality of $X \times X$. We always have $\kappa^2 \geq \kappa$. If $1 < \kappa < \aleph_0$, then $\kappa^2 > \kappa$. If $\kappa$ is infinite, then it is conceivable that $\kappa^2$ always equals $\kappa$; this can be proved using AC (and turns out to be equivalent to AC). Without AC the following can be proved:

(i) $\aleph_0^2 = \aleph_0$ ()

(ii) More generally, if $\aleph$ is any aleph (), then $\aleph^2 = \aleph$.

(iii) $(2^{\aleph_0})^2 = 2^{\aleph_0}$ ()

XII.1.6.18. We now turn to the other cardinal property. Let $\kappa$ be a cardinal, and $X$ a set of cardinality $\kappa$. Let $\Phi$ be the set of all finite subsets of $X$. We have $\text{card}(\Phi) \geq \text{card}(X)$. If $\kappa$ is finite, the inequality is strict. For infinite sets, it is conceivable that equality always holds (it holds for $\aleph_0$ by ()); if $\text{card}(\Phi) = \text{card}(X)$, we say $\kappa$ has the finite subset property.

If $\kappa^2 = \kappa \geq \aleph_0$ and $X$ can be totally ordered, it is not difficult to show that $\kappa$ has the finite subset property (II.5.6.12.). Thus $\aleph_0$ (in fact, any $\aleph$) and $2^{\aleph_0}$ have the finite subset property. So it follows from AC that every infinite cardinal has the finite subset property. It cannot be proved in ZF that every infinite cardinal has the finite subset property (), i.e. some type of Choice is necessary.

XII.1.6.19. Theorem. ([?]; cf. [?], [GJ76, 9.2]) Let $\kappa$ be a (necessarily infinite) cardinal with $\kappa^2 = \kappa$ and the finite subset property, and $X$ a set of cardinality $\kappa$. Then there are $2^{2^\kappa}$ pairwise separated filters on $X$. In particular, there are $2^{2^{\aleph_0}}$ pairwise separated filters on $\mathbb{N}$.

Proof: Let $\Phi$ be the set of all finite subsets of $X$, and $\Lambda$ the set of all finite subsets of $\Phi$. Set $\Omega = \Phi \times \Lambda$. Then from $\kappa^2 = \kappa$ and the finite subset property it follows that $\text{card}(\Phi) = \text{card}(\Lambda) = \text{card}(\Omega) = \kappa$.

We will define $2^{2^\kappa}$ pairwise separated filters on $\Omega$, which will suffice since $\text{card}(\Omega) = \text{card}(X)$. First some notation: for $S \subseteq X$, define a subset $\mathcal{B}_S$ of $\Omega$ as follows:

$$\mathcal{B}_S = \{(F, \mathcal{L}) \in \Phi \times \Lambda : S \cap F \in \mathcal{L}\}.$$
Write $C_S$ for the complement of $B_S$ in $\Omega$. Then, for $S \subseteq \mathcal{P}(X)$, define

$$\mathcal{B}_S = \{B_S : S \in S\} \cup \{C_S : S \notin S\} \subseteq \mathcal{P}(\Omega).$$

We claim that $\mathcal{B}_S$ has the finite intersection property (XII.1.6.9.). If $\mathfrak{A}$ is a finite subset of $\mathcal{B}_S$, the distinct sets in $\mathfrak{A}$ can be numbered as follows:

$$B_{S_1}, \ldots, B_{S_k}, C_{S_{k+1}}, \ldots, C_{S_n}$$

where $S_1, \ldots, S_n$ are distinct subsets of $X$. For $i \neq j$ let $x_{ij}$ be an element of $X$ which is in one of $S_i$ and $S_j$ but not the other (note that there are only finitely many choices to be made here, so no form of AC is needed). Let $F$ be the set of all the $x_{ij}$ for $i \neq j$. Then the sets

$$S_i \cap F, \ldots, S_n \cap F$$

are all distinct since if $i \neq j$, $x_{ij}$ is in one of $S_i \cap F$ and $S_j \cap F$ but not the other. Set

$$\mathcal{L} = \{S_1 \cap F, \ldots, S_k \cap F\} \in \mathfrak{A}.$$  

We have that $S_i \cap F \in \mathcal{L}$ for $1 \leq i \leq k$, and $S_j \cap K \notin \mathcal{L}$ for $k + 1 \leq j \leq n$. So $(F, \mathcal{L}) \in \mathcal{B}_S$, for $1 \leq i \leq k$ and $(F, \mathcal{L}) \in \mathcal{C}_S$, for $k + 1 \leq j \leq n$, and hence the intersection of the sets in $\mathfrak{A}$ is nonempty, proving the claim.

Thus, for each $S \subseteq \mathcal{P}(X)$, there is a filter $\mathfrak{F}_S$ on $\Omega$ containing $\mathcal{B}_S$ by XII.1.6.10. If $S_1$ and $S_2$ are distinct subsets of $\mathcal{P}(X)$, let $S$ be a subset of $X$ which is in one of them (say $S_1$) but not the other. Then $B_S \in \mathfrak{F}_{S_1}$ and $C_S \in \mathfrak{F}_{S_2}$, so $\mathfrak{F}_{S_1}$ and $\mathfrak{F}_{S_2}$ are separated. The number of distinct filters constructed is $\text{card}(\mathcal{P}(\mathcal{P}(X))) = 2^{2^n}$.\[\text{\blacksquare}\]

Note that AC is not used in this result. Using AC, the conclusion holds for all infinite cardinals. The full AC is needed here, not just the BPI (since we need $\kappa^2 = \kappa$ for all $\kappa$). Combining this result with XII.1.6.15., we obtain:

**XII.1.6.20.** Theorem. (AC) Let $X$ be an infinite set, of cardinality $\kappa$. Then there are exactly $2^{2^n}$ free ultrafilters on $X$.

**Proof:** By XII.1.6.15., each $\mathfrak{F}_S$ from the proof of XII.1.6.19. is contained in an ultrafilter. These ultrafilters are all distinct since the $\mathfrak{F}_S$ are pairwise separated. Thus there are at least $2^{2^n}$ ultrafilters on $X$. On the other hand, there are at most $2^{2^n}$ filters on $X$ since there are only $2^{2^n}$ sets of subsets of $X$. Thus there are exactly $2^{2^n}$ ultrafilters on $X$. Only $\kappa$ of these are fixed, so there are $2^{2^n}$ free ones.\[\text{\blacksquare}\]

In particular, there are $2^{2^{\aleph_0}}$ distinct free ultrafilters on $\mathbb{N}$ (assuming AC, or at least BPI).

**XII.1.7.** Exercises

**XII.1.7.1.** Verify the assertions in the examples of XII.1.6.6.

**XII.1.7.2.** Verify the assertion in XII.1.6.7. that every ultrafilter which is not free is of the form $\mathcal{F}_x$ for some $x$.  

1369
XII.1.7.3. (a) Verify that the set $F$ defined in the proof of XII.1.6.10. is a filter.
(b) Show that two filters $F$ and $G$ on a set $X$ are separated if and only if $F \cup G$ fails to have the finite intersection property. Thus if $F$ and $G$ are not separated, they are contained in a common larger filter on $X$.

XII.1.7.4. (a) Show that if $\{F_i : i \in I\}$ is a chain (under inclusion) of filters on a set $X$, i.e. $F_i \subseteq F_j$ or $F_j \subseteq F_i$ for all $i, j \in I$, then $\cup_{i \in I} F_i$ is a filter on $X$.
(b) Let $F$ be a filter on a set $X$, and let $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{P}(X))$ be the set of all filters on $X$ containing $F$. Partially order $\mathcal{Y}$ by inclusion. Show that every chain in $\mathcal{Y}$ has an upper bound in $\mathcal{Y}$.
(c) Conclude from Zorn’s Lemma that $\mathcal{Y}$ contains a maximal element $G$. Show that $G$ is an ultrafilter on $X$.

XII.1.7.5. Let $(\mathcal{X}, \leq)$ be a partially ordered set. Define an order ideal in $(\mathcal{X}, \leq)$ to be a subset $I$ which is hereditary, i.e. if $y \in I$, $x \in \mathcal{X}$, and $x \leq y$, then $x \in I$, and such that any two elements (hence any finite set of elements) of $I$ have an upper bound in $I$. Define a filter on $(\mathcal{X}, \leq)$ to be the complement of a nonempty proper order ideal.
(a) Let $X$ be a set. Show that order ideals and filters in $(\mathcal{P}(X), \subseteq)$ under this definition are exactly the order ideals and filters under the definitions of this section.
(b) Extend the results of this section as far as possible to filters in general partially ordered sets.
XII.2. Borel Sets and Borel Measurable Functions

For measure theory purposes, the Borel sets are the most important subsets of $\mathbb{R}$ or $\mathbb{R}^n$. The class of Borel sets is large enough to contain every subset of $\mathbb{R}$ which can be described “naturally”, and small enough that measure-theoretic operations behave well on them. Similarly, the Borel measurable functions are the natural ones to consider for integration theory. (Later we will expand the classes of Borel sets and Borel measurable functions to the larger classes of Lebesgue measurable sets and functions, primarily for technical reasons.)

We will give two characterizations of the Borel sets in this section, and a third important one in the next section, along with corresponding descriptions of the Borel measurable functions. Most of the work can be done in a general topological space, but we will emphasize the most important case of $\mathbb{R}$ in examples.

XII.2.1. Borel Sets

XII.2.1.1. Definition. Let $(X, T)$ be a topological space. Let $\mathcal{B}_T$ be the $\sigma$-algebra on $X$ generated by $T$, i.e. the smallest $\sigma$-algebra containing the open sets of $X$. $\mathcal{B}_T$ is called the class of Borel sets of $(X, T)$.

When a set $X$ such as $\mathbb{R}$ has a natural topology, we will usually just write $\mathcal{B}$ for the class of Borel sets on $X$, suppressing the subscript $T$.

XII.2.1.2. The following sets are Borel sets in $(X, T)$:

(i) All open sets in $X$.

(i') All closed sets in $X$.

(ii) Any countable intersection of open sets in $X$. Such a subset of $X$ is called a $G_\delta$ set.

(ii') Any countable union of closed sets in $X$. Such a subset of $X$ is called an $F_\sigma$ set.

(iii) Any countable union of $G_\delta$ sets in $X$. Such a subset of $X$ is called a $G_\delta\sigma$ set.

(iii') Any countable intersection of $F_\sigma$ sets in $X$. Such a subset of $X$ is called a $F_\sigma\delta$ set.

(iv) $G_{\delta\sigma\delta}, F_{\sigma\delta\sigma}, G_{\delta\sigma\delta\sigma}, \cdots$ sets can be defined analogously.

XII.2.1.3. In most cases, and in particular if $X = \mathbb{R}$, this process never terminates and must be repeated transfinitely (uncountably many times) to obtain all Borel sets (XII.3.5.4.). (It is not obvious that these classes are never $\sigma$-algebras after finitely or countably many steps for $X = \mathbb{R}$; see () for a proof.) The complement of a $G_\delta$ set is an $F_\sigma$ set, the complement of an $F_\sigma\delta$ is a $G_\delta\sigma$, etc. For measure theoretic purposes, it is rarely necessary to consider sets more complicated than $G_\delta$’s or $F_\sigma$’s (which are complicated enough already!)

The notation $F_\sigma, G_\delta, \text{etc.}$, is due to HAUSDORFF. There is a more modern notation, described in XII.3.4., which is used by descriptive set theorists; but HAUSDORFF’S notation is still commonly used by analysts.

It should be obvious that the class of Borel sets in $\mathbb{R}$ includes a great many subsets. For example:
XII.2.1.4.

(i) Every singleton set \( \{x\} \) in \( \mathbb{R} \) (or, more generally, in a \( T_1 \) space \( X \)) is a closed set, hence a Borel set. Every countable subset of \( \mathbb{R} \) is thus an \( F_\sigma \), hence a Borel set, since it is a countable union of singleton sets. In particular, \( \mathbb{Q} \) is an \( F_\sigma \), and hence \( J \) is a \( G_\delta \). (It turns out, although it is not obvious, that \( \mathbb{Q} \) is not a \( G_\delta \), so \( J \) is not an \( F_\sigma \); see XI.12.1.5.)

(ii) Every open interval in \( \mathbb{R} \) is an \( F_\sigma \); \( (a,b) = \bigcup_{n=1}^\infty [a + \frac{1}{n}, b - \frac{1}{n}] \), and similarly for \( (a,\infty) \) or \( (-\infty,b) \). Since every open set in \( \mathbb{R} \) is a countable union of open intervals, every open set in \( \mathbb{R} \) is an \( F_\sigma \), and thus every closed set is a \( G_\delta \). Of course, every open set is a \( G_\delta \), so every closed set is an \( F_\sigma \).

It is not true that every open set in a general topological space \( X \) is an \( F_\sigma \), but it is true if \( X \) is metrizable ().

(iii) A half-open interval in \( \mathbb{R} \) is neither open nor closed, but is both an \( F_\sigma \) and a \( G_\delta \), and hence a Borel set. The proof is very similar to the proof of (ii), and is left as an exercise.

(iv) An arbitrary (not necessarily countable) union of actual intervals can be written as a countable union of possibly different intervals (), hence is a Borel set (in fact, an \( F_\sigma \)).

An interesting and useful consequence of the construction of the Borel sets is:

XII.2.1.5. PROPOSITION. Let \( \mathcal{A} \) be a collection of subsets of \( \mathbb{R} \) (or any metrizable space) such that

1. \( \mathcal{A} \) contains all closed sets.
2. \( \mathcal{A} \) is closed under countable unions and countable intersections.

Then \( \mathcal{A} \) contains all Borel sets.

PROOF: \( \mathcal{A} \) contains all \( F_\sigma \)'s. Since every open set in \( \mathbb{R} \) is an \( F_\sigma \), \( \mathcal{A} \) contains all open sets, and hence all \( G_\delta \)'s. Hence \( \mathcal{A} \) contains all \( F_{\sigma\delta} \)'s, all \( G_{\delta\sigma} \)'s, \ldots, and by transfinite induction all Borel sets. 

XII.2.1.6. Note that such an \( \mathcal{A} \) need not be a \( \sigma \)-algebra, i.e. need not be closed under complements. An example is the collection of analytic sets (). A similar proof shows that a collection of subsets of a metrizable space which contains all open sets and is closed under countable unions and intersections contains all Borel sets.

XII.2.2. Borel Measurable Functions

Along with the Borel sets themselves, the Borel measurable functions are basic objects for study in measure theory.
XII.2.2.1. Definition. Let \((X, \mathcal{A})\) be a measurable space. A real-valued function \(f : X \to \mathbb{R}\) is \(\mathcal{A}\)-measurable if it is a measurable as a function from \((X, \mathcal{A})\) to \((\mathbb{R}, \mathcal{B})\) in the sense of (\(\circ\)), i.e. if \(f^{-1}(B) \in \mathcal{A}\) for every Borel set \(B\) in \(\mathbb{R}\). Denote by \(\mathcal{B}_{\mathcal{A}}\) the set of \(\mathcal{A}\)-measurable functions from \(X\) to \(\mathbb{R}\).

We can similarly define \(\mathcal{A}\)-measurable functions from a measurable space \((X, \mathcal{A})\) to a topological space \((Y, \mathcal{T})\) to be functions which are \((\mathcal{A}, \mathcal{B}_\mathcal{T})\)-measurable in the sense of (\(\circ\)). Thus we may also define \(\mathcal{A}\)-measurable functions from \(X\) to \(\mathbb{R}^n\).

If \((X, \mathcal{T})\) is a topological space, a function \(f : X \to \mathbb{R}\) is Borel measurable if it is \(\mathcal{B}_\mathcal{T}\)-measurable, i.e. the inverse image of every Borel set in \(\mathbb{R}\) is a Borel set in \((X, \mathcal{T})\). Denote by \(\mathcal{B}_\mathcal{T}\) the set of Borel measurable functions from \(X\) to \(\mathbb{R}\). Write \(\mathcal{B}\) for the set of Borel measurable functions from \(\mathbb{R}\) to \(\mathbb{R}\). We may similarly define Borel measurability for functions from \(X\) to \(\mathbb{R}^n\), and more generally for functions from \(X\) to any topological space \((Y, \mathcal{S})\).

XII.2.2.2. We most commonly use the term “Borel measurable” for functions from \(\mathbb{R}\) to \(\mathbb{R}\) (or from \(\mathbb{R}^n\) to \(\mathbb{R}^m\)). We will show that all “reasonable” functions from \(\mathbb{R}\) to \(\mathbb{R}\) are Borel measurable. In fact, roughly speaking, any function from \(\mathbb{R}\) to \(\mathbb{R}\) which can be defined or described without using the Axiom of Choice is Borel measurable; in particular, any function which can be defined by a countable number of algebraic, order, or pointwise limit operations is Borel measurable. (Strictly speaking, the first part of this statement is not true; it is possible, but quite complicated, to explicitly describe a subset of \(\mathbb{R}\) which is not a Borel set, and hence a function which is not Borel measurable; see (\(\circ\)).

The next proposition is a special case of XII.1.4.3.: 

XII.2.2.3. Proposition. (a) Let \((X, \mathcal{A})\) and \((Y, \mathcal{C})\) be measurable spaces, \(f : X \to Y\) an \((\mathcal{A}, \mathcal{C})\)-measurable function, and \(g : Y \to \mathbb{R}\) a \(\mathcal{C}\)-measurable function. Then \(g \circ f : X \to \mathbb{R}\) is \(\mathcal{A}\)-measurable. (Here \(\mathbb{R}\) may be replaced by any topological space.)
(b) Let \((X, \mathcal{T})\), \((Y, \mathcal{S})\), and \((Z, \mathcal{R})\) be topological spaces, \(f : X \to Y\) and \(g : Y \to Z\) Borel measurable functions. Then \(g \circ f : X \to Z\) is Borel measurable.

It is often convenient to use an apparently weaker criterion for measurability of real-valued functions. Here are some of the most useful ones:

XII.2.2.4. Proposition. Let \((X, \mathcal{A})\) be a measurable space and \(f : X \to \mathbb{R}\) a function. Then \(f\) is \(\mathcal{A}\)-measurable if and only if any of the following conditions are satisfied:

- \((i)\) \(f^{-1}(U) \in \mathcal{A}\) for every open set \(U \subseteq \mathbb{R}\).
- \((i)\) \(f^{-1}(E) \in \mathcal{A}\) for every closed set \(E \subseteq \mathbb{R}\).
- \((ii)\) \(f^{-1}(U) \in \mathcal{A}\) for every open interval \(U \subseteq \mathbb{R}\).
- \((iv)\) \(f^{-1}((-\infty, a)) \in \mathcal{A}\) for every \(a \in \mathbb{R}\).
- \((v)\) \(f^{-1}((a, +\infty)) \in \mathcal{A}\) for every \(a \in \mathbb{R}\).
- \((vi)\) \(f^{-1}([a, +\infty)) \in \mathcal{A}\) for every \(a \in \mathbb{R}\).
- \((vii)\) \(f^{-1}((a, +\infty)) \in \mathcal{A}\) for every \(a \in \mathbb{R}\).
- \((viii)\) \(f^{-1}((\infty, a)) \in \mathcal{A}\) for every \(a \in \mathbb{Q}\).
(ix) $f^{-1}((-\infty, a]) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

(x) $f^{-1}((a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

(xi) $f^{-1}([a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

These criteria apply to Borel measurability for functions from $\mathbb{R}$ to $\mathbb{R}$ (or from a topological space $(X, \mathcal{T})$ to $\mathbb{R}$), with $\mathcal{A} = \mathcal{B}$.

This proposition is an immediate consequence of XII.1.4.9., since each of the collections of sets listed generates $\mathcal{B}$ as a $\sigma$-algebra.

**XII.2.2.5.** Corollary. Every continuous function from a topological space $(X, \mathcal{T})$ to $\mathbb{R}$ is Borel measurable. In particular, every continuous function from $\mathbb{R}$ to $\mathbb{R}$ is Borel measurable.

**Proof:** Apply (i): if $f$ is continuous and $U \subseteq \mathbb{R}$ is open, then $f^{-1}(U)$ is open in $(X, \mathcal{T})$, and every open set is a Borel set.

The next special case of XII.1.4.11. (using XII.2.2.5.) shows that “piecewise-continuous” functions are Borel measurable.

**XII.2.2.6.** Corollary. Let $(X, \mathcal{T})$ be a topological space, and $(E_n)$ a sequence of Borel subsets of $X$ with $\bigcup_{n=1}^{\infty} E_n = X$. Let $f : X \to \mathbb{R}$ be a function. If $f|_{E_n}$ is continuous (regarding $E_n$ as a topological space with the relative topology) for each $n$, then $f$ is Borel measurable.

Many very discontinuous functions are Borel measurable. For example:

**XII.2.2.7.** Proposition. Let $(X, \mathcal{A})$ be a measurable space, and $A \subseteq X$. Then $\chi_A$ is an $\mathcal{A}$-measurable function if and only if $A \in \mathcal{A}$. In particular, if $B \subseteq \mathbb{R}$, then $\chi_B$ is a Borel measurable function if and only if $B$ is a Borel set in $\mathbb{R}$.

**Proof:** Since $\chi_A$ takes only the values 0 and 1, $\chi_A^{-1}((a, +\infty))$ is either $\emptyset$ (if $a \geq 1$), $A$ (if $0 \leq a < 1$), or $X$ (if $a < 0$). Apply XII.2.2.4.(vi).

More generally, we have:

**XII.2.2.8.** Proposition. Let $(X, \mathcal{A})$ be a measurable space, and $\phi = \sum_{i=1}^{n} \alpha_i \chi_{A_i}$ be a simple function from $X$ to $\mathbb{R}$. Then $\phi$ is $\mathcal{A}$-measurable if and only if each $A_i$ is in $\mathcal{A}$.

The proof is almost identical to the proof of XII.2.2.7., and is left as an exercise.

From XII.2.2.3., in combination with XII.2.2.5., we obtain an important property of the set $\mathfrak{B}_A$ of $\mathcal{A}$-measurable functions:
XII.2.2.9. **Proposition.** Let \((X, \mathcal{A})\) be a measurable space, \(f \in \mathcal{B}_A\), and \(g\) a Borel measurable function from \(\mathbb{R}\) to \(\mathbb{R}\). Then \(g \circ f\) is in \(\mathcal{B}_A\). In particular, if \(h : \mathbb{R} \to \mathbb{R}\) is continuous, then \(h \circ f\) is in \(\mathcal{B}_A\).

XII.2.2.10. If \(f\) is a real-valued function, we define the positive and negative parts of \(f\) to be \(f_+\) and \(f_-\), where \(f_+(x) = \max(f(x), 0)\) and \(f_-(x) = \max(-f(x), 0)\). In other words, \(f_+ = f\) wherever \(f\) is nonnegative, and is zero where \(f\) is negative; and \(f_- = |f|\) where \(f\) is negative and zero where \(f\) is nonnegative. Then \(f = f_+ - f_-\), \(|f| = f_+ + f_-\), and \(f_+ f_- = 0\).

XII.2.2.11. **Corollary.** Let \((X, \mathcal{A})\) be a measurable space, and \(f \in \mathcal{B}_A\). Then \(f^2, f_+, f_-\), and \(|f|\) are in \(\mathcal{B}_A\). If \(f \geq 0\), then \(f \in \mathcal{B}_A\).

**Proof:** \(f^2 = g \circ f\), where \(g(t) = t^2\). Similarly, the other functions are the compositions of \(f\) with the functions \(\max(t, 0), \max(-t, 0), |t|, \) and \(\sqrt{|t|}\) respectively.

This corollary can also be easily proved directly from XII.2.4.

Our next goal is to show that \(\mathcal{B}_A\) is closed under the usual algebraic operations: sum, product, and scalar multiple. (We say that \(\mathcal{B}_A\) is an algebra of functions.) The only step which takes some work is showing closure under addition.

XII.2.2.12. **Theorem.** Let \((X, \mathcal{A})\) be a measurable space, \(f, g \in \mathcal{B}_A\). Then \(f + g\) and \(fg\) are in \(\mathcal{B}_A\).

**Proof:** Closure under scalar multiplication is obvious from XII.2.4.(vi)-(viii).

We show closure under addition. Let \(h = f + g\). To show that \(h \in \mathcal{B}_A\), then by XII.2.4.(vi) it suffices to show that \(h^{-1}((a, \infty)) \in \mathcal{A}\) for every \(a \in \mathbb{R}\). But if \(a\) is fixed, then

\[
\begin{align*}
h^{-1}((a, \infty)) &= \{x : f(x) > a - g(x)\} \\
&= \{x : f(x) > r > a - g(x) \text{ for some } r \in \mathbb{Q}\} \\
&= \{x : f(x) > r \text{ and } g(x) > a - r \text{ for some } r \in \mathbb{Q}\} \\
&= \bigcup_{r \in \mathbb{Q}} [f^{-1}((r, \infty)) \cap g^{-1}((a - r, \infty))].
\end{align*}
\]

Since \(f\) and \(g\) are \(\mathcal{A}\)-measurable, and \(\mathbb{Q}\) is countable, the last set is in \(\mathcal{A}\).

A trick using XII.2.2.11. gives closure under multiplication:

\[
fg = \frac{1}{2}((f + g)^2 - f^2 - g^2).
\]
Closure under Pointwise Limits

Measurable Functions to $\mathbb{R}^n$

As previously noted, we can define $\mathcal{A}$-measurability for functions from a measurable space $(X, \mathcal{A})$ to $\mathbb{R}^n$. We do not get anything essentially new by doing this, however:

**XII.2.2.13. Proposition.** Let $(X, \mathcal{A})$ be a measurable space, and $f : X \to \mathbb{R}^n$ a function. Write $f_k : X \to \mathbb{R}$ $(1 \leq k \leq n)$ for the $k$'th coordinate functions of $f$, i.e. $f(x) = (f_1(x), \ldots, f_n(x))$ for $x \in X$. Then $f$ is an $\mathcal{A}$-measurable function from $X$ to $\mathbb{R}^n$ if and only if each $f_k$ is an $\mathcal{A}$-measurable function from $X$ to $\mathbb{R}$.

**Proof:** We have $f_k = \pi_k \circ f$, where $\pi_k : \mathbb{R}^n \to \mathbb{R}$ is the projection onto the $k$'th coordinate. Since $\pi_k$ is continuous, $f_k$ is $\mathcal{A}$-measurable if $f$ is $\mathcal{A}$-measurable by XII.2.2.9. Conversely, suppose each $f_k$ is $\mathcal{A}$-measurable. Let $B_1, \ldots, B_n$ be Borel sets in $\mathbb{R}$, and $B = B_1 \times B_2 \times \cdots \times B_n$ the corresponding measurable rectangle in $\mathbb{R}^n$. Then $f^{-1}(B) = f_1^{-1}(B_1) \cap \cdots \cap f_n^{-1}(B_n)$. Since $f_k^{-1}(B_k) \in \mathcal{A}$ for each $k$, $f^{-1}(B) \in \mathcal{A}$. Measurable rectangles generate the $\sigma$-algebra of Borel sets in $\mathbb{R}^n$, so $f$ is $\mathcal{A}$-measurable by XII.1.4.9.

**XII.2.3. Another Characterization of Borel Sets and Borel Measurable Functions**

It turns out that the Borel measurable functions have a simple and useful alternate characterization as the “sequential completion” of the class of continuous functions. The results of this section can be generalized to functions on arbitrary topological spaces, but we will stick to functions from $\mathbb{R}$ to $\mathbb{R}$ for simplicity.

**XII.2.3.1. Definition.** Let $\mathcal{C}$ be the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Let $\mathcal{C}$ be the smallest class of functions from $\mathbb{R}$ to $\mathbb{R}$ containing $\mathcal{C}$ and which is closed under pointwise sequential limits, i.e. such that if $(f_n)$ is a sequence in $\mathcal{C}$ and $f_n \to f$ pointwise, then $f \in \mathcal{C}$.

A subset $A$ of $\mathbb{R}$ is a $BC$ set if $\chi_A \in \mathcal{C}$.

The abbreviation $BC$ stands for “Baire class.”

The definition of $\mathcal{C}$ makes sense as the intersection of all classes of functions with the two properties. This collection of classes is nonempty since the class of all functions from $\mathbb{R}$ to $\mathbb{R}$ has the properties.

We will show that the functions in $\mathcal{C}$ are precisely the Borel measurable functions, and the $BC$ sets are precisely the Borel sets. Such sets and functions are very natural to consider in measure and integration, where pointwise limits often arise.

**XII.2.3.2. Caution:** It is not true that $\mathcal{C}$ is just the set $\mathcal{BC}_1$ of pointwise limits of sequences of continuous functions. This set $\mathcal{BC}_1$ is contained in $\mathcal{C}$, but is not closed under pointwise sequential limits (although this is not obvious; see Exercise ()). Let $\mathcal{BC}_2$ be the set of pointwise limits of sequences of functions in $\mathcal{BC}_1$, and inductively for each ordinal $\alpha$ let $\mathcal{BC}_{\alpha+1}$ be the set of pointwise limits of sequences of functions in $\mathcal{BC}_\alpha$, with $\mathcal{BC}_\alpha = \cup_{\beta < \alpha} \mathcal{BC}_\beta$ when $\alpha$ is a limit ordinal. Then $\mathcal{C}$ is the union of the $\mathcal{BC}_\alpha$ over all countable $\alpha$, and it turns out that $\mathcal{C} \neq \mathcal{BC}_\alpha$ for any countable $\alpha$ (Appendix B). The functions in $\mathcal{BC}_\alpha$ are said to be of Baire class $\alpha$. 

1376
XII.2.4. Algebraic Properties of $\mathcal{C}$

We first show that, like $\mathcal{B}_A$ (XII.2.2.12), $\mathcal{C}$ is closed under the usual algebraic operations: sum, product, and scalar multiple (i.e. $\mathcal{C}$ is an algebra of functions.) It is a standard fact from Calculus that $\mathcal{C}$ is closed under these operations. It can be easily shown by transfinite induction that each $\mathcal{B}_\mathcal{C}_n$ is also closed under the operations, and hence so is $\mathcal{C}$. A cleaner and more elementary proof can be given, however:

XII.2.4.1. Proposition. Let $f, g \in \mathcal{C}$ and $\alpha \in \mathbb{R}$. Then $f + g$, $fg$, and $\alpha f$ are all in $\mathcal{C}$.

Proof: First fix $g \in \mathcal{C}$, and let $\mathcal{D} = \{ f \in \mathcal{C} : f + g \in \mathcal{C} \}$. We claim that $\mathcal{D} = \mathcal{C}$. Clearly $\mathcal{D} \subseteq \mathcal{C}$, so we need only show that $\mathcal{C} \subseteq \mathcal{D}$ and that $\mathcal{D}$ is closed under pointwise sequential limits. If $f \in \mathcal{C}$, then $f + g \in \mathcal{C} \subseteq \mathcal{C}$, so $f \in \mathcal{D}$. And if $(f_n)$ is a sequence in $\mathcal{D}$ (i.e. $f_n \in \mathcal{C}$ and $f_n + g \in \mathcal{C}$), and $f_n \to f$ pointwise, then $f_n \in \mathcal{C}$ for all $n$, so $f \in \mathcal{C}$; also, $f_n + g \to f + g$ pointwise, and $f_n + g \in \mathcal{C}$, so $f + g \in \mathcal{C}$, i.e. $f \in \mathcal{D}$. This proves the claim, and therefore the sum of a function in $\mathcal{C}$ and a function in $\mathcal{C}$ is in $\mathcal{C}$.

Now suppose $g \in \mathcal{C}$, and repeat the argument with $\mathcal{E} = \{ f \in \mathcal{C} : f + g \in \mathcal{C} \}$. As before, we show that $\mathcal{C} \subseteq \mathcal{E}$ (by the first part of the proof) and that $\mathcal{E}$ is closed under pointwise sequential limits, and thus $\mathcal{E} = \mathcal{C}$. This shows that $\mathcal{C}$ is closed under addition.

The proof that $\mathcal{C}$ is closed under multiplication is virtually identical, replacing “sum” by “product” everywhere, and is left as an exercise ().

The case of scalar multiplication is a special case of the product, regarding the scalar $\alpha$ as the constant function with value $\alpha$.

XII.2.4.2. Corollary. The class of BC sets in $\mathbb{R}$ is a $\sigma$-algebra.

Proof: Suppose $A$ is a BC set, i.e. $\chi_A \in \mathcal{C}$. Then $\chi_A^c = 1 - \chi_A \in \mathcal{C}$, so $A^c$ is a BC set. If $A$ and $B$ are BC sets, then $\chi_A \cap B = \chi_A \chi_B \in \mathcal{C}$, so $A \cap B$ is a BC set. Finally, if $(A_n)$ is a sequence of BC sets and $A = \cap_{n=1}^\infty A_n$, set $f_n = \chi_{A_1 \cap \cdots \cap A_n} = \chi_{A_1} \chi_{A_2} \cdots \chi_{A_n}$. Then $f_n \in \mathcal{C}$ and $f_n \to \chi_A$ pointwise, so $\chi_A \in \mathcal{C}$ and so $A$ is a BC set. Thus the class of BC sets is closed under complements and countable intersections. It clearly contains $\mathbb{R}$ since $\chi_\mathbb{R}$ is the constant function 1, which is in $\mathcal{C}$.

XII.2.4.3. Proposition. Every open interval in $\mathbb{R}$ is a BC set.

Proof: Let $(a, b)$ be an open interval in $\mathbb{R}$. We need to show that $\chi_{(a, b)} \in \mathcal{C}$; we will find a sequence of continuous functions converging pointwise to $\chi_{(a, b)}$. Let $f_n$ be the function with the following graph:

It is easy to check that $f_n \to \chi_{(a, b)}$ pointwise.

XII.2.4.4. Corollary. Every Borel set is a BC set.

Proof: The class of BC sets is a $\sigma$-algebra (XII.2.4.2.) which contains the open intervals (XII.2.4.3.). The smallest such $\sigma$-algebra is the Borel sets.

Now we show that $\mathcal{C}$ is closed under composition:
XII.2.4.5. Proposition. Let \( f, g \in \mathcal{C} \). Then \( f \circ g \in \mathcal{C} \).

Proof: The proof is very similar to the proof of XII.2.4.1. First fix \( f \in \mathcal{C} \), and let \( \mathcal{D} = \{ g \in \mathcal{C} : f \circ g \in \mathcal{C} \} \). As in XII.2.4.1, \( \mathcal{C} \subseteq \mathcal{D} \) since a composition of continuous functions is continuous. Also, if \( g_n \in \mathcal{D} \) and \( g_n \to g \) pointwise, and \( x \in \mathbb{R} \), if \( y_n = g_n(x) \) and \( y = g(x) \) we have that \( y_n \to y \), and thus \( [f \circ g_n](x) = f(y_n) \to f(y) = [f \circ g](x) \) since \( f \) is continuous; thus \( f \circ g_n \to f \circ g \) and so \( f \circ g \in \mathcal{C} \), \( f \in \mathcal{D} \). Thus \( \mathcal{D} = \mathcal{C} \), i.e. \( f \circ g \in \mathcal{C} \) if \( f \in \mathcal{C} \), \( g \in \mathcal{C} \).

Now fix \( g \in \mathcal{C} \) and let \( \mathcal{E} = \{ f \in \mathcal{C} : f \circ g \in \mathcal{C} \} \). By the first part of the proof, \( \mathcal{C} \subseteq \mathcal{E} \). Also, if \( (f_n) \) is a sequence in \( \mathcal{E} \) converging pointwise to \( f \), then \( f \in \mathcal{C} \), and \( f_n \circ g \to f \circ g \) pointwise (this is true with no restriction on \( g \)), and thus \( f \circ g \in \mathcal{C} \), \( f \in \mathcal{E} \). Thus \( \mathcal{C} = \mathcal{E} \) and the proposition is proved. \( \diamond \)

XII.2.4.6. Corollary. If \( g \in \mathcal{C} \), then \( |g| \in \mathcal{C} \).

Proof: We have that \( |g| = f \circ g \), where \( f(x) = |x| \), which is a continuous function. \( \diamond \)

See problem () for another proof of this result.

A very important consequence of this corollary comes from the following lemma.

XII.2.4.7. Lemma. If \( x, y \in \mathbb{R} \), then \( \max(x, y) = \frac{1}{2}(x + y + |x - y|) \) and \( \min(x, y) = \frac{1}{2}(x + y - |x - y|) \). Thus, if \( f \) and \( g \) are functions from \( \mathbb{R} \) to \( \mathbb{R} \), then \( \max(f, g) = \frac{1}{2}(f + g + |f - g|) \) and \( \min(f, g) = \frac{1}{2}(f + g - |f - g|) \).

Proof: The proof is a straightforward case-by-case verification which is left as an exercise. \( \diamond \)

XII.2.4.8. Corollary. If \( f \) and \( g \) are in \( \mathcal{C} \), then so are \( \max(f, g) \) and \( \min(f, g) \).

XII.2.4.9. A sequence \( (f_n) \) of functions from \( \mathbb{R} \) to \( \mathbb{R} \) is pointwise bounded if, for every \( x \in \mathbb{R} \), the set \( \{ f_n(x) : n \in \mathbb{N} \} \) is bounded in \( \mathbb{R} \). If \( (f_n) \) is pointwise bounded, then \( \sup_n f_n \) and \( \inf_n f_n \) are defined as functions from \( \mathbb{R} \) to \( \mathbb{R} \): \( \sup_n f_n(x) = \sup\{ f_n(x) : n \in \mathbb{N} \} \), \( \inf_n f_n(x) = \inf\{ f_n(x) : n \in \mathbb{N} \} \).

XII.2.4.10. Corollary. Let \( (f_n) \) be a pointwise bounded sequence in \( \mathcal{C} \). Then \( \sup_n f_n \) and \( \inf_n f_n \) are in \( \mathcal{E} \).

Proof: Let \( g_n = \max(f_1, \ldots, f_n) \) and \( h_n = \min(f_1, \ldots, f_n) \). Then \( g_n, h_n \in \mathcal{E} \) by XII.2.4.8., and \( g_n \to \sup_n f_n, h_n \to \inf_n f_n \) pointwise.

In fact, it turns out that any algebra of functions which is closed under pointwise sequential limits is also closed under countable suprema and infima (Problem ()). A clever trick yields a converse to this:

1378
XII.2.4.11. **Proposition.** Let $\mathcal{F}$ be a set of real-valued functions on a set $X$. If $\mathcal{F}$ is closed under suprema and infima of pointwise bounded sequences, then $\mathcal{F}$ is closed under pointwise sequential limits.

**Proof:** Suppose $(f_n)$ is a sequence in $\mathcal{F}$, and $f_n \to f$ pointwise. Then, for each $x \in X$, the sequence $(f_n(x))$ is convergent and therefore bounded; thus $(f_n)$ is pointwise bounded. Also, for $x \in X$,

$$f(x) = \lim_n f_n(x) = \limsup_n f_n(x) = \inf_n \{\sup_{k \geq n} f_k(x)\}$$

and so, if $g_n = \sup_{k \geq n} f_k$, then $g_n \in \mathcal{F}$ for all $n$, and therefore $f = \inf_n [\sup_{k \geq n} f_k] = \inf_n g_n \in \mathcal{F}$. $\Diamond$

This applies in particular to the collection of $\mathcal{A}$-measurable real-valued functions on any measurable space $(X, \mathcal{A})$:

XII.2.4.12. **Proposition.** Let $(X, \mathcal{A})$ be a measurable space, and $\mathcal{B}_A$ the set of $\mathcal{A}$-measurable real-valued functions. Then $\mathcal{B}_A$ is closed under suprema and infima of pointwise bounded sequences, and hence under pointwise sequential limits.

**Proof:** Let $(f_n)$ be a locally bounded sequence in $\mathcal{B}_A$, and let $g = \sup_n f_n$. We will show that $g^{-1}((a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{R}$, and use XII.2.4.(vi). If $x \in X$, then $g(x) > a$ if and only if $f_n(x) > a$ for some $n$, and thus $g^{-1}((a, +\infty)) = \bigcup_{n=1}^\infty f_n^{-1}((a, +\infty))$. By assumption, $f_n^{-1}((a, +\infty)) \in \mathcal{A}$ for all $n$, so $g^{-1}((a, +\infty))$ is a countable union of sets in $\mathcal{A}$.

The proof for $\inf_n f_n$ is nearly identical, using XII.2.4.(iv). $\Diamond$

The set of $\mathcal{A}$-measurable functions on $(X, \mathcal{A})$ is also an algebra (XII.2.12.).

As a consequence, we get a relationship between $\mathcal{C}$ and Borel measurability:

XII.2.4.13. **Corollary.** Every function in $\mathcal{C}$ is Borel measurable.

**Proof:** If $\mathcal{B}$ is the set of Borel measurable functions, then $\mathcal{B}$ contains $\mathcal{C}$ by XII.2.5., and is closed under pointwise sequential limits by XII.2.4.12.; thus $\mathcal{B}$ contains $\mathcal{C}$. $\Diamond$

XII.2.4.14. **Corollary.** Every BC set is a Borel set.

**Proof:** If $A$ is a BC set, then $\chi_A \in \mathcal{C}$, and hence $\chi_A$ is Borel measurable. But this implies that $A$ is a Borel set by XII.2.7.. $\Diamond$

Combining XII.2.4.14. with XII.2.4.4., we obtain one of the principal facts we needed to prove:

XII.2.4.15. **Theorem.** A subset of $\mathbb{R}$ is a Borel set if and only if it is a BC set.
**XII.2.4.16. Corollary.** A simple function is Borel measurable if and only if it is a BC function.

It still remains to prove that every Borel measurable function is in $\mathcal{E}$. Since every simple Borel measurable function is in $\mathcal{C}$ (XII.2.4.16.), it suffices to show:

**XII.2.4.17. Lemma.** Every Borel measurable function is a pointwise limit of simple Borel measurable functions.

**Proof:** Suppose $f$ is Borel measurable. Fix $n \in \mathbb{N}$. For each $k \in \mathbb{Z}$, let

$$E_{kn}(f) = f^{-1}\left(\left[\frac{k}{n}, \frac{k+1}{n}\right]\right).$$

Then each $E_{kn}(f)$ is a Borel set since it is the inverse image of a Borel set, and for fixed $n$ we have that $\mathbb{R}$ is the disjoint union of the $E_{kn}(f)$’s. Let $f_n(t) = \frac{k}{n}$ if $t \in E_{kn}(f)$. Since $\frac{k}{n} \leq f(t) < \frac{k+1}{n}$ for $t \in E_{kn}(f)$, we have that $f_n(t) \leq f(t)$ and $f(t) - f_n(t) < \frac{1}{n}$ for every $t$. Thus the sequence $(f_n)$ actually converges uniformly to $f$. However, $f_n$ is not quite simple in general. To obtain simple functions, let $g_n(t) = f_n(t)$ if $t \in E_{kn}(f)$ for $-n \leq k \leq n$ and $g_n(t) = 0$ otherwise. Then $g_n$ is simple, and $g_n \to f$ pointwise on $\mathbb{R}$ (but not uniformly unless $f$ is bounded.)

So we have shown:

**XII.2.4.18. Theorem.** $\mathcal{E} = \mathcal{B}$.

**Egorov’s Theorem**

**XII.2.5. The Multiplicity Function**

In many settings where a function is being analyzed, it is important to consider the “size” of inverse images of points; this provides, among other things, a description of how far the function is from being one-to-one. The most basic way to do this is via the “multiplicity function” (sometimes called the “counting function”):

**XII.2.5.1. Definition.** Let $f : X \to Y$ be a function. The multiplicity function of $f$ is the function $N_f : Y \to \{0\} \cup \mathbb{N} \cup \{+\infty\}$ defined by

$$N_f(y) = \begin{cases} \text{card}(f^{-1}(\{y\})) & \text{if } f^{-1}(\{y\}) \text{ is finite} \\ +\infty & \text{if } f^{-1}(\{y\}) \text{ is infinite} \end{cases}.$$

The Multiplicity Function is most commonly used when $Y = \mathbb{R}$ (or $\mathbb{R}^n$). Note that in this function, we do not distinguish between cases where $f^{-1}(\{y\})$ is countable and cases where it is uncountable. We can sometimes take this into account by a finer invariant; cf. ( ).

1380
XII.2.5.2. **Examples.** (i) \( N_f \) is the constant function 1 if and only if \( f \) is a bijection. If \( f \) is one-to-one, then \( N_f \) takes only the values 0 and 1.

(ii) If \( f : \mathbb{R} \to \mathbb{R} \) is given by \( f(x) = x^2 \), then \( N_f(y) = \begin{cases} 1 & \text{if } y < 0 \\ 1 & \text{if } y = 0 \\ 2 & \text{if } y > 0 \end{cases} \).

(iii) If \( f : [0,1] \to [0,1] \) is a coordinate function of a space-filling curve (.), then \( f \) is continuous and \( N_f(y) = +\infty \) for all \( y \in [0,1] \).

XII.2.5.3. If \((X, A)\) and \((Y, B)\) are measurable spaces and \( f : X \to Y \) is measurable, it is not true in general (cf. XII.2.5.7.) even if \( X = Y = \mathbb{R} \), that \( N_f \) is a measurable function (since its range is a countable set, \( N_f \) will be measurable if and only if, for each \( n \in \{0\} \cup \mathbb{N} \cup \{+\infty\} \) we have that \( \{y \in Y : N_f(y) = n\} \) is in \( B \)). But in some important special cases this can be shown. The next result is the simplest example, and perhaps the most widely used; the proof can be adapted to broader contexts.

XII.2.5.4. **Proposition.** Let \( f \) be a real-valued continuous function on an an interval \( I \) in \( \mathbb{R} \). Then \( N_f \) is an extended real-valued Borel measurable function.

**Proof:** First suppose \( I \) is a bounded interval from \( a \) to \( b \) (with or without endpoints). For each \( n \in \mathbb{N} \), define a function \( \phi_n : \mathbb{R} \to \mathbb{N} \) by taking a partition \( \{x_0 = a, x_1, \ldots, x_{2^n} = b\} \) of \( I \) into \( 2^n \) subintervals \( J_1, \ldots, J_{2^n} \) of equal length, with \( x_k \in J_k \), \( x_k \notin J_{k+1} \) for \( 1 \leq k \leq 2^n - 1 \), \( a \in J_1 \) if \( a \in I \), and \( b \in J_{2^n} \) if \( b \in I \). Then set

\[
\phi_n = \sum_{k=1}^{2^n} \chi_{f(J_k)}.
\]

We have that \( f(J_k) \) is an interval, hence a Borel set; thus \( \phi_n \) is a Borel measurable function from \( \mathbb{R} \) to \( \mathbb{R} \). For any \( y \in \mathbb{R} \), \( \phi_n(y) \) is the number of subintervals whose intersection with \( f^{-1}(\{y\}) \) is nonempty. It is easily seen that \( \phi_n \leq \phi_{n+1} \) for each \( n \), and that \( N_f = \sup_n \phi_n \). Thus \( N_f \) is Borel measurable.

For a general \( I \), write \( I = \cup I_m \), where each \( I_m \) is a bounded interval and \( I_m \subseteq I_{m+1} \). Let \( f_m = f|_{I_m} \). Then \( N_f = \sup_m N_{f_m} \), and each \( f_m \) is Borel measurable, hence so is \( N_f \).

XII.2.5.5. We can try to adapt this argument to a more general setting of a Borel measurable function \( f \) (e.g. a continuous function) between topological spaces \( X \) and \( Y \). There are two things which need to be done:

(i) We need a sequence of finite partitions of \( X \), each refining the previous one, into “nice” subsets (e.g. Borel sets), such that distinct elements of \( X \) are separated by sufficiently fine partitions in the sequence. This can be done systematically using Borel sets in any Polish space (.), hence in any standard Borel space (.), e.g. in any Borel subset of \( \mathbb{R}^n \); for Borel sets in \( \mathbb{R}^n \) there is an obvious approach using grids similar to the approach in the proof of XII.2.5.4..
(ii) We need to know that if \( J_1, \ldots, J_r \) are the sets in the \( n \)'th partition, then \( f(J_k) \) is a Borel subset of \( Y \).
This is false in general (). However, if \( Y \) is a Polish space (e.g. \( \mathbb{R}^n \)), then \( f(J_k) \) is an analytic subset of \( Y \) if \( J_k \) is Borel in \( X \) ()

Combining these observations with (), we obtain the most general result which can be easily stated:

**XII.2.5.6. Theorem.** Let \((X, A)\) be a standard Borel space (e.g. a Borel subset of \( \mathbb{R}^n \)), and \( f : X \to \mathbb{R}^m \) a Borel measurable function. Then \( N_f \) is a Lebesgue measurable extended real-valued function on \( \mathbb{R}^m \).

Lebesgue measurability of \( N_f \) can be replaced by \( \mu \)-measurability, for any measure \( \mu \) which is the completion of a Borel measure on \( \mathbb{R}^m \), e.g. Hausdorff measure \( \mathcal{H}^s \) for any \( s \).

**XII.2.5.7. Example.** This result, which applies to Borel measurable functions from \( \mathbb{R} \) to \( \mathbb{R} \), cannot be extended to Lebesgue measurable functions from \( \mathbb{R} \) to \( \mathbb{R} \):

Let \( K \) be the usual Cantor set, and \( K' \subseteq [0,1] \) a generalized Cantor set of positive measure. Let \( g \) be a strictly increasing continuous function from \([0,1]\) onto \([0,1]\) with \( g(K) = K' \) (). Let \( Y \) be a subset of \( K' \) which is not Lebesgue measurable, and set \( A = g^{-1}(Y) \). Then \( A \) has measure 0 since it is a subset of \( K \).

Define \( f : \mathbb{R} \to \mathbb{R} \) by
\[
f(x) = \begin{cases}  g(x) & \text{if } x \in A \\ 2 & \text{if } x \notin A \end{cases}
\]

Then \( f \) is Lebesgue measurable (it equals the constant function \( 2 \) a.e.), but \( N_f \) is not Lebesgue measurable since \( \{ y \in \mathbb{R} : N_f(y) = 1 \} = Y \).

Note that although \( f = 2 \) a.e., we do not have \( N_f = N_2 \) a.e. In fact, \( N_2 \) (i.e. \( N_h \), where \( h \) is the constant function 2) is Borel measurable (it is constant a.e.)

For continuous real-valued functions on a closed bounded interval, there is a close relationship between the Multiplicity Function and the notion of bounded variation (**XIV.16.3.1.**).

**XII.2.6. Exercises**

**XII.2.6.1.** Let \( \mathfrak{B}_\alpha \) be the set of functions of Baire class \( \alpha \) (). A subset \( A \) of \( \mathbb{R} \) is a **set of Baire class \( \alpha \)** if \( \chi_A \in \mathfrak{B}_\alpha \). Let \( \mathcal{B}_\alpha \) denote the set of subsets of \( \mathbb{R} \) of Baire class \( \alpha \).

(a) Show that \( \cup_{\alpha < \omega_1} \mathcal{B}_\alpha \) is precisely the set of Borel sets.

(b) Show by transfinite induction that \( \mathfrak{B}_\alpha \) is an algebra of functions for each \( \alpha \).

(c) Show that each \( \mathcal{B}_\alpha \) is an algebra of sets. [Hint: \( \chi_{A \cap B} = \chi_A \chi_B \).]

(d) If \( f \in \mathfrak{B}_1 \) and \( U \) is an open set in \( \mathbb{R} \), then \( f^{-1}(U) \) is an \( F_\sigma \). (The converse is also true: if \( f^{-1}(U) \) is an \( F_\sigma \) for every open set \( U \), then \( f \in \mathfrak{B}_1 \)[].) [If \( f_n \to f \) pointwise and \( \alpha \in \mathbb{R} \), then
\[
f^{-1}((\alpha, \infty)) = \bigcup_n \bigcap_{m \geq m} \{ t : f_k(t) \geq \alpha + \frac{1}{n} \}.
\]

(e) Show that \( A \in \mathcal{B}_1 \) if and only if \( A \) is both an \( F_\sigma \) and a \( G_\delta \). [If \( A \) is both an \( F_\sigma \) and a \( G_\delta \), show that there is an increasing sequence \( (F_n) \) of closed sets, and a decreasing sequence \( (U_n) \) of open sets, such that

1382
A = ∪_n F_n = ∩_n U_n. If f_n is a continuous function which is 1 on F_n and 0 outside U_n, then f_n → χ_A. Deduce the converse from (d).] In particular, any open or closed set is in BC_1. In particular, any F_τ or G_δ is in BC_2.

(f) Show that a countable union or countable intersection of sets in BC_α is in BC_{α+1}. In particular, any F_τ or G_δ is in BC_2.

(g) Since Q is not a G_δ, Q ∉ BC_1. But Q ∈ BC_2 since it is an F_τ.

(h) Use the argument of (d) to show that if f ∈ BC_2 and U is open, then f^{-1}(U) is a G_δ. Conclude that if A is in BC_2, then A is both a G_δ and an F_σ. Generalize to BC_n and BC_α.

In fact, if α < β are countable ordinals, then BC_α is a proper subset of BC_β, and hence BC_α is a proper subset of the class of Borel sets (Appendix B).

**XII.2.6.2.** Let f be any function from R to R. Show that the set of points of continuity of f is a G_δ. [For each n, let U_n be the set of all a ∈ R for which there is a δ > 0 such that |f(x) - f(y)| < 1/n whenever |x - a| < δ and |y - a| < δ. Show that U_n is open.]

**XII.2.6.3.** (a) If I is an interval, show that there is a Borel measurable f : R → R such that I is exactly the set of discontinuities for f. [If I is closed, consider χ_QI. If I is open, let g be a continuous function which is nonzero precisely on I, and consider χ_Qg.]

(b) If B is an F_σ containing no interval, show there is a Borel measurable f whose set of discontinuities is precisely B. [Write B = ∪ F_n, with F_n closed, and consider ∑ 2^{-n} χ_{F_n}]

(c) Show that if A is a G_δ, then there is a Borel measurable f : R → R such that A is exactly the set of points where f is continuous.

**XII.2.6.4.** Let f : R → R be continuous. (a) For n, k ∈ N, c ∈ R, let U_{n,k,c} be the set of all a for which there is an x with

0 < |x - a| < 1/n and \[ \frac{f(x) - f(a)}{x - a} > c - \frac{1}{k} \]

Show that U_{n,k,c} is open and ∩_{n,k} U_{n,k,c} is the set

\[ V_c = \left\{ a \in \mathbb{R} : \limsup_{x \to a} \frac{f(x) - f(a)}{x - a} \geq c \right\} \]

So V_c is a G_δ.

(b) Similarly, if d ∈ R, show that

\[ W_d = \left\{ a \in \mathbb{R} : \liminf_{x \to a} \frac{f(x) - f(a)}{x - a} \leq d \right\} \]

is a G_δ.

(c) Show that the set D of points where f is differentiable is an F_σδ. [The complement of D is

\[ \left( \bigcup_{c,d \in \mathbb{Q}} V_c \cap W_d \right) \cup \left( \bigcap_{c \in \mathbb{Q}} V_c \right) \cup \left( \bigcap_{d \in \mathbb{Q}} W_d \right). \]

(d) What if f is not assumed to be continuous? (cf. XII.2.6.2.)
XII.2.6.5. Let $f$ be a real-valued function on an interval $I$ which has only countably many discontinuities. Show that $f$ is Baire class 1. [Let $(x_n)$ be a dense sequence in $I$ containing all discontinuities of $f$, and let $f_n$ be a piecewise-linear continuous function agreeing with $f$ on $\{x_1, \ldots, x_n\}$. Show that $f_n \to f$ pointwise.]

XII.2.6.6. Let $I$ be an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$. Extend $f$ to $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by setting $\tilde{f}$ identically 0 outside $I$. Show that $\tilde{f}$ is Baire class 1 on $\mathbb{R}$ if and only if $f$ is Baire class 1 on $I$.

XII.2.6.7. Let $I$ be an interval in $\mathbb{R}$. Let $D$ be a vector space of real-valued functions on $I$ containing the constant functions, whose uniform closure $\bar{D}$ has the property that if $f \in D$, then $|f| \in D$, and let $D$ be the set of all pointwise limits of sequences in $\bar{D}$. Show that $D$ is closed under uniform limits, as follows.

(a) If $f$ is a pointwise limit of a sequence in $\bar{D}$, show that $f$ is a pointwise limit of a sequence in $D$. Thus we may, and will, assume $\bar{D}$ is uniformly closed.

(b) If $f \in D$, then $f + = f \vee 0 = \frac{1}{2}(f + |f|) \in D$. Show that if $a \leq b$, then $(f \vee a) \wedge b \in D$.

(c) Let $(f_n)$ a sequence in $D$ which converges uniformly to $f$ on $I$. Passing to a subsequence, we may assume that $|f_{n+1}(x) - f_n(x)| \leq 2^{-n-1}$ for all $n$ and all $x \in I$. Set $g_n = f_1$ and $g_{n+1} = f_{n+1} - f_n$ for $n \geq 1$. For each $n$ let $(h_{nk})$ be a sequence in $D$ with $h_{nk} \to g_n$ pointwise. Use (b) to show that we may assume $|h_{nk}(x)| \leq 2^{-n}$ for all $n > 1$, all $k$, and all $x \in I$.

(d) For each $k$, the infinite series $\sum_{n=1}^{\infty} h_{nk}$ converges uniformly. Let $h_k = \sum_{n=1}^{\infty} h_{nk}$. Then $h_k \in D$. Show that $h_k \to f$ pointwise.

(e) The hypotheses are satisfied if $D$ is an algebra of bounded functions on $I$ and contains the constant functions (XV.8.2.9).

(f) The hypotheses are satisfied if $D = \mathcal{C}(I)$. Conclude that $\mathcal{BC}_1(I)$ is uniformly closed.

(g) Show by induction on $n$ that if $g \in \mathcal{BC}_n(I)$ and $f$ is continuous, then $f \circ g$ is in $\mathcal{BC}_n(I)$. Conclude that $D = \mathcal{BC}_n(I)$ satisfies the hypotheses, and hence that $\mathcal{BC}_{n+1}(I)$ is closed under uniform limits.

(h) Show that the result of the problem holds if “uniform” is replaced by “u.c.” throughout. Thus the class $\mathcal{BC}_n(I)$ is closed under u.c. limits for each $n \in \mathbb{N}$.

XII.2.6.8. Let $I$ be an interval in $\mathbb{R}$.

(a) Show that every step function on $I$ is in $\mathcal{BC}_1(I)$. [Use XII.2.6.5.]

(b) Show that every regulated function (V.8.6.7.) on $I$ is in $\mathcal{BC}_1(I)$. [Use XII.2.6.5. and XII.2.6.7., or XII.2.6.5. and V.8.6.7.(a).]
XII.3. Analytic Sets

There is an important class of subsets of \( \mathbb{R} \) (or more general topological spaces) which is slightly larger than the class of Borel sets.

XII.3.1. Standard Borel Spaces

XII.3.1.1. Recall that a topological space is a Polish space if it is homeomorphic to a separable complete metric space.

The next result says that, although there is an enormous variety of topologically distinct uncountable Polish spaces, they are all “identical” from a measure-theoretic standpoint:

Theorem. Let \( X \) and \( Y \) be uncountable Polish spaces. Then the measurable spaces \((X, B_X)\) and \((Y, B_Y)\) are isomorphic. In particular, \((X, B_X)\) is isomorphic to \((\mathbb{R}, \mathcal{B})\) and to \(([0, 1], B_{[0,1]})\).

This theorem is an easy consequence of the following result of Kuratowski [7], which is a sharpening of XI.10.2.7. The proof is rather involved, although fairly elementary, and we omit it; see, for example, [7, 2.4.1].

XII.3.1.2. Theorem. Let \( X \) be an uncountable Polish space which is dense in itself. Then there is an injective continuous function from \( \mathbb{N} \) onto \( X \) such that the image of every open set is an \( F_\sigma \).

Such a function is obviously a Borel isomorphism. To obtain XII.3.1.2., note that by the Cantor-Bendixson Theorem (XI.2.5.4.) every Polish space is a disjoint union of a countable open subset and a closed subset which is dense-in-itself.

XII.3.1.3. Theorem. Let \( X \) be an uncountable Polish space which is dense in itself. Then there is an injective continuous function from \( \mathbb{N} \) onto \( X \) such that the image of every open set is an \( F_\sigma \).

Since every countable subset of a \( T_1 \)-space is a Borel set, it follows that \((\mathbb{N}, B_\mathbb{N})\) is isomorphic as a Borel space to \((\mathbb{N}, B_\mathbb{N}) \oplus (\mathbb{P}(\mathbb{N}))\). Any uncountable Polish space \( X \) decomposes into a countable set and a Borel subset \( Z \) Borel isomorphic to \( \mathbb{N} \) by XII.3.1.3., so

\[ (X, B_X) \cong (Z, B_Z) \oplus (\mathbb{N}, P(\mathbb{N})) \cong (\mathbb{N}, B_{\mathbb{N}}) \oplus (\mathbb{N}, P(\mathbb{N})) \cong (\mathbb{N}, B_{\mathbb{N}}). \]

See problems (–) for an alternate proof of XII.3.1.2.

In particular, we have

XII.3.1.4. So to finish the proof of XII.3.1.2, note that if \( Y \) is \( \mathbb{N} \) with a countable number of points removed, then \( Y \) is a \( G_\delta \) in \( \mathbb{N} \), hence a Polish space, and \( Y \) is dense in itself; so \( Y \) is Borel isomorphic to \( \mathbb{N} \).

In particular, we have

XII.3.1.5. Corollary. Every uncountable Polish space has cardinality \( 2^{\aleph_0} \).

XII.3.1.6. Definition. A measurable space \((X, B)\) is a standard Borel space if it is isomorphic to \(([0, 1], B_{[0,1]}),\) hence to \((Y, B_Y)\) for any uncountable Polish space \( Y \).

Sometimes measurable spaces of the form \((X, P(X))\) for a countable set \( X \) are also called standard Borel spaces. These are the measurable spaces of the form \((X, B_X)\) for countable Polish spaces \( X \).
Subspaces of Standard Borel Spaces

XII.3.1.7. **Proposition.** Let \((X, \mathcal{A})\) be a measurable space which is countably generated and separates points (i.e. if \(x, y \in X\), there is an \(A \in \mathcal{A}\) with \(x \in A \) and \(y \notin A\)). Then there is a subset \(Y\) of \(\mathcal{C}\) such that \((X, \mathcal{A})\) is Borel isomorphic to \((Y, \mathcal{B}_Y)\).

**Proof:** Let \(\{A_n : n \in \mathbb{N}\}\) generate \(\mathcal{A}\) as a \(\sigma\)-algebra. For \(x \in X\), define \(f(x) \in \mathcal{C}\) by \(f(x)(n) = 1\) if \(x \in A_n\), \(f(x) = 0\) if \(x \notin A_n\). It is easily checked that \(f\) is a Borel isomorphism onto \(f(X)\). \(\Diamond\)

Such an \((X, \mathcal{A})\) is not necessarily a standard Borel space, even if uncountable (). But if the set \(Y\) is an uncountable \(G_{\delta}\) in \(\mathcal{C}\), then it follows from XI.10.3.3. that \((X, \mathcal{A})\) is a standard Borel space; We will extend this to show that \((X, \mathcal{A})\) is a standard Borel space if \(Y\) is merely an uncountable Borel set in \(\mathcal{C}\).

XII.3.1.8. **Theorem.** Let \((X, T)\) be a Polish space and \(A\) a Borel set in \((X, T)\). Then there is a Polish topology \(S\) on \(X\) which is stronger than \(T\), in which \(A\) is a clopen set, and such that every Borel set in \((X, S)\) is a Borel set in \((X, T)\).

**Proof:** Let \(\mathcal{A}\) be the set of all Borel subsets of \(X\) for which such an extension of \(T\) is possible. If \(F\) is closed in \(X\), let \(\mathcal{S} = \{U, U \cap F : U \in T\}\). Then \(\mathcal{S}\) is a topology on \(X\) stronger than \(T\), \(F\) is clopen in \(\mathcal{S}\), and \(\mathcal{S} \subseteq B_T\), so \(B_\mathcal{S} = B_T\). Also, \((X, \mathcal{S})\) is homeomorphic to \((F, T_F) \cup (F^c, T_{F^c})\), and \((F, T_F)\) and \((F^c, T_{F^c})\) are Polish by (), so \((X, \mathcal{S})\) is Polish. Thus all closed sets in \(X\) are in \(\mathcal{A}\). \(\mathcal{A}\) is clearly closed under complements.

To show that \(\mathcal{A}\) is closed under countable unions, suppose \((A_n)\) is a sequence in \(\mathcal{A}\). For each \(n\) let \(S_n\) be a Polish topology on \(X\) stronger than \(T\), with the same Borel sets, in which \(A_n\) is clopen. Let \(S_\infty\) be the topology on \(X\) generated by \(\bigcup_n S_n\); then \(S_\infty \subseteq B_T\), so \(B_{\mathcal{S}_\infty} = B_T\). If \(\sigma_n\) is a \([0, 1]\)-valued complete metric giving \(S_n\), then \(\sigma = \sum_n 2^{-n} \sigma_n\) is a complete metric giving \(S_\infty\). If \(D_n\) is a countable dense subset of \((X, S_n)\), then \(D = \bigcup_n D_n\) is a countable dense subset of \((X, S_\infty)\). Thus \((X, S_\infty)\) is Polish. We have that \(A = \bigcup_n A_n\) is open in \((X, S_\infty)\), so by the first part of the proof there is a Polish topology \(\mathcal{S}\) stronger than \(S_\infty\), with the same Borel sets, such that \(A\) is clopen in \((X, \mathcal{S})\). Thus \(A \in \mathcal{A}\) and \(\mathcal{A}\) is a \(\sigma\)-algebra containing \(T\), hence containing \(B_T\). \(\Diamond\)

XII.3.1.9. **Corollary.** An uncountable measurable subspace of a standard Borel space is a standard Borel space.

An interesting consequence is the following, which “settles” the Continuum Hypothesis for Borel sets:

XII.3.1.10. **Corollary.** Let \(X\) be a Polish space. Any uncountable Borel subset of \(X\) has cardinality \(2^{\aleph_0}\).

XII.3.2. **Analytic Sets**

XII.3.2.1. **Definition.** Let \(X\) be a Polish space. A subset \(Y\) of \(X\) is **analytic** if there is a continuous function from \(\mathcal{N}\) onto \(Y\).
XII.3.2.2. The official definition of an analytic set in a general topological space $X$ is a little different: it is a subset of $X$ of the form $\pi_X(F)$, where $F$ is a closed subset of $X \times \mathcal{N}$. If $Y \subseteq X$ and there is a continuous function $f$ from $\mathcal{N}$ onto $Y$, then $Y$ is analytic (let $F$ be the graph of $f$). Conversely, if $X$ is Polish and $Y = \pi_X(F)$ for a closed subset $F$ of $X \times \mathcal{N}$, then $F$ is Polish, so there is a continuous function $g$ from $\mathcal{N}$ onto $F$ by XI.10.2.7., and $\pi_X \circ g$ is a continuous function from $\mathcal{N}$ onto $Y$. Thus the two definitions are equivalent for $X$ Polish.

By XI.10.2.7., if $X$ is a Polish space, then $X$ itself is analytic in $X$; and the property of being analytic is independent of the containing space.

It is obvious from the definition that continuous images of analytic sets are analytic:

XII.3.2.3. Proposition. If $X$ and $Z$ are Polish spaces, $f : X \to Z$ is continuous, and $Y$ is an analytic subset of $X$, then $f(Y)$ is an analytic subset of $Z$.

One of the most important results about analytic sets is:

XII.3.2.4. Theorem. Let $X$ be a Polish space. The collection of analytic subsets of $X$ is closed under countable unions and countable intersections.

Proof: Closure under unions is fairly simple. Let $E_n$ be an analytic set for each $n$, and $f_n : \mathcal{N} \to E_n$ a surjective continuous function. We can regard $f_n$ as being defined on $\{n\} \times \mathcal{N} \subseteq \mathbb{N} \times \mathcal{N}$. Putting these together, we get a map $f$ from $\mathbb{N} \times \mathcal{N} \cong \mathcal{N}$ onto $\bigcup E_n$.

Closure under intersections is trickier, and cannot be done simply by taking complements because it is not clear (and in fact turns out to be false) that the complement of an analytic set is analytic. So we proceed as follows. Let $\psi$ be a homeomorphism from $\mathcal{N}$ onto $\mathcal{N}^{\mathbb{N}}$, with coordinate functions $\psi_n$. (This $\psi$ is an analog of the “ultimate space-filling curve” of ().) Let $E_n$ be an analytic subset of $X$ for each $n$ and $f_n$ a continuous function from $\mathcal{N}$ onto $E_n$. Let

$$Y = \{ s \in \mathcal{N} : f_n(\psi_n(s)) = f_m(\psi_m(s)) \text{ for all } n, m \in \mathbb{N} \}.$$ 

$Y$ is a closed subset of $\mathcal{N}$. For $s \in Y$, set $f(s) = f_1(\psi_1(s)) = f_n(\psi_n(s))$ for all $n \in \mathbb{N}$. We have that $f(s) \in f_n(\mathcal{N}) = E_n$ for all $n$, hence the range of $f$ is contained in $\bigcap E_n$. If $x \in \bigcap E_n$, then for each $n$ there is an $s_n \in \mathcal{N}$ with $f_n(s_n) = x$. Then there is an $s \in \mathcal{N}$ with $s_n = \psi_n(s)$ for all $n$; then $s \in Y$ and $f(s) = x$. So the range of $f$ is exactly $\bigcap E_n$. Since $Y$ is closed in $\mathcal{N}$, it is analytic, and hence $\bigcap E_n$ is analytic by XII.3.2.3.. $\diamondsuit$

XII.3.2.5. Corollary. Let $X$ be a Polish space. Then every Borel subset of $X$ is analytic.

Proof: Every closed subset of $X$ is a Polish space, hence analytic by XI.10.2.7.. The result then follows from XII.3.2.4. and ({}).

This result also follows easily from XII.3.1.8.: if $B$ is a Borel set in $X$, there is a Polish topology $\mathcal{S}$ on $X$ stronger than the given topology $\mathcal{T}$ for which $B$ is clopen. $B$ is thus Polish in the relative topology from $\mathcal{S}$; hence by XI.10.2.7. there is an $\mathcal{S}$-continuous map from $\mathcal{N}$ onto $B$, which is also $\mathcal{T}$-continuous.

1387
The converse of XII.3.2.5. is not quite true, since we will see that the complement of an analytic set is not necessarily analytic. But the converse is not “too far” from being true. To see this, we first show that disjoint analytic sets can be separated by Borel sets:

XII.3.2.6. Theorem. [Luzin Separation Theorem] Let $X$ be a Polish space, and $E$ and $F$ disjoint analytic subsets of $X$. Then there is a Borel subset $B$ of $X$ such that $E \subseteq B$ and $F \subseteq B^c$.

Proof: First note that if $(Y_n)$ and $(Z_m)$ are sequences of subsets of $X$ which can be separated by Borel sets, i.e. for each $n, m$ there is a Borel set $B_{n,m}$ with $Y_n \subseteq B_{n,m}$ and $Z_m \subseteq B_{n,m}^c$, then there is a Borel set $B$ separating $\cup_n Y_n$ and $\cup_m Z_m$: $B = \cup_n \cap_m B_{n,m}$ works.

Write $E = \phi(N)$ and $F = \psi(N)$ for continuous functions $\phi$, $\psi$. If $(k_1, \ldots, k_n) \in \mathbb{N}^n$, set

$$N_{k_1,\ldots,k_n} = \{(s_1, s_2, \ldots) \in N : s_j = k_j \text{ for } 1 \leq j \leq n\}.$$and put $E_{k_1,\ldots,k_n} = \phi(N_{k_1,\ldots,k_n})$, $F_{k_1,\ldots,k_n} = \psi(N_{k_1,\ldots,k_n})$.

Suppose $E$ and $F$ are not separated by a Borel set. Since $E = \cup_k E_k$ and $F = \cup_k F_k$, by the first part of the proof there must be $s_1, t_1$ such that $E_{s_1}$ and $F_{t_1}$ are not separated by a Borel set. Similarly, since $E_{s_1} = \cup_k E_{s_1,k}$ and $F_{t_1} = \cup_k F_{t_1,k}$, there must be $s_2, t_2$ such that $E_{s_2}$, $s_2$ and $F_{t_2}$, $t_2$ are not separated by Borel sets. Continuing inductively, sequences $s$ and $t$ in $N$ are constructed such that $E_{s_1,\ldots,s_n}$ and $F_{t_1,\ldots,t_n}$ cannot be separated by a Borel set. Let $x = \phi(s)$ and $y = \phi(t)$. Then $x \in E$ and $y \in F$, so $x \neq y$. Let $U$ and $V$ be disjoint neighborhoods of $x$ and $y$ in $X$. By continuity, we must have $E_{s_1,\ldots,s_n} = \phi(N_{s_1,\ldots,s_n}) \subseteq U$ and $F_{t_1,\ldots,t_n} = \psi(N_{t_1,\ldots,t_n}) \subseteq V$, so $U$ separates $E_{s_1,\ldots,s_n}$ and $F_{t_1,\ldots,t_n}$, a contradiction.

Note that no form of Choice is needed in this proof: each $s_j$ and $t_j$ can just be chosen to be the smallest one for which the conditions hold (specifically, choose $s_j$ to be the smallest possible for which there is a $t_j$, and then $t_j$ the smallest possible for this $s_j$).

XII.3.2.7. Corollary. Let $E$ be an analytic subset of a Polish space $X$. Then $E$ is a Borel set if and only if $E^c$ is also analytic.

XII.3.2.8. Example. It is not so easy to construct an analytic set which is not a Borel set. Here is a construction of an analytic subset $A$ of $N$ whose complement in $N$ is not analytic, hence $A$ is not Borel. Since $N$ is homeomorphic to the set of irrational numbers $J$, $A$ may be regarded as an analytic subset of $J$ whose complement in $J$ is not analytic (since if it were, $(J \setminus A) \cap J$ would be analytic by XII.3.2.4.). The construction uses a version of Cantor’s diagonalization argument.

We first construct a “universal” open set $W$ in $N^3$ which is a catalog for open sets in $N^2$, i.e. for every open set $V$ in $N^2$ there is a $t \in N$ such that $(r, s) \in V \iff (r, s, t) \in W$. The basic open sets in $N^2$ (products of sets of the form $N_{s_1,\ldots,s_n}$) can be arranged in a sequence $(U_n)$ (this can be done explicitly without any form of AC; cf. Exercise ??). For each $n \in \mathbb{N}$, define

$$W_n = \{(r, s, t) \in N^3 | (r, s) \in U_n\}$$where as usual we write $t = (t_1, t_2, \ldots)$. Then $W_n$ is open: if $(r, s, t) \in W_n$, then

$$\{(r', s', t') \in N^3 | (r', s') \in U_{t_n}, t_n = t_n\}$$
is a neighborhood of $(r, s, t)$ contained in $W_n$. Set $W = \bigcup_n W_n$. Then $W$ is open in $\mathcal{N}^3$. If $V$ is an open set in $\mathcal{N}^2$, let $t = (t_1, t_2, \ldots)$ be the set of $n$’s (in increasing order) such that $U_n \subseteq V$. Then $V = \bigcup_n U_{t_n}$, so

$$(r, s) \in V \iff (r, s) \in U_{t_n} \text{ for some } n \iff (r, s, t) \in W_n \text{ for some } n \iff (r, s, t) \in W$$

so $W$ is a catalog for open sets in $\mathcal{N}^2$.

Let $G = \mathcal{N}^3 \setminus W$. Then $G$ is a catalog for closed sets if $\mathcal{N}^2$: if $F \subseteq \mathcal{N}^2$ is closed, let $V = \mathcal{N}^2 \setminus F$. Then there is a $t \in \mathcal{N}$ such that

$$(r, s) \in F \iff (r, s) \notin V \iff (r, s, t) \notin W \iff (r, s, t) \in G.$$

Now let $C = \pi_{13}(G) \subseteq \mathcal{N}^2$ be the projection of $G$ onto the first and third coordinates. Then $C$ is an analytic subset of $\mathcal{N}^2$ since $G$ is closed in $\mathcal{N}^2 \times \mathcal{N}$. We claim that $C$ is a catalog for analytic subsets of $\mathcal{N}$. If $B$ is an analytic subset of $\mathcal{N}$, there is a closed set $F \subseteq \mathcal{N}^2$ with $B = \pi_1(F)$. There is then a $t \in \mathcal{N}$ such that $(r, s) \in F \iff (r, s, t) \in G$. We then have, for $r \in \mathcal{N}$,

$$r \in B \iff (r, s) \in F \text{ for some } s \iff (r, s, t) \in G \text{ for some } s \iff (r, t) \in C.$$

Finally, let $D = \{(t, t) | (t, t) \in C\} \subseteq \mathcal{N}^2$ and let $A = \pi_1(D)$. $D$ is the intersection of $C$ with the diagonal in $\mathcal{N}^2$, which is a closed, hence analytic, subset of $\mathcal{N}^2$; hence $D$ is analytic in $\mathcal{N}^2$ by XII.3.2.4., and so $A$ is analytic in $\mathcal{N}$. But $A^c = \mathcal{N} \setminus A$ cannot be analytic, since if it were there would be a $t \in \mathcal{N}$ such that $r \in A^c \iff (r, t) \in C$. But then, for this $t$,

$$t \in A^c \iff (t, t) \in C \iff (t, t) \in D \iff t \in A$$

which is a contradiction.

XII.3.3. Suslin Sets and the Suslin Operation

XII.3.3.1. There is a curious operation on sets called the Suslin operation, which is reminiscent of the construction in the proof of XI.10.2.7. For each finite sequence $(s_1, \ldots, s_n)$ of natural numbers, let $A_{s_1, \ldots, s_n}$ be a subset of a set $X$. The collection of finite sequences in $\mathbb{N}$ is countable, so there are countably many such sets in all. Define

$$S(\{A_{s_1, \ldots, s_n}\}) = \bigcup_{s \in \mathbb{N}} \bigcap_{n=1}^{\infty} A_{s_1, \ldots, s_n}.$$

Although the operation begins with a countable collection of sets, the union is over all sequences (uncountably many) in $\mathcal{N}$.

XII.3.3.2. Definition. Let $X$ be a Polish space. A Suslin set is a subset of $X$ obtained by applying the Suslin operation to a collection of closed subsets of $X$. 1389
XII.3.3.3. **Theorem.** A subset of a Polish space is a Suslin set if and only if it is analytic.

**Proof:** Suppose $A$ is a Suslin set in a Polish space $X$. Let $A = \bigcup_{x \in \mathcal{N}} \cap_n F_{s|n}$, where the $F_{s|n}$ are closed sets. Set

$$F = \{(x, s) \in X \times \mathcal{N} : x \in F_{s|n} \text{ for all } n\} \subseteq X \times \mathcal{N}.$$ 

Then $F$ is a closed subset of $X \times \mathcal{N}$, and $A = \pi_X(F)$, so $A$ is analytic.

Conversely, suppose $A$ is analytic. Fix a complete metric $\rho$ on $X$. Let $\phi : \mathcal{N} \to A$ be continuous. For each $(s_1, \ldots, s_n)$ set $F_{s_1, \ldots, s_n} = \phi_\phi(\mathcal{N}_{s_1, \ldots, s_n})$. Then $F_{s_1, \ldots, s_n}$ is closed and for each $s \in \mathcal{N}$, $F_{s|(n+1)} \subseteq F_{s|n}$ for all $n$, and $\text{diam}(F_{s|n}) \to 0$ as $n \to \infty$ since $\phi$ is continuous. Thus for every $s \in \mathcal{N}$ there is a unique $x_s \in \cap_n F_{s|n}$ since $X$ is complete. Fix $s$ and set $x = x_s$. Then for each $n$ there is an $x_n \in \phi(\mathcal{N}_{s_1, \ldots, s_n})$ with $\rho(x_n, x) < 2^{-n}$, and there is an $s(n) \in \mathcal{N}_{s_1, \ldots, s_n}$ with $\phi(s(n)) = x_n$. We have $s(n) \to s$ since $s(n)|n = s|n$ for all $n$; thus $\phi(s(n)) \to \phi(s)$. But $\phi(s(n)) = x_n \to x$, so $x = \phi(s)$, i.e. $x \in A$. So $S(\{F_{s_1, \ldots, s_n}\}) \subseteq A$. The reverse containment is obvious, so $A = S(\{F_{s_1, \ldots, s_n}\}) \in \mathcal{S}(\mathcal{F})$.  

XII.3.3.4. **Theorem.** Let $(X, \mathcal{A}, \mu)$ be a complete measure space. Then $\mathcal{A}$ is closed under the Suslin operation.

XII.3.3.5. **Corollary.** Let $X$ be a Polish space, and $\mu$ a complete measure defined on a $\sigma$-algebra $\mathcal{A}$ of subsets of $X$ containing the Borel sets. Then $\mathcal{A}$ contains all analytic subsets of $X$.

XII.3.3.6. The history of Suslin sets and the Suslin operation is unfortunate and somewhat murky, and is an important facet of the story of analysis in the Soviet Union. SUSLIN died of typhus in 1919 at age 24, having published only one paper during his lifetime. LUZIN, his mentor, then refined, expanded on, and published some of Suslin’s work (with credit to Suslin). Later, ALEKSEANDROV claimed that he was responsible for the Suslin operation and the notion of analytic set, and that in the original names “A-operation” and “A-set”, used by both Suslin and Luzin, the “A” stood for “Aleksandrov”, not “analytic”, the term eventually used by Luzin (Luzin said this name had been proposed by Lebesgue). There is some (inconclusive) evidence to support Aleksandrov’s claim, but also persuasive evidence to the contrary, and most contemporary scholarship attributes the concept of analytic set to Suslin and Luzin (Hausdorff also played a significant and rather independent role); cf. [Lor01]. The relations between Aleksandrov and Luzin began well but soured early, and contained a lot of personal friction and political intrigue during the Stalin era. Aleksandrov was an active participant in the attempted political persecution of Luzin known as the “Luzin affair.” Accounts of these and related events vary widely in point of view, and it is quite difficult to get an unbiased picture of what really happened. The story is certainly a complicated one, involving (according to references) politics, mysticism, and homosexuality. See [GK09], [ZD07], and [Sin03] for pieces of the story.

There are many fascinating anecdotes about these personalities in these references (many of which involve their fanatical and occasionally tragic swimming adventures). For example, the usually reserved Kolmogorov, a close friend of Aleksandrov, once slapped Luzin in the face during a meeting of the Academy of Sciences (according to [GK09], in response to an insult by Luzin); the incident came to the personal attention of Stalin, who shrugged it off, saying, “Well, such things occur with us as well.” It should be noted that both before and after this incident, Kolmogorov had kind things to say about Luzin, as did Aleksandrov earlier in his career.
XII.3.3.7. Universally measurable sets

XII.3.4. Addison’s Notation and the Projective Hierarchy

In 1958 J. Addison [Add59] introduced coherent new notation not only for the sets in the hierarchy of the construction of the Borel sets, but also for analytic sets and a continuing hierarchy of subsets of $\mathbb{R}$ (or a general topological space). Indexing by ordinals had been part of the subject since the work of Baire, Lebesgue, Hausdorff, and the Russian school.

XII.3.4.1. Fix a topological space $X$, usually $\mathbb{R}$ or $\mathcal{N}$. Define the classes $\Sigma^0_\alpha$ and $\Pi^0_\alpha$ of subsets of $X$ inductively for $0 < \alpha < \omega_1$ as follows:

- $\Sigma^0_0$ is the class of open sets.
- $\Pi^0_0$ is the class of closed sets.
- $\Sigma^0_\alpha$ is the class of countable unions of sets in $\bigcup_{\beta < \alpha} \Pi^0_\beta$.
- $\Pi^0_\alpha$ is the class of countable intersections of sets in $\bigcup_{\beta < \alpha} \Sigma^0_\beta$.

Thus $\Sigma^0_2$ is the class of $F_\sigma$'s, $\Pi^0_2$ the class of $G_\delta$'s, $\Sigma^0_3$ the class of $G_{\delta\sigma}$'s, $\Pi^0_3$ the class of $F_{\sigma\delta}$'s, etc.

A set $A$ is called a $\Sigma^0_\alpha$-set (or $\Pi^0_\alpha$-set) if $A \in \Sigma^0_\alpha$ (or $A \in \Pi^0_\alpha$).

XII.3.4.2. If $X$ is Polish, we have that $\Sigma^0_\alpha \subseteq \Pi^0_\beta$ and $\Pi^0_\beta \subseteq \Sigma^0_\alpha$ if $\beta < \alpha$. The complement of a $\Sigma^0_\alpha$-set is a $\Pi^0_\alpha$-set, and vice versa. Since $\Sigma^0_1 \subseteq \Sigma^0_0$ ( ), by induction we have $\Sigma^0_\beta \subseteq \Sigma^0_\alpha$ and $\Pi^0_\alpha \subseteq \Pi^0_\beta$ for $\beta < \alpha$.

The unions $\bigcup_{\alpha < \omega_1} \Sigma^0_\alpha$ and $\bigcup_{\alpha < \omega_1} \Pi^0_\alpha$ are both equal to the class of Borel sets (the Countable AC is needed to show that these unions form a $\sigma$-algebra and hence contain all Borel sets (XII.3.5.4.)).

XII.3.4.3. It follows easily from the results of ( ) that for $X = \mathbb{R}$, these classes strictly increase in $\alpha$. This was first shown by Lebesgue [?], who proved that $\Sigma^0_{\alpha} \nsubseteq \Pi^0_{\alpha}$ for any $\alpha < \omega_1$ (see ( )).

We also define $\Delta^0_{\alpha} = \Sigma^0_{\alpha} \cap \Pi^0_{\alpha}$.

XII.3.4.4. Now for $n \in \mathbb{N} \cup \{0\}$ define $\Sigma^1_n$ and $\Pi^1_n$ inductively as follows.

- $\Sigma^0_0 = \Sigma^1_0$ is the class of open sets.
- $\Pi^0_0$ is the class of complements of $\Sigma^1_0$-sets.
- $\Sigma^1_n$ is the class of countable unions of $\Pi^1_n$-sets in $X \times \mathcal{N}$.
- $\Delta^1_n = \Sigma^1_n \cap \Pi^1_n$.

Thus $\Pi^1_0$ is the class of closed sets, and $\Sigma^1_1$ is the class of analytic sets ( ). $\Delta^1_1$ is the class of Borel sets ( ). It follows that $\Sigma^1_0 \cup \Pi^1_0 \subseteq \Delta^1_1$ (there is also a simple direct proof of this) and hence by induction $\Sigma^1_n \cup \Pi^1_n \subseteq \Delta^1_{n+1}$ for all $n$.

A set in $\bigcup_n \Sigma^1_n = \bigcup_n \Pi^1_n = \bigcup_n \Delta^1_n$ is called a projective set.
To see why the projection operation is natural, and that this hierarchy is analogous to the Borel sets, note that an $F_\sigma$ ($\Sigma^0_2$-set) in $X$ is a $\pi_X$ projection of a closed set ($\Pi^0_1$-set) in $X \times \mathbb{N}$, and more generally a $\Sigma^0_{\alpha+1}$-set in $X$ is the $\pi_X$ projection of a $\Pi^0_{\alpha}$-set in $X \times \mathbb{N}$.

This hierarchy can be extended transfinitely to define $\Sigma^1_\alpha$ and $\Pi^1_\alpha$ for $\alpha < \omega_1$. What about $\Sigma^n_\alpha$, $\Sigma^2_\alpha$, etc.?

ADDISON’s notation is strongly reminiscent of similar notation used in mathematical logic and model theory, and was in fact inspired by analogies and connections with the logic notation.

**Exercises**

**XII.3.5.1.** Let $\mathcal{C} = \{0,1\}^\mathbb{N}$ ($\mathcal{C}$ is homeomorphic to the Cantor set (\)). Define a map $f : \mathcal{N} \to \mathcal{C}$ by

\[
f(n_1, n_2, \ldots) = (0, \ldots, 0, 1, 0, \ldots, 0, 1, \ldots).
\]

(a) Show that $f$ is a homeomorphism of $\mathcal{N}$ with a subset of $\mathcal{C}$ with countable complement.

(b) Show that $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to $(\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$.

(c) Since $\mathcal{N}$ is also homeomorphic to the irrational numbers in $[0,1]$, $([0,1], \mathcal{B}_{[0,1]})$ is Borel isomorphic to $(\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Thus $([0,1], \mathcal{B}_{[0,1]})$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

(d) Show that $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is also Borel isomorphic to $([0,1]^\mathbb{N}, \mathcal{B}_{[0,1]})$. $[\mathcal{C}$ is homeomorphic to $C^{\mathbb{N}}\]$.

(c) Since $(\mathbb{N}, \mathcal{P}(\mathbb{N})) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ is Borel isomorphic to $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, $(\mathcal{C}, \mathcal{B}_\mathcal{C}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$. [Use (b).]

**XII.3.5.2.** If $(X, \rho)$ is a perfect complete metric space, show that there is a homeomorphism from $\mathcal{C}$ onto a (compact) subspace of $X$. Thus $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to a measurable subspace of $(X, \mathcal{B}_X)$.

(a) Let $X$ be a Polish space. Then $X$ is homeomorphic to a $G_\delta$ in the Hilbert cube (\(\)) and (\(\)). Thus $(X, \mathcal{B}_X)$ is Borel isomorphic to a measurable subspace of $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ by (\(\)) (d).

(b) If $X$ is an uncountable Polish space, then by (\(\)) $(X, \mathcal{B}_X)$ is Borel isomorphic to $(Y, \mathcal{B}_Y) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ for some perfect (uncountable) Polish space $Y$. Thus $(X, \mathcal{B}_X)$ is Borel isomorphic to a measurable subspace of $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

(c) If $X$ is an uncountable Polish space, then $X$ has a (compact) subspace homeomorphic to $C$ (\(\)) and (\(\)), and thus $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to a measurable subspace of $(X, \mathcal{B}_X)$.

(d) Apply the Schröder-Bernstein Theorem for measurable spaces (XII.1.5.3.) to conclude that if $X$ is any uncountable Polish space, then $(X, \mathcal{B}_X)$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

**XII.3.5.4.** Explain why the Countable AC is needed to show that the unions $A = \bigcup_{\alpha < \omega_1} \Sigma^0_\alpha = \bigcup_{\alpha < \omega_1} \Pi^0_\alpha$ (XII.3.4.2.) form a $\sigma$-algebra. (See II.8.6.10..)
Chapter XIII

Measure Theory

XIII.1. Measures

It should come as no surprise that the fundamental concept of measure theory is the notion of measure. Measures live on measurable spaces \((X, \mathcal{A})\), and assign to each measurable set a nonnegative extended real-number "size" for the set. The crucial property a measure must have is countable additivity.

XIII.1.1. Definitions

XIII.1.1.1. Definition. Let \((X, \mathcal{A})\) be a measurable space. A measure on \((X, \mathcal{A})\) is a function \(\mu : \mathcal{A} \to [0, \infty]\) satisfying

(i) \(\mu(\emptyset) = 0\).

(ii) If \(A_1, A_2, \ldots\) is a sequence of pairwise disjoint sets in \(\mathcal{A}\), then

\[
\mu \left( \bigcup_{k=1}^{\infty} A_k \right) = \sum_{k=1}^{\infty} \mu(A_k) = \sup_{n} \sum_{k=1}^{n} \mu(A_k) = \lim_{n \to \infty} \sum_{k=1}^{n} \mu(A_k)
\]

(\(\mu\) is countably additive).

A measure space is a triple \((X, \mathcal{A}, \mu)\), where \((X, \mathcal{A})\) is a measurable space and \(\mu\) is a measure on \((X, \mathcal{A})\).

XIII.1.1.2. Note that specification of the \(\sigma\)-algebra is part of the definition of a measure and of a measure space. While we frequently downplay this specification, especially in contexts where there is some natural \(\sigma\)-algebra of measurable sets, it cannot be ignored. Thus, for example, if \((X, \mathcal{A}, \mu)\) is a measure space and \(\mathcal{B}\) is a proper sub-\(\sigma\)-algebra of \(\mathcal{A}\), and \(\nu = \mu|_{\mathcal{B}}\), then \(\nu\) is technically a different measure than \(\mu\), and the measure spaces \((X, \mathcal{A}, \mu)\) and \((X, \mathcal{B}, \nu)\) are different measure spaces. See XIII.1.6.10.

XIII.1.1.3. Condition (i) in the definition is almost redundant: the only function satisfying (ii) but not (i) is the function \(\mu\) with \(\mu(A) = \infty\) for all \(A \in \mathcal{A}\). (To see this, let \(A \in \mathcal{A}\), set \(A_1 = A\) and \(A_n = \emptyset\) for \(n > 1\), and apply (ii).)

By a similar argument, we can see:
XIII.1.1.4. Proposition. Let $\mu$ be a measure on $(X, \mathcal{A})$. Then $\mu$ is finitely additive: if $A_1, \ldots, A_n$ are pairwise disjoint sets in $\mathcal{A}$, then
\[
\mu \left( \bigcup_{k=1}^n A_k \right) = \sum_{k=1}^n \mu(A_k).
\]
In particular, if $A, B \in \mathcal{A}$ are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$.

**Proof:** Set $A_k = \emptyset$ for $k > n$, and apply countable additivity.

XIII.1.1.5. We can define a finitely additive measure on a measurable space $(X, \mu)$ to be a finitely additive function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. (Actually we only need for $\mathcal{A}$ to be an algebra of subsets of $X$ for this definition to make sense.) XIII.1.1.4. says that every measure is a finitely additive measure; but there are many finitely additive measures which are not measures (e.g. XIII.1.1.6.(vii)–(viii)). A few authors call a finitely additive measure a “measure” (measures in our sense are then called “countably additive measures”), but this terminology is nonstandard.

The theory of finitely additive measures is important in some contexts, and the theory of measure and integration can be developed surprisingly far in this generality (cf. [DS88, Chapter 3]). However, countable additivity is really what gives the theory of measure and integration its power. Countable additivity should be regarded as a sort of continuity condition (cf. XIII.1.2.7.).

Although we cannot yet give any really interesting examples of measures (the interesting ones take considerable effort to construct), here are some simple examples which range from uninteresting to moderately interesting:

XIII.1.1.6. Examples. (i) Let $(X, \mathcal{A})$ be any measurable space. Set $\mu(A) = 0$ for all $A \in \mathcal{A}$. Then $\mu$ is a measure on $(X, \mathcal{A})$, called the zero measure.

(ii) Let $X$ be any set. For $A \subseteq X$, set $\mu(A) = 0$ if $A = \emptyset$ and $\mu(A) = \infty$ if $A \neq \emptyset$. Then $\mu$ is a measure on $(X, \mathcal{P}(X))$. (This measure can be restricted to any $\sigma$-algebra on $X$.)

(iii) Let $X$ be an uncountable set. If $A \subseteq X$, set $\mu(A) = 0$ if $A$ is countable and $\mu(A) = \infty$ if $A$ is uncountable. Then $\mu$ is a measure on $(X, \mathcal{P}(X))$.

(iv) Let $X$ be an uncountable set, $\mathcal{A}$ the $\sigma$-algebra of countable/cocountable subsets of $X$. If $A \in \mathcal{A}$, set $\mu(A) = 0$ if $A$ is countable and $\mu(A) = 1$ if $A$ is cocountable. Then $\mu$ is a measure on $(X, \mathcal{A})$.

(v) Let $X$ be any set, and $p$ a fixed element of $X$. For $A \subseteq X$, set $\delta_p(A) = 1$ if $p \in A$ and $\delta_p(A) = 0$ if $p \notin A$. Then $\delta_p$ is a measure on $(X, \mathcal{P}(X))$, called the (unit) point mass (or Dirac measure) at $p$.

(vi) Let $X$ be any set. If $A \subseteq X$, set $\mu(A) = n$ if $A$ is a finite set with $n$ elements, and $\mu(A) = \infty$ if $A$ is infinite. Then $\mu$ is a measure on $(X, \mathcal{P}(X))$, called counting measure on $X$. This measure, while very simple, is important, especially in the case $X = \mathbb{N}$; integration theory for this measure is really exactly the theory of infinite series.

(vii) Let $X$ be an infinite set. If $A \subseteq X$, set $\mu(A) = 0$ if $A$ is finite and $\mu(A) = \infty$ if $A$ is infinite. Then $\mu$ is a finitely additive measure on $(X, \mathcal{P}(X))$ which is not a measure.

(viii) Let $\mathcal{F}$ be a free ultrafilter () on $\mathbb{N}$. For $A \subseteq \mathbb{N}$, set $\mu(A) = 1$ if $A \in \mathcal{F}$ and $\mu(A) = 0$ if $A \notin \mathcal{F}$. Then $\mu$ is a finitely additive measure on $(X, \mathcal{P}(X))$ which is not a measure. The same construction works if $\mathbb{N}$ is replaced by any infinite set whose cardinality is less than the first measurable cardinal.
XIII.1.2. Elementary Properties

Here we develop some of the most basic properties of measures.

XIII.1.2.1. Proposition. Let \((X, \mathcal{A}, \mu)\) be a measure space. Then

(i) \(\mu\) is monotone: if \(A, B \in \mathcal{A}\), \(B \subseteq A\), then \(\mu(B) \leq \mu(A)\).

(ii) \(\mu\) is countably subadditive: if \((A_k)\) is a sequence in \(\mathcal{A}\), then

\[
\mu \left( \bigcup_{k=1}^{\infty} A_k \right) \leq \sum_{k=1}^{\infty} \mu(A_k).
\]

Proof: (i): We have \(A = B \cup (A \setminus B)\), a disjoint union, and \((A \setminus B) \in \mathcal{A}\), so

\[\mu(A) = \mu(B) + \mu(A \setminus B)\]

by finite additivity, and \(\mu(A \setminus B) \geq 0\).

(ii): Disjointize the sequence \((A_n)\) as in (i), i.e. set \(B_1 = A_1\) and \(B_n = A_n \setminus (\bigcup_{k=1}^{n-1} A_k)\) for \(n > 1\). Then the \(B_n\) are in \(\mathcal{A}\) and are pairwise disjoint, \(B_n \subseteq A_n\), and \(\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n\). So, by countable additivity and (i),

\[
\mu \left( \bigcup_{n=1}^{\infty} A_n \right) = \mu \left( \bigcup_{n=1}^{\infty} B_n \right) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n).
\]

As with XIII.1.1.4., we also have

XIII.1.2.2. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space. Then \(\mu\) is finitely subadditive: if \(A_1, \ldots, A_n \in \mathcal{A}\), then

\[
\mu \left( \bigcup_{k=1}^{n} A_k \right) \leq \sum_{k=1}^{n} \mu(A_k).
\]

In particular, if \(A, B \in \mathcal{A}\), then \(\mu(A \cup B) \leq \mu(A) + \mu(B)\).

XIII.1.2.3. Note that by a nearly identical proof, a finitely additive measure is also monotone and finitely subadditive. A finitely additive measure need not be countably subadditive: the measure in XIII.1.1.6.(vii) is a counterexample. In fact, a finitely additive measure which is countably subadditive is countably additive, i.e. a measure (Exercise XIII.1.9.1.).

Decompositions of Measure Spaces

The next observation is an immediate consequence of countable additivity. We will often use this result to break up a measure space into a finite or countable number of pieces.
XIII.1.2.4. PROPOSITION. Let \((X, \mathcal{A}, \mu)\) be a measure space, \(I\) a finite or countable index set, and \(\{X_i : i \in I\}\) a collection of pairwise disjoint measurable subsets of \(X\) with \(X = \bigcup_{i \in I} X_i\). Then for any \(A \in \mathcal{A}\) we have\[
\mu(A) = \sum_{i \in I} \mu(A \cap X_i) .
\]

XIII.1.2.5. Note that countability of the index set \(I\) is essential. In many examples such as in XIII.1.1.6.(iii), or \(\mathbb{R}\) with Lebesgue measure, the entire space \(X\) can be decomposed as a disjoint union of singleton subsets, each of which is measurable and has measure zero.

Measures of Differences

XIII.1.2.6. If \((X, \mathcal{A}, \mu)\) is a measure space, and \(A, B \in \mathcal{A}\) with \(B \subseteq A\), and \(\mu(B) < \infty\) (e.g. if \(\mu(A) < \infty\)), then we have that\[
\mu(A \setminus B) = \mu(A) - \mu(B)
\]where we interpret \(-b = \infty\) for \(b \in \mathbb{R}\). However, if \(\mu(B) = \infty\) (and hence \(\mu(A) = \infty\) also by monotonicity), we cannot draw any conclusion about \(\mu(A \setminus B)\) (cf. Exercise XIII.1.9.2.). Thus we must be careful in subtracting measures of sets when the measures are infinite.

Upward Continuity

The next property is the first instance of how countable additivity can be regarded as a continuity property. It is called the Baby Monotone Convergence Theorem because it is actually the special case of the Monotone Convergence Theorem for indicator functions.

XIII.1.2.7. PROPOSITION. [BABY MONOTONE CONVERGENCE THEOREM] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((A_n)\) an increasing sequence in \(\mathcal{A}\), i.e. \(A_n \subseteq A_{n+1}\) for all \(n\). Then\[
\mu \left( \bigcup_{n=1}^{\infty} A_n \right) = \sup_n \mu(A_n) = \lim_{n \to \infty} \mu(A_n)
\](\(\mu\) is upward continuous).

PROOF: Disjointize the sequence \((A_n)\) by setting \(B_1 = A_1\) and \(B_n = A_n \setminus A_{n-1}\) for \(n > 1\). Then the \(B_n\) are in \(\mathcal{A}\) and are pairwise disjoint, \(A_n = \bigcup_{k=1}^{n} B_k\) for all \(n\), and \(\bigcup_{n=1}^{\infty} A_n = \bigcup_{k=1}^{\infty} B_k\). Then\[
\mu \left( \bigcup_{n=1}^{\infty} A_n \right) = \mu \left( \bigcup_{k=1}^{\infty} B_k \right) = \sum_{k=1}^{\infty} \mu(B_k) = \lim_{n \to \infty} \sum_{k=1}^{n} \mu(B_k) = \lim_{n \to \infty} \mu(A_n) .
\]
XIII.1.2.8. A finitely additive measure $\mu$ is not upward continuous in general, although if $(A_n)$ is an increasing sequence of measurable sets we always have

$$\mu \left( \bigcup_{n=1}^{\infty} A_n \right) \geq \sup_n \mu(A_n) = \lim_{n \to \infty} \mu(A_n)$$

by monotonicity. Example XIII.1.1.6.(vii) is a counterexample to equality. In fact, it is easily seen that upward continuity is equivalent to countable additivity for a finitely additive measure (Exercise XIII.1.9.1.).

**Downward Continuity**

What about downward continuity? The following example shows that it does not hold in general.

XIII.1.2.9. **Example.** Let $\mu$ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Let $A_n = \{ k \in \mathbb{N} : k \geq n \}$. Then $A_n \supseteq A_{n+1}$ for all $n$. We have $\mu(A_n) = \infty$ for all $n$. But $\cap_{n=1}^{\infty} A_n = \emptyset$, so

$$0 = \mu \left( \bigcap_{n=1}^{\infty} A_n \right) < \inf_n \mu(A_n) = \lim_{n \to \infty} \mu(A_n) = \infty .$$

The difficulty stems from the failure of subtraction for measures of sets (cf. XIII.1.2.6.). In fact, we have:

XIII.1.2.10. **Proposition.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and $(A_n)$ a decreasing sequence in $\mathcal{A}$, i.e. $A_n \supseteq A_{n+1}$ for all $n$. If $\mu(A_1) < \infty$, then

$$\mu \left( \bigcap_{n=1}^{\infty} A_n \right) = \inf_n \mu(A_n) = \lim_{n \to \infty} \mu(A_n)$$

($\mu$ is **downward continuous** on sets of finite measure).

**Proof:** Set $B_n = A_1 \setminus A_n$ for each $n$. Then the $B_n$ are in $\mathcal{A}$ and are increasing. If $A = \cap_{n=1}^{\infty} A_n$ and $B = \cup_{n=1}^{\infty} B_n$, we have $B = A_1 \setminus A$. We thus have, using XIII.1.2.6. and XIII.1.2.7.,

$$\mu(A_1) - \mu(A) = \mu(B) = \lim_{n \to \infty} \mu(B_n) = \lim_{n \to \infty} [\mu(A_1) - \mu(A_n)] = \mu(A_1) - \lim_{n \to \infty} \mu(A_n)$$

and the result follows.

Note that, by passing to a tail of the sequence, it suffices that $\mu(A_n) < \infty$ for some $n$.

The First Borel-Cantelli Lemma

We can obtain various more sophisticated results about measures. For example, the next result is particularly important in probability, where it is called the **first Borel-Cantelli Lemma**. Recall () that if $(A_n)$ is a sequence of subsets of a set $X$, then

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \left( \bigcup_{k=n}^{\infty} A_k \right)$$

is the set of $x \in X$ which are in infinitely many of the $A_n$. 

1397
XIII.1.2.11. Proposition. [First Borel-Cantelli Lemma] Let \((X, \mathcal{A}, \mu)\) be a measure space, and 
\((A_n)\) a sequence in \(\mathcal{A}\). If
\[
\sum_{k=1}^{\infty} \mu(A_k) < \infty,
\]
then
\[
\mu(\limsup_{n} A_n) = 0.
\]

Proof: Let \(\epsilon > 0\). Then there is an \(m\) such that
\[
\sum_{k=1}^{\infty} \mu(A_k) - \sum_{k=1}^{m} \mu(A_k) = \sum_{k=m+1}^{\infty} \mu(A_k) < \epsilon
\]
(in fact this holds for all sufficiently large \(m\)). For such an \(m\), we have
\[
\limsup_{n} A_n \subseteq \bigcup_{k=m+1}^{\infty} A_k
\]
so by monotonicity and countable subadditivity we have
\[
\mu(\limsup_{n} A_n) \leq \mu \left( \bigcup_{k=m+1}^{\infty} A_k \right) \leq \sum_{k=m+1}^{\infty} \mu(A_k) < \epsilon.
\]
Since \(\epsilon > 0\) is arbitrary, \(\mu(\limsup_{n} A_n) = 0\).

XIII.1.3. Null Sets and Completeness

XIII.1.3.1. Definition. Let \((X, \mathcal{A}, \mu)\) be a measure space. A set \(A \in \mathcal{A}\) is a null set (or \(\mu\)-null set) if
\[
\mu(A) = 0.
\]

XIII.1.3.2. In general mathematics, the term "null set" is often used as a synonym for "empty set"; but in analysis it is more useful to use the definition we have given. Note that the empty set is always a null set by the analysis definition, but many measure spaces also have other null sets which are nonempty. Of course, whether a set is a null set depends on the measure: if \(\mu\) and \(\nu\) are measures on a measurable space \((X, \mathcal{A})\), then a \(\mu\)-null set is not necessarily a \(\nu\)-null set.

The idea behind the terminology is that null sets are "negligibly small" for measure-theoretic purposes. A general philosophy in measure theory is that sets or functions can be modified on a null set with no effect. In fact, we will eventually feel quite free to ignore what happens on null sets, although at least at first one should be careful about this.

Unions of Null Sets

XIII.1.3.3. Proposition. A countable union of null sets is a null set. (In particular, a finite union of null sets is a null set.)

This is an immediate consequence of countable subadditivity.
Subsets of Null Sets and Completeness

XIII.1.3.4. It may be surprising that a subset of a null set is not necessarily a null set. It is a trivial consequence of monotonicity that any measurable subset of a null set is a null set. However, there are many examples of measure spaces containing null sets with nonmeasurable subsets. Here is perhaps the simplest example:

XIII.1.3.5. Example. Let $\mathcal{A}$ be the $\sigma$-algebra of countable/cocountable subsets of $\mathbb{R}$. If $A \in \mathcal{A}$, let $\mu(A) = 1$ if $1 \in A$ and $\mu(A) = 0$ if $1 \notin A$. It is easily checked that $\mu$ is a measure on $(\mathbb{R}, \mathcal{A})$ (it is the restriction to $\mathcal{A}$ of the Dirac measure $\delta_1$ of XIII.1.3.7. (v)). We have that $\mathbb{R} \setminus \{1\}$ is a null set, but $(1, +\infty)$ and $(-\infty, 1)$ are not null sets since they are not measurable, i.e., not in $\mathcal{A}$ (each is an uncountable set with uncountable complement).

This example may seem artificial since we arbitrarily chose a smaller $\sigma$-algebra than the natural one on which $\mu$ is defined (namely $\mathcal{P}(\mathbb{R})$). There are more natural examples such as Lebesgue measure on the Borel sets $\mathcal{B}$, although in a certain sense every example is unnatural (XIII.1.3.7.).

XIII.1.3.6. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space. Then $(X, \mathcal{A}, \mu)$ is a complete measure space if every subset of a null set is a null set (i.e., measurable).

XIII.1.3.7. Theorem. Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $\mathcal{A}$ be the $\sigma$-algebra of subsets of $X$ generated by $\mathcal{A}$ and all subsets of null sets. Then there is a unique extension $\tilde{\mu}$ of $\mu$ to $(X, \mathcal{A})$, and $(X, \mathcal{A}, \tilde{\mu})$ is complete. In fact, $\mathcal{A}$ is the collection of all subsets of $X$ of the form $A \cup D$, where $A \in \mathcal{A}$ and $D$ is a subset of a null set.

Proof: Let $\mathcal{B}$ be the collection of all subsets of $X$ of the form $A \cup D$, where $A \in \mathcal{A}$ and $D$ is a subset of a null set. Note that $\emptyset$ and $X$ are in $\mathcal{B}$, and it is easy (using XIII.1.3.3.) to check that $\mathcal{B}$ is closed under countable unions. Let $B = A \cup D \in \mathcal{B}$, where $A \in \mathcal{A}$ and $D \subseteq C \subseteq A$ with $\mu(C) = 0$. Then

$$X \setminus B = (X \setminus [A \cup C]) \cup (C \setminus B)$$

and $X \setminus (A \cup C) \in \mathcal{A}$, $C \setminus B \subseteq C$. Thus $\mathcal{B}$ is a $\sigma$-algebra and $A \subseteq B \subseteq \mathcal{A}$; $\mathcal{B}$ contains all subsets of null sets, so $\mathcal{B} = \mathcal{A}$.

For $B = A \cup D$ as above, set $\tilde{\mu}(B) = \mu(A)$. This is well defined: if $B = A_1 \cup D_1 = A_2 \cup D_2$ with $D_j \subseteq C_j$, $\mu(C_j) = 0$, then

$$A_1 \setminus C_2 \subseteq A_2 \subseteq A_1 \cup C_1$$

so $\mu(A_1) = \mu(A_2)$. It is clear that $\tilde{\mu}$ extends $\mu$ and easy to check that $\tilde{\mu}$ is countably additive. For uniqueness, if $\nu$ extends $\mu$ and $B = A \cup D$ with $D \subseteq C$, $\mu(C) = 0$, we must have

$$\mu(A) = \nu(A) \leq \nu(B) \leq \nu(A \cup C) = \mu(A \cup C) \leq \mu(A) + \mu(C) = \mu(A)$$

by monotonicity and subadditivity. Since every subset of a $\tilde{\mu}$-null set is a subset of a $\mu$-null set and hence in $\mathcal{A}$, $(X, \mathcal{A}, \tilde{\mu})$ is complete.

\hfill \Box
XIII.1.3.8. Definition. The measure space \((X, \mathcal{A}, \mu)\) is called the completion of \((X, \mathcal{A}, \mu)\) and, by slight abuse of terminology, \(\bar{\mu}\) is called the completion of \(\mu\).

Almost Everywhere

XIII.1.3.9. It is convenient to use the phrase almost everywhere, abbreviated a.e., to mean “except on a null set.” Thus, for example, if we say “the functions \(f\) and \(g\) agree almost everywhere,” we mean “there is a null set \(A\) such that \(f(x) = g(x)\) for all \(x \notin A\).” When the measure \(\mu\) needs to be specified, we say “\(\mu\)-almost everywhere,” written \(\mu\)-a.e.

However, if we are dealing with an incomplete measure space, there is a subtlety. Saying “the functions \(f\) and \(g\) agree almost everywhere” does not necessarily mean “the set of all \(x\) for which \(f(x) = g(x)\) does not hold (i.e. one or both of \(f(x)\) and \(g(x)\) is undefined, or both are defined but unequal) has measure zero,” because this set could be nonmeasurable (it is only a subset of a set of measure 0). In fact, if \(f\) and \(g\) agree almost everywhere, and \(f\) is measurable, then \(g\) is not necessarily even measurable (consider the case where \(f\) is identically zero and \(g\) is the indicator function of a nonmeasurable subset of a null set). This difficulty disappears when working with complete measure spaces, and is one reason for completing measure spaces (although taking completions also introduces different subtleties; cf. (i) or (ii)).

XIII.1.3.10. We can also say that two measurable sets \(A\) and \(B\) in a measure space \((X, \mathcal{A}, \mu)\) are almost disjoint (or \(\mu\)-almost disjoint) if \(A \cap B\) is a null set.

The next technical result is very useful and is often applied in the later development of measure and integration:

XIII.1.3.11. Proposition. Let \((X, \mathcal{A}, \mu)\) be a measure space, with completion \((X, \bar{\mathcal{A}}, \bar{\mu})\). If \(f\) is an extended real-valued \(\mathcal{A}\)-measurable function, then there is an \(\mathcal{A}\)-measurable function \(g\) with \(g = f\ \mu\)-almost everywhere.

Proof: The proof is a good example of the technique of bootstrapping, proving something for all measurable functions by first proving it for indicator functions, then simple functions, and finally general functions by taking pointwise limits. In this case, suppose \(f = \chi_C\), where \(C \in \mathcal{A}\). Then \(C = A \cup B\), where \(A \in \mathcal{A}\) and \(B \subseteq D\) for some \(\mu\)-null set \(D \in \mathcal{A}\). Then \(\chi_A = \chi_C\) on \(X \setminus B\) and \(a fortiori\) on \(X \setminus D\), and \(\chi_A\) is \(\mathcal{A}\)-measurable. Next, if \(f_1, \ldots, f_n\) are \(\mathcal{A}\)-measurable functions equal a.e. to \(\mathcal{A}\)-measurable functions \(g_1, \ldots, g_n\) respectively, and if \(\alpha_1, \ldots, \alpha_n \in \mathbb{R}\), then \(\alpha_1 f_1 + \cdots + \alpha_n f_n\) is equal a.e. to the \(\mathcal{A}\)-measurable function \(\alpha_1 g_1 + \cdots + \alpha_n g_n\) since a finite union of null sets is a null set. In particular, any \(\mathcal{A}\)-measurable simple function is equal a.e. to an \(\mathcal{A}\)-measurable simple function.

Now let \(f\) be a general \(\mathcal{A}\)-measurable function. There is a sequence \((f_n)\) of \(\mathcal{A}\)-measurable simple functions converging pointwise to \(f\). For each \(n\), there is a \(\mu\)-null set \(D_n\) and an \(\mathcal{A}\)-measurable simple function \(g_n\) such that \(f_n = g_n\) on \(X \setminus D_n\). Let \(D = \bigcup_{n=1}^{\infty} D_n\). Then \(D\) is a \(\mu\)-null set. Let \(h_n = g_n = f_n\) on \(X \setminus D\) and \(h_n = 0\) on \(D\). Then \(h_n\) is \(\mathcal{A}\)-measurable, and \((h_n)\) converges pointwise to a function \(h\) which is \(\mathcal{A}\)-measurable and agrees with \(f\) on \(X \setminus D\) (in fact, \(h_n = g_n \chi_{X \setminus D} = f_n \chi_{X \setminus D}\) and \(h = f \chi_{X \setminus D}\)).

14000
XIII.1.4. Finite, $\sigma$-Finite, and Semifinite Measures

Sets of infinite measure cause difficulties in measure theory. We cannot simply wish them away, however, since many of the most natural and important measure spaces (e.g. $\mathbb{R}$ with Lebesgue measure) have sets of infinite measure. We do want to work with sets of finite measure as much as possible, however. In this subsection we set up some terminology and develop a framework for doing this.

Finite Measure Spaces

The nicest measure spaces are ones with no sets of infinite measure:

XIII.1.4.1. Definition. A measure space $(X, \mathcal{A}, \mu)$ is finite, and the measure $\mu$ is finite, if $\mu(X) < \infty$. The number $\mu(X)$ is the total mass of the measure space. The measure space is a probability measure space, and $\mu$ is a probability measure, if $\mu(X) = 1$.

The term “probability measure” comes of course from probability theory, but the term is used throughout analysis even in contexts unrelated to probability. Any (nonzero) finite measure can be renormalized to a probability measure by taking an appropriate scalar multiple.

XIII.1.4.2. If $(X, \mathcal{A}, \mu)$ is a finite measure space with total mass $m$, then $\mu(A) \leq m$ for any $A \in \mathcal{A}$ by monotonicity. In particular, if $\mu$ is a probability measure, then $\mu(A) \leq 1$ for all $A \in \mathcal{A}$.

$\sigma$-Finite Measures

It turns out to be almost as good for a set to be a countable union of sets of finite measure, and measure spaces which can be decomposed into countably many finite pieces are well behaved.

XIII.1.4.3. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space. A set $A \in \mathcal{A}$ is of $\sigma$-finite measure if there is a sequence $(A_n)$ of sets in $\mathcal{A}$ with $A = \bigcup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for all $n$. The measure space $(X, \mathcal{A}, \mu)$ is $\sigma$-finite, and $\mu$ is a $\sigma$-finite measure, if $X$ is a set of $\sigma$-finite measure, i.e. if there is a sequence $(X_n)$ in $\mathcal{A}$ with $X = \bigcup_{n=1}^{\infty} X_n$ and $\mu(X_n) < \infty$ for all $n$.

Note that in this definition, the sets $A_n$ need not have any relation to each other, except they must of course all be subsets of $A$. In particular, they do not have to be either nested or disjoint. However, if desired it can always be arranged that they are nested or that they are disjoint:

XIII.1.4.4. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $A \in \mathcal{A}$. The following are equivalent:

(i) $A$ has $\sigma$-finite measure.

(ii) There is an increasing sequence $(B_n)$ in $\mathcal{A}$ with $A = \bigcup_{n=1}^{\infty} B_n$ and $\mu(B_n) < \infty$ for all $n$.

(iii) There is a pairwise disjoint sequence $(E_n)$ in $\mathcal{A}$ with $A = \bigcup_{n=1}^{\infty} E_n$ and $\mu(E_n) < \infty$ for all $n$.

Proof: This is just a consequence of standard nesting or disjointizing operations. (ii) $\Rightarrow$ (i) and (iii) $\Rightarrow$ (i) are trivial. To show (i) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (iii), set $B_n = \bigcup_{k=1}^{n} A_k$, $E_1 = A_1$, and $E_n = B_n \setminus B_{n-1}$ for $n > 1$. Note that $\mu(E_n) \leq \mu(B_n) \leq \sum_{k=1}^{n} \mu(A_k) < \infty$ by monotonicity and finite subadditivity. ☺
XIII.1.4.5. **Corollary.** Let \((X, \mathcal{A}, \mu)\) be a \(\sigma\)-finite measure space. Then

(i) There is an increasing sequence \((Y_n)\) of sets in \(\mathcal{A}\) with \(X = \bigcup_{n=1}^{\infty} Y_n\) and \(\mu(Y_n) < \infty\) for all \(n\).

(ii) There is a pairwise disjoint sequence \((Z_n)\) of sets in \(\mathcal{A}\) with \(X = \bigcup_{n=1}^{\infty} Z_n\) and \(\mu(Z_n) < \infty\) for all \(n\).

XIII.1.4.6. **Proposition.** Let \((X, \mathcal{A}, \mu)\) be a measure space. Any measurable subset of a set of \(\sigma\)-finite measure has \(\sigma\)-finite measure. In particular, if the measure space is \(\sigma\)-finite, every measurable set in \(X\) has \(\sigma\)-finite measure.

**Proof:** If \(A \in \mathcal{A}\) has \(\sigma\)-finite measure, let \((A_n)\) be a sequence of sets in \(\mathcal{A}\) with \(A = \bigcup_{n=1}^{\infty} A_n\) and \(\mu(A_n) < \infty\) for all \(n\). If \(B \in \mathcal{A}\), \(B \subseteq A\), set \(B_n = B \cap A_n\) for each \(n\). Then \(B = \bigcup_{n=1}^{\infty} B_n\), and \(\mu(B_n) < \infty\) by monotonicity.

**Decomposable Measures**

The essential feature of \(\sigma\)-finite measure spaces is that the whole space can be chopped up into a countable number of disjoint pieces, each of finite measure. For some purposes it is sufficient that the space can be chopped up in a similar way even if there are uncountably many pieces.

XIII.1.4.7. **Definition.** A measure space \((X, \mathcal{A}, \mu)\) is decomposable (sometimes called finitely decomposable) if there is a collection \(\{X_i : i \in I\}\) of pairwise disjoint sets in \(\mathcal{A}\) with \(X = \bigcup_{i \in I} X_i\), \(\mu(X_i) < \infty\) for all \(i\), and

\[
\mu(A) = \sum_{i \in I} \mu(A \cap X_i)
\]

for all \(A \in \mathcal{A}\) (where the sum is the supremum of the sums over finite subsets of \(I\)). By slight abuse of terminology we will also say that \(\mu\) is decomposable.

XIII.1.4.8. It follows immediately from XIII.1.2.4. that a \(\sigma\)-finite measure space is decomposable; in fact, \((X, \mathcal{A}, \mu)\) is \(\sigma\)-finite if and only if it is decomposable with a countable set \(I\). There are decomposable measures which are not \(\sigma\)-finite, e.g. counting measure on an uncountable set \(X\) (where we can take each \(X_i\) to be a singleton). There are also measures which are not decomposable, e.g. the ones in XIII.1.1.6.(ii)–(iii).

XIII.1.4.9. **Caution:** For a measure space to be decomposable, it is not enough that it decompose into a disjoint union of measurable sets of finite measure. The example of XIII.1.6.(iii) is a simple counterexample: the space is a disjoint union of singletons, each measurable and of finite measure. The problem is not that the subsets have measure 0; see XIII.1.9.7.

**Semifinite Measures**

There is a weaker condition than decomposability that is somewhat useful:
XIII.1.4.10. Definition. Let \((X, \mathcal{A}, \mu)\) be a measure space. A set \(E \in \mathcal{A}\) is semifinite, or has semifinite measure, if, whenever \(A \in \mathcal{A}, A \subseteq E,\) and \(\mu(A) = \infty,\) there is a \(B \in \mathcal{A}, B \subseteq A,\) with \(0 < \mu(B) < \infty.\) \((X, \mathcal{A}, \mu)\) is a semifinite measure space if \(X\) has semifinite measure. By slight abuse of terminology we will also say that \(\mu\) is semifinite.

XIII.1.4.11. It is obvious from the definition that every measurable subset of a set of semifinite measure also has semifinite measure. In particular, every measurable set in a semifinite measure space has semifinite measure.

XIII.1.4.12. Proposition. Every decomposable measure space is semifinite. In particular, every \(\sigma\)-finite measure space is semifinite.

Proof: Let \((X, \mathcal{A}, \mu)\) be decomposable. Let \(\{X_i : i \in I\}\) be a pairwise disjoint collection of sets in \(\mathcal{A}\) with \(X = \bigcup_i X_i, \mu(X_i) < \infty\) for all \(i,\) and
\[
\mu(A) = \sum_{i \in I} \mu(A \cap X_i)
\]
for all \(A \in \mathcal{A}.\) If \(\mu(A) = \infty,\) then for any \(M\) there is a finite subset \(F\) of \(I\) such that \(\sum_{i \in F} \mu(A \cap X_i) > M.\) If \(B = \bigcup_{i \in F} (A \cap X_i),\) then \(B \subseteq A\) and by monotonicity
\[
M < \mu(B) = \sum_{i \in F} \mu(A \cap X_i) \leq \sum_{i \in F} \mu(X_i) < \infty.
\]

XIII.1.4.13. There are semifinite measures which are not decomposable (XIII.1.9.3., XIII.1.9.4.). There are also measures which are not semifinite, e.g. the ones in XIII.1.1.6.(ii)–(iii). Measure spaces which do not at least have many subsets of semifinite measure are so pathological that they are rarely useful; for example, integration theory for such measure spaces is completely degenerate.

XIII.1.4.14. Proposition. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(E \in \mathcal{A}\) of semifinite measure. Then for every \(A \in \mathcal{A}, A \subseteq E,\) we have
\[
\mu(A) = \sup \{ \mu(B) : B \in \mathcal{A}, B \subseteq A, \mu(B) < \infty \}.
\]
Conversely, any \(E\) with this property has semifinite measure.

Proof: The formula is trivial (in any measure space) by monotonicity if \(\mu(A) < \infty,\) so assume \(\mu(A) = \infty.\) Set
\[
M = \sup \{ \mu(B) : B \in \mathcal{A}, B \subseteq A, \mu(B) < \infty \}.
\]
If \(M < \infty,\) for each \(n\) let \(B_n \in \mathcal{A}, B_n \subseteq A,\) and \(M - \frac{1}{n} \leq \mu(B_n) < \infty.\) Set \(D_n = \bigcup_{k=1}^{n} B_k\) for each \(n.\) Then \((D_n)\) is an increasing sequence, \(D := \bigcup_n D_n \subseteq A,\) and
\[
\mu(D_n) \leq \sum_{k=1}^{n} B_k < \infty
\]

1403
and therefore $\mu(D_n) \leq M$ for all $n$. We have $\mu(D) = \sup_n \mu(D_n) \leq M$ by XIII.1.2.7., and since $B_n \subseteq D$ for all $n$, by monotonicity we have 

$$M - \frac{1}{n} \leq \mu(B_n) \leq \mu(D) \leq M$$

for all $n$. It follows that $\mu(D) = M$.

Since $\mu(A) = \infty$ and $\mu(D) < \infty$, we have $\mu(A \setminus D) = \infty$, and thus by semifiniteness of $E$ there is an $C \in A, C \subseteq A \setminus D$, with $0 < \mu(C) < \infty$. But then $C \cup D \subseteq A$, and

$$M < \mu(C \cup D) = \mu(C) + \mu(D) = \mu(C) + M < \infty$$

contradicting the definition of $M$. Thus $M = \infty$.

The converse statement is obvious. 

XIII.1.5. Discrete and Continuous Measures

XIII.1.5.1. Definition. Let $(X, A, \mu)$ be a measure space. An atom in $(X, A, \mu)$ is a set $A \in A$ with the property that $\mu(A) > 0$ and, whenever $B \in A, B \subseteq A$, then either $\mu(B) = 0$ or $\mu(A \setminus B) = 0$.

An atom is a non-null measurable set which is “indivisible” up to null sets. The most obvious examples of atoms are (measurable) singleton subsets of positive measure, when such sets exist, although in general atoms need not be singletons. Measure spaces do not necessarily have any atoms; in fact, most interesting measure spaces do not.

XIII.1.6. Linear Combinations, Restrictions, and Extensions of Measures

We can take sums (finite or infinite) and scalar multiples of measures to obtain new measures, and we can in good circumstances restrict and extend measures. The arguments in this subsection are for the most part virtually trivial, with a few subtleties.

Addition of Measures

XIII.1.6.1. Proposition. Let $\mu$ and $\nu$ be measures on a measurable space $(X, A)$. Then $\mu + \nu$ is a measure on $(X, A)$. [By $\mu + \nu$ we mean the usual pointwise (setwise) sum, i.e. $(\mu + \nu)(A) = \mu(A) + \nu(A)$ for $A \in A$.]

Proof: We have $(\mu + \nu)(\emptyset) = 0 + 0 = 0$. If $(A_k)$ is a sequence of pairwise disjoint sets in $A$, then

$$(\mu+\nu)\left(\bigcup_{k=1}^{\infty} A_k\right) = \mu\left(\bigcup_{k=1}^{\infty} A_k\right) + \nu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) + \sum_{k=1}^{\infty} \nu(A_k) = \sum_{k=1}^{\infty} \left[\mu(A_k) + \nu(A_k)\right] = \sum_{k=1}^{\infty} (\mu+\nu)(A_k)$$

where combining the two infinite sums is justified by () since the terms are nonnegative. 

Note that we can only add measures defined on the same measurable space. We can even add a sequence of measures:
XIII.1.6.2. **Proposition.** Let \( (\mu_n) \) be a sequence of measures on a measurable space \((X, A)\). Define \( \mu = \sum_{n=1}^{\infty} \mu_n \) on \((X, A)\) by

\[
\mu(A) = \left( \sum_{n=1}^{\infty} \mu_n \right)(A) = \sum_{n=1}^{\infty} \mu_n(A)
\]

for \( A \in A \). Then \( \mu \) is a measure on \((X, A)\).

**Proof:** We clearly have \( \mu(\emptyset) = 0 \). If \( (A_k) \) is a sequence of pairwise disjoint sets in \( A \), then

\[
\mu \left( \bigcup_{k=1}^{\infty} A_k \right) = \sum_{n=1}^{\infty} \mu_n \left( \bigcup_{k=1}^{\infty} A_k \right) = \sum_{n=1}^{\infty} \left( \sum_{k=1}^{\infty} \mu_n(A_k) \right)
\]

\[
= \sum_{k=1}^{\infty} \left( \sum_{n=1}^{\infty} \mu_n(A_k) \right) = \sum_{k=1}^{\infty} \mu(A_k)
\]

where the interchange of order in the double summation is justified by () since the terms are nonnegative. \( \diamond \)

In fact, an arbitrary set of measures can be added in the same way. See () for this and more general results about combining measures.

**Scalar Multiples of Measures**

XIII.1.6.3. **Proposition.** Let \( \mu \) be a measure on a measurable space \((X, A)\), and \( 0 \leq \alpha < \infty \). Then \( \alpha \mu \) is also a measure on \((X, A)\). [\( \alpha \mu \) is defined by \( [\alpha \mu](A) = \alpha [\mu(A)] \) for \( A \in A \); here by definition \( 0 \cdot \infty = 0 \).]

**Proof:** We obviously have \( [\alpha \mu](\emptyset) = 0 \). If \( (A_n) \) is a sequence of disjoint sets in \( A \), then

\[
[\alpha \mu] \left( \bigcup_{k=1}^{\infty} A_k \right) = \alpha \left[ \mu \left( \bigcup_{k=1}^{\infty} A_k \right) \right] = \alpha \sum_{k=1}^{\infty} \mu(A_k) = \sum_{k=1}^{\infty} \alpha [\mu(A_k)] = \sum_{k=1}^{\infty} [\alpha \mu](A_k)
\]

where the next-to-last step is justified by (). \( \diamond \)

**Linear Combinations of Measures**

Combining XIII.1.6.1. and XIII.1.6.3., we obtain:

XIII.1.6.4. **Proposition.** Any nonnegative linear combination of measures on a measurable space is a measure. Specifically, if \( \mu_1, \ldots, \mu_n \) are measures on a measurable space \((X, A)\), and \( \alpha_1, \ldots, \alpha_n \) are nonnegative real numbers, then \( \sum_{k=1}^{n} \alpha_k \mu_k \) is a measure on \((X, A)\), where \( [\sum_{k=1}^{n} \alpha_k \mu_k](A) = \sum_{k=1}^{n} \alpha_k \mu_k(A) \) for \( A \in A \).

We have the same for countable sums, using XIII.1.6.2.:
XIII.1.6.5. **Proposition.** If $\mu_1, \mu_2, \ldots$ is a sequence of measures on a measurable space $(X, \mathcal{A})$, and $\alpha_1, \alpha_2, \ldots$ is a sequence of nonnegative real numbers, then $\sum_{k=1}^{\infty} \alpha_k \mu_k$ is a measure on $(X, \mathcal{A})$, where $[\sum_{k=1}^{\infty} \alpha_k \mu_k](A) = \sum_{k=1}^{\infty} \alpha_k \mu_k(A)$ for $A \in \mathcal{A}$.

Restrictions and Compressions of Measures

XIII.1.6.6. **Definition.** Let $\mu$ be a measure on a measurable space $(X, \mathcal{A})$, and $E \in \mathcal{A}$.

(i) The **restriction** of $\mu$ to $E$ is the measure $\mu|_E$ on $(E, \mathcal{A}_E)$ defined by

$$[\mu|_E](A) = \mu(A)$$

for $A \in \mathcal{A}_E$ (i.e. $A \in \mathcal{A}$ and $A \subseteq E$).

(ii) The **compression** of $\mu$ to $E$ is the measure $\mu_{\cdot E}$ on $(X, \mathcal{A})$ defined by

$$[\mu_{\cdot E}](A) = \mu(A \cap E)$$

for $A \in \mathcal{A}$.

The measures $\mu|_E$ and $\mu_{\cdot E}$ are in some sense the “same”, but there is a technical difference: $\mu|_E$ is a measure on $E$, while $\mu_{\cdot E}$ is a measure on all of $X$ which is supported on $E$, i.e. $[\mu_{\cdot E}](X \setminus E) = 0$.

It is a routine exercise (Exercise (i)) to verify that $\mu|_E$ and $\mu_{\cdot E}$ are indeed measures.

Extensions and Dilations of Measures

XIII.1.6.7. **Definition.** Let $(X, \mathcal{A})$ be a measurable space, and $E \in \mathcal{A}$. If $\mu$ is a measure on $(X, \mathcal{A})$, and $\nu$ a measure on $(E, \mathcal{A}_E)$, then $\mu$ is an **extension** of $\nu$ to $X$ if $\nu = \mu|_E$.

There is nothing unique about extensions of measures in general.

There is always a minimal extension of any measure:

XIII.1.6.8. **Definition.** Let $(X, \mathcal{A})$ be a measurable space, and $E \in \mathcal{A}$. If $\nu$ is a measure on $(E, \mathcal{A}_E)$, the **dilation** of $\nu$ to $(X, \mathcal{A})$ is the measure $\nu \uparrow X$ defined by

$$[\nu \uparrow X](A) = \nu(A \cap E)$$

for $A \in \mathcal{A}$.

XIII.1.6.9. It is easily verified that $\nu \uparrow X$ is a measure on $(X, \mathcal{A})$, that $(\nu \uparrow X)$ is an extension of $\nu$ to $X$, that $\nu \uparrow X = (\nu \uparrow X)_{\cdot E}$, and that $\nu \uparrow X$ is the unique measure on $(X, \mathcal{A})$ with these properties and the smallest extension of $\nu$ to $X$ (in the sense that if $\mu$ is any extension, then $\mu \geq (\nu \uparrow X)$.) In fact, if $\mu$ is any measure on $(X, \mathcal{A})$, then $\mu = ([\mu|_E] \uparrow X) + ([\mu|_{X \setminus E}] \uparrow X) = (\mu_{\cdot E}) + (\mu_{\cdot X \setminus E})$.

Restriction of Measures to Sub-$\sigma$-Algebras

There is another type of restriction of measures:
XIII.1.6.10. Definition. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\mathcal{B}\) a sub-\(\sigma\)-algebra of \(\mathcal{A}\). Then the restriction of \(\mu\) to \(\mathcal{B}\) is the measure \(\mu|_{\mathcal{B}}\) on \((X, \mathcal{B})\) defined by
\[
[\mu|_{\mathcal{B}}](B) = \mu(B)
\]
for \(B \in \mathcal{B}\).

Note that \(\mu|_{\mathcal{B}}\) is technically a different measure than \(\mu\) since the specification of the \(\sigma\)-algebra of measurable subsets is part of the definition of a measure. There are occasions, however, where making an explicit distinction between \(\mu\) and \(\mu|_{\mathcal{B}}\) seems overly pedantic, and we will occasionally overlook the distinction when no confusion can occur. The most important example is restriction of Lebesgue measure to the Borel sets.

XIII.1.6.11. We can also regard the measure \(\mu\) as an extension of \(\mu|_{\mathcal{B}}\) to the larger \(\sigma\)-algebra \(\mathcal{A}\). Existence of extensions of measures to larger \(\sigma\)-algebras is a very subtle matter and is not possible in general; see II.10.1.1., XIII.1.3.7., () for some instances of extensions and nonexistence of extensions.

Push-Forward of Measures
We can push measures forward by measurable functions:

XIII.1.6.12. Definition. Let \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) be measure spaces, and \(f : X \to Y\) an \((\mathcal{A}, \mathcal{B})\)-measurable function (). If \(\mu\) is a measure on \((X, \mathcal{A})\), then the push-forward of \(\mu\) by \(f\) is the measure \(f_*(\mu)\) on \((Y, \mathcal{B})\) defined by
\[
[f_*(\mu)](B) = \mu(f^{-1}(B))
\]
for \(B \in \mathcal{B}\).

As usual, it is routine to show that \(f_*(\mu)\) is a measure on \((Y, \mathcal{B})\).

XIII.1.6.13. Examples. (i) If \((X, \mathcal{A})\) is a measurable space and \(E \in \mathcal{A}\), and \(f : E \to X\) is the inclusion map, and \(\nu\) is a measure on \((E, \mathcal{A}_E)\), then \(f_*(\nu) = \nu|_E\).

(ii) Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\mathcal{B}\) a sub-\(\sigma\)-algebra of \(\mathcal{A}\). Let \(f\) be the identity map on \(X\), regarded as a measurable function from \((X, \mathcal{A})\) to \((X, \mathcal{B})\). Then \(f_*(\mu) = \mu|_{\mathcal{B}}\).

XIII.1.6.14. There is no reasonable way to pull back measures from \((Y, \mathcal{B})\) to \((X, \mathcal{A})\) in general, but this can be done under some special circumstances (e.g. ()).

XIII.1.7. Measures Defined on Rings and \(\sigma\)-Rings

XIII.1.8. Egorov’s Theorem

XIII.1.8.1. Theorem. [Egorov’s Theorem] Let \((X, \mathcal{A}, \mu)\) be a finite measure space. Let \((f_n)\) be a sequence of real-valued measurable functions on \(X\) converging pointwise a.e. to a real-valued (measurable) function \(f\). Then for every \(\epsilon > 0\) there is a set \(A \in \mathcal{A}\) with \(\mu(X \setminus A) < \epsilon\) and \(f_n \to f\) uniformly on \(A\).
Proof: Fix \( \epsilon > 0 \). Let \( E \in \mathcal{A} \) with \( f_n(x) \to f(x) \) for all \( x \in E \) and \( \mu(X \setminus E) = 0 \). For each \( n \) and \( m \), set

\[
A_{n,m} = \left\{ x \in X : |f_k(x) - f(x)| < \frac{1}{m} \text{ for all } k \geq n \right\}.
\]

We have \( A_{n,m} \in \mathcal{A} \) for all \( n, m \). For fixed \( m \), we have \( A_{n,m} \subseteq A_{n+1,m} \) for all \( n \) and \( \cup_n A_{n,m} \supseteq E \), i.e. \( \cap_n [X \setminus A_{n,m}] \subseteq X \setminus E \). Thus by (1)

\[
0 \leq \inf_n \mu(X \setminus A_{n,m}) \leq \mu(X \setminus E) = 0
\]

and thus there is an \( n_m \) such that \( \mu(X \setminus A_{n,m}) < \frac{\epsilon}{2m} \).

Let \( A = \cap_{m=1}^{\infty} A_{n,m,m} \). Then

\[
\mu(X \setminus A) = \mu\left( \bigcup_{m=1}^{\infty} [X \setminus A_{n,m,m}] \right) \leq \sum_{m=1}^{\infty} \mu(X \setminus A_{n,m,m}) < \sum_{m=1}^{\infty} \frac{\epsilon}{2m} = \epsilon
\]

and \( f_n \to f \) uniformly of \( A \): if \( \delta > 0 \), fix \( m \) with \( \frac{1}{m} \leq \delta \); then, for all \( k \geq n_m \) and all \( x \in A \subseteq A_{n,m,m} \), we have

\[
|f_k(x) - f(x)| < \frac{1}{m} \leq \delta.
\]

\[\Box\]

XIII.1.9. Exercises

XIII.1.9.1. Let \( \mu \) be a finitely additive measure on a measurable space \((X, \mathcal{A})\). Show that the following are equivalent:

(i) \( \mu \) is countably subadditive (XIII.1.2.1).

(ii) \( \mu \) is upward continuous (XIII.1.2.7).

(iii) \( \mu \) is countably additive, i.e. a measure.

XIII.1.9.2. Let \( \mu \) be counting measure on \((\mathbb{N}, \mathcal{P}([\mathbb{N}]))\). Show that for any \( n \in \{0\} \cup \mathbb{N} \cup \{\infty\} \), there are subsets \( A, B \) of \( \mathbb{N} \) with \( B \subseteq A \), \( \mu(A) = \mu(B) = \infty \), and \( \mu(A \setminus B) = n \).

XIII.1.9.3. Let \( X \) be an uncountable set, and \( x_0 \) a fixed element of \( X \). Let \( \mathcal{A} \) be the collection of all countable subsets of \( X \) not containing \( x_0 \), and their complements.

(a) Show that \( \mathcal{A} \) is a \( \sigma \)-algebra on \( X \).

(b) Let \( \mu \) be the restriction of counting measure to \( \mathcal{A} \). Show that \( \mu \) is semifinite.

(c) The point \( x_0 \) is not contained in any measurable set of finite measure. Conclude that \( \mu \) cannot be decomposable.

(d) If the cardinality of \( X \) is \( \kappa \), then the cardinality of \( \mathcal{A} \) is \( \kappa^{\aleph_0} \). In particular, if \( \text{card}(X) = 2^{\aleph_0} \), then \( \text{card}(\mathcal{A}) = 2^{\aleph_0} \).
XIII.1.9.4.  [Bog07, ] Let $X$ be a set of cardinality $\aleph_1$, $Y$ a set of cardinality $\aleph_2$, and $Z = X \times Y$. If $E \subseteq Z$, $x \in X$, and $y \in Y$, write $E_x$ and $E^y$ for the vertical and horizontal cross sections of $E$ at $x$ and $y$ respectively ().

(a) Let $\mathcal{M}$ be the set of all subsets $E$ of $Z$ such that for every $x \in X$, $E_x$ is either countable or cocountable, and for every $y \in Y$, $E^y$ is either countable or cocountable. Show that $\mathcal{M}$ is a $\sigma$-algebra.
(b) If $E \in \mathcal{M}$, define
\[
\mu(E) = \text{card}\{x \in X : E_x \text{ is cocountable}\} + \text{card}\{y \in Y : E^y \text{ is cocountable}\}
\]
(where $\mu(E) = \infty$ if either set on the right is infinite). Show that $\mu$ is a measure on $(Z, \mathcal{M})$. Every countable subset (in particular, every singleton subset) of $X$ is in $\mathcal{M}$ and has measure $0$.
(c) Show that $\mu$ is semifinite. [If $\mu(E) > 0$, then $E$ contains a set of the form $A \times \{y\}$ for a cocountable subset $A$ of $X$ and some $y \in Y$, or a set of the form $\{x\} \times B$ for cocountable $B \subseteq Y$.]
(d) Suppose $\mu$ is decomposable, and let $\{Z_i : i \in I\}$ be a decomposition. For each $y \in Y$, show that the set $(Z_i)^y$ is cocountable for exactly one $i$.
(e) If $y \in Y$, let $\phi(y)$ be the unique $i$ as in (d). Then $\phi$ is a function from $Y$ to $I$. For any $i \in I$, $\phi^{-1}(\{i\})$ is finite. Conclude that $\text{card}(I) \geq \aleph_2$.
(f) On the other hand, for each $x \in X$ there is an $i$ such that $(Z_i)_x$ is cocountable. Thus there are only countably many $i$ such that $(Z_i)_x$ is nonempty. Conclude that $\text{card}(I) \leq \aleph_1$.
(g) Conclude from (d)–(f) that $\mu$ is not decomposable.

XIII.1.9.5.  Let $I$ be an uncountable set, and for each $i \in I$ let $X_i$ be a copy of $[0, 1]$. Let $X$ be the disjoint union of the $X_i$. Let $\mathcal{M}$ be the $\sigma$-algebra of Lebesgue measurable sets in $[0, 1]$, and let
\[
\mathcal{B} = \{B \subseteq X : B \cap X_i \in \mathcal{M} \text{ for all } i \in I\}
\]
and $\mathcal{A}$ the collection of all $A \in \mathcal{B}$ such that $A \cap X_i$ is either $\emptyset$ or $X_i$ for all but countably many $i$.
(a) Show that $\mathcal{B}$ and $\mathcal{A}$ are $\sigma$-algebras.
(b) For $A \in \mathcal{A}$, set $\mu(A) = \sum_{i \in I} \lambda(A \cap X_i)$, where $\lambda$ is Lebesgue measure on $X_i$. Show that $\mu$ is a complete measure on $(X, \mathcal{A})$.
(c) Show that $\{X_i : i \in I\}$ is a decomposition of $(X, \mathcal{A}, \mu)$, so $\mu$ is decomposable and, in particular, semifinite.
(d) If $B \in \mathcal{B}$, set $\nu(B) = \mu(B)$ if $B \in \mathcal{A}$ and $\nu(B) = \infty$ if $B \in \mathcal{B} \setminus \mathcal{A}$. Show that $\nu$ is a complete measure on $(X, \mathcal{B})$ extending $\mu$.
(e) If $B$ is a subset of $X$ such that $B \cap X_i$ is a singleton for all $i$, then $\nu(B) = \infty$, but $B$ contains no subsets of finite positive $\nu$-measure. Thus $\nu$ is not semifinite.

Note that this problem, including (e), can be done without using the Axiom of Choice by taking $X$ to be the unit square and the $X_i$ to be (say) horizontal cross sections. The set $B$ of (e) can be taken to be a vertical line segment.

XIII.1.9.6.  Let $(X, \mathcal{A}, \mu)$ be a semifinite measure space.
(a) If $\mathcal{A}$ contains only countably many sets of finite measure, show that $\mu$ is $\sigma$-finite.
(b) [R. Israel] If $\mathcal{A}$ contains only $\aleph_1$ sets of finite measure, and every point of $X$ is contained in a measurable subset of finite measure, then $\mathcal{A}$ is decomposable. [Well order the sets of finite measure as $(A_\alpha : \alpha < \omega_1)$, and for each $\alpha$ let $X_\alpha = A_\alpha \setminus (\cup_{\beta < \alpha} A_\beta)$.] The last condition in (b) is not automatic, at least if the continuum hypothesis is assumed (XIII.1.9.3.).

1409
XIII.1.9.7. Let $Y$ be an uncountable set and $\nu$ the measure on $(Y, \mathcal{P}(Y))$ with $\nu(A) = 0$ if $A$ is countable and $\nu(A) = \infty$ if $A$ is uncountable. Let $X = Y \times \{0, 1\}$, and for $A, B \subseteq Y$ set

$$\mu([A \times \{0\}] \cup [B \times \{1\}]) = \nu(A) + |B|$$

where $|B|$ is the number of elements in $B$ (counting measure of $B$).

(a) Show that $\mu$ is a measure on $(X, \mathcal{P}(X))$.

(b) For each $y \in Y$ set $X_y = \{(y, 0), (y, 1)\}$. Then $X$ is the disjoint union of the $X_y$ and $\mu(X_y) = 1$ for all $y$.

(c) Nonetheless, $(X, \mathcal{P}(X), \mu)$ is not decomposable or even semifinite.
XIII.2. Probability

“How can we venture to speak of the laws of chance? Is not chance the antithesis of all law?”

J. Bertrand

XIII.2.1. Introduction

XIII.2.1.1. From its first beginnings, probability theory had some of the flavor of primitive measure theory. In 1933, A. Kolmogorov [1] cemented the connection by giving an axiomatization of probability which serves as the foundation of all modern probability theory. (Of course, Kolmogorov did not come up with this axiomatization out of thin air; the ideas were implicit and in fact increasingly explicit in other work in the field.)

XIII.2.1.2. The two basic ingredients of probability are:

(i) A sample space \( \Omega \), the set of all possible outcomes of a specific process or experiment. A subset of \( \Omega \) is called an event.

(ii) A probability \( P \), an assignment to each event \( E \) a numerical probability or likelihood \( P(E) \) that \( E \) occurs (i.e. that the outcome of the process lies in \( E \)). We should have \( 0 \leq P(E) \leq 1 \) for any \( E \), and \( P(\Omega) = 1 \).

XIII.2.1.3. It is virtually a no-brainer that \( P \) should be finitely additive. Kolmogorov realized that in order to get a satisfactory theory, one needs \( P \) to actually be countably additive.

XIII.2.1.4. Examples show that it is not possible in general to assign a probability to every subset of \( \Omega \) in a countably additive way (the process described in I.3.1.17., along with XIII.3.9.5. and II.10.3.6., is an example where it is not reasonably possible). Thus probability can only be assigned to some collection of well-behaved events. This turns out not to be a serious drawback; indeed, experience shows that it is frequently useful to deliberately restrict the set of events to which a probability is assigned.

Nevertheless, it is important to be able to apply sequences of set-theoretic operations to allowable events and stay within the class of allowable events. This leads naturally to the assumption that the set of allowable events forms a \( \sigma \)-algebra of subsets of \( \Omega \).

Thus Kolmogorov was led to make the following definition:

XIII.2.1.5. Definition. A probability space is a triple \((\Omega, \mathcal{E}, P)\), where

- \( \Omega \) is a set, called the sample space.
- \( \mathcal{E} \) is a \( \sigma \)-algebra of subsets of \( \Omega \), called the set of (allowable) events.
- \( P \) is a function from \( \mathcal{E} \) to \([0, 1]\) satisfying
  
  \[
  (i) \quad P(\Omega) = 1.
  \]

\[\text{References:} \quad [\text{Ber72}]; \text{cf.} \quad [\text{Poi96}, \text{p. 64}].\]

1411
(ii) \( P \) is countably additive, i.e. if \((E_n)\) is a sequence of pairwise disjoint sets in \( \mathcal{E} \), then
\[
P \left( \bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} P(E_n) .
\]

Such a \( P \) is called a probability on \((\Omega, \mathcal{E})\). If \( E \in \mathcal{E} \), \( P(E) \) is called the probability of the event \( E \).

**XIII.2.1.6.** If \((\Omega, \mathcal{E}, P)\) is a probability space, then \((\Omega, \mathcal{E})\) is a measurable space. We have
\[
P(\emptyset) = P(\emptyset \cup \emptyset \cup \cdots) = \sum_{n=1}^{\infty} P(\emptyset) = \infty \cdot P(\emptyset)
\]
and thus \( P(\emptyset) = 0 \) (since it is in \([0,1]\)). Therefore \((\Omega, \mathcal{E}, P)\) is a measure space \((\text{XIII.1.1.1.})\), with \( P \) a probability measure. Conversely, any probability measure space is a probability space. Thus probability spaces are precisely the same thing as probability measure spaces, with slightly different notation and terminology.

Therefore, the results of \((\text{XIII.1.2.})\) apply in probability spaces:

**XIII.2.1.7.** Proposition. Let \((\Omega, \mathcal{E}, P)\) be a probability space. Then

1. **(i) \( P \) is finitely additive:** if \( E_1, \ldots, E_n \in \mathcal{E} \) are pairwise disjoint, then
\[
P \left( \bigcup_{k=1}^{n} E_k \right) = \sum_{k=1}^{n} P(E_k) .
\]

In particular, if \( E, F \in \mathcal{E} \) are disjoint, then \( P(E \cup F) = P(E) + P(F) \).

2. **(ii) \( P \) is monotone:** if \( E, F \in \mathcal{E}, E \subseteq F \), then \( P(E) \leq P(F) \).

3. **(iii) \( P \) is countably subadditive:** if \((E_n)\) is a sequence in \( \mathcal{E} \), then
\[
P \left( \bigcup_{n=1}^{\infty} E_n \right) \leq \sum_{n=1}^{\infty} P(E_n) .
\]

\( P \) is thus also finitely subadditive; in particular, if \( E, F \in \mathcal{E} \), then \( P(E \cup F) \leq P(E) + P(F) \). (In fact, we have \( P(E \cup F) = P(E) + P(F) - P(E \cap F) \).)

4. **(iv) \( P \) is upward continuous:** If \((E_n)\) is an increasing sequence in \( \mathcal{E} \), then
\[
P \left( \bigcup_{n=1}^{\infty} E_n \right) = \sup_n P(E_n) = \lim_{n \to \infty} P(E_n) .
\]

5. **(v) \( P \) is downward continuous:** If \((E_n)\) is a decreasing sequence in \( \mathcal{E} \), then
\[
P \left( \bigcap_{n=1}^{\infty} E_n \right) = \inf_n P(E_n) = \lim_{n \to \infty} P(E_n) .
\]

1412
XIII.2.2. The Dictionary

The mathematical setup of probability is thus measure theory, using somewhat different language. Here is a dictionary of the correspondence of basic terminology between measure theory and probability. Some other terminology such as “distribution” is common to both.

<table>
<thead>
<tr>
<th>Measure Theory</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurable space</td>
<td>Sample space</td>
</tr>
<tr>
<td>Measure</td>
<td>Probability</td>
</tr>
<tr>
<td>Real-valued measurable function</td>
<td>Random variable</td>
</tr>
<tr>
<td>Integral</td>
<td>Expectation</td>
</tr>
<tr>
<td>Almost everywhere (a.e.)</td>
<td>Almost surely (a.s.)</td>
</tr>
<tr>
<td></td>
<td>Independent</td>
</tr>
</tbody>
</table>

In fact, it is sometimes said that probability is “measure theory plus independence” (although this is clearly somewhat of an oversimplification). It is unfair and not really accurate to say that probability is part of measure theory, although modern probability is certainly built on measure theory.

XIII.2.3. Exercises

XIII.2.3.1. A Conditional Probability Paradox?

(a) You have $n$ red balls, $n$ white balls, and $n$ blue balls ($n \geq 3$) in an urn. Your assistant chooses three balls at random from the urn and puts them in a box. What is the probability $p$ that all three balls in the box have the same color?

(b) While the assistant is choosing the balls, your second assistant comes into the room and sees some of what the first assistant is doing. He tells you that he saw one of the balls the first assistant put in the box and it was red. What now is the conditional probability $q$ that all the balls in the box are the same color, given that at least one of the balls is red? Explain why $q < p$, especially since if the second assistant had said that the ball he saw was white, or blue, the conditional probability is also $q$, and one of these situations must occur, with equal probability since the colors are completely symmetric.

(c) Suppose your second assistant doesn’t speak much English, and you know not a word of his native language. He reports to you that the color of ball he saw was gloop, but can’t remember the English word for this color. What is the conditional probability that the three balls are the same color, given that at least one of the balls is gloop? On the one hand, no matter what gloop means, the conditional probability is $q$; on the other hand, he has given you no information, so the conditional probability should just be the probability $p$. Explain and resolve the paradox.

(d) The first assistant speaks the second assistant’s language fluently, but is temporarily unavailable to translate. Anyway, the first assistant can provide more information than just the meaning of gloop; in fact, he knows what the colors of the three balls are. How does this affect the conditional probabilities, if at all?

(e) Do $p$ and/or $q$ depend on $n$? Explain why or why not.
XIII.3. Measures Defined by Extensions

The most important construction of measures is by the Carathéodory extension process. One begins with a “premeasure” defined on a collection $C$ of subsets of a set $X$, and extends the premeasure to a measure on a $\sigma$-algebra containing $C$. The extension is made via an “outer measure” defined on all subsets of $X$.

Let us informally outline the process and try to give some motivation for the procedure, which can otherwise seem strange and nonintuitive at first glance. First consider how to define the “area” of a region in the plane. If $A$ is a bounded subset of the plane $\mathbb{R}^2$, a reasonable and commonly-used way to define the area of $A$ is to cover it with a rectangular grid $G$, and let $\overline{V}(A,G)$ be the sum of the areas of the rectangles in the grid which contain points of $A$, and $\overline{V}(A,G)$ the sum of the areas of the rectangles in the grid entirely contained in $A$. $\overline{V}(A,G)$ and $\underline{V}(A,G)$ can be regarded as an “upper” or “outer” estimate, and a “lower” or “inner” estimate, respectively, for the area of $A$. As $G$ becomes finer, the outer estimates decrease and the inner estimates increase, and for “nice” $A$ they converge to the same value, which is a reasonable definition of the area of $A$. See XIV.3.2.1. and XIV.3.2.10.

It turns out that not enough subsets of the plane are “nice” enough to have an area defined in this way to give a satisfactory theory of measure and integration (i.e. such sets do not form a $\sigma$-algebra). But the idea of approximating from the outside and inside is one which, with suitable care, gives a good theory. We begin with a collection $C$ of sets which have a reasonable “size” already specified, and, to get an “outer estimate” of the “size” of a general set $A$, consider collections $(C_n)$ of sets in $C$ which cover $A$ (whose union contains $A$); the sum of the “sizes” of the $C_n$ is an upper estimate for the “size” of $A$. There are two differences we allow from the grid covering example, which together give us a notion of “size” with considerably more flexibility:

1. We allow countable collections $(C_n)$, not just finite collections.

2. We do not insist that the $C_n$ are disjoint or even “nonoverlapping.”

The infimum of all such upper estimates will be called the “outer measure” of the set $A$, written $\mu^*(A)$.

For the “inner” estimate, it turns out that the allowance of countably infinite covers is incompatible with truly estimating size from the “inside” as in the grid method. But in the grid example, there is a simple change of point of view which does work in the new setting: if $R$ is a large rectangle which is subdivided to make the grid $G$, it is easily seen that if $V(R)$ is the area of $R$, then $\overline{V}(A,G) = V(R) - \overline{V}(R \setminus A, G)$. So, if we are working within subsets of a fixed set $Y$ of finite measure, we can replace inner estimates for the size of our set $A$ by outer estimates for the size of $Y \setminus A$. The criterion that the inner and outer estimates for the size of $A$ should converge to the same limit as the covers get “finer” translates into the requirement that the sum of the outer measure $\mu^*(A)$ and the outer measure $\mu^*(Y \setminus A)$ should be the measure of $Y$. If the whole set $X$ has finite measure, it turns out that for $A$ to be “measurable” (to have a well-behaved notion of “size”) we need only require this condition for $Y = X$; however, in the general setting we must have a technical strengthening: for every $Y \subseteq X$, we require

$$\mu^*(Y \cap A) + \mu^*(Y \cap A^c) = \mu^*(Y).$$

Roughly speaking, the set $A$ is measurable if, for every $Y \subseteq X$, the sets $Y \cap A$ and $Y \cap A^c$ can be finely covered by “almost disjoint” countable families from $C$.

It turns out then that under reasonable conditions on the size function (“premeasure”) on $C$ we can prove that the collection of “measurable” subsets of $X$ is a $\sigma$-algebra containing $C$, and that $\mu^*$ restricts to a measure on this $\sigma$-algebra which agrees with the premeasure on $C$, and in particular the premeasure on $C$ extends to a measure on a $\sigma$-algebra containing $C$. 

1414
We will do most of the construction for “outer measures” defined abstractly, which have the crucial properties of outer measures induced by premeasures. This is essential since there are important outer measures which are not induced (at least not in any useful way) from premeasures.

XIII.3.1. Outer Measures

Outer measures are a crucial technical tool in measure theory. Almost all interesting measures are constructed (at least indirectly) from outer measures.

XIII.3.1.1. Definition. Let \( X \) be a set. An outer measure on \( X \) is a function \( \mu^* : \mathcal{P}(X) \to [0, \infty] \) such that

1. \( \mu^*(\emptyset) = 0 \)
2. If \( A \subseteq B \), then \( \mu^*(A) \leq \mu^*(B) \) (monotonicity)
3. If \( A = \bigcup_{n=1}^{\infty} A_n \), then \( \mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n) \) (countable subadditivity).

There are two important differences between measures and outer measures. First, an outer measure is defined on all subsets of \( X \); second, an outer measure is not even finitely additive, let alone countably additive, in general.

XIII.3.1.2. In the case of a measure, monotonicity (condition (ii)) follows from the other properties (i), but (ii) does not follow from (i) and (iii) for outer measures in general (take Example XIII.3.5.14. with \( \alpha < 1 \)). In the presence of (i), both (ii) and (iii) can be replaced by the single condition

4. If \( A \subseteq \bigcup_{n=1}^{\infty} A_n \), then \( \mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n) \).

[(iii') obviously follows from (ii) and (iii), and (iii') trivially implies (iii); for (iii') \( \Rightarrow \) (ii), if \( A \subseteq B \), set \( A_1 = B \) and \( A_n = \emptyset \) for \( n > 1 \), and use (i).]

XIII.3.1.3. As with measures, countable subadditivity for outer measures implies finite subadditivity: if \( A \subseteq \bigcup_{k=1}^{n} A_k \), then \( \mu^*(A) \leq \sum_{k=1}^{n} \mu^*(A_k) \). This follows from countable subadditivity by taking \( A_k = \emptyset \) for \( k > n \). In particular, if \( A, B \subseteq X \), then \( \mu^*(A \cup B) \leq \mu^*(A) + \mu^*(B) \).

XIII.3.1.4. Note: It must be noted that a substantial number of references, particularly ones concerned with geometric measure theory, call an outer measure a “measure”; in these references, a measure on a set \( X \) is defined on all subsets of \( X \), but is only countably additive on a \( \sigma \)-algebra of “measurable” subsets of \( X \). The distinction is largely a semantic one: every measure has a canonical extension to an outer measure, and conversely every outer measure naturally defines a \( \sigma \)-algebra of measurable subsets on which it is countably additive (a measure in our sense). The two processes are almost, but not quite, inverses of each other, however. The terminology we use is more flexible and superior for most purposes: for example, it is often crucial to keep track of the \( \sigma \)-algebra on which a measure is defined, particularly in probability where it is often necessary to consider and distinguish between two or more \( \sigma \)-algebras.

1415
XIII.3.1.5. Most outer measures are constructed beginning with a “measure” defined on a collection of subsets which is too small to be a \(\sigma\)-algebra, usually a semiring. Suppose \(C\) is a collection of subsets of a set \(X\), with the property that \(\emptyset \in C\). Let \(\iota\) be a function from \(C\) to \([0, \infty]\). We will eventually impose some reasonable measure-like conditions on \(C\) and \(\iota\), but to get an outer measure the only requirement on \(\iota\) is that \(\iota(\emptyset) = 0\).

Define a function \(\mu^*\) from \(\mathcal{P}(X)\) to \([0, \infty]\) as follows. If \(A \subseteq X\), set

\[
\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \iota(C_n) : C_n \in C, \ A \subseteq \bigcup_{n=1}^{\infty} C_n \right\}
\]

(if \(A\) has no countable cover by sets in \(C\), set \(\mu^*(A) = \infty\).)

If \(A \subseteq \bigcup_{n=1}^{\infty} C_n\), we say the sequence \(\{C_n\}\) covers \(A\).

Note that since \(\emptyset \in C\), finite as well as countably infinite covers are allowed in computing \(\mu^*(A)\), since a finite cover \(\{C_1, \ldots, C_m\}\) of \(A\) can be expanded to a countably infinite cover by setting \(C_n = \emptyset\) for \(n > m\) and using \(\iota(\emptyset) = 0\).

XIII.3.1.6. Some authors do not include the requirement that \(\emptyset \in C\) and/or that \(\iota(\emptyset) = 0\), obtaining that \(\mu^*(\emptyset) = 0\) since \(\emptyset\) is covered by the “empty cover” and hence \(\mu^*(\emptyset)\) is less than or equal to the “empty sum,” which is by definition 0. But there is no advantage in this approach (\(\emptyset\) can be added to \(C\) if necessary and \(\iota(\emptyset)\) can be defined or redefined to be 0 with no penalty) and it causes technical problems, e.g. when one wants to use finite covers as well as countable ones.

XIII.3.1.7. Proposition. This function \(\mu^*\) is an outer measure on \(X\).

**Proof:** Since \(\emptyset \in C\), \(\{\emptyset\}\) is a cover of \(\emptyset\), so \(\mu^*(\emptyset) = 0\). If \(A \subseteq B\), any countable cover of \(B\) is a countable cover of \(A\), so the infimum in calculating \(\mu^*(A)\) is over a larger collection of covers, and hence \(\mu^*(A) \leq \mu^*(B)\). Finally, suppose \(A = \bigcup_{n=1}^{\infty} A_n\). If \(\sum_{n=1}^{\infty} \mu^*(A_n) = \infty\), then clearly \(\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)\). In the case that \(\sum_{n=1}^{\infty} \mu^*(A_n) < \infty\), so \(\mu^*(A_n) < \infty\) for all \(n\), let \(\epsilon > 0\), and for each \(n\) let \(\{C_{nk}\}\) be a cover of \(A_n\) with

\[
\sum_{k=1}^{\infty} \iota(C_{nk}) \leq \mu^*(A_n) + 2^{-n} \epsilon.
\]

Then \(\{C_{nk} : n, k \in \mathbb{N}\}\) is a cover of \(A\), so

\[
\mu^*(A) \leq \sum_{n,k} \iota(C_{nk}) \leq \sum_{n=1}^{\infty} \mu^*(A_n) + \epsilon.
\]

Since \(\epsilon\) is arbitrary, \(\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)\).

XIII.3.1.8. Definition. The outer measure \(\mu^*\) is the outer measure induced by \(\iota\).

This construction of an outer measure from a set function is often called **Method I.** See (1) for “Method II.”

Note that \(\mu^*(C) \leq \iota(C)\) for any \(C \in C\), since \(\{C\}\) is a cover of \(C\) by \(C\); but \(\mu^*(C)\) does not equal \(\iota(C)\) in general.
XIII.3.1.9. Example. Let $X = \mathbb{N}$, $C = \mathcal{P}(\mathbb{N})$. Set $\iota(A) = 0$ if $A$ is finite and $\iota(A) = 1$ if $A$ is infinite. Then $\iota$ is monotone and (finitely) subadditive, but $\mu^*(A) = 0$ for all $A \subseteq \mathbb{N}$.

An outer measure induced by such a set function has an outer regularity property. If $C$ is a collection of subsets of $X$ containing $\emptyset$, write $C_\sigma$ for the set of countable unions of sets in $C$, and $C_{\sigma \delta}$ for the set of countable intersections of sets in $C_\sigma$. $C_\sigma$ and $C_{\sigma \delta}$ are contained in the $\sigma$-algebra generated by $C$.

XIII.3.1.10. Proposition. Let $C$ be a set of subsets of $X$ with $\emptyset \in C$, $\iota : C \to [0, \infty]$ with $\iota(\emptyset) = 0$, and $\mu^*$ the induced outer measure on $X$. Let $E \subseteq X$ with $\mu^*(E) < \infty$. Then

(i) For every $\epsilon > 0$, there is an $A \in C_\sigma$ with $E \subseteq A$ and $\mu^*(A) \leq \mu^*(E) + \epsilon$.

(ii) There is a $B \in C_{\sigma \delta}$ with $E \subseteq B$ and $\mu^*(E) = \mu^*(B)$.

Proof: (i) Let $\epsilon > 0$. Take a sequence $(C_n)$ in $C$ with $E \subseteq \bigcup_{n=1}^{\infty} C_n$ and $\sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(E) + \epsilon$. If $A = \bigcup_{n=1}^{\infty} C_n$, then $A \in C_\sigma$, $E \subseteq A$, and

$$\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(C_n) \leq \sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(E) + \epsilon.$$ 

(ii): For each $n$ let $A_n \in C_\sigma$ with $E \subseteq A_n$ and $\mu^*(A_n) \leq \mu^*(E) + \frac{\epsilon}{n}$. Take $B = \bigcap_{n=1}^{\infty} A_n$. Then $B \in C_{\sigma \delta}$, $E \subseteq B$, and

$$\mu^*(E) \leq \mu^*(B) \leq \mu^*(A_n) \leq \mu^*(E) + \frac{1}{n}$$

for every $n$, so $\mu^*(B) = \mu^*(E)$. 

XIII.3.2. Lebesgue Outer Measure on $\mathbb{R}$

By far the most important example of the process of inducing an outer measure from a function on subsets is Lebesgue outer measure, used to construct Lebesgue measure.

XIII.3.2.1. Definition. Let $C$ be the set of all open intervals in $\mathbb{R}$ (including $\emptyset$). For an open interval $I$, let $\ell(I)$ be the length of $I$. The induced outer measure from $\ell$ is Lebesgue outer measure $\lambda^*$. Specifically: for $A \subseteq \mathbb{R}$,

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \ell(I_n) : I_n \in C, A \subseteq \bigcup_{n=1}^{\infty} I_n \right\}.$$ 

The use of open intervals is a matter of taste. It is useful to know that the same outer measure is obtained if arbitrary intervals are used:
III.3.2.2. Proposition. Let $\mathcal{I}$ be the set of all intervals in $\mathbb{R}$, including $\emptyset$ and singleton subsets. For $I \in \mathcal{I}$, let $\ell(I)$ be the length of $I \setminus \{a\} = 0$ for a singleton $\{a\}$). Then the outer measure on $\mathbb{R}$ induced by $\ell$ on $\mathcal{I}$ is Lebesgue outer measure.

Proof: Let $\mu^*$ be the outer measure on $\mathbb{R}$ induced by $\ell$ on $\mathcal{I}$. Since $\mathcal{C} \subseteq \mathcal{I}$, any $\mathcal{C}$-cover of a set $A$ is also an $\mathcal{I}$-cover, so the infimum in computing $\mu^*(A)$ is over more covers than for computing $\lambda^*(A)$ and $\mu^*(A) \leq \lambda^*(A)$. To show the opposite inequality, we may assume $\mu^*(A) < \infty$. Let $\epsilon > 0$, and let $\{I_n\}$ be an $\mathcal{I}$-cover of $A$ with $\sum_{n=1}^{\infty} \ell(I_n) \leq \mu^*(A) + \epsilon/2$. Each $I_n$ is an interval (possibly $\emptyset$ or a singleton); for each $n$, let $I'_n$ be an open interval containing $I_n$ of length $\leq \ell(I_n) + 2^{-n-1}\epsilon$. Then $\{I'_n\}$ is a cover of $A$ by open intervals, so we have

$$\lambda^*(A) \leq \sum_{n=1}^{\infty} \ell(I'_n) \leq \sum_{n=1}^{\infty} \ell(I_n) + \frac{\epsilon}{2} \leq \mu^*(A) + \epsilon.$$ 

Since $\epsilon$ is arbitrary, $\lambda^*(A) \leq \mu^*(A)$.

The next corollary can be proved directly in the same way, or by just applying III.3.2.2.

III.3.2.3. Corollary. Let $\mathcal{H}$ be the set of intervals in $\mathbb{R}$ which are open on the left and closed on the right (i.e. of the form $[a, b)$, $(-\infty, b]$, or $(a, +\infty)$), along with the empty set. For $I \in \mathcal{H}$, let $\ell(I)$ be the length of $I$. Then the outer measure on $\mathbb{R}$ induced by $\ell$ on $\mathcal{H}$ is Lebesgue outer measure.

Proof: Let $\nu^*$ be the outer measure induced by $\ell$ on $\mathcal{H}$. Since $\mathcal{H} \subseteq \mathcal{I}$, we have $\mu^*(A) \leq \nu^*(A)$ for any $A \subseteq \mathbb{R}$. If $A \subseteq \mathbb{R}$ and $\{I_n\}$ is a $\mathcal{C}$-cover of $A$, let $I'_n$ be $I_n$ with right endpoint added (if there is one); then $\{I'_n\}$ is an $\mathcal{H}$-cover of $A$, and $\sum_{n=1}^{\infty} \ell(I'_n) = \sum_{n=1}^{\infty} \ell(I_n)$; thus $\nu^*(A) \leq \lambda^*(A)$.

It is easily shown that the class $\mathcal{C}_f$ of bounded open intervals also induces Lebesgue outer measure, as do the classes $\mathcal{I}_f$ and $\mathcal{H}_f$: for any $\delta > 0$, the classes $\mathcal{C}_\delta$ of intervals of length $\leq \delta$ and $\mathcal{I}_\delta$, $\mathcal{H}_\delta$ also induce Lebesgue outer measure on $\mathbb{R}$ (Exercise ).

The next result is crucial, and shows that Lebesgue outer measure has the expected property of extending the interval length function. (In fact, without this theorem it is not even obvious that Lebesgue outer measure is not identically zero.) This result may seem “obvious” but is not, and the proof is somewhat delicate. Some references omit proving this entirely; the proof in [Fol99] (and in some other references) is described as “fussy.” We give a slick proof from [GP10], attributed to von Neumann.

III.3.2.4. Theorem. If $I$ is an interval in $\mathbb{R}$, then $\lambda^*(I) = \ell(I)$.

Proof: We have $\lambda^*(I) \leq \ell(I)$, since $\{I\}$ is a cover of $I$. To show the reverse inequality, first suppose $I = [a, b]$. Let $\epsilon > 0$, and let $(I_k)$ be a sequence of open intervals which cover $[a, b]$ with $\sum_{k=1}^{\infty} \ell(I_k) \leq \lambda^*(I) + \epsilon$ (so the $I_k$ are necessarily bounded). Since $[a, b]$ is compact, $(I_1, \ldots, I_n)$ covers $I$ for some $n$. Fix such an $n$.

If $J$ is a bounded interval in $\mathbb{R}$, write $\#(J)$ for the number of integers in $J$. We have

$$\ell(J) - 1 \leq \#(J) \leq \ell(J) + 1.$$
If $\alpha > 0$, then $\{\alpha I_1, \ldots, \alpha I_n\}$ covers $\alpha I$, so we have

$$\alpha \ell(I) - 1 = \ell(\alpha I) - 1 \leq \sum_{k=1}^{n} \#(\alpha I_k) \leq \sum_{k=1}^{n} \lfloor \ell(\alpha I_k) + 1 \rfloor = \alpha \left[ \sum_{k=1}^{n} \ell(I_k) \right] + n.$$  

Dividing by $\alpha$, we obtain

$$\ell(I) - \frac{1}{\alpha} \leq \left[ \sum_{k=1}^{n} \ell(I_k) \right] + \frac{n}{\alpha}.$$  

This is true for all $\alpha > 0$, so letting $\alpha \to +\infty$ we obtain

$$\ell(I) \leq \sum_{k=1}^{n} \ell(I_k) \leq \sum_{k=1}^{\infty} \ell(I_k) < \lambda^*(I) + \epsilon.$$  

Since $\epsilon > 0$ is arbitrary, $\ell(I) \leq \lambda^*(I)$.

Now suppose $I = (a, b)$. By subadditivity,

$$\ell(I) = b - a = \lambda^*((a, b)) \leq \lambda^*({a}) + \lambda^*({b}) = 0 + \lambda^*(I) + 0.$$  

If $I = (a, b)$ or $I = [a, b]$, then $(a, b) \subseteq I \subseteq [a, b]$, so by monotonicity $\lambda^*(I) = \ell(I)$.

If $I$ has infinite length, then either $(a, \infty) \subseteq I$ or $(-\infty, a) \subseteq I$ for some $a$. If $(a, \infty) \subseteq I$, then $(a, a+n) \subseteq I$ for all $n$, so by monotonicity

$$n = \lambda^*((a, a+n)) \leq \lambda^*(I)$$

for all $n$, and $\lambda^*(I) = \infty = \ell(I)$. The argument if $(-\infty, a) \subseteq I$ is similar.  

\* \* \*  

XIII.3.3. \ The Measure Induced by an Outer Measure

**XIII.3.3.1. Definition.** Let $\mu^*$ be an outer measure on a set $X$. A subset $A$ of $X$ is $\mu^*$-measurable if, for every $Y \subseteq X$,

$$\mu^*(Y) = \mu^*(Y \cap A) + \mu^*(Y \cap A^c) = \mu^*(Y \cap A) + \mu^*(Y \setminus A).$$

Intuitively in the case where $\mu^*$ comes from a set function by Method I, $A$ splits any set $Y$ into two pieces $Y \cap A$ and $Y \cap A^c$ which have “almost disjoint” covers.

Note that the inequality $\mu^*(Y) \leq \mu^*(Y \cap A) + \mu^*(Y \cap A^c)$ holds for any $Y$ and $A$ by subadditivity, so only the reverse inequality is needed for measurability of $A$. This reverse inequality is automatic if $\mu^*(Y) = \infty$, so we obtain:

**XIII.3.3.2. Proposition.** Let $\mu^*$ be an outer measure on a set $X$, and $A \subseteq X$.

(i) $A$ is $\mu^*$-measurable if and only if

$$\mu^*(Y) \geq \mu^*(Y \cap A) + \mu^*(Y \setminus A)$$

for every $Y \subseteq X$ with $\mu^*(Y) < \infty$.

(ii) $A$ is $\mu^*$-measurable if and only if, for every $Y \subseteq A$ and $Z \subseteq X \setminus A$ with $\mu^*(Y)$, $\mu^*(Z)$ finite,

$$\mu^*(Y \cup Z) \geq \mu^*(Y) + \mu^*(Z).$$
XIII.3.3.3.  **COROLLARY.** Let $\mu^*$ be an outer measure on a set $X$. If $A \subseteq X$ and $\mu^*(A) = 0$, then $A$ is $\mu^*$-measurable.

**Proof:** If $Y \subseteq X$, then by monotonicity we have $\mu^*(Y \cap A) = 0$, and

$$\mu^*(Y) \geq \mu^*(Y \cap A^c) = \mu^*(Y \cap A) + \mu^*(Y \cap A^c).$$

The name “measurable” for such sets explained in the introduction, and justified by the following theorem, which is the main result of this section and one of the fundamental theorems of measure theory:

XIII.3.3.4.  **THEOREM.**  [CARATHÉODORY] Let $\mu^*$ be an outer measure on a set $X$, and $\mathcal{M}$ the set of $\mu^*$-measurable subsets of $X$. Then

(i) $\mathcal{M}$ is a $\sigma$-algebra of subsets of $X$.

(ii) The restriction $\mu$ of $\mu^*$ to $\mathcal{M}$ is a measure (i.e. countably additive).

(iii) The measure space $(X, \mathcal{M}, \mu)$ is complete.

**Proof:** It is obvious from the definition of measurability that $\emptyset, X \in \mathcal{M}$ and that $A \in \mathcal{M}$ implies $A^c \in \mathcal{M}$. Let $A, B \in \mathcal{M}$. If $Y \subseteq X$, then

$$Y \cap (A \cup B) = [Y \cap (A \cap B)] \cup [Y \cap (A \cap B^c)] \cup [Y \cap (A^c \cap B)]$$

and so, by subadditivity,

$$\mu^*(Y \cap (A \cup B)) \leq \mu^*(Y \cap (A \cap B)) + \mu^*(Y \cap (A \cap B^c)) + \mu^*(Y \cap (A^c \cap B)).$$

Thus we have

$$\mu^*(Y) = \mu^*(Y \cap A) + \mu^*(Y \cap A^c)$$

$$= \mu^*(Y \cap A \cap B) + \mu^*(Y \cap A \cap B^c) + \mu^*(Y \cap A^c \cap B) + \mu^*(Y \cap A^c \cap B^c)$$

$$\geq \mu^*(Y \cap (A \cup B)) + \mu^*(Y \cap (A^c \cap B^c)) = \mu^*(Y \cap (A \cup B)) + \mu^*(Y \cap (A \cup B)^c)$$

so $A \cup B$ is $\mu^*$-measurable. Therefore $\mathcal{M}$ is an algebra.

Next suppose $A, B \in \mathcal{M}$ are disjoint. We have

$$\mu^*(A \cup B) = \mu^*((A \cup B) \cap A) + \mu^*((A \cup B) \cap A^c) = \mu^*(A) + \mu^*(B)$$

so $\mu^*$ is finitely additive on $\mathcal{M}$.

Now suppose $(A_n)$ is a pairwise disjoint sequence in $\mathcal{M}$, and set $B_n = \cup_{k=1}^n A_k$. Then $B_n \in \mathcal{M}$ for all $n$.

Set $B = \cup_{n=1}^\infty B_n = \cup_{n=1}^\infty A_n$. If $Y \subseteq X$, we have

$$\mu^*(Y) = \mu^*(Y \cap B_n) + \mu^*(Y \cap B_n^c) = \mu^*(Y \cap B_n \cap A_n) + \mu^*(Y \cap B_n \cap A_n^c) + \mu^*(Y \cap B_n^c)$$

$$= \mu^*(Y \cap A_n) + \mu^*(Y \cap B_n) + \mu^*(Y \cap B_n^c) = \cdots = \sum_{k=1}^n \mu^*(Y \cap A_k) + \mu^*(Y \cap B_n) \geq \sum_{k=1}^n \mu^*(Y \cap A_k) + \mu^*(Y \cap B_n^c)$$

1420
for every \( n \), and hence

\[
\mu^*(Y) \geq \sum_{k=1}^{\infty} \mu^*(Y \cap A_k) + \mu^*(Y \cap B^c) \geq \mu^*(Y \cap B) + \mu^*(Y \cap B^c) \geq \mu^*(Y)
\]

and thus all the inequalities are equalities. So \( B \in \mathcal{M} \) and \( \mathcal{M} \) is a \( \sigma \)-algebra (\( \mathcal{A} \)). Applying this last result to \( Y = B \), we obtain \( \mu^*(B) = \sum_{k=1}^{\infty} \mu^*(A_k) \), so \( \mu^* \) is countably additive on \( \mathcal{M} \).

The fact that \( (X, \mathcal{M}, \mu) \) is complete follows immediately from XIII.3.3.3.

XIII.3.3.5. The set \( \mathcal{M} \) of \( \mu^* \)-measurable subsets is more than just a \( \sigma \)-algebra; for example, unlike a general \( \sigma \)-algebra, it is also closed under the Suslin operation (XIII.3.6.7).

XIII.3.3.6. The definition of measurable set we have used is due to Carathéodory, and is now standard. Lebesgue’s definition of a measurable set (within a bounded interval \( I \) of \( \mathbb{R} \) with Lebesgue outer measure) was different and apparently less restrictive: \( A \subseteq I \) is measurable in Lebesgue’s sense if

\[
\lambda^*(A) + \lambda^*(I \setminus A) = \lambda^*(I) = \ell(I).
\]

The two notions are the same in good cases (XIII.3.5.24.), but not in general. Carathéodory’s definition is more generally useful, and even for finite measure spaces gives simpler proofs.

XIII.3.4. Lebesgue Measure on \( \mathbb{R} \)

[Prerequisite: (\( \mathcal{A} \))]

XIII.3.4.1. Definition. Let \( \mathcal{L} \) be the \( \sigma \)-algebra of \( \lambda^* \)-measurable subsets of \( \mathbb{R} \). \( \mathcal{L} \) is called the set of Lebesgue measurable sets in \( \mathbb{R} \). Write \( \lambda \) for the restriction of \( \lambda^* \) to \( \mathcal{L} \). Then \( \lambda \) is a measure on \( (\mathbb{R}, \mathcal{L}) \), called Lebesgue measure on \( \mathbb{R} \).

XIII.3.4.2. Proposition. \( (\mathbb{R}, \mathcal{L}, \lambda) \) is a complete measure space.

This is an immediate consequence of (\( \mathcal{A} \)).

XIII.3.4.3. Proposition. Every Borel set is \( \lambda^* \)-measurable, i.e. \( B \subseteq \mathcal{L} \).

Proof: This follows immediately from (\( \mathcal{A} \)). However, there is a quicker argument. It suffices to show that \( (a, \infty) \) is measurable for any \( a \in \mathbb{R} \) (actually just for \( a \) in a dense subset of \( \mathbb{R} \)), since such sets generate \( \mathcal{B} \). If \( Y \subseteq \mathbb{R} \) with \( \lambda^*(Y) < \infty \), we need to show that

\[
\lambda^*(Y) \geq \lambda^*(Y \cap (a, \infty)) + \lambda^*(Y \cap (-\infty, a]).
\]

Let \( \epsilon > 0 \), and let \( (I_n) \) be a sequence of intervals covering \( Y \) with

\[
\sum_{n=1}^{\infty} \ell(I_n) \leq \lambda^*(Y) + \epsilon.
\]

For each \( n \), let \( I_n' = I_n \cap (a, \infty) \) and \( I_n'' = I_n \cap (-\infty, a] \). Then the \( I_n' \) and \( I_n'' \) are intervals (some may be empty),

\[
\ell(I_n') + \ell(I_n'') = \ell(I_n), \quad (I_n') \text{ is a cover of } Y \cap (a, \infty), \quad \text{and } (I_n'') \text{ is a cover of } Y \cap (-\infty, a],
\]

so

\[
\lambda^*(Y \cap (a, \infty)) + \lambda^*(Y \cap (-\infty, a]) \leq \sum_{n=1}^{\infty} \ell(I_n') + \sum_{n=1}^{\infty} \ell(I_n'') = \sum_{n=1}^{\infty} \ell(I_n) \leq \lambda^*(Y) + \epsilon.
\]

1421
Proof: For each \( n \), \([-n, n]\) is Lebesgue measurable and \( \lambda([-n, n]) = 2n < \infty \), and \( \mathbb{R} = \bigcup_{n \in \mathbb{Z}} [-n, n] \).

It is not true that every Lebesgue measurable set is a Borel set. In fact, the cardinality of \( \mathcal{B} \) is \( 2^{2^{\aleph_0}} \); but every subset of the Cantor set \( K \) is Lebesgue measurable \((\ast)\), and \( \text{card}(K) = 2^{\aleph_0} \), so it follows that \( \text{card}(\mathcal{L}) = 2^{2^{\aleph_0}} \). For an explicit example of a Lebesgue measurable set which is not a Borel set, see \((\ast)\).

However:

XIII.3.4.4. **Corollary.** (\( \mathbb{R}, \mathcal{L}, \lambda \)) is a \( \sigma \)-finite measure space.

Proof: For every \( n \), \([-n, n]\) is Lebesgue measurable and \( \lambda([-n, n]) = 2n < \infty \), and \( \mathbb{R} = \bigcup_{n \in \mathbb{Z}} [-n, n] \).

For any \( \epsilon > 0 \), there is a sequence \( (I_{nm}) \) of open intervals covering \( A \) with
\[
\sum_{m=1}^{\infty} \ell(I_{nm}) < \lambda(A) + 2^{-\lceil |n| \rceil + 2} \epsilon.
\]

If \( U_n = \bigcup_{m=1}^{\infty} I_{nm} \), then \( U_n \) is open, \( A_n \subseteq U_n \), and
\[
\lambda(A_n) \leq \lambda(U_n) \leq \sum_{m=1}^{\infty} \ell(I_{nm}) < \lambda(A) + 2^{-\lceil |n| \rceil + 2} \epsilon
\]
and hence \( \lambda(U_n \setminus A_n) < 2^{-\lceil |n| \rceil + 2} \epsilon \) since \( \lambda(A_n) < \infty \).

Now let \( U = \bigcup_{n \in \mathbb{Z}} U_n \). Then \( U \) is open, \( A \subseteq U \), and \( U \setminus A \subseteq \bigcup_{n \in \mathbb{Z}} (U_n \setminus A_n) \), so
\[
\lambda(U \setminus A) \leq \sum_{n \in \mathbb{Z}} \lambda(U_n \setminus A_n) < \sum_{n \in \mathbb{Z}} 2^{-\lceil |n| \rceil + 2} \epsilon < \epsilon.
\]

(i): Let \( U \) be an open set containing \( A \) such that \( \lambda(U \setminus A^c) < \epsilon/2 \), and set \( E = U^c \). Then \( E \) is closed, \( E \subseteq A \), and \( A \setminus E = U \setminus A^c \), so \( \lambda(A \setminus E) < \epsilon/2 \). If \( \lambda(A) = \infty \), set \( F = E \). Suppose \( \lambda(A) < \infty \). For \( n \in \mathbb{N} \), set \( F_n = E \cap [-n, n] \). Then \( F_n \) is compact, and \( E = \bigcup_n F_n \), so \( A \setminus E = \bigcap_n A \setminus F_n \). This is a decreasing intersection of sets of finite measure, so \( \lambda(A \setminus E) = \inf_n \{ \lambda(A \setminus F_n) \} < \epsilon/2 \). There is thus an \( n \) so that \( \lambda(A \setminus F_n) < \epsilon \). Set \( F = F_n \) for this \( n \).

(ii): Let \( U \) be an open set containing \( A \) such that \( \lambda(U \setminus A) < \epsilon/2 \), and set \( E = U^c \). Then \( E \) is closed, \( E \subseteq A \), and \( A \setminus E = U \setminus A^c \), so \( \lambda(A \setminus E) < \epsilon/2 \). If \( \lambda(A) = \infty \), set \( F = E \). Suppose \( \lambda(A) < \infty \). For \( n \in \mathbb{N} \), set \( F_n = E \cap [-n, n] \). Then \( F_n \) is compact, and \( E = \bigcup_n F_n \), so \( A \setminus E = \bigcap_n A \setminus F_n \). This is a decreasing intersection of sets of finite measure, so \( \lambda(A \setminus E) = \inf_n \{ \lambda(A \setminus F_n) \} < \epsilon/2 \). There is thus an \( n \) so that \( \lambda(A \setminus F_n) < \epsilon \). Set \( F = F_n \) for this \( n \).

(iii): For each \( n \) let \( U_n \) be an open set containing \( A \) with \( \lambda(U_n \setminus A) < 1/n \), and set \( D = \cap_n U_n \).

(iv): For each \( n \) let \( F_n \) be a closed set contained in \( A \) with \( \lambda(A \setminus F_n) < 1/n \), and set \( E = \cup_n F_n \).
XIII.3.4.6. Thus every Lebesgue measurable set $A$ is “trapped between” a $G_\delta$ set $D$ and an $F_\sigma$ set $E$ (i.e. $E \subseteq A \subseteq D$) with $\lambda(D \setminus E) = 0$, and conversely by completeness any set which is trapped between such $D$ and $E$ is Lebesgue measurable. (Note, however, that for a given $A$ there is nothing unique about $D$ and $E$.)

XIII.3.5. Extending Premeasures from Semirings

XIII.3.5.1. We will now show that under some reasonable restrictions, a set function $\iota : \mathcal{C} \to [0, \infty]$ extends to a measure $\mu$ on a $\sigma$-algebra $\mathcal{A}$ containing $\mathcal{C}$. To motivate the restrictions on $\iota$, suppose an extension $\mu$ exists. Then a few moment’s thought shows that the following must be true:

(i) $\iota$ must be monotone on $\mathcal{C}$.

(ii) If $C_1, \ldots, C_n$ are disjoint sets in $\mathcal{C}$, and $C := \bigcup_{k=1}^n C_k \in \mathcal{C}$, then $\iota(C) = \sum_{k=1}^n \iota(C_k)$. (finite additivity).

(iii) If $C, C_1, C_2, \ldots \in \mathcal{C}$ and $C \subseteq \bigcup_{k=1}^\infty C_k$, then $\iota(C) \leq \sum_{k=1}^\infty \iota(C_k)$. (countable subadditivity).

(Note that (ii') does not necessarily imply (ii) unless $\emptyset \in \mathcal{C}$ and $\iota(\emptyset) = 0$.)

It is also reasonable to require that $\mathcal{C}$ be closed under enough set operations to make these restrictions on $\iota$ have some teeth. It turns out to be sufficient that $\mathcal{C}$ be a semiring; in fact, (i) will follow automatically from (ii) and (iii) in this case, and in the presence of (i), (ii') is equivalent to (ii) and (iii) combined.

XIII.3.5.2. Definition. Let $\mathcal{S}$ be a semiring of subsets of a set $X$. A premeasure on $\mathcal{S}$ is a set function $\iota : \mathcal{S} \to [0, \infty]$ with the properties that

(i) $\iota(\emptyset) = 0$.

(ii) If $C_1, \ldots, C_n$ are pairwise disjoint sets in $\mathcal{S}$, and $C := \bigcup_{k=1}^n C_k \in \mathcal{S}$, then $\iota(C) = \sum_{k=1}^n \iota(C_k)$ (finite additivity).

(iii) If $C, C_1, C_2, \ldots \in \mathcal{S}$ and $C \subseteq \bigcup_{k=1}^\infty C_k$, then $\iota(C) \leq \sum_{k=1}^\infty \iota(C_k)$ (countable subadditivity).

The premeasure $\iota$ is $\sigma$-finite if there are $C_1, C_2, \ldots \in \mathcal{S}$ with $\iota(C_n) < \infty$ for all $n$ and $\bigcup_{n=1}^\infty C_n = X$. If $\iota$ is $\sigma$-finite, then $\mathcal{S}$ is necessarily of countable type.

XIII.3.5.3. Examples. (i) Let $(X, \mathcal{A}, \mu)$ be any measure space, and $\mathcal{S}$ a semiring contained in $\mathcal{A}$. Then $\mu|_{\mathcal{S}}$ is a premeasure on $\mathcal{S}$.

(ii) Let $\mathcal{S}$ be any of the following semirings on $\mathbb{R}$:

- All intervals
  - All bounded intervals
  - For fixed $\delta > 0$, all intervals of length $\leq \delta$

(in all cases, include the degenerate intervals $\emptyset$ and singletons). Then $\iota$ is a $\sigma$-finite premeasure on $\mathcal{S}$ (cf. XIII.3.2.4., Exercise (i)).
XIII.3.5.4. **Note:** There is substantial nonuniformity of use of the term “premeasure,” and no consensus on the proper definition of the term. Some authors use it much more broadly than we have, and others more narrowly. I believe our definition of “premeasure” is the most useful one, since it is directly used not only in defining Lebesgue measure and Radon measures (finite Borel measures on $\mathbb{R}$ defined via cumulative distribution functions), but also naturally in defining product measures.

Measure extensions can be defined by more primitive data than a premeasure on a semiring; see [Kön97] for a study.

XIII.3.5.5. **Proposition.** A premeasure on a semiring is monotone.

**Proof:** Let $\mu$ be a premeasure on a semiring $\mathcal{S}$, and $C, D \in \mathcal{S}, C \subseteq D$. Then $D \setminus C$ is a disjoint union of $C_1, \ldots, C_n \in \mathcal{S}$, and

$$\mu(D) = \mu(C) + \sum_{k=1}^{n} \mu(C_k) \geq \mu(C)$$

since $\mu(C_k) \geq 0$ for all $k$.

Premeasures can be characterized alternatively by countable additivity, which is sometimes easier to check:

XIII.3.5.6. **Proposition.** Let $\mathcal{S}$ be a semiring of subsets of a set $X$, and $\mu : \mathcal{S} \to [0, \infty]$. Then $\mu$ is a premeasure if and only if $\mu$ satisfies

(i) $\mu(\emptyset) = 0$.

(ii') If $C_1, C_2, \ldots$ are pairwise disjoint sets in $\mathcal{S}$, and $C := \bigcup_{k=1}^{\infty} C_k \in \mathcal{S}$, then $\mu(C) = \sum_{k=1}^{\infty} \mu(C_k)$ (countable additivity).

Note that (ii') (countable additivity) does not trivially imply (iii) (countable subadditivity) on semirings, even on algebras, since countable additivity only applies to sequences whose union is in the semiring. The proposition is a consequence of XII.1.1.13. and XII.1.1.14., which insure that “sufficiently many” sets in a semiring are finite or countable disjoint unions of sets in the semiring.

**Proof:** Suppose $\mu$ satisfies (i) and (ii'). Then $\mu$ obviously satisfies (ii). Let $C, C_1, C_2, \ldots \in \mathcal{S}$ with $C \subseteq \bigcup_{n=1}^{\infty} C_n$. Then by XII.1.1.14. there are pairwise disjoint $B_k \in \mathcal{S}$, each contained in a unique $C_n$, only finitely many in each $C_n$, such that $C = \bigcup_{k} B_k$. For each $n$, let $D_n$ be the union of the $B_k$ which are in $C_n$. Then $C_n \setminus D_n$ is a finite disjoint union $A_{n1}, \ldots, A_{nm_n}$ of sets in $\mathcal{S}$ by XII.1.1.13.; hence we have

$$\mu(C_n) = \sum_{B_k \subseteq C_n} \mu(B_k) + \sum_{j=1}^{m_n} \mu(A_{nj}) \geq \sum_{B_k \subseteq C_n} \mu(B_k)$$

by (ii). Therefore we have by (ii')

$$\mu(C) = \sum_{k=1}^{\infty} \mu(B_k) = \sum_{n=1}^{\infty} \sum_{B_k \subseteq C_n} \mu(B_k) \leq \sum_{n=1}^{\infty} \mu(C_n)$$

1424
and thus $\iota$ satisfies (iii).

Conversely, if $\iota$ satisfies (i), (ii), and (iii), let $C_1, C_2, \ldots$ be pairwise disjoint sets in $S$ with $C = \bigcup_{n=1}^{\infty} C_n \in S$. We have $\iota(C) \leq \sum_{n=1}^{\infty} \iota(C_n)$ by (iii). On the other hand, if $D_n = \bigcup_{k=1}^{n} C_k$, we have that $C \setminus D_n$ is a finite disjoint union of sets $A_{n1}, \ldots, A_{nm}$ in $S$, so

$$\iota(C) = \sum_{k=1}^{n} \iota(C_k) + \sum_{j=1}^{m} \iota(A_{nj}) \geq \sum_{k=1}^{n} \iota(C_k)$$

for any $n$, i.e. $\iota(C) = \sum_{k=1}^{\infty} \iota(C_k)$.

Premeasures are technically simpler to work with if the domain is a ring instead of just a semiring. (In fact, some authors such as [Hal50] define a measure to be what we call a premeasure on a ring.) We do not make this restriction in the definition since many naturally occurring premeasures are only naturally defined on semirings. However, a premeasure on a semiring uniquely extends to a premeasure on the ring it generates, so there is no loss of generality for extension theory in considering just premeasures defined on rings (as some authors do):

**XIII.3.5.7.** **Theorem.** Let $\iota$ be a premeasure on a semiring $S$ of subsets of $X$, and let $R$ be the ring of subsets of $X$ generated by $S$. Then $\iota$ extends uniquely to a premeasure on $R$.

**Proof:** Recall () that $R$ consists of all finite disjoint unions of sets in $S$. If $\iota$ extends to a premeasure $\kappa$ on $R$, and $R \in R$ is a disjoint union of sets $C_1, \ldots, C_n \in S$, we must have $\kappa(R) = \sum_{k=1}^{n} \iota(C_k)$. We first need to show that this formula gives a well-defined function on $R$.

Suppose $R \in R$ has two representations $R = \bigcup_{k=1}^{n} C_k = \bigcup_{j=1}^{m} D_j$ as disjoint finite unions of sets in $S$. For each $k$ and $j$ set $E_{kj} = C_k \cap D_j$. Then each $E_{kj} \in S$ since $S$ is closed under finite intersections, and the $E_{kj}$ are pairwise disjoint. Thus

$$\sum_{k=1}^{n} \iota(C_k) = \sum_{k=1}^{n} \sum_{j=1}^{m} \iota(E_{kj}) = \sum_{j=1}^{m} \sum_{k=1}^{n} \iota(E_{kj}) = \sum_{j=1}^{m} \iota(D_j)$$

and so $\kappa$ is well defined.

To show that $\kappa$ is a premeasure, we must show (ii) and (iii). For (ii), suppose $R_1, \ldots, R_n \in R$ are pairwise disjoint and suppose $R := \bigcup_{k=1}^{n} R_k \in R$. For each $k$, write $R_k$ as a disjoint union $\bigcup_{j=1}^{m_k} C_{kj}$ with $C_{kj} \in S$. Then

$$\kappa(R) = \sum_{k=1}^{n} \sum_{j=1}^{m_k} \iota(C_{kj}) = \sum_{k=1}^{n} \kappa(R_k).$$

To show (iii), Suppose $R, R_1, R_2, \ldots \in R$ and $R \subseteq \bigcup_{k} R_k$. Write $R_k = \bigcup_{j=1}^{m_k} C_{kj}$ and $R = \bigcup_{i=1}^{r} D_i$ as disjoint unions with $C_{kj}, D_i \in S$. For each $i, j, k$, set $E_{ijk} = C_{kj} \cap D_i$. Then each $E_{ijk} \in S$ and, for each $i, j, k$,

$$\iota(D_i) \leq \sum_{k,j} \iota(E_{ijk}).$$

Then

$$\kappa(R) = \sum_{i=1}^{r} \iota(D_i) \leq \sum_{i,j,k} \iota(E_{ijk}) = \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \sum_{i=1}^{r} \iota(E_{ijk}) \leq \sum_{k=1}^{\infty} \iota(C_{kj}) = \sum_{k=1}^{\infty} \kappa(R_k).$$

1425
The model for constructing an extension of a premeasure to a measure is the construction of Lebesgue measure: begin with the premeasure \( \iota \), form the induced outer measure \( \mu^* \) and the \( \sigma \)-algebra \( \mathcal{M} \) of \( \mu^* \)-measurable subsets; the restriction of \( \mu^* \) to \( \mathcal{M} \) is the desired extension. The precise statement is the next theorem, which is the main result of this section, and another of the most important facts in measure theory:

**THEOREM. [Extension Theorem]** Let \( \mathcal{S} \) be a semiring on a set \( X \), \( \mathcal{A} \) the \( \sigma \)-algebra generated by \( \mathcal{S} \), and \( \iota \) a premeasure on \( \mathcal{S} \). Let \( \mu^* \) be the outer measure on \( X \) induced by \( \iota \), \( \mathcal{M} \) the \( \sigma \)-algebra of \( \mu^* \)-measurable subsets of \( X \), and \( \mu = \mu^*|_{\mathcal{M}} \). Then

(i) \( \mathcal{S} \subseteq \mathcal{M} \), i.e. every set in \( \mathcal{S} \) is \( \mu^* \)-measurable.

(ii) The restriction of \( \mu^* \) to \( \mathcal{S} \) is \( \iota \), i.e. \( \mu \) extends \( \iota \).

Thus \( \iota \) extends to a complete measure \( \mu \) on \( \mathcal{M} \), and hence to a measure \( \mu|_{\mathcal{A}} \) on \( \mathcal{A} \). If \( \mathcal{B} \) is a \( \sigma \)-algebra with \( \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{M} \), the extension \( \mu|_{\mathcal{B}} \) is the largest extension of \( \iota \) to a measure on \( \mathcal{B} \), i.e. any extension \( \nu \) satisfies \( \nu \leq \mu \), and \( \nu \) agrees with \( \mu \) on sets of \( \sigma \)-finite \( \mu \)-measure. If \( \iota \) is \( \sigma \)-finite, then \( \mu \) is the unique extension of \( \iota \) to a measure on \( \mathcal{B} \), and \( (X, \mathcal{M}, \mu) \) is the completion of \( (X, \mathcal{A}, \mu|_{\mathcal{A}}) \).

**PROOF:** We first prove (ii). Let \( C \in \mathcal{S} \). Since \( \{C\} \) is a cover of \( C \), we have \( \mu^*(C) \leq \iota(C) \). For the opposite inequality, we may assume \( \mu^*(C) < \infty \). Let \( \epsilon > 0 \), and let \( (C_n) \) be a sequence of elements of \( \mathcal{S} \) with \( C \subseteq \bigcup_{n=1}^\infty C_n \) and \( \sum_{n=1}^\infty \iota(C_n) \leq \mu^*(C) + \epsilon \). Then, by (iii),

\[
\iota(C) \leq \sum_{n=1}^\infty \iota(C_n) \leq \mu^*(C) + \epsilon .
\]

Since \( \epsilon > 0 \) is arbitrary, \( \iota(C) \leq \mu^*(C) \).

(i): Now let \( C \in \mathcal{S} \), and let \( A \subseteq X \) with \( \mu^*(A) < \infty \). Fix \( \epsilon > 0 \), and let \( (C_n) \) be a sequence in \( \mathcal{S} \) with \( A \subseteq \bigcup_{n=1}^\infty C_n \) and \( \sum_{n=1}^\infty \iota(C_n) \leq \mu^*(A) + \epsilon \). Then \( A \cap C \subseteq \bigcup_{n=1}^\infty (C_n \cap C) \), and \( C_n \cap C \in \mathcal{S} \), so

\[
\mu^*(A \cap C) \leq \sum_{n=1}^\infty \mu^*(C_n \cap C) = \sum_{n=1}^\infty \iota(C_n \cap C) .
\]

Also, \( A \setminus C \subseteq \bigcup_{n=1}^\infty (C_n \setminus C) \). We have that \( C_n \setminus C \) is a finite disjoint union \( \bigcup_{k=1}^m D_{nk} \) for each \( n \), with \( D_{nk} \in \mathcal{S} \). So

\[
\mu^*(A \setminus C) \leq \sum_{n=1}^\infty \sum_{k=1}^m \mu^*(D_{nk}) = \sum_{n=1}^\infty \sum_{k=1}^m \iota(D_{nk}) .
\]

\[
\mu^*(A \cap C) + \mu^*(A \setminus C) \leq \sum_{n=1}^\infty \iota(C_n \cap C) + \sum_{n=1}^\infty \sum_{k=1}^m \iota(D_{nk}) = \sum_{n=1}^\infty \left( \iota(C_n \cap C) + \sum_{k=1}^m \iota(D_{nk}) \right) .
\]

1426
We have that, for each \( n \), \( C_n \) is the finite disjoint union of \( C_n \cap C \) and the \( D_{nk} \), so by (ii)

\[
\iota(C_n) = \iota(C_n \cap C) + \sum_{k=1}^{m_n} \iota(D_{nk})
\]

and thus

\[
\mu^*(A \cap C) + \mu^*(A \setminus C) \leq \sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(A) + \epsilon .
\]

Since \( \epsilon > 0 \) is arbitrary, \( \mu^*(A \cap C) + \mu^*(A \setminus C) \leq \mu^*(A) \). Thus \( C \) is \( \mu^* \)-measurable.

For the last part, suppose \( \nu \) is a measure extending \( \iota \) to a \( \sigma \)-algebra \( B \) with \( A \subseteq B \subseteq \mathcal{M} \). If \( B \in B \), \( \mu(B) < \infty \), let \( \epsilon > 0 \) and \( B \subseteq \cup_{n=1}^{\infty} C_n \) with \( C_n \in S \) and \( \sum_{n=1}^{\infty} \iota(C_n) \leq \mu(B) + \epsilon \). Then

\[
\nu(B) = \sum_{n=1}^{\infty} \nu(C_n) = \sum_{n=1}^{\infty} \iota(C_n) \leq \mu(B) + \epsilon .
\]

Since \( \epsilon \) is arbitrary, \( \nu(B) \leq \mu(B) \). On the other hand, let \( A = \cup_{n=1}^{\infty} C_n \in A \). Then \( \mu(A) \leq \sum_{n=1}^{\infty} \iota(C_n) < \infty \). Since \( \mu(A \setminus B) < \infty \), we have \( \nu(A \setminus B) \leq \mu(A \setminus B) \). Let \( D_1 = C_1 \) and \( D_n = C_n \setminus \cup_{k=1}^{n-1} C_k \). Then \( D_n \) is a finite disjoint union of sets in \( S \), so \( \nu(D_n) = \mu(D_n) \); and \( A \) is the disjoint union of the \( D_n \), so \( \nu(A) = \mu(A) \). Then

\[
\nu(B) = \nu(A) - \nu(A \setminus B) \geq \mu(A) - \mu(A \setminus B) = \mu(B) .
\]

So \( \nu(B) = \mu(B) \), and thus \( \mu = \nu \) for sets in \( B \) of finite \( \mu \)-measure. Then \( \nu = \mu \) also for sets of \( \sigma \)-finite \( \mu \)-measure, which can be written as countable disjoint unions of sets of finite \( \mu \)-measure; so if \( \iota \) is \( \sigma \)-finite, so is \( \mu \) and \( \nu = \mu \). In general, the only possible difference between \( \mu \) and \( \nu \) is that there could be a set \( B \) which is not of \( \sigma \)-finite \( \mu \)-measure, where \( \mu(B) = \infty \) but \( \nu(B) < \infty \). Thus \( \nu \leq \mu \).

For the last statement, let \( (C_n) \) be a sequence in \( S \) with \( \cup_{n=1}^{\infty} C_n = X \) and \( \iota(C_n) < \infty \) for all \( n \). Let \( M \in \mathcal{M} \). Fix \( n \in \mathbb{N} \); then for each \( m \in \mathbb{N} \) there is a sequence \( (D_{nmk}) \) in \( S \) with

\[
M \cap C_n \subseteq \bigcup_{k=1}^{\infty} D_{nmk}
\]

and

\[
\sum_{k=1}^{\infty} \iota(D_{nmk}) < \mu(M \cap C_n) + \frac{1}{m} .
\]

Set

\[
A_{nm} = \bigcup_{k=1}^{\infty} D_{nmk} .
\]

Then \( A_{nm} \in A \), \( M \cap C_n \subseteq A_{nm} \), and

\[
\mu(A_{nm}) < \mu(M \cap C_n) + \frac{1}{m} .
\]

Thus if

\[
A_n = C_n \cap \left( \bigcap_{m=1}^{\infty} A_{nm} \right)
\]

then
we have $A_n \in \mathcal{A}$, $M \cap C_n \subseteq A_n$, and $\mu(A_n) = \mu(M \cap C_n)$. Thus if $B_n = A_n \setminus (M \cap C_n)$, then $B_n \in \mathcal{M}$ and $\mu(B_n) = 0$. Set $A = \bigcup_{n=1}^{\infty} A_n$. Then $A \in \mathcal{A}$, and we have $M \subseteq A$ and

$$B = A \setminus M = \bigcup_{n=1}^{\infty} [A_n \setminus (M \cap C_n)] = \bigcup_{n=1}^{\infty} B_n$$

so $\mu(B) = 0$. Thus $M$ is the union of a set in $\mathcal{A}$ and a set $B$ of $\mu$-measure 0. Repeating the above argument for $M \cap C_n$, there is an $E \in \mathcal{A}$ with $B \subseteq E$ and $\mu(E) = 0$. Thus $M$ is the union of a set in $\mathcal{A}$ and a subset of a null set in $\mathcal{A}$.

There is an extremely important corollary of the uniqueness part:

**Corollary.** Let $\mathcal{S}$ be a semialgebra on $X$, and $\mathcal{A}$ the $\sigma$-algebra generated by $\mathcal{S}$. If $\mu$ and $\nu$ are measures on $\mathcal{A}$ which agree on $\mathcal{S}$ and are $\sigma$-finite on $\mathcal{S}$, then $\mu = \nu$ on $\mathcal{A}$.

Of course, it is enough to check that one of $\mu, \nu$ is $\sigma$-finite on $\mathcal{S}$. Note that it is not sufficient to check that $\mu = \nu$ on $\mathcal{S}$ and that $\mu$, say, is $\sigma$-finite on $\mathcal{A}$, unless we know that $\mu$ is the measure on $\mathcal{A}$ induced by $\mu|_{\mathcal{S}}$: there is an $\mathcal{S}$ on $\mathbb{R}$ which generates the Borel sets as a $\sigma$-algebra, with the property that every nonempty set in $\mathcal{S}$ has infinite Lebesgue measure (Exercise ()); Lebesgue measure is $\sigma$-finite on $\mathcal{B}$ but not on $\mathcal{S}$, and agrees on $\mathcal{S}$ with counting measure and the measure $\mu$ which is infinite on all nonempty sets ($\mu$ is the measure on $\mathcal{B}$ induced by $\lambda|_{\mathcal{S}}$).

**Theorems XIII.3.3.4. and XIII.3.5.8.** are often combined into a single theorem, which is called the **Carathéodory Extension Theorem**, and the construction called the **Carathéodory extension process**, even though **CARATHÉODORY** only proved XIII.3.3.4. (in addition to giving the definition of measurable set (XIII.3.3.1.)). Theorem XIII.3.5.8. was first proved by **FRÉCHET** []; the proof using XIII.3.3.4. is due independently to **HAHN** [] and **KOLMOGOROV** [].

**Theorem.** ([**CARATHÉODORY EXTENSION THEOREM**]) If $\iota$ is a premeasure on a semiring $\mathcal{S}$ of subsets of a set $X$, then there is an extension of $\iota$ to a complete measure $\mu$ on a $\sigma$-algebra $\mathcal{M}$ containing $\mathcal{S}$. If $\iota$ is $\sigma$-finite on $\mathcal{S}$, then the extension $\mu$ is unique if $\mathcal{M}$ is taken to be the $\sigma$-algebra generated by $\mathcal{S}$ and the null sets.

It is interesting to see what this construction gives when applied to the case where $\mathcal{S}$ is a $\sigma$-algebra and $\iota$ is a measure on $\mathcal{S}$:

**Proposition.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $\mu^*$ be the outer measure induced by $\mu$, $\mathcal{M}$ the $\sigma$-algebra of $\mu^*$-measurable sets, and $\nu = \mu^*|_{\mathcal{M}}$. Then $(X, \mathcal{M}, \nu)$ is the saturation () of the completion (or the completion of the saturation) of $(X, \mathcal{A}, \mu)$. In particular, if $\mu$ is $\sigma$-finite, then $(X, \mathcal{M}, \nu)$ is the completion of $(X, \mathcal{A}, \mu)$.
Regular Outer Measures

Not every outer measure is induced by a premeasure. The ones that are can be characterized in a simple way:

XIII.3.5.13. Definition. An outer measure \( \mu^* \) on a set \( X \) is regular if, for every \( E \subseteq X \), there is a \( \mu^* \)-measurable set \( A \) with \( E \subseteq A \) and \( \mu^*(E) = \mu^*(A) \).

Not every outer measure is regular:

XIII.3.5.14. Examples. (i) If \( X = \{0,1\} \), set \( \mu^*(\emptyset) = 0 \), \( \mu^*(\{0\}) = \mu^*(\{1\}) = 1 \), and \( \mu^*(X) = \alpha \). If \( 1 \leq \alpha \leq 2 \), \( \mu^* \) is an outer measure. If \( \alpha < 2 \), the only \( \mu^* \)-measurable sets are \( \emptyset \) and \( X \); if \( 1 < \alpha < 2 \), \( \mu^* \) is not regular.

(ii) Let \( X = \mathbb{N} \). Set \( \mu^*(\emptyset) = 0 \), \( \mu^*(E) = 1 \) if \( E \) is nonempty and finite, and \( \mu^*(E) = 2 \) if \( E \) is infinite. The only \( \mu^* \)-measurable sets are \( \emptyset \) and \( \mathbb{N} \), and \( \mu^* \) is not regular.

XIII.3.5.15. There is a related property. If \( \mu^* \) is an outer measure on \( X \) and \( (E_n) \) is an increasing sequence of subsets of \( X \), with \( E = \bigcup_{n=1}^{\infty} E_n \), then \( \mu^*(E_1) \leq \mu^*(E_2) \leq \cdots \leq \mu^*(E) \) by monotonicity; thus

\[
\mu^*(E) \geq \lim_{n \to \infty} \mu^*(E_n) = \sup_n \mu^*(E_n).
\]

We do not have equality in general (XIII.3.5.17).

XIII.3.5.16. Definition. Let \( \mu^* \) be an outer measure on a set \( X \). Then \( \mu^* \) is monotone regular if, whenever \( (E_n) \) is an increasing sequence of subsets of \( X \), we have

\[
\mu^*(E) = \lim_{n \to \infty} \mu^*(E_n) = \sup_n \mu^*(E_n).
\]

XIII.3.5.17. Examples. The example of XIII.3.5.14.(ii) is not monotone regular: let \( E_n = \{1,\ldots,n\} \).

The example of XIII.3.5.14.(i) is monotone regular (as is any outer measure on a finite set) but not regular.

XIII.3.5.18. Proposition. A regular outer measure is monotone regular.

Proof: Let \( \mu^* \) be a regular outer measure on \( X \), and \( (E_n) \) an increasing sequence of subsets of \( X \), with \( E = \bigcup_{n=1}^{\infty} E_n \). It suffices to show that \( \mu^*(E) \leq \sup_n \mu^*(E_n) \). Let \( A_n \) be a \( \mu^* \)-measurable set with \( E_n \subseteq A_n \) and \( \mu^*(A_n) = \mu^*(E_n) \) for all \( n \). Set \( B_n = \cap_{k=n}^{\infty} A_k \). We have that \( E_n \subseteq A_k \) whenever \( n \leq k \), and hence \( E_n \subseteq B_n \subseteq B_{n+1} \) for all \( n \). Since \( B_n \subseteq A_n \), we have \( \mu^*(E_n) \leq \mu^*(B_n) \leq \mu^*(A_n) = \mu^*(E_n) \), so we have equality throughout. If \( B = \bigcup_{n=1}^{\infty} B_n \), then \( E \subseteq B \), so by () we have

\[
\mu^*(E) \leq \mu^*(B) = \sup_n \mu^*(B_n) = \sup_n \mu^*(E_n).
\]
**XIII.3.5.19.** Proposition. An outer measure induced by a premeasure is regular.

**Proof:** Suppose the outer measure \( \mu^* \) on \( X \) is induced by \( \nu : S \to [0, \infty] \), and let \( E \subseteq X \). If \( \mu^*(E) = \infty \), take \( A = X \). If \( \mu^*(E) < \infty \), by XIII.3.1.10, let \( A \in \mathcal{S}_\sigma \delta \) with \( E \subseteq A \) and \( \mu^*(A) = \mu^*(E) \).

**XIII.3.5.20.** Corollary. Lebesgue outer measure is a regular outer measure.

For the converse, we prove a more precise statement:

**XIII.3.5.21.** Proposition. Let \( \mu^* \) be an outer measure on \( X \), \( \mathcal{M} \) the set of \( \mu^* \)-measurable sets, and \( \mu^* \) the outer measure induced by \( \mu^*|_M \). Then \( \mu^* \) is a regular outer measure, \( \mu^* \geq \mu^* \), every set in \( \mathcal{M} \) is \( \mu^* \)-measurable, every \( \mu^* \)-measurable set of \( \sigma \)-finite \( \mu^* \)-measure is in \( \mathcal{M} \), and \( \mu^*|_M = \mu^*|_M \); \( \mu^* = \mu^* \) if and only if \( \mu^* \) is regular.

**Proof:** Most of this follows immediately from previous results. By XIII.3.5.8., setting \( \mathcal{S} = \mathcal{M} \) and \( \nu = \mu^*|_M \), every set in \( \mathcal{M} \) is \( \mu^* \)-measurable, and \( \mu^*|_M = \mu^*|_M \). The outer measure \( \mu^* \) is regular by XIII.3.5.19. We have that, for any \( E \subseteq X \),

\[
\mu^*(E) = \inf \{ \mu^*(A) : A \in \mathcal{M}, E \subseteq A \}
\]

and the infimum is attained \( () \) at some \( A \in \mathcal{M} \); for this \( A \) we have \( \mu^*(E) \leq \mu^*(A) = \mu^*(E) \). Thus \( \mu^* \geq \mu^* \).

If \( E \) is \( \mu^* \)-measurable and \( \mu^*(E) < \infty \), then there is a \( \mu^* \)-measurable set \( A \) with \( E \subseteq A \) and \( \mu^*(A) = \mu^*(E) \); then \( \mu^*(A) = \mu^*(A) = \mu^*(E) \), so \( \mu^*(A \setminus E) = 0 \). Since \( \mu^*(A \setminus E) \leq \mu^*(A \setminus E) \), \( \mu^*(A \setminus E) = 0 \) and \( A \setminus E \in \mathcal{M} () \). Thus \( E = A \setminus (A \setminus E) \in \mathcal{M} \). The case where \( E \) has \( \sigma \)-finite \( \mu^* \)-measure follows immediately.

Finally, if \( \mu^* \) is regular, let \( E \subseteq X \), and \( A \in \mathcal{M} \) with \( E \subseteq A \) and \( \mu^*(E) = \mu^*(A) \). Then \( \{ A \} \) is an \( \mathcal{M} \)-cover of \( E \), so \( \mu^*(E) \leq \mu^*(A) = \mu^*(E) \). Thus \( \mu^* \leq \mu^* \). So \( \mu^* = \mu^* \).

**XIII.3.5.22.** Example. [Rog70, p. 14] The \( \sigma \)-algebra of \( \mu^* \)-measurable sets can be strictly larger than \( \mathcal{M} \). Let \( X \) be a set of cardinality \( \aleph_2 \), and define \( \mu^* \) on \( X \) by setting \( \mu^*(E) = 0 \) if \( E \) is countable, \( \mu^*(E) = 1 \) if \( \text{card}(E) = \aleph_1 \), and \( \mu^*(E) = \infty \) if \( \text{card}(E) = \aleph_2 \). Then \( \mu^* \) is an outer measure on \( X \) which is not regular. The \( \sigma \)-algebra \( \mathcal{M} \) of \( \mu^* \)-measurable sets is the \( \sigma \)-algebra of countable and co-countable sets. Construct \( \mu^* \) as above from \( \mu^*|_M \). Then, for any \( E \subseteq X \), \( \mu^*(E) = 0 \) if \( E \) is countable and \( \mu^*(E) = \infty \) if \( E \) is uncountable, and every subset of \( X \) is \( \mu^* \)-measurable.

Putting the last two results together, we obtain:

**XIII.3.5.23.** Proposition. An outer measure is induced by a premeasure if and only if it is regular.

As a corollary of XIII.3.5.21., we can show that Lebesgue’s definition of measurability agrees with Carathéodory’s for finite regular outer measures:
XIII.3.5.24. Theorem. Let $\mu^*$ be a regular outer measure on a set $X$, with $\mu^*(X) < \infty$. Then $A \subseteq X$ is $\mu^*$-measurable if and only if $\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)$.

(Of course, $\mu^*(A) + \mu^*(X \setminus A) \geq \mu^*(X)$ is automatic; so only the reverse inequality needs to be checked.)

Proof: If $A$ is $\mu^*$-measurable, the equality holds by definition. Conversely, suppose $A \subseteq X$ and $\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)$. Let $B$ and $C$ be $\mu^*$-measurable subsets of $X$ with $A \subseteq B$, $\mu^*(B) = \mu^*(A)$, $(X \setminus A) \subseteq C$, $\mu^*(C) = \mu^*(X \setminus A)$. We have

$$\mu^*(B \setminus C) + \mu^*(B \cap C) + \mu^*(C \setminus B) = \mu^*(B \cup C) = \mu^*(X).$$

We also have

$$\mu^*(X) = \mu^*(A) + \mu^*(X \setminus A) = \mu^*(B) + \mu^*(C) = \mu^*(B \setminus C) + 2\mu^*(B \cap C) + \mu^*(C \setminus B)$$

and subtracting we get $\mu^*(B \cap C) = 0$. If $D = B \setminus A$, then $D = B \cap (X \setminus A) \subseteq (B \cap C)$, so $\mu^*(D) = 0$ and $D$ is $\mu^*$-measurable. Hence so is $A = B \setminus D$.

This result can fail for finite outer measures which are not regular (Exercise 1). It also usually fails for outer measures which are not finite: for example, there is a nonmeasurable subset $A$ of $\mathbb{R}$ with $\lambda^*(A) = \lambda^*(\mathbb{R} \setminus A) = \infty$.

Inner Measure

XIII.3.5.25. If $\mu^*$ is a finite regular outer measure on a set $X$, we can define the inner measure of a subset $A$ of $X$ by

$$\mu_+(A) = \mu^*(X) - \mu^*(X \setminus A).$$

Then the result of XIII.3.5.24. can be rephrased:

XIII.3.5.26. Corollary. Let $\mu^*$ be a finite regular outer measure on a set $X$, and $\mu_+$ the corresponding inner measure. Then $A \subseteq X$ is $\mu^*$-measurable if and only if $\mu^*(A) = \mu_+(A)$.

Generalizations of Regularity

XIII.3.5.27. The notion of regularity can be generalized. If $\mathcal{C}$ is a collection of subsets of a set $X$, and $\mu^*$ is an outer measure on $X$, then $\mu^*$ is $\mathcal{C}$-regular if every subset $Y$ of $X$ is contained in some $C \in \mathcal{C}$ with $\mu^*(C) = \mu^*(Y)$ (this implies that $X \in \mathcal{C}$). Normally this definition is applied only when $\mathcal{C} \subseteq \mathcal{M}$, where $\mathcal{M}$ is the set of $\mu^*$-measurable sets. Thus $\mu^*$ is regular if and only if it is $\mathcal{M}$-regular. If $\mu^*$ is a $\sigma$-finite outer measure induced from $\iota : \mathcal{C} \rightarrow [0, \infty]$, then $\mu^*$ is $\mathcal{C}_{\sigma\delta}$-regular (XIII.3.1.10.).

XIII.3.5.28. Proposition. Let $\mu^*$ be an outer measure on a set $X$, $\mathcal{M}$ the set of $\mu^*$-measurable subsets of $X$, and $\mathcal{A}$ a sub-$\sigma$-algebra of $\mathcal{M}$ such that $\mu^*|_{\mathcal{A}}$ is $\sigma$-finite. If $\mu^*$ is $\mathcal{A}$-regular, then $\mathcal{M}$ is the $\sigma$-algebra generated by $\mathcal{A}$ and the $\mathcal{A}$-null sets (subsets of sets in $\mathcal{A}$ of outer measure 0). In particular, if $(X, \mathcal{A}, \mu^*|_{\mathcal{A}})$ is complete and $\sigma$-finite, then $\mathcal{A} = \mathcal{M}$. 

1431
XIII.3.5.29. This does not mean that if \((X, A, \mu)\) is a \(\sigma\)-finite complete measure space, then it is not possible to extend \(\mu\) to a measure on a strictly larger \(\sigma\)-algebra; there are many examples where this is possible. See Exercise () for a simple example, and () for more natural and important ones. What XIII.3.5.28. does say is that if such a \(\mu\) is extended to a measure \(\nu\) on a larger \(\sigma\)-algebra \(B\), and \(\nu^*\) is the outer measure induced by \(\nu\), then \(\nu^*\) is not \(A\)-regular.

XIII.3.5.30. An important special case of regularity is when \(\mu^*\) is an outer measure on a topological space \(X\). Then \(\mu^*\) is Borel regular if it is \(B\)-regular, where \(B\) is the class of Borel sets in \(X\). This definition is usually reserved for Borel outer measures, i.e. ones where all Borel sets are \(\mu^*\)-measurable. The most important example is that Lebesgue outer measure is Borel regular (XIII.3.4.5., XIII.3.5.20.).

XIII.3.6. The Suslin Operation

XIII.3.6.1. There is a curious operation on sets called the Suslin operation (sometimes called the \(A\)-operation), which is reminiscent of the construction in the proof of XI.10.2.7.. For each finite sequence \((s_1, \ldots, s_n)\) of natural numbers, let \(A_{s_1, \ldots, s_n}\) be a subset of a set \(X\). The collection of finite sequences in \(N\) is countable, so there are countably many such sets in all. Define

\[
S(\{A_{s_1, \ldots, s_n}\}) = \bigcup_{s \in \mathcal{N}} \bigcap_{n=1}^{\infty} A_{s_1, \ldots, s_n}
\]

where \(\mathcal{N} = \mathbb{N}^\infty\) is the set of all sequences in \(\mathbb{N}\). Although the operation begins with a countable collection of sets, the union is over all sequences (uncountably many) in \(\mathcal{N}\).

If \(s = (s_1, s_2, \ldots) \in \mathcal{N}\), write \(s|n = (s_1, \ldots, s_n) \in \mathbb{N}^n\). These are called initial segments of \(s\). Take \(s|0\) to be the “empty string.” Thus

\[
S(\{A_{s_1, \ldots, s_n}\}) = \bigcup_{s \in \mathcal{N}} \bigcap_{n=1}^{\infty} A_{s|n}.
\]

XIII.3.6.2. More generally, if \(\mathcal{E}\) is a nonempty collection of subsets of a set \(X\), and all the \(A_{s_1, \ldots, s_n}\) are in \(\mathcal{E}\), the indexed collection \(\{A_{s_1, \ldots, s_n}\}\) is called a Suslin scheme in \(\mathcal{E}\). Denote by \(S(\mathcal{E})\) the collection \(\{S(\{A_{s_1, \ldots, s_n}\})\}\), for all Suslin schemes \(\{A_{s_1, \ldots, s_n}\}\) in \(\mathcal{E}\), along with \(\emptyset\) (we will automatically have \(\emptyset \in S(\mathcal{E})\) in many cases, e.g. if \(\emptyset\) is a countable intersection of sets in \(\mathcal{E}\)). The sets in \(S(\mathcal{E})\) are called the \(\mathcal{E}\)-Suslin sets (sometimes called the \(\mathcal{E}\)-analytic sets, but we reserve this term for sets defined differently in ()); the two classes of sets coincide in most natural settings. It is obvious that if \(\mathcal{E} \subseteq \mathcal{F}\), then \(S(\mathcal{E}) \subseteq S(\mathcal{F})\).

We typically take \(\mathcal{E}\) to be the collection of closed or compact sets in \(\mathbb{R}^n\) (or in a topological space), or a semialgebra (algebra, \(\sigma\)-algebra) of sets such as the Borel sets in \(\mathbb{R}^n\).

The sets in \(S(\mathcal{E})\) have a simpler explicit description than the sets in \(\sigma(\mathcal{E})\) (cf. ()). The collection \(S(\mathcal{E})\) is usually different from \(\sigma(\mathcal{E})\) although it has some of the same properties; it frequently happens that \(S(\mathcal{E})\) contains \(\sigma(\mathcal{E})\) (XIII.3.6.4.(iv)).

XIII.3.6.3. A Suslin scheme \(\{A_{s_1, \ldots, s_n}\}\) is monotone (or regular) if \(A_{s_1, \ldots, s_n, s_{n+1}} \subseteq A_{s_1, \ldots, s_n}\) for all \((s_1, \ldots, s_{n+1})\). If \(\mathcal{E}\) is closed under finite intersections (e.g. in one of the common situations described above), then for any Suslin scheme \(\{A_{s_1, \ldots, s_n}\}\) in \(\mathcal{E}\) there is a monotone Suslin scheme \(\{B_{s_1, \ldots, s_n}\}\) with \(S(\{B_{s_1, \ldots, s_n}\}) = S(\{A_{s_1, \ldots, s_n}\})\); just set

\[
B_{s_1, \ldots, s_n} = A_{s_1} \cap A_{s_1, s_2} \cap \cdots \cap A_{s_1, \ldots, s_n}
\]

1432
for each \((s_1, \ldots, s_n)\).

**XIII.3.6.4. Proposition.** Let \(E\) be a collection of subsets of a set \(X\). Then

(i) \(E \subseteq S(E)\).

(ii) \(S(S(E)) = S(E)\). Thus if \(E \subseteq F \subseteq S(E)\), then \(S(F) = S(E)\).

(iii) \(S(E)\) is closed under countable unions and countable intersections.

(iv) If the complement of every set in \(E\) is in \(S(E)\), then \(\sigma(E) \subseteq S(E)\).

**Proof:** (i): If \(A \in E\), set \(A_{s_1, \ldots, s_n} = A\) for all \((s_1, \ldots, s_n)\). Then \(\{A_{s_1, \ldots, s_n}\}\) is a Suslin scheme in \(E\), and 
\[S(\{A_{s_1, \ldots, s_n}\}) = A.\]

(ii): Let \(\phi : \mathbb{N}^2 \to \mathbb{N}\) be the bijection given by \(\phi(n, m) = 2^{n-1}(2m - 1)\). Let \(\alpha\) and \(\beta\) be the coordinate functions of the inverse map, i.e. \(\alpha(2^{n-1}(2m - 1)) = n\), \(\beta(2^{n-1}(2m - 1)) = m\). Define a bijection \(\psi : \mathbb{N} \times \mathbb{N}^\infty \to \mathbb{N}\) by
\[\psi(s, (t^{(k)})) = \phi(s, (t^{(\alpha(r))}))\]
where \(t^{(k)}_j\) is the \(j\)th term of the sequence \(t^{(k)}\). Note that if \(r = \phi(n, m)\), then \(\psi(s, (t^{(k)}))r\) completely determines \(s\) and \((t^{(n)})m\), i.e. if \((p_1, \ldots, p_r) \in \mathbb{N}^r\), then there are \((s_1, \ldots, s_n)\) and \((t_1, \ldots, t_m)\) such that whenever \(\psi(s, (t^{(k)}))r = (p_1, \ldots, p_r), s = (s_1, \ldots, s_n)\) and \((t^{(n)})m = (t_1, \ldots, t_m)\).

Now let \(B = \bigcup_{s \in \mathbb{N}} \bigcap_{p \in \mathbb{N}} B_{s, p} \in S(S(E))\), where \(\{B_{s, p}\}\) is a Suslin scheme in \(S(E)\), and suppose for each \((s_1, \ldots, s_n)\) we have \(B_{s_1, \ldots, s_n} = \bigcap_{s \in \mathbb{N}} A_{s_1, \ldots, s_n}^{(n)}\) with \(A_{s_1, \ldots, s_n}^{(n)} \in E\). For a finite sequence \((p_1, \ldots, p_r)\), let \((s_1, \ldots, s_n)\) and \((t_1^{(n)}, \ldots, t_m^{(n)})\) be the initial segments determined by \((p_1, \ldots, p_r)\) as above, and define
\[E_{p_1, \ldots, p_r} = A_{s_1, \ldots, s_n}^{(n)} = A_{s_1, \ldots, s_n}^{(n)}.\]
Then \(\{E_{p_1, \ldots, p_r}\}\) is a Suslin scheme in \(E\); set \(E = \bigcup_{p \in \mathbb{N}} \bigcap_{r \in \mathbb{N}} E_{p_1, \ldots, p_r} \in S(E)\).

We have that \(x \in B\) if and only if there is an \(s \in \mathbb{N}\) such that \(x \in B_{s, p}\) for all \(p\), i.e. if and only if for each \(n\) there is \(n \in \mathbb{N}\) such that \(x \in A_{s, n}^{(n)}\) for all \(m\). Set \(p = \psi(s, (t^{(n)}))\). Then, for any \(r\), \(E_{p_1, \ldots, p_r} = A_{s, n}^{(n)}\) where \(r = \phi(n, m)\), so \(x \in B\) if and only if \(x \in E_{p_1, \ldots, p_r}\) for all \(p\), for some \(p \in \mathbb{N}\), i.e. if and only if \(x \in E\). Thus \(B = E \in S(E)\). So \(S(S(E)) \subseteq S(E)\), and the opposite inclusion follows from (i).

(iii): By (ii), it suffices to show that \(S(E)\) contains all countable unions and countable intersections of sets in \(E\). Suppose \((E_n)\) is a sequence of sets in \(E\). For each \((s_1, \ldots, s_n)\), set \(A_{s_1, \ldots, s_n} = E_{s_1}\). Then \(\{A_{s_1, \ldots, s_n}\}\) is a Suslin scheme in \(E\), and \(S(\{A_{s_1, \ldots, s_n}\}) = \bigcup_{n=1}^\infty E_n\). Similarly, set \(A_{s_1, \ldots, s_n} = E_{s_1}\) for all \((s_1, \ldots, s_n)\); then \(S(\{A_{s_1, \ldots, s_n}\}) = \bigcap_{n=1}^\infty E_n\).

(iv): Follows immediately from **XIII.3.10.11.** (cf. **XIII.3.10.11.**), since \(E \subseteq S(E) \cap \widehat{S(E)}\) which is a \(\sigma\)-algebra.

**XIII.3.6.5.** \(S(E)\) is not closed under complements in general, even if \(E\) is a \(\sigma\)-algebra. For example, if \(B\) is the \(\sigma\)-algebra of Borel sets in \(\mathbb{R}\), then \(S(B)\) is not closed under complements ( ).
XIII.3.6.6. Let $\mathcal{K}$ be the collection of compact subsets of $\mathbb{R}^n$, $\mathcal{F}$ the collection of closed subsets, and $\mathcal{B}$ the collection of Borel sets. The complement of every compact set is open, hence a countable union of compact sets; thus by XIII.3.4.4(iv) $\sigma(\mathcal{K}) = \mathcal{B} \subseteq \mathcal{S}(\mathcal{K})$. Thus by (ii) $\mathcal{S}(\mathcal{K}) = \mathcal{S}(\mathcal{F}) = \mathcal{S}(\mathcal{B})$, i.e. every $\mathcal{B}$-Suslin set is $\mathcal{K}$-Suslin. These sets are called the Suslin sets in $\mathbb{R}^n$. See also Exercise XIII.3.10.9.

The next theorem is the main result of this subsection.

XIII.3.6.7. **Theorem.** Let $\mu^*$ be an outer measure on a set $X$, $\mathcal{M}$ the $\sigma$-algebra of $\mu^*$-measurable sets, and $\mu = \mu^*|_\mathcal{M}$. Then

(i) $\mathcal{M}$ is closed under the Suslin operation, i.e. $\mathcal{S}(\mathcal{M}) = \mathcal{M}$.

(ii) Suppose $\mu^*$ is monotone regular. Let $\mathcal{E}$ be a subset of $\mathcal{M}$ which is closed under finite unions and countable intersections, and such that $\mu(E) < \infty$ for all $E \in \mathcal{E}$. Then, for every $A \in \mathcal{S}(\mathcal{E})$ with $\mu(A) < \infty$, we have

$$\mu(A) = \sup \{ \mu(E) : E \in \mathcal{E}, E \subseteq A \} .$$

Combining XIII.3.6.7. with XIII.3.5.12., we obtain:

XIII.3.6.8. **Corollary.** Let $(X, \mathcal{A}, \mu)$ be a complete $\sigma$-finite measure space. Then $\mathcal{A}$ is closed under the Suslin operation. (“$\sigma$-finite” can be replaced by “saturated.”)

XIII.3.6.9. The conclusion of this corollary can be false if the measure is not complete. In fact, the $\sigma$-algebra of Borel sets on $\mathbb{R}$ is not closed under the Suslin operation ().

We now prove Theorem XIII.3.6.7.

**Proof:** ([Rog70], [Bog07]) We first prove (i) and (ii) in the case where $\mu^*$ is monotone regular. Let $A = \mathcal{S}\{\{A_{s_1, \ldots, s_n}\} \in \mathcal{S}(\mathcal{M})$, where $\{A_{s_1, \ldots, s_n}\}$ is a Suslin scheme in $\mathcal{M}$ which we may assume is monotone since $\mathcal{M}$ is closed under finite intersections. For (ii), we take the $A_{s_1, \ldots, s_n}$ to be in $\mathcal{E}$. To show that $A \in \mathcal{M}$, we use the criterion of XIII.3.3.2.(ii). So let $Y \subseteq A$ and $Z \subseteq X \setminus A$ with $\mu^*(Y), \mu^*(Z) < \infty$; in (ii) we take $Y = A$ (and in this case we can take $Z = \emptyset$). For (i) we must show that $\mu^*(Y \cup Z) \geq \mu^*(Y) \cdot \mu^*(Z)$. We will show that for any $\epsilon > 0$ there is an $E \in \mathcal{M}$, $E \subseteq A$, with $\mu^*(Y \cap E) \geq \mu^*(Y) - \epsilon$, and $E \in \mathcal{E}$ in case (ii); (ii) will then follow immediately, and (i) will also follow since then $Z \subseteq X \setminus E$, so

$$\mu^*(Y \cup Z) \geq \mu^*(Y \cap E \cup Z) = \mu^*(Y \cap E) \cdot \mu^*(Z) \geq \mu^*(Y) - \epsilon + \mu^*(Z)$$

and since $\epsilon > 0$ is arbitrary, $\mu^*(Y \cup Z) \geq \mu^*(Y) + \mu^*(Z)$.

Fix $\epsilon > 0$. For $m \in \mathbb{N}$, set

$$B_m = Y \cap \bigcup_{s \in \mathcal{N}_m} \bigcap_{k \geq 1} A_{s_{1:k}}$$

where $\mathcal{N}_m = \{s \in \mathcal{N} : s_1 \leq m\}$. Then the $B_m$ increase with $m$, and $\bigcup_m B_m = Y$. So

$$\mu^*(Y) = \sup_m \mu^*(B_m)$$

1434
since $\mu^*$ is monotone regular. Thus there is an $m_1$ such that

$$\mu^*(B_{m_1}) > \mu^*(Y) - \epsilon.$$  

Proceed inductively. Suppose $m_1, \ldots, m_n$ have been defined and

$$B_{m_1, \ldots, m_n} = Y \cap \bigcup_{s \in \mathcal{N}_{m_1, \ldots, m_n}} \bigcap_k A_{s[k]}$$

where $\mathcal{N}_{m_1, \ldots, m_n} = \{s \in \mathcal{N} : s_1 \leq m_1, \ldots, s_n \leq m_n\}$ and

$$\mu^*(B_{m_1, \ldots, m_n}) > \mu^*(Y) - \epsilon.$$  

For $m \in \mathcal{N}$, set

$$B_{m_1, \ldots, m_n, m} = Y \cap \bigcup_{s \in \mathcal{N}_{m_1, \ldots, m_n, m}} \bigcap_k A_{s[k]}$$

where $\mathcal{N}_{m_1, \ldots, m_n, m} = \{s \in \mathcal{N} : s_1 \leq m_1, \ldots, s_n \leq m_n, s_{n+1} \leq m\}$ Then the $B_{m_1, \ldots, m_n, m}$ increase in $m$, and

$$\mu^*(B_{m_1, \ldots, m,n+1}) > \mu^*(Y) - \epsilon.$$  

Now for each $n$ set

$$E_n = \bigcup_{s \in \mathcal{N}_{m_1, \ldots, m_n}} A_{s[n]}.$$  

This is a finite union, so $E_n \in \mathcal{M}$ for all $n$, and $E_n \in \mathcal{E}$ in case (ii). We have $B_{m_1, \ldots, m_n} \subseteq Y \cap E_n$ for all $n$. We also have $E_{n+1} \subseteq E_n$ for all $n$. Set $E = \cap_n E_n$.  

We claim that $\mu^*(E) \geq \mu^*(Y) - \epsilon$. This is obvious from downward continuity () in case (ii) since $\mu$ is finite on $\mathcal{E}$ and $\mu^*(E_n) \geq \mu^*(B_{m_1, \ldots, m_n}) > \mu^*(Y) - \epsilon$ for all $n$. For case (i), define a new outer measure $\mu_Y$ on $X$ by

$$\mu_Y(S) = \mu^*(Y \cap S)$$

for $S \subseteq X$. Then every set in $\mathcal{M}$ is $\mu_Y$-measurable by (), and we have

$$\mu^*(E) \geq \mu_Y^*(E) = \inf_n \mu^*(E_n) = \inf_n \mu^*(Y \cap E_n) \geq \inf_n \mu^*(B_{m_1, \ldots, m_n}) \geq \mu^*(Y) - \epsilon$$

where the first equality is downward continuity for the finite measure $\mu_Y$.

Finally, we show $E \subseteq A$. Let $x \in E$. Write

$$\mathbb{N}_{m_1, \ldots, m_n}^n = \{(s_1, \ldots, s_n) : s_1 \leq m_1, \ldots, s_n \leq m_n\} \subseteq \mathbb{N}^n$$

for each $n$. If $k > n$, say $(s_1, \ldots, s_n) \in \mathbb{N}_{m_1, \ldots, m_n}^n$ is extendible to $k$ if there are $s_{n+1}, \ldots, s_k$ such that $(s_1, \ldots, s_n, s_{n+1}, \ldots, s_k) \in \mathbb{N}_{m_1, \ldots, m_n}^k$, and $x \in A_{s_1, \ldots, s_k}$. Since $x \in E_k$ for all $k$, there is for every $k$ an $s_1$ for which $(s_1)$ is extendible to $k$. Since there are only finitely many $s_1 \leq m_1$, there is an $s_1$ such that $(s_1)$ is extendible to arbitrarily large $k$. Similarly, for this $s_1$ there is an $s_2$ such that $(s_1, s_2)$ is extendible to arbitrarily large $k$. Continuing inductively, a sequence $s = (s_1, s_2, \ldots)$ is generated such that $x \in A_{s[n]}$ for all $n$, i.e. $x \in \cap_n A_{s[n]} \subseteq A$.

This completes the proof of (ii), and of (i) if $\mu^*$ is monotone regular.
For (i) in the case of general $\mu^*$, let $A, Y, Z$ be as above. Form the outer measure $\mu^*_W$, where $W = Y \cup Z$. Then every set in $\mathcal{M}$ is $\mu^*_W$-measurable. We also have that $\mu^*_W(X \setminus W) = 0$, hence $X \setminus W$ and also $W$ are $\mu^*_W$-measurable. Let $\iota$ be the restriction of $\mu^*_W$ to the $\mu^*_W$-measurable sets, and $\nu^*$ the outer measure on $X$ induced by $\iota$. Then $\nu^*$ is a regular outer measure, hence monotone regular (XIII.3.5.18.), and every $\mu^*_W$-measurable set is $\nu^*$-measurable (XIII.3.5.21.); in particular, $W$ and all sets in $\mathcal{M}$ are $\nu^*$-measurable. So

$$\nu^*(W) = \iota(W) = \mu^*_W(W) = \mu(W).$$

We also have $\nu^*(S) \geq \mu^*_W(S)$ for all $S \subseteq X$ (XIII.3.5.21.); in particular, $\nu^*(Y) \geq \mu^*_W(Y) = \mu^*(Y)$ and $\nu^*(Z) \geq \mu^*_W(Z) = \mu^*(Z)$. By the first part of the proof, $A$ is $\nu^*$-measurable, so

$$\mu^*(Y \cup Z) = \mu^*(W) = \nu^*(W) = \nu^*(W \cap A) + \nu^*(W \cap A^c) = \nu^*(Y) + \nu^*(Z) \geq \mu^*(Y) + \mu^*(Z)$$

and it follows that $A$ is $\mu^*$-measurable.

This completes the proof of XIII.3.6.7.

---

XIII.3.7. Radon Measures on $\mathbb{R}$

XIII.3.8. Lebesgue Measure and Radon Measures on $\mathbb{R}^n$

XIII.3.9. Nonmeasurable Subsets of $\mathbb{R}$

An interesting and rather subtle question is whether every subset of $\mathbb{R}$ is Lebesgue measurable. Using the Axiom of Choice, we construct a family of nonmeasurable subsets. This construction is due to Vitali.

**XIII.3.9.1.** Define a relation $R$ on $\mathbb{R}$ by $xRy$ if $y - x \in \mathbb{Q}$. It is easily checked that $R$ is an equivalence relation. If $a \in \mathbb{R}$, then the equivalence class of $a$ is

$$[a] = a + \mathbb{Q} = \{a + q : q \in \mathbb{Q}\}.$$

Thus the equivalence classes are countable and dense. (They are exactly the cosets of $\mathbb{Q}$ in the additive group $\mathbb{R}$.) As for any equivalence relation, the equivalence classes are pairwise disjoint and their union is $\mathbb{R}$.

**XIII.3.9.2.** **Definition.** (AC) Let $P$ be a subset of $(0, 1)$ which contains exactly one element of each equivalence class of $R$. If $q \in \mathbb{Q}$, set $P_q = P + q$.

Since each equivalence class is dense, it has nonempty intersection with $(0, 1)$. The existence of such a set $P$ is then guaranteed by the Axiom of Choice. (Any other actual interval could be used in place of $(0, 1)$.)

**XIII.3.9.3.** **Proposition.** The sets $\{P_q : q \in \mathbb{Q}\}$ are pairwise disjoint, and $\cup_{q \in \mathbb{Q}} P_q = \mathbb{R}$.

**Proof:** If $x \in (P + q_1) \cap (P + q_2)$, then $x = p_1 + q_1 = p_2 + q_2$ for $p_1, p_2 \in P$. Then $p_2 - p_1 = q_1 - q_2 \in \mathbb{Q}$, so $p_1 R p_2$ and hence $p_1 = p_2$ since $P$ contains only one element from any equivalence class. Thus also $q_1 = q_2$. If $x \in \mathbb{R}$, then $xRp$ for some $p \in P$. Set $q = x - p \in \mathbb{Q}$; then $x = p + q \in P_q$.

Now let $J = \mathbb{Q} \cap (-1, 1)$, and set $N = \cup_{q \in J} P_q$. 1436
XIII.3.9.4. Proposition. We have \((0,1) \subseteq N \subseteq (-1,2)\). Thus \(1 \leq \lambda^*(N) \leq 3\).

Proof: Since \(P \subseteq (0,1)\), it is obvious that \(P_q \subseteq (-1,2)\) for any \(q \in J\). For the other containment, let \(x \in (0,1)\). Then \(x = p + q\) for some \(p \in P\), \(q \in \mathbb{Q}\) by XIII.3.9.3.. Then \(q = x - p \in (-1,1)\), so \(x \in N\). \(\Diamond\)

XIII.3.9.5. Theorem. \(P\) is not Lebesgue measurable.

Proof: Suppose \(P\) is measurable. Then \(P_q\) is measurable for all \(q\), and \(\lambda(P_q) = \lambda(P)\); so \(N\) is measurable. If \(\lambda(P) = 0\), then \(\lambda(N) = \sum_{q \in J} \lambda(P_q) = 0\), contradicting XIII.3.9.4.. And if \(\lambda(P) > 0\), then \(\lambda(N) = \sum_{q \in J} \lambda(P_q) = \infty\) since \(J\) is infinite, again contradicting XIII.3.9.4..

In fact, we can do better:

XIII.3.9.6. Proposition. If \(A\) is a Lebesgue measurable subset of \(P\), then \(\lambda(A) = 0\).

Proof: Let \(B = \bigcup_{q \in J} (A + q)\). The sets \(A + q\) are pairwise disjoint since \(A \subseteq P\). Since \(\lambda(A + q) = \lambda(A)\) for all \(q\), and \(J\) is infinite, we would have \(\lambda(B) = \infty\) if \(\lambda(A) > 0\). But \(B \subseteq N \subseteq (-1,2)\), so \(\lambda(B) \leq 3\). Thus we must have \(\lambda(A) = 0\) (and hence \(\lambda(B) = 0\)). \(\Diamond\)

XIII.3.9.7. Theorem. (AC) Let \(Y \subseteq \mathbb{R}\). If \(\lambda^*(Y) > 0\), then \(Y\) contains a subset which is not Lebesgue measurable.

Proof: For \(q \in \mathbb{Q}\), let \(Y_q = Y \cap P_q\). Then the \(Y_q\) are pairwise disjoint and \(Y = \bigcup_{q \in \mathbb{Q}} Y_q\) by XIII.3.9.3.. If \(Y_q\) is measurable, then \(\lambda(Y_q) = 0\) by XIII.3.9.6.. Thus, if all the \(Y_q\) are measurable, then \(\lambda^*(Y) \leq \sum_{q \in \mathbb{Q}} \lambda^*(Y_q) = 0\). \(\Diamond\)

This result is only interesting if \(Y\) is Lebesgue measurable, since it is trivial otherwise. The important consequence is that every measurable set of positive measure contains a nonmeasurable subset.

There is an interesting corollary for Lebesgue measurable functions:

XIII.3.9.8. Theorem. (AC) The inverse image of a Lebesgue measurable set under a Lebesgue measurable function is not necessarily Lebesgue measurable. In fact, there is a strictly increasing continuous function (homeomorphism) \(f: \mathbb{R} \to \mathbb{R}\) and a measurable subset \(A \subseteq \mathbb{R}\) such that \(f^{-1}(A)\) is not Lebesgue measurable.

Proof: Let \(K\) be the Cantor set, and \(K_\alpha\) the generalized Cantor set of positive measure from (). There is a strictly increasing continuous function \(f: \mathbb{R} \to \mathbb{R}\) with \(f(K_\alpha) = K\). By XIII.3.9.7., there is a nonmeasurable subset \(B\) of \(K_\alpha\). Let \(A = f(B)\); since \(A \subseteq K\) and \(\lambda(K) = 0\), \(A\) is measurable and \(\lambda(A) = 0\). But \(B = f^{-1}(A)\) is nonmeasurable. \(\Diamond\)

1437
XIII.3.9.9. So homeomorphisms of \( \mathbb{R} \) do not necessarily preserve Lebesgue measurability, i.e. Lebesgue measurability of subsets of \( \mathbb{R} \) is not a topological property, although it is a metric property (in fact, see (\).

Note that the property of being a Borel set is a topological property.

XIII.3.9.10. The Axiom of Choice was used in an essential way in the definition of \( P \), so all results of this subsection depend on AC. In fact, R. Solovay showed \([]\) that (assuming that the existence of an inaccessible cardinal is consistent with ZFC; cf. (\)) the existence of a nonmeasurable subset of \( \mathbb{R} \) cannot be proved in ZF, i.e. it is consistent with ZF to assume that every subset of \( \mathbb{R} \) is Lebesgue measurable. Specifically, he showed that if an inaccessible cardinal exists, then there is a model for ZF in which every subset of \( \mathbb{R} \) is Lebesgue measurable; this model satisfies DC, but of course not AC. Thus even ZF+DC is not sufficient to prove the existence of a subset of \( \mathbb{R} \) which is not Lebesgue measurable.

But the full Axiom of Choice is not needed to obtain a set which is not Lebesgue measurable. The argument only requires a choice function for the set of cosets of \( \mathbb{Q} \) in \( \mathbb{R} \), hence existence of a choice function for \( \mathcal{P}(\mathbb{R}) \) suffices, which will be automatic if \( \mathbb{R} \) can be well ordered. Thus the Continuum Hypothesis is sufficient (but not necessary) to obtain Vitali’s example (\). For another example of a Lebesgue nonmeasurable set based on the CH, see XIV.6.3.1.\(]\).

Actually, assuming AC and some weak version of the CH, there is no finite continuous measure at all on \((\{0,1\}, \mathcal{P}([0,1]))\). See (\) for a discussion. It is also known that ZF+(Hahn-Banach Theorem) suffices to prove existence of a Lebesgue nonmeasurable set \([\]\), and the Hahn-Banach Theorem is strictly weaker than the Axiom of Choice (\).

Bernstein’s Nonmeasurable Set

There is another construction of a nonmeasurable subset of \( \mathbb{R} \), due to F. Bernstein. This construction uses more machinery than Vitali’s, but works in great generality.

XIII.3.9.11. The general setting is an outer measure \( \mu^* \) on a set \( X \), for which there is a collection \( \mathcal{C} = \{C_i : i \in I\} \) of subsets of \( X \), with the following properties:

(i) \( |I| \leq 2^{\aleph_0} \).

(ii) \( |C_i| = 2^{\aleph_0} \) for all \( i \).

(iii) \( \mu^*(C_i) > 0 \) for all \( i \).

(iv) If \( A \) is any \( \mu^* \)-measurable subset of \( X \) with \( \mu^*(A) > 0 \), then \( C_i \subseteq A \) for some \( i \).

Actually, \( 2^{\aleph_0} \) can be replaced by any other infinite cardinal here; the case of \( 2^{\aleph_0} \) is just the important one for applications.

XIII.3.9.12. Lebesgue outer measure on \( \mathbb{R} \) has such a \( \mathcal{C} \): the collection of closed subsets of \( \mathbb{R} \) with positive Lebesgue measure. There are \( 2^{\aleph_0} \) closed subsets of \( \mathbb{R} \) in all (\) (and \( 2^{\aleph_0} \) closed intervals, each of positive measure). Each closed set of positive measure is uncountable, hence has cardinality \( 2^{\aleph_0} \) by (\). Condition (iv) is satisfied by XIII.3.4.5..
XIII.3.9.13. Theorem. Let $\mu^*$ be an outer measure on a set $X$ for which there is a $\mathcal{C}$ satisfying (i)–(iv). If $Y$ is any subset of $X$ with $\mu^*(Y) > 0$, there is a subset of $Y$ which is not $\mu^*$-measurable.

Proof: Put a well ordering on $I$ such that the set of predecessors of any $i \in I$ has cardinality strictly less than $2^{80}$ (the Axiom of Choice guarantees that this can be done, by putting $I$ in one-one correspondence with the cardinal $\text{card}(I)$). Inductively choose for each $i$ two distinct elements $x_i$ and $y_i$ of $C_i$ such that $x_i$ and $y_i$ are also distinct from the set $\{x_j, y_j : j < i\}$; this is possible since the cardinality of this set is strictly smaller than the cardinality of $C_i$. Let $A = \{x_i : i \in I\}$.

Let $Y \subseteq X$ with $\mu^*(Y) > 0$. To show that $Y$ contains a subset which is not $\mu^*$-measurable, we may assume that $Y$ is $\mu^*$-measurable (since otherwise $Y$ itself is the desired subset). We claim $A \cap Y$ is not $\mu^*$-measurable. If $A \cap Y$ (and hence also $Y \setminus A$) is $\mu^*$-measurable, we cannot have $\mu^*(A \cap Y) > 0$, since then $C_i \subseteq (A \cap Y)$ for some $i$; but $y_i \in C_i \setminus A$, a contradiction. Thus $\mu^*(A \cap Y) = 0$. Similarly, we cannot have $\mu^*(Y \setminus A) > 0$, so $\mu^*(Y \setminus A) = 0$. But then $\mu^*(Y) = \mu^*(A \cap Y) + \mu^*(Y \setminus A) = 0$, contradicting the assumption that $\mu^*(Y) > 0$.

Note that AC is used twice: well ordering $I$ and choosing the $x_i$ and $y_i$.

XIII.3.9.14. This result applies more generally to any purely nonatomic Borel regular Borel outer measure on a Polish space, whose restriction to the Borel sets is $\sigma$-finite: an infinite Polish space has $2^{80}$ Borel sets ($\mathcal{B}$), and an uncountable Borel set in a Polish space has cardinality $2^{80}$ ($\mathcal{B}$). Compare with XI.13.9.14.

XIII.3.10. Exercises

(a) There is a sequence $(S_i)$ of pairwise disjoint Borel subsets of $[1, \infty)$, each of infinite Lebesgue measure. [Let $(N_i)$ be a sequence of pairwise disjoint infinite subsets of $N$, e.g. by identifying $N$ with $\mathbb{N} \times \mathbb{N}$. Let $S_i = \cup_{n \in N_i} [n, n + 1]$.

(b) Let $\mathcal{F}$ be the set of finite subsets of $\mathbb{N}$. $\mathcal{F}$ is countable; let $i$ be a bijection from $\mathcal{F}$ to $\mathbb{N}$.

(c) Arrange the bounded open intervals in $\mathbb{R}$ with rational endpoints into a sequence $(J_n)$, with each interval appearing infinitely often in the sequence.

(d) For each $n$, let

$$A_n = J_n \cup \left[ \bigcup_{n \in F} S_i(F) \cap [n, \infty) \right]$$

(i.e. the union runs over all sets $F \in \mathcal{F}$ with $n \in F$) and let $\mathcal{S}$ be the set of finite intersections of the $A_n$ and their complements. Then $\mathcal{S}$ is a semi-algebra ($\mathcal{A}$). Show that every nonempty set in $\mathcal{S}$ has infinite Lebesgue measure. [If $E = A_{n_1} \cap \cdots \cap A_{n_k} \cap A_{m_1} \cap \cdots \cap A_{m_l}$, if $n_i = m_j$ for some $i, j$, then $E = \emptyset$, and otherwise $S_{i=(n_1, ..., n_k)} \cap [p, \infty] \subseteq E$, where $p = \max\{n_1, ..., n_k\}$. Note that $S_i \cap [p, \infty]$ has infinite Lebesgue measure for every $i$ and $p$.]

If $\mathcal{A}$ is the algebra generated by $\mathcal{S}$, then $\mathcal{A}$ consists of finite disjoint unions of sets in $\mathcal{S}$ ($\mathcal{A}$), and hence every nonempty set in $\mathcal{A}$ has infinite Lebesgue measure.

(e) If $J$ is a bounded open interval with rational endpoints, then there is an infinite sequence $(n_k)$ such that $J_{n_k} = J$ for all $k$. Then $\bigcap_{k=1}^{\infty} A_{n_k} = J$. Thus $J$ is in the $\sigma$-algebra generated by $\mathcal{S}$ ($\mathcal{B}$). Since such intervals generate $\mathcal{B}$ as a $\sigma$-algebra, the $\sigma$-algebra generated by $\mathcal{S}$ ($\mathcal{A}$) is $\mathcal{B}$.

(f) The restriction of $\lambda$ to $\mathcal{S}$ ($\mathcal{A}$) is not $\sigma$-finite, and $\lambda$ is not the measure induced on $\mathcal{B}$ from $\lambda|_{\mathcal{S}}$ or $\lambda|_{\mathcal{A}}$ by the Carathéodory extension process.
(g) If \((X, \mathcal{B}, \mu)\) is a measure space with \(\mu(X) = \infty\), is there always an algebra \(\mathcal{A}\) in \(\mathcal{B}\) which generates \(\mathcal{B}\) as a \(\sigma\)-algebra, with the property that \(\mu(A) = \infty\) for every nonempty \(A \in \mathcal{A}\)?

**XIII.3.10.2.** Let \(X = \{x_1, x_2, x_3\}\) be a set with three elements. If \(A \subseteq X\), let \(\mu^*(A) = 0\) if \(A = \emptyset\), \(\mu^*(A) = 1\) if \(|A| = 1\) or if \(|A| = 2\), and \(\mu^*(A) = 2\) if \(|A| = 3\).

(a) Show that \(\mu^*\) is an outer measure on \(X\) which is not regular.

(b) If \(A = \{x_1, x_2\}\), show that \(\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)\), but \(A\) is not \(\mu^*\)-measurable. [Consider \(Y = \{x_1, x_3\}\).]

**XIII.3.10.3.** Prove the following generalization of (a). Let \(\mu^*\) be an outer measure on a set \(X\), and \(\mathcal{M}\) the \(\sigma\)-algebra of \(\mu^*\)-measurable sets.

(a) If \((A_n)\) is a pairwise disjoint sequence of sets in \(\mathcal{M}\) with \(A = \bigcup_{n=1}^{\infty} A_n\), and \(Y\) is any subset of \(X\), then \(\mu^*(Y \cap A) = \sum_{n=1}^{\infty} \mu^*(Y \cap A_n)\).

(b) If \(B_1 \subseteq B_2 \subseteq \cdots\) is an increasing sequence in \(\mathcal{M}\) with \(B = \bigcup_{n=1}^{\infty} B_n\), and \(Y\) is any subset of \(X\), then \(\mu^*(Y \cap B) = \sup_n \mu^*(Y \cap B_n)\).

(c) If \(C_1 \supseteq C_2 \supseteq \cdots\) is a decreasing sequence in \(\mathcal{M}\) with \(C = \bigcap_{n=1}^{\infty} C_n\), and \(Y\) is any subset of \(X\) with \(\mu^*(Y \cap C_1) < \infty\), then \(\mu^*(Y \cap C) = \inf_n \mu^*(Y \cap C_n)\).

**XIII.3.10.4.** If \(\mu^*\) is an outer measure on \(X\) and \(A, Y \subseteq X\) with \(A\) \(\mu^*\)-measurable, show that
\[
\mu^*(A \cup Y) + \mu^*(A \cap Y) = \mu^*(A) + \mu^*(Y) .
\]

**XIII.3.10.5.** Let \(\mu^*\) be a regular outer measure on a set \(X\). If \(A, B \subseteq X\) with \(A \cup B\) \(\mu^*\)-measurable and \(\mu^*(A \cup B) = \mu^*(A) + \mu^*(B) < \infty\), show that \(A\) and \(B\) are \(\mu^*\)-measurable.

**XIII.3.10.6.** Let \(X\) and \(Y\) be sets, \(f : X \to Y\) a function, and \(\mu^*\) an outer measure on \(X\). For \(B \subseteq Y\), set \(\nu^* = \mu^* \circ (f^{-1})\). Show that \(\nu^*\) is an outer measure on \(Y\). This outer measure is usually denoted \(f_*(\mu^*)\).

**XIII.3.10.7.** Let \(\{\mu^*_i\}\) be a collection of outer measures on a set \(X\), and let \(\mu^* = \sup \mu^*_i\). Show that \(\mu^*\) is an outer measure.

**XIII.3.10.8.** Let \(X = (0, 2], \mathcal{A} = \{\emptyset, [0, 1], (1, 2], (0, 2]\}\). Let \(\mu\) be the restriction of Lebesgue measure to \(\mathcal{A}\), and \(\mu^*\) the induced outer measure on \(X\).

(a) Describe \(\mu^*(Y)\) for a general \(Y \subseteq X\), and compare it to \(\lambda^*(Y)\). Find the \(\sigma\)-algebra of \(\mu^*\)-measurable subsets of \(X\).

(b) Show that \((X, \mathcal{A}, \mu)\) is a finite complete measure space, but that \(\mu\) extends to a measure (e.g. \(\lambda\)) on a strictly larger \(\sigma\)-algebra. Show that \(\lambda^*\) is not \(\mathcal{A}\)-regular.

**XIII.3.10.9.** Let \(\mathcal{U}\) be the collection of open sets in \(\mathbb{R}^n\), and let \(\mathcal{Q}\) be the set of bounded open intervals with rational endpoints. Show that \(\mathcal{S}(\mathcal{U}) = \mathcal{S}(\mathcal{Q})\) and that this collection of sets is the same as \(\mathcal{S}(\mathcal{K}), \mathcal{S}(\mathcal{F})\), and \(\mathcal{S}(\mathcal{B})\) of XIII.3.6.6. These sets are usually just called the Suslin subsets of \(\mathbb{R}^n\).
XIII.3.10.10. Show that the collection $S(Q)$ of Suslin subsets of $\mathbb{R}^n$ (Exercise XIII.3.10.9.) has cardinality $2^{\aleph_0}$. [Show that there are exactly $2^{\aleph_0}$ Suslin schemes in $Q$.] Since $B \subseteq S(Q)$, this gives a cleaner proof than ( ) that $B$ has cardinality $2^{\aleph_0}$.

XIII.3.10.11. Let $E$ be a nonempty collection of subsets of a set $X$.

(a) If $A = S(E) \cap \widehat{S(E)}$ is nonempty, it is a $\sigma$-algebra and the largest $\sigma$-algebra contained in $S(E)$ (cf. XII.1.5.2.).

(b) Show that $S(E) \cap \widehat{S(E)}$ can be empty.

In probability, $P$ is usually used for a measure instead of $\mu$. Thus $P(E)$ is the probability (measure) of $E$, and $E(X) = \int X \, dP$ is the expectation (integral) of $X$. 
XIII.4. Measures on Metric Spaces

Many important outer measures (e.g. Lebesgue outer measure) are defined on metric spaces, and behave nicely relative to the topology of the space.

XIII.4.1. Metric Outer Measures and Borel Measures

XIII.4.1.1. Definition. Let \((X, \rho)\) be a metric space, and let \(\mu^*\) be an outer measure on \(X\). The outer measure \(\mu^*\) is a metric outer measure on \(X\) (with respect to \(\rho\)) if, whenever \(A, B \subseteq X\) and \(\rho(A, B) > 0\), we have \(\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)\).

XIII.4.1.2. Example. By Exercise (), Lebesgue outer measure is a metric outer measure on \(\mathbb{R}\) with the usual metric.

XIII.4.1.3. Proposition. Let \((X, \rho)\) be a metric space, and \(\mu^*\) an outer measure on \(X\) for which every closed set is \(\mu^*\)-measurable. Then \(\mu^*\) is a metric outer measure.

Proof: If \(A, B \subseteq X\) with \(\rho(A, B) > 0\), then \(B \cap \bar{A} = \emptyset\). Since \(\bar{A}\) is closed, we have

\[
\mu^*(A \cup B) = \mu^*((A \cup B) \cap \bar{A}) + \mu^*((A \cup B) \cap (\bar{A})^c) = \mu^*(A) + \mu^*(B) .
\]

The main theorem of this subsection is the converse:

XIII.4.1.4. Theorem. Let \((X, \rho)\) be a metric space, and \(\mu^*\) a metric outer measure on \(X\) with respect to \(\rho\). Then every Borel set in \(X\) is \(\mu^*\)-measurable.

Proof: Since the collection of \(\mu^*\)-measurable sets is a \(\sigma\)-algebra, it suffices to show that every closed set \(F\) in \(X\) is \(\mu^*\)-measurable. Fix \(E \subseteq X\) with \(\mu^*(E) < \infty\), and let \(A = E \cap F\), \(B = E \cap F^c\). It suffices to show that \(\mu^*(E) \geq \mu^*(A) + \mu^*(B)\).

For \(n \in \mathbb{N}\), set

\[
B_n = \left\{ x \in B : \rho(x, F) \geq \frac{1}{n} \right\} .
\]

We have \(B = \bigcup_{n=1}^\infty B_n\) since \(F\) is closed; and we have \(\mu^*(A \cup B_n) = \mu^*(A) + \mu^*(B_n)\) for each \(n\) since \(\mu^*\) is a metric outer measure.

If \(\mu^*\) were known to be monotone regular, the proof would be trivial:

\[
\mu^*(E) = \mu^*(A \cup B) = \sup_n \mu^*(A \cup B_n) = \sup_n \left[ \mu^*(A) + \mu^*(B_n) \right] = \mu^*(A) + \sup_n \mu^*(B_n) = \mu^*(A) + \mu^*(B) .
\]

In general, we must show that \(\mu^*\) has enough vestiges of monotone regularity to get the conclusion.
For each \( n \) let \( D_n = B_{n+1} \setminus B_n \). We have \( \rho(D_{n+1}, B_n) \geq \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)} \) by the triangle inequality, and thus \( \mu^*(B_n \cup D_{n+1}) = \mu^*(B_n) + \mu^*(D_{n+1}) \) for all \( n \); since \( B_n \cup D_{n+1} \subseteq B_{n+2} \), we have \( \mu^*(B_n) + \mu^*(D_{n+1}) \leq \mu^*(B_{n+2}) \) for all \( n \) (this also holds for \( n = 0 \) if we set \( B_0 = \emptyset \)). Inductively, we get

\[
\sum_{k=1}^{n} \mu^*(D_{2k-1}) \leq \mu^*(B_{2n}) \leq \mu^*(E) < \infty 
\]

for all \( n \). Similarly,

\[
\sum_{k=1}^{n} \mu^*(D_{2k}) \leq \mu^*(B_{2n+1}) \leq \mu^*(E) < \infty 
\]

for all \( n \). Thus \( \sum_{k=1}^{\infty} \mu^*(D_k) \) converges; hence, for any \( \epsilon > 0 \) there is an \( n \) such that \( \sum_{k=1}^{\infty} \mu^*(D_k) < \epsilon \). By countable subadditivity we then have for this \( n \) that

\[
\mu^*(B) \leq \mu^*(B_n) + \sum_{k=n+1}^{\infty} \mu^*(D_k) < \mu^*(B_n) + \epsilon 
\]

Thus \( \mu^*(B) = \sup_n \mu^*(B_n) \). We then have

\[
\mu^*(E) \geq \sup_n \mu^*(A \cup B_n) = \sup_n [\mu^*(A) + \mu^*(B_n)] = \mu^*(A) + \sup_n \mu^*(B_n) = \mu^*(A) + \mu^*(B) 
\]

Borel Measures

**XIII.4.1.5.** DEFINITION. Let \( X \) be a topological space. A Borel measure on \( X \) is a measure on \((X, B)\), where \( B \) is the \( \sigma \)-algebra of Borel subsets of \( X \).

**XIII.4.1.6.** Suppose \( X \) is a metrizable topological space, and \( \rho \) is a metric on \( X \) giving the topology. If \( \mu^* \) is an outer measure on \( X \) which is a metric outer measure with respect to \( \rho \), then \( \mu^* \) defines a Borel measure on \( X \) by **XIII.4.1.4.** Conversely, a Borel measure \( \mu \) on \( X \) defines a (regular) outer measure \( \mu^* \) on \( X \) by (), which is a metric outer measure for \( \rho \) by **XIII.4.1.3.** Thus every Borel measure arises this way from a (regular) metric outer measure.

**XIII.4.2.** Method II Construction of Outer Measures

The standard construction of an outer measure from a set function (e.g., a premeasure) described in () is sometimes called the Method I construction, particularly in Geometric Measure Theory. On a metric space, there is another construction of outer measures which is more roundabout, but gives interesting examples such as Hausdorff measures.

It is customary in Geometric Measure Theory to use upper-case script letters such as \( \mathcal{M} \) to denote outer measures. (The reader is also cautioned that in much of the literature for Geometric Measure Theory, an outer measure is called a “measure.” The outer measures of Geometric Measure Theory are typically metric outer measures and are applied only to Borel sets, where they are indeed countably additive.)

Method II uses covers of sets which are “fine” in the sense that every point is contained in a set in the cover of arbitrarily small diameter:

1443
XIII.4.2.1. Definition. Let \((X, \rho)\) be a metric space, and \(E \subseteq X\). A Vitali cover of \(E\) is a collection \(\mathcal{V}\) of subsets of \(X\) such that, for every \(x \in E\) and \(\epsilon > 0\), there is a \(V \in \mathcal{V}\) with \(x \in V\) and \(\text{diam}(V) < \epsilon\).

XIII.4.2.2. The notion of Vitali cover depends on \(\rho\). It is easily checked that if \(\sigma\) is another metric on \(X\) uniformly equivalent to \(\rho\), then a collection \(\mathcal{V}\) of subsets of \(X\) is a Vitali cover for \(E\) with respect to \(\rho\) if and only if it is a Vitali cover with respect to \(\sigma\). However, this is not necessarily true if \(\rho\) and \(\sigma\) are just equivalent. To denote the dependence on \(\rho\) where necessary, we will use the terminology "\(\rho\)-Vitali cover."

XIII.4.2.3. The sets in a Vitali cover are not required to be of any particular form. But note that if \(\mathcal{V}\) is a Vitali cover of \(E\), then

\[
\widehat{\mathcal{V}} = \{ \widehat{V} : V \in \mathcal{V} \}
\]

is a Vitali cover of \(E\) consisting of closed sets. Thus we may work with Vitali covers consisting of closed sets, which is often convenient.

XIII.4.2.4. Examples. (i) The power set of \(X\) is a Vitali cover of any subset of \(X\).
(ii) The collection of open balls in \((X, \rho)\) is a Vitali cover of any subset \(E\) of \(X\).
(iii) The collection of closed balls in \((X, \rho)\) is a Vitali cover of any subset \(E\) of \(X\).
(iv) In (ii) and (iii), we may restrict to balls of rational radius and/or with centers in \(E\).
(v) If \(\mathcal{V}\) is a Vitali cover of \(E\) and \(\delta > 0\), then

\[
\mathcal{V}_\delta = \{ V \in \mathcal{V} : \text{diam}(V) < \delta \}
\]

are also Vitali covers of \(E\).
(vi) A Vitali cover of \(E\) is also a Vitali cover of any subset of \(E\). In particular, a Vitali cover of \(X\) is a Vitali cover of any subset of \(X\).

XIII.4.2.5. Fix a metric space \((X, \rho)\) and a Vitali cover \(\mathcal{V}\) of \(X\). We may assume \(\emptyset \in \mathcal{V}\). Let \(\iota\) be a function from \(\mathcal{V}\) to \([0, +\infty]\) with \(\iota(\emptyset) = 0\). For any \(\delta > 0\), define \(\mathcal{M}_\delta^\iota\) to be the outer measure on \(X\) defined by the restriction of \(\iota\) to \(\mathcal{V}_\delta\) (XIII.4.2.4.(v)) as in (i), i.e. for \(E \subseteq X\) define

\[
\mathcal{M}_\delta^\iota(E) = \inf \left\{ \sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \bigcup_{k=1}^{\infty} V_k, V_k \in \mathcal{V}, \text{diam}(V_k) < \delta \right\}
\]

(with \(\mathcal{M}_\delta^\iota(E) = +\infty\) if there are no countable covers of \(E\) by sets in \(\mathcal{V}_\delta\)). As usual, since \(\emptyset \in \mathcal{V}\) and \(\iota(\emptyset) = 0\), finite as well as countably infinite covers are allowed in computing \(\mathcal{M}_\delta^\iota\).

As \(\delta\) decreases, there are fewer such covers, so \(\mathcal{M}_\delta^\iota\) increases. Define

\[
\mathcal{M}^\iota = \sup_{\delta > 0} \mathcal{M}_\delta^\iota
\]

and note that \(\mathcal{M}^\iota\) is an outer measure by (i). \(\mathcal{M}^\iota\) is called the outer measure defined by \(\mathcal{V}\) and \(\iota\) by Method II.

The outer measure \(\mathcal{M}^\iota\) does depend on \(\mathcal{V}\) in general, but the standard notation does not reflect this since in reasonable cases (e.g. (i)) it is independent of \(\mathcal{V}\). It is most common to take \(\mathcal{V}\) to be the entire power set of \(X\).
XI.4.2.6. Use of a strict inequality in the diameters of the sets used is arbitrary. We could have also used the covers $\mathcal{V}_\delta$ to obtain outer measures $\mathcal{M}_\delta^\ast$, where

$$\mathcal{M}_\delta^\ast(E) = \inf \left\{ \sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \bigcup_{k=1}^{\infty} V_k, V_k \in \mathcal{V}, diam(V_k) \leq \delta \right\}$$

and then obtained

$$\mathcal{M}^\ast = \sup_{\delta>0} \mathcal{M}_\delta^\ast$$

since we have

$$\mathcal{M}_\delta^\ast(E) \leq \mathcal{M}_\delta^\ast(E) \leq \mathcal{M}_\delta^\ast(E)$$

for any $E$ and any $0 < \delta < \delta'$. 

XI.4.2.7. In the most commonly used special case where $\mathcal{V} = \mathcal{P}(X)$, we may also restrict to covers by closed sets:

$$\mathcal{M}_\delta^\ast(E) = \inf \left\{ \sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \bigcup_{k=1}^{\infty} V_k, V_k \text{ closed}, diam(V_k) < \delta \right\}$$

since $diam(A) = diam(A)$ for any $A \subseteq X$. In some cases, one can also restrict to covers by open sets (cf. XIII.4.3.6.).

XI.4.2.8. It is the passage to the limit as $\delta \to 0$ which gives Method II outer measures some nice properties making them useful. The idea is very similar to the way arc length for curves is defined: short chords and a passage to a limit are necessary to truly give an appropriate notion of length for curves. See () for more on this.

The first example of a nice property of Method II outer measures is:

XI.4.2.9. Proposition. For any $\mathcal{V}$ and $\iota$, $\mathcal{M}^\ast$ is a metric outer measure.

Proof: Suppose $A, B \subseteq X$ and $\rho(A, B) > 0$. If $\delta < \frac{\rho(A, B)}{2}$, then covers of $A$ and $B$ by sets in $\mathcal{V}_\delta$ are disjoint, so any cover of $A \cup B$ by sets in $\mathcal{V}_\delta$ breaks into a disjoint union of a cover of $A$ and a cover of $B$, and every pair of covers of $A$ and $B$ arises this way. It follows that

$$\mathcal{M}_\delta^\ast(A \cup B) = \mathcal{M}_\delta^\ast(A) + \mathcal{M}_\delta^\ast(B)$$

by taking infima. Then taking supremum over such small $\delta$ gives

$$\mathcal{M}^\ast(A \cup B) = \mathcal{M}^\ast(A) + \mathcal{M}^\ast(B) .$$

XI.4.2.10. So by XIII.4.1.4., $\mathcal{M}^\ast$ defines a Borel measure on $X$, also denoted $\mathcal{M}^\ast$. 

1445
XIII.4.3. Hausdorff Measure

Perhaps the most important example of the Method II construction of outer measures is Hausdorff measure. Hausdorff measure is most commonly considered on \( \mathbb{R}^n \), but it can be equally well defined on any metric space. It gives a metric notion of “dimension” for subsets of the space, which need not be an integer. Actually, we will get a more refined and flexible notion of metric dimension than just a number.

We first consider a class of functions:

XIII.4.3.1. Definition. Let \( \mathcal{H} \) be the set of all nondecreasing, right continuous functions from \([0, +\infty]\) to \([0, +\infty]\). Let \( \mathcal{H}_0 \) be the set of functions \( h \in \mathcal{H} \) with \( h(0) = 0 \) (and hence \( \lim_{t \to 0^+} h(t) = 0 \)).

If \((X, \rho)\) is any metric space, we will define a Hausdorff measure on \((X, \rho)\) corresponding to each \( h \in \mathcal{H} \). (Unfortunately, there are many \( h \)'s in the standard notation for Hausdorff measure, which must be carefully distinguished; it will be impossible to stick to the recommendations of II.1.3.14. if standard notation is used.)

XIII.4.3.2. Definition. Let \((X, \rho)\) be a metric space, and \( h \in \mathcal{H} \). Define the Hausdorff measure \( \mathcal{H}^h \) on \( X \) by \( \mathcal{H}^h = \mathcal{M} \), the Method II outer measure on \( X \) using \( V = \mathcal{P}(X) \) and \( \iota(V) = h(\text{diam}(V)) \) (set \( \iota(\emptyset) = 0 \)). Thus \( \mathcal{H}^h = \sup_{\delta > 0} \mathcal{H}_\delta^h \), where for \( E \neq \emptyset \) we have

\[
\mathcal{H}_\delta^h(E) = \inf \left\{ \sum_{k=1}^{\infty} h(\text{diam}((V_k)) : E \subseteq \bigcup_{k=1}^{\infty} V_k, V_k \subseteq X, \text{diam}(V_k) < \delta \right\}
\]

(with \( \mathcal{H}_\delta^h(E) = +\infty \) if there are no countable covers of \( E \) by sets of diameter \( < \delta \)).

We will use the notation \( \mathcal{H}^h \) for both the outer measure and the associated Borel measure. When the dependence on the metric \( \rho \) is relevant, we will write \( \mathcal{H}_\rho^h \).

The most important case is when \( h \) is a power function:

XIII.4.3.3. Definition. Let \( s > 0 \). Then \( s \)-dimensional Hausdorff measure \( \mathcal{H}^{(s)} \) on a metric space \((X, \rho)\) is \( \mathcal{H}^h \), where \( h(t) = t^s \). In other words, \( \mathcal{H}^{(s)} = \sup_{\delta > 0} \mathcal{H}_\delta^{(s)} \), where \( E \neq \emptyset \) we have

\[
\mathcal{H}_\delta^{(s)}(E) = \inf \left\{ \sum_{k=1}^{\infty} (\text{diam}((V_k))^s : E \subseteq \bigcup_{k=1}^{\infty} V_k, V_k \subseteq X, \text{diam}(V_k) < \delta \right\}
\]

(with \( \mathcal{H}_\delta^{(s)}(E) = +\infty \) if there are no countable covers of \( E \) by sets of diameter \( < \delta \)).

We also define \( \mathcal{H}^{(0)} \) to be \( \mathcal{H}^h \), where \( h \) is the constant function 1.

XIII.4.3.4. Proposition. The measure \( \mathcal{H}^{(0)} \) is counting measure on \( X \).

Proof: Suppose \( E \) is a finite subset of \( X \) with \( n \) elements. If \( \delta \) is smaller than the minimum distance between distinct points of \( E \), any cover of \( E \) by sets of diameter \( < \delta \) must have at least \( n \) elements, so \( \mathcal{H}_\delta^{(0)}(E) = n \). A similar argument applies in the case where \( E \) is infinite: for any \( n \), we have \( \mathcal{H}_\delta^{(0)}(E) > n \) for sufficiently small \( \delta \).
In the definition of Hausdorff measure, we can also use the outer measures $H^h$, where

$$H^h(E) = \inf \left\{ \sum_{k=1}^{\infty} h(\text{diam}(V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X, \text{diam}(V_k) \leq \delta \right\}$$

(cf. XIII.4.2.6.).

In computing $H^h(E)$, we may restrict to covers by closed sets by XIII.4.2.7. We may also use open covers (this is the place where right continuity of $h$ is used):

**Proposition.** Let $(X, \rho)$ be a metric space, $h \in \mathcal{H}$, and $\delta > 0$. Then

$$H^h(E) = \inf \left\{ \sum_{k=1}^{\infty} h(\text{diam}(V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X \text{ open}, \text{diam}(V_k) < \delta \right\}$$

for any $E \subseteq X$.

**Proof:** We clearly have

$$H^h(E) \leq \inf \left\{ \sum_{k=1}^{\infty} h(\text{diam}(V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X \text{ open}, \text{diam}(V_k) < \delta \right\}.$$

For the reverse inequality, there is nothing to prove if $H^h(E) = +\infty$, so assume $H^h(E) < +\infty$. Let $\epsilon > 0$, and fix a cover $\{V_k\}$ of $E$ with $\text{diam}(V_k) < \delta$ for all $k$ and

$$\sum_{k=1}^{\infty} h(\text{diam}(V_k)) < H^h(E) + \frac{\epsilon}{2}.$$

For each $k$, let $U_k$ be an open set containing $V_k$ such that $\text{diam}(U_k) < \delta$ and $h(\text{diam}(U_k)) \leq h(\text{diam}(V_k)) + \frac{\epsilon}{2k^r}$; $U_k$ can be taken to be $U_k = \{x \in X : \rho(x, V_k) < r\}$ for some sufficiently small $r > 0$, using the right continuity of $h$. Then

$$\sum_{k=1}^{\infty} h(\text{diam}(U_k)) \leq \sum_{k=1}^{\infty} h(\text{diam}(V_k)) + \frac{\epsilon}{2} < H^h(E) + \epsilon$$

and since $\epsilon > 0$ is arbitrary, we obtain the reverse inequality.

The Hausdorff measures $\mathcal{H}^h$, even the $s$-dimensional Hausdorff measures $\mathcal{H}^{(s)}$, depend strongly on the choice of metric. If $\sigma$ is uniformly equivalent to $\rho$, . . .
Many slight variations on Hausdorff measure are possible. One variation is to take $V$ to be the set of balls in $(X, \rho)$ (in metric spaces such as $\mathbb{R}^n$ where each closed ball is the closure of an open ball of the same diameter, it does not matter whether one uses open balls, closed balls, or arbitrary balls in between; note that in a general metric space the diameter of a ball of radius $r$ may be strictly smaller than $2r$). The measures made in the same way are sometimes called spherical measures and denoted $S^h$ or $S^{(s)}$.

The measure $S$ is not the same as $H^h$ since a set of diameter $d$ is not necessarily contained in a ball of diameter $d$ (cf. $();$ it is clearly contained in a ball of diameter $\leq 2d$). We have $H^h(E) \leq S^h(E)$ for any $h$ and $E$. For power functions we have $H^s(E) \leq S^s(E)$ for any $s$ and $E$.

Hausdorff Dimension

As $s$ increases, the Hausdorff measure $H^{(s)}(E)$ of a subset $E$ of a metric space $(X, \rho)$ jumps suddenly down from $+\infty$ to $0$. The critical value where this occurs is called the Hausdorff dimension of $E$. More generally, there are typically functions $h \in H$ whose rate of convergence to 0 describe the boundary between infinite and zero Hausdorff measure.

**Proposition.** Suppose $g, h \in H$ and $\lim_{t \to 0^+} \frac{g(t)}{h(t)} = 0$. Let $(X, \rho)$ be a metric space, and $E \subseteq X$.

(i) If $H^h(E) < +\infty$, then $H^g(E) = 0$.

(ii) If $H^g(E) > 0$, then $H^h(E) = +\infty$.

**Proof:** Statement (ii) is the contrapositive of (i). To prove (i), let $\epsilon > 0$, and choose $\eta > 0$ such that $g(t) \leq \frac{\epsilon}{H^h(E) + 1} h(t)$ for $t < \eta$. If $0 < \delta < \eta$, there is a cover $(V_k)$ of $E$ with $\text{diam}(V_k) < \delta$ for all $k$, such that

$$\sum_{k=1}^{\infty} h(\text{diam}(V_k)) < \frac{\epsilon}{H^h(E) + 1} \sum_{k=1}^{\infty} h(\text{diam}(V_k)) < \epsilon$$

and for this cover we have

$$\sum_{k=1}^{\infty} g(\text{diam}(V_k)) \leq \frac{\epsilon}{H^h(E) + 1} < \epsilon$$

so we have $H^g(E) < \epsilon$. Taking the supremum over such $\delta$, we have $H^g(E) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $H^g(E) = 0$. $\Box$
XIII.4.3.11. Corollary. Let $0 \leq s < t$. Let $(X, \rho)$ be a metric space, and $E \subseteq X$. Then

(i) If $\mathcal{H}^s(E) < +\infty$, then $\mathcal{H}^t(E) = 0$.

(ii) If $\mathcal{H}^t(E) > 0$, then $\mathcal{H}^s(E) = +\infty$.

XIII.4.3.12. Corollary. Let $(X, \rho)$ be a metric space, and $E$ an infinite subset of $X$. Then

$$\sup\{s \geq 0 : \mathcal{H}^s(E) = \infty\} = \inf\{t > 0 : \mathcal{H}^t(E) = 0\}$$

(set $\inf(\emptyset) = +\infty$).

XIII.4.3.13. Definition. The common value in XIII.4.3.12. is called the Hausdorff dimension of $E$ in $(X, \rho)$, denoted $\dim_H(E)$. We also define the Hausdorff dimension of a finite set to be 0.

XIII.4.3.14. So the Hausdorff dimension of $E$ is the critical value $s$ where the Hausdorff measure $\mathcal{H}^s(E)$ jumps from $+\infty$ to 0. Examples show that if $s = \dim_H(E)$, then $\mathcal{H}^s(E)$ can be 0, $+\infty$, or any finite positive number.

XIII.4.3.15. If only power functions are considered, for a given $E$ there is an exact boundary between the functions $h$ for which $\mathcal{H}^h(E) = +\infty$ and the $h$ for which $\mathcal{H}^h(E) = 0$. But if all functions in $\mathcal{H}$ are considered, there is not such a clean boundary. Still, if there is an $h \in \mathcal{H}$ for which $0 < \mathcal{H}^h(E) < +\infty$, it can be considered a reasonable boundary: for any $g \in \mathcal{H}$ with $\lim_{t \to 0+} \frac{g(t)}{h(t)} = 0$, we have $\mathcal{H}^g(E) = 0$, and for any $g \in \mathcal{H}$ with $\lim_{t \to 0+} \frac{g(t)}{h(t)} = +\infty$, we have $\mathcal{H}^g(E) = +\infty$. Such an $h$ is not generally unique, however; but dividing $\mathcal{H}$ into those $h$ for which $\mathcal{H}^h(E) < +\infty$ and those $h$ for which $\mathcal{H}^h(E) = +\infty$ (or, almost equivalently, those $h$ for which $\mathcal{H}^h(E) = 0$ and those $h$ for which $\mathcal{H}^h(E) > 0$) is an illuminating partition.

XIII.4.3.16. Because of XIII.4.3.8., the same phenomenon works for spherical measures and gives the same $s$ as the transition point. Thus Hausdorff dimension can be calculated using spherical measures, i.e. covering with balls. For more general $h$ the phenomenon is not quite the same or so simple.

XIII.4.4. Hausdorff Measure on $\mathbb{R}^n$

XIII.4.4.1. Hausdorff measures and Hausdorff dimension were originally defined and studied for subsets of $\mathbb{R}^n$, and are still of most interest there.

Unfortunately, there is no consistency in references about the exact definition of Hausdorff measure in $\mathbb{R}^n$. It appears that most references define the measures in the same way up to a constant factor, which can vary from reference to reference. There are two predominant definitions used in the literature: the “unnormalized Hausdorff measure,” primarily used in the theory of fractals, and the “normalized Hausdorff measure,” primarily used in Geometric Measure Theory (only $s$-dimensional Hausdorff measure can be reasonably normalized, and usually only integer values of $s$ are used). The normalized version has nicer properties for most purposes in analysis, since it exactly coincides with Lebesgue measure and arc length/surface area/surface measure in appropriate dimensions. Thus we will concentrate on the normalized case. The
distinction is not important for many purposes: the crucial distinction is often simply whether a set has zero, infinite, or finite nonzero measure. For example, both versions give the same Hausdorff dimension for subsets of \( \mathbb{R}^n \).

**Unnormalized Hausdorff Measure**

**XIII.4.4.2.** **Definition.** Let \( h \in \mathcal{H} \). Unnormalized Hausdorff \( h \)-measure \( \mathcal{H}^h \) on \( \mathbb{R}^n \) is \( \mathcal{H}^h \) on \( \mathbb{R}^n \) with the usual Euclidean metric \( \rho_2 \). If \( 0 \leq s \), unnormalized \( s \)-dimensional Hausdorff measure \( \mathcal{H}^s \) on \( \mathbb{R}^n \) is \( \mathcal{H}^s \) for \( (\mathbb{R}^n, \rho_2) \).

As usual, we will use the same notation for the outer measure and the Borel measure.

**Normalized Hausdorff Measure**

**XIII.4.4.3.** **Definition.** If \( 0 \leq s \), normalized \( s \)-dimensional Hausdorff measure \( \mathcal{L}^s \) on \( \mathbb{R}^n \) is \( \mathcal{H}^s \) for \( (\mathbb{R}^n, \rho_2) \), where

\[
\alpha_s = \frac{\pi^{n/2}}{2^s \Gamma(\frac{n}{2} + 1)}
\]

is a normalizing constant (\( \Gamma \) is the Gamma function).)

**XIII.4.4.4.** If \( s \) is an integer, \( \alpha_s \) is the \( s \)-dimensional “volume” of a ball in \( \mathbb{R}^s \) of diameter 1 (radius \( 1/2 \)). Thus \( \alpha_1 = 1 \), \( \alpha_2 = \frac{\pi}{4} \), \( \alpha_3 = \frac{1}{8} \left( \frac{4}{3} \pi \right) = \frac{\pi}{6} \), etc.; see (n).

**XIII.4.4.5.** The notation \( \mathcal{L}^s \) for normalized Hausdorff measure is nonstandard; it is most commonly called \( \mathcal{H}^s \). But this notation is also used by some authors to mean unnormalized Hausdorff measure, which is consistent with its use in general metric spaces, so we use \( \mathcal{H}^s \) to mean unnormalized Hausdorff measure. We have chosen the notation \( \mathcal{L}^s \) for normalized Hausdorff measure since for appropriate \( s \) it coincides with usual Lebesgue measure or standard variants. Note that we do not define normalized Hausdorff measure for general \( h \in \mathcal{H} \).

**XIII.4.4.6.** Hausdorff measure on \( \mathbb{R}^n \) can be defined in other ways too. For example, in [7] it is defined using the metric \( \rho_\infty \). Hausdorff measure defined in this way is not a constant multiple of the usual Hausdorff measure in most cases, but it is an equivalent measure and gives the same theory of Hausdorff dimension, and is technically simpler in some respects.

**XIII.4.5. Hausdorff Measure and Hausdorff Dimension for Self-Similar Sets**

**XIII.4.6. Packing Measure and other Geometric Measures**

Hausdorff measure, or more precisely spherical measure (XIII.4.3.8.), which is equivalent and hence gives the same dimension, uses good coverings by balls. There is a “dual” notion which uses dense packings by disjoint balls, not surprisingly called packing measure. Packing measure is a much more recent development than Hausdorff measure, being defined by Tricot in 1982 [1].

The definition of packing measure involves yet another level of approximation:
XIII.4.6.1. **Definition.** Let \((X, \rho)\) be a metric space, and \(s \geq 0\). Let \(E \subseteq X\). For each \(\delta > 0\) set
\[
\mathcal{P}_\delta^{(s)}(E) = \sup \left\{ \sum_{k=1}^{\infty} (\text{diam}(B_k))^s \right\}
\]
where the supremum is over all sequences \((B_k)\) of pairwise disjoint open balls of radii \(\leq \delta\) with centers in \(E\) (empty balls are allowed, i.e. finite sequences of \(B_k\) are allowed). Set
\[
\mathcal{P}_0^{(s)}(E) = \inf_{\delta > 0} \mathcal{P}_\delta^{(s)}(E)
\]
(note that \(\mathcal{P}_\delta^{(s)}(E)\) decreases as \(\delta\) decreases) and
\[
\mathcal{P}^{(s)}(E) = \inf \left\{ \sum_{k=1}^{\infty} \mathcal{P}_0^{(s)}(E_k) : E \subseteq \bigcup_{k=1}^{\infty} E_k \right\}
\]
i.e. \(\mathcal{P}^{(s)}\) is the outer measure defined by the set function \(\mathcal{P}_0^{(s)}()\). \(\mathcal{P}^{(s)}\) is called \(s\)-dimensional packing measure on \((X, \rho)\).

XIII.4.6.2. Note that “packing measure” is actually an outer measure. But it is easily seen to be a metric outer measure, so all Borel sets are measurable and it is truly a measure on Borel sets. \(\mathcal{P}_\delta^{(s)}\) is an outer measure for any \(\delta > 0\), but \(\mathcal{P}_0^{(s)}\) is not an outer measure in general, so the last step is necessary even though it makes packing measure harder to compute.

A routine argument shows that disjoint closed balls can be used instead of disjoint open balls in the definition of packing measure, since the diameter of an open ball is the supremum of the diameters of the closed balls contained in it, and two open balls are disjoint if and only if all pairs of closed balls contained in them are disjoint.

It is also easily verified that \(\mathcal{P}^{(0)}\) is counting measure on \((X, \mathcal{P}(X))\).

XIII.4.6.3. **Proposition.** Let \((X, \rho)\) be a metric space, and \(s \geq 0\). Then \(\mathcal{H}^{(s)} \leq \mathcal{P}^{(s)}\) on \(X\). Thus \(\mathcal{H}^{(s)} \leq \mathcal{P}^{(s)}\) on \(X\).

Packing measure behaves in a similar way to Hausdorff measure as \(s\) varies:

XIII.4.6.4. **Proposition.** Let \(0 \leq s < t\). Let \((X, \rho)\) be a metric space, and \(E \subseteq X\). Then
(i) If \(\mathcal{P}^{(s)}(E) < +\infty\), then \(\mathcal{P}^{(t)}(E) = 0\).
(ii) If \(\mathcal{P}^{(t)}(E) > 0\), then \(\mathcal{P}^{(s)}(E) = +\infty\).

**Proof:** If \((B_k)\) is a sequence of pairwise disjoint balls of radius \(\leq \delta\) with centers in \(E\), then \(\sum_{k=1}^{\infty} (\text{diam}(B_k))^t \leq (2\delta)^{t-s} \sum_{k=1}^{\infty} (\text{diam}(B_k))^s\). Taking the infimum over all such sequences, we obtain \(\mathcal{P}^{(t)}(E) \leq (2\delta)^{t-s} \mathcal{P}^{(s)}(E)\). Taking the infimum over all \(\delta\), if \(\mathcal{P}_0^{(s)}(E) < +\infty\), then \(\mathcal{P}_0^{(t)}(E) = 0\). Now suppose \(\mathcal{P}^{(s)}(E) < +\infty\). Then
there is a sequence \((E_k)\) with \(E \subseteq \bigcup_{k=1}^{\infty} E_k\) and \(\sum_{k=1}^{\infty} \mathcal{P}_0^{(s)}(E_k) < +\infty\). Then \(\mathcal{P}_0^{(s)}(E_k) < +\infty\) for all \(k\), so \(\mathcal{P}_0^{(1)}(E_k) = 0\) for all \(k\), and thus
\[
\mathcal{P}^{(1)}(E) \leq \sum_{k=1}^{\infty} \mathcal{P}_0^{(1)}(E) = 0.
\]

Statement (ii) is the contrapositive of (i).

XIII.4.6.5. **COROLLARY.** Let \((X, \rho)\) be a metric space, and \(E\) an infinite subset of \(X\). Then
\[
\sup\{s \geq 0 : \mathcal{P}^{(s)}(E) = \infty\} = \inf\{t > 0 : \mathcal{P}^{(t)}(E) = 0\}
\]
(set \(\inf(\emptyset) = +\infty\)).

XIII.4.6.6. **DEFINITION.** The common value in XIII.4.6.5. is called the **packing dimension** of \(E\) in \((X, \rho)\), denoted \(\text{dim}_P(E)\). We also define the packing dimension of a finite set to be 0.

XIII.4.6.7. So the packing dimension of \(E\) is the critical value \(s\) where the packing measure \(\mathcal{P}^{(s)}(E)\) jumps from \(+\infty\) to 0. Examples show that if \(s = \text{dim}_P(E)\), then \(\mathcal{P}^{(s)}(E)\) can be 0, \(+\infty\), or any finite positive number.

XIII.4.6.8. It is an immediate corollary of XIII.4.6.3. that \(\text{dim}_H(E) \leq \text{dim}_P(E)\) for any \(E \subseteq X\). The Hausdorff and packing dimensions of a subset can differ (), although for well-behaved subsets they frequently coincide; when they do, the “fractal dimension” of the set is well defined as the common value.

XIII.4.6.9. A variation of the diameter-based packing measure we have considered is **radius-based packing measure**: use twice the radius of \(B_n\) in place of \(\text{diam}(B_n)\) in the definition. This makes no difference in Euclidean space since the diameter of any ball is exactly twice the radius, but this is false in general metric spaces, even subsets of Euclidean space with the restricted metric. The radius-based \(s\)-dimensional packing measure of a subset of a general metric space can differ significantly from the diameter-based packing measure (the radius-based one is larger), and even the radius-based and diameter-based packing dimensions of a subset can be different. Some references use radius-based packing measure. See [] for a discussion.

XIII.4.6.10. There is also a question of interpretation of the diameter of a ball: is it the diameter in \(X\), or the diameter of the intersection of the ball with the subset? In fact, there is a potential ambiguity in packing measure: is the packing measure or packing dimension of a subset \(E\) of \(X\) the same whether \(E\) is regarded as a subset of \(X\) or just as a metric space with the restricted metric? More generally, if \(Y\) is a subset of \(X\) with the restricted metric, and \(E \subseteq Y\), is the packing measure or packing dimension of \(E\) as a subset of \(Y\) the same as it is as a subset of \(X\)? Note that this problem does not occur with radius-based packing measure or dimension.

XIII.4.6.11. There are some other notions of fractal dimension generally intermediate between Hausdorff and packing dimension, e.g. upper and lower box-counting dimension and modified upper and lower box-counting dimensions, which do not arise from measures and which are not as nicely behaved, and some others which are more subtle and variable. These are defined and discussed in [Fal14].
Integralgeometric Measure

$\mathcal{H}$ is Hausdorff measure
$\mathcal{P}$ is packing measure
$\mathcal{I}$ is integralgeometric measure
$\mathcal{L}$ is Lebesgue measure

XIII.4.7. Radon Measures on Metric Spaces

XIII.4.8. Exercises

XIII.4.8.1. (a) Let $\rho$ be a metric on $\mathbb{R}^n$ coming from a norm. Show that $\mathcal{H}_\rho^{(n)}$ is a scalar multiple $\alpha \mathcal{L}^{(n)}$, where $\alpha$ is the reciprocal of the Lebesgue measure of the unit ball for $\rho$.

(b) Show that unless $\rho$ is a constant multiple of the usual Euclidean distance, $\mathcal{H}_\rho^{(s)}$ for $0 < s < n$ is not a constant multiple of $\mathcal{L}^{(s)}$. [For $s = 1$ consider horizontal and diagonal line segments.]
XIII.5. Measures on Product Spaces

In this section, we will discuss a natural way of constructing a Cartesian product of two or more measure spaces.

XIII.5.1. Product $\sigma$-Algebras

XIII.5.1.1. Suppose $(X, A)$ and $(Y, B)$ are measurable spaces. If $A \subseteq X$ and $B \subseteq Y$, then $A \times B$ is called a rectangle in $X \times Y$. If $A \in A$ and $B \in B$, then $A \times B$ is called a measurable rectangle (technically, $(A, B)$-measurable rectangle). It is natural that such subsets of $X \times Y$ should be regarded as measurable sets.

XIII.5.1.2. Definition. The product $\sigma$-algebra of $(A, B)$, denoted $A \otimes B$, is the $\sigma$-algebra on $X \times Y$ generated by the measurable rectangles.

Although some authors use the notation $AB$ for the product $\sigma$-algebra, we will use the symbol $A \otimes B$ since $A \otimes B$ is much larger than just the set of measurable rectangles in general; we will use $AB$ to denote the set of measurable rectangles. (The symbol $\otimes$ is normally used for “tensor product” in abstract algebra; there is a close analogy between our use of the symbol here and algebraic tensor products.)

XIII.5.1.3. Examples.

(i) If $X$ and $Y$ are sets, and at least one is countable, then it is easily checked that $\mathcal{P}(X) \otimes \mathcal{P}(Y) = \mathcal{P}(X \times Y)$ (Exercise (i)). In particular, $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N}) = \mathcal{P}(\mathbb{N}^2)$. But this result does not hold if $X$ and $Y$ are of large cardinality (Exercise (i)).

(ii) If $B$ is the $\sigma$-algebra of Borel sets in $\mathbb{R}$, then $B \otimes B = \mathcal{B}^{(2)}$, the $\sigma$-algebra of Borel sets in $\mathbb{R}^2$. Neither inclusion is completely obvious. To show that $B \otimes B \subseteq \mathcal{B}^{(2)}$, it suffices to show that every measurable rectangle is in $\mathcal{B}^{(2)}$. But $\mathcal{B}_x = \{ B \times \mathbb{R} : B \in \mathcal{B} \}$ is a $\sigma$-algebra which is generated by $\{ U \times \mathbb{R} : U \text{ open} \}$, and $U \times \mathbb{R}$ is open in $\mathbb{R}^2$, hence in $\mathcal{B}^{(2)}$; thus $\mathcal{B}_x \subseteq \mathcal{B}^{(2)}$. Similarly, $\mathcal{B}^{(2)} = \{ \mathbb{R} \times B : B \in \mathcal{B} \} \subseteq \mathcal{B}^{(2)}$. So if $A, B \in \mathcal{B}$, then $A \times B = (A \times \mathbb{R}) \cap (\mathbb{R} \times B) \in \mathcal{B}^{(2)}$, so $B \otimes B \subseteq \mathcal{B}^{(2)}$. For the other inclusion, it suffices to show that every open set in $\mathbb{R}^2$ is in $\mathcal{B} \otimes \mathcal{B}$. But it is a simple exercise (i) that every nonempty open set in $\mathbb{R}^2$ is a (countable) union of open rectangles of the form $(a, b) \times (c, d)$ with $a, b, c, d \in \mathbb{Q}$. (This example generalizes to the Borel sets in any product of second countable topological spaces; see (ii).)

Although the set of measurable rectangles is not a $\sigma$-algebra or even an algebra because it is not closed under complements or finite unions (except in trivial cases), we have:

XIII.5.1.4. Proposition. Let $(X, A)$ and $(Y, B)$ be measurable spaces. Then the set

$A \times B = \{ A \times B : A \in A, B \in B \}$

of measurable rectangles is a semialgebra.

Proof: We have $X \times Y$ and $\emptyset = \emptyset \times \emptyset$ in $A \times B$. If $\{ A_k \times B_k : 1 \leq k \leq n \}$ is a finite collection in $A \times B$, then

$\bigcap_{k=1}^{n} (A_k \times B_k) = \left( \bigcap_{k=1}^{n} A_k \right) \times \left( \bigcap_{k=1}^{n} B_k \right) \in A \times B$. 

1454
so $A \times B$ is closed under finite intersections (in fact, the same argument shows that $A \times B$ is closed under countable intersections). And if $A \times B \in A \times B$, then

$$(A \times B)^c = (A \times B^c) \cup (A^c \times Y)$$

so $(A \times B)^c$ is a disjoint union of two sets in $A \times B$.

\[\Box\]

XIII.5.1.5. PROPOSITION. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces. Let $\pi_X$ and $\pi_Y$ be the projections from $X \times Y$ to $X$ and $Y$ respectively, i.e. $\pi_X(x, y) = x$, $\pi_Y(x, y) = y$. Then

(i) $\pi_X$ is an $(\mathcal{A} \otimes \mathcal{B}, \mathcal{A})$-measurable function and $\pi_Y$ is an $(\mathcal{A} \otimes \mathcal{B}, \mathcal{B})$-measurable function.

(ii) If $f : Z \to (X \times Y)$ is a function, then $f$ is $(\mathcal{C}, \mathcal{A} \otimes \mathcal{B})$-measurable if and only if $\pi_X \circ f : Z \to X$ is $(\mathcal{C}, \mathcal{A})$-measurable and $\pi_Y \circ f : Z \to Y$ is $(\mathcal{C}, \mathcal{B})$-measurable.

PROOF: (i): If $A \in \mathcal{A}$, then $\pi_X^{-1}(A) = A \times Y \in \mathcal{A} \otimes \mathcal{B}$, so $\pi_X$ is measurable. The argument for $\pi_Y$ is similar.

(ii): If $f$ is measurable, then $\pi_X \circ f$ and $\pi_Y \circ f$ are measurable by XIII.1.4.3. Conversely, if $\pi_X \circ f$ is measurable, then $f^{-1}(A \times Y) = [\pi_X \circ f]^{-1}(A) \in \mathcal{C}$ for all $A \in \mathcal{A}$. Similarly, if $\pi_Y \circ f$ is measurable, then $f^{-1}(X \times B) \in \mathcal{C}$ for all $B \in \mathcal{B}$. Since $\{A \times Y : A \in \mathcal{A}\} \cup \{X \times B : B \in \mathcal{B}\}$ generates $\mathcal{A} \otimes \mathcal{B}$ as a $\sigma$-algebra, $f$ is measurable by XIII.1.4.10. \[\Box\]

XIII.5.2. Product Measures

Now suppose that $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ are measure spaces. We want to obtain a reasonable measure $\mu \times \nu$ on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$. It is natural to want the measure of a measurable rectangle $A \times B$ to be $\mu(A)\nu(B)$. It turns out that it is possible to obtain such a measure through the Carathéodory extension process.

XIII.5.2.6. THEOREM. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces. For $A \times B$ in the semialgebra $\mathcal{C}$ of measurable rectangles (XIII.5.1.4.), set $\iota(A \times B) = \mu(A)\nu(B)$ (where $0 \cdot \infty = \infty \cdot 0 = 0$ by convention). Then $\iota$ is a premeasure on $\mathcal{C}$, and hence $\iota$ extends to a measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ denoted $\mu \times \nu$ by the Carathéodory extension process. If $\mu$ and $\nu$ are $\sigma$-finite, then $\mu \times \nu$ is the unique extension of $\iota$ to a measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$, and is also $\sigma$-finite.

PROOF: We verify that $\iota$ satisfies (i), (ii'), and (iii) of XIII.3.5.2. and XIII.3.5.6. (note that (ii') $\Rightarrow$ (ii) since $\emptyset \in \mathcal{A} \times \mathcal{B}$). Condition (i) is obvious. We prove (ii') and (iii) simultaneously since the arguments are almost identical (we could simplify the argument slightly by using the somewhat nontrivial result XIII.3.5.6. to avoid proving (iii), but this is not necessary). Suppose $A_1 \times B_1, A_2 \times B_2, \ldots$ is a sequence of measurable rectangles and

$$\bigcup_{k=1}^{\infty} (A_k \times B_k) \supseteq A \times B \in \mathcal{A} \times \mathcal{B}$$

1455
(for (ii'), assume the $A_k \otimes B_k$ are pairwise disjoint and $\bigcup_{k=1}^{\infty} (A_k \times B_k) = A \times B$). We must show $\mu(A)\nu(B) \leq \sum_{k=1}^{\infty} \mu(A_k)\nu(B_k)$, with equality for (ii'). We may assume $A$ and $B$ are nonempty. Set

$$\phi = \mu(A)\chi_B, \quad \psi = \sum_{k=1}^{\infty} \mu(A_k)\chi_{B_k}$$

(where the convention $\infty \cdot 0 = 0$ is followed). Then $\phi$ and $\psi$ are extended real-valued measurable functions on $(Y, \mathcal{B})$. We claim $\phi \leq \psi$, with equality in (ii'). If $y \notin B$, then $\phi(y) = 0$, and in case (ii') also $\psi(y) = 0$. If $y \in B$, set

$$I_y = \{k \in I : y \in B_k\} \subseteq \mathbb{N}.$$  

If $x \in A$, then $(x, y) \in A \times B$, so $(x, y) \in A_k \times B_k$ for at least one $k$ (exactly one $k$ in case (ii')), i.e. $x \in A_k$ for at least one $k \in I_y$ (exactly one in case (ii')). Thus $A$ is contained in the union of $\{A_k : k \in I_y\}$. In case (ii'), if $x \notin A$, then for every $k \in I_y$, $(x, y) \notin A_k \times B_k$; since $y \in B_k$, we have $x \notin A_k$ for every $k \in I_y$, i.e. $A$ is the disjoint union of $\{A_k : k \in I_y\}$. Thus

$$\phi(y) = \mu(A) \leq \sum_{k \in I_y} \mu(A_k) = \sum_{k \in I_y} \mu(A_k)\chi_{B_k}(y)$$

$$= \sum_{k \in \mathbb{N}} \mu(A_k)\chi_{B_k}(y) = \psi(y)$$

(with equality in case (ii')), where the first equality in the last line is because the additional terms are zero. Thus $\phi \leq \psi$, with equality in case (ii'). So

$$\mu(A)\nu(B) = \int_Y \phi \, d\nu \leq \int_Y \psi \, d\nu = \int_Y \left[ \sum_{k=1}^{\infty} \mu(A_k)\chi_{B_k} \right] \, d\nu$$

$$= \sum_{k=1}^{\infty} \mu(A_k) \left[ \int_Y \chi_{B_k} \, d\nu \right] = \sum_{k=1}^{\infty} \mu(A_k)\nu(B_k)$$

(with equality in case (ii')), where interchange of the sum and integral is justified by the MCT (XIV.4.5.5.). Thus $\iota$ satisfies (ii') and (iii) and is a premeasure.

The other conclusions then follow from XIII.3.5.8.. }

XIII.5.2.7. **DEFINITION.** The measure $\mu \times \nu$ is called the product measure of $\mu$ and $\nu$.

Completeness

XIII.5.2.8. There is a slight ambiguity in XIII.5.2.7. By the Extension Theorem, the $\sigma$-algebra $\mathcal{M}$ of $(\mu \times \nu)$-measurable sets contains the $\sigma$-algebra $\mathcal{A} \otimes \mathcal{B}$. But $(X \times Y, \mathcal{M}, \mu \times \nu)$ is a complete measure space and $\mathcal{M}$ is strictly larger than $\mathcal{A} \otimes \mathcal{B}$ in general. By slight abuse of notation, we will write $\mu \times \nu$ for the measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ and its extension on $(X \times Y, \mathcal{M})$ when a distinction is unimportant.

1456
If $\mu$ and $\nu$ are $\sigma$-finite, then $(X \times Y, \mathcal{M}, \mu \times \nu)$ is precisely the completion of $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ XIII.3.5.8. If $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ are the completions of $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ respectively, then the outer measures induced by $\iota$ on $\mathcal{A} \times \mathcal{B}$ and $\iota$ on $\mathcal{A} \times \mathcal{B}$ coincide (Exercise XIV.6.3.2); so if $\mathcal{M}$ is the $\sigma$-algebra of $(\mu \times \nu)$-measurable sets, then $\mathcal{M} = \mathcal{M}$. Thus in the $\sigma$-finite case we can also identify the completion of $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ with $(X \times Y, \mathcal{M}, \mu \times \nu)$, i.e. the completions of $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ and $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ coincide, and in particular $\mathcal{A} \otimes \mathcal{B} = \mathcal{A} \otimes \mathcal{B}$.

**Products of Finitely Many Spaces**

**XIII.5.2.9.** The results about products of two measure spaces can easily be generalized to products of finitely many. If $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ are measurable spaces, then a set of the form $A_1 \times \cdots \times A_n$ with $A_i \subseteq X_i$, is called a rectangle, a measurable rectangle if $A_i \in \mathcal{A}_i$ for each $i$. Then $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$ is the $\sigma$-algebra generated by the measurable rectangles in $X_1 \times \cdots \times X_n$. If $I = \{1, \ldots, n\}$, $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$ is often denoted $\bigotimes_{i \in I} \mathcal{A}_i$ or $\bigotimes_{i=1}^n \mathcal{A}_i$.

The next proposition shows that the product $\sigma$-algebra can be defined also by successive products of two $\sigma$-algebras, using any convenient grouping:

**XIII.5.2.10.** Proposition. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces. If $X \times Y \times Z$ is identified with $(X \times Y) \times Z$ and $X \times (Y \times Z)$ in the standard way, we have $\mathcal{A} \otimes \mathcal{B} \otimes \mathcal{C} = (\mathcal{A} \otimes \mathcal{B}) \otimes \mathcal{C} = \mathcal{A} \otimes (\mathcal{B} \otimes \mathcal{C})$.

The proof is routine, and is left as an exercise.

**XIII.5.2.11.** Corollary. Let $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ be measurable spaces, and let $I_1, I_2$ be disjoint subsets of $I = \{1, \ldots, n\}$ with $I = I_1 \cup I_2$. If $\prod_{i \in I} X_i$ is identified with $\left(\prod_{i \in I_1} X_i\right) \times \left(\prod_{i \in I_2} X_i\right)$ in the standard way, then

$$\bigotimes_{i \in I} \mathcal{A}_i = \left(\bigotimes_{i \in I_1} \mathcal{A}_i\right) \otimes \left(\bigotimes_{i \in I_2} \mathcal{A}_i\right).$$

There is then an analog of XIII.5.1.5. This can be proved directly in the same way as XIII.5.1.5., or by repeated application of XIII.5.1.5. and XIII.5.2.10.

**XIII.5.2.12.** Theorem. Let $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ and $(Z, \mathcal{C})$ be measurable spaces. Let $\pi_k$ $(1 \leq k \leq n)$ be the projections from $X_1 \times \cdots \times X_n$ to $X_k$, i.e. $\pi_k(x_1, \ldots, x_n) = x_k$. Then

(i) $\pi_k$ is an $(\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n, \mathcal{A}_k)$-measurable function.

(ii) If $f : Z \to (X_1 \times \cdots \times X_n)$ is a function, then $f$ is $(\mathcal{C}, \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n)$-measurable if and only if $\pi_k \circ f : Z \to X$ is $(\mathcal{C}, \mathcal{A}_k)$-measurable for each $k$, $1 \leq k \leq n$.

We also have the analog of XIII.5.1.4., which can also either be proved directly in the same way as XIII.5.1.4. or by repeated application of XIII.5.1.4. and XIII.5.2.10.
XIII.5.2.13. Proposition. Let \((X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)\) be measurable spaces. Then the set \(\mathcal{C} = \{ A_1 \times \cdots \times A_n : A_i \in \mathcal{A}_i \}\) of measurable rectangles is a semialgebra.

XIII.5.2.14. If \((X_1, \mathcal{A}_1), \cdots, (X_n, \mathcal{A}_n, \mu_n)\) are measure spaces, we can form a product measure \(\mu_1 \times \cdots \times \mu_n\) on \((\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)\) such that \((\mu_1 \times \cdots \times \mu_n)(A_1 \times \cdots \times A_n) = \mu_1(A_1) \mu_2(A_2) \cdots \mu_n(A_n)\) for any \(A_i \in \mathcal{A}_i\). The proof of the next theorem is an adaptation of the proof of (\(\)) and is left as an exercise (\(\)).

XIII.5.2.15. Theorem. Let \(\{(X_i, \mathcal{A}_i, \mu_i) : 1 \leq i \leq n\}\) be measure spaces. For \(\prod_{i=1}^n \mathcal{A}_i\) in the semialgebra \(\mathcal{C}\) of measurable rectangles (XIII.5.2.13.), set \(\tau(\prod_{i=1}^n A_i) = \mu_1(A_1) \cdots \mu_n(A_n)\). Then \(\tau\) is a premeasure on \(\mathcal{C}\), and hence \(\tau\) extends to a measure on \((\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)\) denoted \(\mu_1 \times \cdots \times \mu_n\) or \(\chi^n_{i=1} \mu_i\), by the Caratheodory extension process. If all \(\mu_i\) are \(\sigma\)-finite, then \(\chi^n_{i=1} \mu_i\) is the unique extension of \(\tau\) to a measure on \((\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)\).

The next fact is immediate from the uniqueness in the \(\sigma\)-finite case, but holds in general. The proof in the general case is routine but somewhat tedious, and the result in the non-\(\sigma\)-finite case is of limited interest, so we omit the details.

XIII.5.2.16. Proposition. Let \((X_1, \mathcal{A}_1, \mu_1), \ldots, (X_n, \mathcal{A}_n, \mu_n)\) be measure spaces, and let \(I_1, I_2\) be disjoint subsets of \(I = \{1, \ldots, n\}\) with \(I = I_1 \cup I_2\). If \(\prod_{i \in I} X_i\) is identified with \((\prod_{i \in I_1} X_i) \times (\prod_{i \in I_2} X_i)\) in the standard way, then \(\chi_{i \in I_1} \mu_i = (\chi_{i \in I_1} \mu_i) \times (\chi_{i \in I_2} \mu_i)\).

XIII.5.3. Lebesgue Measure on \(\mathbb{R}^n\)

XIII.5.4. Infinite Product Spaces

It is also possible to define an infinite Cartesian product of measure spaces provided that all (or all but finitely many) of the measures are probability measures.

XIII.5.4.1. Suppose that \(\{(X_i, \mathcal{A}_i) : i \in I\}\) is an indexed collection of measurable spaces. A rectangle in \(\prod_i X_i\) is a finite set \(I_0 = \{i_1, \ldots, i_n\}\) of coordinates, and for each \(i_k\) a subset \(A_{i_k}\) of \(X_{i_k}\). The corresponding set

\[\{\cdots x_i \cdots \} \in \prod_i X_i : x_{i_k} \in A_{i_k} \text{ for all } k, 1 \leq k \leq n\]

is often denoted \(A_{i_1} \times \cdots \times A_{i_n}\), although it should more correctly be denoted \(\prod_{i \in I_0} A_i \times \prod_{i \not\in I_0} X_i\). Such a rectangle is a measurable rectangle if each \(A_{i_k}\) is in \(\mathcal{A}_{i_k}\).

If \(A_i\) is a subset of \(X_i\) for each \(i\), then the subset

\[\prod_i A_i = \{\cdots x_i \cdots \} \in \prod_i X_i : x_i \in A_i \text{ for all } i \in I\]

is called a box, and, if each \(A_i \in \mathcal{A}_i\), a measurable box. Note that a [measurable] box is not a [measurable] rectangle in general if \(I\) is infinite, unless \(A_i = X_i\) for all but finitely many \(i\).
XIII.5.4.2. Definition. The product $\sigma$-algebra $\bigotimes_{i \in I} A_i$ is the $\sigma$-algebra on $\prod_i X_i$ generated by the measurable rectangles.

XIII.5.4.3. Proposition. If $I$ is countable, then every measurable box is in $\bigotimes_{i \in I} A_i$.

XIII.5.4.4. This statement is false in general if $I$ is uncountable. We say that a subset $A$ of $\prod_i X_i$ depends only on the coordinates in $J \subseteq I$ if, whenever $x = (\cdots x_i \cdots)$ and $y = (\cdots y_i \cdots)$ are elements of $\prod_i X_i$, and $x_j = y_j$ for all $j \in J$, then $x \in A$ if and only if $y \in A$. A subset $A$ of $\prod_i X_i$ depends on finitely many countably many coordinates if it depends only on some finite countable set of coordinates. Thus every rectangle depends only on finitely many coordinates.

XIII.5.4.5. Proposition. Every set in $\bigotimes_{i \in I} A_i$ depends only on countably many coordinates.

Proof: It is easily checked that the collection of all sets which depend on only countably many coordinates is a $\sigma$-algebra.

XIII.5.4.6. Corollary. A nonempty measurable box $\prod_i A_i$ is in $\bigotimes_{i \in I} A_i$ if and only if $A_i = X_i$ for all but countably many $i$.

Infinite product measures are particularly important in probability. The next example is a good illustration of an infinite product measure which has a nice probabilistic interpretation.

XIII.5.4.7. Example. Let $X = \{0, 1\}$, and for each $n$ let $(X_n, A_n) = (X, \mathcal{P}(X))$. Put a measure $\mu$ on $(X, \mathcal{P}(X))$ by $\mu(\{0\}) = \mu(\{1\}) = 1/2$, and let $\mu_n$ be the corresponding measure on $(X_n, A_n)$. Let $X^n = \prod_{n \in \mathbb{N}} X_n$, $A^n = \bigotimes_{n \in \mathbb{N}} A_n$, and $\mu^n = \times_{n \in \mathbb{N}} \mu_n$. $(X, \mathcal{P}(X), \mu)$ is not a very interesting or complicated measure space, but $(X^n, A^n, \mu^n)$ is quite interesting and nontrivial. Elements of $X^n$ are, of course, all sequences of 0’s and 1’s.

If instead of $\{0, 1\}$ we take $X = \{H, T\}$, we may interpret $(X, \mathcal{P}(X), \mu)$ as the sample space of the experiment of tossing a fair coin once. Then $(X^n, A^n, \mu^n)$ can be interpreted as the sample space of the experiment of tossing a fair coin infinitely many times. Probabilists use this measure space to analyze such things as the likelihood of long runs and other details in a sequence of coin tosses. This is a good example of a stochastic process, the study of which is an entire branch of probability theory.

This example also has other interpretations not closely connected with probability. The product measure can be interpreted as a measure on the Cantor set (or as Haar measure on a certain compact topological group $(\cdot)$). There is also a close connection with Lebesgue measure on the unit interval $(\cdot)$.

The example could be modified by taking $\mu_\alpha(\{0\}) = \alpha$, $\mu_\alpha(\{1\}) = 1 - \alpha$ for some fixed $\alpha$, $0 < \alpha < 1$ (the original $\mu$ is $\mu_{1/2}$). The measure $\mu_\alpha$ could be interpreted probabilistically (using $\{H, T\}$ for $\{0, 1\}$) as tossing an unfair coin with probability of heads equal to $\alpha$. It turns out that the $\mu^n_\alpha$ are mutually singular continuous probability measures on $(X^n, A^n)$ as $\alpha$ ranges over $(0, 1)$ $(\cdot)$. One could even similarly define $\mu_0$ and $\mu_1$; $\mu^n_0$ is then the point mass at the constant sequence of 0’s, and $\mu^n_1$ the point mass at the constant sequence of 1’s.
XIII.5.4.8. One might try to construct the infinite product of \( \{ (X_i, A_i, \mu_i) : i \in I \} \) (where \( \mu_i(X_i) = 1 \) for all \( i \)) as follows. For each finite subset \( F \) of \( I \), let \( A_F \) be the \( \sigma \)-algebra on \( X = \prod X_i \) generated by the measurable rectangles depending only on \( F \), and \( \mu_F \) the measure on \( (X, A_F) \) which is effectively \( \times_{i \in F} \mu_i \) using the obvious identification of \( A_F \) with \( \otimes_{i \in F} A_i \) on \( \prod_{i \in F} X_i \). Then \( B = \cup_F A_F \) is an algebra of subsets of \( X \), and the \( \mu_F \) give a well-defined function \( \iota \) from \( B \) to \( [0, 1] \). The \( \sigma \)-algebra generated by \( B \) is \( \otimes_{i \in I} A_i \), and \( \times_{i \in I} \mu_i \) is an extension of \( \iota \) to \( \otimes_{i \in I} A_i \), so one might try to obtain \( \times_{i \in I} \mu_i \) directly from \( \iota \) by the Caratheodory extension process.

The difficulty in this approach is showing that \( \iota \) is a premeasure on \( B \). This is not automatic: it is not true in general that if \( (X, A_n, \nu_n) \) is a compatible family of probability measures on an increasing sequence \( (A_n) \) of \( \sigma \)-algebras, that the function \( \nu \) defined by the \( \nu_n \) on the algebra \( \cup_n A_n \) is a premeasure. For example, let \( X = \mathbb{N} \), \( A_n \) the \( \sigma \)-algebra of all subsets of \( \{1, \ldots, n\} \) and their complements in \( \mathbb{N} \), and \( \nu_n \) the measure on \( (\mathbb{N}, A_n) \) with \( \nu_n(A) = 0 \) if \( A \subseteq \{1, \ldots, n\} \) and \( \nu_n(A) = 1 \) if \( \mathbb{N} \setminus A \subseteq \{1, \ldots, n\} \). Then the function \( \nu \) is not a premeasure on \( \cup_n A_n \).

XIII.5.5. Exercises

XIII.5.5.1. Let \( X \) and \( Y \) be sets, and \( E \subseteq X \times Y \). For each \( x \in X \), set

\[
E_x = \{ y \in Y : (x, y) \in E \} \subseteq Y
\]

\[S_X(E) = \{ E_x : x \in X \} \subseteq \mathcal{P}(Y) .\]

\( S_X(E) \) is the set of “vertical cross-sections” of the set \( E \). The \( X \)-variety of \( E \) is \( var_X(E) = card(S_X(E)) \), the “number” of different sets occurring as vertical cross sections of \( E \). Similarly, for \( y \in Y \) define

\[
E^y = \{ x \in X : (x, y) \in E \} \subseteq X
\]

\[S_Y(E) = \{ E^y : y \in Y \} \subseteq \mathcal{P}(X) \]

and the \( Y \)-variety of \( E \) to be \( var_Y(E) = card(S_Y(E)) \).

(a) Prove that \( C_X = \{ E \subseteq X \times Y : var_X(E) \leq 2^{|Y|} \} \) is a \( \sigma \)-algebra of subsets of \( X \times Y \) containing \( \mathcal{P}(X) \otimes \mathcal{P}(Y) \). Similarly, \( C_Y = \{ E \subseteq X \times Y : var_Y(E) \leq 2^{|X|} \} \) is a \( \sigma \)-algebra containing \( \mathcal{P}(X) \otimes \mathcal{P}(Y) \).

(b) If \( card(X) > 2^{|X|} \) and \( card(Y) \geq 2^{|Y|} \), then \( \mathcal{P}(X) \otimes \mathcal{P}(Y) \neq \mathcal{P}(X \times Y) \). [Let \( \kappa = min(card(X), 2^{|X|}) \). Construct an \( E \subseteq X \times Y \) with \( var_X(E) = \kappa \). (What is really needed is a function from \( X \) to \( \mathcal{P}(Y) \) whose range has cardinality greater than \( 2^{|X|} \); AC is needed in general to insure that \( card(X) \) is comparable with \( 2^{|X|} \).)]

(c) Is it true that \( \mathcal{P}(X) \otimes \mathcal{P}(Y) = C_X \cap C_Y \) for all \( X, Y \)?

(d) Is \( \mathcal{P}(\mathbb{R}) \otimes \mathcal{P}(\mathbb{R}) = \mathcal{P}(\mathbb{R}^2) \)? [Unfair question: it cannot be answered in ZFC! The CH, or, more generally, Martin’s Axiom (II.10.5.6.), implies a positive answer, but (if ZFC is consistent) there are models of ZFC where the answer is negative.]

XIII.5.5.2. ([GP10], attributed to von Neumann) Let \( R = \prod_{k=1}^n [a_k, b_k] \) be a closed rectangle in \( \mathbb{R}^n \). Note that \( \lambda(R) \leq \prod_{k=1}^n (b_k - a_k) \) is immediate from the definition. The aim of this problem is to show the opposite inequality.
(a) Suppose \( R_1, \ldots, R_m \) are open rectangles covering \( R \) (note that any cover of \( R \) by open rectangles has a finite subcover). Let \( R_j = \prod_{k=1}^{n} (a_{k}^{(j)}, b_{k}^{(j)}) \). If \( S \subseteq \mathbb{R}^n \), let \( \#(S) \) be the number of points with integer coordinates in \( S \). If \( \alpha > 0 \) is large enough that \( \alpha(b_k - a_k) > 1 \) and \( \alpha(b_{k}^{(j)} - a_{k}^{(j)}) > 1 \) for all \( k \) and \( j \), then

\[
\prod_{k=1}^{n} (\alpha b_k - \alpha a_k - 1) \leq \#(\alpha R) \leq \sum_{j=1}^{m} \#(\alpha R_j) \leq \sum_{j=1}^{m} \prod_{k=1}^{n} (\alpha b_{k}^{(j)} - \alpha a_{k}^{(j)} + 1).
\]

(b) Let \( \alpha \to +\infty \) to conclude that

\[
\prod_{k=1}^{n} (b_k - a_k) \leq \sum_{j=1}^{m} \prod_{k=1}^{n} (b_{k}^{(j)} - a_{k}^{(j)}).
\]

Taking the infimum over all such open covers, conclude that

\[
\prod_{k=1}^{n} (b_k - a_k) \leq \lambda(R).
\]

XIII.5.3. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be measure spaces, with completions \((X, \bar{\mathcal{A}}, \bar{\mu})\) and \((Y, \bar{\mathcal{B}}, \bar{\nu})\) respectively, and let \(\iota\) and \(\bar{\iota}\) be the induced premeasures on \(\bar{\mathcal{A}} \times \bar{\mathcal{B}}\) and \(\mathcal{A} \times \mathcal{B}\), i.e. \(\iota(A \times B) = \mu(A)\mu(B)\) for \(A \in \mathcal{A}, B \in \mathcal{B}\) and \(\bar{\iota}(\bar{A} \times \bar{B}) = \bar{\mu}(\bar{A})\bar{\nu}(\bar{B})\) for \(\bar{A} \in \bar{\mathcal{A}}, \bar{B} \in \bar{\mathcal{B}}\). Let \(\theta^*\) and \(\bar{\theta}^*\) be the outer measures on \(X \times Y\) induced by \(\iota\) and \(\bar{\iota}\) respectively.

(a) If \(\bar{A} \in \bar{\mathcal{A}}\) and \(\bar{B} \in \bar{\mathcal{B}}\), show that there are \(A \in \mathcal{A}\) and \(B \in \mathcal{B}\) such that \(\bar{A} \subseteq A, \bar{B} \subseteq B\), and \(\iota(A \times B) = \bar{\iota}(\bar{A} \times \bar{B})\).

(b) Show that \(\bar{\theta}^* = \theta^*\).
XIII.6. Signed and Complex Measures

In this section we examine the theory of measures which are not necessarily nonnegative or even real-valued. It turns out that the theory of such measures can be nicely reduced to the theory of ordinary (nonnegative) measures.

The only insurmountable difficulty comes with trying to make sense of the expression $\infty - \infty$. It will be implicit in our definitions that this possibility will never arise, which will entail that our measures do not take both values $\pm \infty$. In most applications, we will only consider real-valued measures, i.e. ones not taking either value $\pm \infty$. For complex measures we will not be able to satisfactorily allow any type of infinite value.

XIII.6.1. Signed Measures

XIII.6.1.1. Definition. Let $(X, A)$ be a measurable space. A signed measure on $(X, A)$ is a function $\nu : A \to [-\infty, +\infty]$ such that

(i) $\nu(\emptyset) = 0$.

(ii) If $\{A_n\}$ is a sequence of pairwise disjoint sets in $A$, then either

$$\sum_{\nu(A_n) > 0} \nu(A_n) < +\infty \quad \text{or} \quad \sum_{\nu(A_n) < 0} \nu(A_n) > -\infty$$

(so $\sum_n \nu(A_n)$ is well defined), and

$$\nu \left( \bigcup_{n=1}^\infty A_n \right) = \sum_{n=1}^\infty \nu(A_n).$$

The signed measure $\nu$ is finite if $\nu(A) \in \mathbb{R}$ for all $A \in A$.

Note that, as with measures, the countable additivity of (ii) includes finite additivity (taking the remaining sets to be $\emptyset$). In particular, there cannot be disjoint sets $A_1, A_2 \in A$ with $\nu(A_1) = +\infty$ and $\nu(A_2) = -\infty$.

XIII.6.1.2. Examples. (i) Any measure is a signed measure.

(ii) If $\mu_1$ and $\mu_2$ are (ordinary) measures on a measurable space $(X, A)$, at least one of which is finite, then $\mu_1 - \mu_2$ is a well-defined signed measure on $(X, A)$. More generally, any (finite) linear combination of measures, all but at most one of which are finite, is a signed measure. Any (finite) linear combination of finite signed measures is a (finite) signed measure.

(iii) Let $(X, A, \mu)$ be an (ordinary) measure space, and $f \in L^1(X, A, \mu)$. For $A \in A$, let

$$\nu(A) = \int_A f \, d\mu.$$

It is easy to check (cf. (i)) that $\nu$ is a finite signed measure.

A signed measure of the form in (ii) can be regarded as “trivial”, since its structure is completely determined and easy to describe in terms of the properties of $\mu_1$ and $\mu_2$. The main result of this section is that every signed measure is of this form, with the positive and negative pieces supported on disjoint sets:
XIII.6.1.3. Theorem. [Jordan Decomposition Theorem] Let \( \nu \) be a signed measure on a measure space \((X, \mathcal{A})\). Then there are unique measures \( \nu^+ \) and \( \nu^- \) on \((X, \mathcal{A})\), with \( \nu = \nu^+ - \nu^- \) and \( \nu^+ \perp \nu^- \). At least one of \( \nu^+ \) and \( \nu^- \) is a finite measure.

Thus the theory of signed measures turns out to be “trivial.” (However, the proof that the theory is actually trivial is definitely nontrivial!)

It is useful to rephrase this theorem in a somewhat different way. We first make some definitions:

XIII.6.1.4. Definition. Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\). A set \( P \in \mathcal{A} \) is a positive set for \( \nu \) if \( \nu(A) \geq 0 \) for every \( A \in \mathcal{A}, A \subseteq P \). Similarly, \( N \in \mathcal{A} \) is a negative set for \( \nu \) if \( \nu(A) \leq 0 \) for all \( A \in \mathcal{A}, A \subseteq N \). A set \( E \in \mathcal{A} \) is a null set for \( \nu \) if it is both positive and negative, i.e. \( \nu(A) = 0 \) for all \( A \in \mathcal{A}, A \subseteq E \).

Note that a set \( E \in \mathcal{A} \) with \( \nu(E) = 0 \) is not necessary a null set: such an \( E \) could be a disjoint union \( A \cup B \), where \( \nu(A) > 0 \) and \( \nu(B) = -\nu(A) \). Similarly, an \( E \in \mathcal{A} \) with \( \nu(E) > 0 \) is not necessarily a positive set.

It is obvious that any (measurable) subset of a positive set is positive, and a finite or countable union of positive sets is positive (Exercise ()). The same holds for negative sets and null sets.

XIII.6.1.5. Definition. Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\). A Hahn decomposition for \( \nu \) is a pair \((P, N)\), where \( P, N \in \mathcal{A}, N = X \setminus P \), \( P \) is positive for \( \nu \), and \( N \) is negative for \( \nu \).

We can then rephrase the Jordan Decomposition Theorem in the following way:

XIII.6.1.6. Theorem. [Hahn Decomposition Theorem] Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\). Then there is a Hahn decomposition for \( \nu \). If \((P, N)\) and \((P', N')\) are Hahn decompositions for \( \nu \), then \( P \triangle P' \) and \( N \triangle N' \) are null sets for \( \nu \), i.e. the Hahn decomposition of \( \nu \) is unique up to null sets.

XIII.6.1.7. To see that the Hahn Decomposition Theorem is equivalent to the Jordan Decomposition Theorem (we will only need one direction of this equivalence), suppose \((P, N)\) is a Hahn decomposition for \( \nu \). For \( A \in \mathcal{A} \), set \( \nu^+(A) = \nu(A \cap P) \) and \( \nu^-(A) = -\nu(A \cap N) \). Then it is easily checked that \( \nu^+ \) and \( \nu^- \) are measures, \( \nu^+ \perp \nu^- \), and \( \nu = \nu^+ - \nu^- \). These measures do not change if \( P \) and \( N \) are changed by null sets. At least one of \( P \) and \( N \) must have finite measure. Conversely, let \( \nu = \mu_1 - \mu_2 \), where \( \mu_1 \) and \( \mu_2 \) are measures with \( \mu_1 \perp \mu_2 \). Let \( P \in \mathcal{A} \) with \( \mu_2(P) = 0 \) and \( \mu_1(X \setminus P) = 0 \); such a \( P \) is unique up to null sets. Then \((P, X \setminus P)\) is a Hahn decomposition for \( \nu \). Another Hahn decomposition for \( \nu \) not differing from this one by null sets would give a pair \((\nu^+, \nu^-)\) different from \((\mu_1, \mu_2)\).

XIII.6.1.8. Example. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \( f \in L^1(X, \mathcal{A}, \mu) \), and for \( A \in \mathcal{A} \) set \( \nu(A) = \int_A f \, d\mu \) as in XIII.6.1.2. A Hahn decomposition for \( \nu \) is precisely a pair \((P, N)\), where \( N = X \setminus P \), \( f \geq 0 \) a.e. on \( P \), and \( f \leq 0 \) a.e. on \( N \). One Hahn decomposition is given by \( P = \{ x : f(x) \geq 0 \} \) and \( N = \{ x : f(x) < 0 \} \). The Jordan decomposition of \( \nu \) is \( \nu = \nu^+ - \nu^- \), where \( \nu^+(A) = \int_A f_+ \, d\mu \) and \( \nu^-(A) = \int_A f_- \, d\mu \) for \( A \in \mathcal{A} \) and \( f_+ \), \( f_- \) the positive and negative parts of \( f \).
The proof of Theorem XIII.6.1.6. (and therefore Theorem XIII.6.1.3.) will consist of a series of lemmas, all of which would be easy corollaries of the theorem. Our proof is based on [Dos80] [note that the argument of [Dos80] is incomplete as written since it assumes the conclusion of XIII.6.1.14. without comment.]

The first result about signed measures is that a signed measure cannot take both the values $+\infty$ and $-\infty$:

**XIII.6.1.9. Lemma.** Let $\nu$ be a signed measure on $(X, \mathcal{A})$. Then there do not exist $A, B \in \mathcal{A}$ with $\nu(A) = +\infty$ and $\nu(B) = -\infty$.

**Proof:** It follows immediately from XIII.6.1.1(ii) that there do not exist disjoint sets $A_1, A_2 \in \mathcal{A}$ with $\nu(A_1) = +\infty$ and $\nu(A_2) = -\infty$. Suppose $\nu(A) = +\infty$ and $\nu(B) = -\infty$. Then $A \cap B$, $A \setminus B$, and $B \setminus A$ are pairwise disjoint, and:

- At least one of $A \cap B$ and $A \setminus B$ has measure $+\infty$.
- At least one of $A \cap B$ and $B \setminus A$ has measure $-\infty$.

Thus there are disjoint sets with measure $+\infty$ and $-\infty$, a contradiction.

**XIII.6.1.10. Lemma.** Let $\nu$ be a signed measure on $(X, \mathcal{A})$, and $A, B \in \mathcal{A}$ with $B \subseteq A$. If $\nu(A) \in \mathbb{R}$ (i.e. $|\nu(A)| < \infty$), then $\nu(B) \in \mathbb{R}$.

**Proof:** We have $\nu(A) = \nu(B) + \nu(A \setminus B)$, and the sum on the right is well defined. If $\nu(B) = +\infty$, then we cannot have $\nu(A \setminus B) = -\infty$, and hence we would have to have $\nu(A) = +\infty$. The argument if $\nu(B) = -\infty$ is similar.

The next result is an exact analog of (i).

**XIII.6.1.11. Lemma.** Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$.

(i) If $E_1, E_2, \cdots \in \mathcal{A}$ with $E_1 \subseteq E_2 \subseteq \cdots$, then $\lim_{n \to \infty} \nu(E_n)$ exists as an extended real number and equals $\nu(\bigcup_{n=1}^{\infty} E_n)$.

(ii) If $E_1, E_2, \cdots \in \mathcal{A}$ with $E_1 \supseteq E_2 \supseteq \cdots$, and $\nu(E_n) \in \mathbb{R}$ for some $n$, then $\lim_{n \to \infty} \nu(E_n)$ exists as a real number and equals $\nu(\bigcap_{n=1}^{\infty} E_n)$.

**Proof:** (i): Let $A_1 = E_1$ and $A_n = E_n \setminus E_{n-1}$ for $n > 1$. Then the $A_n$ are pairwise disjoint, $E_n = \bigcup_{k=1}^{n} A_k$ for all $n$, $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} E_n$, and the series $\sum_{k=1}^{\infty} \nu(A_k)$ converges to $\nu(\bigcup_{n=1}^{\infty} E_n)$. So

$$\nu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{k=1}^{\infty} \nu(A_k) = \lim_{n \to \infty} \sum_{k=1}^{n} \nu(A_k) = \lim_{n \to \infty} \nu\left(\bigcup_{k=1}^{n} A_k\right) = \lim_{n \to \infty} \nu(E_n).$$

(ii): We may assume $\nu(E_1) \in \mathbb{R}$. Set $E = \bigcap_{n=1}^{\infty} E_n$, $B_n = E_1 \setminus E_n$, and $B = \bigcup_{n=1}^{\infty} B_n = E_1 \setminus E$. Then by (i)

$$\nu(B) = \lim_{n \to \infty} \nu(B_n) = \lim_{n \to \infty} \nu(E_1 \setminus E_n).$$

1464
Thus, since \( \nu(E_1) \in \mathbb{R} \),
\[
\lim_{n \to \infty} \nu(E_1 \setminus E_n) = \lim_{n \to \infty} [\nu(E_1) - \nu(E_n)]
\]
exists and hence \( \lim_{n \to \infty} \nu(E_n) \) exists and equals \( \nu(E_1) - \lim_{n \to \infty} \nu(B_n) \), and we have
\[
\nu(E) = \nu(E_1) - \nu(B) = \nu(E_1) - \lim_{n \to \infty} \nu(B_n) = \lim_{n \to \infty} \nu(E_n).
\]

Furthermore, since \( E \subseteq E_1 \) and \( \nu(E_1) \in \mathbb{R} \), we have \( \nu(E) \in \mathbb{R} \) by XIII.6.1.10..  

**XIII.6.1.12.** Lemma. Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\) not taking the value \(+\infty\), and let \( \epsilon > 0 \). Then for every \( A \in \mathcal{A} \) there is a set \( A_\epsilon \in \mathcal{A} \) with \( A_\epsilon \subseteq A \), \( \nu(A_\epsilon) \geq \nu(A) \), and \( \nu(B) > -\epsilon \) for all \( B \in \mathcal{A}, \ B \subseteq A_\epsilon \).

**Proof:** Fix \( \epsilon > 0 \) and \( A \in \mathcal{A} \). If \( \nu(A) \leq 0 \), we may take \( A_\epsilon = \emptyset \). Assume \( \nu(A) > 0 \) (in particular, \( \nu(A) \in \mathbb{R} \)). Suppose the conclusion is false. Then there is a \( B_1 \in \mathcal{A}, \ B_1 \subseteq A \), with \( \nu(B_1) \leq -\epsilon \) (otherwise we could take \( A_\epsilon = A \)). Then \( \nu(A \setminus B_1) = \nu(A) - \nu(B_1) \geq \nu(A) + \epsilon \). There is a \( B_2 \in \mathcal{A}, \ B_2 \subseteq (A \setminus B_1) \), with \( \nu(B_2) \leq -\epsilon \) (otherwise we could take \( A_\epsilon = (A \setminus B_1) \)). Proceeding in this way, we generate a sequence \((B_n)\) of pairwise disjoint sets in \( \mathcal{A} \) with \( B_n \subseteq A \) and \( \nu(B_n) \leq -\epsilon \) for all \( n \). If \( B = \bigcup_{n=1}^\infty B_n \), then \( B \in \mathcal{A}, \ B \subseteq A \), and \( \nu(B) = -\infty \), contradicting XIII.6.1.10..  

**XIII.6.1.13.** Lemma. Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\) not taking the value \(+\infty\). If \( A \in \mathcal{A} \), then \( A \) contains a subset \( P \) with \( P \in \mathcal{A}, \ \nu(P) \geq \nu(A) \), and \( P \) a positive set for \( \nu \).

**Proof:** Let \( A \in \mathcal{A} \). If \( \nu(A) \leq 0 \) (in particular, if \( \nu(A) = -\infty \)), we may take \( P = \emptyset \). Assume \( \nu(A) > 0 \) (in particular, \( \nu(A) \in \mathbb{R} \)). Generate a decreasing sequence \((A_n)\) of subsets of \( A \) by setting \( A_0 = A \) and \( A_{n+1} = (A_n)_{c} \) (XIII.6.1.12) with \( \epsilon = \frac{1}{n+1} \). Set \( P = \bigcap_{n=1}^\infty A_n \). Then \( P \in \mathcal{A}, \ P \subseteq A \), and
\[
\nu(P) = \lim_{n \to \infty} \nu(A_n) \geq \nu(A)
\]
by XIII.6.1.11 (ii) and the fact that \( \nu(A_n) \geq \nu(A) \) for all \( n \). Also, if \( B \in \mathcal{A}, \ B \subseteq P \), then \( B \subseteq A_n \) for each \( n \), and so \( \nu(B) > -\frac{1}{n} \) for all \( n \geq 1 \). Hence \( \nu(B) \geq 0 \). Thus \( P \) is positive for \( \nu \).

We can now prove that a signed measure which does not take an infinite value is actually bounded on that side.

**XIII.6.1.14.** Lemma. Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\). If \( \nu \) does not take the value \(+\infty\), then \( \{\nu(A) : A \in \mathcal{A}\} \) is bounded above. If \( \nu \) does not take the value \(-\infty\), then \( \{\nu(A) : A \in \mathcal{A}\} \) is bounded below.

**Proof:** We prove only the first statement; the second follows by replacing \( \nu \) by \(-\nu \).

Suppose for each \( n \) there is a set \( A_n \in \mathcal{A} \) with \( \nu(A_n) > n \). By XIII.6.1.13 there is a set \( P_n \in \mathcal{A} \) with \( \nu(P_n) \geq \nu(A_n) > n \) which is positive for \( \nu \). Set \( P = \bigcup_{n=1}^\infty P_n \). Then \( P \in \mathcal{A} \) and is positive for \( \nu \); thus \( \nu(P) \geq \nu(P_n) > n \) for all \( n \), i.e. \( \nu(P) = +\infty \).  

1465
We can now prove Theorem XIII.6.1.6.

Proof: Replacing $\nu$ by $-\nu$ if necessary, we may assume that $\nu$ does not take the value $+\infty$. Let $M = \sup\{\nu(A) : A \in \mathcal{A}\}$. Then $0 \leq M < +\infty$ by XIII.6.1.14. (M $\geq 0$ since $\nu(\emptyset) = 0$). Let $(A_n)$ be a sequence of sets in $\mathcal{A}$ with $\sup_n \nu(A_n) = M$. By XIII.6.1.13., for each $n$ there is a set $P_n \in \mathcal{A}$ with $\nu(P_n) \geq \nu(A_n)$ which is positive for $\nu$. Then $\sup_n \nu(P_n) = M$. Set $P = \bigcup_{n=1}^{\infty} P_n$. Then $P$ is a positive set for $\nu$, and $\nu(P) \geq \nu(P_n)$ for all $n$; thus $\nu(P) = M$. Let $N = X \setminus P$. If $B \in \mathcal{A}$, $B \subseteq N$, then

$$M + \nu(B) = \nu(P \cup B) \leq M$$

and hence $\nu(B) \leq 0$; thus $N$ is a negative set for $\nu$, and $(P, N)$ is a Hahn decomposition for $\nu$.

To show uniqueness, if $(P', N')$ is another Hahn decomposition, let $A$ be a subset of $P \setminus P'$. Then $\nu(A) \geq 0$ since $A \subseteq P$, and $\nu(A) \leq 0$ since $A \subseteq N'$. Hence $P \setminus P'$ is a null set. Similarly, $P' \setminus P$, $N \setminus N'$, and $N' \setminus N$ are null sets.

This completes the proof of XIII.6.1.6.

The proof actually shows the following:

XIII.6.1.15. Corollary. Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$. If $L = \inf\{\nu(A) : A \in \mathcal{A}\}$ and $M = \sup\{\nu(A) : A \in \mathcal{A}\}$, then:

(i) At least one of $L$ and $M$ is finite (note that $L \leq 0$ and $M \geq 0$ since $\nu(\emptyset) = 0$).

(ii) A pair $(P, N)$ of sets in $\mathcal{A}$ is a Hahn decomposition for $\nu$ if and only if $N = X \setminus P$, $\nu(P) = M$, and $\nu(N) = L$.

(iii) There exists at least one pair $(P, N)$ as in (ii).

See [Kön97] for a detailed discussion of signed premeasures and versions of the Carathéodory Extension Theorem.

The Total Variation of a Signed Measure

XIII.6.1.16. Definition. Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$, with Jordan decomposition $\nu = \nu^+ - \nu^-$. The measure $|\nu| = \nu^+ + \nu^-$ is called the total variation measure, or total variation, of $\nu$.

The next proposition is nearly trivial (given our previous results). Part (ii) is preparation for the definition of the total variation of a complex measure (XIII.6.2.6.), which is less trivial. This part is strongly reminiscent of the definition of the total variation of a function (). (This similarity is no accident! See ()).

XIII.6.1.17. Proposition. Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$. If $E \in \mathcal{A}$, then

(i) $|\nu(E)| \leq |\nu|(E)$.

(ii) $|\nu|(E) = \sup(\sum_{n=1}^{\infty} |\nu(E_n)|)$, where the supremum runs over all measurable partitions $\{E_n\}$ of $E$.  

1466
Proof: (i) is obvious. For (ii), let $S$ be the supremum. If $\{E_n\}$ is a measurable partition of $E$, then since $|\nu|$ is a measure,

$$
\sum_{n=1}^{\infty} |\nu(E_n)| \leq \sum_{n=1}^{\infty} |\nu|(E_n) = |\nu| \left( \bigcup_{n=1}^{\infty} E_n \right) = |\nu|(E)
$$

using (i), so $S \leq |\nu|(E)$. On the other hand, if $(P, N)$ is a Hahn decomposition for $\nu$, we have

$$
|\nu|(E) = \nu^+(E) + \nu^-(E) = |\nu(E \cap P)| + |\nu(E \cap N)| \leq S.
$$

An easy consequence of (ii) is:

**XIII.6.1.18. Corollary.** Let $\nu_1$ and $\nu_2$ be signed measures on a measurable space $(X, \mathcal{A})$. Suppose $\nu_1 + \nu_2$ is well defined as a signed measure (i.e. both $\nu_1$ and $\nu_2$ are bounded above or both are bounded below, e.g. if at least one is finite). Then $|\nu_1 + \nu_2| \leq |\nu_1| + |\nu_2|$ (i.e. $|\nu_1 + \nu_2|(E) \leq |\nu_1|(E) + |\nu_2|(E)$ for all $E \in \mathcal{A}$).

Note that there is no simple way to obtain a Hahn decomposition for $\nu_1 + \nu_2$ from Hahn decompositions for $\nu_1$ and $\nu_2$ in general.

**XIII.6.2. Complex Measures**

**XIII.6.2.1. Definition.** If $(X, \mathcal{A})$ is a measurable space, a *complex measure* on $(X, \mathcal{A})$ is a function $\nu : \mathcal{A} \to \mathbb{C}$ such that

(i) $\nu(\emptyset) = 0$.

(ii) If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then $\sum_{n=1}^{\infty} \nu(A_n)$ converges to $\nu(\bigcup_{n=1}^{\infty} A_n)$.

**XIII.6.2.2. Examples.** (i) Any finite signed measure (in particular, any finite measure) is a complex measure.

(ii) If $\nu_1$ and $\nu_2$ are finite signed measures on $(X, \mathcal{A})$, then $\nu = \nu_1 + i\nu_2$ is a complex measure on $(X, \mathcal{A})$. More generally, any (finite) complex linear combination of complex measures on $(X, \mathcal{A})$ is a complex measure on $(X, \mathcal{A})$.

The following converse to (ii) essentially reduces the theory of complex measures to the theory of finite signed measures, and hence to the theory of ordinary finite measures:

**XIII.6.2.3. Proposition.** Let $\nu$ be a complex measure on a measurable space $(X, \mathcal{A})$. For $E \in \mathcal{A}$, define $\nu_{re}(E) = \text{Re}(\nu(E))$ and $\nu_{im}(E) = \text{Im}(\nu(E))$. Then $\nu_{re}$ and $\nu_{im}$ are finite signed measures on $(X, \mathcal{A})$, called the *real* and *imaginary parts* of $\nu$ respectively.

Conversely, if $\nu_1$ and $\nu_2$ are finite signed measures on $(X, \mathcal{A})$, and $\nu = \nu_1 + i\nu_2$, then $\nu_{re} = \nu_1$, $\nu_{im} = \nu_2$.  

1467
XIII.6.2.4. To a complex measure $\nu$ there are thus four canonically associated finite measures $\nu_{re}^+$, $\nu_{re}^-$, $\nu_{im}^+$, and $\nu_{im}^-$. Note, however, that there may be overlap between the supports of $\nu_{re}^+$ and $\nu_{im}^+$, i.e. the measures $|\nu_{re}|$ and $|\nu_{im}|$ may not be mutually singular. Thus there is no analog of a Hahn decomposition of $X$ into “real” and “imaginary” subsets for $\nu$ in general.

XIII.6.2.5. There is a slight difficulty in making a good definition of the total variation of a complex measure. A reasonable approach is to take the measure

$$|\nu_{re}| + |\nu_{im}| = \nu_{re}^+ + \nu_{re}^- + \nu_{im}^+ + \nu_{im}^-$$

but this measure does not have all the desirable properties of a total variation. Instead it is customary to take the following definition:

XIII.6.2.6. **Definition.** Let $\nu$ be a complex measure on a measurable space $(X, A)$. If $E \in A$, define $|\nu|(E) = \sup(\sum_{n=1}^{\infty} |\nu(E_n)|)$, where the supremum runs over all measurable partitions $\{E_n\}$ of $E$.

XIII.6.2.7. **Theorem.** Let $\nu$ be a complex measure on a measurable space $(X, A)$. Then

(i) $|\nu|$ is a measure on $(X, A)$.
(ii) If $E \in A$, then $|\nu(E)| \leq |\nu|(E)$.
(iii) If $E \in A$, then

$$\max(|\nu_{re}(E)|, |\nu_{im}(E)|) \leq |\nu|(E) \leq |\nu_{re}(E)| + |\nu_{im}(E)|.$$ 

In particular, $|\nu|$ is a finite measure, and since

$$\frac{1}{2} \left[ |\nu_{re}(E)| + |\nu_{im}(E)| \right] \leq \max(|\nu_{re}(E)|, |\nu_{im}(E)|)$$

we have

$$\frac{1}{2} \left[ |\nu_{re}(E)| + |\nu_{im}(E)| \right] \leq |\nu|(E) \leq |\nu_{re}(E)| + |\nu_{im}(E)|.$$ 

**Proof:** (ii): This is trivial since $\{E\}$ is a measurable partition of $E$.
(iii): Note that if $x, y \in \mathbb{R}$, then

$$\frac{1}{2} \left[ |x| + |y| \right] \leq \max(|x|, |y|) \leq |x + iy| \leq |x| + |y|$$

(). If $\{E_n\}$ is a measurable partition of $E$, then

$$\max \left( \sum_{n=1}^{\infty} |\nu_{re}(E_n)|, \sum_{n=1}^{\infty} |\nu_{im}(E_n)| \right) \leq \sum_{n=1}^{\infty} \left[ \max(|\nu_{re}(E_n)|, |\nu_{im}(E_n)|) \right]$$

$$\leq \sum_{n=1}^{\infty} |\nu(E_n)| \leq \sum_{n=1}^{\infty} \left[ |\nu_{re}(E_n)| + |\nu_{im}(E_n)| \right] = \sum_{n=1}^{\infty} |\nu_{re}(E_n)| + \sum_{n=1}^{\infty} |\nu_{im}(E_n)|.$$ 

1468
and the result follows by taking the supremum over all measurable partitions of $E$ and applying XIII.6.17 (ii).

In particular, since $|\nu(E)|$ and $|\nu_m|(E)$ are finite, so is $|\nu|(E)$.

(i): It is obvious that $|\nu|(\emptyset) = 0$. Let $\{A_m\}$ be a pairwise disjoint sets in $\mathcal{A}$, and let $A = \cup_{m=1}^{\infty} A_m$. Fix $\epsilon > 0$. Let $\{E_n\}$ be a measurable partition of $A$ with $\sum_{n=1}^{\infty} |\nu(E_n)| > |\nu|(A) - \epsilon$ ($|\nu|(A)$ is finite by (iii)). Set $A_{n,m} = E_n \cap A_m$ for each $n, m$. Then

$$|\nu|(A) - \epsilon \leq \sum_{n=1}^{\infty} |\nu(E_n)| = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \nu(A_{n,m}) \leq \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\nu|(A_m)$$

(the equality in the last line uses (i)). Since $\epsilon > 0$ is arbitrary, $|\nu|(A) \leq \sum_{m=1}^{\infty} |\nu|(A_m)$. Conversely, if $\epsilon > 0$, for each $m$ let $\{E_{n,m}\}$ be a measurable partition of $A_m$ with

$$\sum_{n=1}^{\infty} |\nu(E_{n,m})| > |\nu|(A_m) - 2^{-m}\epsilon .$$

Then $\{E_{n,m}\}$ is a measurable partition of $A$, so

$$\left(\sum_{m=1}^{\infty} |\nu|(A_m)\right) - \epsilon = \sum_{m=1}^{\infty} |\nu|(A_m) - 2^{-m}\epsilon < \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |\nu(E_{n,m})| \leq |\nu|(A) .$$

Since $\epsilon > 0$ is arbitrary, $\sum_{m=1}^{\infty} |\nu|(A_m) \leq |\nu|(A)$.

XIII.6.2.8. PROPOSITION. Let $\nu_1$ and $\nu_2$ be complex measures on $(X, \mathcal{A})$, and let $\alpha \in \mathbb{C}$. Then

(i) $|\nu_1 + \nu_2| \leq |\nu_1| + |\nu_2|$ (i.e. $|\nu_1 + \nu_2|(E) \leq |\nu_1|(E) + |\nu_2|(E)$ for all $E \in \mathcal{A}$).

(ii) $|\alpha \nu_1| = |\alpha||\nu_1|$.

XIII.6.3. Integration With Respect to Signed and Complex Measures

There is nothing difficult or complicated about extending integration to the case of signed or complex measures; we need only make the right definitions, which are fairly obvious, guided by the desire to preserve linearity.

The only slight difficulty is deciding what should be meant by an integrable function. But the definition here is fairly obvious too, because of the following fact:

XIII.6.3.1. PROPOSITION. (i) Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$, and $f : X \to \mathbb{R}$ an $\mathcal{A}$-measurable function. Then $f \in L^1(X, \mathcal{A}, |\nu|)$ if and only if $f \in L^1(X, \mathcal{A}, \nu^+)$ and $f \in L^1(X, \mathcal{A}, \nu^-)$.

(ii) Let $\nu$ be a complex measure on a measurable space $(X, \mathcal{A})$, and $f : X \to \mathbb{C}$ an $\mathcal{A}$-measurable function. Then $f \in L^1_c(X, \mathcal{A}, |\nu|)$ if and only if $f \in L^1_c(X, \mathcal{A}, |\nu_+|)$ and $f \in L^1_c(X, \mathcal{A}, |\nu_-|)$.

PROOF: Part (i) is obvious since $\int_X |f| \, d|\nu| = \int_X |f| \, d\nu^+ + \int_X |f| \, d\nu^-$, and (ii) follows easily from the same argument and XIII.6.2.7 (iii).
**XIII.6.3.2.** Definition. Let \( \nu \) be a signed measure or complex measure on a measurable space \((X, \mathcal{A})\), and \( f \) a real- or complex-valued \( \mathcal{A} \)-measurable function. Then \( f \in \mathcal{L}^1(X, \mathcal{A}, \nu) \) if \( f \) is complex-valued and \( |f| \in \mathcal{L}^1(X, \mathcal{A}, |\nu|) \).

We then make the following obvious definition:

**XIII.6.3.3.** Definition. (i) Let \( \nu \) be a signed measure on a measurable space \((X, \mathcal{A})\), and \( f \in \mathcal{L}^1(X, \mathcal{A}, \nu) \) or \( \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \). Then

\[
\int_X f \, d\nu = \int_X f \, d\nu^+ - \int_X f \, d\nu^- .
\]

(ii) Let \( \nu \) be a complex measure on a measurable space \((X, \mathcal{A})\), and \( f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \). Then

\[
\int_X f \, d\nu = \int_X f \, d\nu_{re} + i \int_X f \, d\nu_{im} .
\]

(iii) If \( A \in \mathcal{A} \), then in either of the above cases \( \int_A f \, d\nu = \int_X f \chi_A \, d\nu \).

**XIII.6.3.4.** The integrals defined in this way have almost all the properties of integrals with respect to ordinary measures; the only obvious exceptions are positivity and monotonicity. It thus only makes sense to integrate \( \mathcal{L}^1 \)-functions, not arbitrary nonnegative measurable functions.

We list the important properties of this extended integral. The proofs are simple applications of the results for nonnegative measures, and are mostly omitted.

**XIII.6.3.5.** Proposition. Let \( \nu \) be a signed or complex measure on a measurable space \((X, \mathcal{A})\). If \( f, g \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \), then

\[
\int_A (f + g) \, d\nu = \int_A f \, d\nu + \int_A g \, d\nu
\]

for all \( A \in \mathcal{A} \). If \( f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \) and \( \alpha \in \mathbb{C} \), then

\[
\int_A \alpha f \, d\nu = \alpha \int_A f \, d\nu
\]

for all \( A \in \mathcal{A} \).

For the proof of the last statement, see Exercise XIII.6.4.2.

**XIII.6.3.6.** If \( \nu \) is a signed or complex measure and \( f \in \mathcal{L}^1(\nu) \), then in general

\[
\left| \int_X f \, d\nu \right| \leq \int |f| \, d\nu
\]

(the right side is not necessarily nonnegative or even real). However, we have:
XIII.6.3.7. PROPOSITION. Let \( \nu \) be a signed or complex measure on a measurable space \((X, \mathcal{A})\), and let \( f \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \). Then
\[
\left| \int_A f \, d\nu \right| \leq \int_A |f| \, d|\nu|
\]
for any \( A \in \mathcal{A} \).

Because of the lack of monotonicity, there are no versions of the Monotone Convergence Theorem or Fatou’s Lemma for signed or complex measures. But the Dominated Convergence Theorem holds with an identical statement:

XIII.6.3.8. THEOREM. [DOMINATED CONVERGENCE THEOREM] Let \( \nu \) be a signed or complex measure on a measurable space \((X, \mathcal{A})\). Let \((f_n)\) be a sequence of functions in \(L^1_{\mathbb{C}}(X, \mathcal{A}, \nu)\) converging pointwise a.e. to a function \( f \). If there is a \( g \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \) with \( |f_n| \leq g \) a.e. for all \( n \), then
\[
\int_X f \, d\nu = \lim_{n \to \infty} \int_X f_n \, d\nu .
\]

There is a version for infinite series. We omit the general statement, but note the following important special case:

XIII.6.3.9. PROPOSITION. Let \( \nu \) be a signed or complex measure on a measurable space \((X, \mathcal{A})\), and let \( f \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \). Let \( \{A_n\} \) be a sequence of pairwise disjoint sets in \( \mathcal{A} \), with \( A = \bigcup_{n=1}^{\infty} A_n \). Then the series
\[
\sum_{n=1}^{\infty} \int_{A_n} f \, d\nu
\]
converges absolutely to \( \int_A f \, d\nu \).

XIII.6.3.10. As a consequence, if \( \nu \) is a complex measure on \((X, \mathcal{A})\), and \( f \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu) \), and \( \sigma \) is defined on \( \mathcal{A} \) by \( \sigma(A) = \int_A f \, d\nu \), then \( \sigma \) is also a complex measure on \( \mathcal{A} \). In this case, we have that \( |\sigma|(A) = \int_A |f| \, d|\nu| \) (XIV.9.1).

XIII.6.4. Exercises

XIII.6.4.1. (a) Let \((X, \mathcal{A}, \mu)\) be a measure space, and \( f \in L^1_{\mathbb{R}}(X, \mathcal{A}, \mu) \). Define \( \nu(A) = \int_A f \, d\mu \) as in XIII.6.1.2. and XIII.6.1.8. Show that, for all \( A \in \mathcal{A} \), \( |\nu|(A) = \int_A |f| \, d\mu \).

(b) Show the same for \( f \in L^1_{\mathbb{C}}(X, \mathcal{A}, \mu) \). [Hint: for \( n \in \mathbb{N} \) and \( 1 \leq k \leq n \), consider the set
\[
A_{n,k} = \left\{ x \in A : \frac{2\pi(k-1)}{n} \leq \arg(f(x)) < \frac{2\pi k}{n} \right\}.
\]

1471
XIII.6.4.2. (a) Let $\nu$ be a complex measure on a measurable space $(X, \mathcal{A})$, and $f \in L^1_C(X, \mathcal{A}, \nu)$. Show that
\[
\int_A f \, d\nu = \left[ \int_A \text{Re}(f) \, d\nu + i \int_A \text{Im}(f) \, d\nu \right]
\]
for any $A \in \mathcal{A}$.
(b) Show that if $\alpha \in \mathbb{C}$, then
\[
\int_X \alpha f \, d\nu = \alpha \int_X f \, d\nu.
\]

XIII.6.4.3. Let $\nu$ be a signed measure on a measurable space $(X, \mathcal{A})$, and $f \in L^1_{\mathbb{R}}(X, \mathcal{A}, \nu)$. Define a signed measure $\sigma$ by $\sigma(A) = \int_A f \, d\nu$ for $A \in \mathcal{A}$.
(a) Find a Hahn decomposition for $\sigma$, and the Jordan decomposition.
(b) Show that $|\sigma|(A) = \int_A |f| \, d|\nu|$ for $A \in \mathcal{A}$.

XIII.6.4.4. Try to extend the theory of complex measures to include measures of the form $\nu_1 + i\nu_2$, where $\nu_1$ and $\nu_2$ are signed measures, not necessarily finite. Try to give a definition as in XIII.6.1.1. and XIII.6.2.1. which covers this case (clumsy and unnatural, isn’t it?) Discuss the properties which can and cannot be established in this case, and the difficulties and limitations of such a theory.

XIII.6.4.5. A finitely additive signed measure on an algebra can be finite but unbounded, and can fail to have a Hahn decomposition.
(a) Let $\mathcal{A}$ be the algebra of finite/cofinite subsets of $\mathbb{N}$. If $A \in \mathcal{A}$ is finite, let $\nu(A)$ be the number of elements in $A$; if $A$ is cofinite, let $\nu(A) = -n$, where $n$ is the number of elements in $A^c$. Show that $\nu$ is a finitely additive signed measure on $(\mathbb{N}, \mathcal{A})$.
(b) Show that there is no Hahn decomposition for $\nu$. 

1472
XIII.7. Limits of Measures

In this section, we discuss the following question: if \((\mu_i)\) is a sequence (or net) of measures, and \(\mu_i \to \mu\), is \(\mu\) a measure?

XIII.7.1. General Considerations

We must first make sense of what it means to say that \(\mu_i \to \mu\). There is one obvious minimal sense to interpret this:

XIII.7.1.1. Definition. Let \(X\) be a set, \(\mathcal{A}\) a collection of subsets of \(X\), and \(Y\) a topological space. Then a sequence (or net) \((f_i)\) of functions from \(\mathcal{A}\) to \(Y\) converges setwise to the function \(f : \mathcal{A} \to Y\) if \(f_i(A) \to f(A)\) for all \(A \in \mathcal{A}\). In particular, if \((\mu_i)\) is a sequence (or net) of measures on a measurable space \((X, \mathcal{A})\), and \(\mu : \mathcal{A} \to [0, \infty]\), then \(\mu_i \to \mu\) setwise if \(\mu_i(A) \to \mu(A)\) for all \(A \in \mathcal{A}\). Setwise convergence of a sequence (or net) \((\nu_i)\) of complex measures to a function \(\nu : \mathcal{A} \to \mathbb{C}\) is defined in the same way.

In other words: a measure on \((X, \mathcal{A})\) may be thought of as an element of the product space \([0, \infty]^\mathcal{A}\), the set of all functions from \(\mathcal{A}\) to \([0, \infty]\). Then \(\mu_i \to \mu\) setwise if \(\mu_i \to \mu\) in the product topology. Similarly, a complex measure on \((X, \mathcal{A})\) is an element of \(\mathbb{C}^\mathcal{A}\), and setwise convergence is convergence in the product topology.

There are stronger notions of convergence; see, for example, Exercise XIII.7.4.1.

So the basic question is:

XIII.7.1.2. Question. If \((\mu_i)\) is a sequence (or net) of measures (ordinary or complex), and \(\mu_i \to \mu\) setwise, is \(\mu\) is also a measure (or complex measure)?

This question can be rephrased: is the set of measures a closed subset of \([0, \infty]^\mathcal{A}\)? Is the set of complex measures a closed subset of \(\mathbb{C}^\mathcal{A}\)?

The following simple examples show that the answer is no in general.

XIII.7.1.3. Examples. (i) On \((\mathbb{N}, \mathcal{P}(\mathbb{N}))\), let \(\mu_n\) be counting measure on \(\{m : m > n\}\), regarded as a measure on \((\mathbb{N}, \mathcal{P}(\mathbb{N}))\) by setting \(\mu_n(E) = 0\) if \(E \subseteq \{1, \ldots, n\}\). Then each \(\mu_n\) is a \(\sigma\)-finite measure, \((\mu_n(E))\) is decreasing for each \(E\), and \(\mu(E) = \lim_{n \to \infty} \mu_n(E)\) exists for each \(E \subseteq \mathbb{N}\). But \(\mu(E) = 0\) if \(E\) is finite, and \(\mu(\mathbb{N}) = \infty\) (in fact, \(\mu(E) = \infty\) if \(E\) is any infinite set), so \(\mu\) is not a measure.

(ii) Again on \((\mathbb{N}, \mathcal{P}(\mathbb{N}))\), let \(\delta_n\) be the unit point mass at \(n\), i.e. for \(E \subseteq \mathbb{N}\), \(\delta_n(E) = 1\) if \(n \in E\) and \(\delta_n(E) = 0\) if \(n \notin E\). Each \(\delta_n\) gives a point of \(\{0, 1\}^\mathcal{P}(\mathbb{N})\), which is compact in the product topology; thus there is a subnet of the sequence \((\delta_n)\) which converges setwise to a function \(\mu : \mathcal{P}(\mathbb{N}) \to \{0, 1\}\). But \(\mu(E) = 0\) for all finite \(E\), while \(\mu(\mathbb{N}) = 1\), and thus \(\mu\) cannot be a measure.

Example (ii) is rather generic: there are grave difficulties in taking limits of nets of measures which are not sequences. But, rather remarkably, often (i.e. under mild hypotheses) it will turn out that limits of sequences of measures are measures.

A setwise limit of measures is always at least a finitely additive measure:
XIII.7.1.4. Proposition. Let \((X, \mathcal{A})\) be a measurable space. Then the set of finitely additive measures on \((X, \mathcal{A})\) is closed under setwise limits (i.e. is a closed subset of \([0, \infty]^\mathcal{A}\)). The set of finitely additive complex measures is a closed subset of \(\mathbb{C}^\mathcal{A}\).

This is an immediate consequence of ()

So the whole difficulty is with countable additivity.

XIII.7.2. Monotone and Dominated Convergence

The first result is an analog of the Monotone Convergence Theorem:

XIII.7.2.1. Theorem. Let \((\mu_i)\) be a net of measures on a measurable space \((X, \mathcal{A})\). If \((\mu_i(A))\) is increasing for each \(A \in \mathcal{A}\), then

\[
\mu = \lim_{i} \mu_i = \sup_{i} \mu_i \quad \text{(defined by } \mu(A) = \sup_{i} \mu_i(A) \text{ for } A \in \mathcal{A})
\]

is a measure on \((X, \mathcal{A})\).

Proof: We only need to show countable additivity. Suppose \((A_n)\) is a pairwise disjoint sequence in \(\mathcal{A}\), and set \(A = \bigcup_{n=1}^{\infty} A_n\). We have

\[
\sum_{n=1}^{\infty} \mu(A_n) = \sup_{N} \sum_{n=1}^{N} \mu(A_n) = \sup_{N} \left[ \mu \left( \bigcup_{n=1}^{N} A_n \right) \right] \leq \mu(A)
\]

by finite additivity and monotonicity. To show the reverse inequality, first suppose \(\mu(A) < \infty\). Let \(\epsilon > 0\) and choose \(i_0\) such that \(\mu_i(A) > \mu(A) - \epsilon\). Then we have

\[
\mu(A) - \epsilon < \mu_{i_0}(A) = \sum_{n=1}^{\infty} \mu_i(A_n) \leq \sum_{n=1}^{\infty} \mu(A_n).
\]

Since \(\epsilon > 0\) is arbitrary, \(\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)\).

Now suppose \(\mu(A) = \infty\). Fix \(M > 0\), and choose \(i_0\) such that \(\mu_{i_0}(A) > M\). We then have

\[
M < \mu_{i_0}(A) = \sum_{n=1}^{\infty} \mu_i(A_n) \leq \sum_{n=1}^{\infty} \mu(A_n),
\]

so \(\sum_{n=1}^{\infty} \mu(A_n) = +\infty\).

This implies that arbitrary sums of measures are measures (finite sums of measures are obviously measures ()).

XIII.7.2.2. Corollary. Let \(\{\mu_i : i \in I\}\) be a set of measures on a measurable space \((X, \mathcal{A})\). For each \(A \in \mathcal{A}\), define

\[
\mu(A) = \sum_{i \in I} \mu_i(A) = \sup \left\{ \sum_{i \in F} \mu_i(A) : F \subseteq I \text{ finite} \right\}.
\]

Then \(\mu\) is a measure on \((X, \mathcal{A})\).

There is also a version of the Dominated Convergence Theorem:
XIII.7.2.3. Theorem. Let $(\nu_i)$ be a net of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\nu$. If there is a finite measure $\mu$ such that $|\nu_i| \leq \mu$ for all $i$, then $\nu$ is a complex measure.

(Of course, this theorem applies in the case the $\nu_i$ are ordinary measures uniformly dominated by a finite measure $\mu$.)

In fact, this result holds in greater generality. We make some definitions:

XIII.7.2.4. Definition. Let $(\nu_i)_{i \in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$, and $\mu$ a measure on $(X, \mathcal{A})$. Then $(\nu_i)$ is uniformly absolutely continuous with respect to $\mu$ if for every $\epsilon > 0$ there is a $\delta > 0$ and an $i_0 \in I$ such that, for all $A \in \mathcal{A}$ with $\mu(A) < \delta$,

$$|\nu_i(A)| < \epsilon \text{ for all } i \geq i_0.$$ 

The term “uniformly absolutely continuous” is appropriate for this notion, in analogy with the characterization of absolute continuity of measures in ()

Under the hypotheses of XIII.7.2.5., $(\nu_i)$ is uniformly absolutely continuous with respect to $\mu$ (take $\delta = \epsilon$). Thus Theorem XIII.7.2.5. is a special case of the following result:

XIII.7.2.5. Theorem. Let $(\nu_i)$ be a net of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\nu : \mathcal{A} \to \mathbb{C}$. If there is a finite measure $\mu$ such that $(\nu_i)$ is uniformly absolutely continuous with respect to $\mu$, then $\nu$ is a complex measure.

We can even go to one step greater generality:

XIII.7.2.6. Definition. Let $(\nu_i)_{i \in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$. Then $(\nu_i)$ is uniformly countably additive if, whenever $(A_n)$ is a sequence of pairwise disjoint sets in $\mathcal{A}$ with $A = \bigcup_{n=1}^{\infty} A_n$, and $\epsilon > 0$, there is an $N_0 \in \mathbb{N}$ and an $i_0 \in I$ such that

$$|\nu_i(A) - \nu_i\left(\bigcup_{n=1}^{N} A_n\right)| < \epsilon$$

for all $N \geq N_0$ and $i \geq i_0$.

XIII.7.2.7. Proposition. Let $(\nu_i)_{i \in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$. Then $(\nu_i)$ is uniformly countably additive if and only if, whenever $(B_n)$ is a decreasing sequence in $\mathcal{A}$ with $\bigcap_{n=1}^{\infty} B_n = \emptyset$, we have $\lim_{n \to \infty} \nu_i(B_n) = 0$ uniformly in $i$, i.e. for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ and $i_0 \in I$ such that

$$|\nu_i(B_n)| < \epsilon$$

for all $n \geq n_0$ and $i \geq i_0$.

Proof: If $(\nu_i)$ is uniformly countably additive and $(B_n)$ is decreasing with empty intersection, set $A_n = B_n \setminus B_{n+1}$. Conversely, if $(A_n)$ is a pairwise disjoint sequence, let $B_n = \bigcup_{k \geq n} A_k$. ☑
XIII.7.2.8. **Corollary.** Let \((\nu_i)_{i \in I}\) be a net of complex measures on a measurable space \((X, \mathcal{A})\). If \((\nu_i)\) is uniformly countably additive, then whenever \((A_n)\) is a pairwise disjoint sequence in \(\mathcal{A}\) we have 
\[
\lim_{n \to \infty} \nu_i(A_n) = 0 \text{ uniformly in } i.
\]

The converse is also true for a sequence of complex measures, and is a key step in the direct proof of the Nikodým Convergence Theorem (Exercise XIII.7.4.2.).

**Proof:** Set \(B_n = \bigcup_{k=n}^{\infty} A_k\). Then \(\nu_i(B_n) \to 0 \text{ uniformly in } i\), and 
\[
|\nu_i(A_n)| \leq |\nu_i(B_n)| + |\nu_i(B_{n+1})| \text{ for all } n \text{ and } i.
\]


XIII.7.2.9. **Proposition.** Let \((\nu_i)\) be a net of complex measures on a measurable space \((X, \mathcal{A})\). If there is a finite measure \(\mu\) such that \((\nu_i)\) is uniformly absolutely continuous with respect to \(\mu\), then \((\nu_i)\) is uniformly countably additive.

The converse is true for a setwise convergent sequence of complex measures (XIII.7.3.8.).

**Proof:** Use XIII.7.2.7. Let \(\epsilon > 0\), and fix \(\delta_0\) and \(i_0\) as in Definition XIII.7.2.4. Let \((B_n)\) be a decreasing sequence in \(\mathcal{A}\) with \(\cap_{n=1}^{\infty} B_n = \emptyset\). Fix \(n_0\) such that \(\mu(B_n) < \delta\) for all \(n \geq n_0\). Then, for all \(n \geq n_0\) and \(i \geq i_0\), 
\[
|\nu_i(B_n)| < \epsilon.
\]

We then obtain the most general version of the Dominated Convergence Theorem for measures (note that there is no actual domination in this version!)

XIII.7.2.10. **Theorem.** Let \((\nu_i)\) be a net of complex measures on a measurable space \((X, \mathcal{A})\), converging setwise to a function \(\nu : \mathcal{A} \to \mathbb{C}\). If \((\nu_i)\) is uniformly countably additive, then \(\nu\) is a complex measure.

**Proof:** We again need only show countable additivity. Since \(\nu\) is finite-valued, it suffices to show () that if \((B_n)\) is a decreasing sequence in \(\mathcal{A}\) with \(\cap_{n=1}^{\infty} B_n = \emptyset\), we have \(\lim_{n \to \infty} \nu(B_n) = 0\). Let \(\epsilon > 0\), and fix a corresponding \(n_0\) and \(i_0\) as in the definition of uniform countable additivity (XIII.7.2.6.). Then, if \(n \geq n_0\), then \(|\nu_i(B_n)| < \epsilon\) for all \(i \geq i_0\), and thus \(|\nu(A_n)| \leq \epsilon\).

XIII.7.3. **The Nikodým and Vitali-Hahn-Saks Theorems**

We now restrict attention to sequences of measures. There is a circle of results due to NIKODÝM, VITALI, HAHN, and SAKS (there were other contributors too) which say that under certain mild restrictions, the setwise limit of a sequence of complex measures is a complex measure, and that the convergence is “almost uniform.”

These results are closely interconnected, and can be largely deduced from each other; various schemes for proving them are found in the literature. Even the names of the results are not uniform in references; what is called the “Vitali-Hahn-Saks Theorem” in some references is called the “Nikodým Convergence Theorem” in others, reflecting the close connections. We have tried to use names accurately reflecting the actual work of the originators.

The first observation is that a sequence of complex measures is dominated by a single measure:
XIII.7.3.1. PROPOSITION. Let \((\nu_n)\) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\). Then there is a finite measure \(\mu\) on \((X, \mathcal{A})\) such that \(\nu_n \ll \mu\) for all \(n\).

PROOF: We may assume all the \(\nu_n\) are nonzero. For \(A \in \mathcal{A}\), set
\[
\mu(A) = \sum_{n=1}^{\infty} \frac{|\nu_n|(A)}{2^n|\nu_n|(X)}.
\]
By XIII.7.2.2., \(\mu\) is a measure. It is obvious that \(\mu\) is finite and \(\nu_n \ll \mu\) for all \(n\).

The next theorem is the main result of this section. Vitali proved the first version as a special case of his convergence theorem (), and Hahn obtained the result for Lebesgue-Stieltjes measures. The general result is due to Saks. Our proof, also due to Saks (cf. [DS88]), uses the Baire Category Theorem and is strongly reminiscent of the proof of the Uniform Boundedness Theorem ().

XIII.7.3.2. THEOREM. [VITALI-HAHN-SAKS] Let \(\mu\) be a measure, and \((\nu_n)\) a sequence of complex measures, on a measurable space \((X, \mathcal{A})\). If \((\nu_n)\) converges setwise to a function \(\nu: \mathcal{A} \to \mathbb{C}\), and \(\nu_n \ll \mu\) for all \(n\), then \((\nu_n)\) is uniformly absolutely continuous with respect to \(\mu\).

PROOF: Let \(\Sigma(\mu)\) be the metric space of \(\mu\). Then each \(\nu_n\) gives a continuous function from \(\Sigma(\mu)\) to \(\mathbb{C}\), converging pointwise to \(\nu: \Sigma(\mu) \to \mathbb{C}\). Fix \(\epsilon > 0\). For each \(m\) and \(n\), set
\[
E_{m,n} = \{ A \in \mathcal{A} : |\nu_n(A) - \nu_m(A)| \leq \epsilon \}
\]
(considered as a subset of \(\Sigma(\mu)\)) and set \(F_k = \bigcap_{m,n \geq k} E_{m,n}\). Then \(F_k\) is a closed subset of \(\Sigma(\mu)\) for each \(k\), and since \((\nu_n)\) converges pointwise, \(\bigcup_{k=1}^{\infty} F_k = \Sigma(\mu)\). Thus, since \(\Sigma(\mu)\) is complete (), \(F_k\) has a nonempty interior for some \(k\) by the Baire Category Theorem, i.e. there is a \(k\), an \(A \in \mathcal{A}\), and an \(r > 0\) such that
\[
\{ B \in \mathcal{A} : \mu(A \Delta B) < r \} \subseteq F_k.
\]
There is a \(\delta, 0 < \delta < r\), such that \(|\nu_n(B)| < \epsilon\) for \(1 \leq n \leq k\) whenever \(B \in \mathcal{A}\) and \(\mu(B) < \delta\) (). Then, if \(B \in \mathcal{A}, \mu(B) < \delta\), we have that \(A \cup B\) and \(A \setminus B\) are in \(F_k\), so for \(n \geq k\) we have
\[
\nu_n(B) = \nu_k(B) + (\nu_n(B) - \nu_k(B))
\]
\[
= \nu_k(B) + (\nu_n(A \cup B) - \nu_k(A \cup B)) - (\nu_n(A \setminus B) - \nu_k(A \setminus B))
\]
and thus \(|\nu_n(B)| < 3\epsilon\) for all \(n\).

XIII.7.3.3. COROLLARY. [NIKODYM CONVERGENCE THEOREM] Let \((\nu_n)\) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\), converging setwise to a function \(\nu: \mathcal{A} \to \mathbb{C}\). Then \((\nu_n)\) is uniformly countably additive, and \(\nu\) is a complex measure.

PROOF: There is a finite measure \(\mu\) on \(\mathcal{A}\) with \(\nu_n \ll \mu\) for all \(n\) by XIII.7.3.1. Then \((\nu_n)\) is uniformly absolutely continuous with respect to \(\mu\) by XIII.7.3.2., and hence is uniformly countably additive by XIII.7.2.9.. Then \(\nu\) is a measure by XIII.7.2.10..
XIII.7.3.4. The Nikodým Convergence Theorem was proved before the general theorem XIII.7.3.2. was obtained. See [Swa94] for a direct proof of the Nikodým Convergence Theorem using a matrix version (cf. Exercise ()). The Vitali-Hahn-Saks Theorem can then be derived from the Nikodým Convergence Theorem (cf.). There is a direct proof of the Nikodým Convergence Theorem for ordinary (finite) measures not using the Baire Category Theorem; see Exercise XIII.7.4.3.

There are versions of these results for finitely additive measures and for vector measures; see [DU77].

A nice consequence of the Nikodým Convergence Theorem is the following result, which is a close analog of the Uniform Boundedness Principle:

XIII.7.3.5. **Theorem. [Nikodým Boundedness Theorem]** Let \( (\nu_n) \) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\). If \( \{\nu_n(A) : n \in \mathbb{N}\} \) is bounded for each \( A \in \mathcal{A} \), then the \( \nu_n \) are uniformly bounded, i.e. there is an \( M \) such that \( |\nu_n(A)| \leq M \) for all \( n \) and for all \( A \in \mathcal{A} \).

For the proof, we need a lemma:

XIII.7.3.6. **Lemma.** Let \( (\nu_n) \) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\). If \( \{\nu_n(A) : n \in \mathbb{N}\} \) is bounded for each \( A \in \mathcal{A} \), then \( (\nu_n) \) is uniformly bounded if and only if, for every sequence \( (A_n) \) of pairwise disjoint sets in \( \mathcal{A} \), the set \( \{\nu_n(A_n)\} \) is bounded.

**Proof:** The condition is obviously necessary. For sufficiency, suppose \( (\nu_n) \) is not uniformly bounded. Let \( N_0 = \sup_n \{ |\nu_n(X)| \} \). Let \( E_1 \in \mathcal{A} \) with \( |\nu_n(E_1)| > 1 + N_0 \) for some \( n_1 \); then, if \( F_1 = X \setminus E_1 \), \( |\nu_n(F_1)| \geq |\nu_n(E_1)| - |\nu_n(X)| > 1 \). Either \( \{\nu_n(A \cap E_1) : A \in \mathcal{A}, n \in \mathbb{N}\} \) or \( \{\nu_n(A \cap F_1) : A \in \mathcal{A}, n \in \mathbb{N}\} \) is unbounded; set \( B_1 = E_1 \) in the first case, and \( B_1 = F_1 \) in the second case, and let \( A_{n_1} \) be the other (i.e. \( A_{n_1} = X \setminus B_1 \)). Set \( N_1 = \sup_n \{|\nu_n(B_1)|\} \). Let \( E_2 \in \mathcal{A}, E_2 \subseteq B_1 \), with \( |\nu_n(E_2)| > 2 + N_1 \) for some \( n_2 \); we may assume \( n_2 \neq n_1 \) since \( \{\nu_n(A) : A \in \mathcal{A}\} \) is bounded \( \{\nu_n(|X|)\} \). Then, if \( F_2 = B_1 \setminus E_2, |\nu_n(F_2)| > 2 \). Let \( B_2 = E_2 \) if \( \{\nu_n(A \cap E_2) : A \in \mathcal{A}, n \in \mathbb{N}\} \) is unbounded and \( B_2 = F_2 \) if \( \{\nu_n(A \cap F_2) : A \in \mathcal{A}, n \in \mathbb{N}\} \) is unbounded, and set \( A_{n_2} = B_1 \setminus B_2 \). Continue inductively to obtain a sequence \( (n_k) \) in \( \mathbb{N} \) and a pairwise disjoint sequence \( (A_{n_k}) \) in \( \mathcal{A} \) such that \( |\nu_{n_k}(A_{n_k})| > k \). Set \( A_n = \emptyset \) for \( n \) not equal to any \( n_k \).

XIII.7.3.7. **Remark:** The hypothesis that \( (\nu_n) \) be setwise bounded is omitted from the statement of this result in [Swa94, p. 68]. This hypothesis is easily seen to be necessary: let \( X \) be a singleton and \( \nu_n(X) = n \). Also, the proof only requires that \( \mathcal{A} \) be an algebra and that the \( \nu_n \) be finitely additive if the additional hypothesis is added that each \( \nu_n \) is bounded, which is not automatic in this case (Exercise XIII.6.4.5.). (Without this hypothesis, the result fails: let \( \nu_1 \) be a finitely additive signed measure which is finite but not bounded, and \( \nu_n = 0 \) for \( n > 1 \).)

We now prove XIII.7.3.5..

**Proof:** By XIII.7.3.6. and (), it suffices to show that if \( (A_n) \) is a pairwise disjoint sequence in \( \mathcal{A} \) and \( (\alpha_n) \) is a sequence in \( \mathbb{C} \) with \( \alpha_n \to 0 \), that \( \alpha_n \nu_n(A_n) \to 0 \). But \( (\alpha_n \nu_n) \) is a sequence of complex measures on \((X, \mathcal{A})\) converging setwise to the zero measure; hence by XIII.7.3.3. the sequence \( (\alpha_n \nu_n) \) is uniformly countably additive, giving the conclusion from XIII.7.2.8. 

As a related matter, we have the following converse to XIII.7.2.9.:
XIII.7.3.8. Theorem. Let \((\nu_n)\) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\), and \(\mu\) a measure on \((X, \mathcal{A})\) such that \(\nu_n \ll \mu\) for all \(n\). If \((\nu_n)\) is uniformly countably additive, then \((\nu_n)\) is uniformly absolutely continuous with respect to \(\mu\).

**Proof:** If \((\nu_n)\) is not uniformly absolutely continuous with respect to \(\mu\), we will construct a pairwise disjoint sequence \((A_k)\) in \(\mathcal{A}\) and indices \(n_k\) such that \(\nu_{n_k}(A_k)\) does not approach 0, contradicting XIII.7.2.8.

If \((\nu_n)\) is not uniformly absolutely continuous with respect to \(\mu\), there is an \(\epsilon > 0\) such that for all \(\delta > 0\), there is a \(B \in \mathcal{A}\) with \(\mu(B) < \delta\) and \(|\nu_n(B)| \geq \epsilon\) for infinitely many \(n\). Fix such an \(\epsilon\). Choose \(B_1 \in \mathcal{A}\) and \(n_1\) with \(|\nu_{n_1}(B_1)| \geq \epsilon\) and \(\mu(B_1) < 1\). Since \(\nu_{n_1} \ll \mu\), there is a \(\delta_1 > 0\) such that \(|\nu_{n_1}(B)| < \epsilon/2\) whenever \(\mu(B) < \delta_1\). There is a \(B_2 \in \mathcal{A}\) and an \(n_2 > n_1\) such that \(|\nu_{n_2}(B_2)| \geq \epsilon\) and \(\mu(B_2) < \delta_1/2\). Continuing inductively, construct sets \(B_k \in \mathcal{A}\), numbers \(\delta_k\), and a strictly increasing sequence \((n_k)\) of indices, such that, for all \(k\), \(\delta_{k+1} < \delta_k/2\), \(|\nu_{n_k}(B_k)| \geq \epsilon\), \(\mu(B_{k+1}) < \delta_k/2\), and \(|\nu_{n_k}(B)| < \epsilon/2\) whenever \(\mu(B) < \delta_k\).

Set \(C_k = \cup_{n=k+1}^{\infty} B_n\) for each \(k\). Then

\[
\mu(C_k) \leq \sum_{n=k+1}^{\infty} \mu(B_n) < \sum_{n=1}^{\infty} \frac{\delta_k}{2^n} = \delta_k
\]

so that \(|\nu_{n_k}(B_k \cap C_k)| < \epsilon/2\) for each \(k\). Thus, if \(A_k = B_k \setminus C_k\), the \(A_k\) are pairwise disjoint and, for all \(k\),

\[
|\nu_{n_k}(A_k)| \geq |\nu_{n_k}(B_k)| - |\nu_{n_k}(B_k \cap C_k)| > \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2}.
\]

\(\Box\)

XIII.7.4. Exercises

XIII.7.4.1. Let \((\nu_n)\) be a sequence of complex measures on a measurable space \((X, \mathcal{A})\) which is a Cauchy sequence with respect to the total variation norm \((\cdot)\). Give a direct proof (not using the Nikodým Convergence Theorem) that \((\nu_n)\) converges setwise to a complex measure.

XIII.7.4.2. (a) Prove the converse of XIII.7.2.8. If \((B_n)\) is a decreasing sequence with \(\cap_{n=1}^{\infty} B_n = \emptyset\) but \(|\nu_n(B_n)| > \delta\) for all \(n\), for some \(\delta\), choose \(n_1\) with \(\nu_1(B_{n_1}) < \delta/2\), and for each \(k\) choose \(n_{k+1}\) with \(\nu_{n_k}(B_{n_{k+1}}) < \delta/2\). If \(A_{n_k} = B_{n_k} \setminus B_{n_{k+1}}\), the \(A_{n_k}\) are pairwise disjoint and \(|\nu_{n_k}(A_{n_k})| > \delta/2\) for all \(k\).

(b) Use (a) and (b) to give a direct proof of XIII.7.3.3. (cf. [Swa94, p. 68]).

XIII.7.4.3. [Doo94, III.10] Give a direct proof that if \((\mu_n)\) is a sequence of finite measures on a measurable space \((X, \mathcal{A})\), converging setwise to a function \(\mu : \mathcal{A} \to [0, \infty)\), then \(\mu\) is a measure (i.e. countably additive), as follows:

(a) Suppose \((B_n)\) is a decreasing sequence in \(\mathcal{A}\) with \(\cap_{n=1}^{\infty} B_n = \emptyset\) but

\[
\lim_{n \to \infty} \mu(B_n) = \epsilon > 0
\]

(why does the limit exist?) Let \(n_1 = m_1 = 1\) and inductively choose \(n_{k+1} > n_k\) and \(m_{k+1} > m_k\) with \(\mu_{n_{k+1}}(B_{m_k}) > \frac{\epsilon}{8}\) and \(\mu_{n_{k+1}}(B_{m_{k+1}}) < \frac{\epsilon}{8}\).

1479
(b) Let $A_k = B_{m_k} \setminus B_{m_{k+1}}$ for each $k$. Show that if $j > k$ is odd, then

$$
\mu_{n_j}(\cup \{A_n : n \text{ even}, n \geq k\}) > \frac{3\epsilon}{4}
$$

and if $j > k$ is even, then

$$
\mu_{n_j}(\cup \{A_n : n \text{ odd}, n \geq k\}) > \frac{3\epsilon}{4}.
$$

(c) Conclude that $\mu(B_{m_k}) \geq \frac{3\epsilon}{2}$ for all $k$, a contradiction.
Chapter XIV

Integration

In this chapter, we will explore the four general approaches to integration: the Riemann integral and the "modern" approaches of Lebesgue, Daniell, and Denjoy-Perron-Henstock-Kurzweil. The idea is to associate to a real- (or complex)-valued function \( f \) defined on a set \( X \), and a subset \( A \) of \( X \), a number \( \mathcal{I}_A(f) \), usually denoted by some version of the notation \( \int_A f \), reflecting the "contribution" of the function \( f \) on the set \( A \).

XIV.1. Introduction

Integration is a basic component of real analysis. The goal of integration is to systematically assign to a real-valued function \( f \) on a set \( X \), and a subset \( A \) of \( X \), a number \( \mathcal{I}_A(f) \) called the integral of \( f \) over \( A \) (the notation \( \int_A f \), or some variation, is usually used; the symbol \( \int \) is called the integral sign), which is usually thought of as a "normalized sum" of the values of \( f \) on the points of the subset \( A \). This number has various interpretations in applied situations as total distance (regarding \( f \) as giving instantaneous velocity), total mass (regarding \( f \) as giving density), etc. Integration in analysis is thus almost always (and always in this book) a generalization or analog of what in elementary calculus is called the "definite integral." (We will also have occasion to consider analogs of the "indefinite integral" of calculus, but these will have different names and will not be called integrals.)

XIV.1.1. Origin of Integration Theories

XIV.1.1.1. All modern theories of integration are in a sense outgrowths of the so-called Riemann integral (actually developed first by Cauchy for continuous functions and generalized by Riemann to also handle moderately discontinuous functions), the type of integration treated in elementary calculus. The Riemann integral was the first reasonably satisfactory theory of integration, and is still by far the theory most widely used in applications.

XIV.1.1.2. For theoretical purposes, Riemann integration has two serious inadequacies:

1. The class of Riemann integrable functions is too small. For a function to be Riemann integrable on an interval, it must be bounded and "almost continuous" (see () for a precise statement). It is particularly useful to be able to integrate certain unbounded functions. This is reasonably, but not entirely, satisfactorily
achieved by using various notions of improper Riemann integrals and principal values. It is also important to be able to integrate over unbounded intervals, which can be done through other types of improper integrals, although not in an entirely satisfactory way. At least for theoretical purposes, it is also important to be able to integrate functions which are too discontinuous to be Riemann integrable, and to be able to integrate functions over subsets of the real numbers more general than intervals or finite unions of intervals (although the practical value of this type of integration is sometimes somewhat overemphasized as an inadequacy of Riemann integration).

2. It is crucial in many settings to have limit and approximation theorems, i.e. to know that under reasonable conditions, if \( f_n \to f \) in some sense, we have \( \mathcal{I}_A(f_n) \to \mathcal{I}_A(f) \). It is quite difficult to establish good theorems of this sort for Riemann integration which are applicable even if we know that the limit function is Riemann integrable (which is often not the case due to the limitations described above). It turns out that to get reasonable convergence theorems, more general integration theories are needed.

XIV.1.1.3. Another limitation of Riemann integration is that it only (directly) works on \( \mathbb{R} \), or on \( \mathbb{R}^n \). It is very important to have integration theories which work on more general spaces, and it turns out that modern approaches give this flexibility with essentially no additional work.

XIV.1.1.4. There are three principal modern approaches to integration, each with many variations. We first discuss the basics of Riemann integration in a form allowing the various generalizations to be made naturally.

XIV.1.1.5. The Riemann integral has gotten somewhat of a bad rap: it is fashionable in some circles to downplay or even ignore Riemann integration in favor of more modern integration theories. There are certainly inadequacies in Riemann integration which must be addressed, and there will be no suggestion here that modern integration theories are not important. However, Riemann integration is far from obsolete! Here are some senses in which Riemann integration (and even pre-Riemann Cauchy integration) is still central:

(i) Although there are versions of the Fundamental Theorem of Calculus for more general types of integration, it is still really a theorem about Riemann (or Cauchy) integration, and most applications are to computing Riemann integrals.

(ii) All numerical integration is based on the Riemann integral.

(iii) Integrals in vector calculus, complex analysis, and differential geometry (e.g. line and surface integrals, contour integrals, integrals of differential forms) are Riemann integrals.

(iv) Many, if not most, integrals arising in physics and engineering are Riemann integrals.

(v) It is often pointed out that one rarely if ever actually computes a Lebesgue (or other modern) integral; the integral is approximated by a Riemann integral which is computed either by the Fundamental Theorem of Calculus or by numerical integration techniques.

Thus it is worthwhile to give a full treatment of the Riemann integral as well as more modern theories.
XIV.1.2. The Cauchy Integral for Continuous Functions

XIV.1.2.1. Actually, we will begin with a sort of “pre-Riemann integral,” due to Cauchy, in which there seems to be only one natural approach, which is fairly simple: the case of integrating a continuous function on a closed bounded interval. All theories of integration essentially reduce to the same thing in this case. Actually this case, although quite special, is by far the most important one in a practical sense – most integrals encountered in practice are of this form, or slight variations of it which can be treated on an ad hoc basis. In fact, although the integral in this form was not given a rigorous approach until Cauchy’s work, it had been used by mathematicians dating back at least to Archimedes in ancient Greece; the fundamental and brilliant insight of both Newton and Leibniz which got the whole subject of calculus off the ground was the inverse relation between integration and differentiation.

XIV.1.2.2. The integral is intended to denote the “area” of the region under the graph and above the x-axis, between the lines \( x = a \) and \( x = b \) (Figure XIV.1). To obtain an integral with good properties, the parts of the region below the axis, where the function is negative, must be interpreted as “negative area”; such regions are shown in tan. The integral thus denotes a “signed area.”

\[
\begin{align*}
&y = f(x) \\
&a & b
\end{align*}
\]

Figure XIV.1: The Integral of a Continuous Function

XIV.1.2.3. Suppose \( f \) is a continuous real-valued function on a closed bounded interval \( J = [a, b] \). Write \( \ell(J) = b - a \) for the length of \( J \). We want to define the integral \( \mathcal{L}_J(f) \) of \( f \) over \( J \). There are three fundamental properties the integral ought to have:
(i) If \( f \leq g \) on \( J \), then \( \mathcal{I}_J(f) \leq \mathcal{I}_J(g) \).

(ii) If \( f \) takes the constant value \( c \) on \( J \), then \( \mathcal{I}_J(f) = c \cdot \ell(J) = c(b-a) \).

(iii) If \( J \) is partitioned into nonoverlapping intervals \( J_1, \ldots, J_n \) (“nonoverlapping” means the subintervals only have endpoints in common), then

\[
\mathcal{I}_J(f) = \sum_{k=1}^{n} \mathcal{I}_{J_k}(f) .
\]

These three properties are enough to completely characterize the integral in this case, and to suggest an explicit method of defining and calculating its value.

**XIV.1.2.4.** The continuity of \( f \) means that if \( P = \{J_1, \ldots, J_n\} \) is a sufficiently fine partition of \( J \) into nonoverlapping subintervals, then \( f \) is almost constant on each \( J_k \); thus, if \( t_k \) is any number (tag) in \( J_k \), we must have

\[
\mathcal{I}_J(f) \approx \sum_{k=1}^{n} f(t_k)\ell(J_k)
\]

(such a sum is called a *Riemann sum* for \( \mathcal{I}_J(f) \), although they were introduced by Cauchy, and Riemann did not even make extensive use of them) with the approximation becoming better and better as the partition becomes finer no matter how the \( t_k \) are chosen (within \( J_k \)). See Figure XIV.2.

**XIV.1.2.5.** This can be described more carefully in a slightly different way (this was in effect Riemann’s general approach, although he phrased things somewhat differently; this formulation was explicitly introduced by Darboux after Riemann’s death): if \( m_k \) and \( M_k \) are the infimum and supremum of \( f \) on \( J_k \) (in this case actually the minimum and maximum on \( J_k \) respectively), and \( \omega(f, J_k) = M_k - m_k \) is the oscillation of \( f \) over \( J_k \), then we define the upper and lower estimates \( \overline{\mathcal{I}}(f, P) \) and \( \underline{\mathcal{I}}(f, P) \) by

\[
\overline{\mathcal{I}}(f, P) = \sum_{k=1}^{n} M_k\ell(J_k) \quad \text{and} \quad \underline{\mathcal{I}}(f, P) = \sum_{k=1}^{n} m_k\ell(J_k)
\]

and we have that

\[
\underline{\mathcal{I}}(f, P) \leq \mathcal{I}_J(f) \leq \overline{\mathcal{I}}(f, P)
\]

and

\[
\overline{\mathcal{I}}(f, P) - \underline{\mathcal{I}}(f, P) = \sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) .
\]

(See Figures XIV.3 and XIV.4.) As the partition \( P \) becomes finer, we have that the \( \underline{\mathcal{I}}(f, P) \) increase, the \( \overline{\mathcal{I}}(f, P) \) decrease, and the \( \omega(f, J_k) \) approach 0, so the upper and lower estimates approach the same limiting value (there are of course some technicalities to check here), which is the only reasonable value for \( \mathcal{I}_J(f) \). We also have

\[
\underline{\mathcal{I}}(f, P) \leq \sum_{k=1}^{n} f(x_k)\ell(J_k) \leq \overline{\mathcal{I}}(f, P)
\]

1484
For any Riemann sum, so we have

\[ I_J(f) - \frac{1}{n} \sum_{k=1}^{n} f(x_k) \ell(J_k) \leq T(f, P) - I(f, P) = \sum_{k=1}^{n} \omega(f, J_k) \ell(J_k) \]

for any Riemann sum, and thus the Riemann sums converge to \( I_J(f) \) as the partitions become arbitrarily fine (in the sense that the mesh of the partition, defined as the maximum of the lengths of the \( J_k \), goes to 0).

**XIV.1.2.6.** Before the rigour introduced in the nineteenth century, and informally still to this day, instead of an actual (finite) partition of \( J \) one would think of “partitioning” the interval \( J \) into an infinite number of subintervals of “infinitesimal” length \( dx \) (Figure XIV.5). Then on the infinitesimal interval containing the number \( x \), the function value would differ only infinitesimally from the number \( f(x) \), so the integral over this infinitesimal subinterval would be \( f(x) \cdot dx \) (up to a “second-order infinitesimal”). The total integral would then be the “sum” of these. This approach was taken by Leibniz, who developed the standard notation used for the integral even today: he denoted this “sum” \( S \), and wrote

\[ I_J(f) = S_a^b f(x) \, dx \]

He actually used a “long \( S \)” and almost immediately elongated it further to the familiar integral sign \( \int \). (Summation notation using \( \sum \) was introduced by Euler, and did not exist in Leibniz’s time.)

1485
XIV.1.3. The Riemann Integral

XIV.1.3.1. Cauchy’s approach works in somewhat greater generality, as developed by Riemann. In order to have the upper and lower estimates converge together, we do not actually have to have the oscillations $\omega(f, J_k)$ approach zero as the partitions become fine; we need only that the sum (“weighted local oscillation”)

$$\sum_{k=1}^{n} \omega(f, J_k) \ell(J_k)$$

becomes small, which only requires that the oscillation $\omega(f, J_k)$ be small for “most” $k$. See Figures XIV.6 and XIV.7. If $f$ has isolated discontinuities, the weighted local oscillation is still small for fine partitions (Figure XIV.8).

XIV.1.3.2. Thus, if $f$ is bounded, and continuous except at finitely many points (or, more generally, except on a set of Lebesgue measure 0), the Cauchy approach works perfectly well, either via the weighted local oscillation idea or via Riemann sums. (If $f$ is unbounded, the weighted local oscillation approach does not make sense, and it is also easily seen that the Riemann sum approach cannot work either, at least the way we have done it using the mesh of the partitions.) The functions for which the approach works are called Riemann-integrable functions on $J$; it turns out that they are precisely the bounded functions which are continuous “almost everywhere” (except on a set of Lebesgue measure 0).
XIV.1.4. The Lebesgue Approach

XIV.1.4.1. The first and most widely used modern approach began with Lebesgue’s work in the early twentieth century. Lebesgue’s idea was to consider more general partitions of the interval $J$ on which the integral is to be defined, where the subsets $A_k$ are not necessarily subintervals but can be more complicated, but where as in the Cauchy approach the function is almost constant on each $A_k$. The integral $\int_J (f)_{P}$ is then approximated by a sum over the sets in the partition, where the term corresponding to $A_k$ is the almost constant value of $f$ on $A_k$ times the “size” of $A_k$.

XIV.1.4.2. It is easy to find such sets $A_k$ for any bounded function. (For motivational purposes, we will consider only bounded functions, although the actual development of the theory outlined below works equally well for unbounded functions and even extended real-valued functions.) In fact, it is not important that the domain of the function be an interval or even a set of real numbers; we can consider a bounded function $f$ from any set $X$ to $\mathbb{R}$ (bounded just means the range is a bounded subset of $\mathbb{R}$). Suppose $|f(x)| \leq M$ for all $x \in X$. Then, for any natural number $n$, we set

$$A_k = \left\{ x \in X : \frac{k-1}{n} < f(x) \leq \frac{k}{n} \right\}$$

for integers $k$ with $|k| \leq nM$. This is a finite partition of $X$ into pairwise disjoint subsets, and $\omega(f, A_k) \leq \frac{1}{n}$.
XIV.1.4.3. We need one more ingredient, though: a “size” for each $A_k$. The reasonable way to do this is to require the $A_k$ to be measurable sets for a measure $\mu$ on $X$; then $\mu(A_k)$ can be taken to be the “size” of $A_k$, and we get an approximation

$$I_X(f, \mu) \approx \sum_{k} \frac{k}{n} \mu(A_k)$$

(the notation reflects the fact that the integral is with respect to the measure $\mu$) which should become exact in the limit as $n \to \infty$. The restriction on $f$ is then that the sets $A_k$ must be $\mu$-measurable for any choices of $n$ and $k$; such an $f$ is called a $\mu$-measurable function. In most applications, including Lebesgue measure on $\mathbb{R}$, $\mu$-measurability is a very mild restriction on $f$ and is satisfied by any function which can be explicitly described. (There is an additional technicality: the sum may not be well defined if $\mu(A_k)$ is infinite for both positive and negative $k$’s. This technicality forces a slightly different approach to the actual integral.)

XIV.1.4.4. So in this approach, one begins with a measure space (LEBESGUE at first only considered Lebesgue measure on $\mathbb{R}$, but the general theory is nearly identical). As mentioned, for technical reasons, we will not proceed exactly as above, but the motivational description is an accurate reflection of the essence of the process.

XIV.1.4.5. If $\mu$ is a measure on a set $X$, we build up the integral corresponding to $\mu$ in stages. We first consider indicator functions (often called characteristic functions, but this term also has other uses in...

Figure XIV.5: Partition into Rectangles of “Infinitesimal” Width
analysis). If $A \subseteq X$, let $\chi_A$ denote the function from $X$ to $\mathbb{R}$ defined by

\begin{align*}
\chi_A(x) &= \begin{cases} 
1 & \text{if } x \in A \\
0 & \text{if } x \notin A
\end{cases}.
\end{align*}

This $\chi_A$ is called the \textit{indicator function} of the set $A$. It is reasonable to define the integral of $\chi_A$ to be the measure $\mu(A)$ of the set $A$ (provided, of course, that $A$ is a $\mu$-measurable set, or equivalently that $\chi_A$ is a $\mu$-measurable function).

\textbf{XIV.1.4.6.} Next, we consider \textit{simple functions}, finite nonnegative linear combinations of indicator functions of $\mu$-measurable sets. Such a function has a representation

\begin{equation}
 s = \sum_{k=1}^{n} a_k \chi_{A_k}
\end{equation}

where the $A_k$ are $\mu$-measurable sets and the $a_k$ are nonnegative real numbers. These are precisely the nonnegative $\mu$-measurable functions taking only finitely many values. It then makes sense to define the integral by

\begin{equation}
\mathcal{I}_X(s, \mu) = \sum_{k=1}^{n} a_k \mu(A_k)
\end{equation}

and indeed this is the only possible definition to obtain a linear integral. (There is a technicallity here: it must be shown that this number is well defined by $s$, since the representation $s = \sum a_k \chi_{A_k}$ is not unique.)
XIV.1.4.7. If $f$ is a nonnegative $\mu$-measurable function, we then define

$$\mathcal{I}_X(f, \mu) = \sup \{ \mathcal{I}_X(s, \mu) : s \text{ simple}, 0 \leq s \leq f \} .$$

This definition even makes sense for extended real-valued functions. Even if $f$ is real-valued, $\mathcal{I}_X(f, \mu)$ can equal $+\infty$. We say a nonnegative function $f$ is $\mu$-integrable if $f$ is $\mu$-measurable and $\mathcal{I}_X(f, \mu) < \infty$.

XIV.1.4.8. Finally, if $f$ is a real-valued $\mu$-measurable function, and there are nonnegative $\mu$-integrable functions $g$ and $h$ with $f = g - h$, we say $f$ is $\mu$-integrable and set

$$\mathcal{I}_X(f, \mu) = \mathcal{I}_X(g, \mu) - \mathcal{I}_X(h, \mu)$$

(it must be shown that this is well defined by $f$ independent of the choices of $g$ and $h$).

XIV.1.4.9. We can extend the notion of integration over the whole measure space $X$ to integration over any $\mu$-measurable set $A$ by setting

$$\mathcal{I}_A(f, \mu) = \mathcal{I}_X(f \chi_A, \mu)$$

where $f \chi_A$ is the function which is equal to $f$ on $A$ and is identically zero outside $A$; this definition even makes sense if $f$ is only defined on the subset $A$. 

1490
XIV.1.4.10. It turns out that this integral $I_X$ has all the nice properties one would expect, and almost all one might hope for, in an integral. In the case where $\mu$ is Lebesgue measure $\lambda$ on $\mathbb{R}$, this integral extends the Riemann integral in the sense that any Riemann-integrable function $f$ on a closed bounded interval $J$ is Lebesgue-integrable ($\lambda$-integrable) on $J$ and the the Riemann integral $I_J(f)$ of $f$ over $J$ agrees with the Lebesgue integral $L_J(f, \lambda)$ of $f$ over $J$. But many more functions are Lebesgue integrable than are Riemann-integrable, including many unbounded functions.

XIV.1.4.11. The Lebesgue approach to integration on a general measure space is also very important. For example, if $(\Omega, \mathcal{E}, P)$ is a probability space, a $P$-measurable real-valued function $X$ on $\Omega$ is called a random variable, and if $X$ is nonnegative or $P$-integrable, its integral is called the expected value of the random variable $X$.

XIV.1.4.12. There are a couple of drawbacks to the Lebesgue approach, however. In order to define the Lebesgue integral on $\mathbb{R}$, Lebesgue measure must first be constructed, which is a moderately complicated task. The Lebesgue integral on a general measure space is somewhat simpler, but preliminaries are not completely avoided: besides the simple basics of the theory of measure spaces from ( ), one needs to know about measurable functions, where the details ( ) are somewhat more challenging.

XIV.1.4.13. The other drawback of Lebesgue integration from a certain point of view is that it is an “absolute” theory of integration: if $f$ is measurable, then $f$ is integrable if and only if $|f|$ is integrable.
Roughly, this means that both the positive and negative parts of $f$ are not too large. There are situations where the improper Riemann integral of a function $f$ taking both positive and negative values is defined, yet $f$ is not Lebesgue integrable. On the other hand, the “absolute” property is what makes Lebesgue integration well-behaved; it is a sort of analog of absolute convergence of infinite series, which forces all sorts of nice behavior which does not carry over to conditionally convergent series. (The connection between integration and infinite series is much more than just an analog; the theory of infinite series can be interpreted as a special case of the theory of measure and integration – see ( ).)

XIV.1.5. The Daniell Approach

XIV.1.5.1. Either measure or integration can be regarded as the primitive notion, and the other one can be derived from it. We have seen how integration can be derived from measure via the Lebesgue approach. In Euclidean space, there is a more direct approach simply consisting of increasing dimension by one: if $f$ is a nonnegative real-valued function on a subset $E$ of $\mathbb{R}^n$, the integral $\int_E f \, d\lambda^{(n)}$ is simply the $(n+1)$-dimensional Lebesgue measure of the “subgraph”

$$\{(x_1, \ldots, x_n, y) : (x_1, \ldots, x_n) \in E, 0 \leq y \leq f(x_1, \ldots, x_n)\} \subseteq \mathbb{R}^{n+1}$$

(in particular, the integral of a nonnegative measurable function over a subset of $\mathbb{R}$ is just the measure of a corresponding subset of the plane). Conversely, measure can be recovered from integration: the measure of a subset $E$ of $\mathbb{R}^n$ (or of a general measure space) is just the integral of the indicator function $\chi_E$ over $\mathbb{R}^n$.

XIV.1.5.2. Taking integration as the primitive concept and defining measure from it is the approach taken by Daniell. In this approach, we begin with a sufficiently large supply of real-valued functions on a set $X$ for which there is an “integral” defined having reasonable linearity, positivity, and continuity properties. The actual mechanism by which the integral is defined is unimportant; it can even just be an abstract real-valued functional defined on the set of functions. If necessary, the integral is then extended in a systematic way to a larger class of real-valued functions on $X$ including indicator functions of enough subsets of $X$. A measure $\mu$ on $X$ can then be defined by setting $\mu(E)$ for a subset $E$ of $X$ to be the integral of $\chi_E$. A good motivating example (which was Daniell’s own starting point) is to take $X$ to be an interval $[a, b]$ in $\mathbb{R}$, the functions to be the real-valued continuous functions on $[a, b]$, and the integral the Cauchy integral. The measure obtained by the Daniell process is then just Lebesgue measure on $[a, b]$. This turns out to actually be a rather painless way to define Lebesgue measure.

XIV.1.5.3. An important generalization of the motivating example is the Riesz Representation Theorem: if $X$ is a compact Hausdorff space and $\phi$ is a positive linear functional on the set $C(X)$ of real-valued (or complex-valued) continuous functions on $X$, i.e. an assignment of a number $\phi(f)$ to each continuous function $f$ with the properties that

1. $\phi(f + g) = \phi(f) + \phi(g)$ for all $f, g \in C(X)$.
2. $\phi(\alpha f) = \alpha \phi(f)$ for every $f \in C(X)$ and every scalar $\alpha$.
3. $\phi(f) \geq 0$ if $f$ takes only nonnegative real values.

then there is a finite measure $\mu$ on the $\sigma$-algebra of subsets of $X$ generated by the open sets (the Borel sets) such that $\phi(f) = \int_X f \, d\mu$ for all $f$. The measure $\mu$ is constructed by the Daniell process of extending the integral (linear functional) to a larger class of functions including indicator functions of Borel sets.
XIV.1.6.  Direct Generalizations of Riemann Integration

XIV.1.6.1.  A third approach to generalizing Riemann integration is to directly modify the definition of the Riemann integral to obtain a more-inclusive integral on $\mathbb{R}$. This approach was notably used by Denjoy and Perron, and culminated relatively recently (1957) in the so-called Henstock-Kurzweil (H-K) integral, which agrees with the integrals of Denjoy and Perron but has a much better definition.

XIV.1.6.2.  The H-K integral is easier to define than the Lebesgue integral (Lebesgue measure is not needed), and almost as easy as the Riemann integral; and it remedies both shortcomings of the Riemann integral. In fact, it turns out to be almost the same as the Lebesgue integral, and even slightly more general: a nonnegative function is H-K integrable if and only if it is Lebesgue integrable, but a function taking both positive and negative values can be H-K integrable but not Lebesgue integrable if its positive and negative parts are too big but “cancel out” in a suitable way.

XIV.1.6.3.  The H-K integral is defined using the Riemann sum approach. The main difference from Riemann integration is a much more delicate notion of the “fineness” of a partition into subintervals than just its mesh. We will not give a precise description here (see ()), but the fineness of a partition depends not only on the intervals $J_k$ but also on which tags $t_k \in J_k$ are chosen. It is then a somewhat subtle matter to show that arbitrarily fine partitions exist. Once this has been done, the rest of the development formally follows the theory of Riemann integration very closely. But many more functions turn out to be H-K integrable.

XIV.1.6.4.  The H-K integral is nice to work with in many ways. For example, integrals which would naturally be called “improper H-K integrals” turn out to be honest H-K integrals. One of the nicest properties of H-K integration is that the following version of the Fundamental Theorem of Calculus, which fails spectacularly for Riemann integration and even fails for Lebesgue integration, holds for H-K integration:

XIV.1.6.5.  **Theorem.** Let $f$ be a differentiable function on a closed bounded interval $[a, b]$. Then $f'$ is H-K integrable on $[a, b]$ (and hence on any subinterval), and for every $x \in [a, b]$ we have

$$f(x) = f(a) + \int_a^x f'(t) \, dt$$

where the integral is the H-K integral of $f'$ over $[a, x]$.

XIV.1.6.6.  On the other hand, the H-K integral is strictly an extension of Riemann integration on $\mathbb{R}$, and does not generalize the way Lebesgue integration does to other measure spaces. (There is a version of H-K integration on $\mathbb{R}^n$, which is not nearly as satisfactory as on $\mathbb{R}$, but nothing more general.)
XIV.1.7. Oriented and Unoriented Integration

XIV.1.7.1. There are two general kinds of integration: oriented integration and unoriented integration (they are not very different, differing only in sign, but the sign is often crucial in applications). The two types both appear in Riemann integration: if \( I \) is a closed bounded interval \([a, b]\), with \( a < b \), then the integral

\[
\int_a^b f(x) \, dx
\]

of a function \( f \) from \( I \) to \( \mathbb{R} \) can be regarded either as an integral over \( I \) ("unoriented") or as an integral from \( a \) to \( b \) ("oriented"). The expression

\[
\int_b^a f(x) \, dx
\]

appears to be rather artificial and intrinsically meaningless, and the definition

\[
\int_b^a f(x) \, dx = - \int_a^b f(x) \, dx
\]

is seemingly arbitrary, although it is useful in a number of natural contexts, for example making the formula

\[
\int_a^c f(x) \, dx = \int_a^b f(x) \, dx + \int_b^c f(x) \, dx
\]

valid for any \( a, b, c \) no matter what their order (so long as they are all contained in some interval on which \( f \) is integrable). Another important use of this convention is that if \( f \) is continuous on an interval \( I \) and \( x_0 \in I \), not necessarily the left endpoint, then the formula

\[
F(x) = \int_{x_0}^x f(t) \, dt
\]

defines the unique antiderivative of \( f \) on \( I \) satisfying \( F(x_0) = 0 \).

XIV.1.7.2. Integrating from right to left in \( \mathbb{R} \) seems unnatural since there is a natural orientation of \( \mathbb{R} \). However, in Euclidean space there is no natural orientation; there are two equally natural possibilities (choice of a coordinate system specifies one of the two orientations; cf. VI.3.1.1., []). Complicated subsets of \( \mathbb{R}^n \) do not have any obvious intrinsic notion of orientation. Thus in the Lebesgue approach to integration it is natural to use unoriented integration. And when the generalization to abstract measure spaces is made, there is no notion of orientation available at all, so unoriented integration is the only possibility. Thus the theory of integration in measure theory is (almost) exclusively unoriented integration.

XIV.1.7.3. On the other hand, in vector calculus and analysis such integrals as line and surface integrals are important, and in Complex Analysis contour integrals are basic; such integrals incorporate a natural orientation, and the orientation is crucial in many results. In fact, any result relating integration and differentiation along the lines of the Fundamental Theorem of Calculus, e.g. the various versions of Stokes' Theorem, necessarily involves oriented integration, since differentiation itself requires orientation (this is obscured for functions from \( \mathbb{R} \) to \( \mathbb{R} \) since \( \mathbb{R} \) has a natural orientation, but orientations must be chosen to define the derivative of a real-valued function on a one-dimensional Euclidean space or a function between two one-dimensional Euclidean spaces). It turns out that many deep results in geometry and topology involve oriented integration.
XIV.1.7.4. As a general (but not universal) rule, oriented integration is usually done only for smooth objects and is primarily used in geometry, while unoriented integration is done for not necessarily smooth or even continuous objects and is primarily used in analysis. That these statements are not categorically true is partly due to the fact that the boundary between analysis and geometry is quite blurred. We will examine important examples of both types of integration.
XIV.2. The Riemann Integral

We first examine the integral commonly called the Riemann integral. This integral was actually developed informally by Newton and Leibniz (based in part on ideas going back to ancient Greece), and formally developed by Cauchy for continuous functions and expanded and put on a “modern” foundation by Riemann, with important contributions by others, notably Dirichlet and Darboux. See [?] for a detailed historical account of the development of this integral.

It is a curious fact that there is no consensus among mathematicians at the present time about the proper definition of the Riemann integral and of Riemann-integrable functions. Several rather different-looking (but equivalent) definitions can be found in standard references and texts, with other variations given as theorems. While it is pretty easy to see that any two of these definitions give the same value for the integral of any function satisfying both definitions, it is not entirely obvious that they apply to exactly the same functions (i.e. give the same class of integrable functions). We will discuss the various definitions and show that they are actually all equivalent. A normal pedagogical approach would be to pick one definition as the definition, and prove that the others are equivalent characterizations; but the lack of a standard definition makes the several-definition approach more reasonable in this case.

The multitude of definitions can be somewhat of a conceptual obstacle to focusing on the essential nature of the Riemann integral. However, the situation can, and should, be viewed positively: the fact that Riemann-integrability can be viewed in a number of ways is of theoretical importance and an indication that it is a natural concept. And the large supply of equivalent definitions is very useful in practice: in any situation, we are free to use whichever definition is most convenient to work with at the moment.

The various definitions are all precise versions of the following informal definition:

**Definition. (Informal)** Let \( f \) be a real-valued function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) if the sums
\[
\sum_{k=1}^{n} c_k \ell(J_k)
\]
where \( \{J_1, \ldots, J_n\} \) is a partition of \( J \) into subintervals and \( c_k \) is an approximation to the value of \( f \) on \( J_k \), converge to a limiting value \( L \) as the partition becomes sufficiently fine. The limiting value \( L \) is called the \textit{(Riemann) integral of} \( f \) \textit{over} \( J \), denoted
\[
\int_{a}^{b} f(x) \, dx.
\]

There are, of course, several aspects of this “definition” which must be made precise to have a real definition. It is not entirely clear how to make some parts of the definition precise, particularly the part about the partition becoming “sufficiently fine.” (With the right interpretation of this phrase, the statement actually becomes a definition of the Henstock-Kurzweil integral.)

The integral can be interpreted as the “area” of the region under the graph and above the \( x \)-axis, between the lines \( x = a \) and \( x = b \). Parts of the region below the axis, where the function is negative, are interpreted as “negative area”; such regions are shown in tan. The integral thus denotes a “signed area.”

XIV.2.1. Partitions, Tags, Riemann and Darboux Sums

In this subsection we make a number of definitions and set up the notation we will use for Riemann integration.
**Partitions**

**XIV.2.1.1. Definition.** Let \([a, b]\) be a closed bounded interval in \(\mathbb{R}\). A *partition* of \([a, b]\) is a division of \([a, b]\) into a finite number of subintervals \(\{[x_{k-1}, x_k] : 1 \leq k \leq n\}\) with only endpoints in common. A partition \(\mathcal{P}\) is usually described by its set of endpoints \(\{x_0, x_1, \ldots, x_n\}\), where \(a = x_0 < x_1 < \cdots < x_n = b\). We can also write \(\mathcal{P} = \{J_1, \ldots, J_n\}\), where \(J_k = [x_{k-1}, x_k]\) is the \(k\)’th subinterval. We sometimes write \(\Delta x_k\) for \(\ell(J_k) = x_k - x_{k-1}\), the length of the \(k\)’th subinterval of the partition.

The *norm* or *mesh* of a partition \(\mathcal{P} = \{x_0, x_1, \ldots, x_n\}\) is

\[
\|
\mathcal{P}\| = \max\{x_k - x_{k-1} : 1 \leq k \leq n\}.
\]

If \(\mathcal{P} = \{x_0, \ldots, x_n\}\) and \(\mathcal{Q} = \{x_0', \ldots, x_m'\}\) are partitions of \([a, b]\), we say that \(\mathcal{Q}\) is *finer* than \(\mathcal{P}\), or a *refinement* of \(\mathcal{P}\), if \(\mathcal{P} \subseteq \mathcal{Q}\), i.e. if each \(x_k\) is one of the \(x_j'\), or equivalently if each interval in the partition \(\mathcal{Q}\) is a subinterval of one of the intervals in \(\mathcal{P}\).

A *tag* in an interval \([c, d]\) is a specified number in \([c, d]\). A *tagged partition* is a partition \(\{x_0, x_1, \ldots, x_n\}\) with a specified tag \(t_i\) in \([x_{i-1}, x_i]\) for each \(i\). A tagged partition is usually denoted by

\[
\mathcal{P} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}.
\]

The norm or mesh of a partition is a measure of how fine the partition is. There is a more sophisticated and flexible notion of fineness of a partition based on gauges (\().

1497
Note that if $P$ and $Q$ are two partitions of $[a; b]$, neither is a refinement of the other in general. But:

**XIV.2.1.2. Proposition.** Two partitions $P$ and $Q$ of an interval $[a; b]$ have a common refinement.

**Proof:** Just take $P \cup Q$, the partition obtained by using all the endpoints in both partitions. ☑

**Riemann Sums**

**XIV.2.1.3. Definition.** Let $f$ be a real-valued function on an interval $[a, b]$, and let

$$\hat{P} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}$$

be a tagged partition of $[a, b]$. The *Riemann sum* for $f$ with respect to $\hat{P}$ is

$$R(f, \hat{P}) = \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) = \sum_{k=1}^{n} f(x_k) \Delta x_k .$$

Note that Riemann sums are only defined for tagged partitions, and the numerical value depends on the choice of tags as well as the partition.

1498
Darboux Sums and Oscillation

Two closely related sums which do not depend on tags are the upper and lower Darboux sums:

**XIV.2.1.4. Definition.** Let $f$ be a bounded function on an interval $[a, b]$. If $\mathcal{P} = \{x_0, \ldots, x_n\}$ is a partition of $[a, b]$, for $1 \leq k \leq n$ set

$$M_k = \sup_{x_{k-1} \leq x \leq x_k} f(x) \quad \text{and} \quad m_k = \inf_{x_{k-1} \leq x \leq x_k} f(x).$$

Then the upper and lower Darboux sums for $f$ with respect to $\mathcal{P}$ are

$$\mathcal{S}(f, \mathcal{P}) = \sum_{k=1}^{n} M_k (x_k - x_{k-1}) = \sum_{k=1}^{n} M_k \Delta x_k$$

$$\underline{S}(f, \mathcal{P}) = \sum_{k=1}^{n} m_k (x_k - x_{k-1}) = \sum_{k=1}^{n} m_k \Delta x_k$$

(written $U(f, \mathcal{P})$ and $L(f, \mathcal{P})$ in many references).

The relation between Darboux and Riemann sums is:
XIV.2.1.5. **Proposition.** Let $f$ be a bounded function on an interval $[a, b]$, and let $P$ be a partition of $[a, b]$. Then

(i) If $\hat{P}$ is any tagging of $P$, then
\[
\underline{S}(f, P) \leq R(f, \hat{P}) \leq \overline{S}(f, P).
\]

(ii) For any $\epsilon > 0$ there are taggings $\hat{P}$ and $\bar{P}$ of $P$ such that
\[
R(f, \hat{P}) > \overline{S}(f, P) - \epsilon \quad \text{and} \quad R(f, \bar{P}) < \underline{S}(f, P) + \epsilon.
\]

**Proof:** Part (i) is obvious. For (ii), let $P = \{x_0, \ldots, x_n\}$. For each $k$, choose $t_k \in [x_{k-1}, x_k]$ with
\[
f(t_k) > M_k - \frac{\epsilon}{n \Delta x_k}
\]
and set $\hat{P} = \{x_0, \ldots, x_k; t_1, \ldots, t_k\}$. The construction of $\hat{P}$ is similar.

As the partition $P$ becomes finer, the upper Darboux sums decrease and the lower Darboux sums decrease.
XIV.2.1.6.  **Proposition.** Let \( f \) be a bounded function on an interval \([a, b]\), and \( P \) and \( Q \) partitions of \([a, b]\). If \( Q \) is finer than \( P \), then

\[
\mathcal{L}(f, P) \leq \mathcal{L}(f, Q) \leq \mathcal{U}(f, Q) \leq \mathcal{U}(f, P).
\]

There is no such simple relation for Riemann sums.

**Interior Darboux Sums**

For technical purposes, there is a useful variation on the notion of a Darboux sum:

XIV.2.1.7.  **Definition.** Let \( f \) be a bounded function on an interval \([a, b]\). If \( P = \{x_0, \ldots, x_n\} \) is a partition of \([a, b]\), for \( 1 \leq k \leq n \) set

\[
M^o_k = \sup_{x_{k-1} < x < x_k} f(x) \quad \text{and} \quad m^o_k = \inf_{x_{k-1} < x < x_k} f(x).
\]

Then the **upper and lower interior Darboux sums** for \( f \) with respect to \( P \) are

\[
\mathcal{L}_o(f, P) = \sum_{k=1}^{n} M^o_k (x_k - x_{k-1}) = \sum_{k=1}^{n} M^o_k \Delta x_k
\]

1501
The difference is that $M_k^o$ and $m_k^o$ are the supremum and infimum of $f$ on the interior of the $k$'th subinterval.

XIV.2.1.8. For any $f$ and $P$, we obviously have $m_k \leq m_k^o \leq M_k^o \leq M_k$, so

$$\underline{S}(f, P) \leq \underline{S}_o(f, P) \leq \overline{S}_o(f, P) \leq \overline{S}(f, P).$$

If $f$ is continuous on $[a, b]$, then $M_k^o = M_k$ and $m_k^o = m_k$ for all $k$, so we have equality at the ends, but the interior Darboux sums for step functions often differ from the usual Darboux sums. We will obtain an approximate equivalence between the notions in (\).

**The Metaprinciple**

There is a simple but useful observation about the behavior of Riemann sums, Darboux sums, and interior Darboux sums:

XIV.2.1.9. **Proposition.** Let $f$ be a bounded function on an interval $[a, b]$, $P$ a partition of $[a, b]$, and $\_P$ any tagging of $P$. We have the following relations for any constant $c$:

\[
\begin{align*}
R(f + c, \_P) &= R(f, \_P) + c(b - a) \\
\overline{S}(f + c, P) &= \overline{S}(f, P) + c(b - a) \\
\underline{S}(f + c, P) &= \underline{S}(f, P) + c(b - a) \\
\overline{S}_o(f + c, P) &= \overline{S}_o(f, P) + c(b - a) \\
\underline{S}_o(f + c, P) &= \underline{S}_o(f, P) + c(b - a) \\
R(-f, \_P) &= -R(f, \_P) \\
\overline{S}(-f, P) &= -\overline{S}(f, P) \\
\underline{S}(-f, P) &= -\underline{S}(f, P) \\
\overline{S}_o(-f, P) &= -\overline{S}_o(f, P) \\
\underline{S}_o(-f, P) &= -\underline{S}_o(f, P).
\end{align*}
\]

XIV.2.1.10. Thus, in analyzing and comparing such sums, it suffices to consider nonnegative functions by adding a suitable constant; and results about upper Darboux sums have an exact analog for lower Darboux sums which do not have to be proved separately, but simply follow by taking negations.

**Relation between Darboux Sums and Interior Darboux Sums**

The next result is a good example of the metaprinciple in action.
**XIV.2.1.11.** Proposition. Let $f$ be a bounded function on an interval $[a, b]$, and $\mathcal{P}$ a partition of $[a, b]$. Then for every $\epsilon > 0$ there is a partition $\mathcal{Q}$ of $[a, b]$ such that

$$\mathcal{S}_o(f, \mathcal{P}) - \epsilon \leq \mathcal{S}(f, \mathcal{Q}) \leq \mathcal{S}(f, \mathcal{Q}) \leq \mathcal{S}_o(f, \mathcal{P}) + \epsilon .$$

**Proof:** By the Metaprinciple, we need show there is a $\mathcal{Q}$ giving the last inequality: then by negation, there is a $\mathcal{Q}'$ giving the first inequality, and a common refinement of $\mathcal{Q}$ and $\mathcal{Q}'$ gives both. Furthermore, in finding $\mathcal{Q}$ for the last inequality we may assume that $f$ is nonnegative.

Let $\mathcal{P} = \{x_0, \ldots, x_n\}$. We will obtain $\mathcal{Q}$ by shortening each interval in $\mathcal{P}$ and adding additional small intervals around $x_k$ for each $k$: given $\epsilon > 0$, choose $\delta > 0$ small enough that $\delta < \ell(J_k)/2$ for all $k$, and so that $\delta < \frac{\epsilon}{2M(n+1)}$, where $M$ is a bound for $f$ on $[a, b]$, and let

$$\mathcal{Q} = \{x_0, x_0 + \delta, x_1 - \delta, x_1 + \delta, \ldots, x_n - \delta, x_n\} .$$

Then for each $k$ we have

$$\sup_{x_{k-1} + \delta \leq x \leq x_k - \delta} f(x) \leq M_k^\circ$$

and so by adding up separately over the truncated original intervals and the new intervals we have

$$\mathcal{S}(f, \mathcal{Q}) \leq \sum_{k=1}^{n} M_k^\circ [\ell(J_k) - 2\delta] + \sum_{k=0}^{n} M(2\delta) \leq \mathcal{S}_o(f, \mathcal{P}) + \epsilon .$$

**Oscillation**

The oscillation of a bounded real-valued function on any set can be defined:

**XIV.2.1.12.** Definition. Let $f$ be a real-valued function on a set $X$, and $S \subseteq X$. If $f$ is bounded on $S$, the number

$$\omega(f, S) = \sup_S f - \inf_S f$$

is called the oscillation of $f$ on $S$.

**XIV.2.1.13.** More generally, if $f$ is a function from a set $X$ to a metric space $(Y, \rho)$, and $f$ is bounded on a subset $S$ of $X$, then we can define

$$\omega(f, S) = \sup\{\rho(f(s), f(t)) : s, t \in S\} .$$

If $X$ is a topological space and $x_0 \in X$, then $f$ is continuous at $x_0$ if and only if

$$\lim_{\mathcal{U}} \omega(f, \mathcal{U}) = 0$$

as $\mathcal{U}$ ranges over the directed set of neighborhoods of $x_0$. In particular, if $X \subseteq \mathbb{R}$, then $f : X \to \mathbb{R}$ is continuous at $x_0 \in X$ if and only if

$$\lim_{\epsilon \to 0} \omega(f, (x - \epsilon, x + \epsilon)) = 0 .$$
XIV.2.1.14. The connection with integration is that if $f$ is a bounded function on an interval $[a, b]$, and $\mathcal{P}$ is a partition of $[a, b]$, with $M_k$ and $m_k$ defined as in XIV.2.1.4., then

$$\omega(f, J_k) = M_k - m_k$$

and so we have

$$\overline{S}(f, \mathcal{P}) - \underline{S}(f, \mathcal{P}) = \sum_{k=1}^{n} \omega(f, J_k) \ell(J_k).$$

The last number is called the weighted local oscillation of $f$ with respect to $\mathcal{P}$.

![Figure XIV.14: The Weighted Local Oscillation for a Partition](image)

XIV.2.1.15. The weighted local oscillation does not increase, and typically decreases, as the partition becomes finer (Figure XIV.15).

XIV.2.2. Step Functions and their Integrals

XIV.2.3. Definitions of the Riemann Integral

We will give five definitions of Riemann integrability. All have slight rephrasings which are easily seen to be equivalent to the basic definition. There are other definitions which can be found in the literature, but the ones we give are the “extreme” ones and others can be seen to fit between our definitions.
The basic definitions are in terms of (1) Riemann sums, (2) upper and lower Darboux sums (or oscillation), or (3) approximation above and below by step functions. The other variation is (A) requiring existence of one partition forcing good approximation or (B) requiring good approximation for all sufficiently fine partitions.

**Definition 1**

**XIV.2.3.1.** Definition. [Definition 1A] Let $f$ be a function on a closed bounded interval $J = [a, b]$. Then $f$ is Riemann-integrable on $J$ in sense 1A, with Riemann integral $R \in \mathbb{R}$, if for every $\epsilon > 0$ there is a partition $P = \{x_0, \ldots, x_n\}$ of $J$ such that, for every choice of tags $t_1, \ldots, t_n$ for $P$, we have
\[
\left| R - \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) \right| < \epsilon .
\]

**XIV.2.3.2.** Definition. [Definition 1B] Let $f$ be a function on a closed bounded interval $J = [a, b]$. Then $f$ is Riemann-integrable on $J$ in sense 1B, with Riemann integral $R \in \mathbb{R}$, if for every $\epsilon > 0$ there is a $\delta > 0$ such that, for every tagged partition $P = \{x_0, \ldots, x_n; t_1, \ldots, t_n\}$ of $J$ with $\|P\| < \delta$, we have
\[
\left| R - \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) \right| < \epsilon .
\]

Figure XIV.15: The Weighted Local Oscillation for a Finer Partition
XIV.2.3.3. Definition. [Definition 2A] Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 2A if for every \( \epsilon > 0 \) there is a partition \( P \) of \( J \) such that
\[
\mathfrak{S}(f, P) - \mathfrak{s}(f, P) < \epsilon.
\]

XIV.2.3.4. Definition. [Definition 2B] Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 2B if for every \( \epsilon > 0 \) there is a \( \delta > 0 \) such that, for every partition \( P \) of \( J \) with \( ||P|| < \delta \), we have
\[
\mathfrak{S}(f, P) - \mathfrak{s}(f, P) < \epsilon.
\]

Definitions 2A and 2B can be rephrased in terms of “weighted local oscillations”:

XIV.2.3.5. Definition. [Definition 2A’] Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 2A’ if for every \( \epsilon > 0 \) there is a partition \( P \) of \( J \) such that
\[
\sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) < \epsilon
\]
where \( J_k \) is the \( k \)’th subinterval of \( P \).

XIV.2.3.6. Definition. [Definition 2B’] Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 2B’ if for every \( \epsilon > 0 \) there is a \( \delta > 0 \) such that, for every partition \( P \) of \( J \) with \( ||P|| < \delta \), we have
\[
\sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) < \epsilon
\]
where \( J_k \) is the \( k \)’th subinterval of \( P \).

The fact that Definitions 2A’ and 2B’ are equivalent to 2A and 2B respectively is obvious.

There is a less obvious rephrasing of Definition 2A using interior Darboux sums (XIV.2.1.7) which also turns out to be equivalent and technically useful:

XIV.2.3.7. Definition. [Definition 2A’’] Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 2A’’ if for every \( \epsilon > 0 \) there is a partition \( P \) of \( J \) such that
\[
\mathfrak{S}_o(f, P) - \mathfrak{s}_o(f, P) < \epsilon.
\]
XIV.2.3.8. Since for any \( f \) and any partition \( P \) we have
\[
\underline{S}(f, P) \leq \underline{S}_e(f, P) \leq \overline{S}_e(f, P) \leq \overline{S}(f, P)
\]
it follows that if \( f \) is Riemann-integrable in sense 2A, then it is Riemann-integrable in sense 2A”. The converse implication follows easily from XIV.2.1.11., so we have:

XIV.2.3.9. Proposition. Let \( f \) be a bounded function on a closed bounded interval \([a, b]\). Then \( f \) is Riemann-integrable in sense 2A” if and only if \( f \) is Riemann-integrable in sense 2A.

XIV.2.3.10. We could also give an obvious similar version 2B” of 2B. However, although 2B  2B” is obvious, it is difficult to prove directly that 2B”  2B. After the fact it will be seen to be true since we will have 2B”  2A”  2A  2B.

Definition 3

XIV.2.3.11. Definition. [Definition 3A] Let \( f \) be a function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 3A if for every \( \epsilon > 0 \) there are step functions \( \phi \) and \( \psi \) on \( J \) such that \( \phi \leq f \leq \psi \) on \( J \) and
\[
\int_a^b (\psi - \phi) = \int_a^b \psi - \int_a^b \phi < \epsilon .
\]

XIV.2.3.12. One could give a corresponding Definition 3B, but the statement is awkward and does not appear in the literature to my knowledge. But there is a commonly used rephrasing of Definition 3A, which is fairly obviously equivalent to 3A. First note that since a step function is bounded, any function satisfying Definition 3A is automatically bounded. Then make the following definition:

XIV.2.3.13. Definition. Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Define
\[
\int_a^b f(x) \, dx = \sup \left\{ \int_a^b \phi : \phi \text{ step, } \phi \leq f \right\}
\]
\[
\int_a^b f(x) \, dx = \inf \left\{ \int_a^b \psi : \psi \text{ step, } f \leq \psi \right\}.
\]

XIV.2.3.14. Definition. [Definition 3A’] Let \( f \) be a function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) in sense 3A’ if
\[
\int_a^b f(x) \, dx = \int_a^b f(x) \, dx .
\]
The common value is called the \( \text{Riemann integral of } f \) from \( a \) to \( b \).
XIV.2.3.15. To see that Definitions 3A and 3A’ are equivalent, note first that if \( f \) is any bounded function on \( J \) and \( \phi, \psi \) are step functions with \( \phi \leq f \leq \psi \), then \( \int_a^b \phi \leq \int_a^b f(x) \, dx \leq \int_a^b \psi \), and thus by (\() we have

\[
\int_a^b \phi \leq \int_a^b f(x) \, dx \leq \int_a^b \psi.
\]

Thus, if \( f \) is Riemann integrable in sense 3A, then for any \( \epsilon > 0 \) we have

\[
0 \leq \int_a^b f(x) \, dx - \int_a^b f(x) \, dx < \epsilon
\]

so \( f \) is Riemann integrable in sense 3A’. Conversely, if \( f \) is Riemann integrable in sense 3A’, and \( \epsilon > 0 \), choose step functions \( \phi \) and \( \psi \) on \( J \) with \( \phi \leq f \leq \psi \) and

\[
\int_a^b \phi > \int_a^b f(x) \, dx - \int_a^b \phi \, dx = \frac{\epsilon}{2} \quad \text{and} \quad \int_a^b \psi < \int_a^b f(x) \, dx + \int_a^b \phi \, dx = \frac{\epsilon}{2}.
\]

Then

\[
\int_a^b \psi - \int_a^b \phi < \epsilon
\]

so \( f \) is Riemann integrable in sense 3A.

One useful technical fact is the invariance of all the definitions under adding constants:

XIV.2.3.16. Proposition. Let \( f \) be a function on a closed bounded interval \( J = [a, b] \), let \( c \) be a constant, and \( g = f + c \). If \( f \) is Riemann-integrable in any of the senses, then \( g \) is also Riemann-integrable in the same sense, and the Riemann integral of \( g \) is the Riemann integral of \( f \) plus \( c(b - a) \). Riemann and Darboux sums for \( g \) are also the same as for \( f \), plus \( c(b - a) \).

Using this proposition, the theory of Riemann integration can be reduced to the case of nonnegative functions, since any function which is bounded below can be converted to a nonnegative function by adding a constant.

XIV.2.4. Comparing the Definitions

The main result is:

XIV.2.4.1. Theorem. Let \( f \) be a function on a closed bounded interval \( J = [a, b] \). Then

(i) If \( f \) is Riemann-integrable on \( J \) in sense 1A or 1B, then \( f \) is bounded on \( J \).

(ii) If \( f \) is Riemann-integrable in sense 2A or 2B, there is a unique real number \( I \) such that

\[
\mathcal{S}(f, \mathcal{P}) \leq I \leq \mathcal{S}(f, \mathcal{P})
\]

for every partition \( \mathcal{P} \) of \( J \).
If \( f \) is Riemann-integrable on \( J \) in any of the senses 1A, 1B, 2A, 2B, 2A', 2B', 3A, or 3A', it is Riemann-integrable on \( J \) in all these senses, and the Riemann integral of \( f \) over \( J \) is well defined. In particular, the number \( R \) in definitions 1A and 1B is unique, and coincides with the number \( I \) of (ii) and with \( f_a^b f(x) \, dx \) of 3A'.

**XIV.2.4.2.** A few implications are obvious. If \( f \) is Riemann-integrable in sense 1B, it is clearly Riemann-integrable in sense 1A. Similarly, if \( f \) is Riemann-integrable in sense 2B, it is also Riemann-integrable in sense 2A. And the obvious equivalence of 2A and 2A', and of 2B and 2B', has already been noted, as was the slightly less obvious equivalence between 3A and 3A'.

Thus, to prove (i), it suffices to consider an \( f \) which is Riemann-integrable in sense 1A.

**XIV.2.4.3.** **Lemma.** Suppose \( f \) is Riemann-integrable on \( J \) in sense 1A. Then \( f \) is bounded on \( J \).

**Proof:** Let \( R \) be the Riemann integral of \( f \), and let \( \mathcal{P} = \{J_1, \ldots, J_n\} \) be a partition such that for any choice of tags \( t_k \in J_k \), we have

\[
\left| R - \sum_{k=1}^{n} f(t_k) \ell(J_k) \right| < 1.
\]

If \( f \) is not bounded above, then there is a \( k_0 \) such that \( f \) is not bounded above on \( J_{k_0} \). Let \( t_1, \ldots, t_n \) be a choice of tags, and choose \( s_{k_0} \in J_{k_0} \) with \( f(s_{k_0}) > f(t_{k_0}) + \frac{2}{\ell(J_{k_0})} \). Set \( s_k = t_k \) for \( k \neq k_0 \). We then have

\[
\left| R - \sum_{k=1}^{n} f(t_k) \ell(J_k) \right| < 1
\]

\[
\left| R - \sum_{k=1}^{n} f(s_k) \ell(J_k) \right| < 1
\]

so we have

\[
[f(s_{k_0}) - f(t_{k_0})] \ell(J_{k_0}) = \left| \sum_{k=1}^{n} f(s_k) \ell(J_k) - \sum_{k=1}^{n} f(t_k) \ell(J_k) \right|
\]

\[
\leq \left| \sum_{k=1}^{n} f(s_k) \ell(J_k) \right| + \left| \sum_{k=1}^{n} f(t_k) \ell(J_k) \right| - R < 2
\]

which is a contradiction. The proof that \( f \) is bounded below is almost identical. 

Next we compare definitions 1A and 2A, and 1B and 2B.
XIV.2.4.4. **Lemma.** Suppose $f$ is a bounded function on $J$, and $\mathcal{P} = \{J_1, \ldots, J_n\}$ is a partition of $J$. Then $\underline{S}(f, \mathcal{P})$ and $\overline{S}(f, \mathcal{P})$ are the infimum and supremum, respectively, of the Riemann sums corresponding to all choices of tags for $\mathcal{P}$.

**Proof:** It is clear that if $t_1, \ldots, t_n$ is any choice of tags for $\mathcal{P}$, then

$$\underline{S}(f, \mathcal{P}) \leq \sum_{k=1}^{n} f(t_k)\ell(J_k) \leq \overline{S}(f, \mathcal{P}).$$

Fix $\epsilon > 0$. For each $k$, choose $t_k \in J_k$ with $f(t_k) > M_k - \frac{\epsilon}{\ell(J)}$. Then

$$\sum_{k=1}^{n} f(t_k)\ell(J_k) > \sum_{k=1}^{n} M_k\ell(J_k) - \frac{\epsilon}{\ell(J)}\sum_{k=1}^{n} \ell(J_k) = \overline{S}(f, \mathcal{P}) - \epsilon.$$

The argument for the infimum is almost identical.

XIV.2.4.5. Now let $f$ be a (necessarily bounded) function which is Riemann-integrable on $J$ in sense 1A with Riemann integral $R$, $\epsilon > 0$, and $\mathcal{P} = \{J_1, \ldots, J_k\}$ a partition of $J$ such that

$$\left| R - \frac{1}{n} \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < \frac{\epsilon}{3},$$

for any choice of tags $t_k \in J_k$. Then

$$\left| R - \overline{S}(f, \mathcal{P}) \right| \leq \frac{\epsilon}{3} \quad \text{and} \quad \left| R - \underline{S}(f, \mathcal{P}) \right| \leq \frac{\epsilon}{3}$$

so we have

$$\overline{S}(f, \mathcal{P}) - \underline{S}(f, \mathcal{P}) \leq |\overline{S}(f, \mathcal{P}) - R| + |R - \underline{S}(f, \mathcal{P})| \leq \frac{2\epsilon}{3} < \epsilon$$

and so $f$ is Riemann-integrable in sense 2A. An almost identical argument shows that if $f$ is Riemann-integrable in sense 1B, then it is also Riemann-integrable in sense 2B.

For the other direction, note that there is no explicitly given value for the Riemann integral in definitions 2A or 2B, or in 2A′ or 2B′ (or in 3A, but there is one in the equivalent definition 3A′.) The essential point of (ii) is to provide a value for the Riemann integral in case 2A or 2B. To prove (ii), we first make an important observation:

XIV.2.4.6. **Lemma.** If $f$ is a bounded function on $J$, and $\mathcal{P}$ and $\mathcal{Q}$ are any partitions of $J$, then

$$\underline{S}(f, \mathcal{P}) \leq \overline{S}(f, \mathcal{Q}).$$

**Proof:** The key is that there is a partition $\mathcal{R}$ which is a refinement of both $\mathcal{P}$ and $\mathcal{Q}$ (XIV.2.1.2.). Then, by XIV.2.1.6.,

$$\underline{S}(f, \mathcal{P}) \leq \underline{S}(f, \mathcal{R}) \leq \overline{S}(f, \mathcal{R}) \leq \overline{S}(f, \mathcal{Q}).$$

We then conclude from () that

1510
XIV.2.4.7. **Corollary.** Let \( f \) be a bounded function on \( J \). Then

\[
\sup_{\mathcal{P}} \{S(f, \mathcal{P})\} \leq \inf_{\mathcal{P}} \{\overline{S}(f, \mathcal{P})\}
\]

where the supremum and infimum are taken over all partitions of \( J \).

XIV.2.4.8. Thus, if \( f \) is Riemann-integrable in sense 2A, then we have

\[
\sup_{\mathcal{P}} \{S(f, \mathcal{P})\} = \inf_{\mathcal{P}} \{\overline{S}(f, \mathcal{P})\}.
\]

Call this number \( I \). Then we have

\[
\int_a^b f(x) \, dx \leq I \leq \overline{S}(f, \mathcal{P})
\]

for any partition \( \mathcal{P} \), and \( I \) is clearly the only number with this property, proving (ii). Also, if \( \epsilon > 0 \) and \( \mathcal{P} = \{J_1, \ldots, J_n\} \) is a partition with \( \overline{S}(f, \mathcal{P}) - S(f, \mathcal{P}) < \epsilon \), then for any choice of tags \( t_1, \ldots, t_n \) with \( t_k \in J_k \) we have

\[
\left| I - \sum_{k=1}^n f(t_k) \ell(J_k) \right| < \epsilon
\]

so \( f \) is Riemann-integrable in sense 1A with Riemann integral \( I \). If \( f \) is Riemann-integrable in sense 2B, it follows almost identically that \( f \) is Riemann-integrable in sense 1B.

XIV.2.4.9. Now let us compare definitions 2A and 3A. Let \( f \) be a bounded function on \( J = [a, b] \), which we may assume to be nonnegative by XIV.2.3.16.. We want to show that

\[
\inf_{\mathcal{P}} \{\overline{S}(f, \mathcal{P})\} = \inf \left\{ \int_a^b \psi : \psi \text{ step, } f \leq \psi \right\}
\]

and similarly for lower estimates. (We have \( \geq \) by the previous argument.) Let \( \psi \) be a step function for which \( f \leq \psi \), and \( \mathcal{P} \) the partition consisting of \( x_0 = a, x_n = b \), and the jump discontinuities \( x_1, \ldots, x_{n-1} \).
of $\psi$ in $(a, b)$. Let $c_k$ be the value of $\psi$ on $(x_{k-1}, x_k)$, and $d_k = f(x_k)$. If $M_k$ is the supremum of $f$ on $J_k = [x_{k-1}, x_k]$, we have $M_k \leq \max(c_k, d_k, d_{k-1})$. So if $d_{k-1}, d_k \leq c_k$ for all $k$, we have

$$\mathcal{I}(f, \mathcal{P}) \leq \int_a^b \psi$$

but this fails if $d_k > c_k$ or $d_{k-1} > c_k$ for some $k$. To get around this problem, use definition $2A''$ instead: we have

$$\mathcal{I}_o(f, \mathcal{P}) \leq \int_a^b \psi$$

for any such $\psi$. We thus get the desired inequality

$$\inf_{\mathcal{P}} \{\mathcal{I}(f, \mathcal{P})\} = \inf_{\mathcal{P}} \{\mathcal{I}_o(f, \mathcal{P})\} \leq \inf \left\{ \int_a^b \psi : \text{step, } f \leq \psi \right\}$$

where the first equality is from XIV.2.11. By a similar argument we obtain

$$\sup_{\mathcal{P}} \{\mathcal{I}(f, \mathcal{P})\} \geq \sup \left\{ \int_a^b \phi : \phi \text{ step, } \phi \leq f \right\}$$

(the opposite inequality follows from the previous argument).

**XIV.2.4.12.** Thus, if $f$ is Riemann-integrable in sense $3A$, it is Riemann-integrable in sense $2A$, and the number $I$ of (ii) is exactly $\int_a^b f(x) \, dx$.

**XIV.2.4.13.** We have now shown that definitions $1A$, $2A$, and $3A$ (hence also $2A'$ and $3A'$) define the same class of Riemann-integrable functions, and the Riemann integrals coincide (using (ii) in case $2A$). We will say that a function Riemann-integrable in any of these senses is Riemann-integrable in sense $A$. Similarly, definitions $1B$ and $2B$ (hence also $2B'$) define the same class of Riemann-integrable functions, with a well-defined Riemann integral, called the functions Riemann-integrable in sense $B$. Also, a function Riemann-integrable in sense $B$ is Riemann-integrable in sense $A$, and the Riemann integrals coincide. Thus, finish the proof of the theorem, we need to show that a function which is Riemann-integrable in sense $A$ is Riemann-integrable in sense $B$. This is the least obvious part of the proof of the theorem.

**XIV.2.4.14.** So suppose $f$ is Riemann-integrable on $J$ in sense $A$, and let $\epsilon > 0$. Let $\mathcal{P} = \{x_0, x_1, \ldots, x_n\}$ be a partition satisfying

$$\mathcal{I}(f, \mathcal{P}) - \mathcal{I}_o(f, \mathcal{P}) < \frac{\epsilon}{2}.$$ 

Let $M$ be a bound for $f$ on $J$, and set $g = f + M$, $h = f - M$. Then $0 \leq g \leq 2M$ and $-2M \leq h \leq 0$. Let $\delta < \frac{\epsilon}{4(n-1)M}$. If $Q = \{I_1, \ldots, I_n\}$ is a partition of $J$ with $\|Q\| < \delta$, then $Q$ might not quite be a refinement of $\mathcal{P}$ since some or all of the $x_k$ might be interior points of intervals of $Q$. But there are at most $n - 1$ of these intervals; let $N_1$ be the set of $r$ such that $I_r$ is one of these, and let $N_2$ be the rest of the indices. Then

$$\mathcal{I}(g, Q) = \sum_{r \in N_1} M_r \ell(I_r) + \sum_{r \in N_2} M_r \ell(I_r)$$

1512
where \( M_r \) is the supremum of \( g \) on \( I_r \). Since for each \( r \in N_2 \) we have that \( I_r \) is contained in one of the \( J_k \), and the supremum of \( g \) over \( I_r \) is less than or equal to the supremum of \( g \) over the corresponding \( J_k \), we have that

\[
\sum_{r \in N_2} M_r \ell(I_r) \leq \mathcal{S}(g, \mathcal{P}) .
\]

Since each \( I_r \) has length \( < \delta \), and \( N_1 \) contains at most \( n - 1 \) indices, we have

\[
\sum_{r \in N_1} M_r \ell(I_r) \leq 2(n - 1)M \delta < \frac{\epsilon}{2}.
\]

Thus we have

\[
\mathcal{S}(g, \mathcal{Q}) < \mathcal{S}(g, \mathcal{P}) + \frac{\epsilon}{2}.
\]

By translation (XIV.2.3.16.), we thus have

\[
\mathcal{S}(f, \mathcal{Q}) < \mathcal{S}(f, \mathcal{P}) + \frac{\epsilon}{2}.
\]

Applying a similar argument to \( h \) using infima, we have

\[
\mathcal{S}(h, \mathcal{Q}) > \mathcal{S}(h, \mathcal{P}) - \frac{\epsilon}{2}
\]

and so by linearity we have

\[
\mathcal{S}(f, \mathcal{Q}) > \mathcal{S}(f, \mathcal{P}) - \frac{\epsilon}{2}.
\]

Combining the two estimates, we get

\[
\mathcal{S}(f, \mathcal{Q}) - \mathcal{S}(f, \mathcal{Q}) < \epsilon
\]

and so \( f \) is Riemann-integrable in sense 2B (i.e. in sense B).

**XIV.2.4.15.** There is one last loose end. We must show that the number \( R \) in Definition 1A is unique. Let \( f \) be Riemann-integrable in sense 1A, and \( R \) a number satisfying Definition 1A. Fix \( \epsilon > 0 \), and let \( \mathcal{P} \) be a partition such that

\[
\left|R - \sum_{k=1}^{n} f(t_k) \ell(J_k)\right| < \epsilon
\]

for any choice of tags. Then by XIV.2.4.4., we have

\[
|R - \mathcal{S}(f, \mathcal{P})| \leq \epsilon
\]

and so, since the number \( I \) of (ii) is between \( \mathcal{S}(f, \mathcal{P}) \) and \( \mathcal{S}(f, \mathcal{P}) \), we have

\[
|R - I| \leq \epsilon.
\]

since \( \epsilon > 0 \) is arbitrary, we have \( R = I \) and \( R \) is uniquely determined.

This completes the proof of XIV.2.4.1.

In light of Theorem XIV.2.4.1., we can make the definitive definition of Riemann-integrability:
XIV.2.4.16. Definition. Let \( f \) be a bounded function on a closed bounded interval \( J = [a, b] \). Then \( f \) is Riemann-integrable on \( J \) if it is Riemann-integrable in any of the senses 1A-B, 2A-B, 2A'-B', or 3A-A'. The Riemann integral of \( f \) over \( J \), or from \( a \) to \( b \), denoted \( \int_a^b f(x) \, dx \), is the number defined in XIV.2.3.14.

XIV.2.5. Integrals of Continuous and Monotone Functions

Let \( \chi_Q \) be the indicator function of \( Q \).

XIV.2.6. Properties of the Riemann Integral

The following result summarizes some of the most important properties of the Riemann integral:

XIV.2.6.1. Theorem. Let \( [a, b] \) be a closed bounded interval in \( \mathbb{R} \), and \( f \) and \( g \) real-valued functions on \( [a, b] \), and \( c \in \mathbb{R} \). Then

(i) (Linearity) If \( f \) and \( g \) are Riemann-integrable on \( [a, b] \), then \( f + g \) and \( cf \) are Riemann-integrable on \( [a, b] \), and

\[
\int_a^b [f(x) + g(x)] \, dx = \int_a^b f(x) \, dx + \int_a^b g(x) \, dx
\]

\[
\int_a^b (cf)(x) \, dx = c \int_a^b f(x) \, dx
\]

(ii) (Monotonicity) If \( f \) and \( g \) are Riemann-integrable on \( [a, b] \), and if \( f(x) \leq g(x) \) for all \( x \in [a, b] \), then

\[
\int_a^b f(x) \, dx \leq \int_a^b g(x) \, dx
\]

(iii) The constant function \( c \) is Riemann-integrable on \( [a, b] \), and

\[
\int_a^b c \, dx = c(b - a)
\]

Proof: For (i), note that for any tagged partition \( \hat{P} \) of \( [a, b] \), we have

\[
\mathcal{R}(f + g, \hat{P}) = \mathcal{R}(f, \hat{P}) + \mathcal{R}(g, \hat{P})
\]

\[
\mathcal{R}(cf, \hat{P}) = c\mathcal{R}(f, \hat{P})
\]

For (ii), if \( f \leq g \), for any tagged partition \( \hat{P} \) of \( [a, b] \) we have

\[
\mathcal{R}(f, \hat{P}) \leq \mathcal{R}(g, \hat{P})
\]

For (iii), for any tagged partition \( \hat{P} \) of \( [a, b] \) we have

\[
\mathcal{R}(c, \hat{P}) = c(b - a)
\]

Apply Definition 1.

The following crude estimate is surprisingly useful:
**XIV.2.6.2.** Corollary. Let $f$ be Riemann-integrable on $[a, b]$. If $m \leq f(x) \leq M$ for all $x \in [a, b]$, then

$$m(b-a) \leq \int_a^b f(x) \, dx \leq M(b-a).$$

In particular, if $|f(x)| \leq M$ for all $x \in [a, b]$, then

$$\left| \int_a^b f(x) \, dx \right| \leq M(b-a).$$

This Corollary can be proved more easily by simply noting that if $\mathcal{P}$ is the partition of $[a, b]$ consisting of one interval, then

$$m(b-a) \leq S(f, \mathcal{P}) \leq \int_a^b f(x) \, dx \leq S(f, \mathcal{P}) \leq M(b-a).$$

Another closely related inequality, an integral version of the Triangle Inequality ($\omega$), is also used regularly:

**XIV.2.6.3.** Corollary. Let $f$ be Riemann-integrable on $[a, b]$. Then $|f|$ is also Riemann-integrable on $[a, b]$, and

$$\left| \int_a^b f(x) \, dx \right| \leq \int_a^b |f(x)| \, dx.$$

**Proof:** We only need to see that $|f|$ is Riemann-integrable. But if $\mathcal{P}$ is any partition of $[a, b]$, then

$$\mathcal{S}(|f|, \mathcal{P}) - \mathcal{S}(|f|, \mathcal{P}) \leq \mathcal{S}(f, \mathcal{P}) - \mathcal{S}(f, \mathcal{P})$$

since $\omega(|f|, I) \leq \omega(f, I)$ for any interval $I$. \(\blacksquare\)

**XIV.2.6.4.** In the context of Figure XIV.9, this is really a triangle inequality: if $A_1$ and $A_3$ are the areas of the two blue regions and $A_2$ the area of the tan region, then $\left| \int_a^b f(x) \, dx \right| = |A_1 - A_2 + A_3|$ and $\int_a^b |f(x)| \, dx = A_1 + A_2 + A_3$.

**Continuity of the Integral**

One must be quite careful about interchanging integrals and limits: the Riemann integral is not continuous for all types of limits. In fact, obtaining good limit theorems was one of the main goals of developing modern integration theories.

There is one simple but important situation where Riemann integration is continuous:
XIV.2.6.5. PROPOSITION. Let \([a, b]\) be a closed bounded interval, and \((f_n)\) a sequence of functions on \([a, b]\) converging uniformly on \([a, b]\) to a function \(f\). If each \(f_n\) is Riemann-integrable on \([a, b]\), then \(f\) is Riemann-integrable on \([a, b]\) and
\[
\int_a^b f_n(x) \, dx \to \int_a^b f(x) \, dx .
\]
(Informally, the integral of a uniform limit is the limit of the integrals.)

PROOF: If we know \(f\) is Riemann-integrable (e.g. if the \(f_n\) are continuous), this follows immediately from XIV.2.6.1., XIV.2.6.2., and XIV.2.6.3.: if \(\epsilon > 0\), choose \(N\) so that \(|f_n(x) - f(x)| < \frac{\epsilon}{b-a}\) for all \(n \geq N\) and all \(x \in [a, b]\); then
\[
\left| \int_a^b f_n(x) \, dx - \int_a^b f(x) \, dx \right| = \left| \int_a^b (f_n(x) - f(x)) \, dx \right| \leq \int_a^b |f_n(x) - f(x)| \, dx \leq \frac{\epsilon}{b-a}(b-a) = \epsilon
\]
for all \(n \geq N\).

The argument that \(f\) is Riemann-integrable is very similar to the proof that a uniform limit of continuous functions is continuous (). Note that if \(|f_n(x) - f(x)| < \epsilon\) for all \(x \in [a, b]\), then for any partition \(P\) we have
\[
|S(f_n, P) - S(f, P)| \leq \epsilon(b-a)
\]
so, if \(\epsilon > 0\), and \(N\) is such that
\[
|f_n(x) - f(x)| < \frac{\epsilon}{3(b-a)}
\]
for all \(n \geq N\) and \(x \in [a, b]\), and \(P\) is a partition such that
\[
S(f_n, P) - S(f, P) < \frac{\epsilon}{3}
\]
we have that
\[
S(f, P) - S(f, P) \leq |S(f, P) - S(f_n, P)| + |S(f_n, P) - S(f, P)| + |S(f_n, P) - S(f_n, P)|
\]
\[
< \frac{\epsilon}{3(b-a)}(b-a) + \frac{\epsilon}{3} + \frac{\epsilon}{3(b-a)}(b-a) = \epsilon .
\]
Since \(\epsilon > 0\) is arbitrary, \(f\) is Riemann-integrable by Definition 2A.

Dense Sets and Integrability

XIV.2.6.6. PROPOSITION. If \(f_1\) and \(f_2\) are Riemann-integrable on \([a, b]\) and agree on a dense subset \(D\), then
\[
\int_a^b f_1(x) \, dx = \int_a^b f_2(x) \, dx .
\]

PROOF: If \(P\) is any partition of \([a, b]\), then Riemann sums for \(f_1\) and \(f_2\) with respect to \(P\) can be chosen with the tags in \(D\), so these Riemann sums coincide.

We can use this fact to give a definition of Riemann-integrability and the Riemann integral for certain functions which are not defined everywhere.
XIV.2.6.7. Definition. Let \( f \) be a real-valued function defined on a dense subset \( D \) of \([a, b]\). Then \( f \) is Riemann-integrable on \([a, b]\) if there is a Riemann-integrable function \( g \) on \([a, b]\) whose restriction to \( D \) is \( f \). The Riemann integral \( \int_a^b f(x) \, dx \) is defined to be \( \int_a^b g(x) \, dx \).

XIV.2.6.8. The Riemann integral of \( f \) is well defined independent of the choice of \( g \) by XIV.2.6.6., and any two choices for \( g \) actually agree almost everywhere (XIV.3.11.1.). An \( f \) which is Riemann-integrable in this extended sense is necessarily bounded.

The simplest situation is when \( f \) is defined except at finitely many points. Such a function can be changed, defined, or left undefined at finitely many points without changing its Riemann-integrability or its Riemann integral:

XIV.2.6.9. Proposition. Let \( f \) be a function defined at all but finitely many points of \([a, b]\). If \( f \) is Riemann-integrable, then any function \( g \) defined on all but finitely many points and agreeing with \( f \) at all but finitely many points is also Riemann-integrable, with

\[
\int_a^b g(x) \, dx = \int_a^b f(x) \, dx.
\]

Proof: If \( P \) is a partition fine enough that it includes all the points where \( f \) or \( g \) is undefined and the points where they have unequal values, we have

\[
\mathcal{S}_o(f, P) = \mathcal{S}_o(g, P) \text{ and } \mathcal{S}_o(f, P) = \mathcal{S}_o(g, P)
\]

and we can apply Definition 2A'.
XIV.2.7.2. The term “null set” is often used as a synonym for “empty set”, but in analysis it is more useful to use the definition we have given. (Note that the empty set is a null set by the analysis definition, but so are many nonempty sets, such as any finite set; in fact finite sets have zero content.)

Any subset of a null set (resp. a set of zero content) is a null set (resp. a set of zero content).

A set with zero content is obviously a null set (since $\emptyset$ is a degenerate interval); the converse is false:

XIV.2.7.3. **Examples.** (i) Any countable set is a null set. [If $A$ is countable, say $A = \{x_n : n \in \mathbb{N}\}$, and $\epsilon > 0$, let $I_k$ be the open interval centered at $x_k$ of length $\frac{\epsilon}{2^k}$.] Thus, for example, $\mathbb{Q}$ is a null set. But $\mathbb{Q}$ does not have zero content since a set with zero content is clearly bounded.

(ii) Even a bounded null set need not have zero content. For example, $A = \mathbb{Q} \cap [0,1]$ is a null set by (i), but if $\{I_1, \ldots, I_n\}$ is a finite set of intervals containing $A$, then $\bigcup_{k=1}^n I_k$ contains all but finitely many points of $[0,1]$. It follows that $\sum_{k=1}^n \ell(I_k) \geq 1$ (this seems obvious, but is actually somewhat tricky to prove; see XIV.2.7.12). This argument can be used to show that a true interval cannot be a null set, but this can be otherwise proved; see XIV.2.7.12. Thus the complement of a null set is dense.

(iii) There are uncountable sets which have measure zero, e.g. the Cantor set $K()$. In fact, $K$ has zero content.

In testing whether a set is a null set or has zero content, it suffices to use open intervals:

XIV.2.7.4. **Proposition.** Let $A$ be a subset of $\mathbb{R}$. Then $A$ is a null set if and only if there is a sequence $(J_k)$ of bounded open intervals with $A \subseteq \bigcup_{k=1}^\infty J_k$ and $\sum_{k=1}^\infty \ell(J_k) < \epsilon$. $A$ has zero content if and only if for every $\epsilon > 0$ there is a finite sequence $(J_k)$ of bounded open intervals with the same properties.

**Proof:** Suppose $A$ has measure zero, and $\epsilon > 0$. There is a sequence $(I_k)$ of bounded intervals such that $A \subseteq \bigcup_{k=1}^\infty I_k$ and $\sum_{k=1}^\infty \ell(I_k) < \epsilon$. For each $k$ let $J_k$ be the open interval with the same midpoint as $I_k$ with $\ell(J_k) = 2\ell(I_k)$; then $I_k \subseteq J_k$, and the $J_k$ work. The same argument works for finitely many intervals in the zero content case. The converse implication is trivial.

The converse implication is trivial.

XIV.2.7.5. **Corollary.** A subset $A$ has zero content if and only if it is bounded and its closure has measure zero. In particular, a closed bounded null set has zero content.

**Proof:** Suppose $A$ has zero content. Then $A$ is bounded. Let $\epsilon > 0$, and let $\{I_1, \ldots, I_n\}$ be a finite set of bounded intervals with $A \subseteq \bigcup_{k=1}^n I_k$ and $\sum_{k=1}^n \ell(I_k) < \epsilon$. Replacing each $I_k$ by its closure if necessary, we may assume each $I_k$ is closed. But then $\bigcup_{k=1}^n I_k$ is closed, so $A \subseteq \bigcup_{k=1}^n I_k$ and $A$ is a null set (in fact, has zero content).

Conversely, suppose $A$ is bounded and $\bar{A}$ is a null set. Fix $\epsilon > 0$, and let $(I_k)$ be a sequence of bounded open intervals such that $A \subseteq \bigcup_{k=1}^\infty I_k$ and $\sum_{k=1}^\infty \ell(I_k) < \epsilon$. The $I_k$ cover $A$, and since $A$ is compact there is a finite subcover, i.e. $A \subseteq \bigcup_{k=1}^n I_k$ for some $n$. Since $\sum_{k=1}^n \ell(I_k) < \epsilon$, it follows that $\bar{A}$, and hence $A$, has zero content.
Unions of Null Sets

A finite union of null sets is a null set, and a finite union of sets of zero content has zero content. For null sets, we have a better result:

**XIV.2.7.6. Proposition.** A countable union of null sets is a null set.

**Proof:** Let \( A_n \) be a null set for each \( n \), and \( A = \bigcup_{n=1}^{\infty} A_n \). Fix \( \epsilon > 0 \). For each \( n \), let \( \{I_{n,k}\} \) be a sequence of bounded intervals such that \( A_n \subseteq \bigcup_{k=1}^{\infty} I_{n,k} \) and \( \sum_{k=1}^{\infty} \ell(I_{n,k}) < \frac{\epsilon}{2^n} \). Then the set \( \{I_{n,k} : n, k \in \mathbb{N}\} \) is a countable set of bounded intervals with \( A \subseteq \bigcup_{n,k} I_{n,k} \) and \( \sum_{n,k} \ell(I_{n,k}) < \epsilon \). Since \( \epsilon > 0 \) is arbitrary, \( A \) is a null set.

**XIV.2.7.7.** This result gives another proof that every countable set is a null set, since it is a countable union of singletons.

It is not true that countable unions of sets of zero content have zero content (e.g. \( \mathbb{Q} \)). Closure of the null sets under countable unions indicates that the notion of null set is a more flexible one than that of a set of zero content. Good behavior under countable unions turns out to be necessary for a good theory of measure and advanced integration.

Almost Everywhere

**XIV.2.7.8.** It is convenient to use the phrase *almost everywhere*, abbreviated a.e., to mean “except on a null set.” Thus, for example, if we say “the functions \( f \) and \( g \) agree almost everywhere,” we mean “the set of all \( x \) for which \( f(x) = g(x) \) does not hold (i.e. one or both of \( f(x) \) and \( g(x) \) is undefined, or both are defined but unequal) has measure zero.”

Nonnegative Riemann-Integrable Functions

Using the language of null sets and almost everywhere, we can give an important fact about nonnegative Riemann-integrable functions:

**XIV.2.7.9. Proposition.** Let \( f \) be a nonnegative function on an interval \([a, b]\). If \( f \) is Riemann-integrable on \([a, b]\) and \( \int_{a}^{b} f(x) \, dx = 0 \), then \( f = 0 \) a.e. on \([a, b]\).

**Proof:** For each \( n \), let

\[
A_n = \left\{ x \in [a, b] : f(x) \geq \frac{1}{n} \right\}.
\]

We will show that each \( A_n \) has zero content, hence is a null set. This will suffice, since then

\[
\{ x \in [a, b] : f(x) \neq 0 \} = \bigcup_{n=1}^{\infty} A_n
\]

will be a null set by XIV.2.7.6.
Fix $n$, and then fix $\epsilon > 0$. Let $\mathcal{P} = \{x_0, \ldots, x_m\}$ be a partition with $\mathcal{S}(f, \mathcal{P}) < \frac{\epsilon}{n}$. Let $K$ be the set of all $k$ such that $A_n \cap [x_{k-1}, x_k]$ is nonempty. Then $A_n \subseteq \bigcup_{k \in K} [x_{k-1}, x_k]$. If $M_k$ is the supremum of $f$ on $[x_{k-1}, x_k]$, we have

$$
\frac{1}{n} \sum_{k \in K} (x_k - x_{k-1}) \leq \sum_{k \in K} M_k (x_k - x_{k-1}) \leq \sum_{k = 1}^m M_k (x_k - x_{k-1}) = \mathcal{S}(f, \mathcal{P}) < \frac{\epsilon}{n}
$$

and so $\sum_{k \in K} (x_k - x_{k-1}) < \epsilon$. Since $\epsilon > 0$ is arbitrary, $A_n$ has zero content.

**XIV.2.7.10.** XIV.2.6.6. gives a strong converse: if $f$ is Riemann-integrable on $[a, b]$ and equal to zero on a dense subset, then $\int_a^b f(x) \, dx = 0$.

One direction of the next proposition is a corollary of the proof of XIV.2.7.9.:

**XIV.2.7.11.** **Proposition.** Let $A$ be a subset of $[a, b]$. Then $A$ has zero content if and only if the indicator function $\chi_A$ is Riemann-integrable on $[a, b]$ and $\int_a^b \chi_A(x) \, dx = 0$.

**XIV.2.7.12.** **Corollary.** No interval is a null set.

**Proof:** First let $[a, b]$ be a closed bounded interval (with $a < b$). If $[a, b]$ is a null set, it has zero content by XIV.2.7.5.. But $\int_a^b \chi_{[a,b]}(x) \, dx = b - a > 0$, so $[a, b]$ does not have zero content by XIV.2.7.11.. Now let $I$ be an arbitrary interval, and let $a, b \in I$, $a < b$. If $I$ is a null set, then its subset $[a, b]$ is also a null set, a contradiction.

Almost the same result as XIV.2.6.9. holds if “finite set” is replaced by “set of zero content,” but boundedness of the modified function is not automatic and must be assumed:

**XIV.2.7.13.** **Proposition.** Let $f$ be a function defined on $[a, b]$ except on a set of zero content. If $f$ is Riemann-integrable, then any bounded function $g$ defined except on a set of zero content and agreeing with $f$ except on a set of zero content is also Riemann-integrable, with

$$
\int_a^b g(x) \, dx = \int_a^b f(x) \, dx.
$$

**Proof:** Let $h$ be a Riemann-integrable function on $[a, b]$ agreeing with $f$ where $f$ is defined. Extend $g$ in any manner to a bounded function defined everywhere on $[a, b]$: we will from now on assume $g$ is defined everywhere.

Let $A$ be a set of zero content such that $f$ and $g$ are defined and equal on $[a, b] \setminus A$, and let $M > 0$ be a bound for both $g$ and $h$ on $[a, b]$. Fix $\epsilon > 0$. Let $\{I_1, \ldots, I_n\}$ be a set of bounded intervals with $A \subseteq \bigcup_{k=1}^n I_k$.
and $\sum_{k=1}^{n} \ell(I_k) < \frac{\epsilon}{4M}$. Let $P$ be a partition of $[a, b]$ such that $\overline{S}_n(h, P) - \underline{S}_n(h, P) < \frac{\epsilon}{4}$ and such that $P$ includes all endpoints of $I_1, \ldots, I_n$ which are in $[a, b]$. Let $M_j$ and $m_j$ be the supremum and infimum of $g$ on $(x_{j-1}, x_j)$.

Let $J$ be the set of all indices $j$ such that $(x_{j-1}, x_j) \subseteq \bigcup_{k=1}^{n} I_k$. Then

$$\overline{S}_n(g, P) - \underline{S}_n(g, P) = \sum_{j=1}^{m} (M_j - m_j)(x_j - x_{j-1})$$

$$= \sum_{j \in J} (M_j - m_j)(x_j - x_{j-1}) + \sum_{j \notin J} (M_j - m_j)(x_j - x_{j-1}) < 2M \frac{\epsilon}{4M} + \frac{\epsilon}{2} = \epsilon.$$ 

Since $\epsilon > 0$ is arbitrary, $g$ is Riemann-integrable on $[a, b]$ by Definition 2A'', and

$$\int_a^b g(x) \, dx = \int_a^b h(x) \, dx.$$ 

\

**XIV.2.8. The Upper and Lower Envelope of a Riemann-Integrable Function**

We continue somewhat in the spirit of XIV.2.3.11., in a sense carrying it to its logical conclusion.

**XIV.2.8.1. Definition.** Let $f$ be a bounded function on an interval $[a, b]$. Define

$$\overline{f} = \sup \{ \phi : [a, b] \to \mathbb{R} : \phi \text{ lower semicontinuous, } \phi \leq f \}$$

$$\underline{f} = \inf \{ \psi : [a, b] \to \mathbb{R} : \psi \text{ upper semicontinuous, } f \leq \psi \}.$$ 

The functions $\overline{f}$ and $\underline{f}$ are called the upper and lower envelopes of $f$ on $[a, b]$.

We have that $\overline{f}$ and $\underline{f}$ are bounded, and $\underline{f} \leq f \leq \overline{f}$. By (4), $\overline{f}$ is upper semicontinuous and $\underline{f}$ is lower semicontinuous.

**XIV.2.8.2. Proposition.** Let $f$ be a bounded function on $[a, b]$, and $\overline{f}$ and $\underline{f}$ its upper and lower envelopes. If $J$ is any open subinterval of $[a, b]$, then

$$\sup_{x \in J} \overline{f}(x) = \sup_{x \in J} f(x) \quad \text{and} \quad \inf_{x \in J} \underline{f}(x) = \inf_{x \in J} f(x).$$ 

**Proof:** It is clear that $\sup_{x \in J} \overline{f}(x) \geq \sup_{x \in J} f(x)$. Conversely, let $M = \sup_{x \in J} f(x)$. Then there is an upper semicontinuous step function $\psi$ on $[a, b]$, which is constant on $J$ and two other subintervals of $[a, b]$, with $\psi = M$ on $J$ and $f \leq \psi$. Thus

$$\sup_{x \in J} \overline{f}(x) \leq \sup_{x \in J} \psi(x) = M = \sup_{x \in J} f(x).$$ 

The argument for $\underline{f}$ is nearly identical.
**XIV.2.8.3.** COROLLARY. Let $f$ be a bounded function on $[a, b]$, and $ar{f}$ and $\underline{f}$ its upper and lower envelopes. For each $x \in [a, b]$, we have

$$\bar{f}(x) = \max\{f(x), \limsup_{y \to x} f(y)\} \quad \text{and} \quad \underline{f}(x) = \min\{f(x), \liminf_{y \to x} f(y)\} .$$

**XIV.2.8.4.** COROLLARY. Let $f$ be a bounded function on $[a, b]$, and $ar{f}$ and $\underline{f}$ its upper and lower envelopes. If $\mathcal{P}$ is any partition of $[a, b]$, then

$$\mathfrak{S}_o(\bar{f}, \mathcal{P}) = \mathfrak{S}_o(f, \mathcal{P}) \quad \text{and} \quad \mathfrak{S}_\delta(f, \mathcal{P}) = \mathfrak{S}_\delta(f, \mathcal{P}) .$$

Thus we have

$$\int_a^b \bar{f}(x) \, dx = \int_a^b f(x) \, dx \quad \text{and} \quad \int_a^b \underline{f}(x) \, dx = \int_a^b f(x) \, dx .$$

Using these results, we can prove the fundamental characterization of Riemann integrability:

**XIV.2.8.5.** THEOREM. Let $f$ be a function on an interval $[a, b]$. Then $f$ is Riemann-integrable on $[a, b]$ if and only if it is bounded, and continuous almost everywhere.

PROOF: We have already shown that a Riemann-integrable function must be bounded. Suppose $f$ is Riemann-integrable on $[a, b]$. Let $\bar{f}$ and $\underline{f}$ be its upper and lower envelopes, and set $g = \bar{f} - \underline{f}$. Then $g \geq 0$. Fix $\epsilon > 0$, and choose a partition $\mathcal{P}$ such that $\mathfrak{S}_o(f, \mathcal{P}) - \mathfrak{S}_\delta(f, \mathcal{P}) < \epsilon$. Then we have

$$\mathfrak{S}_o(g, \mathcal{P}) = \mathfrak{S}_o(\bar{f}, \mathcal{P}) - \mathfrak{S}_\delta(f, \mathcal{P}) = \mathfrak{S}_o(f, \mathcal{P}) - \mathfrak{S}_\delta(f, \mathcal{P}) < \epsilon$$

and, since $g$ is nonnegative and hence $0 \leq \mathfrak{S}_o(g, \mathcal{P}) \leq \mathfrak{S}_o(g, \mathcal{P}) < \epsilon$, we have that $g$ is Riemann-integrable on $[a, b]$ and $\int_a^b g(x) \, dx = 0$. Thus $g = 0$ a.e. (XIV.2.7.9.), i.e. $f = \bar{f} = \underline{f}$ a.e. If $x$ is such that $\bar{f}(x) = f(x) = \underline{f}(x)$, then $f$ is both upper semicontinuous and lower semicontinuous at $x$, hence continuous at $x$. (The conclusion also follows immediately from XIV.2.8.3.)

Conversely, suppose $f$ is bounded and continuous a.e. on $[a, b]$. Let $M > 0$ be a bound for $f$ on $[a, b]$, and $A$ the set of discontinuities of $f$ in $[a, b]$. Fix $\epsilon > 0$. Let $(I_k)$ be a sequence of open intervals with $A \subseteq \bigcup_{k=1}^{\infty} I_k$ and $\sum_{k=1}^{\infty} \ell(I_k) < \frac{\epsilon}{4M}$. For each $x \in [a, b] \setminus A$ there is an open interval $J_x$ around $x$ with $|f(y) - f(z)| < \frac{\epsilon}{2(b-a)}$ for all $y, z \in J_x$, since $f$ is continuous at $x$. The collection $\{I_k : k \in N\} \cup \{J_x : x \in [a, b] \setminus A\}$ is an open cover of $[a, b]$. By compactness, there is a finite subcover, so there is a partition $\mathcal{P} = \{x_0, \ldots, x_n\}$ of $[a, b]$ such that each interval $[x_{j-1}, x_j]$ is contained either in one of the $I_k$ or one of the $J_x$. Let $K$ be the set of $j$ such that $[x_{j-1}, x_j]$ is contained in one of the $I_k$. Then, if $M_j$ and $m_j$ are the supremum and infimum of $f$ on $[x_{j-1}, x_j]$, we have

$$\mathfrak{S}(f, \mathcal{P}) - \mathfrak{S}(f, \mathcal{P}) = \sum_{j=1}^{n} (M_j - m_j)(x_j - x_{j-1})$$

$$= \sum_{j \in K} (M_j - m_j)(x_j - x_{j-1}) + \sum_{j \notin K} (M_j - m_j)(x_j - x_{j-1}) < 2M \frac{\epsilon}{4M} + \frac{\epsilon}{2(b-a)}(b-a) = \epsilon .$$

Since $\epsilon > 0$ is arbitrary, $f$ is Riemann-integrable on $[a, b]$. \(\blacksquare\)
XIV.2.9. Differentiation and the Fundamental Theorem of Calculus

In this section, we establish the fundamental results that integration and differentiation are inverse processes.

XIV.2.9.1. Let \( f \) be a Riemann-integrable function on \([a, b]\). Then, for any \( x \in [a, b] \), \( f \) is Riemann-integrable on \([a, x]\) . For \( a \leq x \leq b \), define

\[
A(x) = \int_a^x f(t) \, dt .
\]

\( A(x) \) is called the cumulative area of \( f \) from \( a \) to \( x \); this can be regarded as a function \( A \) from \([a, b]\) to \( \mathbb{R} \), called the cumulative area function of \( f \) from \( a \) to \( b \). (If \( f \) is nonnegative, \( A(x) \) can be truly regarded as the “area” of the region below the graph of \( f \) and above the \( x \)-axis, between the vertical lines at \( a \) and \( x \). Rigorously, this is backward: the area of this region can be defined to be this integral – see (). For a general \( f \), \( A(x) \) gives the “signed area” under the graph from \( a \) to \( x \), with regions below the \( x \)-axis having negative “area.”)

XIV.2.9.2. Proposition. If \( f \) is a Riemann-integrable function on \([a, b]\), then its cumulative area function is Lipschitz () on \([a, b]\), and hence (uniformly) continuous on \([a, b]\).

Proof: Since \( f \) is Riemann-integrable on \([a, b]\), it is bounded; let \( M \) be a bound for \( |f| \) on \([a, b]\). Then, for every \( x, y \in [a, b] \), \( x < y \), we have

\[
|A(y) - A(x)| = \left| \int_a^y f(t) \, dt - \int_a^x f(t) \, dt \right| = \left| \int_x^y f(t) \, dt \right| \leq \int_x^y |f(t)| \, dt \leq M(y - x)
\]

by () .

The next theorem is the main result of this section and, as the name suggests, one of the most important theorems in elementary analysis:

XIV.2.9.3. Theorem. [Fundamental Theorem of Calculus, Version 1] Let \( f \) be Riemann-integrable on \([a, b]\), and \( A \) its cumulative area function. If \( f \) is continuous at \( x_0 \in [a, b] \), then \( A \) is differentiable at \( x_0 \) and \( A'(x_0) = f(x_0) \).

Before giving the proof of this theorem, we give some simple consequences. The first is the same as V.8.5.9.:

XIV.2.9.4. Corollary. If \( f \) is continuous on \([a, b]\), then there is a function \( F \), unique up to addition of a constant, which is differentiable on \([a, b]\) and for which \( F''(x) = f(x) \) for all \( x \in [a, b] \).

The uniqueness of the antiderivative up to a constant follows immediately from V.8.2.5.:

The next variation gives the common procedure for calculating the value of the integral:
XIV.2.9.5. **Corollary.** If \( f \) is continuous on \([a, b]\), and \( F \) is any antiderivative for \( f \) on \([a, b]\), then
\[
\int_a^b f(x) \, dx = F(b) - F(a) .
\]

**Proof:** Let \( A \) be the cumulative area function of \( f \). Then we have that \( A \) is an antiderivative for \( f \) by XIV.2.9.3., and that
\[
\int_a^b f(x) \, dx = A(b) - A(a) = \int_a^b f(x) \, dx .
\]
and \( A(a) = 0 \). Also, by XIV.2.9.4., there is a constant \( C \) such that \( F(x) = A(x) + C \) for all \( x \in [a, b] \). We then have
\[
F(b) - F(a) = [A(b) + C] - [A(a) + C] = A(b) - A(a) = \int_a^b f(x) \, dx .
\]

XIV.2.9.6. Extensive and detailed applications of this version of the Fundamental Theorem of Calculus to evaluating definite integrals can be found in any Calculus text. Since this is not a Calculus text, we omit discussion of these applications.

XIV.2.9.7. We now give the proof of XIV.2.9.3.

**Proof:** We have, for \( x \in [a, b] \), \( x \neq x_0 \),
\[
A(x) - A(x_0) = \int_a^x f(t) \, dt - \int_a^{x_0} f(t) \, dt = \int_{x_0}^x f(t) \, dt .
\]
We also have
\[
f(x_0) = \frac{1}{x-x_0} \int_{x_0}^x f(x_0) \, dt \]
(since \( f(x_0) \) is just a constant), and so
\[
\frac{A(x) - A(x_0)}{x-x_0} - f(x_0) = \frac{1}{x-x_0} \int_{x_0}^x [f(t) - f(x_0)] \, dt .
\]
Let \( \epsilon > 0 \). Since \( f \) is continuous at \( x_0 \), there is a \( \delta > 0 \) such that, whenever \( x \in [a, b] \) and \( |x-x_0| < \delta \), \( |f(x) - f(x_0)| < \epsilon \). If \( x \in [a, b] \) and \( |x-x_0| < \delta \), then \( |t-x_0| < \delta \) for all \( t \) between \( x_0 \) and \( x \), and hence
\[
\left| \int_{x_0}^x [f(t) - f(x_0)] \, dt \right| < \epsilon |x-x_0|
\]
by (1); so we have
\[
\left| \frac{A(x) - A(x_0)}{x-x_0} - f(x_0) \right| = \frac{1}{|x-x_0|} \left| \int_{x_0}^x [f(t) - f(x_0)] \, dt \right| < \epsilon
\]
1524
and therefore

\[ A'(x_0) = \lim_{x \to x_0} \frac{A(x) - A(x_0)}{x - x_0} = f(x_0) \, . \]

\(\blacksquare\)

**XIV.2.9.8. Examples.** If \( f \) is Riemann-integrable on \([a, b]\), \( A \) is its cumulative area function, and \( f \) is not continuous at \( x_0 \in [a, b] \), then \( A \) may or may not be differentiable at \( x_0 \), and if it is we may or may not have \( A'(x_0) = f(x_0) \):

(i) Let \( f(x) = 0 \) for \( x < 0 \) and \( f(x) = 1 \) for \( x \geq 0 \). Then \( f \) is Riemann-integrable on \([-1, 1]\), and its cumulative area function \( A \) is given by \( A(x) = 0 \) if \(-1 \leq x \leq 0\) and \( A(x) = x \) if \( 0 \leq x \leq 1 \). \( A \) is not differentiable at 0.

(ii) Let \( f(x) = 0 \) for \( x \neq 0 \) and \( f(0) = 1 \). Then \( f \) is Riemann-integrable on \([-1, 1]\), and \( A \) is identically 0. We have that \( A \) is differentiable at 0, but \( A'(0) \neq f(0) \).

(iii) [vRS82, p. 3-4] Let \( f(x) = \sin \left( \frac{1}{x} \right) \) for \( x \neq 0 \) and \( f(0) = 0 \). Then \( f \) is bounded, and continuous except at 0, so \( f \) is Riemann-integrable on \([-1, 1]\) (or any other closed bounded interval). If \( A \) is its cumulative area function, we will show that \( A \) is differentiable at 0 and \( A'(0) = 0 = f(0) \) even though \( f \) is discontinuous at 0. This can be shown directly by brute force, but we can get it more easily indirectly.

We first show that \( f \) has an antiderivative \( F \). Let

\[ G(x) = \begin{cases} x^2 \cos \left( \frac{1}{x} \right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \, . \]

Then \( G \) is differentiable everywhere, and

\[ G'(x) = g(x) = \begin{cases} 2x \cos \left( \frac{1}{x} \right) + \sin \left( \frac{1}{x} \right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} = h(x) + f(x) \, , \]

where

\[ h(x) = \begin{cases} 2x \cos \left( \frac{1}{x} \right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \, . \]

Since \( h \) is continuous, it has an antiderivative \( H \) (XIV.2.9.4.). Then \( f = g - h \) also has an antiderivative \( F = G - H \).

We then have that \( A - F \) is continuous everywhere, differentiable everywhere except possibly at 0, and \( (A - F)'(x) = 0 \) for all \( x \neq 0 \). Thus by V.8.4.1.(iii) \( A - F \) is constant, hence differentiable also at 0; so \( A \) is differentiable at 0 and \( A'(0) = F'(0) = 0 = f(0) \).

More dramatic versions of this example exist: see Exercise V.12.3.3.. For a related example, if \( f \) is a Lebesgue integrable function on \([a, b]\) and \( A \) is its cumulative area function (i.e. \( A(x) \) is the Lebesgue integral of \( f \) from \( a \) to \( x \)), then \( A \) is continuous, differentiable almost everywhere, and \( A' = f \) almost everywhere (). Such an \( f \) need not be continuous anywhere and need not be equal almost everywhere to a function which is Riemann integrable on any subinterval ()

**XIV.2.9.9.** Version XIV.2.9.3. of the Fundamental Theorem of Calculus roughly says that the derivative of the integral is the original function. What about the integral of the derivative? One problem is that the derivative of a differentiable function on an interval need not be Riemann-integrable:
XIV.2.9.10. **Examples.** (i) Let \( F \) be the function
\[
F(x) = \begin{cases} 
  x^2 \sin\left(\frac{1}{x^2}\right) & \text{if } x \neq 0 \\
  0 & \text{if } x = 0
\end{cases}.
\]
Then \( F \) is differentiable everywhere, and
\[
F'(x) = f(x) = \begin{cases} 
  2x \sin\left(\frac{1}{x^2}\right) - \frac{2}{x} \cos\left(\frac{1}{x^2}\right) & \text{if } x \neq 0 \\
  0 & \text{if } x = 0
\end{cases}
\]
which is unbounded near 0 and hence not Riemann-integrable in any interval containing 0 (in fact, it is not even Lebesgue integrable in any interval containing 0).

(ii) Let \( F \) be the function of \((\)\). Then \( F \) is differentiable on \([0;1]\), and \( F' = f \) is bounded there. Since \( f \) is 0 on a dense subset of \([0,1]\), the Riemann integral of \( f \) on any subinterval, if it existed, would have to be 0 (XIV.2.6.6.). Since \( F \) is not constant on any subinterval, it follows from XIV.2.9.11. below that \( f \) is not Riemann-integrable on any subinterval. It even turns out that \( f \) is not equal almost everywhere to a Riemann-integrable function \((.)\).

We do, however, have the following theorem, which extends XIV.2.9.5.:  

**XIV.2.9.11. Theorem.** [**Fundamental Theorem of Calculus, Version 2**] Let \( F \) be a function which is continuous on \([a;b]\) and differentiable except possibly at finitely many points, with derivative \( f \). If \( f \) is Riemann-integrable on \([a;b]\), then for all \( x \in [a,b] \) we have
\[
F(x) = F(a) + \int_a^x f(t) \, dt
\]
and so, in particular,
\[
\int_a^b f(t) \, dt = F(b) - F(a) .
\]

**Proof:** It suffices to prove the second statement; the first statement then results from applying the second to the interval \([a,x]\).

Let \( \epsilon > 0 \), and let \( \mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\} \) be a partition satisfying Definition 1A for this \( \epsilon \), and fine enough that \( \mathcal{P} \) contains all points where \( F \) is not differentiable (such a partition exists by Definition 1B). By the Mean Value Theorem (.), for each \( k \) choose \( t_k \in (x_{k-1},x_k) \) so that
\[
F(x_k) - F(x_{k-1}) = f(t_k)(x_k - x_{k-1}) .
\]
Then, by telescoping, we have
\[
F(b) - F(a) = \sum_{k=1}^n (F(x_k) - F(x_{k-1})) = \sum_{k=1}^n f(t_k)(x_k - x_{k-1})
\]
and so, since the right side is a Riemann sum for \( f \) with respect to \( P \), we have

\[
\left| (F(b) - F(a)) - \int_a^b f(t) \, dt \right| < \epsilon.
\]

Since \( \epsilon > 0 \) is arbitrary, the result follows.

This result generalizes to the case in which \( f \) is only required to be Lebesgue integrable ().

The hypothesis can be relaxed to allow \( F \) to be nondifferentiable at countably many points (if it is continuous ()), but not just differentiable almost everywhere ().

More general versions of the Fundamental Theorem of Calculus will be discussed in (), (), ...

The Average Value of a Function

XIV.2.9.12. Definition. Let \( f \) be Riemann-integrable on \( J = [a;b] \). Then the average value of \( f \) on \( J \) is

\[
\text{av}(f, J) = \frac{1}{b-a} \int_a^b f(x) \, dx.
\]

(Note that this definition also makes sense if \( b < a \), and gives the same value if \( a \) and \( b \) are interchanged.)

It is not hard to motivate this definition as a reasonable version of the average value of a function on an interval, e.g. by considering Riemann sums corresponding to a subdivision of the interval into a large number of subintervals of equal length. One interpretation which is reasonable motivation is that \( \text{av}(f, J) \) is the value of a constant function on \( J \) which has the same integral over \( J \) as \( f \).

XIV.2.9.13. Using this notion of average value, we have an interesting relation between average and instantaneous rates of change. If \( f \) is differentiable on \( [a;b] \) and \( f' \) is Riemann integrable on \( [a;b] \), then by XIV.2.9.11. we have

\[
\frac{1}{b-a} \int_a^b f'(t) \, dt = \frac{f(b) - f(a)}{b-a}
\]

i.e. the average rate of change of \( f \) over \( [a;b] \) (the right-hand side) is equal to the average value of the instantaneous rate of change of \( f \) over \( [a;b] \) (the left-hand side). This can be thought of as a variant of the Mean Value Theorem: If \( f \) is differentiable on \( [a;b] \) and \( f' \) is Riemann integrable on \( [a,b] \), the difference quotient of \( f \) over \( [a,b] \) is the average value of \( f' \) over \( [a,b] \) (the MVT asserts it is the value of \( f' \) at some particular point of \( [a,b] \)).

We then have the following analog of the Mean Value Theorem (actually more than an analog, in light of XIV.2.9.3.).

XIV.2.9.14. Proposition. [Mean Value Theorem for Integrals] Let \( f \) be a continuous function on a closed bounded interval \( J = [a,b] \). Then there is at least one \( c \in (a,b) \) with \( f(c) = \text{av}(f, J) \).
Proof: Let \( m \) and \( M \) be the minimum and maximum of \( f \) on \([a,b]\). Then, if \( \mathcal{P} = \{a,b\} \) is the partition with one subinterval, we have \( \underline{S}(f,\mathcal{P}) = m(b-a) \) and \( \overline{S}(f,\mathcal{P}) = M(b-a) \), so we have

\[
m(b-a) \leq \int_a^b f(x) \, dx \leq M(b-a)
\]

so, dividing through by \( b-a \), we have

\[
m \leq av(f,J) \leq M.
\]

If \( m = M \), then \( f \) is constant on \([a,b]\), and we can take \( c \) to be any number in \((a,b)\). If \( m < M \), we have \( m < av(f,J) < M \) [if \( m = av(f,J) \), set \( g(x) = f(x) - m \); then \( g \geq 0 \) is continuous and \( \int_a^b g(x) \, dx = 0 \), so \( g \) is identically \( 0 \) (), a contradiction. A similar argument using \( h(x) = M - f(x) \) shows that we cannot have \( av(f,J) = M \).] Choose \( x_1, x_2 \in [a,b] \) with \( f(x_1) = m \), \( f(x_2) = M \). Then by the Intermediate Value Theorem () there is a \( c \) in the open interval between \( x_1 \) and \( x_2 \) with \( f(c) = av(f,J) \).

Note that this result can fail if \( f \) is not continuous, e.g. if \( f \) is a step function taking only the values 0 and 1.

Derivatives and Limits

We can use the FTC plus properties of the Riemann integral to give a simpler proof of a crucial special case of V.8.5.5:

XIV.2.9.15. Theorem. Let \((F_n)\) and \((f_n)\) be sequences of functions on an interval \(I\). Suppose

(a) Each \( f_n \) is continuous on \( I \).

(b) Each \( F_n \) is differentiable on \( I \), with \( F_n' = f_n \).

(c) The sequence \((f_n)\) converges u.c. on \( I \) to a function \( f \).

(d) There is an \( x_0 \in I \) such that the sequence \((F_n(x_0))\) converges.

Then:

(i) The sequence \((F_n)\) converges u.c. on \( I \) to a function \( F \).

(ii) \( F \) is differentiable on \( I \) and \( F' = f \).

Proof: Set \( L = \lim_{n \to \infty} F_n(x_0) \). By the FTC2, we have, for each \( n \) and each \( x \in I \),

\[
F_n(x) = F_n(x_0) + \int_{x_0}^x f_n(t) \, dt .
\]

For each fixed \( x \in I \), since \( f_n \to f \) uniformly on the interval between \( x_0 \) and \( x \),

\[
F_n(x) = F_n(x_0) + \int_{x_0}^x f_n(t) \, dt \to L + \int_{x_0}^x f(t) \, dt.
\]
by XIV.2.6.5.; so, if we define, for $x \in I$,

$$F(x) = L + \int_{x_0}^{x} f(t) \, dt$$

we have by the FTC1 that $F$ is differentiable on $I$ and $F' = f$; and $F_n \to F$ pointwise on $I$. It remains to show that $F_n \to F$ u.c. on $I$. Let $[a, b]$ be a closed bounded subinterval of $I$ containing $x_0$. Fix $\epsilon > 0$, and choose $N$ so that $|f_n(x) - f(x)| < \frac{\epsilon}{2(b-a)}$ for all $n \geq N$ and all $x \in [a, b]$, and $|F_n(x_0) - L| < \frac{\epsilon}{2}$ for all $n \geq N$. Then, for $n \geq N$ and $x \in [a, b]$,

$$|F_n(x) - F(x)| \leq |F_n(x_0) - L| + \left| \int_{x_0}^{x} [f_n(t) - f(t)] \, dt \right| \leq |F_n(x_0) - L| + \int_{x_0}^{x} |f_n(t) - f(t)| \, dt$$

$$< \frac{\epsilon}{2} + \frac{\epsilon}{2(b-a)} |x - x_0| \leq \epsilon.$$  

\[ \diamond \]

XIV.2.10. Integration by Substitution and Integration by Parts
XIV.2.11. Improper Riemann Integration

The fact that unbounded functions cannot be Riemann integrable is partially compensated by the various forms of improper Riemann integration.

XIV.2.11.1. Definition. If $f$ is a real-valued function defined on $(a, b]$, and Riemann integrable on $[a + \epsilon, b]$ for every $\epsilon > 0$, then $f$ is improperly Riemann integrable from $a$ to $b$ if

$$\lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right]$$

exists (in the usual sense, as a real number). We then define the improper Riemann integral to be

$$\int_{a}^{b} f(x) \, dx = \lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right].$$

If $f$ is nonnegative on $(a, b]$, then

$$\lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right] = \sup_{\epsilon > 0} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right]$$

always exists (in the regular or extended sense), and in this case, we define

$$\int_{a}^{b} f(x) \, dx = \lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right].$$

A nonnegative function is then improperly Riemann integrable from $a$ to $b$ if $\int_{a}^{b} f(x) \, dx < \infty$. 

1530
XIV.2.12. Numerical Integration and Simpson’s Rule

Usually the best way to compute the exact value of a Riemann integral is via the Fundamental Theorem of Calculus. But sometimes integrals are encountered where this approach is not possible or feasible, and to calculate the integral we need to return to the definition and compute the integral directly as a limit.

We know in theory that if a function is Riemann integrable on an interval, then the Riemann sums for the integral converge to the exact integral as the partitions become sufficiently fine. But how can we choose a Riemann sum, or similar sum, which gives the integral to within a specified accuracy, and how do we know how accurate an approximation we get from a given sum? Numerical integration theory is the study of these problems.

Numerical methods for integration are essentially based on Riemann integration, and do not work satisfactorily unless the function being integrated is well-behaved (piecewise-continuous at the least, and usually with some smoothness). Thus numerical integration is really a part of the theory of Riemann integration and not more advanced integration theories. Numerical integration is frequently used in situations where function values can only be calculated easily at certain points, and even then sometimes only to a certain accuracy, or to functions whose values are known only from data sets.

Numerical integration is an important part of the subject of numerical analysis, a part of mathematics (usually regarded as part of “applied mathematics”) which seeks to develop efficient ways to numerically calculate or approximate various important mathematical quantities, e.g. integrals, sums of infinite series, zeroes of functions, eigenvalues and eigenvectors, solutions to differential equations. Much of numerical analysis has a very different character than the type of analysis developed in this book, but the word analysis is appropriately used for this subject since (among other things) analytic techniques are used in determining the accuracy and efficiency of the numerical algorithms developed. We will only scratch the surface of numerical integration here; books such as [Hil87] or [IK94] have a far more thorough treatment.

Quadrature Procedures

XIV.2.12.1. Suppose we have a fixed interval, which we will for simplicity take to be a closed bounded interval \([a, b]\), and we want to have a fixed procedure for estimating integrals

\[
\int_{a}^{b} f(x) \, dx
\]

for continuous functions \(f\). The obvious procedure is to fix points \(x_0, \ldots, x_n\) in \([a, b]\) (which we will always take to be in increasing order). These will typically be a partition of \([a, b]\), but not necessarily, i.e. the endpoints need not be among the \(x_k\). We will then fix real constants \(\alpha_0, \ldots, \alpha_n\), usually positive (a nonnegative procedure, cf. XIV.2.12.4.). The procedure \(Q = Q(x_0, \ldots, x_n; \alpha_0, \ldots, \alpha_n)\) then gives the approximation

\[
\int_{a}^{b} f(x) \, dx \approx Q(f) = \sum_{k=0}^{n} \alpha_k f(x_k) .
\]

XIV.2.12.2. The first obvious property we should have is that the approximation should be exact for constant functions. This translates into the restriction

\[
\sum_{k=0}^{n} \alpha_k = b - a .
\]
XIV.2.12.3. Another more restrictive condition which is natural is that for any continuous $f$,

$$S(f, \mathcal{P}) \leq Q(f) \leq \mathcal{S}(f, \mathcal{P})$$

for some partition $\mathcal{P}$ of $[a,b]$, where $S(f, \mathcal{P})$ and $\mathcal{S}(f, \mathcal{P})$ are the lower and upper Darboux sums. A procedure satisfying this condition is called regular with respect to $\mathcal{P}$. If $Q$ is regular with respect to the partition of $[a,b]$ given by $x_0, \ldots, x_n$ and the endpoints of the interval if not included among the $x_k$, then $Q$ is called regular.

Some quadrature procedures considered in numerical analysis are regular, but many are not, although most are regular with respect to a coarser partition. The condition $\sum \alpha_k = b - a$ insures that all nonnegative procedures are regular with respect to the trivial partition, since any nonnegative procedure is monotone.

XIV.2.12.4. In numerical work, subtraction, or addition of numbers of different magnitudes, can lead to a serious loss of precision; thus the most useful quadrature procedures are ones where the $\alpha_n$ are all the same sign (positive) and approximately the same size, although these restrictions are not necessary in principle.

XIV.2.12.5. Definition. A quadrature procedure, or numerical integration procedure, on an interval $[a,b]$ is a choice of points $x_0, \ldots, x_n \in [a,b]$ and real numbers $\alpha_0, \ldots, \alpha_n$ with $\sum_{k=0}^n \alpha_k = b - a$. For any real-valued function $f$, the application of the procedure $Q = Q(x_0, \ldots, x_n; \alpha_0, \ldots, \alpha_n)$ to $f$ is the sum

$$Q(f) = \sum_{k=0}^n \alpha_k f(x_k).$$

$Q(f)$ is called the approximation to the integral of $f$ over $[a,b]$ by the procedure $Q$.

Definite integration, especially numerical integration, is often called quadrature because of the standard interpretation of integrals as areas.

XIV.2.12.6. We could also consider more general quadrature procedures of the form

$$Q(f) = \sum_{k=0}^n \alpha_k f(x_k) + \sum_{k=0}^n \beta_k f'(x_k) + \sum_{k=0}^n \gamma_k f''(x_k) + \cdots$$

which are sometimes used in numerical analysis (cf. XIV.2.12.19., XIV.2.12.39., XIV.2.12.51.).

XIV.2.12.7. A procedure by this definition must be carefully distinguished from a numerical integration process or scheme, which is a sequence of quadrature procedures converging to the exact integral for all, or at least nicely behaved, functions. For example, the standard procedures like Simpson’s rule begin by partitioning the interval into $n$ subintervals in some specific way. There is thus one procedure for each fixed $n$, and they form a process as $n \to \infty$.

Note that any sequence $(Q_n)$ of quadrature procedures regular with respect to partitions $\mathcal{P}_n$ with $\|\mathcal{P}_n\| \to 0$ as $n \to \infty$ will converge to the exact integral for all continuous functions (in fact for all Riemann-integrable functions), i.e. will form a numerical integration process. However, for practical reasons the rate of convergence is a critical consideration in numerical analysis.
XIV.2.12.8. It is easily checked that any quadrature procedure $Q$ is linear, i.e.

$$Q(cf + g) = cQ(f) + Q(g)$$

for any functions $f, g$ and scalar $c$; a procedure with all $\alpha_k > 0$ is also monotone, i.e. $Q(f) \leq Q(g)$ if $f \leq g$.

XIV.2.12.9. Examples. (i) The simplest quadrature procedures are the left and right estimates for the Riemann integral $\int_a^b$: if $\{x_0, \ldots, x_n\}$ is a partition of $[a, b]$, set $\alpha_k = x_{k+1} - x_k$ for $0 \leq k < n$ and $\alpha_n = 0$. Then the estimate

$$\sum_{k=0}^{n} \alpha_k f(x_k)$$

is the left estimate $L_n$ for the integral of $f$. If we instead take $\alpha_0 = 0$ and $\alpha_k = x_k - x_{k-1}$ for $1 \leq k \leq n$, we get the right estimate $R_n$.

(ii) More generally, if $\{x_0, \ldots, x_n; t_1, \ldots, t_n\}$ is a tagged partition of $[a, b]$, set $\alpha_k = \Delta x_k = x_k - x_{k-1}$ for $1 \leq k \leq n$. Then $Q(t_1, \ldots, t_n; \alpha_1, \ldots, \alpha_n)$ gives the Riemann sum approximation to the integral for this tagged partition:

$$Q(f) = \sum_{k=1}^{n} f(t_k) \Delta x_k .$$

A numerically good procedure is to take the $x_k$ to be evenly spaced and $t_k$ the midpoint of the $k$'th interval. This procedure is called the *midpoint rule* with $n$ intervals, denoted $M_n$. Note that in this procedure the points where the function is evaluated do not form a partition of the interval, i.e. the endpoints are not included.

(iii) Here is a quadrature procedure which is not just a Riemann sum. We average the left and right estimates from (i), normally using $n$ subintervals of equal length $h = \frac{b-a}{n}$. In this case $\alpha_0 = \alpha_n = \frac{h}{2}$ and $\alpha_k = h$ for $1 \leq k \leq n - 1$.

This procedure is called the *trapezoid rule* with $n$ intervals, for the following reason. It is also obtained by approximating the function within each subinterval by a linear (affine) function, so that the integral is approximated by a sum of areas of trapezoids (Figure XIV.16, from http://tutorial.math.lamar.edu/Classes/CalcII/approximatingDefIntegrals.aspx). We call this procedure $T_n$.

An alternate formula for $T_n$ is

$$T_n(f) = \frac{h}{2} \left[ f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n) \right]$$

where $h = \frac{b-a}{n}$ and $x_k = x_0 + kh$ for $0 \leq k \leq n$.

All these quadrature procedures are regular, as is easily verified.

Simpson’s Rule

XIV.2.12.10. If the graph of $f$ is concave up on $[a, b]$, it is clear that the trapezoid estimate is too large and the midpoint estimate is too small, for any fixed $n$, and vice versa if the graph of $f$ is concave down. Thus we would expect the errors to tend to have opposite sign. And experiments with calculable integrals suggest that the absolute error from the midpoint estimate for a given $n$ is typically about half the absolute
error of the trapezoid estimate for the same \( n \) – see XIV.2.12.44. for a precise statement along these lines. (Some unsophisticated references assert or suggest the absolute midpoint error is exactly half the absolute trapezoid error, but in fact it is rarely exactly half, and in some cases is not even approximately half; cf. Exercise XIV.3.11.5.)

**XIV.2.12.11.** These statements suggest that in many cases, twice the midpoint error should largely cancel with the trapezoid error; thus a weighted average of the two should often be much better than either. So we define

\[
S_{2n}(f) = \frac{2}{3}M_n(f) + \frac{1}{3}T_n(f)
\]

which is *Simpson’s rule* with \( 2n \) intervals.

To put Simpson’s rule into the standard form of a quadrature procedure and explain the use of the index \( 2n \), note that in evaluating \( M_n(f) \) and \( T_n(f) \), the function \( f \) must be evaluated at \( 2n + 1 \) points, the \( n + 1 \) points of the partition and the \( n \) midpoints. These points form a partition of the interval \([a, b]\) into \( 2n \) subintervals of equal length.

**XIV.2.12.12.** Changing notation slightly, if \( n \) is an even number and \( \{x_0, \ldots, x_n\} \) is a partition of \([a, b]\) into \( n \) subintervals of equal length \( h = \frac{b-a}{n} \), we have \( \alpha_0 = \alpha_n = \frac{h}{3} \), \( \alpha_k = \frac{2h}{3} \) for \( k \) even, \( 0 < k < n \), and
\[ \alpha_k = \frac{4h}{3} \text{ for } k \text{ odd. Thus} \]
\[ S_n(f) = \frac{h}{3} \left[ f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n) \right] \]

where \( x_k = x_0 + kh \) for \( 0 \leq k \leq n \).

**XIV.2.12.13.** The Simpson’s rule approximation with \( 2n \) intervals is indeed typically (but not always!) *much* better than the trapezoid or midpoint approximations with \( n \) intervals. The function must, of course, be evaluated at about twice as many points. But \( S_{2n}(f) \) is typically even considerably more accurate than \( T_{2n}(f) \) or \( M_{2n}(f) \). See XIV.2.12.36.–XIV.2.12.52. for an error analysis.

Simpson’s rule \( S_{2n} \) is not a regular quadrature procedure, but it is regular with respect to the partition of \([a, b]\) into \( n \) equal subintervals; hence \( S_{2n}(f) \to \int_a^b f(x) \, dx \) as \( n \to \infty \) for any continuous \( f \) (even every Riemann-integrable \( f \)), as does \( T_n(f) \) and \( M_n(f) \).

**XIV.2.12.14.** There is a quite different way of motivating and arriving at Simpson’s rule; see XIV.2.12.24. and XIV.2.12.26.

**Is There a “Best” Quadrature Procedure?**

**XIV.2.12.15.** We would like to find and use the “best” quadrature procedure. However, it appears obvious that we run into a tradeoff between accuracy and complexity, since it stands to reason (and is “generally” correct) that procedures using more points would tend to be more accurate; thus we should not expect a universal procedure which is “best” in all respects.

But there is a more fundamental sense in which there is no “best” procedure. Given any two quadrature procedures \( Q_1 \) and \( Q_2 \), a continuous function \( f_1 \), even a polynomial, can be constructed such that \( Q_1(f_1) \) is a better approximation to the true integral of \( f_1 \) than \( Q_2(f_1) \), and another continuous function (polynomial) \( f_2 \) such that \( Q_1(f_2) \) is better than \( Q_2(f_2) \). So even beyond questions of complexity of computation there is no universally “best” quadrature procedure.

**XIV.2.12.16.** We will appear to argue that Simpson’s rule is generally the best rule to use for numerical integration, and within the shallow limits of our numerical analysis work here this statement is defensible for many typically encountered integrals, and Simpson’s rule is the most widely used quadrature procedure. However, the reader should be aware that it is much too simplistic to tout Simpson’s rule as the ultimate quadrature procedure; there are many others (e.g. Romberg integration, cf. XIV.2.13.7., or Clenshaw-Curtis Integration, cf. () which are better in certain circumstances. See XIV.2.12.61. for further comments.

**Simple and Composite Quadrature Procedures**

One fairly good measure of the accuracy of a quadrature procedure is how well it approximates the exact integral for polynomials of small order.

**XIV.2.12.17.** Definition. Let \( Q \) be a quadrature procedure on \([a, b] \). Then \( Q \) has *order of precision* \( m \) if
\[
Q(f) = \int_a^b f(x) \, dx
\]
(i.e. $Q$ is exactly correct) whenever $f$ is a polynomial of degree $\leq m$, but not for polynomials of degree $m + 1$.

The term order of precision is standard, but some references use order of exactness which seems more descriptive.

We have built into our definition that a quadrature procedure always has order of precision at least 0.

XIV.2.12.18. A quadrature procedure is simple if it has high order of precision for the number of nodes used. A composite quadrature procedure is one which consists of first dividing the interval $[a, b]$ into a number of smaller intervals and applying a simple quadrature procedure on each subinterval (the simple procedures used on the subintervals are usually, but not always, the same on all subintervals). The order of precision of a composite procedure is at least as great as (usually exactly equal to) the minimum of the order of precision of the simple procedures used on the subintervals. A composite procedure is always regular with respect to the partition into the subintervals on which the simple procedures are applied.

For reasons described in XIV.2.12.4., as well as considerations of efficiency, composite procedures using a simple procedure of lower order on each subinterval are often better in practice than simple procedures of higher order.

Simple quadrature procedures are almost always interpolatory or osculatory:

XIV.2.12.19. Definition. Let $x_0, \ldots, x_n$ be points of $[a, b]$. The interpolatory quadrature procedure on $[a, b]$ for the nodes $x_0, \ldots, x_n$ is the procedure $Q$ defined by

$$Q(f) = \int_a^b p_{0, \ldots, n}(x) \, dx$$

where $p_{0, \ldots, n}$ is the interpolating polynomial of $f$ on $[a, b]$ for these nodes (V.11.1.1.). It follows easily from the Lagrange or Newton form of the interpolating polynomial that an interpolatory quadrature procedure is indeed a quadrature procedure of the form of XIV.2.12.5.

This definition can be extended to the situation where the $x_0, \ldots, x_n$ are not necessarily distinct by using the osculating polynomial (V.11.6.4.): such a quadrature procedure is called osculatory. (An osculatory procedure is of the generalized form of XIV.2.12.6., and can only be used for functions which are sufficiently differentiable.)

XIV.2.12.20. It is immediate that an interpolatory or osculatory quadrature procedure using $n + 1$ nodes has order of precision at least $n$. (Conversely, a quadrature procedure using $n + 1$ nodes of order of precision at least $n$ must by interpolatory or osculatory; but cf. XIV.2.13.6.) If the nodes are well chosen, the order of precision can be greater than $n$ (XIV.2.12.30., XIV.2.12.46.).

XIV.2.12.21. There are thus two general strategies for developing simple quadrature procedures:

(1) First pick nodes $x_0, \ldots, x_n$ in a convenient way, e.g. evenly spaced, and use the interpolatory or osculatory procedure for these nodes.

(2) Fix $n$ and choose the $x_k$ carefully to maximize the order of precision of the corresponding interpolatory or osculatory procedure.
The Newton-Cotes Formulas

**XIV.2.12.22.** We first examine the procedure obtained by choosing a partition of \([a, b]\) into \(n\) subintervals of equal length \(h = \frac{b-a}{n}\), i.e. \(x_k = a + kh\) for \(0 \leq k \leq n\). The corresponding interpolatory procedure is called the Newton-Cotes procedure (or Newton-Cotes formula) of order \(n\), denoted \(NC_n\).

Since an explicit formula for the interpolating polynomial can be obtained, a formula for the procedure \(NC_n\) can be found by integrating this polynomial. The calculation is straightforward but tedious, and we carry it out only for \(n = 1\) and \(n = 2\).

**XIV.2.12.23.** First consider the case \(n = 1\), i.e. \(x_0 = a, x_1 = b\). The interpolating polynomial is

\[ p_{0,1}(x) = f(a) + \frac{f(b) - f(a)}{b-a}(x-a) \]

so we have

\[ NC_1(f) = \int_a^b \left[ f(a) + \frac{f(b) - f(a)}{b-a}(x-a) \right] \, dx \]

\[ = f(a)(b-a) + \frac{1}{2} \frac{f(b) - f(a)}{b-a}(b-a)^2 = \frac{1}{2} [f(a) + f(b)] \]

which is just the trapezoid rule for \(n = 1\), i.e. \(NC_1 = T_1\).

**XIV.2.12.24.** Now consider the case \(n = 2\), i.e. \(x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b\). The interpolating polynomial is

\[ p_{0,1,2}(x) = f(a) + \frac{f(b) - f(a)}{b-a}(x-a) + \frac{f[b, x_1] - f[a, b]}{x_1 - a}(x-a)(x-b) \]

\[ = f(a) + \frac{f(b) - f(a)}{2h}(x-a) + \frac{1}{2h} \left[ \frac{f(b) - f(x_1)}{h} - \frac{f(b) - f(a)}{2h} \right] (x-a)(x-b) \]

\[ = f(a) + \frac{f(b) - f(a)}{2h}(x-a) + \frac{1}{4h^2} [f(b) - 2f(x_1) + f(a)] (x-a)(x-b) \]

so we have

\[ NC_2(f) = \int_a^b \left[ f(a) + \frac{f(b) - f(a)}{2h}(x-a) + \frac{1}{4h^2} [f(b) - 2f(x_1) + f(a)] (x-a)(x-b) \right] \, dx . \]

We have

\[ \int_a^b (x-a)(x-b) \, dx = -\frac{(2h)^3}{6} \]

by making the substitution \(x = a + th\); thus

\[ NC_2(f) = f(a)(2h) + \frac{f(b) - f(a)}{2h} \left( \frac{1}{2} (2h)^2 \right) - \frac{8h^3}{6} \frac{1}{4h^2} [f(b) - 2f(x_1) + f(a)] = \frac{h}{3} [f(a) + 4f(x_1) + f(b)] \]

which is exactly \(S_2(f)\) (XIV.2.12.12.), i.e. \(NC_2\) is exactly the simple Simpson’s rule for \(n = 2\): Simpson’s rule is obtained by approximating the graph of \(f\) by a parabola through the endpoints and midpoint, cf. Figure XIV.17, from http://en.wikipedia.org/wiki/Simpson%27s_rule. (In fact, this is how Simpson came up with the rule in the first place in 1743; the simple Simpson’s rule was previously essentially known to others, including Newton and Kepler, and was apparently first described by Cavalieri in 1639.)
XIV.2.12.25. We list without proof a few more of the Newton-Cotes formulas:

\[ NC_3(f) = \frac{3h}{8} [f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] \]

\[ NC_4(f) = \frac{2h}{45} [7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] \]

\[ NC_5(f) = \frac{5h}{288} [19f(x_0) + 75f(x_1) + 50f(x_2) + 50f(x_3) + 75f(x_4) + 19f(x_5)] \]

\[ NC_8(f) = \frac{4h}{1417} [989f(x_0) + 5888f(x_1) + 928f(x_2) + 10496f(x_3) - 4540f(x_4) + 10496f(x_5) - 928f(x_6) + 5888f(x_7) + 989f(x_8)] \]

XIV.2.12.26. As \( n \) increases, the coefficients in \( NC_n \) vary more greatly. More problematically, if \( n \geq 8 \) the coefficients are not all positive. Thus, for reasons described in XIV.2.12.4., higher Newton-Cotes formulas are not well suited for numerical work. Composite procedures using Newton-Cotes formulas of smaller degree on the subintervals generally work better numerically.

It turns out that if \( n \) is even, \( NC_n \) generally gives almost as good an approximation as \( NC_{n+1} \) (and sometimes better), for reasons described in XIV.2.12.57. In particular, \( NC_2 \) (simple Simpson’s rule) is generally nearly as good as \( NC_3 \) (Simpson’s three-eighths rule), which requires more arithmetic. (In fact, the
composite Simpson’s rule is generally even better than the composite three-eighths rule for the same overall number of subintervals; cf. Exercise XIV.2.13.4.)

The ordinary Simpson’s rule $S_{2n}$ of XIV.2.12.12. is nothing but the composite procedure consisting of dividing $b - a$ into $n$ subintervals of equal length and then applying the simple Simpson’s rule $S_2$ on each subinterval; for this reason, Simpson’s rule is sometimes called the parabolic rule. In part for the reasons in the last two paragraphs, Simpson’s rule is a popular and generally very good procedure.

**XIV.2.12.27.** As a variation, given $n$ we can divide $[a, b]$ into $n + 2$ subintervals of equal length $h = \frac{b - a}{n + 2}$, and not use the two endpoints but only the $n + 1$ interior points $x_1, \ldots, x_{n+1}$, where $x_k = a + kh$ for $1 \leq k \leq n + 1$. The resulting formulas are the open Newton-Cotes formulas, written $NC^n_0$. The only one of interest to us is $NC^n_0$, which is just the simple midpoint rule $M_1$. We list a couple more:

$$NC^n_1(f) = \frac{3h}{2} [f(x_1) + f(x_2)]$$

$$NC^n_2(f) = \frac{4h}{3} [2f(x_1) - f(x_2) + 2f(x_3)]$$

Already the coefficients in $NC^n_2$ are not all positive. $NC^n_0$ is again of order of precision at least $n$.

**Gaussian Quadrature**

**XIV.2.12.28.** We now change indexing convention slightly, beginning with $k = 1$ instead of $k = 0$. The quadrature formula

$$Q(f) = \sum_{k=1}^{n} \alpha_k f(x_k)$$

has $2n$ unknowns, the coefficients $\alpha_k$ and the nodes $x_k$. If we fix the $x_k$, there is a unique choice of the $\alpha_k$ giving a $Q$ of maximum order of precision, which is necessarily the interpolatory procedure and has order of precision $\geq n - 1$ (these $\alpha_k$ can be found by solving a system of linear equations; the technique is called the method of undetermined coefficients). The Newton-Cotes approach was of this form. We could alternately fix the $\alpha_k$ and find the $x_k$ for which the procedure has maximum order of precision, or we could vary both the $\alpha_k$ and $x_k$, keeping only $n$ fixed.

**XIV.2.12.29.** First consider fixing the $\alpha_k$. For reasons discussed in XIV.2.12.4. (a more detailed analysis justifies this further), a good choice is to take all the $\alpha_k$ equal (to $\frac{h}{n}$). This problem can be solved and gives a procedure of order of precision $\geq n$ only for $n \leq 7$ and $n = 9$; if $n$ is even the order of precision is actually $n + 1$. The case $n = 1$ gives the midpoint rule. If $[a, b] = [-1, 1]$ and $n = 2$, the optimal notes are at $\pm \frac{1}{\sqrt{3}} = 0.577 \cdots$. Tables giving the nodes for other $n$ can be found in [IK94] and other references.

**XIV.2.12.30.** Now consider varying both the $\alpha_k$ and the $x_k$. It can be shown that the maximum order of precision is obtained by taking the nodes to be the roots of the unique (up to constant multiple) polynomial $p_n$ of degree $n$ which is orthogonal to the subspace of polynomials of degree $\leq n - 1$ with respect to the usual real inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x) \, dx$$
If \([a, b] = [-1, 1]\), the \(p_n\) are the Legendre polynomials \(\cdot\); other intervals can be handled via affine transformations. The \(\alpha_k\) are then chosen so the procedure is interpolatory. The resulting procedure is called Gaussian quadrature, or Legendre-Gauss quadrature, with \(n\) nodes, denoted \(G_n\). The procedure \(G_n\) has order \(2n - 1\).

Here are formulas for small \(n > 1\) for the interval \([-1, 1]\) (the case \(n = 1\) is just the midpoint rule):

\[
G_2(f) = f \left( -\frac{1}{\sqrt{3}} \right) + f \left( \frac{1}{\sqrt{3}} \right) = f(-.577350 \cdots) + f(.577350 \cdots)
\]

\[
G_3 = \frac{5}{9} f(-\sqrt{0.6}) + \frac{8}{9} f(0) + \frac{5}{9} f(\sqrt{0.6}) = \frac{1}{9} [5f(-0.774597 \cdots) + 8f(0) + 5f(0.774597 \cdots)]
\]

\[
G_4(f) = (0.347855 \cdots) f(-0.861136 \cdots) + (0.652145 \cdots) f(-0.339981 \cdots) + (0.652145 \cdots) f(0.339981 \cdots) + (0.347855 \cdots) f(0.861136 \cdots)
\]

The coefficients and nodes are algebraic numbers, but irrational in general and must be computed by a numerical procedure; see e.g. http://mathworld.wolfram.com/Legendre-GaussQuadrature.html. There are tables available online for at least up to \(n = 64\).

**XIV.2.12.31.** Advantages of Gaussian quadrature are that it is usually very accurate for modest \(n\) if \(f\) is sufficiently smooth (cf. XIV.2.12.60.), and the coefficients are all positive, although not too close to the same magnitude if \(n\) is large (cf. XIV.2.12.4.). Disadvantages are that it can be difficult to calculate the nodes and coefficients for \(n\) large (although there are tables for reasonable \(n\)), and if the process is repeated for a larger \(n\) all the coefficients and nodes must be recalculated (unlike, say, Simpson’s rule if \(n\) is doubled). A method like this also cannot be easily applied to functions arising from data sets.

**Weight Functions and Chebyshev Quadrature**

**XIV.2.12.32.** It is often important to consider a more general version of quadrature, to estimate integrals of the form

\[
I(f) = \int_{a}^{b} f(x) w(x) \, dx
\]

where \(w\) is a fixed weight function on \([a, b]\). Here \(w\) can be any nonnegative function which is Riemann-integrable or improperly Riemann-integrable (or, more generally, Lebesgue integrable) on \([a, b]\), but in applications \(w\) is usually a simple smooth function. We want a quadrature procedure \(Q\) of the form

\[
Q(f) = \sum_{k=1}^{n} \alpha_k f(x_k)
\]

for constants \(\alpha_k\) and nodes \(x_k\). The basic restriction now is

\[
\sum_{k=1}^{n} \alpha_k = \int_{a}^{b} w(x) \, dx.
\]

Usual quadrature is the special case where \(w\) is the constant function 1.
XIV.2.12.33. Gaussian quadrature works similarly in this setting: the optimal nodes are the roots of the polynomial of degree $n$ which is orthogonal to all polynomials of smaller degree with respect to the real inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x)w(x) \, dx$$

and the weights are then found by the method of undetermined coefficients. This works because of the result of Exercise XIV.2.13.5.

XIV.2.12.34. More generally, a Radon measure $\mu$ can be used on $[a, b]$, and the integrals

$$\int_{[a,b]} f \, d\mu$$

approximated by quadrature procedures as above. The optimal polynomial of degree $n$ is the one orthogonal to the polynomials of smaller degree for the inner product

$$\langle f, g \rangle = \int_{[a,b]} fg \, d\mu .$$

It can be shown using the Krein-Milman theorem that such quadrature procedures must converge for any continuous $f$ as $n \to \infty$. However, this generality is of little or no practical use in numerical analysis.

XIV.2.12.35. A particularly important case is the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$ on $[-1, 1]$. Making the substitution $x = \cos \theta$, we have

$$\int_{-1}^1 f(x) \, dx = \int_0^\pi f(\cos(\theta)) \, d\theta$$

and the optimal nodes are the Chebyshev nodes. This case is called Chebyshev quadrature.

**Error Analysis of the Trapezoid Rule**

XIV.2.12.36. We begin with an analysis of the accuracy of the simple trapezoid rule

$$T_1(f) = \frac{h}{2}[f(a) + f(b)]$$

where $h = b - a$. The technique in the other error analyses will be similar. We want to calculate the error

$$E_1^T(f) = \int_a^b f(x) \, dx - T_1(f) = \int_a^b [f(x) - p_{0,1}(x)] \, dx = \int_a^b (x - a)(x - b)f[a, b, x] \, dx$$

by (1). Assume $f$ is $C^2$ on $[a, b]$. Then $g(x) = f[a, b, x]$ is continuous on $[a, b]$ (3), and for each $x \in (a, b)$ we can write

$$f[a, b, x] = \frac{f''(\xi(x))}{2}$$

1541
for some $\xi(x) \in (a, b)$ (V.11.5.9.). Since $(x-a)(x-b)$ does not change sign on $[a, b]$, and $f[a, b, x]$ is continuous on $[a, b]$, we can apply the Mean Value Theorem for integrals (1) to obtain
\[
E_T^T(f) = \int_a^b (x-a)(x-b) f[a, b, x] \, dx = f[a, b, c] \int_a^b (x-a)(x-b) \, dx = -\frac{h^3}{6}
\]
for some $c \in (a, b)$, where $\xi = \xi(c) \in (a, b)$. We calculate
\[
\int_a^b (x-a)(x-b) = -\frac{h^3}{6}
\]
(cf. the proof of XIV.2.12.24.), so we get
\[
E_T^T(f) = -\frac{h^3}{12}f''(\xi)
\]
for some $\xi \in (a, b)$.

**XIV.2.12.37.** Now suppose we do the composite trapezoid rule with $n$ subintervals, each of length $h = \frac{b-a}{n}$. Using the previous analysis on each subinterval, we have
\[
E_n^T(f) = -\frac{h^3}{12}f''(\xi_1) - \frac{h^3}{12}f''(\xi_2) - \cdots - \frac{h^3}{12} = -\frac{(b-a)^3}{12n^3} [f''(\xi_1) + \cdots + f''(\xi_n)]
\]
for some $\xi_k \in (x_{k-1}, x_k)$. If $f$ is $C^2$, we can apply the Intermediate Value Theorem to $f''$ to obtain a $\xi \in (a, b)$ with
\[
f''(\xi_1) + \cdots + f''(\xi_n) = nf''(\xi)
\]
so we obtain a simplified form of the error:
\[
E_n^T(f) = -\frac{(b-a)^3}{12n^2}f''(\xi)
\]
for some $\xi \in (a, b)$. If $|f''|$ is bounded by $K$ on $[a, b]$, we have
\[
|E_n^T(f)| \leq \frac{(b-a)^3}{12n^2} K.
\]

**XIV.2.12.38.** Thus, as $n$ increases, the error in the trapezoid rule is $O(n^{-2})$, so decreases fairly rapidly. The error bound at the end of XIV.2.12.37. is actually conservative in many cases: the sum $|f''(\xi_1) + \cdots + f''(\xi_n)|$ is often much less than $n \cdot \max_{[a, b]} |f''|$, especially if $f''$ varies significantly or changes sign on $[a, b]$; in fact,
\[
f''(\xi_1) + \cdots + f''(\xi_n)
\]
is close to $n$ times the average value of $f''$ on $[a, b]$, i.e.
\[
\frac{n f'(b) - f'(a)}{b-a}
\]
for $n$ large. Thus
\[
E_n^T(f) \approx -\frac{(b-a)^2}{12n^2} |f'(b) - f'(a)|
\]
for $n$ large.

1542
XIV.2.12.39. We can thus add a correction term to the trapezoid rule: set

\[ T_n(f) = T_n(f) + \frac{(b-a)^2}{12n^2} [f'(b) - f'(a)]. \]

The error for this quadrature rule is \( o(n^{-2}) \), and \( O(n^{-4}) \) if \( f \) is \( C^4 \). (It is not always practical to compute \( f'(a) \) and \( f'(b) \), e.g. if \( f \) is given only by a data set. There are approximations possible using higher-order differences which can be used as a substitute.)

XIV.2.12.40. If \( f'(a) = f'(b) \), e.g. if \( f \) is periodic and the integration is over an entire period, then the correction term is zero and we have that the error \( E_n^T(f) \) is \( o(n^{-2}) \) if \( f \) is \( C^2 \), and \( O(n^{-4}) \) if \( f \) is \( C^4 \). Thus the trapezoid rule is especially accurate for periodic functions. In fact, using the Euler-Maclaurin formula () it can be shown that if \( f \) is \( C^r \) and periodic with period \( b-a \), then there is a constant \( C_r \) depending only on \( r \) such that

\[ E_n^T(f) = C_r \frac{(b-a)^{r+1}}{n^r} f^{(r)}(\xi) \]

for some \( \xi \in (a,b) \), and thus \( E_n^T(f) = O(n^{-r}) \). See [IK94] for details.

**Error Analysis of the Midpoint Rule**

XIV.2.12.41. Let us examine the error in the simple midpoint rule

\[ M_1(f) = f(x_1)h \]

where \( h = b-a \) and \( x_1 \) is the midpoint of \([a,b]\). The key trick in getting a good error formula is to note that not only is \( M_1 \) the interpolatory quadrature rule for the set \( \{x_1\} \), but, if \( f \) is \( C^1 \), it is also the osculatory quadrature rule for the nodes \( (x_1,x_1) \): the osculating polynomial is \( p(x) = f(x_1) + f'(x_1)(x-x_1) \), and

\[ \int_a^b [f(x_1) + f'(x_1)(x-x_1)] \, dx = f(x_1)(b-a) + \frac{1}{2} f'(x_1)((b-x_1)^2 - (a-x_1)^2) \]

and the second term is zero. Thus the error is

\[ E_1^M(f) = \int_a^b f(x) \, dx - M_1(f) = \int_a^b (x-x_1)^2 f[x_1, x_1, x] \, dx. \]

Assume \( f \) is \( C^2 \) on \([a,b]\). Then \( g(x) = f[x_1, x_1, x] \) is continuous on \([a,b]\) so by the MVT for integrals, there is a \( c \in (a,b) \) such that

\[ \int_a^b (x-x_1)^2 f[x_1, x_1, x] \, dx = f[x_1, x_1, c] \int_a^b (x-x_1)^2 \, dx. \]

We also have by () that

\[ f[x_1, x_1, c] = \frac{f''(\xi)}{2} \]

for some \( \xi \in (a,b) \). We calculate

\[ \int_a^b (x-x_1)^2 \, dx = \frac{1}{3} [(b-x_1)^3 - (a-x_1)^3] = \frac{1}{3} \cdot 2 \left( \frac{h}{2} \right)^3 = \frac{h^3}{12} \]

1543
so we have
\[ E_1^M(f) = \frac{h^3}{24} f''(\xi) \]
for some \( \xi \in (a, b) \).

**XIV.2.12.42.** Now consider the composite midpoint rule with \( n \) subintervals of length \( h = \frac{b-a}{n} \). If \( f \) is \( C^2 \) on \([a, b]\), we have
\[
E_n^M(f) = \frac{h^3}{24} f''(\xi_1) + \cdots + \frac{h^3}{24} f''(\xi_n) = \frac{(b-a)^3}{24n^3} \left[ f''(\xi_1) + \cdots + f''(\xi_n) \right]
\]
for some \( \xi_k \in (x_{k-1}, x_k) \). By the Intermediate Value Theorem there is a \( \xi \in (a, b) \) with \( f''(\xi_1) + \cdots + f''(\xi_n) = nf''(\xi) \), so we get the clean formula
\[
E_n^M(f) = \frac{(b-a)^3}{24n^2} f''(\xi)
\]
for some \( \xi \in (a, b) \). If \( |f''| \) is bounded by \( K \) on \([a, b]\), we get
\[
|E_n^M(f)| \leq \frac{(b-a)^3}{24n^2} K .
\]

**XIV.2.12.43.** As with the trapezoid rule, the absolute error estimate for the midpoint rule is conservative, and we may add a correction term: set
\[
\tilde{M}_n(f) = M_n(f) - \frac{(b-a)^2}{24n^2} [f'(b) - f'(a)] .
\]
This has error \( O(n^{-4}) \) if \( f \) is \( C^4 \). If \( f'(a) = f'(b) \), e.g. if \( f \) is periodic and the integration is over an entire period, then the correction term is zero and we have that the error \( E_n^M(f) \) is \( O(n^{-4}) \) if \( f \) is \( C^4 \). (In the periodic case, the midpoint rule, the trapezoid rule, and the left and right approximations are effectively all the same procedure.)

**XIV.2.12.44.** Comparing XIV.2.12.41. with XIV.2.12.36. and XIV.2.12.42. with XIV.2.12.37., we see that if \( f \) is \( C^2 \), the error in the midpoint rule is indeed approximately of opposite sign and half the absolute value of the error in the trapezoid rule for the same \( n \) (cf. XIV.2.12.10.). The statement is not exactly true since the \( \xi \)'s are not the same in general; but if \( n \) is chosen large enough that \( f'' \) varies little over each subinterval, the \( f''(\xi_k) \)'s are almost equal for each \( k \), so the statement is very nearly true. Thus for such an \( n \) we have that Simpson's rule \( S_{2n} \) is an extremely good approximation to the integral. See XIV.2.12.49..

**XIV.2.12.45.** Another consequence of the error formula (which is pretty obvious just from the definition) is that even though the simple midpoint rule is an interpolatory procedure with only one node, it has order of precision 1, not 0; it gives the exact answer for linear functions, not just constant functions. The simple trapezoid rule also has order of precision 1, as expected since it is an interpolatory procedure with two nodes. The composite procedures also have order of precision 1.
Error Analysis of Simpson’s Rule

**XIV.2.12.46.** Consider the simple Simpson’s rule $S_2$, where $h = \frac{b-a}{3}$. We first observe that $S_2$ actually has order of precision $\geq 3$ instead of the expected 2. To show this, by linearity it suffices to show that $S_2(f) = \int_a^b f(x) \, dx$ for $f(x) = x^3$, which is a simple calculation:

$$S_2(f) = \frac{h}{3} [a^3 + 4x_1^3 + b^3] = \frac{b-a}{6} \left[ a^3 + 4 \left( \frac{a+b}{2} \right)^3 + b^3 \right]$$

$$= \frac{b-a}{6} \left[ a^3 + \frac{1}{2}(a+b)^3 + b^3 \right] = \frac{1}{6} (b-a) \left[ a^3 + \frac{1}{2}a^3 + \frac{3}{2}a^2b + \frac{3}{2}ab^2 + \frac{1}{2}b^3 + b^3 \right]$$

$$= \frac{1}{4} (b-a)(a^3 + a^2b + ab^2 + b^3) = \frac{1}{4} (b^4 - a^4) = \int_a^b x^3 \, dx .$$

(It is easily checked the same way that $S_2$ is not exact on $f(x) = x^4$, so the order of precision of $S_2$ is exactly 3.)

**XIV.2.12.47.** Now suppose $f$ is $C^1$, and let $p$ be the osculating polynomial for $f$ with nodes $x_0, x_1, x_1, x_2$. Then the interpolating polynomial for $p$ for the nodes $x_0, x_1, x_2$ is the same polynomial $q$ as the interpolating polynomial for $f$ with these same nodes. Since $p$ is cubic, we have

$$\int_a^b p(x) \, dx = S_2(p) = \int_a^b q(x) \, dx = S_2(f) .$$

Thus, for $C^1$ functions, $S_2$ is the osculating quadrature procedure with nodes $x_0, x_1, x_1, x_2$.

**XIV.2.12.48.** As a result, if $f$ is $C^1$ we get a formula for the error in $S_2$:

$$E_2^S(f) = \int_a^b [f(x) - p(x)] \, dx = \int_a^b (x-x_0)(x-x_1)^2(x-x_2)f[x_0, x_1, x_1, x_2, x] \, dx .$$

We proceed as before. Suppose $f$ is $C^1$ on $[a, b]$. The function $g(x) = f[x_0, x_1, x_1, x_2, x]$ is continuous on $[a, b]$, and $(x-x_0)(x-x_1)^2(x-x_2)$ does not change sign on $[a, b]$, so by the MVT for integrals we have

$$E_2^S(f) = \int_a^b (x-x_0)(x-x_1)^2(x-x_2)f[x_0, x_1, x_1, x_2, x] \, dx = f[x_0, x_1, x_1, x_2, c] \int_a^b (x-x_0)(x-x_1)^2(x-x_2) \, dx$$

for some $c \in (a, b)$.

Next we calculate

$$\int_a^b (x-a)(x-x_1)^2(x-b) \, dx .$$

Make the substitution $x = x_0 + th$; then

$$\int_a^b (x-a)(x-x_1)^2(x-b) \, dx = h^5 \int_0^2 t(t-1)^2(t-2) \, dt = -\frac{4h^5}{15} .$$

1545
Now suppose $f$ is $C^4$ on $[a, b]$. Then

$$f[x_0, x_1, x_1, x_2, c] = \frac{f^{(4)}(\xi)}{4!}$$

for some $\xi \in (a, b)$. Thus we obtain

$$E_n^S(f) = -\frac{4h^5}{15} \frac{f^{(4)}(\xi)}{24} = -\frac{h^5}{90} f^{(4)}(\xi)$$

for some $\xi \in (a, b)$.

**XIV.2.12.49.** Now consider the composite Simpson's rule with $n$ subintervals of length $h = \frac{b-a}{n}$, where $n = 2m$ is even. If $f$ is $C^4$ on $[a, b]$, we have, applying $S_2$ on each pair of successive intervals,

$$E_n^S(f) = -\frac{h^5}{90} f^{(4)}(\xi_1) - \cdots - \frac{h^5}{90} f^{(4)}(\xi_m) = -\frac{(b-a)^5}{90n^5} [f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)]$$

for some $\xi_k \in (x_{2k-2}, x_{2k})$. By the Intermediate Value Theorem there is a $\xi \in (a, b)$ with $f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m) = m f^{(4)}(\xi)$, so we get the clean formula

$$E_n^M(f) = -\frac{(b-a)^5}{180n^4} f^{(4)}(\xi)$$

for some $\xi \in (a, b)$. If $|f^{(4)}|$ is bounded by $K$ on $[a, b]$, we get

$$|E_n^S(f)| \leq \frac{(b-a)^5}{180n^4} K .$$

**XIV.2.12.50.** Thus, as $n$ increases, the error in Simpson’s rule is $O(n^{-4})$, so decreases rapidly. The error bound at the end of XIV.2.12.49. is actually conservative in many cases: the sum

$$|f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)|$$

is often much less than $m \cdot \max_{[a, b]} |f^{(4)}|$, especially if $f^{(4)}$ varies significantly or changes sign on $[a, b]$; in fact,

$$f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)$$

is close to $m$ times the average value of $f^{(4)}$ on $[a, b]$, i.e.

$$m \frac{f^{(3)}(b) - f^{(3)}(a)}{b - a}$$

for $n$ large. Thus

$$E_n^S(f) \approx -\frac{(b-a)^4}{180n^4} [f^{(3)}(b) - f^{(3)}(a)]$$

for $n$ large. See also XIV.2.12.44.
XIV.2.12.51. We can thus add a correction term to Simpson’s rule: if \( n \) is even, set
\[
\tilde{S}_n(f) = S_n(f) + \frac{(b-a)^4}{180n^4}[f^{(3)}(b) - f^{(3)}(a)] .
\]
The error for this quadrature rule is \( o(n^{-4}) \) if \( f \) is \( C^4 \), and \( O(n^{-6}) \) if \( f \) is \( C^6 \). (It is not always practical to compute \( f^{(3)}(a) \) and \( f^{(3)}(b) \), e.g. if \( f \) is given only by a data set. There are approximations possible using higher-order differences which can be used as a substitute.)

XIV.2.12.52. If \( f^{(3)}(a) = f^{(3)}(b) \), e.g. if \( f \) is periodic and the integration is over an entire period, then the correction term is zero and we have that the error \( E_{n, S}^S(f) \) is \( O(n^{-6}) \) if \( f \) is \( C^6 \). Thus, like the trapezoid rule, Simpson’s rule is especially accurate for periodic functions. But because of XIV.2.12.40., there is little practical reason to use Simpson’s rule instead of the trapezoid rule for periodic functions.

XIV.2.12.53. For functions which are not smooth, the trapezoid rule is actually often a little better than Simpson’s rule for the same number of subintervals. See [CUN02] for a detailed analysis of the errors for both methods for classes of nonsmooth functions.

**Error Analysis for Newton-Cotes Procedures**

XIV.2.12.54. Error formulas for the higher-order Newton-Cotes formulas and open Newton-Cotes formulas can be obtained by essentially the same arguments, although the details are more complicated beyond Simpson’s rule; see e.g. [IK94]. The result is:

XIV.2.12.55. **Theorem.** Let \( f \) be a function on \([a, b]\), and \( n \in \mathbb{N} \). Set \( h = \frac{b-a}{n} \).

(i) Suppose \( n \) is odd and \( f \) is \( C^{n+1} \) on \([a, b]\). Then
\[
E_{n, NC}^C(f) = \int_a^b f(x) \, dx - NC_n(f) = \frac{M_n}{(n+1)!} h^{n+2} f^{(n+1)}(\xi)
\]
for some \( \xi \in (a, b) \), where
\[
M_n = \int_0^n \pi_n(t) \, dt = \int_0^n t(t-1) \cdots (t-n) \, dt < 0 .
\]

(ii) Suppose \( n \) is even and \( f \) is \( C^{n+2} \) on \([a, b]\). Then
\[
E_{n, NC}^C(f) = \int_a^b f(x) \, dx - NC_n(f) = \frac{M_n}{(n+2)!} h^{n+3} f^{(n+2)}(\xi)
\]
for some \( \xi \in (a, b) \), where
\[
M_n = \int_0^n t\pi_n(t) = \int_0^n t^2(t-1) \cdots (t-n) \, dt < 0 .
\]

These formulas are consistent with the cases \( n = 1 \) (trapezoid rule) and \( n = 2 \) (Simpson’s rule) developed in XIV.2.12.36. and XIV.2.12.46.
XIV.2.12.56. **Corollary.** If \( n \) is even, the Newton-Cotes procedure \( NC_n \) has order of precision \( n + 1 \); if \( n \) is odd, \( NC_n \) has order of precision \( n \).

XIV.2.12.57. Thus, if \( n \) is even, The Newton-Cotes procedure \( NC_n \) has essentially the same accuracy as \( NC_{n+1} \) (the accuracy of \( NC_{n+1} \) is roughly better by a constant factor). Thus Newton-Cotes procedures are “better” for even \( n \). The “explanation” is that if \( n \) is even, the procedure \( NC_n \), which is the interpolatory procedure for the evenly spaced nodes \( x_0, \ldots, x_n \), turns out to be also the osculatory procedure for the \( n + 2 \) nodes \( x_0, \ldots, x_{n/2}, x_{n/2}, \ldots, x_n \), with \( x_{n/2} \) repeated, as in the case of the midpoint rule and Simpson’s rule.

XIV.2.12.58. There are similar results for the open Newton-Cotes formulas \( NC_o^n \), which also gain an extra order of precision for \( n \) even (e.g. for the midpoint rule \( NC_0^0 \)). In this case the constants are positive.

XIV.2.12.59. These error formulas are for simple Newton-Cotes procedures. Similar error formulas for compound Newton-Cotes procedures can be obtained in exactly the same manner as for the trapezoid, midpoint, and Simpson’s rule cases.

**Error Analysis for Gaussian Quadrature**

XIV.2.12.60. The error for Gaussian quadrature \( G_n \) on \([a, b]\) may be calculated in a similar manner: if \( f \) is \( C^{2n} \) on \([a, b]\), then

\[
E_n^G(f) = \int_a^b f(x) \, dx - G_n(f) = \frac{2}{(2n + 1)!} \left[ \frac{2^n(n!)^2}{(2n)!} \right]^2 (b - a)^{2n+1} f^{(2n)}(\xi)
\]

for some \( \xi \in (a, b) \). Using Stirling’s formula (\() to approximate the factorials, we get, for \( n \) large,

\[
E_n^G(f) \approx \frac{\pi n(b - a)^{2n+1}}{2^{2n-1}(2n + 1)!} f^{(2n)}(\xi)
\]

for some \( \xi \in (a, b) \). Thus if the derivatives of \( f \) do not grow too rapidly (e.g. if \( f \) is complex-analytic on a large enough region containing \([a, b]\)), the Gaussian quadrature approximations converge rapidly as \( n \to \infty \). See e.g. [IK94] for details.

XIV.2.12.61. In practice, knowing when a procedure gives a sufficiently accurate estimate is not easy. There are procedures which carefully combine results of a standard procedure done twice for different partitions to greatly improve accuracy and make it easier to know when to stop. A theoretically and computationally appealing such improvement of the trapezoid rule is Romberg integration (Exercise XIV.2.13.7.). A widely-used similar variation of Gaussian quadrature is Gauss-Kronrod integration.

There are also adaptive procedures for numerical integration, where the interval lengths and procedures are not fixed in advance but the details vary according to properties of the function being integrated. These are generally the best numerical procedures when feasible. Most software packages for numerical integration use some combination or variation of these procedures; further discussion is beyond our scope, but can be found in numerical analysis texts.
XIV.2.13. Exercises

XIV.2.13.1. Here are some additional relations among standard integration procedures on a fixed interval \([a, b]\).
(a) Show that \(T_{2n} = \frac{1}{2} (T_n + M_n)\) for any \(n\).
(b) Show that \(S_{2n} = \frac{1}{3} (4T_n - T_n)\) for any \(n\).

XIV.2.13.2. (a) Consider 
\[
\int_{-m}^{m} \frac{1}{x^2 + 1} \, dx.
\]
Compute the exact value, and the trapezoid and midpoint estimates using one subinterval. Show that the absolute trapezoid error is less than the absolute midpoint error if \(m > 1.44\), and the trapezoid estimate is much better than the midpoint estimate for \(m\) large.
(b) Consider 
\[
\int_{0}^{1} x^m \, dx.
\]
Compute the exact value, and the trapezoid and midpoint estimates using one subinterval. Show that the midpoint estimate is much better than the trapezoid estimate for \(m\) large.

Actually, none of the estimates are very good in these cases, as may be expected when only one interval is used. In (b), at least the midpoint errors approach 0 as \(m \to +\infty\), while the absolute trapezoid error increases as \(m\) increases. In (a), both the absolute trapezoid and midpoint errors increase as \(m\) increases, but the trapezoid errors are bounded and the midpoint errors unbounded as \(m \to +\infty\).

XIV.2.13.3. Consider the integral 
\[
\int_{-1}^{1} |x| \, dx.
\]
Show that either the trapezoid rule or the midpoint rule for \(n\) subintervals gives the exact answer if \(n\) is even (even \(n = 2\)), and not if \(n\) is odd. Thus the approximations do not necessarily become monotonically better as \(n\) increases (although they do converge to the exact integral as \(n \to \infty\)).

This function is not differentiable everywhere. But it can be uniformly approximated arbitrarily closely by a polynomial () for which the same statements are approximately true; in particular, approximations using a relatively small even number of intervals, even 2, are better than approximations of the same kind using a larger odd number of subintervals.

It can be shown that if \(f\) is a polynomial, then the absolute error \(E_T^n(f)\) monotonically decreases for sufficiently large \(n\), but this is not always true if \(f\) is merely smooth. See http://mathoverflow.net/questions/180542/monotonicity-of-trapezoid-approximations.

XIV.2.13.4. Simpson’s Three-Eighths Rule. Fix an interval \([a, b]\).
(a) Show that if \(f\) is a function on \([a, b]\), then
\[
NC_3(f) = \frac{3h}{8} [f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)]
\]
where \(h = \frac{b-a}{3}\).
(b) Show that if \( f \) is \( C^4 \) on \([a, b]\), then

\[
E^{NC}_3(f) = -\frac{3}{80} h^5 f^{(4)}(\xi)
\]

for some \( \xi \in (a, b) \).

(c) Compare \( E^{NC}_3(f) \) with \( E^{NC}_2(f) = E^{S}_2(f) \) (XIV.2.12.46). [Caution: the \( h \)'s are different in the two cases, and the \( \xi \)'s are also normally different.]

(d) If \( n = 3m \) is divisible by 3, let \( R_n \) be the composite quadrature procedure obtained by dividing \([a, b]\) into \( m \) equal subintervals and applying \( NC_3 \) on each (so \([a, b]\) is divided into \( n \) equal subintervals of length \( h = \frac{b-a}{n} \)). Show that if \( f \) is \( C^4 \) on \([a, b]\), then

\[
E_R^n(f) = \int_a^b f(x) \, dx - R_n(f) = -\frac{3}{80} \frac{(b-a)^5}{n^5} [f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)] = -\frac{1}{80} \frac{(b-a)^5}{n^4} f^{(4)}(\xi)
\]

for some \( \xi_k \in (x_{3k-3}, x_{3k}) \) and some \( \xi \in (a, b) \).

(e) Suppose \( f \) is \( C^4 \) on \([a, b]\), and \( n \) is divisible by 6 and large enough that \( f^{(4)} \) is almost constant on each subinterval. Show that \( |E^{S}_n(f)| \) is “typically” smaller than \( |E^n_R(f)| \).

XIV.2.13.5. Let \( w \) be a nonnegative Riemann-integrable function on \([a, b]\) with \( \int_a^b w(x) \, dx > 0 \). Define a real inner product on the set of polynomials (with real coefficients) by

\[
\langle p, q \rangle = \int_a^b p(x)q(x)w(x) \, dx .
\]

(a) Show that for each \( n \in \mathbb{N} \) there is a unique (up to scalar multiple) nonzero polynomial \( p \) of degree \( n \) which is orthogonal to all polynomials of degree \( \leq n - 1 \) under this inner product.

(b) Suppose \( p \) changes sign exactly \( m \) times in \([a, b]\), at \( x_1, \ldots, x_m \) in \((a, b)\), i.e. \( x_1, \ldots, x_m \) are the distinct roots of \( p \) of odd order in \((a, b)\). Show that \( p \) cannot be orthogonal to

\[
q(x) = (x - x_1)(x - x_2)\cdots(x - x_m)
\]

(\( q(x) = 1 \) if \( m = 0 \)) since \( pq \) does not change sign on \([a, b]\). Thus \( m = \deg(q) \geq n \).

(c) Conclude that all the roots of \( p \) are real and distinct and lie in \((a, b)\).

XIV.2.13.6. [Ham86, 17.7] Consider the following quadrature procedure for twice-differentiable functions on \([-1, 1] \):

\[
Q(f) = \frac{1}{21} [5f(-1) + 32f(0) + 5f(1)] - \frac{1}{315} [f''(-1) + 32f''(0) + f''(1)] .
\]

(a) Show that \( Q \) has order of precision 6.

(b) Show that \( Q \) is not an osculating procedure.
XIV.2.13.7. Romberg Integration. Fix an interval $[a, b]$. Let $T^n_0$ be the trapezoid rule $T_n$ for any $n$. Recursively define

$$T^n_k = \frac{1}{4^k - 1} (4^k T^{k-1}_n - T^{k-1}_n)$$

for all $n$ and $k$.

(a) Show that $T^n_1 = S_{2n}$ for any $n$.

(b) Show that $T^n_4$ is $NC_{4, n}$, the process of dividing $[a, b]$ into $n$ subintervals of equal length and applying $NC_4$ on each subinterval.

(c) The simple relation of (a) and (b) with the Newton-Cotes formulas does not persist. Show that

$$T^n_3 = \frac{4h}{2835} (217f(x_0) + 1024f(x_1) + 352f(x_2) + 1024f(x_3) + 436f(x_4) + 1024f(x_5) + 352f(x_6) + 1024f(x_7) + 217f(x_8))$$

($h = \frac{b-a}{8}$) which is not the same as $NC_8$ (however, the coefficients of $T^n_3$ are all positive, which is desirable computationally, and the sizes of the coefficients do not vary greatly).

The simple relation of (a) and (b) with the Newton-Cotes formulas does not persist. Show that

$$T^n_3 = \frac{4h}{2835} (217f(x_0) + 1024f(x_1) + 352f(x_2) + 1024f(x_3) + 436f(x_4) + 1024f(x_5) + 352f(x_6) + 1024f(x_7) + 217f(x_8))$$

($h = \frac{b-a}{8}$) which is not the same as $NC_8$ (however, the coefficients of $T^n_3$ are all positive, which is desirable computationally, and the sizes of the coefficients do not vary greatly).

The Romberg integration procedure is as follows. Fix an $n$ (in practice $n = 1$ is often used). Successively calculate $T^n_k$ for $k \leq 2^m$. The results can be conveniently collected in a triangular array

$$\begin{array}{cccc}
T^0_0 & T^1_0 & T^2_0 & T^3_0 \\
T^n_0 & T^n_1 & T^n_2 & T^n_3 \\
\vdots & \vdots & \vdots & \vdots \\
\end{array}$$

where each entry is computed from the one to the left and the one diagonally above to the left (the entries in the first column are computed from the trapezoid rule using additional function evaluations; note that $T^n_{2m+1}$ can be calculated from $T^n_{2m}$ by only $2^m n$ additional function evaluations). In practice the entries are rounded to the desired accuracy and the process stops when two identical numbers to the desired accuracy are obtained. The process is computationally efficient and frequently produces the desired accuracy with fewer function evaluations than Simpson’s Rule. The convergence rate is comparable to the rate using correction terms (XIV.2.12.39.), and is more computationally efficient and avoids making a careful estimate of the starting $n$ needed. See books on Numerical Analysis for details and error analyses.
XIV.3. Jordan Content and Riemann Integration in $\mathbb{R}^n$

Generalization of Riemann integration to $\mathbb{R}^n$ is in many ways straightforward, but there are some significant technicalities to overcome. One is to nail down what kinds of sets functions can be reasonably integrated over, and the answer is considerably more complicated than in the one-dimensional case where intervals are the obvious answer. There is a natural collection of subsets called Jordan regions which work nicely for Riemann integration and which are general enough to cover most cases of practical interest. As a special case of integration, we also have a natural notion of “volume,” called Jordan content, for Jordan regions.

XIV.3.1. Integration over Rectangles

It is most straightforward to integrate over rectangles in $\mathbb{R}^n$ whose edges are parallel to the axes. The following definition will be used throughout this section (a more general definition is used in (1)).

**XIV.3.1. Definition.** A rectangle in $\mathbb{R}^n$ is a set $R$ of the form $I_1 \times \cdots \times I_n$, where each $I_k$ is a closed bounded interval $[a_k, b_k]$ or degenerate interval (if $a_k = b_k$) in $\mathbb{R}$, i.e.

$$ R = \{ (x_1, \ldots, x_n) : a_k \leq x_k \leq b_k \text{ for all } k \} . $$

The rectangle is degenerate if $I_k$ is degenerate for at least one $k$. The content (n-dimensional content), or volume (n-dimensional volume) of $R$ is

$$ V(R) = \prod_{k=1}^{n} \ell(I_k) = \prod_{k=1}^{n} (b_k - a_k) . $$

Such rectangles should more properly be called closed coordinate rectangles or closed aligned rectangles, since there are more general sets which can be called rectangles. However, rectangles as defined in XIV.3.1. are exclusively used in developing Riemann integration and it is pedantic to have to constantly qualify them.

XIV.3.1.2. We use the term “rectangle” in any number of dimensions even though it is only geometrically accurate if $n = 2$; in $\mathbb{R}^3$ a rectangle by our definition is what is usually called a rectangular solid, and for $n > 3$ there is not a standard geometric name. Note that the case $n = 1$ is not excluded; a rectangle in $\mathbb{R}$ is just a closed bounded interval or single point. Similarly, “volume” is only strictly appropriate if $n = 3$, which is why we prefer the term “content”: if $n = 2$ content is usually called “area” (and for $n = 1$ it is “length”). We will, however, use the notation $V(R)$ which suggests the term “volume” no matter what $n$ is. To emphasize the dimension, we can (but rarely will) write $V_n$ in place of $V$.

Note that the $n$-dimensional content of a degenerate rectangle is always 0. A degenerate rectangle is geometrically $m$-dimensional for some $m < n$ (the number of nondegenerate coordinate intervals) and so has an $m$-dimensional “content”; thus we must exercise some care in fixing $n$ and interpreting “content” to be $n$-dimensional content for this $n$.

It is a matter of taste whether to regard $\emptyset$ as a rectangle; whether we do or not necessarily leads to some notational complications, which are greater if we do. We will not.

The next observation is obvious from the fact that the intersection of two intervals in $\mathbb{R}$ is an interval or degenerate interval.
XIV.3.1.3. **Proposition.** If $R$ and $S$ are rectangles in $\mathbb{R}^n$, and $R \cap S \neq \emptyset$, then $R \cap S$ is also a rectangle.

**Proof:** If $R = I_1 \times \cdots \times I_n$ and $S = J_1 \times \cdots \times J_n$, then $R \cap S = (I_1 \cap J_1) \times \cdots \times (I_n \cap J_n)$.

XIV.3.1.4. **Definition.** Let $R$ and $S$ be rectangles in $\mathbb{R}^n$. Then $R$ and $S$ are nonoverlapping if either $R \cap S = \emptyset$ or $R \cap S$ is a degenerate rectangle.

### Partitions and Grids

XIV.3.1.5. **Definition.** Let $R$ be a rectangle in $\mathbb{R}^n$. A partition of $R$ is a set $\mathcal{P} = \{R_1, \ldots, R_m\}$ of pairwise nonoverlapping rectangles in $\mathbb{R}^n$ with $R = \bigcup_{k=1}^m R_k$.

A partition $\mathcal{Q}$ is a refinement of $\mathcal{P}$ if every rectangle in $\mathcal{Q}$ is contained in a rectangle of $\mathcal{P}$.

A grid in $R$ is a partition of $R = I_1 \times \cdots \times I_n$ of the form $\mathcal{G} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$, where $\mathcal{P}_k$ is a partition of $I_k$.

Note that a partition of a rectangle is not necessarily a grid. See Figure XIV.18. It is convenient to allow degenerate rectangles in partitions, as our definition does; thus a partition of an interval in $\mathbb{R}$ by this definition is slightly more general than the definition in (1), since repeated points are allowed. This will make no practical difference but simplifies some notation. We could develop the whole theory just using grids, but it is convenient to allow partitions which are not grids. Every partition has a refinement which is a grid (see the proof of XIV.3.1.6).

By XIV.3.1.3., any two partitions of a rectangle have a common refinement obtained by intersecting the subrectangles. If the partitions are grids, so is the common refinement constructed in this way.

Note that there is no natural ordering of the rectangles in a partition, even a grid, if $n > 1$. We have notationally fixed an ordering, but the ordering is not part of the definition of the partition.

Here is a technicality which may be considered geometrically “obvious” (but see the discussion in I.5.1.1.); however, it needs proof (if $n > 1$). This is one of the fundamental observations which gets Jordan content and Riemann integration theory in $\mathbb{R}^n$ off the ground.

XIV.3.1.6. **Proposition.** Let $R$ be a rectangle in $\mathbb{R}^n$. Suppose $R$ is a union of a finite number of pairwise nonoverlapping subrectangles $R = \bigcup_{k=1}^m R_k$. Then

$$V(R) = \sum_{k=1}^m V(R_k).$$

To prove XIV.3.1.6., we will first prove a special case where the subdivision is a grid:
XIV.3.1.7. Lemma. Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be a rectangle in $\mathbb{R}^n$. For each $k$ let $\mathcal{P}_k$ be a partition of $[a_k, b_k]$, and let $\mathcal{G} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the corresponding grid. Let $\{R_1, \ldots, R_p\}$ be the subrectangles in the grid. Then

$$V(R) = \sum_{k=1}^p V(R_k).$$

Proof: We write the proof for $n = 2$; the proof for general $n > 1$ is nearly identical but notationally offensive. So we have $R = [a_1, b_1] \times [a_2, b_2]$, $\mathcal{P}_1 = \{c_0, \ldots, c_r\}$, $\mathcal{P}_2 = \{d_0, \ldots, d_s\}$. Set

$$R_{jk} = [c_{j-1}, c_j] \times [d_{k-1}, d_k]$$

so

$$V(R_{jk}) = (c_j - c_{j-1})(d_k - d_{k-1}).$$

We then have

$$\sum_{j,k} V(R_{jk}) = \sum_{j=1}^r \left[ \sum_{k=1}^s (c_j - c_{j-1})(d_k - d_{k-1}) \right]$$

$$= \sum_{j=1}^r \left[ (c_j - c_{j-1}) \sum_{k=1}^s (d_k - d_{k-1}) \right] = \sum_{j=1}^r (c_j - c_{j-1})(b_2 - a_2)$$
\[
(b_2 - a_2) \sum_{j=1}^{r} (c_j - c_{j-1}) = (b_2 - a_2)(b_1 - a_1) = V(R) .
\]

We now give the proof of XIV.3.1.6.

**Proof:** Let \( R = [a_1, b_1] \times \cdots \times [a_n, b_n] \) and \( R_k = [a_{k,1}, b_{k,1}] \times \cdots \times [a_{k,n}, b_{k,n}] \). For each \( j \), let \( \mathcal{P}_j \) be the partition of \([a_j, b_j]\) consisting of all the \( a_{k,j} \) and \( b_{k,j} \). Let \( \mathcal{G} \) be the corresponding grid, and let \( \{R_{j_1, \ldots, j_n}\} \) be the rectangles in the grid. Then, by XIV.3.1.7.,

\[
V(R) = \sum_{j_1, \ldots, j_n} V(R_{j_1, \ldots, j_n}) .
\]

For each \( k \), set

\[
I_k = \{ (j_1, \ldots, j_n) : R_{j_1, \ldots, j_n} \subseteq R_k \} .
\]

Every \((j_1, \ldots, j_n)\) for which \( R_{j_1, \ldots, j_n} \) is nondegenerate is in exactly one \( I_k \). For each \( k \), the \( R_{j_1, \ldots, j_n} \) for \((j_1, \ldots, j_n) \in I_k\) form a grid subdivision of \( R_k \), so again by XIV.3.1.7. we have

\[
V(R_k) = \sum_{(j_1, \ldots, j_n) \in I_k} V(R_{j_1, \ldots, j_n})
\]

for each \( k \), and thus we have

\[
V(R) = \sum_{j_1, \ldots, j_n} V(R_{j_1, \ldots, j_n}) = m \left[ \sum_{k=1}^{m} \sum_{(j_1, \ldots, j_n) \in I_k} V(R_{j_1, \ldots, j_n}) \right] = m \sum_{k=1}^{m} V(R_k) .
\]

See Figure XIV.19.

**Upper and Lower Integrals over a Rectangle**

We now proceed just as we did to define the integral of a bounded function over an interval in \( \mathbb{R} \):

**XIV.3.1.8.** **Definition.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \to \mathbb{R} \) a bounded function. If \( \mathcal{P} = \{R_1, \ldots, R_m\} \) is a partition of \( R \), for each \( k \) set \( M_k = \sup \{ f(x) : x \in R_k \} \) and \( m_k = \inf \{ f(x) : x \in R_k \} \). Then the **upper and lower sums for \( f \) with respect to \( \mathcal{P} \)** are

\[
\mathcal{S}(f, \mathcal{P}) = \sum_{k=1}^{m} M_k \cdot V(R_k)
\]

\[
\mathcal{L}(f, \mathcal{P}) = \sum_{k=1}^{m} m_k \cdot V(R_k) .
\]

This definition agrees with the one in () if \( n = 1 \). Note that it does not make sense unless \( f \) is bounded.
XIV.3.1.9. It is clear that if $Q$ is a refinement of $P$, then

$$\mathcal{S}(f, P) \leq \mathcal{S}(f, Q) \leq \mathcal{S}(f, Q) \leq \mathcal{S}(f, P)$$

for any bounded function $f$. Thus the following definition is reasonable:

XIV.3.1.10. **Definition.** Let $R$ be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. The *upper and lower integrals* of $f$ over $R$ are

$$\int_R f \, dV = \inf_P \mathcal{S}(f, P)$$

$$\int_R f \, dV = \sup_P \mathcal{S}(f, P) .$$

The function $f$ is *(Riemann)* integrable over $R$ if $\int_R f \, dV = \int_R f \, dV$; in this case, the common value is the integral of $f$ over $R$, denoted $\int_R f \, dV$.

XIV.3.1.11. If $n = 1$, this definition is consistent with (): If $R = [a, b]$ and $f : R \to \mathbb{R}$ is bounded, then $f$ is integrable over $R$ if and only if it is Riemann integrable in the sense of (), and we have

$$\int_R f \, dV = \int_{[a, b]} f \, dV = \int_a^b f(x) \, dx .$$
XIV.3.1.12. Of course we have \( \int_R f \, dV \leq \overline{\int}_R f \, dV \) for any \( R \) and \( f \). As in (\() we have:

XIV.3.1.13. Proposition. Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \rightarrow \mathbb{R} \) a bounded function. Then \( f \) is integrable on \( R \) if and only if, for every \( \epsilon > 0 \), there is a partition \( \mathcal{P} \) of \( R \) such that
\[
\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) < \epsilon .
\]
Reasonable functions turn out to be Riemann integrable. Most importantly, we have the following higher-dimensional version of (\(), which is proved almost identically. We omit the proof since we will obtain a more general result in XIV.3.2.21. with almost the same proof.

XIV.3.1.14. Theorem. Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \rightarrow \mathbb{R} \) a continuous function. Then \( f \) is integrable over \( R \).

However, functions which are too discontinuous are not integrable:

XIV.3.1.15. Example. Let \( A \) be a subset of \( \mathbb{R}^n \) which is dense and has dense complement, e.g. the set of points with rational coordinates. If \( R \) is any rectangle in \( \mathbb{R}^n \) and \( \mathcal{P} \) is any partition of \( R \), then \( \overline{\mathcal{S}}(\chi_A, \mathcal{P}) = V(R) \) and \( \underline{\mathcal{S}}(\chi_A, \mathcal{P}) = 0 \). Thus \( \int_R \chi_A \, dV = V(R) \) and \( \int_R \chi_A \, dV = 0 \), so \( \chi_A \) is not integrable on \( R \) unless \( R \) is degenerate.

Additivity of Upper and Lower Integrals

XIV.3.1.16. Proposition. Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \rightarrow \mathbb{R} \) a bounded function. If \( \mathcal{P} = \{R_1, \ldots, R_m\} \) is a partition of \( R \), then we have
\[
\int_R f \, dV = \sum_{k=1}^{m} \int_{R_k} f \, dV \quad \text{and} \quad \int_R f \, dV = \sum_{k=1}^{m} \int_{R_k} f \, dV .
\]

Proof: In calculating \( \int_R f \, dV \), only partitions of \( R \) refining \( \mathcal{P} \) need be considered. Each such partition \( \mathcal{Q} \) splits into partitions \( \mathcal{Q}_k \) of \( R_k \) for each \( k \), so we obtain
\[
\overline{\mathcal{S}}(f, \mathcal{Q}) = \sum_{k=1}^{m} \overline{\mathcal{S}}(f, \mathcal{Q}_k) .
\]
Conversely, if \( \mathcal{Q}_k \) is a partition of \( R_k \) for each \( k \), the union of the \( \mathcal{Q}_k \) is a partition of \( R \) refining \( \mathcal{P} \). So the infimum over partitions can be taken separately in each term. The argument for \( \underline{\mathcal{S}}(f, \mathcal{Q}) \) is essentially identical.
XIV.3.1.17. Also note that if \( R \) is a degenerate rectangle and \( f \) is any bounded function on \( R \), then \( f \) is integrable on \( R \) and \( \int_R f \, dV = 0 \).

**Linearity of the Integral**

**XIV.3.1.18. Proposition.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), \( f : R \to \mathbb{R} \) a bounded function, and \( \alpha \in \mathbb{R} \). Then

\[
\int_R \alpha f \, dV = \begin{cases} 
\alpha \int_R f \, dV & \text{if } \alpha \geq 0 \\
\alpha \int_R f \, dV & \text{if } \alpha < 0 
\end{cases}
\]

\[
\int_R \alpha f \, dV = \begin{cases} 
\alpha \int_R f \, dV & \text{if } \alpha \geq 0 \\
\alpha \int_R f \, dV & \text{if } \alpha < 0 
\end{cases}
\]

In particular, if \( f \) is integrable over \( R \), then so is \( \alpha f \), and

\[
\int_R \alpha f \, dV = \alpha \int_R f \, dV .
\]

**Proof:** This is obvious from the fact that if \( P = \{R_1, \ldots, R_m\} \) is any partition of \( R \), then, for \( 1 \leq k \leq m \),

\[
M_k(\alpha f) = \alpha M_k(f) \text{ if } \alpha \geq 0 \quad \text{and} \quad M_k(\alpha f) = \alpha m_k(f) \text{ if } \alpha < 0,
\]

so that \( \mathcal{S}(\alpha f, P) = \alpha \mathcal{S}(f, P) \) if \( \alpha \geq 0 \) and \( \mathcal{S}(\alpha f, P) = \alpha \mathcal{S}(f, P) \) if \( \alpha < 0 \), and similarly for the other.

**XIV.3.1.19. Proposition.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) and \( g \) bounded functions on \( R \). Then

\[
\int_R (f + g) \, dV \leq \int_R f \, dV + \int_R g \, dV
\]

\[
\int_R (f + g) \, dV \geq \int_R f \, dV + \int_R g \, dV
\]

In particular, if \( f \) and \( g \) are integrable over \( R \), then so is \( f + g \), and

\[
\int_R (f + g) \, dV = \int_R f \, dV + \int_R g \, dV .
\]

**Proof:** This is similar to the previous proposition, noting that if \( P = \{R_1, \ldots, R_m\} \) is any partition of \( R \), then, for \( 1 \leq k \leq m \),

\[
M_k(f + g) \leq M_k(f) + M_k(g) \quad \text{and} \quad m_k(f + g) \geq m_k(f) + m_k(g)
\]

\( \)
Thus the integral over a fixed rectangle is linear: if \( f \) and \( g \) are integrable over a rectangle \( R \), and \( \alpha, \beta \in \mathbb{R} \), then \( \alpha f + \beta g \) is integrable over \( R \) and
\[
\int_R (\alpha f + \beta g) \, dV = \alpha \int_R f \, dV + \beta \int_R g \, dV.
\]
However, the upper and lower integrals are not linear for functions which are not integrable. For example, if \( A \) is a dense set with dense complement, then for any nondegenerate \( R \) we have
\[
V(R) = \int_R 1 \, dV = \int_R (\chi_A + \chi_{A^c}) \, dV < \int_R \chi_A \, dV + \int_R \chi_{A^c} \, dV = 2V(R)
\]
and similarly for the lower integral for \( -\chi_A - \chi_{A^c} \).

**Order Properties of the Integral**
The order properties of the integral are obvious (consider partitions):

**Proposition.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) and \( g \) bounded functions on \( R \). If \( f \leq g \) on \( R \), then \( \int_R f \, dV \leq \int_R g \, dV \) and \( \int_R f \, dV \leq \int_R g \, dV \).

**Corollary.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) a bounded function on \( R \). If \( m \leq f(x) \leq M \) for all \( x \in R \), then
\[
m \cdot V(R) \leq \int_R f \, dV \leq \int_R f \, dV \leq M \cdot V(R) .
\]
In particular, if \( f \geq 0 \) on \( R \), then \( \int_R f \, dV \geq 0 \).

**Integrability of Products**

Although the result is a special case of the general characterization of integrable functions in (), we can give a simple proof that a product of integrable functions is integrable.

**Lemma.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) a bounded function on \( R \). If \( |f(x)| \leq B \) for all \( x \in R \), then, for any partition \( \mathcal{P} = \{R_1, \ldots, R_m\} \) of \( R \),
\[
\mathbb{S}(f,\mathcal{P}) - \mathbb{S}(f^2,\mathcal{P}) \leq 2B[\mathbb{S}(f,\mathcal{P}) - \mathbb{S}(f,\mathcal{P})] .
\]

**Proof:** For \( 1 \leq k \leq m \), we have, for any \( x, y \in R_k \),
\[
f(x)^2 - f(y)^2 = (f(x) + f(y))(f(x) - f(y))
\]
\[
|f(x)^2 - f(y)^2| = |(f(x) + f(y))(f(x) - f(y))| \leq 2B|f(x) - f(y)|
\]
and therefore \( M_k(f^2) - m_k(f^2) \leq 2B(M_k(f) - m_k(f)) \), and the result follows by summing over \( k \).
XIV.3.1.24. **Corollary.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) a bounded function on \( R \). If \( |f(x)| \leq B \) for all \( x \in R \), then
\[
\int_R f^2 \, dV - \int_R f^2 \, dV \leq 2B \left[ \int_R f \, dV - \int_R f \, dV \right].
\]
In particular, if \( f \) is integrable over \( \mathbb{R} \), then so is \( f^2 \).

XIV.3.1.25. **Proposition.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f \) and \( g \) bounded functions on \( R \). If \( f \) and \( g \) are integrable over \( R \), then so is \( fg \).

**Proof:** We have \( f + g \) integrable over \( R \) by XIV.3.1.19., and thus \( f^2, g^2, \) and \( (f + g)^2 \) are integrable over \( R \) by XIV.3.1.24.. Since
\[
fg = \frac{1}{2} [(f + g)^2 - f^2 - g^2]
\]
we have that \( fg \) is integrable over \( R \) by XIV.3.1.20..

**Expanding the Rectangle**

We now obtain a technical result which will be important when extending the notion of Riemann integral from rectangles to more general regions. If \( R \) is a rectangle, \( S \) is a subrectangle, and \( f \) is integrable on \( S \), extend \( f \) to \( \mathbb{R}^n \) by setting \( f(x) = 0 \) for \( x \in \mathbb{R}^n \setminus S \). Then the integral of the extended function over \( R \) should be the same as the integral of \( f \) over \( S \). It is:

XIV.3.1.26. **Theorem.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( S \) a subrectangle of \( R \). Let \( f : S \rightarrow \mathbb{R} \) be a bounded function. Define \( g : R \rightarrow \mathbb{R} \) by
\[
g(x) = \begin{cases} f(x) & \text{if } x \in S \\ 0 & \text{if } x \in R \setminus S \end{cases}.
\]
Then
\[
\int_R g \, dV = \int_S f \, dV \text{ and } \int_R g \, dV = \int_S f \, dV.
\]
In particular, \( g \) is integrable over \( R \) if and only if \( f \) is integrable over \( S \), and in this case
\[
\int_R g \, dV = \int_S f \, dV.
\]

There would seem to be a simple proof: consider partitions of \( R \) consisting of a partition of \( S \) and some additional rectangles on which \( f \) is zero. But the problem is that the rectangles in a partition are only nonoverlapping, not disjoint, so some of the additional rectangles will contain points of \( \partial S \) at which \( f \) may well be nonzero. Thus we must work harder to prove the theorem, taking partitions for which all the rectangles covering \( \partial S \) but not in \( S \) have a small total content.
The following construction will be used in the proof. Let \( S = [a_1, b_1] \times \cdots \times [a_n, b_n] \) be a rectangle in \( \mathbb{R}^n \). If \( \delta > 0 \), set
\[
S_\delta = [a_1 - \delta, b_1 + \delta] \times \cdots \times [a_n - \delta, b_n + \delta].
\]
Then \( S_\delta \) is a rectangle in \( \mathbb{R}^n \) containing \( S \) in its interior. \( V(S_\delta) \) is a polynomial in \( \delta \), so by continuity, for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that \( V(S) < V(S_\delta) < V(S) + \epsilon \).

We now give the proof of XIV.3.1.26.

**Proof:** We may assume \( R \) contains \( S \) in its interior, since if not two applications of the theorem for a larger rectangle containing both \( R \) and \( S \) in its interior give the result. Then \( S_\delta \subseteq R \) for all sufficiently small \( \delta > 0 \). Note that if \( S_\delta \subseteq R \), any partition of \( S \) can be extended to a partition of \( R \) in which \( S_\delta \) as well as \( S \) is a union of rectangles in the partition.

Fix \( \epsilon > 0 \). Let \( B \) be a bound for \( |f| \) on \( S \). Choose \( \delta > 0 \) so that \( S_\delta \subseteq R \) and \( V(S_\delta) < V(S) + \frac{\epsilon}{3B} \). If \( \mathcal{P} \) is any partition of \( S \), extend \( \mathcal{P} \) to a partition \( \mathcal{Q} \) of \( R \) for which \( S_\delta \) is a union of rectangles in \( \mathcal{Q} \). Let \( \mathcal{Q}_1 \) be the set of rectangles in \( \mathcal{Q} \) contained in \( S_\delta \) but not in \( S \), and \( \mathcal{Q}_2 \) the set of rectangles in \( \mathcal{Q} \) not contained in \( S_\delta \). Then
\[
\mathcal{S}(g, \mathcal{Q}) = \sum_{k=1}^m M_k \cdot V(R_k) = \sum_{R_k \in \mathcal{P}} M_k \cdot V(R_k) + \sum_{R_k \in \mathcal{Q}_1} M_k \cdot V(R_k) + \sum_{R_k \in \mathcal{Q}_2} M_k \cdot V(R_k).
\]
The first sum is exactly \( \mathcal{S}(f, \mathcal{P}) \), and the last sum is 0 since \( g \) is identically 0 on the rectangles in \( \mathcal{Q}_2 \). If \( R_k \in \mathcal{Q}_2, |M_k| \leq B \), so the second sum in absolute value is dominated by
\[
\sum_{R_k \in \mathcal{Q}_2} B \cdot V(R_k) = B(V(S_\delta) - V(S)) < \frac{\epsilon}{3}
\]
and therefore
\[
|\mathcal{S}(f, \mathcal{P}) - \mathcal{S}(g, \mathcal{Q})| < \frac{\epsilon}{3}
\]
i.e. for every partition \( \mathcal{P} \) of \( S \) there is a partition \( \mathcal{Q} \) of \( R \) for which this holds. On the other hand, if \( \mathcal{Q} \) is any refinement of \( \mathcal{Q} \), then \( \mathcal{Q} \) is of this form for some refinement \( \mathcal{P} \) of \( \mathcal{P} \). We can thus choose \( \mathcal{P} \) and \( \mathcal{Q} \) so that
\[
\mathcal{S}(f, \mathcal{P}) - \int_S f \, dV < \frac{\epsilon}{3}
\]
and
\[
\mathcal{S}(g, \mathcal{Q}) - \int_R g \, dV < \frac{\epsilon}{3}
\]
from which it follows that
\[
\left| \int_S f \, dV - \int_S g \, dV \right| < \epsilon.
\]
Since \( \epsilon > 0 \) is arbitrary, the upper integrals are equal. The argument for the lower integrals is almost identical.

See XIV.3.11.8. for an alternate version of the proof.
XIV.3.2. Content and Jordan Regions

Using the integral over rectangles, we can say what the “volume” is of more general subsets of Euclidean space.

XIV.3.2.1. Definition. Let $E$ be a bounded subset of $\mathbb{R}^n$. The upper $n$-dimensional (Jordan) content of $E$ is

$$V_n(E) = \int_R \chi_E \, dV$$

where $R$ is a rectangle in $\mathbb{R}^n$ containing $E$.

The lower $n$-dimensional (Jordan) content of $E$ is

$$\underline{V}_n(E) = \int_R \chi_E \, dV$$

where $R$ is a rectangle in $\mathbb{R}^n$ containing $E$.

$E$ is a Jordan region if $V_n(E) = \underline{V}_n(E)$; the (Jordan) content $V_n(E)$ is the common value.

XIV.3.2.2. Note that the upper and lower content are well defined and independent of the containing rectangle $R$ by XIV.3.1.26. We usually suppress the words “$n$-dimensional” and the subscript $n$, but the dimension must be carefully kept in mind. Some authors use the term “volume” instead of (Jordan) content. See XIV.3.1.2.

XIV.3.2.3. A bounded subset $E$ of $\mathbb{R}^n$ is a Jordan region if and only if $\chi_E$ is integrable over some (hence every by XIV.3.1.26.) rectangle $R$ containing $E$, and $V(E) = \int_R \chi_E \, dV$.

XIV.3.2.4. Proposition. If $R$ is a rectangle in $\mathbb{R}^n$, then $R$ is a Jordan region and its Jordan content is $V(R)$, as defined in XIV.3.1.1.

Proof: We can take the containing rectangle to be $R$ itself. Then $\chi_R$ is the constant function 1 on $R$. By XIV.3.1.22. (or by simple direct calculation), $\chi_R$ is integrable on $R$ and $\int_R \chi_R \, dV = \int_R 1 \, dV = V(R)$. \(\Box\)

Thus our notation and terminology is consistent.

XIV.3.2.5. A subset of $\mathbb{R}^n$ is a Jordan region if and only if it boundary has zero content (XIV.3.2.24.). “Geometrically reasonable” subsets of $\mathbb{R}^n$ are thus Jordan regions. But a set like the intersection of the set $A$ of XIV.3.1.15. with a rectangle is not a Jordan region. Note that Jordan regions are by definition bounded.

XIV.3.2.6. If $A$ and $B$ are bounded subsets of $\mathbb{R}^n$ with $A \subseteq B$, then $\overline{V}(A) \leq \overline{V}(B)$ and $\underline{V}(A) \leq \underline{V}(B)$ (cf. XIV.3.1.22.).

The next observation will be considerably improved in XIV.3.2.25., but this form is needed to logically develop the theory and characterization of Jordan regions:
**XIV.3.2.7. Lemma.** Let $A$ and $B$ be Jordan regions in $\mathbb{R}^n$. Then $A \cup B$ is a Jordan region if and only if $A \cap B$ is a Jordan region. In this case, we have

$$V(A \cup B) = V(A) + V(B) - V(A \cap B).$$

**Proof:** We have $\chi_{A \cup B} = \chi(A) + \chi(B) - \chi(A \cap B)$. Let $R$ be a rectangle containing both $A$ and $B$. By assumption, $\chi(A)$ and $\chi(B)$ are integrable over $R$, so it follows from XIV.3.1.20. that $\chi_{A \cup B}$ is integrable over $R$ if and only if $\chi_{A \cap B}$ is. The volume formula also follows from XIV.3.1.20.

**XIV.3.2.8. Corollary.** If \{R_1, \ldots, R_m\} is a finite set of rectangles in $\mathbb{R}^n$, not necessarily nonoverlapping, then $\bigcup_{k=1}^m R_k$ is a Jordan region and

$$V\left(\bigcup_{k=1}^m R_k\right) \leq \sum_{k=1}^m V(R_k)$$

with equality if and only if the $R_k$ are nonoverlapping.

**Proof:** For $m = 2$ the result follows immediately from XIV.3.2.4. and XIV.3.2.7. since the intersection of two rectangles is a rectangle. The result for general $m$ follows by induction.

Along the same lines, the next observation is needed for the logical development although it will be improved in XIV.3.2.25.:

**XIV.3.2.9. Proposition.** Let $E$ be a Jordan region, and $R$ a rectangle containing $E$. Then $R \setminus E$ is also a Jordan region, and $V(R \setminus E) = V(R) - V(E)$.

**Proof:** We have that $\chi_{R \setminus E} = \chi_R - \chi_E$: $\chi_E$ is integrable over $R$ by assumption, and $\chi_R$ is integrable over $R$ by XIV.3.2.4., so $\chi_{R \setminus E}$ is integrable over $R$ and the content formula holds by XIV.3.1.20.

**An Alternate Description of Content**

We can give an illuminating direct description of content by explicitly describing upper and lower sums for the integral of an indicator function.

**XIV.3.2.10.** Let $E$ be a bounded subset of $\mathbb{R}^n$. If $R$ is a rectangle in $\mathbb{R}^n$ containing $E$ and $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of $R$, we have

$$\mathcal{S}(\chi_E, \mathcal{P}) = \sum_{R_k \cap E \neq \emptyset} V(R_k)$$

and

$$\mathcal{S}(\chi_E, \mathcal{P}) = \sum_{R_k \subseteq E} V(R_k)$$

1563
i.e. $\overline{V}(E)$ is estimated as the content of the union of the rectangles in $P$ containing points of $E$, and $\overline{V}(E)$ is estimated as the content of the union of the rectangles in $P$ entirely contained in $E$; in fact, they are the infimum and supremum of such contents over all partitions of $R$.

Combining with XIV.3.2.8, we obtain:

**XIV.3.2.11. Proposition.** If $E$ is a bounded subset of $\mathbb{R}^n$, then

$$\overline{V}(E) = \sup \left( \sum_{k=1}^{m} V(R_k) \right)$$

where the supremum is over all finite sets $\{R_1, \ldots, R_m\}$ of pairwise nonoverlapping rectangles contained in $E$.

For upper sums, since any finite set of rectangles has a subdivision into nonoverlapping rectangles, we can relax the requirement that the rectangles be nonoverlapping:

**XIV.3.2.12. Proposition.** If $E$ is a bounded subset of $\mathbb{R}^n$, then

$$\underline{V}(E) = \inf \left( \sum_{k=1}^{m} V(R_k) \right)$$

where the infimum is over all finite sets $\{R_1, \ldots, R_m\}$ of rectangles covering $E$.

Sets of Zero Content

**XIV.3.2.13. Definition.** A subset $E$ of $\mathbb{R}^n$ has *zero content* if $\overline{V}(E) = 0$.

Two subsets $A$ and $B$ of $\mathbb{R}^n$ are *nonoverlapping* if $A \cap B$ has zero content.

**XIV.3.2.14.** If $E$ has zero content, it is a Jordan region since $0 \leq \overline{V}(E) \leq \overline{V}(E) = 0$. Any subset of a set of zero content also has zero content. A rectangle has zero content if and only if it is degenerate. Thus two rectangles are nonoverlapping in the sense of XIV.3.2.13, if and only if they are nonoverlapping in the sense of XIV.3.1.4.

The next result is then an immediate consequence of XIV.3.2.7:

**XIV.3.2.15. Proposition.** Let $E_1, \ldots, E_m$ be pairwise nonoverlapping Jordan regions. Then $\bigcup_{k=1}^{m} E_k$ is a Jordan region and

$$\overline{V} \left( \bigcup_{k=1}^{m} E_k \right) = \sum_{k=1}^{m} V(E_k) .$$
XIV.3.2.16. **COROLLARY.** A finite union of sets of zero content has zero content.

The next result gives a large supply of sets of zero content:

XIV.3.2.17. **PROPOSITION.** Let $E$ be a compact Jordan region in $\mathbb{R}^{n-1}$, $\phi : E \to \mathbb{R}$ a continuous function, and $G$ the graph of $\phi$, i.e.,

$$G = \{(x_1, \ldots, x_n) : (x_1, \ldots, x_{n-1}) \in E, x_n = \phi(x_1, \ldots, x_{n-1})\} \subseteq \mathbb{R}^n.$$  

Then $G$ is a set of zero content in $\mathbb{R}^n$.

**Proof:** Let $R$ be a rectangle in $\mathbb{R}^n$ containing $E$. Let $\epsilon > 0$. Since $E$ is compact, $\phi$ is uniformly continuous on $E$, so there is a $\delta > 0$ such that $|\phi(x) - \phi(y)| < \frac{\epsilon}{\sqrt{V(R)}}$ whenever $x, y \in E$ and $|x - y| < \delta$. There is a partition $P = \{R_1, \ldots, R_m\}$ of $R$ such that the diameter of each $R_k$ is less than $\delta$; thus if $M_k$ and $m_k$ are the supremum and infimum of $\phi$ over $R_k \cap E$ for each $R_k$ containing points of $E$, we have $M_k - m_k \leq \frac{\epsilon}{\sqrt{V(R)}}$ for all such $k$. Then if $\tilde{R}_k = R_k \times [m_k, M_k]$ for each such $k$, then the $\tilde{R}_k$ are nonoverlapping rectangles covering $G$, and

$$\sum_k V(\tilde{R}_k) \leq \sum_k V(R_k) \cdot \frac{\epsilon}{\sqrt{V(R)}} \leq \epsilon$$

(the sum is over all $k$ for which $R_k \cap E \neq \emptyset$). Thus $\overline{V_n}(G) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $\overline{V_n}(G) = 0$.  

XIV.3.2.18. If $E$ is a compact set in $\mathbb{R}^n$ and every point of $E$ has a neighborhood in $E$ which is a set of zero content in $\mathbb{R}^n$, then a simple compactness argument and XIV.3.2.16. shows that $E$ is a set of zero content. In particular, if $E$ is compact and locally the graph of a continuous function of some $n - 1$ of the coordinates, then $E$ is a set of zero content. Many sets can be shown to be of this form by the Implicit Function Theorem.

XIV.3.2.19. It is crucial to specify the $n$ when saying that a set has zero content in $\mathbb{R}^n$. If we regard $\mathbb{R}^n$ as a subset of $\mathbb{R}^{n+1}$ by identifying it with an affine subspace, every bounded subset of $\mathbb{R}^n$ has zero content in $\mathbb{R}^{n+1}$. In fact, the notion of Jordan region in $\mathbb{R}^n$ depends on $n$: any bounded subset of $\mathbb{R}^n$ is thus a Jordan region in $\mathbb{R}^{n+1}$. So the property of $E$ being a Jordan region is not intrinsic to $E$, but depends on the choice of Euclidean space containing $E$ (and also how $E$ is embedded in that space).

XIV.3.2.20. **PROPOSITION.** Let $A$ and $B$ be bounded regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. If $A$ has zero content in $\mathbb{R}^p$, then $A \times B$ has zero content in $\mathbb{R}^{p+q}$.

**Proof:** Let $S$ be a bounded rectangle in $\mathbb{R}^q$ containing $B$. Let $\epsilon > 0$. There is a finite set $\{R_1, \ldots, R_n\}$ in $\mathbb{R}^p$ covering $A$ with

$$\sum_{k=1}^n V(R_k) < \frac{\epsilon}{V(S) + 1}.$$  

1565
Then \( \{R_1 \times S, \ldots, R_n \times S\} \) is a finite set of rectangles covering \( A \times B \), and \( V(R_k \times S) = V(R_k)V(S) \), so

\[
V(A \times B) \leq \sum_{k=1}^{n} V(R_k \times S) = V(S) \sum_{k=1}^{n} V(R_k) < \epsilon .
\]

The next theorem is a key result about integrability, generalizing XIV.3.14. This is, however, not the ultimate characterization of integrability, which will be given in ( ).

**Theorem.** Let \( R \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \to \mathbb{R} \) a bounded function. If \( f \) is continuous except on a set of zero content, then \( f \) is integrable over \( R \).

**Proof:** Suppose \( f \) is continuous except on a set \( A \) with \( V(A) = 0 \). We will use the following notation. If \( \mathcal{P} = \{R_1, \ldots, R_m\} \) is a partition of \( R \), set \( J = \{j : R_j \cap A \neq \emptyset\} \) and \( K = \{k : R_k \cap A = \emptyset\} \).

Let \( M \) be a bound for \( |f| \) on \( R \). Let \( \epsilon > 0 \). Then, since \( A \) has volume 0, there is a partition \( \mathcal{P} \) of \( R \) such that

\[
\sum_{j \in J} V(R_j) < \epsilon .
\]

Let \( B = \bigcup_{k \in K} R_k \). Then \( B \) is compact and \( f \) is continuous on \( B \), hence uniformly continuous, so there is a \( \delta > 0 \) such that \( |f(x) - f(y)| < \epsilon \frac{1}{2V(R)} \) whenever \( x, y \in B \) and \( |x - y| < \delta \). By subdividing the \( R_k \) for \( k \in K \) if necessary, we may assume that the diameter of each \( R_k \) is less than \( \delta \). Thus if \( M_k \) and \( m_k \) are the supremum and infimum of \( f \) over \( R_k \), we have

\[
M_k - m_k \leq \frac{\epsilon}{2V(R)} .
\]

Furthermore, for each \( j \in J \) we have \( M_j - m_j \leq 2M \). Thus

\[
\mathcal{S}(f, \mathcal{P}) - \mathcal{S}(f, \mathcal{P}) = \sum_{j \in J} (M_j - m_j)V(R_j) + \sum_{k \in K} (M_k - m_k)V(R_k)
\]

\[
\leq 2M \sum_{j \in J} V(R_j) + \frac{\epsilon}{2V(R)} \sum_{k \in K} V(R_k) \leq 2M \frac{\epsilon}{4M} + \frac{\epsilon}{2V(R)} V(R) = \epsilon .
\]

Since \( \epsilon > 0 \) is arbitrary, \( f \) is integrable over \( R \).

**Characterization of Jordan Regions**

The first result is a corollary of XIV.3.2.21.

**Corollary.** Let \( E \) be a bounded set in \( \mathbb{R}^n \). If \( \partial E \) has zero content, then \( E \) is a Jordan region.

**Proof:** The indicator function \( \chi_E \) is discontinuous exactly on \( \partial E \).
XIV.3.2.23. Proposition. If $E$ is a Jordan region in $\mathbb{R}^n$, then $\bar{E}$ and $E^o$ are also Jordan regions, and

$$V(\bar{E}) = V(E^o) = V(E).$$

Proof: Let $R$ be a rectangle containing $E$. Let $\epsilon > 0$, and let $P = \{R_1, \ldots, R_m\}$ be a partition of $R$ such that

$$\sum_{j \in J} V(R_j) < V(E) + \epsilon$$

where $J$ is the set of $j$ for which $R_j \cap E \neq \emptyset$. Let $B = \bigcup_{j \in J} R_j$. Then $B$ is a closed set and $E \subseteq B$, so $\bar{E} \subseteq B$, and $B$ is a Jordan region by XIV.3.2.24. We have

$$V(E) = V(\bar{E}) \leq V(E) \leq V(B) = \sum_{j \in J} V(R_j) < V(E) + \epsilon.$$

Since $\epsilon > 0$ is arbitrary, we have

$$V(E) = V(\bar{E}) = V(E) = V(E)$$

so $\bar{E}$ is a Jordan region and $V(\bar{E}) = V(E)$.

By XIV.3.2.9, $R \setminus E$ is a Jordan region, so $\overline{R \setminus E} = R \setminus E^o$ is a Jordan region, and hence $E^o$ is a Jordan region; and we have

$$V(R) - V(E^o) = V(R \setminus E^o) = V(R \setminus E) = V(R \setminus E) = V(R) - V(E)$$

and it follows that $V(E^o) = V(E)$.

We now come to the main theorem characterizing Jordan regions.

XIV.3.2.24. Theorem. Let $E$ be a bounded subset of $\mathbb{R}^n$. Then $E$ is a Jordan region in $\mathbb{R}^n$ if and only if $\partial E$ has zero content.

Proof: If $\partial E$ has zero content, then $E$ is a Jordan region by XIV.3.2.22. Conversely, suppose $E$ is a Jordan region. Then $\bar{E}$ and $E^o$ are also Jordan regions, and $V(\bar{E}) = V(E^o) = V(E)$ by XIV.3.2.23. Let $R$ be a rectangle containing $E$. If $P = \{R_1, \ldots, R_m\}$ is a partition of $R$, we can partition $\{1, \ldots, m\}$ into three sets $I, J, K$, where $I$ is the set of $i$ such that $R_i \subseteq E^o$, $J$ is the set of $j$ such that $R_j \subseteq R \setminus E$ (i.e. such that $R_j \cap \bar{E} = \emptyset$), and $K$ is the set of $k$ such that $R_k \cap \partial E \neq \emptyset$.

Let $\epsilon > 0$. If $P$ is sufficiently fine, then by XIV.3.2.11. and XIV.3.2.9, we have

$$\sum_{i \in I} V(R_i) > V(E^o) - \frac{\epsilon}{2} = V(E) - \frac{\epsilon}{2}$$

$$\sum_{j \in J} V(R_j) > V(R \setminus \bar{E}) - \frac{\epsilon}{2} = V(R) - V(E) - \frac{\epsilon}{2}.$$

$$\sum_{i \in I} V(R_i) + \sum_{j \in J} V(R_j) > \left[ V(E) - \frac{\epsilon}{2} \right] + \left[ V(R) - V(E) - \frac{\epsilon}{2} \right] = V(R) - \epsilon.$$
By XIV.3.1.6., we have
\[ V(R) = \sum_{i \in I} V(R_i) + \sum_{j \in J} V(R_j) + \sum_{k \in K} V(R_k) \]
so we conclude
\[ \sum_{k \in K} V(R_k) < \epsilon . \]

If we let \( B = \bigcup_{k \in K} R_k \), then \( \partial E \subseteq B \), so by XIV.3.2.8.
\[ \overline{V}(\partial E) \leq V(B) = \sum_{k \in K} V(R_k) < \epsilon \]
and since \( \epsilon > 0 \) is arbitrary, \( \overline{V}(\partial E) = 0. \)

**XIV.3.2.25. Corollary.** Let \( A \) and \( B \) be Jordan regions in \( \mathbb{R}^n \). Then \( A \cup B, A \cap B, \) and \( A \setminus B \) are Jordan regions, and
\[ V(A \cup B) = V(A) + V(B) - V(A \cap B) \]
\[ V(A \setminus B) = V(A) - V(A \cap B) . \]

**Proof:** By XIV.3.2.24., \( \partial A \) and \( \partial B \) have zero content. We have \( \partial(A \cup B) \subseteq \partial A \cup \partial B \), and \( \partial A \cup \partial B \) has zero content by XIV.3.2.16.; thus \( \partial(A \cup B) \) has zero content and \( A \cup B \) is a Jordan region by XIV.3.2.24.. It then follows from XIV.3.2.7. that \( A \cap B \) is a Jordan region and the first content formula holds. If \( R \) is a rectangle containing \( A \) and \( B \), then \( R \setminus B \) is a Jordan region by XIV.3.2.9., so \( A \setminus B = A \cap (R \setminus B) \) is a Jordan region. We have
\[ V(A) = V(A \cap B) + V(A \setminus B) \]
since \( A \cap B \) and \( A \setminus B \) are disjoint.

**XIV.3.2.26. Corollary.** Let \( A \) and \( B \) be Jordan regions in \( \mathbb{R}^p \) and \( \mathbb{R}^q \) respectively. Then \( A \times B \) is a Jordan region in \( \mathbb{R}^{p+q} \).

**Proof:** We have \( \partial(A \times B) = (\partial A \times B) \cup (A \times \partial B) \), which is a set of zero content by XIV.3.2.20..

**XIV.3.3. Integration over Jordan Regions**

It is now a simple matter to extend the integral to general Jordan regions.

**XIV.3.3.1. Definition.** Let \( E \) be a Jordan region in \( \mathbb{R}^n \), and \( f : E \to \mathbb{R} \). Denote by \( f \chi_E \) the function from \( \mathbb{R}^n \) to \( \mathbb{R} \) defined by
\[ f \chi_E(x) = \begin{cases} f(x) & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases} . \]

This definition is a slight abuse of notation since \( f \) is only defined on \( E \), so \( f \chi_E \) is technically not a product; but this ambiguity will cause no trouble.
XIV.3.3.2. Definition. Let $E$ be a Jordan region in $\mathbb{R}^n$, and $f : E \to \mathbb{R}$ a function. Then $f$ is (Riemann) integrable over $E$ if $f \chi_E$ is integrable over some (hence every) rectangle containing $E$. If $f$ is integrable over $E$, define

$$\int_E f \, dV = \int_R f \chi_E \, dV$$

where $R$ is any rectangle containing $E$ (cf. XIV.3.1.26.).

Note that an integrable function is necessarily bounded.

XIV.3.3.3. Comparing definitions, the constant function 1 is integrable over $E$, and $\int_E 1 \, dV = V(E)$. (This example indicates that it is not reasonable to try to similarly define the integral of a function over a set which is not a Jordan region, since even constant functions will not then be integrable).

XIV.3.3.4. Proposition. Let $A$ and $B$ be Jordan regions in $\mathbb{R}^n$, with $A \subseteq B$. If $f : A \to \mathbb{R}$ is integrable on $A$, then $f|_B$ is integrable on $B$.

Proof: We have $f \chi_B = (f \chi_A) \chi_B$ (this is an honest product). If $R$ is a rectangle containing $A$, then $f \chi_A$ and $\chi_B$ are integrable over $R$ by assumption, so $(f \chi_A) \chi_B$ is integrable over $R$ by XIV.3.1.25..)

The next extension of XIV.3.2.21. gives a large supply of integrable functions including almost all functions of interest in applications:

XIV.3.3.5. Theorem. Let $E$ be a Jordan region in $\mathbb{R}^n$, and $f : E \to \mathbb{R}$ a bounded function. If $f$ is continuous except on a set of zero content, then $f$ is integrable over $E$.

Proof: Suppose $f$ is continuous on $E$ except on a subset $A$ of zero content. If $R$ is a rectangle containing $E$, then $f \chi_E$ is continuous on $R$ except (possibly) on $A \cup \partial E$, which is a set of zero content (XIV.3.2.16.). Thus $f \chi_E$ is integrable over $R$ by XIV.3.2.21..)

Note that even if $f$ is continuous on $E$, XIV.3.2.21. is needed to show that $f$ is integrable on $E$; XIV.3.1.14. is not enough.

Properties of the Integral

The standard properties of the integral over Jordan regions then follow immediately from the properties of integrals over rectangles (note e.g. that $(f + g) \chi_E = f \chi_E + g \chi_E$, etc.)

XIV.3.3.6. Proposition. The integral over a fixed Jordan region is linear: if $f$ and $g$ are integrable over a Jordan region $E$, and $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g$ is integrable over $E$ and

$$\int_E (\alpha f + \beta g) \, dV = \alpha \int_E f \, dV + \beta \int_E g \, dV$$.
XIV.3.3.7. Proposition. Let $E$ be a Jordan region in $\mathbb{R}^n$, and $f$ and $g$ bounded functions on $E$. If $f \leq g$ on $E$, then

$$\int_E f\,dV \leq \int_E g\,dV.$$ 

XIV.3.3.8. Corollary. Let $E$ be a Jordan region in $\mathbb{R}^n$, and $f$ a bounded function on $E$. If $m \leq f(x) \leq M$ for all $x \in E$, then

$$m \cdot V(E) \leq \int_E f\,dV \leq M \cdot V(E).$$

In particular, if $f \geq 0$ on $E$, then $\int_E f\,dV \geq 0$.

XIV.3.3.9. Proposition. Let $E$ be a set of zero content, and $f : E \to \mathbb{R}$ a bounded function. Then $f$ is integrable over $E$, and $\int_E f\,dV = 0$.

Proof: Suppose $m \leq f(x) \leq M$ for all $x \in E$. If $R$ is a rectangle containing $E$, then $m\chi_E \leq f\chi_E \leq M\chi_E$ on $R$, so

$$0 = mV(E) = \int_R m\chi_E\,dV \leq \int_R f\chi_E\,dV \leq \int_R M\chi_E\,dV = 0.$$

XIV.3.3.10. Proposition. Let $A$ and $B$ be Jordan regions in $\mathbb{R}^n$, and $f : A \cup B \to \mathbb{R}$ an integrable function. Then $f$ is integrable over $A$, $B$, and $A \cap B$, and

$$\int_{A \cup B} f\,dV = \int_A f\,dV + \int_B f\,dV - \int_{A \cap B} f\,dV.$$

Proof: We have $f$ integrable over $A$, $B$, $A \cap B$ by XIV.3.3.4., and the formula then follows by linearity.

XIV.3.3.11. Corollary. Let $E_1, \ldots, E_m$ be pairwise nonoverlapping Jordan regions in $\mathbb{R}^n$, and $E = E_1 \cup \cdots \cup E_n$. If $f : E \to \mathbb{R}$ is a function, then $f$ is integrable over $E$ if and only if $f|_{E_k}$ is integrable over $E_k$ for each $k$. If this is the case, then

$$\int_E f\,dV = \sum_{k=1}^m \int_{E_k} f\,dV.$$
XIV.3.4. Fubini’s Theorem

XIV.3.4.1. The definition of the integral works more or less the same for \( \mathbb{R}^n \), \( n > 1 \), as it does for \( \mathbb{R} \). But there is one dramatic difference: so far we have no effective way of calculating integrals in the multivariable case, i.e. there is no Fundamental Theorem of Calculus in the vector case (actually there is, Stokes’ Theorem \(), but this is usually not useful for computing the integral).

XIV.3.4.2. Many authors use different notation for multivariable integrals, e.g. writing

\[
\int_R f \, dA \quad (n = 2)
\]

\[
\int_R \int_R f \, dV \quad (n = 3)
\]

\[
\cdots \int_R f \, dV \quad (n \text{ integral signs})
\]

in general for \( \int_R f \, dV \). I prefer the single integral sign notation, since the integral is a single integral over a multidimensional set; to emphasize the dimension, we can write \( \int_R f \, dV_n \). But the multiple integral notation is suggestive of a possible way to compute the integral: as an \( n \)-fold repeated one-dimensional integral, each of which can (potentially) be calculated using the Fundamental Theorem of Calculus. Fubini’s Theorem (the Riemann integral form) says that this can be done under reasonable hypotheses.

Iterated Integrals

XIV.3.4.3. If \( A \subseteq \mathbb{R}^p \) and \( B \subseteq \mathbb{R}^q \), we can regard \( A \times B \) as a subset of \( \mathbb{R}^{p+q} \cong \mathbb{R}^p \times \mathbb{R}^q \). If \( f \) is a bounded real-valued function on \( A \times B \), for each \( x \in A \) we have a function \( f[x] : B \to \mathbb{R} \) defined by \( f[x](y) = f(x, y) \). If \( A \) and \( B \) are Jordan regions and we set

\[
g(x) = \int_B f[x] \, dV
\]

then \( g \) is a bounded function from \( A \) to \( \mathbb{R} \); we symbolically write

\[
\int_A g \, dV = \int_A \left[ \int_B f(x, y) \, dV(y) \right] \, dV(x) = \int_A \int_B f(x, y) \, dV(y) \, dV(x)
\]

(the brackets are typically omitted), and similarly

\[
h(x) = \int_B f[x] \, dV
\]

\[
\int_A h \, dV = \int_A \int_B f(x, y) \, dV(y) \, dV(x)
\]

In the same way, for each \( y \in B \) we have a function \( f[y] : A \to \mathbb{R} \) defined by \( f[y](x) = f(x, y) \), and if

\[
g(y) = \int_A f[y] \, dV
\]
we write
\[ \int_B g \, dV = \int_B \int_A f(x, y) \, dV(x) \, dV(y) \]
and if
\[ h(y) = \int_A f(y) \, dV \]
\[ \int_B h \, dV = \int_B \int_A f(x, y) \, dV(x) \, dV(y) . \]
These are called \textit{iterated upper and lower integrals} (we could in principle also define \( \int_A \int_B f(x, y) \, dV(y) \, dV(x) \), etc., but these are rarely if ever used). If all the functions involved are integrable, we can write
\[ \int_A \int_B f(x, y) \, dV(y) \, dV(x) \text{ or } \int_B \int_A f(x, y) \, dV(x) \, dV(y) . \]
These are \textit{iterated integrals}.

\textbf{Fubini’s Theorem for Rectangles}

\textbf{XIV.3.4.4. Theorem. [Fubini’s Theorem, Upper and Lower Form]} Let \( R \) be a rectangle in \( \mathbb{R}^p \) and \( S \) a rectangle in \( \mathbb{R}^q \), so that \( R \times S \) is a rectangle in \( \mathbb{R}^{p+q} \). Let \( f : R \times S \to \mathbb{R} \) be a bounded function. Then
\[ \int_{R \times S} f \, dV \leq \int_R \int_S f(x, y) \, dV(y) \, dV(x) \leq \int_{R \times S} f \, dV . \]

To emphasize the dimensions of the integrals, we can write
\[ \int_{R \times S} f \, dV_{p+q} \leq \int_R \int_S f(x, y) \, dV_p(y) \, dV_q(x) \leq \int_{R \times S} f \, dV_{p+q} . \]

\textbf{XIV.3.4.5. Note that the order of \( R \) and \( S \) is arbitrary: we equally obtain}
\[ \int_{R \times S} f \, dV \leq \int_S \int_R f(x, y) \, dV(x) \, dV(y) \leq \int_{R \times S} f \, dV . \]

\textbf{Proof:} We prove the third inequality; the proof of the first is almost identical, and the second inequality is trivial. So, for \( x \in R \), set
\[ g(x) = \int_S f(x, y) \, dV = \int_S f(x, y) \, dV(y) . \]
We must prove
\[ \int_R g \, dV \leq \int_{R \times S} f \, dV . \]
Since every grid partition of $R \times S$ is of the form $\mathcal{P} \times \mathcal{Q}$ for some partitions $\mathcal{P}$ of $R$ and $\mathcal{Q}$ of $S$, it suffices to show that
\[ \mathcal{S}(g, \mathcal{P}) \leq \mathcal{S}(f, \mathcal{P} \times \mathcal{Q}) \]
for any such $\mathcal{P}$ and $\mathcal{Q}$, since then the desired inequality is obtained by taking the infimum over all $\mathcal{P}$ and $\mathcal{Q}$.

So let $\mathcal{P} = \{R_1, \ldots, R_m\}$ and $\mathcal{Q} = \{S_1, \ldots, S_n\}$ be partitions of $R$ and $S$. For $1 \leq j \leq m$ and $1 \leq k \leq n$, let $M_{jk}(f)$ be the supremum of $f$ over $R_j \times S_k$, and let $M_j(g)$ be the supremum of $g$ over $R_j$. We then have
\[ \mathcal{S}(f, \mathcal{P} \times \mathcal{Q}) = \sum_{j=1}^{m} \sum_{k=1}^{n} M_{jk}(f) V(R_j) V(S_k) \]
\[ \mathcal{S}(g, \mathcal{P}) = \sum_{j=1}^{m} M_j(g) V(R_j) \]
so it suffices to show that for each $j$,
\[ M_j(g) \leq \sum_{k=1}^{n} M_{jk}(f) V(S_k) . \]

Fix $j$. For $x \in R_j$,
\[ g(x) = \int_{S} f_{[x]} \, dV \leq \mathcal{S}(f_{[x]}, \mathcal{Q}) = \sum_{k=1}^{n} M_k(f_{[x]}) V(S_k) \]
where
\[ M_k(f_{[x]}) = \sup_{y \in S_k} f(x, y) \leq M_{jk}(f) \]
since $x \in R_j$. Thus
\[ g(x) \leq \sum_{k=1}^{n} M_{jk}(f) V(S_k) \]
and, taking the supremum $M_j(g)$ over $x \in R_j$, the desired inequality follows.

The next result, an immediate corollary of XIV.3.4.4., is the case of most interest:

**XIV.3.4.6.** Corollary. [Fubini’s Theorem, Riemann Integral Version] Let $R$ be a rectangle in $\mathbb{R}^p$ and $S$ a rectangle in $\mathbb{R}^q$, so that $R \times S$ is a rectangle in $\mathbb{R}^{p+q}$. Let $f : R \times S \to \mathbb{R}$ be a bounded function.

(i) Suppose $f$ is integrable over $R \times S$ and, for each $x \in R$, $f_{[x]}$ is integrable over $S$. If
\[ g(x) = \int_{S} f_{[x]} \, dV \]
for $x \in R$, then $g$ is integrable over $R$, and
\[ \int_{R \times S} f \, dV = \int_{R} g \, dV = \int_{R} \int_{S} f(x, y) \, dV(y) \, dV(x) . \]
(ii) Suppose \( f \) is integrable over \( R \times S \) and, for each \( y \in S \), \( f[y] \) is integrable over \( R \). If
\[
h(y) = \int_{R} f[y] \, dV
\]
for \( y \in S \), then \( h \) is integrable over \( S \), and
\[
\int_{R \times S} f \, dV = \int_{S} h \, dV = \int_{S} \int_{R} f(x, y) \, dV(x) \, dV(y) .
\]

(iii) If both (i) and (ii) hold (i.e. if \( f \) and each \( f[x] \) and \( f[y] \) are integrable), then
\[
\int_{R} \int_{S} f(x, y) \, dV(y) \, dV(x) = \int_{S} \int_{R} f(x, y) \, dV(x) \, dV(y) \quad \left( = \int_{R \times S} f \, dV \right) .
\]

**Proof:** (i): From the Theorem, we have
\[
\int_{R \times S} f \, dV \leq \int_{R} \int_{S} f(x, y) \, dV(y) \, dV(x) \leq \int_{R} \int_{S} f(x, y) \, dV(y) \, dV(x) \leq \int_{R \times S} f \, dV
\]
so the inequalities are equalities, and the middle equality exactly says that \( g \) is integrable over \( R \). The proof of (ii) is almost identical, and (iii) is just a combination of (i) and (ii).

A clean and especially important special case (cf. XIV.3.14., XV.8.3.5.(c)) is:

**XIV.3.4.7.** Corollary. [Fubini’s Theorem, Continuous Version] Let \( R \) be a rectangle in \( \mathbb{R}^p \) and \( S \) a rectangle in \( \mathbb{R}^q \), so that \( R \times S \) is a rectangle in \( \mathbb{R}^{p+q} \). Let \( f : R \times S \to \mathbb{R} \) be a continuous function. Then
\[
\int_{R \times S} f \, dV = \int_{R} \int_{S} f(x, y) \, dV(y) \, dV(x) = \int_{S} \int_{R} f(x, y) \, dV(x) \, dV(y) .
\]
In particular, the two iterated integrals are equal.

**XIV.3.4.8.** This result (XIV.3.4.6.) can be interpreted, and used, in two different ways. On the one hand, it says that (under certain conditions) the integral over \( R \times S \) can be computed by an iterated integral. On the other hand, (iii) says that (under certain conditions) the order of integration can be reversed in an iterated integral. Both interpretations are important in different contexts; for example, it is sometimes possible to evaluate an iterated integral in one order but not the other using the FTC (cf. XIV.3.11.11.).

**XIV.3.4.9.** Actually these versions of “Fubini’s Theorem” may have been “known” before Fubini, who proved a general version for Lebesgue integration (). All theorems involving changing the order of iterated summation or integration are now generically called “Fubini’s Theorem” since they are special cases of (). See XIV.3.4.34. for further comments.

If \( p = q = 1 \), XIV.3.4.7. can be rephrased, changing notation slightly:
XIV.3.4.10. **Theorem.** [Fubini’s Theorem in \( \mathbb{R}^2 \), Rectangle Continuous Version] Let \( R = [a, b] \times [c, d] \) be a rectangle in \( \mathbb{R}^2 \), and \( f : R \to \mathbb{R} \) a continuous function. Then

\[
\int_R f \, dV = \int_a^b \int_c^d f(x, y) \, dy \, dx = \int_c^d \int_a^b f(x, y) \, dx \, dy.
\]

In particular, the two iterated integrals are equal.

XIV.3.4.11. In calculating the inner integral (e.g. by the FTC), the outer variable is treated as a constant. Thus the inner integral can be regarded as a “partial antiderivative.”

XIV.3.4.12. If \( p \) or \( q \) is greater than 1, the iterated integrals can be further broken into iterated integrals of smaller dimension by repeatedly applying the various versions of Fubini’s Theorem. It is cumbersome to explicitly state the appropriate hypotheses for the higher iterated version of XIV.3.4.6., but XIV.3.4.7. can be cleanly stated. We only state the ultimate reduced version as in XIV.3.4.10.:

XIV.3.4.13. **Theorem.** [Fubini’s Theorem in \( \mathbb{R}^n \), Rectangle Continuous Version] Let \( R = [a_1, b_1] \times \cdots \times [a_n, b_n] \) be a rectangle in \( \mathbb{R}^n \), and \( f : R \to \mathbb{R} \) a continuous function. Then

\[
\int_R f \, dV = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) \, dx_n \cdots dx_2 dx_1.
\]

The iterated integrals can be done in any order.

**Iterated Integrals over Other Regions**

In applications we frequently want to integrate over regions more general than rectangles. In good cases such integrals can also be calculated as iterated integrals. Note that even to do this when integrating a continuous function requires the version of Fubini’s Theorem in XIV.3.4.6., i.e. XIV.3.4.7. is not enough.

We will not try to do this in the utmost generality, which requires cumbersome notation and hypotheses; instead, we will restrict to clean forms which are adequate for almost all applications. The first version is straightforward:

XIV.3.4.14. **Theorem.** Let \( A \) and \( B \) be Jordan regions in \( \mathbb{R}^p \) and \( \mathbb{R}^q \) respectively (so \( A \times B \) is a Jordan region in \( \mathbb{R}^{p+q} \)). If \( f : A \times B \to \mathbb{R} \) is bounded and continuous, then

\[
\int_{A \times B} f \, dV = \int_A \int_B f(x, y) \, dV(y) \, dV(x) = \int_B \int_A f(x, y) \, dV(x) \, dV(y)
\]

(in particular, the iterated integrals exist and are equal).

**Proof:** \( A \times B \) is a Jordan region (XIV.3.2.26.). Let \( R \) and \( S \) be rectangles containing \( A \) and \( B \) respectively, and apply XIV.3.4.6. to \( f \chi_{A \times B} \) on \( R \times S \) (it is easily checked that the hypotheses apply).  

\[\Box\]
XIV.3.4.15. Corollary. Let $A$ and $B$ be Jordan regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively (so $A \times B$ is a Jordan region in $\mathbb{R}^{p+q}$). Then $V(A \times B) = V(A)V(B)$.

To handle regions more general than products, we begin with subsets of $\mathbb{R}^2$:

XIV.3.4.16. Definition. A subset $E$ of $\mathbb{R}^2$ is projectable (with respect to the order $x, y$) if there are intervals $[a, b]$ and continuous functions $\phi, \psi : [a, b] \to \mathbb{R}$ with $\phi \leq \psi$ and

$$E = \{(x, y) : a \leq x \leq b, \ \phi(x) \leq y \leq \psi(x)\}.$$  

XIV.3.4.17. See Figure XIV.20 (adapted from http://tutorial.math.lamar.edu/Classes/CalcIII/DIGeneralRegion.aspx). We can also define projectability in the order $y, x$: $E$ is projectable in the order $y, x$ if there is an interval $[c, d]$ and functions $\phi, \psi : [c, d] \to \mathbb{R}$ such that

$$E = \{(x, y) : c \leq y \leq d, \ \phi(y) \leq x \leq \psi(y)\}.$$  

A set $E$ is projectable if it is projectable in at least one order; some simple regions are projectable in either order, and others projectable in one order but not the other. Most regions in $\mathbb{R}^2$ one would want to integrate over are projectable, or at least a finite union of nonoverlapping projectable regions. Note that a projectable region is compact.

![Projectable Regions](http://tutorial.math.lamar.edu/Classes/CalcIII/DIGeneralRegion.aspx)

Figure XIV.20: Projectable Regions.

XIV.3.4.18. The boundary of a projectable region

$$E = \{(x, y) : a \leq x \leq b, \ \phi(x) \leq y \leq \psi(x)\}$$

consists of the graphs of $\phi$ and $\psi$, and the vertical line segments

$$\{(a, y) : \phi(a) \leq y \leq \psi(a)\} \text{ and } \{(b, y) : \phi(b) \leq y \leq \psi(b)\}$$

(one or both vertical segments can be degenerate), all of which have volume 0 in $\mathbb{R}^2$ by XIV.3.2.17. Thus $E$ is a Jordan region in $\mathbb{R}^2$. 

1576
If \( E \) is the projectable region above, fix \( c \) and \( d \) with \( c \leq \phi(x) \leq \psi(x) \leq d \) for all \( x \in [a, b] \).

If \( f : E \to \mathbb{R} \) is a function, then \( f \) can be considered a function on \( R = [a, b] \times [c, d] \) in the usual way by setting it 0 on \( R \setminus E \). If \( f \) is continuous on \( E \), then for each \( x \in [a, b] \) we have that \( f(x) \) is piecewise-continuous on \([c, d]\), supported on \([\phi(x), \psi(x)]\), hence integrable on \([c, d]\). Fubini’s Theorem (XIV.3.4.6. (i) form) applies to give

\[
\int_E f \, dV = \int_R f \, dV = \int_a^b \int_c^d f(x, y) \, dy \, dx = \int_a^b \int_{\phi(x)}^{\psi(x)} f(x, y) \, dy \, dx .
\]

Thus we obtain:

**Theorem.** [Fubini’s Theorem for Projectable Sets in \( \mathbb{R}^2 \)] Let

\( E = \{(x, y) : a \leq x \leq b, \ \phi(x) \leq y \leq \psi(x)\} \)

be a projectable region (with respect to the order \( x, y \)) in \( \mathbb{R}^2 \), and \( f : E \to \mathbb{R} \) a continuous function. Then

\[
\int_E f \, dV = \int_a^b \int_{\phi(x)}^{\psi(x)} f(x, y) \, dy \, dx
\]

(in particular, the iterated integral exists).

Of course, the same result holds with \( x \) and \( y \) interchanged: if

\( E = \{(x, y) : c \leq y \leq d, \ \phi(y) \leq x \leq \psi(y)\} \)

and \( f \) is continuous on \( E \), then

\[
\int_E f \, dV = \int_c^d \int_{\phi(y)}^{\psi(y)} f(x, y) \, dx \, dy .
\]

If \( E \) is projectable in either order, we thus get a result about interchanging order of integration:

**Corollary.** [Fubini’s Theorem for Iterated Integrals on Projectable Regions] Let

\( E = \{(x, y) : a \leq x \leq b, \ \phi_1(x) \leq y \leq \psi_1(x)\} = \{(x, y) : c \leq y \leq d, \ \phi_2(y) \leq x \leq \psi_2(y)\} \)

be a region in \( \mathbb{R}^2 \) which is projectable in either order. If \( f : E \to \mathbb{R} \) is continuous, then

\[
\int_a^b \int_{\phi_1(x)}^{\psi_1(x)} f(x, y) \, dy \, dx = \int_c^d \int_{\phi_2(y)}^{\psi_2(y)} f(x, y) \, dx \, dy
\]

(in particular, the iterated integrals exist).
XIV.3.4.23. However, note that in general $[c, d] \neq [a, b]$, $\phi_2 \neq \phi_1$, $\psi_2 \neq \psi_1$, so the limits of integration must be carefully modified. It is frequently easy to do an iterated integral in one order and difficult or even impossible in the other order; cf. XIV.3.11.11.

XIV.3.4.24. We can do the same thing in higher dimensions, although the notation becomes cumbersome. Fix an ordering $x_1, \ldots, x_n$ of the coordinates of $\mathbb{R}^n$. To streamline notation, we will write things out with $j_k = k$ for all $k$; the masochist reader can easily adapt to the general situation. Inductively define $E_k \subseteq \mathbb{R}^k$ as follows. Let $E_1 = [a, b]$ be an interval in $\mathbb{R}$, and $\phi_1, \psi_1 : E_1 \to \mathbb{R}$ continuous functions with $\phi_1 \leq \psi_1$ on $E_1$. Suppose $E_1, \ldots, E_k$ have been defined, and $\phi_k, \psi_k : E_k \to \mathbb{R}$ continuous functions with $\phi_k \leq \psi_k$ on $E_k$. Set

$$E_{k+1} = \{(x_1, \ldots, x_{k+1}) \in \mathbb{R}^{k+1} : (x_1, \ldots, x_k) \in E_k, \phi_k(x_1, \ldots, x_k) \leq x_{k+1} \leq \psi_k(x_1, \ldots, x_k)\}.$$ 

By inductively applying XIV.3.2.17. as in XIV.3.4.18., we obtain that $E_k$ is a Jordan region in $\mathbb{R}^k$ for each $k$.

XIV.3.4.25. Definition. A set $E \subseteq \mathbb{R}^n$ is called a projectable region with respect to the order $x_1, \ldots, x_n$ if $E = E_n$ for some set of continuous functions as in XIV.3.4.24..

Let $E$ be a projectable region (for the order $x_1, \ldots, x_n$), and $f : E \to \mathbb{R}$ a continuous function. Use the notation of XIV.3.4.24. By successively applying the argument of XIV.3.4.19., we obtain (note that all the integrals exist as Riemann integrals):

XIV.3.4.26. Theorem. [Fubini’s Theorem for Projectable Sets in $\mathbb{R}^n$] Let $E$ be a projectable region (for the order $x_1, \ldots, x_n$), and $f : E \to \mathbb{R}$ a continuous function. Then

$$\int_E f \, dV = \int_a^b \int_{\phi_1(x_1)}^{\psi_1(x_1)} \cdots \int_{\phi_{n-1}(x_1, \ldots, x_{n-1})}^{\psi_{n-1}(x_1, \ldots, x_{n-1})} f(x_1, \ldots, x_n) \, dx_n \cdots dx_2 \, dx_1.$$ 

XIV.3.4.27. If $E$ is also projectable in another ordering of the coordinates, the integral $\int_E f \, dV$ can be computed by an iterated integral in that order, although the $\phi_k$ and $\psi_k$ will change in general when the ordering of the coordinates is changed. Thus we obtain as a corollary that the iterated integrals in the two orders are equal. The precise statement along the lines of XIV.3.4.22. is left to the reader.

The case $n = 3$ is important enough for applications that it is worth stating explicitly:

XIV.3.4.28. Theorem. [Fubini’s Theorem for Projectable Sets in $\mathbb{R}^3$] Let $[a, b]$ be an interval in $\mathbb{R}$, $\phi_1, \psi_1$ continuous functions from $[a, b]$ to $\mathbb{R}$ with $\phi_1 \leq \psi_1$ on $[a, b]$, $E_2 = \{(x, y) : a \leq x \leq b, \phi_1(x) \leq y \leq \psi_1(x)\}$, $\phi_2, \psi_2$ continuous functions from $E_2$ to $\mathbb{R}$ with $\phi_2 \leq \psi_2$, and $E = \{(x, y, z) : a \leq x \leq b, \phi_1(x) \leq y \leq \psi_1(x), \phi_2(x, y) \leq z \leq \psi_2(x, y)\}$.
Let \( f : E \to \mathbb{R} \) be a continuous function. Then
\[
\int_E f \, dV = \int_a^b \int_{\phi_1(x)}^{\phi_2(x,y)} f(x, y, z) \, dz \, dy \, dx .
\]

If \( E \) can be similarly written with respect to a different ordering of \( x, y, z \), \( \int_E f \, dV \) can be also calculated as an iterated integral in that order, and in particular the iterated integrals in the two orders are equal.

Most calculus texts have extensive examples of computing areas, volumes, and multiple integrals.

**Counterexamples**

We now give some examples showing that the hypotheses of Fubini’s Theorem are necessary.

**XIV.3.4.29. Examples.** Existence and equality of the iterated integrals does not imply integrability of \( f \):

(i) Let \( R = [0, 1]^2 \subseteq \mathbb{R}^2 \). Define \( f : R \to \mathbb{R} \) by
\[
f(x, y) = \begin{cases} 
2^n & \text{if } 2^{-n} < x, y \leq 2^{-n+1} \\
0 & \text{otherwise}
\end{cases} .
\]

Then \( f\{x\} \) and \( f\{y\} \) are integrable for all \( x \) and \( y \), and all have integral 1, so the iterated integrals exist and
\[
\int_0^1 \int_0^1 f(x, y) \, dy \, dx = \int_0^1 \int_0^1 f(x, y) \, dx \, dy = 1.
\]

But \( f \) is not integrable over \( R \) since it is unbounded.

(ii) A bounded example is more complicated. Let \( R = [0, 1]^2 \subseteq \mathbb{R}^2 \). Define \( f : R \to \mathbb{R} \) by setting \( f(x, y) = 1 \) if \( x \) and \( y \) are rational numbers with the same denominator when written in lowest terms, and \( f(x, y) = 0 \) otherwise. Then for fixed \( x \), \( f(x, y) = 1 \) for only finitely many \( y \), hence \( f\{y\} \) is integrable and has integral 0. Similarly, any \( f\{x\} \) is integrable and has integral 0. Thus the iterated integrals both exist and are 0. However, if \( \mathcal{P} \) is any partition of \( R \), it is easily seen that \( S(f, \mathcal{P}) = 1 \) and \( S(f, \mathcal{P}) = 0 \), so \( f \) is not integrable over \( R \). In fact, \( f \) is discontinuous everywhere on \( R \).

**XIV.3.4.30. Example.** Existence of the iterated integrals does not imply they are equal if \( f \) is not integrable: let \( R = [0, 1]^2 \subseteq \mathbb{R}^2 \). Define \( f : R \to \mathbb{R} \) by
\[
f(x, y) = \begin{cases} 
2^{2n} & \text{if } 2^{-n} < x, y \leq 2^{-n+1} \\
-2^{2n+1} & \text{if } 2^{-n} < x \leq 2^{-n+1}, 2^{-n-1} < y \leq 2^{-n} \\
0 & \text{otherwise}
\end{cases} .
\]

Then \( f\{x\} \) is integrable on \([0, 1]\) and \( \int_0^1 f\{x\}(y) \, dy = 0 \) for all \( x \); thus
\[
\int_0^1 \int_0^1 f(x, y) \, dy \, dx \text{ exists and equals 0 .}
\]
Also, \( f[y] \) is integrable on \([0, 1]\) for all \( y \): \( \int_0^1 f[y](x) \, dx = 0 \) if \( y \leq \frac{1}{2} \) but \( \int_0^1 f[y](x) \, dx = 2 \) if \( y > \frac{1}{2} \); thus
\[
\int_0^1 \int_0^1 f(x, y) \, dy \, dx \text{ exists and equals 1}.
\]

So the iterated integrals are not equal. Note that \( f \) is unbounded and hence not integrable on \( \mathbb{R} \).

This example is an adaptation of the double sum example IV.2.10.6.

XIV.3.4.31. \textbf{Example.} Integrability of \( f \) does not imply integrability of \( f[x] \) or \( f[y] \) for all \( x, y \): let
\( R = [0, 1]^2 \subset \mathbb{R}^2 \), \( B \) any finite subset of \([0, 1]\), and set

\[
f(x, y) = \begin{cases} 1 & \text{if } x \in B, \ y \in \mathbb{Q} \\ 0 & \text{otherwise} \end{cases}.
\]

Then \( f \) is bounded, and continuous except on \( B \times [0, 1] \), a set of content 0, so \( f \) is integrable on \( R \). However, for \( x \in B \), \( f[x] \) is not integrable on \([0, 1]\). The example can be easily modified to also make \( f[y] \) nonintegrable for finitely many \( y \). Replacing \( B \) with the Cantor set \( () \) gives examples where \( f[x] \) and/or \( f[y] \) are nonintegrable for infinitely many \( x, y \).

XIV.3.4.32. The full measure-theoretic form of Fubini’s Theorem implies that if \( f \) is Riemann integrable on \( R \times S \), hence continuous except on a set \( E \) of measure 0, then the vertical cross sections \( E_x \) have measure 0 for almost all \( x \), and similarly \( f[y] \) is Riemann integrable on \( R \) for almost all \( y \). Thus XIV.3.4.31. is nearly as bad as the situation can get. But \( f[x] \) and/or \( f[y] \) can fail to be Riemann integrable for a dense set of \( x \) and/or \( y \) (Exercise XIV.3.11.10.).

XIV.3.4.33. The example XIV.3.4.30. of an \( f \) where the iterated integrals exist but are unequal has an unbounded \( f \). In fact, a bounded example is impossible. For the full Fubini Theorem () implies that if \( f : R \times S \to \mathbb{R} \) is bounded and Lebesgue measurable, a very mild restriction satisfied by any function which can be explicitly described, the iterated integrals (which automatically exist as Lebesgue integrals) are equal. Whether or not \( f \) is Lebesgue measurable (but bounded), if \( f[x] \) and \( f[y] \) are Riemann integrable for all \( x \) and \( y \), the iterated integrals automatically exist as Riemann integrals and are equal ([?], [?]; cf. [Bog07, 3.10.57]). Tonelli’s Theorem () also rules out an example with an unbounded nonnegative (measurable) \( f \), hence an example with (measurable) \( f \) bounded either above or below.

XIV.3.4.34. Some authors have objected to the use of the name “Fubini’s Theorem” for results like XIV.3.4.10. or even XIV.3.4.6., on the grounds that they were “known” to people like Euler, long before Fubini. But the multidimensional Riemann integral was not even defined until the 1870s (by Jordan), so XIV.3.4.10. as stated could not have been known to Euler. The part of this result about interchanging order of integration in an iterated integral (the “Weak Fubini Theorem” XIV.3.5.2.), or its equivalent XIV.3.5.1., might have been a folklore result in Euler’s time, but it is highly doubtful it had been proved (for one thing, at some point in the proof compactness or uniform continuity is needed, conceptions not understood until the nineteenth century). Even justifying interchanging order of summation in a double infinite sum (IV.2.10.8.) requires careful analysis not nailed down until the nineteenth century. Some results which can be viewed as
special cases of Fubini’s Theorem, or more accurately of Cavalieri’s Principle, such as computing the volume of a solid of revolution, had been proved (by the standards of the time), but such results only occasionally involved iterated integrals. I have yet to see clear evidence that anyone proved XIV.3.4.10. or any other true version of Fubini’s Theorem before FUBINI. Thus I see no reason not to continue to use the term “Fubini’s Theorem” generically for these results.

**XIV.3.5. Interchanging Differentiation and Integration**

From Fubini’s Theorem we can obtain an important result on interchanging the order of differentiation and integration: for a function of two or more variables, differentiation with respect to one variable commutes with integration with respect to another variable. We will obtain a more general version in XIV.5.1.1., but from XIV.3.4.10. we can easily get the continuous version, which is adequate for many (perhaps most) applications.

**XIV.3.5.1. Theorem.** Let $I$ be an open interval in $\mathbb{R}$, $[c, d]$ a closed bounded interval in $\mathbb{R}$, and $R = I \times [c, d]$ the corresponding rectangle in $\mathbb{R}^2$. Let $f : R \to \mathbb{R}$ be a function such that $f$ and $\frac{\partial f}{\partial x}$ are continuous on $R$. For $x \in I$ define

$$g(x) = \int_c^d f(x, y) \, dy .$$

Then $g$ is $C^1$ on $I$ and, for $x \in I$,

$$g'(x) = \int_c^d \frac{\partial f}{\partial x}(x, y) \, dy .$$

Slightly informally,

$$\frac{d}{dx} \left( \int_c^d f(x, y) \, dy \right) = \int_c^d \frac{\partial f}{\partial x}(x, y) \, dy$$

(often called “differentiation under the integral sign.”)

In fact, for the proof we only need the part of Fubini’s Theorem asserting that iterated integrals are equal:

**XIV.3.5.2. Theorem.** [Weak Fubini Theorem] Let $R = [a, b] \times [c, d]$ be a rectangle in $\mathbb{R}^2$, and $f : R \to \mathbb{R}$ a continuous function. Then

$$\int_a^b \int_c^d f(x, y) \, dy \, dx = \int_c^d \int_a^b f(x, y) \, dx \, dy .$$

In particular, the two iterated integrals exist as Riemann integrals.

**XIV.3.5.3.** Actually, Theorem XIV.3.5.1. is equivalent to the Weak Fubini Theorem in the sense that each is an easy corollary of the other.

For the proofs, we need the following uniform continuity result, which is a special case of ():
XIV.3.5.4. **Proposition.** Let $I$ be an interval and $[c, d]$ a closed bounded interval in $\mathbb{R}$, and let $f : I \times [c, d] \to \mathbb{R}$ be a continuous function. Then for every $x_0 \in I$ and $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x, y) - f(x_0, y)| < \epsilon$ for any $x \in I$, $|x - x_0| < \delta$, and every $y \in [c, d]$.

XIV.3.5.5. **Corollary.** Let $I$ be an interval and $[c, d]$ a closed bounded interval in $\mathbb{R}$, and let $f : I \times [c, d] \to \mathbb{R}$ be a continuous function. For $x \in I$, define $$\phi(x) = \int_c^d f(x, y) \, dy.$$ Then $\phi$ is a continuous function from $I$ to $\mathbb{R}$.

**Proof:** Let $x_0 \in I$. If $\epsilon > 0$, choose $\delta > 0$ such that $|f(x, y) - f(x_0, y)| < \frac{\epsilon}{d-c}$ for any $x \in I$, $|x - x_0| < \delta$, and every $y \in [c, d]$. Then, for $x \in I$, $|x - x_0| < \delta$,

$$|\phi(x) - \phi(x_0)| = \left| \int_c^d [f(x, y) - f(x_0, y)] \, dy \right| \leq \int_c^d |f(x, y) - f(x_0, y)| \, dy < \frac{\epsilon}{d-c}(d-c) = \epsilon .$$

XIV.3.5.6. We now give the proof of XIV.3.5.1. from XIV.3.5.2.:

**Proof:** For $x \in I$, define $$\phi(x) = \int_c^d \frac{\partial f}{\partial x}(x, y) \, dy .$$ Then $\phi$ is a continuous function from $I$ to $\mathbb{R}$ (XIV.3.5.5). Fix $x_0 \in I$, and fix $a \in I$, $a < x_0$. For $x \in I$, $x > a$, define $$h(x) = \int_a^x \phi(t) \, dt = \int_a^x \int_c^d \frac{\partial f}{\partial x}(t, y) \, dy \, dt .$$ Then, by the Fundamental Theorem of Calculus (FTC1), $h$ is differentiable and $h'(x) = \phi(x)$ for $x \in I$, $x > a$, and in particular $h'(x_0) = \phi(x_0)$. On the other hand, by XIV.3.5.2. and the FTC2, for $x \in I$, $x > a$ we have

$$h(x) = \int_c^d \int_a^x \frac{\partial f}{\partial x}(t, y) \, dt \, dy = \int_c^d [f(x, y) - f(a, y)] \, dy$$

$$= \int_c^d f(x, y) \, dy - \int_c^d f(a, y) \, dy = g(x) - g(a)$$

so $g$ is differentiable and $g'(x) = h'(x) = \phi(x)$ for $x \in I$, $x > a$, and in particular $g$ is differentiable at $x_0$ and $g'(x_0) = \phi(x_0)$.

\[\Box\]
XIV.3.5.7. Conversely, we can easily deduce XIV.3.5.2. from XIV.3.5.1.:

**Proof:** Let $B > b$ and extend $f$ to a continuous function from $[a, B] \times [c, d]$ to $\mathbb{R}$ by setting $f(x, y) = f(b, y)$ for $x \in [b, B]$, $y \in [c, d]$. For $x \in [a, B]$, define

$$g(x) = \int_{c}^{d} \int_{a}^{x} f(t, y) \, dt \, dy.$$ 

Then $g(a) = 0$, and by XIV.3.5.1. and the FTC1 $g$ is differentiable on $(a, B)$ and

$$g'(x) = \int_{c}^{d} \frac{\partial}{\partial x} \left[ \int_{a}^{x} f(t, y) \, dt \right] \, dy = \int_{c}^{d} f(x, y) \, dy$$

for $x \in (a, B)$. Since $g'$ is continuous, by the FTC2 we have, for $x \in (a, B)$,

$$g(x) = g(a) + \int_{a}^{x} g'(t) \, dt = \int_{c}^{d} f(t, y) \, dy \, dt.$$ 

Then XIV.3.5.2. follows by setting $x = b$.

XIV.3.5.8. We can give a direct proof of Theorem XIV.3.5.1., which by XIV.3.5.7. gives an alternate proof of XIV.3.5.2.:

**Proof:** Fix $x_{0} \in I$. Then, for $h$ small enough that $x_{0} + h \in I$, we have

$$\frac{g(x_{0} + h) - g(x_{0})}{h} - \int_{c}^{d} \frac{\partial f}{\partial x}(x_{0}, y) \, dy = \int_{c}^{d} \left[ f(x_{0} + h, y) - \frac{f(x_{0}, y)}{h} \right] \, dy.$$ 

By the Mean Value Theorem, if the interval between $x_{0}$ and $x_{0} + h$ is contained in $I$, then for each $y \in [c, d]$ there is a $u_{y}$ between $x_{0}$ and $x_{0} + h$ (depending on $y$ and $h$) with

$$\frac{f(x_{0} + h, y) - f(x_{0}, y)}{h} = \frac{\partial f}{\partial x}(u_{y}, y).$$

Let $\epsilon > 0$. Since $\frac{\partial f}{\partial x}$ is continuous, by XIV.3.5.4. there is a $\delta > 0$ such that

$$\left| \frac{\partial f}{\partial x}(x, y) - \frac{\partial f}{\partial x}(x_{0}, y) \right| < \frac{\epsilon}{d - c}$$

whenever $x \in I$, $|x - x_{0}| < \delta$, and $y \in [c, d]$. Thus, if $h$ is sufficiently small,

$$\left| \frac{g(x_{0} + h) - g(x_{0})}{h} - \int_{c}^{d} \frac{\partial f}{\partial x}(x_{0}, y) \, dy \right| = \left| \int_{c}^{d} \left[ \frac{\partial f}{\partial x}(u_{y}, y) - \frac{\partial f}{\partial x}(x_{0}, y) \right] \, dy \right|$$

$$\leq \int_{c}^{d} \left| \frac{\partial f}{\partial x}(u_{y}, y) - \frac{\partial f}{\partial x}(x_{0}, y) \right| \, dy < \frac{\epsilon}{d - c}(d - c) = \epsilon.$$ 

Note, however, that this does not give a proof of the full Fubini Theorem, even XIV.3.4.10., without additional argument.

Using the Chain Rule, we can extend XIV.3.5.1. to integration over more general regions:
XIV.3.5.9. Theorem. Let $U$ be an open set in $\mathbb{R}^2$ containing the closed set
\[ S = \{(x, y) : a \leq x \leq b, u(x) \leq y \leq v(x)\} \]
where $u$ and $v$ are differentiable functions from $I = [a, b]$ to $\mathbb{R}$ with $u \leq v$. Let $f : U \rightarrow \mathbb{R}$ be a function such that $f$ and $\frac{\partial f}{\partial x}$ are continuous on $U$. For $x \in I$ define
\[ g(x) = \int_{u(x)}^{v(x)} f(x, y) dy. \]
Then $g$ is differentiable on $I$ and, for $x \in I$,
\[ g'(x) = \int_{u(x)}^{v(x)} \frac{\partial f}{\partial x}(x, y) dy + f(x, v(x))v'(x) - f(x, u(x))u'(x). \]
If $u$ and $v$ are $C^1$, so is $g$.

Proof: Set $\phi(x, u, v) = \int_u^v f(x, y) dy$. Then $\phi$ is defined on a suitable neighborhood $W$ of
\[ \{(x, u, v) : a \leq x \leq b, u(x) \leq u \leq v \leq v(x)\} \]
in $\mathbb{R}^3$, and $g(x) = \phi(x, u(x), v(x))$. By XIV.3.5.1,
\[ \frac{\partial \phi}{\partial x}(x, u, v) = \int_u^v \frac{\partial f}{\partial x}(x, y) dy. \]
In particular, $\frac{\partial \phi}{\partial x}$ exists and is continuous on $W$. By the FTC, $\frac{\partial \phi}{\partial u}(x, u, v) = f(x, v)$ and $\frac{\partial \phi}{\partial v}(x, u, v) = -f(x, u)$ and are thus continuous on $W$. Therefore $\phi$ is differentiable ($C^1$) on $W$. By the Chain Rule, $g$ is differentiable on $I$ and, for $x \in I$,
\[
\begin{align*}
g'(x) &= \frac{\partial \phi}{\partial x}(x, u(x), v(x)) \cdot 1 + \frac{\partial \phi}{\partial u}(x, u(x), v(x))u'(x) + \frac{\partial \phi}{\partial v}(x, u(x), v(x))v'(x) \\
&= \int_{u(x)}^{v(x)} \frac{\partial f}{\partial x}(x, y) dy - f(x, u(x))u'(x) + f(x, v(x))v'(x).
\end{align*}
\]

Multidimensional versions can be obtained similarly, such as (cf. Exercise XIV.3.11.13):
XIV.3.6. Change of Variable

XIV.3.7. Improper Integrals
XIV.3.8. Step Functions

Riemann integration can be alternately described via step functions. There are some subtleties which must be addressed in this approach concerning values on boundaries of rectangles; these subtleties are “wasted” in the sense that they have no ultimate effect, but they are necessary for a logically complete development, and they can be handled cleanly if done right.

XIV.3.8.1. Definition. A step function on $\mathbb{R}^p$ is a finite linear combination of indicator functions of rectangles (degenerate rectangles are allowed).

XIV.3.8.2. Any (finite) linear combination of step functions is a step function.

XIV.3.8.3. If $\phi = \sum_{k=1}^{n} \alpha_k \phi_{R_k}$ is a step function, there is a rectangle $R$ such that $R_k \subseteq R$ for all $k$. A partition $\mathcal{P}$ of $R$ is $\phi$-fine if each $R_k$ is a union of rectangles in $\mathcal{P}$. By (), every step function $\phi$ has a $\phi$-fine partition of any sufficiently large rectangle. Any refinement of a $\phi$-fine partition is $\phi$-fine.

XIV.3.8.4. In the step function approach to Riemann integration, it is very convenient to allow partitions which include degenerate rectangles. In fact, it is most convenient to work with full partitions $\mathcal{P}$, in which any nonempty intersection of rectangles in $\mathcal{P}$ is also in $\mathcal{P}$ (all such intersections are, of course, degenerate since the rectangles in $\mathcal{P}$ are nonoverlapping). A grid in which every one-dimensional factor has all transition points included as degenerate intervals is a full partition, so any partition has a full refinement.

If $\phi$ is a step function and $\mathcal{P}$ is a full $\phi$-fine partition, then $\phi$ can be written as a linear combination of indicator functions of rectangles in $\mathcal{P}$. This is an immediate consequence of the following fact, which has a straightforward but notationally complex proof by induction on $n$ left to the reader:

XIV.3.8.5. Proposition. Let $X$ be a set, and $A_1, \ldots, A_n$ subsets of $X$. Then

$$\chi_{A_1 \cup \cdots \cup A_n} = \sum_{k=1}^{n} \chi_{A_k} - \sum_{j<k} \chi_{A_j \cap A_k} + \sum_{i<j<k} \chi_{A_i \cap A_j \cap A_k} - \cdots + (-1)^{n+1} \chi_{A_1 \cap \cdots \cap A_n}$$

where each sum is over all strictly increasing sets of distinct indices.

XIV.3.8.6. Each rectangle $R$ has an “interior” $R^o$ consisting of all points of $R$ not contained in a boundary rectangle of smaller dimension. If $R$ is a singleton, set $R^o = R$. Note that $R^o$ is not the topological interior of $R$ in $\mathbb{R}^p$ if $R$ is degenerate, but is rather the topological interior relative to the affine subspace of $\mathbb{R}^p$ generated by $R$. $R^o$ is a relatively open subset of $R$. If $\mathcal{P}$ is a full partition of a rectangle $R$, for each $R_k \in \mathcal{P}$ we have that $R^o_k$ is the set of points of $R_k$ not contained in any $R_j \in \mathcal{P}$ properly contained in $R_k$. Then $R$ is the disjoint union of the $R^o_k$. Any linear combination $\phi$ of the indicator functions of the $R^o_k$ is a step function (use XIV.3.8.5.) for which $\mathcal{P}$ is $\phi$-fine, and any step function $\psi$ for which $\mathcal{P}$ is $\psi$-fine can be written uniquely as a linear combination of these indicator functions.

XIV.3.8.7. It follows immediately that any finite maximum or minimum of step functions is a step function. [Consider a full partition which is fine for all the step functions.]
Integrals of Step Functions

XIV.3.8.8. If $\phi = \sum_{k=1}^{m} \alpha_k \chi_{R_k}$ is a step function, then $\phi$ is Riemann integrable over any rectangle $R$ containing all the $R_k$ and

$$\int_R \phi \, dV = \sum_{k=1}^{m} \alpha_k V(R_k).$$

It is fairly easily verified using XIV.3.8.5. that the sum on the right depends only on $\phi$ and not on how it is written as a linear combination of indicator functions; thus it can be taken as the definition of the integral of $\phi$ over $R$. The integral of step functions so defined is easily verified to be linear, additive, and monotone. It also has a continuity property (XIV.3.8.10.).

We first note a way of showing certain sets have measure 0 using step functions. This will be generalized to a characterization of measure 0 (XIV.3.8.14.).

XIV.3.8.9. Lemma. Let $R$ be a rectangle in $\mathbb{R}^p$, and $A \subseteq R$. If, for every $\epsilon > 0$, there is a nonnegative step function $\phi$ such that $\phi(x) \geq 1$ for all $x \in A$ and $\int_R \phi \, dV < \epsilon$, then $A$ has measure 0.

Proof: Let $\epsilon > 0$ and $\phi$ the corresponding step function. Let $\phi = \sum \alpha_k \chi_{R_k}$ be the canonical representation of $\phi$ (XIV.3.8.6.), and let $k_1, \ldots, k_n$ be the $k$ for which $\alpha_k \geq 1$. Then $A \subseteq \bigcup_{j=1}^{n} R_{k_j}$ and

$$\sum_{j=1}^{n} V(R_{k_j}) \leq \sum_{j=1}^{n} \alpha_{k_j} V(R_{k_j}) \leq \int_R \phi \, dV < \epsilon.$$

Since $\epsilon > 0$ is arbitrary, $A$ has measure 0.

XIV.3.8.10. Proposition. Let $(\psi_n)$ be a nonincreasing sequence of nonnegative step functions supported on a fixed rectangle $R$. If

$$\lim_{n \to \infty} \int_R \psi_n \, dV = \inf_n \int_R \psi_n \, dV = 0$$

Then $\psi_n \to 0$ pointwise a.e.

Proof: It suffices to show that $A_m = \{x \in R : \inf_n \psi_n(x) \geq \frac{1}{m}\}$ has measure 0 for all $m \in \mathbb{N}$, since $\psi_n \to 0$ pointwise on the complement of $\bigcup_m A_m$. Fix $m$. For any $\epsilon > 0$, there is an $n$ such that $\int_R \psi_n \, dV < \frac{\epsilon}{m}$. Then $\phi = m \psi_n$ is a step function with $\phi(x) \geq 1$ for all $x \in A_m$ and $\int_R \phi \, dV < \epsilon$. Thus $A_m$ has measure 0 by XIV.3.8.9..

We will prove an important converse in ().

XIV.3.8.11. The Riemann integral can be alternately defined and constructed using this integral for step functions, without using the upper and lower sum definitions previously used ().
Maximum and Minimum Step Functions

XIV.3.8.12. Let $R$ be a rectangle and $f$ a bounded function on $R$. If $\mathcal{P}$ is a full partition of $R$, then there is a unique largest step function $\phi$ for which $\mathcal{P}$ is $\phi$-fine and $\phi \leq f$. If $\mathcal{P} = \{R_1, \ldots, R_m\}$, then $\phi$ is the function whose value on $R_k$ is the infimum of $f$ on $R_k$. Similarly, there is a smallest step function $\psi$ for which $\mathcal{P}$ is $\psi$-fine and $\psi \leq f$. We have

$$S(f, \mathcal{P}) \leq \int_R \phi \, dV \leq \int_R f \, dV \leq \int_R \psi \, dV \leq S(f, \mathcal{P}).$$

Write $\phi(f, \mathcal{P})$ and $\psi(f, \mathcal{P})$ for $\phi$ and $\psi$. If $Q$ is a refinement of $\mathcal{P}$, then $\phi(f, \mathcal{P}) \leq \phi(f, Q)$ and $\psi(f, Q) \leq \psi(f, \mathcal{P})$.

XIV.3.8.13. Proposition. Let $R$ be a rectangle in $\mathbb{R}^p$, and $f : \mathbb{R} \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) $f$ is Riemann integrable on $R$.

(ii) There are sequences $(\phi_n)$ and $(\psi_n)$ of step functions supported on $R$ with

$$\phi_n \leq \phi_{n+1} \leq f \leq \psi_{n+1} \leq \psi_n$$

on $R$ for all $n$ and

$$\lim_{n \to \infty} \int_R \phi_n \, dV = \sup_n \int_R \phi_n \, dV = \inf_n \int_R \psi_n \, dV = \lim_{n \to \infty} \int_R \psi_n \, dV.$$

Proof: (ii) $\Rightarrow$ (i) since

$$\int_R \phi_n \, dV \leq \int_R f \, dV \leq \int_R \psi_n \, dV$$

for each $n$ by monotonicity.

For (i) $\Rightarrow$ (ii), choose a sequence $\mathcal{P}_n$ of full partitions of $R$ such that $\mathcal{P}_{n+1}$ refines $\mathcal{P}_n$ for all $n$ and

$$\lim_{n \to \infty} |S(f, \mathcal{P}_n) - S(f, \mathcal{P}_n)| = 0.$$ Set $\phi_n = \phi(f, \mathcal{P}_n)$, $\psi_n = \psi(f, \mathcal{P}_n)$ and apply XIV.3.8.12.

Two more equivalent conditions will be added in XIV.3.10.6.

Characterization of Sets of Measure Zero

Sets of measure 0 can be characterized using integrals of step functions:

XIV.3.8.14. Theorem. Let $R$ be a rectangle in $\mathbb{R}^p$, and $A \subseteq R$. Then $A$ has measure 0 if and only if, for every $\epsilon > 0$, there is a nondecreasing sequence $(\phi_n)$ of step functions such that $\sup_n \phi_n(x) \geq 1$ for all $x \in A$ and $\int_R \phi_n \, dV < \epsilon$ for all $n$.  

1588
XIV.3.9. The Monotone Convergence Theorem

The Monotone Convergence Theorem is an important continuity property of integrals:

XIV.3.9.1. Theorem. [Monotone Convergence Theorem for Riemann Integration] Let \( R \) be a rectangle in \( \mathbb{R}^p \), and \( (f_n) \) a nondecreasing sequence of nonnegative Riemann integrable functions on \( R \) converging pointwise a.e. on \( R \) to a Riemann integrable function \( f \). Then

\[
\int_R f \, dV = \lim_{n \to \infty} \int_R f_n \, dV = \sup_n \int_R f_n \, dV.
\]

Note that it is the sequence which is nondecreasing, i.e. \( f_n(x) \leq f_{n+1}(x) \) for all \( n \) and all \( x \in R \).

XIV.3.9.2. There are two difficulties with this statement. First, the hypothesis that the limit function \( f \) is Riemann integrable on \( R \) is not automatic, even if the \( f_n \) are uniformly bounded. For example, let \( (r_n) \) be an enumeration of the rational numbers in \( [0, 1] \), and define \( f_n(x) \) to be 1 if \( x = r_k \) for some \( k, 1 \leq k \leq n \), and \( f_n(x) = 0 \) otherwise. Then \( f_n \) is 0 except at finitely many points, hence is Riemann integrable on \( [0, 1] \), but \( f_n \to \chi_{[0,1]} \) pointwise on \( [0,1] \), and \( \chi_{[0,1]} \) is not Riemann integrable on \( [0,1] \). Thus the Riemann integrability of \( f \) must be included as a hypothesis. To avoid this difficulty, replacing \( f_n \) by \( f - f_n \) and using linearity of the Riemann integral the theorem may be more cleanly stated:

XIV.3.9.3. Theorem. Let \( R \) be a rectangle in \( \mathbb{R}^p \), and \( (f_n) \) a nonincreasing sequence of nonnegative Riemann integrable functions on \( R \) converging pointwise a.e. on \( R \) to 0. Then

\[
\lim_{n \to \infty} \int_R f_n \, dV = \inf_n \int_R f_n \, dV = 0.
\]

XIV.3.9.4. There is a second difficulty: proving this theorem. Although a direct proof can be given, it is quite complicated to do so. The simplest approach is to work via the Lebesgue integral and prove a more general theorem applicable to arbitrary Lebesgue measurable functions (XIV.4.4.1). In fact, overcoming both the difficulties with the Riemann integration statement of the theorem was one of the principal motivations for developing Lebesgue integration theory.

XIV.3.9.5. We will content ourselves here with proving only the following special case, which will be crucial in developing the Daniell integral ()

XIV.3.9.6. Theorem. [Monotone Convergence Theorem for Step Functions] Let \( R \) be a rectangle in \( \mathbb{R}^p \), and \( (\psi_n) \) a nonincreasing sequence of nonnegative step functions on \( R \) converging pointwise a.e. on \( R \) to 0. Then

\[
\lim_{n \to \infty} \int_R \psi_n \, dV = \inf_n \int_R \psi_n \, dV = 0.
\]
Proof: Let $A$ be the set of $x \in R$ for which $(\psi_n(x))$ fails to converge to 0, and $B$ the set of $x$ at which at least one $\psi_n$ is discontinuous, and $C = A \cup B$. Then $C$ has measure 0. Let $M$ be a bound for $\phi_1$ on $R$.

Let $\epsilon > 0$. There is a sequence $(R^0_k)$ of open rectangles covering $C$ for which

$$\sum_{k=1}^{\infty} V(R^0_k) < \frac{\epsilon}{2M}.$$ 

If $E = R \setminus \cup_k R^0_k$, then for each $x \in E$ there is an $n_x$ and an open rectangle $R^*_x$ containing $x$ such that $\psi_{n_x}$ takes a constant value less than $\frac{\epsilon}{2M}$ on $R^*_x$. The $R^0_k$ and $R^*_x$ cover $R$, so by compactness finitely many $R^0_k$ and $R^*_x$ cover $E$, say $R^0_{k_1}, \ldots, R^0_{k_r}, R^*_x, \ldots, R^*_m$. Let $R_{x_k}$ be the closure of $R^*_x$. By subdividing and eliminating overlaps we may assume the $R_{x_k}$ are nonoverlapping, so $\sum_{k=1}^{m} V(R_{x_k}) \leq V(R)$.

Set $N = \max \{ n_{x_1}, \ldots, n_{x_m} \}$. Then if $n \geq N$, we have

$$\int_R \psi_n dV \leq \sum_{j=1}^{r} \int_{R^0_{k_j}} \psi_n dV + \sum_{k=1}^{m} \int_{R_{x_k}} \psi_n dV < \frac{\epsilon}{2M} \sum_{k=1}^{\infty} M V(R^0_k) + \sum_{k=1}^{m} \frac{\epsilon}{2V(R)} V(R_{x_k}) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$ 

\[ \square \]

XIV.3.9.7. What this argument really shows is a special case of Egorov’s Theorem (XIII.1.8.1.): under the hypotheses, $\psi_n$ for $n$ sufficiently large is actually uniformly close to 0 on the complement of a set of small measure. This is also reminiscent of Dini’s Theorem ( ).
XIV.3.10. Characterization of Riemann Integrable Functions

Using the Monotone Convergence Theorem (XIV.3.9.6.), we can add a third equivalent condition to the characterization of Riemann integrability:

Theorem. Let $R$ be a rectangle in $\mathbb{R}^p$, and $f : R \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) $f$ is Riemann integrable on $R$.

(ii) There are sequences $(\phi_n)$ and $(\psi_n)$ of step functions supported on $R$ with

$$\phi_n \leq \phi_{n+1} \leq f \leq \psi_{n+1} \leq \psi_n$$

on $R$ for all $n$ and

$$\lim_{n \to \infty} \int_R \phi_n dV = \sup_n \int_R \phi_n dV = \inf_n \int_R \psi_n dV = \lim_{n \to \infty} \int_R \psi_n dV.$$

(iii) There are sequences $(\phi_n)$ and $(\psi_n)$ of step functions supported on $R$ with

$$\phi_n \leq \phi_{n+1} \leq f \leq \psi_{n+1} \leq \psi_n$$

on $R$ for all $n$ and $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. on $R$.

Proof: (i) $\iff$ (ii) by XIV.3.8.13.. For (ii) $\Rightarrow$ (iii), let $\phi_n, \psi_n$ be as in (ii). Then $(\psi_n - \phi_n)$ is a nonincreasing sequence of nonnegative step functions with

$$\lim_{n \to \infty} \int_R (\psi_n - \phi_n) dV = 0$$

and hence $\psi_n - \phi_n \to 0$ pointwise a.e., i.e. $\liminf_n \psi_n(x) = \sup_n \phi_n(x)$ for almost all $x \in R$. Since $\sup_n \phi_n(x) \leq f(x) \leq \inf_n \psi_n(x)$ for all $x \in R$, it follows that $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. (XIV.3.8.10.).

Conversely, suppose $\phi_n$ and $\psi_n$ are as in (iii). Then $(\psi_n - \phi_n)$ is a nonincreasing sequence of nonnegative step functions on $R$ and $\psi_n - \phi_n \to 0$ pointwise a.e. Then by the Monotone Convergence Theorem we have

$$\lim_{n \to \infty} \int_R (\psi_n - \phi_n) dV = 0$$

so (ii) is satisfied.

Upper and Lower Envelopes

There is an alternate way of describing condition (iii) of XIV.3.10.1. which leads to Lebesgue’s characterization of Riemann integrability. Let $f$ be a bounded function on a rectangle $R$. For each $a \in R$, define

$$f'(a) = \min \left( f(a), \liminf_{x \to a} f(x) \right)$$
XIV.3.10.3. Let $\mathcal{P}$ be a partition of $R$, and let $\phi(f, \mathcal{P})$ and $\psi(f, \mathcal{P})$ be the step functions defined in XIV.3.8.12. Then $\phi_n(f, \mathcal{P}) \leq f$ on the interior of any nondegenerate rectangle in $\mathcal{P}$, hence $\phi_n(f, \mathcal{P}) \leq f$ a.e. on $R$. Thus, if $\phi$ is any step function with $\phi \leq f$ a.e. on $R$, then $\phi \leq f$ a.e. on $R$. Similarly, if $\psi$ is any step function with $f \leq \psi$ a.e. on $R$, then $\overline{f} \leq \psi$ a.e. on $R$.

PROOF: The union of the degenerate rectangles in $\mathcal{P}_n$ has measure 0 (zero content) for each $n$. Hence if $A$ is the set of points of $R$ which are interior points of nondegenerate rectangles of $\mathcal{P}_n$ for every $n$, then $R \setminus A$ has measure 0. Fix $a \in A$, and let $\epsilon > 0$. There is a $\delta > 0$ such that $f(x) > f(a) - \epsilon$ for all $x \in R \cap B_\delta(a)$. Then there is a $N$ with $\|\mathcal{P}_n\| < \delta$ for $n \geq N$; for such an $n$ the nondegenerate rectangle in $\mathcal{P}_n$ containing $a$ is contained in $B_\delta(a)$ and thus $f(a) - \epsilon \leq \phi(f, \mathcal{P}_n)(a) \leq f(a)$. Therefore $\phi(f, \mathcal{P}_n)(a) \to f(a)$. So $\phi(f, \mathcal{P}_n) \to f$ pointwise on $A$. Similarly, $\psi(f, \mathcal{P}_n) \to \overline{f}$ pointwise on $A$.

XIV.3.10.5. In fact, it is easily verified () that $\underline{f}$ is precisely the supremum of all step functions $\phi$ satisfying $\phi \leq f$, and $\overline{f}$ the infimum of all step functions $\psi$ satisfying $f \leq \psi$.

We can then add another equivalent condition for Riemann integrability:

XIV.3.10.6. Theorem. Let $R$ be a rectangle in $\mathbb{R}^p$, and $f : \mathbb{R} \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) $f$ is Riemann integrable on $R$.

(ii) There are sequences $(\phi_n)$ and $(\psi_n)$ of step functions supported on $R$ with

\[ \phi_n \leq \phi_{n+1} \leq f \leq \psi_{n+1} \leq \psi_n \]

on $R$ for all $n$ and

\[ \lim_{n \to \infty} \int_R \phi_n \, dV = \sup_n \int_R \phi_n \, dV = \inf_n \int_R \psi_n \, dV = \lim_{n \to \infty} \int_R \psi_n \, dV. \]

(iii) There are sequences $(\phi_n)$ and $(\psi_n)$ of step functions supported on $R$ with

\[ \phi_n \leq \phi_{n+1} \leq f \leq \psi_{n+1} \leq \psi_n \]

on $R$ for all $n$ and $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. on $R$. 

1592
(iv) We have \( \underline{f} = \overline{f} \) a.e. on \( R \) (and in particular \( \underline{f} = f \) a.e. and \( \overline{f} = f \) a.e. on \( R \)).

**Proof:** (i)–(iii) are equivalent by XIV.3.10.1. If \( \phi_n, \psi_n \) are as in (iii), we have
\[
\phi_n \leq f \leq \psi_n \quad \text{a.e.}
\]
for each \( n \); since \( \phi_n \to f \) a.e. and \( \psi_n \to f \) a.e. we have \( f = \underline{f} = \overline{f} \) a.e.

Conversely, if (iv) is satisfied, let \( (P_n) \) be a sequence of full partitions of \( R \) such that \( \|P_n\| \to 0 \) as \( n \to \infty \), such that \( P_{n+1} \) refines \( P_n \) for all \( n \). Then by XIV.3.10.4. \( \phi(f, P_n) \not\to f \) a.e. and \( \psi(f, P_n) \not\to f \) a.e. on \( R \), so (iii) is satisfied.

Noting that \( \underline{f} = f = \overline{f} \) precisely at the points of continuity of \( f \), we obtain Lebesgue’s characterization:

**XIV.3.10.7. Corollary.** Let \( R \) be a rectangle in \( \mathbb{R}^p \), and \( f : R \to \mathbb{R} \). Then \( f \) is Riemann integrable on \( R \) if and only if \( f \) is continuous almost everywhere and bounded on \( R \).
XIV.3.11. Exercises

XIV.3.11.1. (a) Show that a function which is continuous almost everywhere and zero on a dense set is zero almost everywhere. [Show it is zero wherever it is continuous.]

(b) If \( f \) and \( g \) are functions which are each continuous almost everywhere, and \( f \) and \( g \) agree on a dense set, then \( f = g \) a.e.

XIV.3.11.2. Show that a finite union of sets of zero content has zero content. [You can mimic the proof of XIV.2.7.6., but there is an easier argument.]

XIV.3.11.3. Prove Proposition XIV.2.7.11.

XIV.3.11.4. This problem is a continuation of V.8.6.7.

(a) Show that every regulated function on an interval \( I \) is Riemann-integrable on any closed bounded subinterval of \( I \). Then deduce V.8.6.7.(c) from XIV.2.9.3.

(b) Let \( g(x) = \sin(1/x) \) for \( x \neq 0 \), and \( g(0) = 0 \). Let \( (r_n) \) be an enumeration of \( \mathbb{Q} \), and set

\[
 f(x) = \sum_{n=1}^{\infty} 2^{-n} g(x - r_n) .
\]
Show that $f$ is discontinuous at every point of $\mathbb{Q}$ and continuous at every point of $\mathbb{J}$.

(c) Show that $f$ is Riemann-integrable over every closed bounded subinterval of $\mathbb{R}$ (give two proofs, one using XIV.2.8.5. and one using ()), but $f$ is not regulated on any interval.

(d) Show using XIV.2.9.8.(iii) and V.8.5.5. that $f$ has an antiderivative on $\mathbb{R}$.

XIV.3.11.5. Let $f$ be a bounded function on an interval $[a, b]$, and $\overline{f}$ its upper envelope (XIV.2.8.1.).

(a) Show that

$$\overline{f} = \inf \{ \psi : \psi \geq f, \psi \text{ an upper semicontinuous step function} \} .$$

[Use the argument of the proof of XIV.2.8.2.]

(b) Show that, for every $x_0 \in [a, b],$

$$\overline{f}(x_0) = \max \left( f(x_0), \limsup_{x \to x_0} f(x) \right).$$

(c) Give similar characterizations of $f$.

XIV.3.11.6. Let $E$ be a subset of $[a, b], and consider the indicator function $\chi_E$.

(a) Show that the upper envelope (XIV.2.8.1.) $\overline{\chi_E}$ is $\chi_E$, and the lower envelope $\underline{\chi_E}$ is $\chi_E^\circ$.

(b) Show that $\chi_E$ is Riemann-integrable on $[a, b]$ if and only if $\partial E$ has zero content (), i.e. $E$ is a Jordan region () in $\mathbb{R}$.

XIV.3.11.7. Use the continuous version of Fubini’s Theorem (XIV.3.4.7.; XIV.3.5.2. suffices) to prove the following version of the equality of mixed partials VIII.2.2.4. (compare with VIII.2.2.8.):

**Theorem.** Let $U$ be an open set in $\mathbb{R}^2$, and $f : U \to \mathbb{R}$ a function such that $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist and are continuous on $U$. Then $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ on $U$.

(a) Show using the Fundamental Theorem of Calculus that if $R = [a, b] \times [c, d] \subseteq U$, then

$$\int_c^b \int_c^d \frac{\partial^2 f}{\partial y \partial x}(x, y) \, dy \, dx = \int_c^d \int_a^b \frac{\partial^2 f}{\partial x \partial y}(x, y) \, dx \, dy.$$

(b) If there is an $(x_0, y_0) \in U$ with $\frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) > \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$, then there is a rectangle $R$ around $(x_0, y_0)$ contained in $U$ such that $\frac{\partial^2 f}{\partial y \partial x} - \frac{\partial^2 f}{\partial x \partial y} > 0$ on $R$, and hence

$$\int_R \left[ \frac{\partial^2 f}{\partial y \partial x} - \frac{\partial^2 f}{\partial x \partial y} \right] \, dV > 0.$$

Draw a similar conclusion if there is an $(x_0, y_0) \in U$ with $\frac{\partial^2 f}{\partial y \partial x}(x_0, y_0) < \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$.

(c) Use Fubini’s Theorem to obtain a contradiction from (a) and (b).

(d) Extend to functions of several variables and higher-order mixed partials.
XIV.3.11.8. Explain why Theorem XIV.3.1.26. is not an immediate corollary of Proposition XIV.3.1.16. Identify what else in addition to XIV.3.1.16. is needed, and prove it.

XIV.3.11.9. Let $A$ be a Jordan region in $\mathbb{R}^n$, and $f : A \to \mathbb{R}$ a nonnegative continuous function. Let

$$B = \{(x, y) : x \in A, 0 \leq y \leq f(x)\} \subseteq \mathbb{R}^{n+1}.$$ 

Show that $B$ is a Jordan region in $\mathbb{R}^{n+1}$, and that

$$V_{n+1}(B) = \int_A f \, dV_n.$$

XIV.3.11.10. Let $E$ be the subset of $R = [0, 1]^2 \subseteq \mathbb{R}^2$ consisting of the points on the $x$-axis and, for each rational $x = \frac{p}{q}$ in lowest terms, the points in a Cantor set of positive measure () on the vertical segment

$$\left\{ (x, y) : 0 \leq y \leq \frac{1}{q}\right\}.$$

(i) Show that $E$ is a closed set of measure 0, hence of content 0.
(ii) Show that $f = \chi_E$ is Riemann integrable over $R$, but that $f[x]$ is not Riemann integrable on $[0, 1]$ if $x$ is rational.
(iii) Modify this example to one for which $f[y]$ is also not Riemann integrable for $y$ rational.

XIV.3.11.11. Evaluate

$$\int_0^1 \int_x^1 e^{-y^2} \, dy \, dx$$

by applying Fubini’s Theorem to interchange the order of integration.


XIV.3.11.13. (a) Prove XIV.3.5.10.
(b) Generalize to the case where $S$ is replaced by an arbitrary Jordan region in $\mathbb{R}^q$ (cf. XIV.3.4.14.).
XIV.4. The Lebesgue Integral

This section treats the heart of modern integration theory, the general integral first defined by Lebesgue and subsequently refined by other authors. The term “Lebesgue integral” is commonly used to mean two different things:

(i) Lebesgue’s generalization of the Riemann integral for functions from \( \mathbb{R} \) to \( \mathbb{R} \).

(ii) The general integral for real-valued functions on a measure space, often called the abstract Lebesgue integral.

As the integral is normally done today, the Lebesgue integral on \( \mathbb{R} \) is just the special case of the general abstract Lebesgue integral for the measure space \((\mathbb{R}, \mathcal{M}, \lambda)\), and it is no easier and even no different to develop this special case than the general integral. Thus we will work generally and point out the few important special properties of the Lebesgue integral on \( \mathbb{R} \) as we go. We will usually suppress the name “Lebesgue” from the general abstract integral, except when this integral must be distinguished from integrals defined in other ways, and mostly reserve it for the special case of Lebesgue measure on \( \mathbb{R} \) (or \( \mathbb{R}^n \)).

If \((X, \mathcal{A}, \mu)\) is a measure space, we want to assign an extended real number to any extended real-valued measurable function \( f \), and \( E \in \mathcal{A} \), called the integral of \( f \) over \( E \) and denoted \( \int_E f \, d\mu \), in such a way that the integral has standard nice properties we would like and expect. It turns out not to be possible to do this for every extended real-valued measurable function in general (because of the obstacle of defining \( \infty - \infty \)), but only for nonnegative functions and general functions whose positive and negative parts are not “too large” (“integrable” functions). We will build up the integral in three steps (the process is called bootstrapping):

(i) Simple functions.

(ii) Nonnegative extended real-valued measurable functions.

(iii) Integrable functions.

XIV.4.1. Properties of the Integral

We can give a list of properties we want the integral to have. This list might be regarded as a “wish list” at the moment, but we will be able to get almost everything on the list.

XIV.4.1.1. Let \((X, \mathcal{A}, \mu)\) be a measure space. We would like the assignment \((f, E) \mapsto \int_E f \, d\mu\) to have the following properties:

(i) [MONOTONICITY] If \( f \leq g \), then \( \int_E f \, d\mu \leq \int_E g \, d\mu \) for any \( E \).

(ii) [LINEARITY] \( \int_E (f + g) \, d\mu = \int_E f \, d\mu + \int_E g \, d\mu \) for any \( f, g, E \).

(iii) [HOMOGENEITY] \( \int_E \alpha f \, d\mu = \alpha \int_E f \, d\mu \) for any \( f, E, \alpha \in \mathbb{R} \) (where we always interpret \( 0 \cdot \infty = 0 \)).

(iv) [ADDITIVITY] If \((E_k)\) are pairwise disjoint in \( \mathcal{A} \) and \( E = \bigcup_k E_k \), then \( \int_E f \, d\mu = \sum_{k=1}^{\infty} \int_{E_k} f \, d\mu \).

(v) If \( f = g \text{ a.e. (} \mu\text{-a.e.)} \), then \( \int_E f \, d\mu = \int_E g \, d\mu \) for any \( E \).
(vi) If $\mu(E) = 0$, then $\int_E f \, d\mu = 0$ for any $f$.

(vii) [Continuity] If $f_n \to f$ pointwise, then $\int_E f_n \, d\mu \to \int_E f \, d\mu$ for any $E$.

(viii) [Normalization] $\int_E 1 \, d\mu = \mu(E)$ for any $E$.

**XIV.4.1.2.** Actually, (ii) and (iii) together are usually called “linearity,” and can be replaced by the single condition

$$(ii') \, \int_E (\alpha f + g) \, d\mu = \alpha \int_E f \, d\mu + \int_E g \, d\mu$$

for any $f, g, E$, and $\alpha \in \mathbb{R}$.

We have used “linearity” for (ii) instead of “additivity” to distinguish it from (iii), although the way we will proceed it will turn out that (iii), at least for finitely many sets, is a consequence of (ii).

**XIV.4.1.3.** We will be able to get everything on this list except Continuity, although in several others we will have to make modest restrictions to avoid the $\infty - \infty$ problem. Continuity in the general form of (vii) is simply not possible in general (XIV.4.7.1.); however, the major convergence theorems of integration theory (MCT, DCT) will show that continuity holds under some reasonable additional assumptions.

**XIV.4.2.** Integral of Simple Functions

Recall the definition of a simple function $\phi$. To avoid verbosity, we will restrict the definition to only include functions which are nonnegative and measurable:

**XIV.4.2.1.** Definition. Let $(X, \mathcal{A})$ be a measurable space. A simple function on $(X, \mathcal{A})$ is a nonnegative real-valued measurable function which takes only finitely many values.

If $\phi$ is a simple function on $(X, \mathcal{A})$ taking exactly the values $\alpha_1, \ldots, \alpha_n$, for each $k$ set

$$A_k = \{ x \in X : \phi(x) = \alpha_k \} .$$

Then $\{A_1, \ldots, A_n\}$ is a partition of $X$ into nonempty measurable subsets, and

$$\phi = \sum_{k=1}^{n} \alpha_k \chi_{A_k}$$

is the canonical representation of $\phi$.

In some references, a term with $\alpha_k = 0$ is deleted from the canonical representation, but we prefer to keep such a term to force $\{A_1, \ldots, A_n\}$ to be a partition of $X$. This will have no practical effect.

**XIV.4.2.2.** Any nonnegative (finite) linear combination of indicator functions of measurable sets is a simple function. A sum

$$\phi = \sum_{j=1}^{m} \beta_j \chi_{B_j}$$

may not be a canonical representation of $\phi$ for four reasons:

1. The $B_j$ may not be disjoint.
The $\beta_j$ may not all be distinct.

We may not have $\cup_j B_j = X$.

Some of the $B_j$ may be empty.

(3) and (4) are easy to remedy: for (3), set $B_{m+1} = X \setminus \cup_{j=1}^m B_j$ and $\beta_{m+1} = 0$, and for (4), just delete any terms with $B_j = \emptyset$. The others can also be systematically fixed: for (1), let $A_1, \ldots, A_n$ consist of all nonempty finite intersections of the $B_j$ and their complements (i.e. the nonempty sets in the partition of $\cup_{j=1}^m B_j$ generated by the $B_j$), and if $A_k = B_{j_1} \cap \cdots \cap B_{j_r} \cap B_{j_{r+1}} \cap \cdots B_{j_c}^c$, set $\alpha_k = \beta_{j_1} + \cdots + \beta_{j_c}$. Then for (2) just combine together terms with equal $\beta_j$, using that $\chi_{A\cup B} = \chi_A + \chi_B$ if $A$ and $B$ are disjoint. If these corrections are done in the order (1), (3), (4), (2), each only needs to be done once and the representation is converted into canonical form.

A representation in which all the $B_j$ are disjoint, but (2), (3), (4) are allowed, is said to be in semicanonical form.

XIV.4.2.3. Note that if $\phi$ is simple and $E \in \mathcal{A}$, then $\phi \chi_E$ is also simple: if $\phi = \sum_{j=1}^m \beta_j \chi_{B_j}$, then $\phi \chi_E = \sum_{j=1}^m \beta_j \chi_{B_j \cap E}$. If the representation of $\phi$ is in semicanonical form, so is this representation of $\phi \chi_E$; but the representation of $\phi \chi_E$ is not necessarily in canonical form even if the representation of $\phi$ is.

The Integral of a Simple Function

We can define the integral of a simple function in the only way possible if we hope to have linearity and normalization:

XIV.4.2.4. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\phi$ a simple function on $(X, \mathcal{A})$. Write $\phi = \sum_{k=1}^n \alpha_k \chi_{A_k}$ in canonical form. Then

$$\int_X \phi \, d\mu = \sum_{k=1}^n \alpha_k \mu(A_k)$$

where $0 \cdot \infty$ is always interpreted to be 0, and $r \cdot \infty = \infty$ if $r > 0$. If $E \in \mathcal{A}$, set

$$\int_E \phi \, d\mu = \int_X \phi \chi_E \, d\mu .$$

A few simple properties are immediate from the definition:

XIV.4.2.5. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $E \in \mathcal{A}$.

(i) If $\phi$ is a simple function on $(X, \mathcal{A})$, and $r \geq 0$, then

$$\int_E r \phi \, d\mu = r \int_E \phi \, d\mu .$$
(ii) If $\phi$ and $\psi$ are simple functions and $\phi = \psi$ a.e. on $E$ (i.e. $\phi \chi_E = \psi \chi_E$ a.e.), then

$$\int_E \phi \, d\mu = \int_E \psi \, d\mu .$$

(iii) If $\mu(E) = 0$, then $\int_E \phi \, d\mu = 0$ for any simple function $\phi$.

**Proof:** (i): If $r > 0$ and $\phi \chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$ is the canonical representation, then the canonical representation of $r \phi \chi_E$ is $\sum_{k=1}^n \alpha_k \chi_{A_k}$. And if $r = 0$, both sides are 0 (recall the convention that $0 \cdot \infty = 0$).

(ii): If $\phi \chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$ and $\psi \chi_E = \sum_{j=1}^m \beta_j \chi_{B_j}$ are the canonical representations, then each $\alpha_k$ corresponds to a unique $\beta_j$ unless $\mu(A_k) = 0$ in which case the term $\alpha_k \mu(A_k)$ can be deleted from the sum, and conversely: thus once zero terms are deleted the terms in the sums are in one-one correspondence. For each $k$ the sets $A_k$ and $B_{j_k}$ differ by a null set, so $\mu(A_k) = \mu(B_{j_k})$. Thus the remaining terms are numerically identical.

(iii): If $\mu(E) = 0$, then in the canonical representation $\phi \chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$, in any term with $\alpha_k > 0$ we have $A_k \subseteq E$, so $\mu(A_k) = 0$. Thus all the terms in $\sum_{k=1}^n \alpha_k \mu(A_k)$ are zero. \(\diamondsuit\)

The key technical fact giving reasonable properties of the integral defined this way is that the formula in the definition holds for any representation of $\phi$, not just the canonical one:

**XIV.4.2.6. Proposition.** Let $(X, A, \mu)$ be a measure space, and $\phi$ a simple function on $(X, A)$. If $\phi = \sum_{j=1}^m \beta_j \chi_{B_j}$ is any representation of $\phi$, then

$$\int_X \phi \, d\mu = \sum_{j=1}^m \beta_j \mu(B_j) .$$

**Proof:** We must observe that the sum does not change as the four corrections of XIV.4.2.2. are made to convert the given representation of $\phi$ to the canonical one. Corrections (3) and (4) obviously do not affect the sum (using the convention $0 \cdot \infty = 0$ for (3)), and correction (2) is easily seen to also leave the sum unchanged if the $B_j$ are disjoint (using finite additivity of $\mu$).

Correction (1) requires a little more work. We disjointize $B_1$ and $B_2$ into $B_1 \cap B_2$, $B_1 \setminus B_2$, and $B_2 \setminus B_1$. Observe that

$$\beta_1 \chi_{B_1} + \beta_2 \chi_{B_2} = \beta_1 \chi_{B_1 \setminus B_2} + \beta_2 \chi_{B_2 \setminus B_1} + (\beta_1 + \beta_2) \chi_{B_1 \cap B_2}$$

and that

$$\beta_1 \mu(B_1 \setminus B_2) + \beta_2 \mu(B_2 \setminus B_1) + (\beta_1 + \beta_2) \mu(B_1 \cap B_2)$$

$$= \beta_1 [\mu(B_1 \setminus B_2) + \mu(B_1 \cap B_2)] + \beta_2 [\mu(B_2 \setminus B_1) + \mu(B_1 \cap B_2)] = \beta_1 \mu(B_1) + \beta_2 \mu(B_2)$$

by finite additivity of $\mu$; thus the sum for the beginning representation equals the sum obtained by disjointing $B_1$ and $B_2$. We can then successively disjointize any two non-disjoint remaining sets in the new representation in the same way without affecting the sum or the previous disjointizations. After a finite number of steps the representation is converted to semicanonical form and the sum is unchanged. \(\diamondsuit\)

Most of the important properties of the integral for simple functions follow easily. First is a formula for the integral of a simple function over a subset:
XIV.4.2.7. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi = \sum_{k=1}^{n} \alpha_k \chi_{A_k}\) a simple function on \((X, \mathcal{A})\) (not necessarily in canonical form). If \(E \in \mathcal{A}\), then

\[
\int_E \phi \, d\mu = \int_X \phi \chi_E \, d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k \cap E).
\]

Linearity and (finite) additivity also follow easily:

XIV.4.2.8. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi\) and \(\psi\) simple functions on \((X, \mathcal{A})\). If \(E \in \mathcal{A}\), then

\[
\int_E (\phi + \psi) \, d\mu = \int_E \phi \, d\mu + \int_E \psi \, d\mu.
\]

Proof: If \(\phi \chi_E = \sum_{k=1}^{n} \alpha_k \chi_{A_k}\) and \(\psi \chi_E = \sum_{j=1}^{m} \beta_j \chi_{B_j}\) are representations (not necessarily canonical), then

\[
(\phi + \psi) \chi_E = \sum_{k=1}^{n} \alpha_k \chi_{A_k} + \sum_{j=1}^{m} \beta_j \chi_{B_j}
\]

is a representation.

Note that the representation of \((\phi + \psi) \chi_E\) in the proof is almost necessarily non-canonical, so the proposition is necessary to establish linearity of the integral for simple functions.

XIV.4.2.9. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi\) a simple function on \((X, \mathcal{A})\). If \(D_1, \ldots, D_n \in \mathcal{A}\) are pairwise disjoint and \(E = D_1 \cup \cdots \cup D_n\), then

\[
\int_E \phi \, d\mu = \sum_{k=1}^{n} \left[ \int_{D_k} \phi \, d\mu \right].
\]

Proof: We have \(\phi \chi_E = \phi \chi_{D_1} + \cdots + \phi \chi_{D_n}\).

Finite additivity can be extended to countable additivity (XIV.4.2.12.).

We can also obtain monotonicity:
XIV.4.2.10. COROLLARY. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi\) and \(\psi\) simple functions on \((X, \mathcal{A})\). If \(E \in \mathcal{A}\) and \(\phi \leq \psi\) a.e. on \(E\), then
\[
\int_E \phi \, d\mu \leq \int_E \psi \, d\mu .
\]

PROOF: By modifying \(\phi\) on a null set (cf. XIV.4.2.5 (ii)), we may assume \(\phi \chi_E \leq \psi \chi_E\) everywhere. Let \(\phi \chi_E = \sum_{k=1}^{m} \alpha_k \chi_{A_k} \) and \(\psi \chi_E = \sum_{j=1}^{n} \beta_j \chi_{B_j} \) be canonical representations. Then
\[
\phi \chi_E = \sum_{j=1}^{m} \sum_{k=1}^{n} \alpha_k \chi_{A_k \cap B_j} \\
\psi \chi_E = \sum_{j=1}^{m} \sum_{k=1}^{n} \beta_j \chi_{A_k \cap B_j}
\]
are semicanonical representations; and whenever \(A_k \cap B_j \neq \emptyset\), we have \(\alpha_k \leq \beta_j\) since \(\phi \chi_E \leq \psi \chi_E\). Thus
\[
\int_E \phi \, d\mu = \sum_{j=1}^{m} \sum_{k=1}^{n} \alpha_k \mu(A_k \cap B_j) \leq \sum_{j=1}^{m} \sum_{k=1}^{n} \beta_j \mu(A_k \cap B_j) = \int_E \psi \, d\mu .
\]

We also have the following slight extension of the Baby Monotone Convergence Theorem (), which will be used to establish the full Monotone Convergence Theorem and other continuity properties of the integral:

XIV.4.2.11. PROPOSITION. [JUNIOR MONOTONE CONVERGENCE THEOREM] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi\) a simple function on \((X, \mathcal{A})\). If \(E_1, E_2, \ldots\) is an increasing sequence of sets in \(\mathcal{A}\), and \(E = \bigcup_{n=1}^{\infty} E_n\), then
\[
\int_E \phi \, d\mu = \lim_{n \to \infty} \left[ \int_{E_n} \phi \, d\mu \right] = \sup_n \left[ \int_{E_n} \phi \, d\mu \right] .
\]

PROOF: Let \(\phi = \sum_{k=1}^{m} \alpha_k \chi(A_k)\) be a representation (not necessarily canonical). Fix \(k\). Then we have \(A_k \cap E_1 \subseteq A_k \cap E_2 \subseteq \cdots\) and \(A_k \cap E = \bigcup_{n=1}^{\infty} [A_k \cap E_n]\), so by the Baby MCT we have
\[
\alpha_k \mu(A_k \cap E) = \lim_{n \to \infty} \alpha_k \mu(A_k \cap E_n) = \sup_n \alpha_k \mu(A_k \cap E_n) .
\]

This is true for all \(k\), and since the sum is finite, we have
\[
\sum_{k=1}^{m} \alpha_k \mu(A_k \cap E) = \lim_{n \to \infty} \left[ \sum_{k=1}^{m} \alpha_k \mu(A_k \cap E_n) \right] = \sup_n \left[ \sum_{k=1}^{m} \alpha_k \mu(A_k \cap E_n) \right] .
\]
XIV.4.2.12. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(\phi\) a simple function on \((X, \mathcal{A})\). If \(D_1, \ldots, D_n, \ldots \in \mathcal{A}\) are pairwise disjoint and \(E = \bigcup_{k=1}^{\infty} D_k\), then

\[
\int_E \phi \, d\mu = \sum_{k=1}^{\infty} \left[ \int_{D_k} \phi \, d\mu \right].
\]

Proof: For each \(n\) set \(E_n = D_1 \cup \cdots \cup D_n\). Then the \(E_n\) are increasing and \(E = \bigcup_{n} E_n\), so by XIV.4.2.11. and XIV.4.2.9. we have

\[
\int_E \phi \, d\mu = \lim_{n \to \infty} \left( \int_{E_n} \phi \, d\mu \right) = \lim_{n \to \infty} \left( \sum_{k=1}^{n} \left[ \int_{D_k} \phi \, d\mu \right] \right) = \sum_{k=1}^{\infty} \left[ \int_{D_k} \phi \, d\mu \right].
\]

XIV.4.3. The Integral of a Nonnegative Measurable Function

We can now define the integral of a general nonnegative measurable function:

XIV.4.3.1. Definition. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). Define

\[
\int_X f \, d\mu = \sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple, } \phi \leq f \right\}.
\]

If \(E \in \mathcal{A}\), define

\[
\int_E f \, d\mu = \int_X f \chi_E \, d\mu.
\]

Note that if \(f\) itself is a simple function, then this definition of \(\int_E f \, d\mu\) agrees with the previous definition (cf. XIV.4.2.10.); thus the definitions are consistent and unambiguous.

XIV.4.3.2. The following properties of the integral are immediate:

(i) \(0 \leq \int_E f \, d\mu \leq \infty\) for any \(f\) and \(E\).

(ii) If \(f \leq g\) on \(E\), then \(\int_E f \, d\mu \leq \int_E g \, d\mu\).

(iii) If \(\alpha \geq 0\), then \(\int_E \alpha f \, d\mu = \alpha \int_E f \, d\mu\) (using \(0 \cdot \infty = 0\) and \(\alpha \cdot \infty = \infty\) if \(\alpha > 0\)). In particular, \(\int_E 0 \, d\mu = 0\) for any \(E\).

(iv) If \(\mu(E) = 0\), then \(\int_E f \, d\mu = 0\) for any \(f\) (cf. XIV.4.2.5.).
XIV.4.3.3. **Proposition.** Let \((X,\mathcal{A},\mu)\) be a measure space, and \(f\) and \(g\) nonnegative extended real-valued measurable functions on \((X,\mathcal{A})\). If \(f = g\) a.e. (\(\mu\)-a.e.) on \(E\), then
\[
\int_E f \, d\mu = \int_E g \, d\mu .
\]

**Proof:** If \(f\chi_E = g\chi_E\) on \(X \setminus C\), where \(C \in \mathcal{A}\) and \(\mu(C) = 0\), and \(\phi\) is a simple function with \(\phi \leq f\), then \(\psi = \phi\chi_{X \setminus C}\) is a simple function with \(\psi \leq g\chi_E\), and \(\psi = \phi\) a.e., so \(\int_X \psi \, d\mu = \int_X \phi \, d\mu\) by XIV.4.2.5.. Thus \(\int_E f \, d\mu \leq \int_E g \, d\mu\). Symmetrically, \(\int_E g \, d\mu \leq \int_E f \, d\mu\).

Combining this with XIV.4.3.2 (ii), we obtain:

XIV.4.3.4. **Proposition.** Let \((X,\mathcal{A},\mu)\) be a measure space, and \(f\) and \(g\) nonnegative extended real-valued measurable functions on \((X,\mathcal{A})\). If \(f \leq g\) a.e. (\(\mu\)-a.e.) on \(E\), then
\[
\int_E f \, d\mu \leq \int_E g \, d\mu .
\]

XIV.4.3.5. We have thus established all the desired properties of the integral (for nonnegative functions) from XIV.4.1.1. except for the crucial properties of linearity, additivity, and continuity. These are trickier to show; we will first prove one of the important continuity properties, the Monotone Convergence Theorem, and use it to obtain linearity and additivity (and much more).

**Integrable Functions**

The integral of a nonnegative extended real-valued measurable function can be \(\infty\) (even the integral of a simple function can be infinite). But functions with finite integral are the most interesting ones:

XIV.4.3.6. **Definition.** Let \((X,\mathcal{A},\mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X,\mathcal{A})\). Then \(f\) is \(\mu\)-integrable if
\[
\int_X f \, d\mu < \infty .
\]

If \(E \in \mathcal{A}\), then \(f\) is \(\mu\)-integrable on \(E\) if \(\int_E f \, d\mu < \infty\).

XIV.4.3.7. Note that a \(\mu\)-integrable function is by definition measurable. We often write “integrable” in place of “\(\mu\)-integrable” when the \(\mu\) is understood. But note that, although measurability of a function \(f\) depends only on \(\mathcal{A}\) and not on \(\mu\), integrability of \(f\) depends on \(\mu\).

When \(\mu\) is Lebesgue measure \(\lambda\) on \(\mathbb{R}\) (or a subset), we will say \(f\) is Lebesgue integrable if it is \(\lambda\)-integrable. We usually, but not always, write “Lebesgue integrable” in this context to distinguish the notion from Riemann integrability (or other types of integrability on \(\mathbb{R}\) such as H-K integrability).
XIV.4.3.8. A nonnegative measurable function dominated by an integrable function is integrable. A nonnegative measurable function equal almost everywhere to an integrable function is integrable (XIV.4.3.3).

If \( \mu \) is finite, any bounded nonnegative measurable function on \((X, \mathcal{A})\) is integrable; but if \( \mu \) is infinite, even a (nonzero) constant function is not integrable.

An integrable function can be unbounded. It can even take the value \(1\) on a null set, e.g. \(f = \infty \cdot \chi_A\), where \(\mu(A) = 0\). But an integrable function must be finite a.e., hence equal a.e. to a real-valued measurable function:

XIV.4.3.9. **Proposition.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). Set \(A = \{x \in X : f(x) = \infty\}\). If \(f\) is integrable, then \(\mu(A) = 0\).

**Proof:** If \(\phi_n = n \chi_A\), then \(\phi_n\) is a simple function and \(\phi_n \leq f\). Thus
\[
n\mu(A) = \int_X \phi_n \, d\mu \leq \int_X f \, d\mu
\]
for all \(n\), which implies \(\mu(A) = 0\). \(\Box\)

Similarly, if \(f\) is integrable, the support \((\cdot)\) of \(f\) has “almost finite measure” (it need not actually have finite measure; cf. ()):  

XIV.4.3.10. **Proposition.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). Fix \(\epsilon > 0\), and set \(A_\epsilon = \{x \in X : f(x) \geq \epsilon\}\). Then
\[
\mu(A_\epsilon) \leq \frac{1}{\epsilon} \int_X f \, d\mu.
\]
In particular, if \(f\) is integrable, then \(\mu(A_\epsilon) < \infty\) for all \(\epsilon > 0\).

**Proof:** Set \(\phi_\epsilon = \epsilon \chi_{A_\epsilon}\). Then \(\phi_\epsilon\) is a simple function and \(\phi_\epsilon \leq f\). Thus
\[
\epsilon \mu(A_\epsilon) = \int_X \phi_\epsilon \, d\mu \leq \int_X f \, d\mu.
\]
\(\Box\)

XIV.4.3.11. **Corollary.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). If \(E \in \mathcal{A}\) and \(\int_E f \, d\mu = 0\), then \(f = 0\) a.e. on \(E\).

**Proof:** If \(A = \{x \in E : f(x) > 0\}\), then \(A = \cup_{n=1}^\infty A_n\), where \(A_n = \{x \in E : f(x) \geq \frac{1}{n}\}\), and
\[
\mu(A_n) \leq n \int_E f \, d\mu = 0.
\]
\(\Box\)
XIV.4.3.12. Proposition XIV.4.3.10. is not only interesting for $\epsilon$ small; it is of interest for arbitrary positive $\epsilon$. In fact, as $\epsilon \to \infty$ we obtain XIV.4.3.9. The proposition says that an integrable function is small except on a set of reasonable finite measure, and not too large except on a set of small measure.

XIV.4.3.13. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a nonnegative integrable function on $(X, \mathcal{A})$. Then the support of $f$ is a set of $\sigma$-finite measure.

Proof: If $E = \{x \in X : f(x) > 0\}$ is the support of $f$, then $E = \cup_{n=1}^{\infty} A_{1/n}$.

XIV.4.3.14. Thus, although integration is defined on arbitrary measure spaces, it is only a nontrivial theory if there are sufficiently many sets of finite measure (e.g. if the measure is semifinite). If there are no sets of nonzero finite measure, then the only nonnegative integrable functions are ones equal to 0 almost everywhere.

Integration for Counting Measure

By far the most important example of the integral we have constructed is the Lebesgue integral on $\mathbb{R}$, discussed in Section XIV.4.6. But another illuminating example is counting measure on $\mathbb{N}$.

XIV.4.3.15. Consider the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where $\mu$ is counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. A real-valued function on $\mathbb{N}$ is just a sequence, and every such function is measurable since the $\sigma$-algebra is $\mathcal{P}(\mathbb{N})$. An extended real-valued function is one where some terms of the sequence are allowed to be $\infty$. A simple function is a sequence of nonnegative real numbers where only finitely many distinct numbers occur; such a function is integrable if and only if only finitely many terms are nonzero, and the integral is just the sum of the terms.

Thus if $x = (x_1, x_2, \ldots)$ is a nonnegative sequence, then $\int_{\mathbb{N}} x \, d\mu$ is just the unordered sum

$$\sum_{n \in \mathbb{N}} x_n$$

which converges if and only if all terms are finite and the infinite series

$$\sum_{n=1}^{\infty} x_n$$

converges ($\cdot$).

We will see ($\cdot$) that a general function (sequence) $(x_n)$ is integrable with respect to $\mu$ if and only if the infinite series

$$\sum_{n=1}^{\infty} x_n$$

converges absolutely. Thus the theory of absolutely convergent infinite series is the special case of the integral for counting measure. In fact, it is worthwhile to view integration in general as some sort of “continuous” version of absolute summation.
XIV.4.4. The Monotone Convergence Theorem

The Monotone Convergence Theorem (MCT) is one of the three main convergence or continuity theorems of integration theory, the others being Fatou’s Lemma and the Dominated Convergence Theorem.

XIV.4.4.1. Theorem. [Monotone Convergence Theorem] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((f_n)\) a nondecreasing sequence of nonnegative extended real-valued measurable functions converging pointwise a.e. to a (nonnegative) extended real-valued measurable function \(f\). If \(E \in \mathcal{A}\), then

\[
\int_E f \, d\mu = \lim_{n \to \infty} \int_E f_n \, d\mu = \sup_n \int_E f_n \, d\mu .
\]

Note that it is the \(\text{sequence} (f_n)\) which is nondecreasing, i.e. \(f_1 \leq f_2 \leq \cdots\), not the individual functions which are nondecreasing. In fact, since the \(f_n\) are real-valued functions on a general set \(X\), it does not even make sense to ask whether such a function is nondecreasing. Note also that the hypothesis that \(f\) is measurable is almost, but not quite, automatic (if the measure space is not complete, \(f\) might have to be modified on a subset of a null set to make it measurable).

Proof: By modifying \(f\) on a null set, we may assume that \(f_n \to f\) pointwise everywhere. Since \(f_n \leq f\) for all \(n\), we have \(\int_E f_n \, d\mu \leq \int_E f \, d\mu\) for all \(n\). Thus if

\[
L := \lim_{n \to \infty} \int_E f_n \, d\mu = \sup_n \int_E f_n \, d\mu
\]

(the limit exists since the sequence of integrals is nondecreasing), we have \(L \leq \int_E f \, d\mu\).

So we need to show \(\int_E f \, d\mu \leq L\). If \(L = \infty\) there is nothing to show, so we may assume \(L < \infty\).

Fix \(r\), \(0 < r < 1\). Let \(\phi\) be a simple function with \(\phi \leq f\chi_E\), and set

\[
E_n = \{x \in E : r\phi(x) \leq f_n(x)\} .
\]

Then \(E_n\) is measurable \([E_n = E \cap \{x \in X : f_n(x) - r\phi(x) \geq 0\}\) and \(f_n - r\phi\) is a measurable function], and \(E_n \subseteq E_{n+1}\) for all \(n\) since \(f_n \leq f_{n+1}\). If \(x \in E\) and \(f(x) > 0\), then \(f_n(x) \to f(x) > r\phi(x)\), so there is an \(n\) such that \(f_k(x) > r\phi(x)\) for all \(k \geq n\); thus \(x \in E_n\). And if \(f(x) = 0\), then \(\phi(x) = f_n(x) = 0\) for all \(n\), so \(x \in E_n\) for all \(n\). Thus \(\bigcup_n E_n = E\). And, for each \(n\), \(r\phi \chi_{E_n} \leq f_n \chi_{E_n}\), so \(\int_{E_n} r\phi \, d\mu \leq \int_{E_n} f_n \, d\mu \leq \int_E f_n \, d\mu \leq L\) by monotonicity.

By the Junior Monotone Convergence Theorem (XIV.4.2.11.),

\[
\int_E r\phi \, d\mu = \sup_n \int_{E_n} r\phi \, d\mu \leq \sup_n \int_{E_n} f_n \, d\mu \leq L
\]

Thus

\[
\int_E \phi \, d\mu \leq \frac{1}{r} L .
\]

This is true for all simple \(\phi \leq f\chi_E\), so we have

\[
\int_E f \, d\mu = \sup_{\phi \leq f\chi_E} \int_E \phi \, d\mu \leq \frac{1}{r} L
\]

Since this is true for all \(r < 1\), we conclude \(\int_E f \, d\mu \leq L\).
XIV.4.4.2. In some references, the MCT or its series version XIV.4.5.5. is called the Beppo Levi Theorem. While there is some historical justification for this, and we do like Beppo’s name (perhaps the main reason mathematicians use this terminology?), we prefer the more descriptive MCT term.

XIV.4.5. Consequences of the Monotone Convergence Theorem

The Monotone Convergence Theorem is the key to several more important properties of the integral.

Linearity

XIV.4.5.1. Theorem. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ and $g$ nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$. If $E \in \mathcal{A}$, then

$$\int_E (f + g) \, d\mu = \int_E f \, d\mu + \int_E g \, d\mu.$$  

Proof: Let $(\phi_n)$ and $(\psi_n)$ be nondecreasing sequences of simple functions converging pointwise to $f \chi_E$ and $g \chi_E$ respectively. Then $(\phi_n + \psi_n)$ is an increasing sequence of simple functions converging pointwise to $(f + g) \chi_E$, and by XIV.4.2.8. we have

$$\int_E (\phi_n + \psi_n) \, d\mu = \int_E \phi_n \, d\mu + \int_E \psi_n \, d\mu$$

for all $n$. Thus by three applications of the MCT, along with (),

$$\int_E (f + g) \, d\mu = \lim_{n \to \infty} \int_E (\phi_n + \psi_n) \, d\mu = \lim_{n \to \infty} \left[ \int_E \phi_n \, d\mu + \int_E \psi_n \, d\mu \right]$$

$$= \left[ \lim_{n \to \infty} \int_E \phi_n \, d\mu \right] + \left[ \lim_{n \to \infty} \int_E \psi_n \, d\mu \right] = \int_E f \, d\mu + \int_E g \, d\mu.$$  

Additivity

We next establish finite additivity:

XIV.4.5.2. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. If $D_1, \ldots, D_n \in \mathcal{A}$ are pairwise disjoint and $E = D_1 \cup \cdots \cup D_n$, then

$$\int_E f \, d\mu = \sum_{k=1}^n \int_{D_k} f \, d\mu.$$  

Proof: We have $f \chi_E = f \chi_{D_1} + \cdots + f \chi_{D_n}$.

Before extending to countable additivity, we note the following special case of the MCT which extends the Junior MCT:
XIV.4.5.3. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). If \(E_1, E_2, \ldots\) is an increasing sequence of sets in \(\mathcal{A}\), and \(E = \bigcup_{n=1}^{\infty} E_n\), then

\[
\int_E f \, d\mu = \lim_{n \to \infty} \left[ \int_{E_n} f \, d\mu \right] = \sup_n \left[ \int_{E_n} f \, d\mu \right].
\]

Proof: The sequence \((f_{\chi_{E_n}})\) is nondecreasing and converges pointwise to \(f_{\chi_E}\).

We then obtain countable additivity:

XIV.4.5.4. Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(f\) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). If \(D_1, \ldots, D_n, \ldots \in \mathcal{A}\) are pairwise disjoint and \(E = \bigcup_{k=1}^{\infty} D_k\), then

\[
\int_E f \, d\mu = \sum_{k=1}^{\infty} \left[ \int_{D_k} f \, d\mu \right].
\]

Proof: For each \(n\) set \(E_n = D_1 \cup \cdots \cup D_n\). Then the \(E_n\) are increasing and \(E = \bigcup_{n} E_n\), so by XIV.4.5.3. and XIV.4.5.2. we have

\[
\int_E f \, d\mu = \lim_{n \to \infty} \left( \int_{E_n} f \, d\mu \right) = \lim_{n \to \infty} \left( \sum_{k=1}^{n} \left[ \int_{D_k} f \, d\mu \right] \right) = \sum_{k=1}^{\infty} \left[ \int_{D_k} f \, d\mu \right].
\]

The Monotone Convergence Theorem for Series

More generally, we have the following series version of the MCT:

XIV.4.5.5. Theorem. [MONOTONE CONVERGENCE THEOREM FOR SERIES] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((g_k)\) a sequence of nonnegative extended real-valued measurable functions on \((X, \mathcal{A})\). Set

\[
f = \sum_{k=1}^{\infty} g_k
\]

(the series converges pointwise to an extended real-valued function since the terms are nonnegative). If \(E \in \mathcal{A}\), then

\[
\int_E f \, d\mu = \sum_{k=1}^{\infty} \left[ \int_E g_k \, d\mu \right].
\]

Proof: For each \(n\) set

\[
f_n = \sum_{k=1}^{n} g_k.
\]
Then \( (f_n) \) is a nondecreasing sequence of nonnegative extended real-valued measurable functions converging pointwise to \( f \). By the MCT and XIV.4.5.1,

\[
\int_E f \, d\mu = \lim_{n \to \infty} \int_E f_n \, d\mu = \lim_{n \to \infty} \left[ \sum_{k=1}^{n} \int_E g_k \, d\mu \right] = \sum_{k=1}^{\infty} \left[ \int_E g_k \, d\mu \right].
\]

The Downward MCT

If \( (f_n) \) is a nonincreasing sequence of nonnegative real-valued measurable functions on a measure space \((X, \mathcal{A}, \mu)\), it is not necessarily true that \( \int_X f_n \, d\mu \to \int_X f \, d\mu \), even if the \( f_n \) are indicator functions of measurable sets (XIV.4.7.1). However, if the \( f_n \) are integrable, downward continuity holds:

**XIV.4.5.6. Theorem. [Downward Monotone Convergence Theorem]** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \( (f_n) \) a nonincreasing sequence of nonnegative extended real-valued measurable functions converging pointwise a.e. to a (nonnegative) extended real-valued measurable function \( f \). If \( E \in \mathcal{A} \), and \( \int_E f_1 \, d\mu < \infty \), then

\[
\int_E f \, d\mu = \lim_{n \to \infty} \int_E f_n \, d\mu = \inf_n \int_E f_n \, d\mu.
\]

**Proof:** If \( A = \{ x \in X : f_1(x) = \infty \} \), then \( \mu(A \cap E) = 0 \) by XIV.4.3.9. Thus we may replace \( f_n \) by \( f_n \chi_{A^c} \) for each \( n \) without affecting the integrals over \( E \), i.e. we may assume the \( f_n \) are real-valued.

For each \( n \), set \( g_n = f_1 - f_n \) (this is well defined since the functions are real-valued). Then \( (g_n) \) is a nondecreasing sequence of nonnegative real-valued measurable functions converging pointwise a.e. to \( f_1 - f \); thus by the MCT

\[
\int_E (f_1 - f) \, d\mu = \lim_{n \to \infty} \int_E g_n \, d\mu = \lim_{n \to \infty} \int_E (f_1 - f_n) \, d\mu.
\]

For each \( n \) we have

\[
\int_E f_1 \, d\mu = \int_E f_n \, d\mu + \int_E (f_1 - f_n) \, d\mu
\]

and, since the integrals are finite,

\[
\int_E (f_1 - f_n) \, d\mu = \int_E f_1 \, d\mu - \int_E f_n \, d\mu.
\]

Similarly,

\[
\int_E (f_1 - f) \, d\mu = \int_E f_1 \, d\mu - \int_E f \, d\mu.
\]

Thus we have

\[
\int_E f_1 \, d\mu - \int_E f \, d\mu = \lim_{n \to \infty} \left[ \int_E f_1 \, d\mu - \int_E f_n \, d\mu \right] = \int_E f_1 \, d\mu - \lim_{n \to \infty} \int_E f_n \, d\mu
\]

and the result follows.
XIV.4.5.7. The requirement that \( \int_E f_1 \, d\mu < \infty \) may be relaxed to \( \int_E f_n \, d\mu < \infty \) for some \( n \) by passing to a subsequence. The Downward MCT is also an immediate corollary of the Dominated Convergence Theorem.

Approximation From Above

We now discuss when the integral of a nonnegative real-valued function can be approximated from above by integrals of simple functions. Fix a measure space \((X, \mathcal{A}, \mu)\).

XIV.4.5.8. An unbounded nonnegative real-valued function cannot be dominated by a simple function, since simple functions are bounded. And if \( \mu \) is infinite, an integrable \( f \) whose support has infinite measure cannot be dominated by an integrable simple function, since the support of an integrable simple function has finite measure. Thus the only functions for which there is a hope of approximating the integral from above by integrals of simple functions are bounded (nonnegative) functions whose support has finite measure.

XIV.4.5.9. Theorem. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \( f \) a nonnegative extended real-valued measurable function on \((X, \mathcal{A})\). Let \( S \) be the support of \( f \). If \( E \in \mathcal{A} \) and \( \mu(E \cap S) < \infty \), and \( f \) is bounded on \( E \), then

\[
\int_E f \, d\mu = \inf \left\{ \int_E \psi \, d\mu : \psi \text{ simple}, f \chi_E \leq \psi \right\}.
\]

Proof: Suppose \( f \leq M \) on \( E \). Then \( f \chi_E \leq \theta := M \chi_{E \cap S} \), \( \theta \) is a simple function, and \( \int_E \theta \, d\mu < \infty \). We have

\[
\int_E (\theta - f) \, d\mu = \sup \left\{ \int_E \phi \, d\mu : \phi \text{ simple}, \phi \leq \theta - f \chi_E \right\}.
\]

If \( \phi \) is a simple function with \( \phi \leq \theta - f \chi_E \), then \( \psi = \theta - \phi \) is a simple function with \( f \chi_E \leq \psi \leq \theta \), and conversely if \( f \chi_E \leq \psi \leq \theta \); thus

\[
\int_E (\theta - f) \, d\mu = \sup \left\{ \int_E (\theta - \psi) \, d\mu : \phi \chi_E \leq \psi \right\}.
\]

If \( \psi \leq \theta \), we have

\[
\int_E f \, d\mu = \int_E \psi \, d\mu + \int_E (\theta - \psi) \, d\mu,
\]

and, since the integrals are finite, we have

\[
\int_E (\theta - \psi) \, d\mu = \int_E \theta \, d\mu - \int_E \psi \, d\mu.
\]

Similarly, we have

\[
\int_E f \, d\mu = \int_E \theta \, d\mu - \int_E (\theta - f) \, d\mu,
\]

\[
= \int_E \theta \, d\mu - \sup \left\{ \left[ \int_E \theta \, d\mu - \int_E \psi \, d\mu \right] : \phi \text{ simple}, f \chi_E \leq \psi \leq \theta \right\}.
\]
Here is a simpler version of XIV.4.5.9. which captures its essence:

**XIV.4.5.10.** Corollary. Let \((X, \mathcal{A}, \mu)\) be a finite measure space, and \(f\) a bounded nonnegative real-valued measurable function on \((X, \mathcal{A})\). Then

\[
\int_X f \, d\mu = \inf \left\{ \int_X \psi \, d\mu : \psi \text{ simple}, f \leq \psi \right\}.
\]

In particular,

\[
\sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple}, \phi \leq f \right\} = \inf \left\{ \int_X \psi \, d\mu : \psi \text{ simple}, f \leq \psi \right\}.
\]

**XIV.4.5.11.** With more work, this theorem can be proved directly without use of the MCT or linearity of the integral. It then easily implies linearity of the integral for bounded functions with finite support (\(\ast\)). From this one can obtain general additivity with some additional work (\(\ast\)). One can then readily deduce Fatou’s Lemma (\(\ast\)), and obtain the MCT as a corollary of Fatou’s Lemma (\(\ast\)). This is the approach taken in [Roy88], for example. The approach we have taken, which is used e.g. in [MW99], seems a little better pedagogically.

**The Role of Measurability**

One can only reasonably define the integral directly for an indicator function or simple function if it is measurable. But the definition

\[
\int_X f \, d\mu = \sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple}, \phi \leq f \right\}
\]

makes sense for any nonnegative extended real-valued function, measurable or not. Why not integrate arbitrary nonnegative functions this way? Since measurability is used in the MCT, and hence for linearity, it might be suspected that something goes wrong; in fact, the following converse of XIV.4.5.10. shows that additivity must fail if we do not restrict to “almost measurable” functions.
XIV.4.5.12. PROPOSITION. Let \((X, \mathcal{A}, \mu)\) be a finite measure space, and \(f\) a bounded nonnegative real-valued function on \((X, \mathcal{A})\). If
\[
\sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple, } \phi \leq f \right\} = \inf \left\{ \int_X \psi \, d\mu : \psi \text{ simple, } f \leq \psi \right\}
\]
then \(f\) is equal a.e. to an \(\mathcal{A}\)-measurable function.

PROOF: The supremum/infimum \(I\) is finite. Choose a sequence \((\phi_n)\) of (measurable) simple functions with \(\phi_n \leq f\) for all \(n\) and \(\int_X \phi_n \, d\mu \to I\). Replacing \(\phi_n\) by \(\max_{1 \leq k \leq n} \phi_k\), we may assume \((\phi_n)\) is nondecreasing. Similarly, let \((\psi_n)\) be a nonincreasing sequence of simple functions with \(f \leq \psi_n\) for all \(n\) and \(\int_X \psi_n \, d\mu \to I\). Then \((\phi_n)\) and \((\psi_n)\) converge pointwise to measurable functions \(g\) and \(h\) respectively, and \(g \leq f \leq h\). By monotonicity,
\[
\int_X g \, d\mu = \int_X h \, d\mu = I
\]
so \(\int_X (h - g) \, d\mu = 0\), and hence \(h - g = 0\) a.e. (XIV.4.3.11.). Thus \(f = g\) a.e. 

XIV.4.5.13. This result does not quite imply that \(f\) is measurable in general if the measure space is not complete (), only that \(f\) is equal a.e. to a measurable function. But if \((X, \mathcal{A}, \mu)\) is complete, e.g. if it is Lebesgue measure on \((X, \mathcal{L})\), then \(f\) must be measurable.

Restricting and Extending Measures

In this connection, we obtain a (hardly unexpected) relation between integrals with respect to restrictions of a measure, and in particular between a measure and its completion:

XIV.4.5.14. PROPOSITION. Let \((X, \mathcal{A}, \mu)\) be a measure space, \(\mathcal{B}\) a sub-\(\sigma\)-algebra of \(\mathcal{A}\), and \(\nu = \mu|_\mathcal{B}\). If \(f\) is a nonnegative extended real-valued \(\mathcal{B}\)-measurable function, and \(E \in \mathcal{B}\), then
\[
\int_E f \, d\nu = \int_E f \, d\mu .
\]

PROOF: This is a good example of bootstrapping. First suppose \(f\) is \(\chi_E\) for some \(E \in \mathcal{B}\). Then \(B\) and \(B \cap E\) are in \(\mathcal{A}\), so we have
\[
\int_E \chi_B \, d\nu = \nu(B \cap E) = \mu(B \cap E) = \int_E \chi_B \, d\mu .
\]

By linearity, if \(\phi\) is a \(\mathcal{B}\)-measurable simple function, then \(\int_E \phi \, d\nu = \int_E \phi \, d\mu\). Now, if \(f\) is a general nonnegative extended real-valued \(\mathcal{B}\)-measurable function, there is an increasing sequence \((\phi_n)\) of \(\mathcal{B}\)-measurable simple functions converging pointwise to \(f\). Then by two applications of the MCT,
\[
\int_E f \, d\nu = \sup_n \int_E \phi_n \, d\nu = \sup_n \int_E \phi_n \, d\mu = \int_E f \, d\mu .
\]

\(\Box\)
XIV.4.5.15. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, with completion $(X, \tilde{\mathcal{A}}, \tilde{\mu})$. If $f$ is a nonnegative extended real-valued $\mathcal{A}$-measurable function, and $E \in \mathcal{A}$, then

$$\int_E f \, d\mu = \int_E f \, d\tilde{\mu}.$$ 

XIV.4.5.16. Thus when a measure is extended to a larger $\sigma$-algebra, more functions become measurable and hence potentially integrable, but the integrals of functions which were already measurable do not change.
XIV.4.6. The Lebesgue Integral and Riemann Integral

We now examine the Lebesgue integral, which is by definition the integral on the measure space \((\mathbb{R}, \mathcal{L}, \lambda)\), and its relation with the Riemann integral. We only consider nonnegative functions for now; this is not a serious restriction for bounded functions on finite intervals since a constant function can be added to make them nonnegative. The Lebesgue integral for general functions is discussed in (). In this section, “measurable” will always mean “Lebesgue measurable.”

XIV.4.6.1. If \(E\) is a measurable subset of \(\mathbb{R}\) and \(f\) a nonnegative extended real-valued measurable function on \(\mathbb{R}\), the Lebesgue integral of \(f\) over \(E\) is
\[
\int_E f \, d\lambda = \sup \left\{ \int_E \phi \, d\lambda : \phi \text{ simple, } \phi \leq f \right\}
\]
where the integral of a simple function \(\phi = \sum_{k=1}^n \alpha_k \chi_{A_k}\) is
\[
\int_E \phi \, d\lambda = \sum_{k=1}^n \alpha_k \lambda(A_k \cap E)
\]
as in XIV.4.2.4.. If \(E = [a, b]\) and \(f\) is Riemann integrable on \(E\), write the Riemann integral of \(f\) over \(E\) as
\[
\int_a^b f(x) \, dx
\]
to distinguish it notationally from the Lebesgue integral.

XIV.4.6.2. If an extended real-valued measurable function is defined only on a measurable subset \(E\) of \(\mathbb{R}\), we will always canonically extend it to a function on \(\mathbb{R}\) by setting \(f(x) = 0\) for \(x \notin E\). The extended function is also measurable, and (if nonnegative) has the same integral over any (measurable) subset of \(E\). Note the following fact which is used constantly:

XIV.4.6.3. Proposition. Let \(I\) be an interval in \(\mathbb{R}\), and \(f\) a nonnegative extended real-valued real-valued measurable function on \(\mathbb{R}\) (or on a measurable subset of \(\mathbb{R}\), extended canonically to \(\mathbb{R}\)). Then
\[
\int_I f \, d\lambda = \int_I f \, d\lambda .
\]

Proof: We have \(f \chi_I = f \chi_I\) a.e. \(\Diamond\)

XIV.4.6.4. More generally, if \(A\) and \(B\) are measurable subsets of \(\mathbb{R}\) differing by a null set (i.e. \(\lambda(A \Delta B) = 0\)), then \(f \chi_A = f \chi_B\) a.e., so \(\int_A f \, d\lambda = \int_B f \, d\lambda\).

The following result shows that the Lebesgue integral agrees with the Riemann integral for Riemann integrable functions, and thus the Lebesgue integral extends and generalizes the Riemann integral:
**Theorem.** Let \([a, b]\) be a closed bounded interval in \(\mathbb{R}\), and \(f : [a, b] \to \mathbb{R}\) a function. Extend \(f\) to \(\mathbb{R}\) by setting \(f(x) = 0\) for \(x \notin [a, b]\). If \(f\) is Riemann integrable on \([a, b]\), then \(f\) is Lebesgue integrable on \(\mathbb{R}\) and on \([a, b]\), and

\[
\int_{\mathbb{R}} f \, d\lambda = \int_{[a, b]} f \, d\lambda = \int_{a}^{b} f(x) \, dx .
\]

**Proof:** It is almost transparent that the Riemann integral equals the Lebesgue interval for nonnegative step functions (cf. ()), which are nothing but piecewise-continuous simple functions. A general (necessarily bounded) nonnegative function \(f\) is Riemann integrable on \([a, b]\) if and only if there are an increasing sequence of (nonnegative) step functions \((\phi_n)\) (corresponding to lower Darboux sums) and a decreasing sequence of step functions \((\psi_n)\) such that \(\phi_n \leq f \leq \psi_n\) for all \(n\) and

\[
\sup_n \left[ \int_{a}^{b} \phi_n(x) \, dx \right] = \sup_n \left[ \int_{[a, b]} \phi_n \, d\lambda \right] = \inf_n \left[ \int_{[a, b]} \psi_n \, d\lambda \right] = \inf_n \left[ \int_{a}^{b} \phi_n(x) \, dx \right] .
\]

Thus by (), \(f\) is equal a.e. to a Lebesgue measurable function and, since Lebesgue measure is complete, \(f\) is Lebesgue measurable. By monotonicity, we have that

\[
\int_{[a, b]} \phi_n \, d\lambda \leq \int_{[a, b]} f \, d\lambda \leq \int_{[a, b]} \psi_n \, d\lambda
\]

\[
\int_{a}^{b} \phi_n(x) \, dx \leq \int_{a}^{b} f(x) \, dx \leq \int_{a}^{b} \psi_n(x) \, dx
\]

and since the suprema of the integrals on the left and the infima of the integrals on the right are all the same, it follows that

\[
\int_{[a, b]} f \, d\lambda = \int_{a}^{b} f(x) \, dx .
\]

**XIV.4.6.6.** It follows easily from the Characterization Theorem () that a Riemann integrable function is Lebesgue measurable. However, we did not need to use this theorem to obtain the result.

**Lebesgue Integrable Functions Which Are Not Riemann Integrable**

There are many additional functions on \(\mathbb{R}\) which are Lebesgue Integrable, as the following examples show.

**Examples.** (i) The indicator function \(\chi_{\mathbb{Q}}\) is not Riemann Integrable over any interval () because it is “too discontinuous.” But it is Lebesgue integrable over \(\mathbb{R}\), or any measurable subset \(E\) of \(\mathbb{R}\) (in particular, over any interval), and

\[
\int_{E} \chi_{\mathbb{Q}} \, d\lambda = 0
\]

1616
for any \( E \in \mathcal{L} \).

(ii) Set

\[
f(x) = \begin{cases} 
\frac{1}{\sqrt{x}} & \text{if } x > 0 \\
0 & \text{if } x \leq 0 
\end{cases}
\]

Then \( f \) is not Riemann integrable on \([0,1]\) even though it is piecewise-continuous, since it is unbounded. But \( f \) is measurable, and Riemann integrable on \([\frac{1}{n}, 1]\) for any \( n \), so

\[
\int_{[0,1]} f \, d\lambda = \int_{[0,1]} f \, d\lambda = \sup_n \int_{[\frac{1}{n}, 1]} f \, d\lambda 
\]

\[
= \sup_n \int_{\frac{1}{n}}^1 \frac{1}{\sqrt{x}} \, dx = \sup_n \left[ 2\sqrt{x} \right]_{\frac{1}{n}}^1 = \sup_n \left[ 2 - \frac{2}{\sqrt{n}} \right] = 2
\]

by XIV.4.5.3., so \( f \) is Lebesgue integrable on \([0,1]\). In fact, \( f \) is Lebesgue integrable on any bounded interval (more generally, on any bounded measurable set) by a similar argument. Note that \( f \) is improperly Riemann integrable on \([0,1]\) (XIV.4.6.8.).

(iii) Set

\[
f(x) = \begin{cases} 
\frac{x}{r} & \text{if } x > 0 \\
0 & \text{if } x \leq 0 
\end{cases}
\]

Then \( f \) is measurable, and Riemann integrable on \([\frac{1}{n}, 1]\) for any \( n \), so

\[
\int_{[0,1]} f \, d\lambda = \int_{[0,1]} f \, d\lambda = \sup_n \int_{[\frac{1}{n}, 1]} f \, d\lambda 
\]

\[
= \sup_n \int_{\frac{1}{n}}^1 \frac{1}{x} \, dx = \sup_n \left[ \log x \right]_{\frac{1}{n}}^1 = \sup_n \left[ -\log \frac{1}{n} \right] = \infty
\]

so \( f \) is not Lebesgue integrable over \([0,1]\). In fact, if

\[
f(x) = \begin{cases} 
x^{-r} & \text{if } x > 0 \\
0 & \text{if } x \leq 0 
\end{cases}
\]

then \( f \) is Lebesgue integrable on \([0,1]\) if and only if \( r < 1 \). Note that these are exactly the power functions which are improperly Riemann integrable on \([0,1]\).

(iii) Set

\[
f(x) = \begin{cases} 
\frac{1}{x} & \text{if } x > 0 \\
0 & \text{if } x \leq 0 
\end{cases}
\]

Then \( f \) is measurable, but not Lebesgue integrable on \([0,1]\). It is Riemann integrable on \([1,b]\) for any \( b > 1 \), so

\[
\int_{[1,\infty)} f \, d\lambda = \sup_n \int_{[1,n]} f \, d\lambda 
\]

\[
= \sup_n \int_1^n \frac{1}{x^2} \, dx = \sup_n \left[ -\frac{1}{x} \right]_1^n = \sup_n \left[ 1 - \frac{1}{n} \right] = 1
\]

1617
so \( f \) is Lebesgue integrable over \([1, \infty)\). In fact, if
\[
f(x) = \begin{cases} 
  x^{-r} & \text{if } x > 0 \\
  0 & \text{if } x \leq 0
\end{cases}
\]
then \( f \) is Lebesgue integrable on \([1, \infty)\) if and only if \( r > 1 \). Note that these are exactly the power functions which are improperly Riemann integrable on \([1, \infty)\).

**The Lebesgue Integral and the Improper Riemann Integral**

The fact that unbounded functions cannot be Riemann integrable is partially compensated by the various forms of improper Riemann integration.

**XIV.4.6.8.** If \( f \) is a nonnegative function defined on \((a, b]\), and Riemann integrable on \([a + \epsilon, b]\) for every \( \epsilon > 0 \), recall (XIV.2.11.1.) that \( f \) is *improperly Riemann integrable* from \( a \) to \( b \) if
\[
\lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right] = \sup_{\epsilon > 0} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right] < \infty
\]
and in this case, we define
\[
\int_{a}^{b} f(x) \, dx = \lim_{\epsilon \to 0^+} \left[ \int_{a+\epsilon}^{b} f(x) \, dx \right].
\]

**XIV.4.6.9.** PROPOSITION. If \( f \) is a nonnegative function defined on \((a, b]\), and Riemann integrable on \([a + \epsilon, b]\) for every \( \epsilon > 0 \), then \( f \) is improperly Riemann integrable from \( a \) to \( b \) if and only if \( f \) is Lebesgue integrable on \([a, b]\), and
\[
\int_{a}^{b} f(x) \, dx = \int_{[a, b]} f \, d\lambda.
\]

PROOF: We have that \( f \chi_{[a + \frac{1}{n}, b]} \) is measurable for all \( n \) and \( f \chi_{[a + \frac{1}{n}, b]} \to f \chi_{(a, b]} \), so \( f \chi_{(a, b]} \) is measurable. Thus
\[
\int_{[a, b]} f \, d\lambda = \int_{(a, b)} f \, d\lambda = \sup_{n} \left[ \int_{[a + \frac{1}{n}, b]} f \, d\lambda \right] = \sup_{n} \left[ \int_{a + \frac{1}{n}}^{b} f(x) \, dx \right] = \int_{a}^{b} f(x) \, dx
\]
by XIV.4.5.3., and in particular the suprema are either both finite or both infinite.

**XIV.4.6.10.** There are unbounded nonnegative functions which are continuous a.e. and Lebesgue integrable, but not improperly Riemann integrable in any reasonable sense. Let \((r_{k})\) be an enumeration of the left endpoints of the intervals removed in the construction of the Cantor set (), and
\[
g(x) = \sum_{k=1}^{\infty} 2^{-k} f(x - r_{k})
\]
where \( f \) is the function in XIV.4.6.7.(ii). Then \( g \) is Lebesgue integrable on \([0, 1]\) by XIV.4.5.5., and continuous at all points not in the Cantor set, but not improperly Riemann integrable on \([0, 1]\) in any of the usual senses.
XIV.4.7. Continuity and Fatou’s Lemma

To put the Monotone Convergence Theorem in perspective, we now note in more detail the failure of Continuity for integration (XIV.4.1.1.) in general.

XIV.4.7.1. Examples. (i) Let \( \mu \) be counting measure on \((\mathbb{N}, \mathcal{P}(\mathbb{N}))\). Let \( E_n = \{ k : k \geq n \} \). Then \( \cap_n E_n = \emptyset \), so \( \chi_{E_n} \to 0 \) pointwise. But \[
\infty = \int_{\mathbb{N}} \chi_{E_n} \, d\mu \not\to \int_{\mathbb{N}} 0 \, d\mu = 0 .
\]

Maybe the problem is just that the \( \chi_{E_n} \) are not integrable. But:

(ii) Consider \((\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)\) again. Let \( f_n = \frac{1}{n} \chi_{\{1, \ldots, n\}} \). Then \( f_n \to 0 \), even uniformly; however,

\[
1 = \int_{\mathbb{N}} f_n \, d\mu \not\to \int_{\mathbb{N}} 0 \, d\mu = 0 .
\]

(iii) Consider Lebesgue measure on \([0, 1]\). Let \( f_n \) be the continuous function with the following graph:

Figure XIV.22: A continuous function
If \( x \in (0, 1] \), then for sufficiently large \( N \) we have \( \frac{2}{N} \leq x \), so \( f_n(x) = 0 \) for all \( n \geq N \). And \( f_n(0) = 0 \) for all \( n \). So \( f_n \to 0 \) pointwise. However,

\[
1 = \int_{[0,1]} f_n \, d\lambda \rightarrow \int_{[0,1]} 0 \, d\lambda = 0.
\]

Maybe the problem is that the functions are not uniformly bounded. But it is not quite so simple (cf. also (ii)):

(iv) Suppose we take \( f_n \) as in (iii), but the height of the triangle is \( n^r \) for fixed \( r \). Then \( f_n \to 0 \) pointwise as before. If \( r > 0 \), the \( f_n \) are not uniformly bounded. But

\[
\int_{[0,1]} f_n \, d\lambda = n^{r-1}
\]

so \( \int_{[0,1]} f_n \, d\lambda \to \int_{[0,1]} 0 \, d\lambda \) if and only if \( r < 1 \).

**XIV.4.7.2.** One might note that in (iv), the difference between the case \( r < 1 \) and the case \( r \geq 1 \) is that in the case \( r < 1 \), there is an integrable function \( g \) such that \( f_n \leq g \) for all \( n \), namely \( g(x) = x^{-r} \) (with \( g(0) = 0 \)). (It may not be obvious that there is no such \( g \) if \( r \geq 1 \).) This is also the root of the difficulty in (ii). In fact, the Dominated Convergence Theorem (cf. XIV.4.7.10.) states that existence of such a \( g \) is sufficient for convergence of the integrals.

**Fatou’s Lemma**

**XIV.4.7.3.** Another observation which might be made in these examples is that the integral of the limit is always less than or equal to the limit of the integrals. Actually this is a no-brainer in these examples since the functions are nonnegative and the integral of the limit function is zero; but experiments with less trivial similar examples suggest this is a general principle.

**XIV.4.7.4.** Another motivation for why this should be so is that if \( (a_n) \) and \( (b_n) \) are (nonnegative) sequences of real numbers, then

\[
\lim \inf_{n \to \infty} a_n + \lim \inf_{n \to \infty} b_n \leq \lim \inf_{n \to \infty} (a_n + b_n)
\]

\( (\cdot) \), and equality does not hold in general, i.e. “the lim inf of a sum is greater than or equal to the sum of the lim inf’s.” Since integration is a continuous version of summation, we should have that “the integral of a lim inf is less than or equal to the lim inf of the integrals.”

The second of the three main convergence theorems of measure theory asserts this:

**XIV.4.7.5.** **Theorem. [Fatou’s Lemma]** Let \( (X, \mathcal{A}, \mu) \) be a measure space, and \( (f_n) \) a sequence of nonnegative extended real-valued measurable functions on \( (X, \mathcal{A}) \). Set

\[
f(x) = \lim \inf_{n \to \infty} f_n(x) = \lim_{n \to \infty} \left[ \inf_{k \geq n} f_k(x) \right] = \sup_n \left[ \inf_{k \geq n} f_k(x) \right]
\]
(note that \( f \) is a well defined nonnegative extended real-valued measurable function). If \( E \in \mathcal{A} \), then
\[
\int_E f \, d\mu \leq \liminf_{n \to \infty} \int_E f_n \, d\mu .
\]

The following special case is the one most commonly used:

**XIV.4.7.6. Corollary.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((f_n)\) a sequence of nonnegative extended real-valued measurable functions on \((X, \mathcal{A})\), converging pointwise a.e. to a measurable function \( f \). If \( E \in \mathcal{A} \), then
\[
\int_E f \, d\mu \leq \liminf_{n \to \infty} \int_E f_n \, d\mu .
\]

**XIV.4.7.7.** Fatou’s Lemma has the most general hypotheses of the three major convergence theorems, but also not surprisingly the weakest conclusion.

The proof of Fatou’s Lemma is easy given the Monotone Convergence Theorem:

**Proof:** For each \( n \) set
\[
g_n = \inf_{k \geq n} f_k .
\]

Then \((g_n)\) is a nondecreasing sequence of nonnegative extended real-valued measurable functions, and \(g_n \to f\) pointwise. We also have \(g_n \leq f_k\) for all \( k \geq n \), and thus
\[
\int_E g_n \, d\mu \leq \int_E f_k \, d\mu
\]
for \( k \geq n \) by monotonicity. Thus by the MCT we have
\[
\int_E f \, d\mu = \sup_n \left[ \int_E g_n \, d\mu \right] \leq \sup_n \left[ \inf_{k \geq n} \int_E f_k \, d\mu \right] = \liminf_{n \to \infty} \int_E f_n \, d\mu .
\]

**XIV.4.7.8.** Note that the MCT also follows almost immediately from Fatou’s Lemma: if the sequence \((f_n)\) is nondecreasing, we get \( \int_E f \, d\mu \geq \sup_n \int_E f_n \, d\mu \) by monotonicity and \( \int_E f \, d\mu \leq \liminf_n \int_E f_n \, d\mu \) by Fatou, so \( \int_E f \, d\mu = \lim_n \int_E f_n \, d\mu \).

Of course, we used the MCT to prove Fatou’s Lemma; but there are approaches where Fatou’s Lemma is proved first and the MCT obtained as a consequence (cf. ()).

**XIV.4.7.9.** The examples of XIV.4.7.1. show that the inequality in Fatou’s Lemma cannot be replaced by an equality in general.
The Dominated Convergence Theorem for Nonnegative Functions

We can then obtain almost immediately the nonnegative version of the Dominated Convergence Theorem:

**Theorem.** [Dominated Convergence Theorem, Nonnegative Version] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((f_n)\) a sequence of nonnegative extended real-valued measurable functions on \((X, \mathcal{A})\), converging pointwise a.e. to a measurable function \(f\). Suppose there is a nonnegative integrable function \(g\) with \(f_n \leq g\) a.e. for all \(n\). If \(E \in \mathcal{A}\), then \(f\) is integrable on \(E\) and

\[
\int_E f \, d\mu = \lim_{n \to \infty} \int_E f_n \, d\mu
\]

(in particular, the limit exists).

**Proof:** Note that \(f\) is automatically integrable since \(f \leq g\) a.e. We have that \(g f_n\) is nonnegative a.e. for all \(n\) and \(g f_n \to g f\) a.e. By two applications of Fatou’s Lemma, we have

\[
\int_E f \, d\mu \leq \liminf_{n \to \infty} \int_E f_n \, d\mu
\]

\[
\int_E g \, d\mu - \int_E f \, d\mu = \int_E (g - f) \, d\mu \leq \liminf_{n \to \infty} \int_E (g - f_n) \, d\mu
\]

\[
= \liminf_{n \to \infty} \left[ \int_E g \, d\mu - \int_E f_n \, d\mu \right] = \int_E g \, d\mu - \limsup_{n \to \infty} \int_E f_n \, d\mu
\]

(the integrals are finite so subtraction is valid). Rearranging the last inequality, we obtain

\[
\int_E f \, d\mu \geq \limsup_{n \to \infty} \int_E f_n \, d\mu
\]

\(\blacksquare\)

**Corollary.** [Bounded Convergence Theorem, Nonnegative Version] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((f_n)\) a uniformly bounded sequence of nonnegative real-valued measurable functions on \((X, \mathcal{A})\), converging pointwise a.e. to a measurable function \(f\). If \(E \in \mathcal{A}\) with \(\mu(E) < \infty\), then \(f\) is integrable on \(E\) and

\[
\int_E f \, d\mu = \lim_{n \to \infty} \int_E f_n \, d\mu
\]

(in particular, the limit exists).

**Proof:** If \(f_n(x) \leq M\) for all \(n\) and all \(x \in E\), then \(M \chi_E\) is a dominating function for \((f_n \chi_E)\).

\(\blacksquare\)

**The restriction that \(\mu(E) < \infty\) is necessary in general, even if the \(f_n\) and \(f\) are assumed to be integrable on \(E\) and the convergence is uniform (XIV.4.7.1(ii)).**
XIV.4.8. The Integral For General Functions

We can extend the integral to many measurable functions which take both positive and negative values, as long as we carefully avoid the $\infty - \infty$ problem. The simplest and most satisfactory way to proceed is:

XIV.4.8.1. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a real-valued function on $X$. Then $f$ is $\mu$-integrable if $f = g - h$ for nonnegative real-valued integrable functions $g$ and $h$. Define the integral of $f$ to be

$$\int_X f \, d\mu = \int_X g \, d\mu - \int_X h \, d\mu.$$ 

If $E \in \mathcal{A}$, then $f \chi_E = g \chi_E - h \chi_E$ is also $\mu$-integrable; define

$$\int_E f \, d\mu = \int_X f \chi_E \, d\mu = \int_E g \, d\mu - \int_E h \, d\mu.$$ 

XIV.4.8.2. Note that we are here only considering real-valued functions, since if $g$ and $h$ are nonnegative extended real-valued functions, $g - h$ might not be well defined. We consider extended real-valued functions in $()$.

A $\mu$-integrable function is a difference of two measurable functions, hence automatically measurable. We often write “integrable” for “$\mu$-integrable” if the $\mu$ is understood.

We have defined the integral of an integrable function in the only possible way to obtain a linear integral; however, it is not entirely obvious that the integral is independent of the choice of $g$ and $h$. But it is:

XIV.4.8.3. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a real-valued integrable function on $X$. Then the integral of $f$ is well defined.

Proof: Suppose $f = g_1 - h_1 = g_2 - h_2$ with $g_i, h_i$ nonnegative real-valued integrable functions. Then $g_1 + h_2 = g_2 + h_1$, so, for any $E \in \mathcal{A}$,

$$\int_E g_1 \, d\mu + \int_E h_2 \, d\mu = \int_E (g_1 + h_2) \, d\mu = \int_E (g_2 + h_1) \, d\mu = \int_E g_2 \, d\mu + \int_E h_1 \, d\mu$$

and since the integrals are finite we may rearrange the equation by subtraction to obtain

$$\int_E g_1 \, d\mu - \int_E h_1 \, d\mu = \int_E g_2 \, d\mu - \int_E h_2 \, d\mu.$$ 

XIV.4.8.4. Although the definition in XIV.4.8.1. is logically satisfactory in light of XIV.4.8.3., it is perhaps uncomfortably indefinite because the $g$ and $h$ are (usually) not unique. It is more satisfying to have a canonical way to write a general integrable function as a difference of two nonnegative integrable functions. This is possible:
**XIV.4.8.5. Definition.** Let $X$ be a set, and $f : X \to \mathbb{R}$ a function. The positive part of $f$ is $f_+ = \max(f, 0)$ and the negative part of $f$ is $f_- = -\min(f, 0) = |\min(f, 0)|$.

We have

$$f_+(x) = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ 0 & \text{if } f(x) < 0 \end{cases}$$

$$f_-(x) = \begin{cases} |f(x)| & \text{if } f(x) \leq 0 \\ 0 & \text{if } f(x) > 0 \end{cases}$$

so $f_+$ and $f_-$ are nonnegative functions with the properties $f_+ - f_- = f$, $f_+ + f_- = |f|$, $f_+ f_- = 0$.

Conversely, if $g$ and $h$ are nonnegative functions with $f = g - h$ and $gh = 0$, then necessarily $g = f_+$ and $h = f_-$.

If $(X, \mathcal{A})$ is a measurable space and $f$ is measurable, then $f_+$ and $f_-$ are also measurable. (And if $(X, \mathcal{T})$ is a topological space and $f$ is continuous, then $f_+$ and $f_-$ are also continuous.)

**XIV.4.8.6. Proposition.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a real-valued measurable function on $X$. Then $f$ is integrable if and only if $f_+$ and $f_-$ are integrable. If $f$ is integrable, then

$$\int_E f \, d\mu = \int_E f_+ \, d\mu - \int_E f_- \, d\mu$$

for any $E \in \mathcal{A}$.

**Proof:** If $f_+$ and $f_-$ are integrable, then $f$ is integrable and by definition

$$\int_E f \, d\mu = \int_E f_+ \, d\mu - \int_E f_- \, d\mu$$

for any $E \in \mathcal{A}$. Conversely, if $f$ is integrable, write $f = g - h$ for nonnegative integrable functions $g$ and $h$. If $x \in X$ and $f(x) \geq 0$, then $f_+(x) = f(x) = g(x) - h(x) \leq g(x)$; if $f(x) < 0$, then $0 = f_+(x) \leq g(x)$. Thus $f_+ \leq g$. Similarly, $f_- \leq h$. Thus $f_+$ and $f_-$ are integrable.

**XIV.4.8.7. Corollary.** Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ a real-valued function on $X$. If $f$ is integrable, then $|f|$ is integrable. If $f$ is measurable and $|f|$ is integrable, then $f$ is integrable.

**Proof:** If $f$ is integrable, then $f_+$ and $f_-$ are integrable, so $|f| = f_+ + f_-$ is integrable. Conversely, if $f$ is measurable and $|f|$ is integrable, then $f_+, f_- \leq |f|$ so $f_+$ and $f_-$ are integrable.

**XIV.4.8.8.** Note that it is not quite correct to say that $f$ is integrable if and only if $|f|$ is integrable: if $|f|$ is integrable, $f$ might not be measurable. For example, if $A$ is a subset of $[0, 1]$ which is not Lebesgue measurable, set

$$f(x) = \begin{cases} 1 & \text{if } x \in A \\ -1 & \text{if } x \in [0, 1] \setminus A \\ 0 & \text{if } x \notin [0, 1] \end{cases}$$

Then $|f| = \chi_{[0,1]}$ is Lebesgue integrable, but $f$ is not Lebesgue measurable.

It is correct to say that if $f$ is measurable, then $f$ is integrable if and only if $|f|$ is integrable. (Note that if $f$ is measurable, then $|f|$ is also measurable.)
XIV.4.8.9. Thus, except for the measurability problem, a function is integrable if and only if its absolute value is integrable. So the theory of abstract Lebesgue integration is an “absolute” integration theory, analogous to the theory of absolutely convergent infinite series. In fact:

XIV.4.8.10. Corollary. Let $\mu$ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. If $x = (x_1, x_2, \ldots)$ is a sequence in $\mathbb{R}$ (a real-valued function on $\mathbb{N}$), then $x$ is $\mu$-integrable if and only if the infinite series

$$\sum_{k=1}^{\infty} x_k$$

converges absolutely. We then have

$$\int_{\mathbb{N}} x \, d\mu = \sum_{k=1}^{\infty} x_k .$$

A Vanishing Criterion

There is a convenient and useful criterion for an integrable function to vanish almost everywhere:

XIV.4.8.11. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f$ an integrable extended real-valued function on $X$. If $\int_A f \, d\mu = 0$ for all $A \in \mathcal{A}$, then $f = 0$ a.e.

Proof: Let $A = \{x \in X : f(x) \geq 0\}$ and $B = \{x \in X : f(x) \leq 0\}$. Since $\int_A f \, d\mu = \int_B f \, d\mu = 0$, we have that $f|_A$ and $f|_B$ are zero a.e. by XIV.4.3.11.

Actually we only need to assume that $\int_A f \, d\mu$ for a set of $A$’s which generate $\mathcal{A}$ as a $\sigma$-algebra, since it then follows that $\int_X f \, d\mu = 0$, and under this assumption it is easily checked that $\{B \in \mathcal{A} : \int_B f \, d\mu = 0\}$ is a $\sigma$-algebra. The set of such $A$ even only needs to generate $\mathcal{A}$ up to null sets. The following special case is especially useful:

XIV.4.8.12. Corollary. Let $[a, b]$ be an interval in $\mathbb{R}$, and $f$ a Lebesgue integrable function on $[a, b]$. If

$$\int_a^x f \, d\lambda = 0$$

for all $x \in [a, b]$, then $f = 0$ a.e. on $[a, b]$. 

1625
XIV.5. The Dominated Convergence Theorem

Except for the Fundamental Theorem of Calculus (1) which is the basis for the actual calculation of elementary integrals, the Dominated Convergence Theorem is probably the most important theorem in integration theory in the sense that it is the result used most often in applications, although the Fubini-Tonelli Theorem is a close second.

XIV.5.1. Applications of the DCT

Here is a general result about interchanging the order of differentiation and integration. It is phrased in terms of Lebesgue measure, but can be adapted to general measure spaces with the same proof (Exercise (1)).

XIV.5.1.1. Theorem. Let $I$ and $J$ be intervals in $\mathbb{R}$, and let $F : I \times J \to \mathbb{R}$. Assume that

(a) For each $y \in J$, the function $f[y] : I \to \mathbb{R}$ defined by $f[y](x) = F(x, y)$ is Lebesgue measurable.

(b) There is a $y_0 \in J$ such that $f[y_0]$ is Lebesgue integrable on $I$.

(c) For almost all $x \in I$, $\frac{\partial F}{\partial y}(x, y)$ exists for all $y \in J$.

(d) There is a nonnegative Lebesgue integrable function $g$ on $I$ such that, for almost all $x \in I$,

$$\left| \frac{\partial F}{\partial y}(x, y) \right| \leq g(x)$$

for all $y \in J$.

Then we have

(i) $f[y]$ is Lebesgue integrable on $I$ for all $y \in J$.

(ii) For every $y \in J$, the function $\phi^y$ defined by $\phi^y(x) = \frac{\partial F}{\partial y}(x, y)$ (defined a.e.) is integrable on $I$.

(iii) For every $y_1 \in J$,

$$\frac{d}{dy} \left[ \int_I F(x, y) \, d\lambda(x) \right] (y_1) = \frac{d}{dy} \left[ \int_I f[y](x) \, d\lambda(x) \right] (y_1) \text{ exists and equals } \int_I \frac{\partial F}{\partial y}(x, y_1) \, d\lambda(x) .$$

Proof: Let $A$ be a subset of $I$ such that (c) and (d) hold for all $x \in A$ and $\lambda(I \setminus A) = 0$.

(i): Fix $y \in J$, and set $\psi(x) = F(x, y) - F(x, y_0).$ Thus $f[y] = f[y_0] + \psi,$ so to show that $f[y]$ is Lebesgue integrable it suffices to show that $\psi$ is Lebesgue integrable. Note that it is Lebesgue measurable by (a). For any $x \in A$, there is a $c$ (depending on $x$) such that

$$\psi(x) = \frac{\partial F}{\partial y}(x, c)(y - y_0)$$

by the Mean Value Theorem (1), and hence we have

$$|\psi(x)| \leq g(x)|y - y_0|$$

1626
for almost all \( x \in I \) by (d). Since \( y - y_0 \) is a constant, \( \psi \) is Lebesgue integrable.

(ii): Fix \( y_1 \in J \). If \((h_n)\) is any sequence converging to 0 such that \( y_1 + h_n \in J \) for all \( n \), set

\[ \phi_n(x) = \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n} \]

for each \( n \) and note that each \( \phi_n \) is Lebesgue measurable by (a). We have \( \phi_n \to \phi^{y_1} \) pointwise a.e., so \( \phi^{y_1} \) is Lebesgue measurable. Since \( |\phi^{y_1}(x)| \leq g(x) \) for almost all \( x \) by (d), \( \phi^{y_1} \) is Lebesgue integrable.

(iii): We use the sequential criterion. Fix \( y_1 \in J \), and let

\[ L = \int_I \frac{\partial F}{\partial y}(x, y_1) \, d\lambda(x) = \int_I \phi^{y_1}(x) \, d\lambda(x) \]

(note that this integral is defined by (ii)). To prove (iii), it suffices to show that whenever \((h_n)\) is a sequence of nonzero numbers such that \( y_1 + h_n \in J \) for all \( n \), we have

\[ \lim_{n \to \infty} \int_I \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n} \, d\lambda(x) = L \]

We have

\[ \int_I \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n} \, d\lambda(x) = \int_I \phi_n(x) \, d\lambda(x) \]

where \( \phi_n \) is as in the proof of (ii). For each \( n \) and \( x \in A \), by the MVT there is a \( c_n \) (depending on \( x \) as well as \( n \)) between \( y_1 \) and \( y_1 + h_n \) such that

\[ \phi_n(x) = \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n} = \frac{\partial F}{\partial y}(x, c_n) \]

and thus \( |\phi_n(x)| \leq g(x) \) for all \( n \) and almost all \( x \) by (d). Thus by the DCT,

\[ \lim_{n \to \infty} \int_I \phi_n(x) \, d\lambda(x) = \int_I \phi^{y_1}(x) \, d\lambda(x) = L \]

since \( \phi_n \to \phi^{y_1} \) pointwise a.e., and the result follows.

XIV.5.1.2. Hypothesis (a) is not automatic, as the following example shows. For (a) to hold, it suffices that \( F \) be a Borel measurable function from \( I \times J \) to \( \mathbb{R} \). It does not quite suffice for \( F \) to be Lebesgue measurable (), but in the presence of (c) Lebesgue measurability might be sufficient. Hypotheses (a) and (b) are obviously necessary to even make sense of the left side of the equation in (iii), and hypothesis (c) is necessary to make sense of the right side (technically what we need to define the right side is that for each \( y \), \( \frac{\partial F}{\partial y}(x, y) \) is defined for almost all \( x \)).

XIV.5.1.3. Example. Let \( I = J = [-1, 1] \). Let \( A \) be a nonmeasurable subset of \( I \), and set

\[ F(x, y) = \begin{cases} y & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \]

Then \( F \) satisfies (b), (c), and (d), but not (a).
XIV.6. Integration on Product Spaces

XIV.6.1. The Fubini-Tonelli Theorems

XIV.6.1.1. The Fubini-Tonelli Theorems concern calculation of integrals over a product space by iterated integrals, and/or interchange of order of integration in iterated integrals. There are two distinct but closely related results, one concerning integration of nonnegative measurable functions and the other concerning integration of integrable functions which are not necessarily nonnegative (in line with the development of Lebesgue integration itself, first done for nonnegative functions and then for general integrable functions).

It is conventional to call the nonnegative version “Tonelli’s Theorem” and the integrable version “Fubini’s Theorem.” But the actual history is more complicated, and these names are not quite accurate historically: both FUBINI and TONELLI (who worked roughly simultaneously but independently) were partially responsible for both results. We will nonetheless follow convention and use the names “Tonelli’s Theorem” and “Fubini’s Theorem” since it is convenient to have distinct names for the two results, with just the warning that the names are not well justified historically.

Fubini’s Theorem is one of the two most important theorems of integration theory, along with the Dominated Convergence Theorem, in the sense that these two theorems are by far the results used most often in applications.

XIV.6.1.2. Some versions of Fubini’s Theorem were previously “known” (if not proved); cf. XIV.3.4.34. But it is conventional to call all versions “Fubini’s Theorem” and we will follow the convention.

Preliminaries

XIV.6.1.3. Proposition. Let \((X, A, \mu)\) and \((Y, B, \nu)\) be measure spaces, \(C \subseteq A \otimes B\). Then \(C\) is of \(\sigma\)-finite measure if and only if there are \(A \in A, B \in B\) of \(\sigma\)-finite measure, and \(M \in A, N \in B\) of measure 0, such that \(C \subseteq (A \times B) \cup (M \times Y) \cup (X \times N)\).

Proof: It is obvious that if there are such \(A, B, M,\) and \(N\), then \(C\) is \(\sigma\)-finite (see Exercise ()). Conversely, let \(C = \bigcup_{n=1}^{\infty} C_n\), with \((\mu \times \nu)(C_n) < \infty\). For each \(n\) there is a sequence \(R_{n,k} = A_{n,k} \times B_{n,k}\) of rectangles with \(A_{n,k} \in A, B_{n,k} \in B, C_n \subseteq \bigcup_k R_{n,k}\), and \(\sum_k \mu(A_{n,k})\nu(B_{n,k}) < (\mu \times \nu)(C_n) + 1 < \infty\). For each \(k\), \(\mu(A_{n,k}) = \infty\) implies \(\nu(B_{n,k}) = 0\) and \(\nu(B_{n,k}) = \infty\) implies \(\mu(A_{n,k}) = 0\). Let \(M_n\) be the union of the \(A_{n,k}\) with \(\mu(A_{n,k}) = 0\), and \(N_n\) the union of the \(B_{n,k}\) with \(\nu(B_{n,k}) = 0\); renumbering, we may assume \(\nu(A_{n,k}) < \infty\) for all \(k\) and \(C_n \subseteq (M_n \times Y) \cup (X \times N_n) \cup \bigcup_k R_{n,k}\). Set \(A = \bigcup_{n,k} A_{n,k}, B = \bigcup_{n,k} B_{n,k}, M = \bigcup_n M_n, N = \bigcup_n N_n\). Then \(\mu(M) = \nu(N) = 0\), \(A\) and \(B\) have \(\sigma\)-finite measure, and \(C \subseteq (A \times B) \cup (M \times Y) \cup (X \times N)\).

XIV.6.1.4. Note: it is not true that a set of \(\sigma\)-finite measure in \(A \otimes B\) is contained in a product of sets of \(\sigma\)-finite measure in general. For example, if \(\nu\) is not \(\sigma\)-finite, and \(x \in X\) is not an atom, then \(\{x\} \times Y\) has \(\mu \times \nu\)-measure zero, hence is \(\sigma\)-finite, but is not contained in a rectangle \(A \times B\) with \(B\) of \(\sigma\)-finite measure.

We will use the following notation, which was partially introduced in ():
XIV.6.1.5. Definition. Let $X$ and $Y$ be sets.

(i) If $C \subseteq X \times Y$, for each $x \in X$ the vertical cross section of $C$ at $x$ is

$$C_x = \{y \in Y : (x, y) \in C\}$$

and, for each $y \in Y$, the horizontal cross section of $C$ at $y$ is

$$C^y = \{x \in X : (x, y) \in C\}.$$  

(ii) If $f : X \times Y \to \mathbb{R}$ (or to $\mathbb{C}$) is a function, for each $x \in X$ define $f_{[x]} : Y \to \mathbb{R}$ by

$$f_{[x]}(y) = f(x, y)$$

and for each $y \in Y$, define $f^{[y]} : X \to \mathbb{R}$ by

$$f^{[y]}(x) = f(x, y).$$

The functions $f_{[x]}$ and $f^{[y]}$ are called the marginal functions of $f$.

Cross sections and marginal functions are well-behaved under set-theoretic and algebraic operations:

XIV.6.1.6. Proposition. Let $X$ and $Y$ be sets.

(i) If $C \subseteq X \times Y$, then $[(X \times Y) \setminus C]_x = Y \setminus C_x$ for all $x \in X$ and $[(X \times Y) \setminus C]^y = X \setminus C^y$ for all $y \in Y$ (informally, $(C^c)_x = (C_x)^c$ and $(C^c)^y = (C^y)^c$ for all $x$ and $y$).

(ii) If $\{C_i : i \in I\}$ is a collection of subsets of $X \times Y$, then $(\bigcup_{i \in I} C_i)_x = \bigcup_{i \in I} (C_i)_x$ for all $x \in X$ and $(\bigcup_{i \in I} C_i)^y = \bigcup_{i \in I} (C_i)^y$ for all $y \in Y$.

(iii) If $f$ and $g$ are functions from $X \times Y$ to $\mathbb{R}$, and $\alpha \in \mathbb{R}$, then $(f + g)_{[x]} = f_{[x]} + g_{[x]}$ and $(\alpha f)_{[x]} = \alpha f_{[x]}$ for all $x \in X$ and $(f + g)^{[y]} = f^{[y]} + g^{[y]}$ and $(\alpha f)^{[y]} = \alpha f^{[y]}$ for all $y \in Y$.

(iv) If $C \subseteq X \times Y$, then $(\chi_C)_x = \chi_{C_x}$ and $(\chi_C)^y = \chi_{C^y}$ for all $x \in X$, $y \in Y$.

The straightforward proof is left to the reader.

XIV.6.1.7. Now let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces. Form the product $\sigma$-algebra $\mathcal{A} \otimes \mathcal{B}$ and the product measure $\mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$. We need several technical results guaranteeing that various sets and functions arising in iterated integrals are measurable. These results might be expected, but it turns out that some $\sigma$-finite hypotheses are necessary.

The first results are that cross sections of measurable sets are measurable, and margins of measurable functions are measurable. No measures are needed for these results, and in particular no $\sigma$-finite hypotheses are needed (or even make sense!)
XIV.6.1.8. Lemma. Let \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) be measurable spaces, and \(C \in \mathcal{A} \otimes \mathcal{B}\). Then \(C_x \in \mathcal{B}\) for all \(x \in X\) and \(C^y \in \mathcal{A}\) for all \(y \in Y\).

Proof: Let \(\mathcal{C}\) be the collection of all \(C \in \mathcal{A} \otimes \mathcal{B}\) for which the conclusion holds. If \(C = A \times B \in \mathcal{A} \times \mathcal{B}\), then
\[
C_x = \begin{cases} B & \text{if } x \in A \\ \emptyset & \text{if } x \notin A \end{cases}
\]
and thus \(C_x \in \mathcal{B}\) for any \(x\), and similarly \(C^y \in \mathcal{A}\) for any \(y\). Thus \(C \in \mathcal{C}\), i.e. \(\mathcal{A} \times \mathcal{B} \subseteq \mathcal{C}\). In particular, \(\emptyset\) and \(X \times Y\) are in \(\mathcal{C}\). If \(C \in \mathcal{C}\), then \((C^n)_x = (C_x)^c \in \mathcal{B}\) for any \(x \in X\) since \(\mathcal{B}\) is closed under complements, and similarly \((C^n)^y \in \mathcal{A}\) for all \(y\), so \(C^n \in \mathcal{C}\), i.e. \(\mathcal{C}\) is closed under complements.

If \((C_n)\) is a sequence in \(\mathcal{C}\), then \((\cup_n C_n)_x = \cup_n C_x \in \mathcal{B}\) for all \(x\) since \(\mathcal{B}\) is closed under countable unions, and similarly \((\cup_n C_n)^y \in \mathcal{A}\) for all \(y\), so \(\cup_n C_n \in \mathcal{C}\) and \(\mathcal{C}\) is a \(\sigma\)-algebra containing \(\mathcal{A} \times \mathcal{B}\), and thus \(\mathcal{C}\) contains (and therefore equals) \(\mathcal{A} \otimes \mathcal{B}\). \(\Box\)

XIV.6.1.9. Lemma. Let \((X, \mathcal{A})\) and \((Y, \mathcal{B})\) be measurable spaces, and \(f : X \times Y \to \mathbb{R}\). If \(f\) is \((\mathcal{A} \otimes \mathcal{B})\)-measurable, then \(f[x]\) is \(\mathcal{B}\)-measurable for all \(x \in X\) and \(f[y]\) is \(\mathcal{A}\)-measurable for all \(y \in Y\).

Proof: We bootstrap. If \(f = \chi_C\) for \(C \in \mathcal{A} \otimes \mathcal{B}\), then \(f\) is measurable by XIV.6.1.6.(iv) and XIV.6.1.8.. If \(f\) is a simple function, then \(f\) is measurable by XIV.6.1.6.(iii). And if \(f\) is a general measurable function, there is a sequence \((s_n)\) of simple functions converging pointwise to \(f\) \((\ast)\). For each \(x\) the sequence \(((s_n)[x])\) converges pointwise to \(f[x]\), and thus \(f[x]\) is measurable. Similarly, \(f[y]\) is measurable. \(\Box\)

We now need measures and a \(\sigma\)-finite hypothesis. The next result is the most delicate part of the proof of the Fubini-Tonelli Theorems.

XIV.6.1.10. Lemma. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be \(\sigma\)-finite measure spaces, and \(C \in \mathcal{A} \otimes \mathcal{B}\). For \(x \in X\) set \(g(x) = \nu(C_x)\) and for \(y \in Y\) set \(h(y) = \mu(C^y)\). Then \(g\) is \(\mathcal{A}\)-measurable and \(h\) is \(\mathcal{B}\)-measurable.

Proof: Let \(\mathcal{D}\) be the set of all \(C\) in \(\mathcal{A} \otimes \mathcal{B}\) for which the functions \(g\) and \(h\) are measurable. We will show that \(\mathcal{D}\) equals \(\mathcal{A} \otimes \mathcal{B}\) by the Monotone Class Theorem \((\ast)\).

First note that if \(C = A \times B \in \mathcal{A} \times \mathcal{B}\), then \(g = \nu(B)\chi_A\) and \(h = \mu(A)\chi_B\), so \(g\) and \(h\) are measurable and \(C \in \mathcal{D}\), i.e. \(\mathcal{A} \times \mathcal{B} \subseteq \mathcal{D}\). Now note that \(\mathcal{E}\) is the set of finite disjoint unions of sets in \(\mathcal{A} \times \mathcal{B}\). If \(C\) is the disjoint union of \(A_k \times B_k\), \(1 \leq k \leq n\), then
\[
g = \sum_{k=1}^{n} \nu(B_k)\chi_{A_k}, \quad h = \sum_{k=1}^{n} \mu(A_k)\chi_{B_k}
\]
so \(g\) and \(h\) are again measurable and \(C \in \mathcal{D}\), i.e. \(\mathcal{E} \subseteq \mathcal{D}\).

Now suppose \((C_n)\) is an increasing sequence in \(\mathcal{D}\) with union \(C\). Let \(g_n, h_n\) be the functions corresponding to \(C_n\), and \(g\) and \(h\) the functions corresponding to \(C\). For each \(x \in X\), we have that \(((C_n)_x)\) is an increasing sequence in \(\mathcal{B}\) with union \(C_x\), and thus \(g_n(x) = \nu((C_n)_x)\) increases to \(g(x) = \nu(C_x)\) by the Baby Monotone Convergence Theorem \((\ast)\), i.e. \(g_n \to g\) pointwise. Since \(C_n \in \mathcal{D}\) for all \(n\), the \(g_n\) are measurable, hence \(g\) is
also measurable. By an almost identical argument, \( h \) is also measurable, so \( C \in \mathcal{D} \), i.e. \( \mathcal{D} \) is closed under increasing sequential unions.

Next we handle decreasing intersections. This is the trickiest part, and we must use the hypothesis that \( \mu \) and \( \nu \) are \( \sigma \)-finite. Let \( R_m = A_m \times B_m \) be an increasing sequence of rectangles with \( A_m \in \mathcal{A} \), \( \bigcup_{m=1}^{\infty} A_m = X \), \( B_m \in \mathcal{B} \), \( \bigcup_{m=1}^{\infty} B_m = Y \), \( \mu(A_m) < \infty \), \( \nu(B_m) < \infty \) for all \( m \).

Set

\[
\mathcal{F} = \{ C \in \mathcal{A} \otimes \mathcal{B} : C \cap (A_m \times B_m) \in \mathcal{D} \text{ for all } m \}.
\]

By arguments identical to the ones above (or by just applying them to the product space obtained by replacing \( X \) and \( Y \) by \( A_m \) and \( B_m \)), we have that \( \mathcal{E} \subseteq \mathcal{F} \) and \( \mathcal{F} \) is closed under increasing sequential unions. To show that \( \mathcal{F} \) is closed under decreasing intersections, let \( (F_n) \) be a decreasing sequence in \( \mathcal{F} \) with intersection \( F \). Fix \( m \), and let \( g_n \) and \( h_n \) be the functions corresponding to \( Z_n := F_n \cap (A_m \times B_m) \), and \( g \) and \( h \) the functions corresponding to \( Z := F \cap (A_m \times B_m) \). Then \( (Z_n) \) is a decreasing sequence in \( \mathcal{D} \) with intersection \( Z \). If \( x \in X \), then \( ((Z_n)_x) \) is a decreasing sequence in \( Y \) with intersection \( Z_x \), and since each \( (Z_n)_x \) is contained in \( B_m \) and thus has finite \( \nu \)-measure, we have \( g_n(x) = \nu((Z_n)_x) \rightarrow \nu(Z_x) = g(x) \) by downward continuity \( (\cdot) \), i.e. \( g_n \rightarrow g \) pointwise. Since each \( Z_n \) is in \( \mathcal{D} \), \( g_n \) is measurable, so \( g \) is measurable. Similarly, \( h \) is measurable. So \( Z \in \mathcal{D} \). This is true for every \( m \), so \( F \in \mathcal{F} \). Thus \( \mathcal{F} \) is a monotone class containing \( \mathcal{E} \), so it contains (and equals) \( \sigma(\mathcal{E}) = \mathcal{A} \otimes \mathcal{B} \) by the Monotone Class Theorem.

Finally, let \( C \in \mathcal{A} \otimes \mathcal{B} \). Set \( C_m := C \cap (A_m \times B_m) \). Then \( (C_m) \) is an increasing sequence with union \( C \). We have that \( C_m \in \mathcal{D} \) for all \( m \) since \( C \in \mathcal{F} \). Since \( \mathcal{D} \) is closed under increasing sequential unions, \( C \in \mathcal{D} \). Thus \( \mathcal{D} = \mathcal{A} \otimes \mathcal{B} \).

**XIV.6.1.11. ****L**e**m**a**. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be \( \sigma \)-finite measure spaces, and \( f : X \times Y \rightarrow \mathbb{R} \) a nonnegative extended real-valued \((\mathcal{A} \otimes \mathcal{B})\)-measurable function. Set

\[
g(x) = \int_Y f_{[x]} \, d\nu = \int_Y f(x, y) \, d\nu(y) \]

\[
h(y) = \int_X f^{[y]} \, d\mu = \int_X f(x, y) \, d\mu(x) .
\]

Then \( g \) is \( \mathcal{A} \)-measurable and \( h \) is \( \mathcal{B} \)-measurable.

**Proof:** Bootstrap. If \( f = \chi_C \) for \( C \in \mathcal{A} \otimes \mathcal{B} \), then \( f_{[x]} = \chi_{C_x} \) and \( f^{[y]} = \chi_{C^y} \) for \( x \in X \), \( y \in Y \), so

\[
g(x) = \int_Y f_{[x]} \, d\nu = \int_Y \chi_{C_x} \, d\nu = \nu(C_x)
\]

\[
h(y) = \int_X f^{[y]} \, d\mu = \int_X \chi_{C^y} \, d\mu = \mu(C^y)
\]

so \( g \) and \( h \) are measurable by Lemma XIV.6.1.10. Note also that if \( f_1, \ldots, f_n \) are nonnegative extended real-valued measurable functions with corresponding \( g_1, \ldots, g_n \) and \( h_1, \ldots, h_n \), then \( f = f_1 + \cdots + f_n \) has corresponding functions \( g = g_1 + \cdots + g_n \) and \( h = h_1 + \cdots + h_n \), so if all \( g_k \) and \( h_k \) are measurable, so are \( g \) and \( h \). Similarly, if \( f \) has measurable \( g \) and \( h \), and \( \alpha > 0 \), then \( \alpha f \) has measurable \( \alpha g \) and \( \alpha h \). In particular, if \( f \) is a simple function, then the corresponding \( g \) and \( h \) are measurable.

1631
For a general $f$ with corresponding functions $g$ and $h$, let $(s_n)$ be an increasing sequence of simple functions converging pointwise to $f$, with corresponding functions $g_n$ and $h_n$. Then, for each $x \in X$, $((s_n)_x)$ is an increasing sequence of $\mathcal{B}$-measurable functions converging pointwise to $f_x$, so

$$g_n(x) = \int_Y (s_n)_x \; d\nu \to \int_Y f_x \; d\nu = g(x)$$

by the Monotone Convergence Theorem, i.e. $g_n \to g$ pointwise. Each $g_n$ is measurable, so $g$ is measurable. Similarly, $h$ is measurable.

The Junior Tonelli Theorem

The next result is called the Junior Tonelli Theorem since it is Tonelli's theorem for characteristic or indicator functions. It is a measure-theoretic version of Cavalieri's Principle: to find the volume of a solid, integrate the cross-sectional area over the length of the solid (cf. []).

**XIV.6.1.12. Theorem.** [Junior Tonelli Theorem] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be $\sigma$-finite measure spaces, and $C \in \mathcal{A} \otimes \mathcal{B}$. Then $C_x \in \mathcal{B}$ for each $x \in X$, $C^y \in \mathcal{A}$ for all $y \in A$, $g(x) = \nu(C_x)$ is $\mu$-measurable, $h(y) = \mu(C^y)$ is $\nu$-measurable, and

$$(\mu \times \nu)(C) = \int_X \nu(C_x) \; d\mu(x) = \int_Y \mu(C^y) \; d\nu(y).$$

**Proof:** If $C \in \mathcal{A} \otimes \mathcal{B}$, the first four assertions hold by Lemmas XIV.6.1.8. and XIV.6.1.10., so the integrals make sense. For $C \in \mathcal{A} \otimes \mathcal{B}$, define

$$\theta(C) = \int_X \nu(C_x) \; d\mu = \int_X g \; d\mu.$$

We show $\theta$ is a measure. Obviously $\theta(\emptyset) = 0$. If $(C_n)$ is a sequence of pairwise disjoint sets in $\mathcal{A} \otimes \mathcal{B}$ with union $C$, then for each $x \in X$, $((C_n)_x)$ is a sequence of pairwise disjoint sets in $\mathcal{B}$, so we have

$$\theta(C) = \theta \left( \bigcup_{n=1}^{\infty} C_n \right) = \int_X \nu \left( \left[ \bigcup_{n=1}^{\infty} C_n \right]_x \right) \; d\mu = \int_X \nu \left( \bigcup_{n=1}^{\infty} (C_n)_x \right) \; d\mu = \int_X \left[ \sum_{n=1}^{\infty} \nu((C_n)_x) \right] \; d\mu$$

$$= \sum_{n=1}^{\infty} \int_X \nu((C_n)_x) \; d\mu = \sum_{n=1}^{\infty} \theta(C_n)$$

where the interchange of integral and sum is justified by the Monotone Convergence Theorem. So $\theta$ is a measure. If $A \times B \in \mathcal{A} \times \mathcal{B}$, then

$$\theta(A \times B) = \int_X \nu((A \times B)_x) \; d\mu = \int_X \nu(B)_A \; d\mu = \mu(A) \nu(B) = (\mu \times \nu)(A \times B).$$

Thus by the uniqueness part of the Extension Theorem in the $\sigma$-finite case, $\theta = \mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$, and the first equality is proved. The proof that $(\mu \times \nu)(C) = \int_Y \mu(C^y) \; d\nu(y)$ for $C \in \mathcal{A} \otimes \mathcal{B}$ is analogous. □
Tonelli’s Theorem

Tonelli’s Theorem concerns calculating the integral of a nonnegative measurable function over a product space by iterated integrals, and/or interchanging the order of integration in iterated integrals of nonnegative measurable functions.

XIV.6.1.13. Theorem. [Tonelli’s Theorem] Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be \(\sigma\)-finite measure spaces, and \(f\) a nonnegative extended real-valued \((\mathcal{A} \otimes \mathcal{B})\)-measurable function on \(X \times Y\), with marginal functions \(f_x\) and \(f_y\). Then

(i) \(f_x\) is \(\mathcal{B}\)-measurable for all \(x \in X\).

(i') \(f_y\) is \(\mathcal{A}\)-measurable for all \(y \in Y\).

(ii) \(g\) defined by \(g(x) = \int_Y f_x \, d\nu\) is \(\mathcal{A}\)-measurable.

(ii') \(h\) defined by \(h(y) = \int_X f_y \, d\mu\) is \(\mathcal{B}\)-measurable.

(iii) We have

\[
\int_{X \times Y} f \, d(\mu \times \nu) = \int_X g \, d\mu = \int_X \left[ \int_Y f_x \, d\nu \right] \, d\mu = \int_X \left[ \int_Y f(x, y) \, d\nu(y) \right] \, d\mu(x)
\]

\[
= \int_Y h \, d\nu = \int_Y \left[ \int_X f_y \, d\mu \right] \, d\nu = \int_Y \left[ \int_X f(x, y) \, d\mu(x) \right] \, d\nu(y).
\]

In particular, the iterated integrals

\[
\int_X \left[ \int_Y f(x, y) \, d\nu(y) \right] \, d\mu(x) = \int_Y \left[ \int_X f(x, y) \, d\mu(x) \right] \, d\nu(y)
\]

exist and are equal.

Proof: We have (i) and (i') by Lemma XIV.6.1.9., and (ii) and (ii') by Lemma XIV.6.1.11.. For (iii), we bootstrap. If \(f = \chi_C, C \in \mathcal{A} \otimes \mathcal{B}\), the equalities hold by the Junior Tonelli Theorem XIV.6.1.12.. It is also clear from the (nonnegative) linearity of integrals that if the equalities hold for \(f_1, \ldots, f_n\), and if \(\alpha_1, \ldots, \alpha_n \geq 0\), then the equalities hold for \(\alpha_1 f_1 + \cdots + \alpha_n f_n\). In particular, they hold if \(f\) is a simple function.

Now let \(f\) be a general nonnegative \((\mathcal{A} \otimes \mathcal{B})\)-measurable function. Let \((s_n)\) be a sequence of simple functions increasing pointwise to \(f\). Then \((s_n)[x]\) increases pointwise to \(f[x]\) for all \(x \in X\), so by the Monotone Convergence Theorem we have

\[
g_n(x) := \int_Y (s_n)[x] \, d\nu \to \int_Y f[x] \, d\nu = g(x)
\]

for all \(x \in X\), i.e. \(g_n\) increases pointwise to \(g\). Thus, by two more applications of the MCT,

\[
\int_{X \times Y} f \, d(\mu \times \nu) = \lim_{n \to \infty} \int_{X \times Y} s_n \, d(\mu \times \nu) = \lim_{n \to \infty} \int_Y g_n \, d\mu = \int_Y g \, d\mu
\]

and the first equality is proved for \(f\). The proof of the second equality is analogous. \(\Box\)
Fubini’s Theorem

Fubini’s Theorem is analogous to Tonelli’s Theorem, but instead of nonnegative measurable functions it deals with integrable functions which are not necessarily nonnegative.

XIV.6.1.14. Theorem. [Fubini’s Theorem] Let \((X, A, \mu)\) and \((Y, B, \nu)\) be \(\sigma\)-finite measure spaces, and \(f\) a complex-valued \((A \otimes B)\)-measurable function with marginal functions \(f_{[x]}\) and \(f^{[y]}\) for \(x \in X, y \in Y\). Assume at least one of the following integrals exists and is finite:

\[
\int_{X \times Y} |f| \, d(\mu \times \nu), \quad \int_Y \left[ \int_X |f(x, y)| \, d\mu(x) \right] \, d\nu(y), \quad \int_X \left[ \int_Y |f(x, y)| \, d\nu(y) \right] \, d\mu(x).
\]

Then all three integrals exist and are finite and equal, hence \(f\) is \((\mu \times \nu)\)-integrable, and

(i) \(f_{[x]}\) is \(\nu\)-integrable for \(\mu\)-almost all \(x \in X\).

(ii) \(f^{[y]}\) is \(\mu\)-integrable for \(\nu\)-almost all \(y \in Y\).

(iii) \(g\) defined by \(g(x) = \int_Y f_{[x]} \, d\nu\) (which is defined \(\mu\)-a.e.) is equal \(\mu\)-a.e. to a \(\mu\)-integrable function.

(ii) \(h\) defined by \(h(y) = \int_X f^{[y]} \, d\mu\) (which is defined \(\nu\)-a.e.) is equal \(\nu\)-a.e. to a \(\nu\)-integrable function.

(iii) We have

\[
\int_{X \times Y} f \, d(\mu \times \nu) = \int_X g \, d\mu = \int_X \left[ \int_Y f_{[x]} \, d\nu \right] \, d\mu = \int_X \left[ \int_Y f(x, y) \, d\nu(y) \right] \, d\mu(x)
\]

= \int_Y h \, d\nu = \int_Y \left[ \int_X f^{[y]} \, d\mu \right] \, d\nu = \int_Y \left[ \int_X f(x, y) \, d\mu(x) \right] \, d\nu(y).

In particular, the iterated integrals

\[
\int_X \left[ \int_Y f(x, y) \, d\nu(y) \right] \, d\mu(x) = \int_Y \left[ \int_X f(x, y) \, d\mu(x) \right] \, d\nu(y)
\]

exist and are equal.

Proof: We prove the theorem for \(f\) real-valued; the complex-valued case follows by taking real and imaginary parts.

The first integral in the hypotheses automatically exists, and by Tonelli’s Theorem the other two also exist and equal the first one; thus if one is finite, they all are, and are equal. For each \(x \in X\), we have that \(f_{[x]}\) is \(B\)-measurable by Lemma XIV.6.1.9., and \(|f_{[x]}| = |f|_{[x]}\); and since

\[
\int_{X \times Y} |f| \, d(\mu \times \nu) = \int_Y \left[ \int_X |f_{[x]}| \, d\nu \right] \, d\mu = \int_X \left[ \int_Y |f_{[x]}| \, d\nu \right] \, d\mu < \infty
\]

it follows that

\[
\int_X |f_{[x]}| \, d\nu < \infty
\]
for almost all \( x \), i.e. \( f_{[x]} \) is \( \nu \)-integrable for almost all \( x \in X \). Similarly, \( f^{[y]} \) is \( \mu \)-integrable for almost all \( y \in Y \).

Let \( f_+ \) and \( f_- \) be the positive and negative parts of \( f \); then \( f_+, f_- \leq |f| \), so are \((\mu \times \nu)\)-integrable. We have \((f_+)_x = (f_{[x]})_+ \) and \((f_-)_x = (f_{[x]})_- \) for all \( x \in X \), and if

\[
   g_+(x) = \int_Y (f_+)_x \, d\nu = \int_Y (f_{[x]})_+ \, d\nu
\]

\[
   g_-(x) = \int_Y (f_-)_x \, d\nu = \int_Y (f_{[x]})_- \, d\nu
\]

we have that

\[
   g_+(x) \leq \bar{g}(x) := \int_Y |f_{[x]}| \, d\nu = \int_Y |f|_{[x]} \, d\nu
\]

and since

\[
   \int_X \bar{g} \, d\mu = \int_{X \times Y} |f| \, d(\mu \times \nu) < \infty
\]

by Tonelli’s Theorem, \( g_+ \) is \( \mu \)-integrable. Similarly, \( g_- \) is \( \mu \)-integrable. In particular, \( g_+ \) and \( g_- \) are finite \( \mu \)-a.e., and if they are both finite at \( x \), then \( f_{[x]} = (f_{[x]})_+ - (f_{[x]})_- \) and

\[
   g(x) = \int_Y f_{[x]} \, d\mu = \int_Y [(f_{[x]})_+ - (f_{[x]})_-] \, d\mu = \int_Y (f_{[x]})_+ \, d\mu - \int_Y (f_{[x]})_- \, d\mu = g_+(x) - g_-(x) .
\]

Thus \( g \) is defined \( \mu \)-a.e. and is equal \( \mu \)-a.e. to a \( \mu \)-integrable function (note though that \( g_+ \) and \( g_- \) are not necessarily the positive and negative parts of \( g \)). We also have

\[
   \int_X g \, d\mu = \int_X g_+ \, d\mu - \int_X g_- \, d\mu = \int_X \left[ \int_Y (f_+)_x \, d\nu \right] \, d\mu - \int_X \left[ \int_Y (f_-)_x \, d\nu \right] \, d\mu
\]

\[
   = \int_{X \times Y} f_+ \, d(\mu \times \nu) - \int_{X \times Y} f_- \, d(\mu \times \nu) = \int_{X \times Y} f \, d(\mu \times \nu)
\]

by Tonelli’s Theorem applied to \( f_+ \) and \( f_- \). The proof of the second equality is analogous.

Examples where \( \sigma \)-finite hypothesis is necessary

The Complete Case

Extension to Sets of \( \sigma \)-Finite Measure

Fubini-Tonelli

We have the following corollary of the Fubini-Tonelli Theorem, which can be used to justify the usual iterated integrals encountered in vector calculus problems:
XIV.6.1.15. Corollary. Let \((X, \mathcal{A}, \mu)\) and \((Y, \mathcal{B}, \nu)\) be measure spaces, \(C \in \mathcal{A} \otimes \mathcal{B}\) of \(\sigma\)-finite measure, and \(f : C \to \mathbb{R}\) a function which is either nonnegative and \((\mu \times \nu)\)-measurable, or integrable over \(C\) (i.e. \(f\chi_C\) is integrable over \(X \times Y\)). Then

(i) The function \(g(x) = \int_{C_x} f(x, y) \, d\nu(y)\) is \(\mathcal{A}\)-measurable, and is \(\mu\)-integrable for almost all \(x\) if \(f\) is \((\mu \times \nu)\)-integrable over \(C\). ***FIX

(ii) The function \(h(y) = \int_{C_y} f(x, y) \, d\mu(x)\) is \(\mathcal{B}\)-measurable, and is \(\nu\)-integrable for almost all \(x\) if \(f\) is \((\mu \times \nu)\)-integrable over \(C\). ***FIX

(iii) We have that

\[
\int_C f \, d(\mu \times \nu) = \int_X \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x) = \int_Y \left[ \int_{C_y} f(x, y) \, d\mu(x) \right] d\nu(y).
\]

Proof: Let \(A, M \in \mathcal{A}\), \(B, N \in \mathcal{B}\) be as in the proposition. Then

\[
\int_C f \, d(\mu \times \nu) = \int_{X \times Y} f \chi_C \, d(\mu \times \nu) = \int_{A \times B} f \chi_C \, d(\mu \times \nu)
\]

(since \(f\chi_C = f\chi_C \chi_{A \times B}\) a.e.) Then, by the Fubini-Tonelli Theorem applied to \(f\chi_C\) on \(A \times B\), \(g_0(x) = \int_B f(x, y) \chi_C(x, y) \, d\nu(y) = \int_{B \cap C_x} f(x, y) \, d\nu(y)\) is \(\mathcal{A}\)-measurable, and integrable for almost all \(x\) if \(f\) is integrable. Since \(C_x \setminus B \subseteq N\), we have \(g_0(x) = \int_{C_x} f(x, y) \, d\nu(y)\), i.e. \(g(x) = g_0(x)\) for \(x \in A\). Also, \(g(x) = 0\) for \(x \notin A \cup M\) since \(C_x \subseteq N\), so \(g(x) = 0\) a.e. on \(X \setminus A\). Then, by Fubini-Tonelli on \(A \times B\),

\[
\int_{A \times B} f \chi_C \, d(\mu \times \nu) = \int_A \left[ \int_B f(x, y) \chi_C(x, y) \, d\nu(y) \right] d\mu(x)
\]

\[
= \int_A \left[ \int_{C_x} f(x, y) \chi_{C_x}(x, y) \, d\nu(y) \right] d\mu(x) = \int_A \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x)
\]

\[
= \int_A \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x) + \int_M \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x) +
\]

\[
+ \int_{X \setminus (A \cup M)} \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x) = \int_X \left[ \int_{C_x} f(x, y) \, d\nu(y) \right] d\mu(x)
\]

since \(\mu(M) = 0\) and \(\nu(C_x) = 0\) for \(x \notin A \cup M\). The proof for the iterated integral in the opposite order is nearly identical.

\(\diamondsuit\)

XIV.6.1.16. As a sample application, suppose \(C\) is a region in \(\mathbb{R}^2\) of the form

\[
\{(x, y) : g(x) \leq y \leq h(x), a \leq x \leq b\}
\]

where \(g, h\) are continuous functions with \(g(x) \leq h(x)\) for \(a \leq x \leq b\), and suppose \(f\) is a Lebesgue measurable function from \(C\) to \(\mathbb{R}\) which is either nonnegative or integrable over \(C\) (e.g. a continuous function.) Then \(C_x = [g(x), h(x)]\) for each \(x \in [a, b]\), and so

\[
\int_C f \, d\lambda^2 = \int_{[a,b]} \left[ \int_{[g(x), h(x)]} f(x, y) \, d\lambda(y) \right] \, d\lambda(x) = \int_a^b \left[ \int_{g(x)}^{h(x)} f(x, y) \, dy \right] \, dx.
\]

1636
XIV.6.2. Lebesgue Integration on $\mathbb{R}^n$

XIV.6.3. Exercises

XIV.6.3.1. Another Nonmeasurable Set. [?] Assume the Continuum Hypothesis, and let $\preceq$ be a well-ordering on $[0, 1]$ according to the first uncountable ordinal, i.e. such that the set of predecessors of any element is countable. Set

$$E = \{(x, y) \in [0, 1]^2 : x \preceq y\}.$$ 

(a) For fixed $y \in [0, 1]$, the set $E^y = \{x \in [0, 1] : x \preceq y\}$ is countable, hence Lebesgue measurable, and $\lambda(E^y) = 0$. Thus, if $E$ is Lebesgue measurable, by the Junior Tonelli Theorem (XIV.6.1.12.)

$$\lambda^2(E) = \int_0^1 \lambda(E^y) \, d\lambda(y) = 0.$$ 

(b) For fixed $x \in [0, 1]$, the set $E_x = \{y \in [0, 1] : x \preceq y\}$ is cocountable, hence Lebesgue measurable, and $\lambda(E_x) = 1$. Thus, if $E$ is Lebesgue measurable, by the Junior Tonelli Theorem

$$\lambda^2(E) = \int_0^1 \lambda(E_x) \, d\lambda(x) = 1.$$ 

(c) Conclude that $E$ is not a Lebesgue measurable subset of $[0, 1]^2$.

This suggests why the CH cannot be proved in ZF: there can be no “measurable” bijection from $\mathbb{R}$ to $\omega_1$.

XIV.6.3.2. [?] Let $X$ be the set of ordinals less than the first uncountable ordinal, $\mathcal{A}$ the $\sigma$-algebra of countable-cocountable subsets of $X$, and $\mu$ the measure on $(X, \mathcal{A})$ with $\mu(A) = 0$ if $A$ is countable and $\mu(A) = 1$ if $A$ is cocountable. Let $E = \{(\alpha, \beta) : \alpha \leq \beta\} \subseteq X \times X$.

(a) Show that the iterated integrals

$$\int_X \int_X \chi_E(\alpha, \beta) \, d\mu(\alpha) \, d\mu(\beta) \text{ and } \int_X \int_X \chi_E(\alpha, \beta) \, d\mu(\beta) \, d\mu(\alpha)$$

both exist but are not equal.

(b) Conclude that $E$ is not $(\mathcal{A} \otimes \mathcal{A})$-measurable, i.e. $\chi_E$ is not $(\mu \times \mu)$-integrable. (It is not hard to see directly that $E \notin \mathcal{A} \otimes \mathcal{A}$.)

This is a variation of XIV.6.3.1. which does not use the Continuum Hypothesis or the AC.
XIV.7. The Henstock-Kurzweil Integral

The Henstock-Kurzweil (HK) integral on \( \mathbb{R} \) is a beautiful and rather simple extension of the Riemann integral. The only modification which needs to be made is to consider a more general and flexible notion of “fineness” of a partition than its mesh. Once some preliminary technicalities are handled, the theory proceeds essentially identically to the Riemann integral case. One ends up with an integral which not only extends the Riemann integral, but even (slightly) the Lebesgue integral, with nice properties such as an unrestricted Fundamental Theorem of Calculus.

XIV.7.1. Definition and Basic Properties

Recall the definition of gauges and fine partitions ().

XIV.7.1.1. Definition. Let \([a, b]\) be a closed bounded interval and \(f : [a, b] \to \mathbb{R}\) a function. Then \(f\) is \(HK\)-integrable on \([a, b]\), with integral \(I\), if for every \(\epsilon > 0\) there is a gauge \(\delta\) on \([a, b]\) such that \(|\mathcal{R}(f, \mathcal{P}) - I| < \epsilon\) for every \(\delta\)-fine partition \(\mathcal{P}\) of \([a, b]\).

It is obvious that the integral \(I\) is unique if it exists.

We first observe that the HK-integral extends the Riemann integral:

XIV.7.1.2. Proposition. Let \(f : [a, b] \to \mathbb{R}\) be Riemann integrable on \([a, b]\). Then \(f\) is HK-integrable on \([a, b]\), with integral \(I = \int_a^b f(t) \, dt\).

Proof: Let \(\epsilon > 0\), and choose \(\delta > 0\) such that whenever \(\mathcal{P}\) is a tagged partition with \(\|\mathcal{P}\| < \delta\), we have \(|\mathcal{R}(f, \mathcal{P}) - I| < \epsilon\). Then the constant gauge \(\delta\) works in the definition of HK-integration. \(\Box\)

XIV.7.1.3. Thus we may justifiably use the notation \(\int_a^b f(t) \, dt\) for the HK-integral of an HK-integrable function.

Conversely, if we can always choose a constant gauge for any \(\epsilon\), then \(f\) is Riemann integrable. Since there are many nonconstant gauges with varying properties which can be used in the definition of HK-integration, it is much easier to satisfy the HK definition for a given \(f\) and one might expect that many more functions are HK-integrable. The next result illustrates this dramatically:

XIV.7.1.4. Theorem. Let \([a, b]\) be a closed bounded interval, and \(A\) a Lebesgue measurable subset of \([a, b]\). Then \(\chi_A\) is HK-integrable over \([a, b]\) with integral \(\lambda(A)\).

Proof: Set \(B = [a, b] \setminus A\). Let \(\epsilon > 0\). There is a sequence \((I_n)\) of open intervals such that \(A \subseteq \bigcup_n I_n\) and \(\sum_n \ell(I_n) < \lambda(A) + \epsilon\), and a sequence \((J_n)\) of open intervals such that \(B \subseteq \bigcup_n J_n\) and \(\sum_n \ell(J_n) < \lambda(B) + \epsilon\). If \(t \in A\), choose an \(n\) (say, the smallest \(n\)) for which \(t \in I_n\), and set \(\delta(t)\) equal to the distance from \(t\) to \(I_n^c\), and similarly if \(t \in B\) choose an \(n\) for which \(t \in J_n\) and set \(\delta(t)\) equal to the distance from \(t\) to \(J_n^c\). Then \(\delta\) is a gauge on \([a, b]\).

1638
Let \( \mathcal{P} = \{a = x_0, \ldots, x_m = b; t_1, \ldots, t_m\} \) be a \( \delta \)-fine partition of \([a, b]\). Then each \( t_k \) is either in \( A \) or \( B \). Let \( N_1 \) be the set of \( k \) for which \( t_k \in A \), and \( N_2 \) the set of \( k \) for which \( t_k \in B \). For each \( k \in N_1 \), the interval \((x_{k-1}, x_k)\) is contained in \( I_n \) for some \( n \); thus we have
\[
\sum_{k \in N_1} \ell([x_{k-1}, x_k]) \leq \sum_n \ell(I_n) < \lambda(A) + \epsilon .
\]
Similarly, we have
\[
\sum_{k \in N_2} \ell([x_{k-1}, x_k]) \leq \sum_n \ell(I_n) < \lambda(B) + \epsilon = (b - a) - [\lambda(A) - \epsilon] .
\]
Since
\[
\sum_{k \in N_1} \ell([x_{k-1}, x_k]) + \sum_{k \in N_2} \ell([x_{k-1}, x_k]) = \sum_{k=1}^m \ell([x_{k-1}, x_k]) = b - a
\]
we have
\[
\lambda(A) - \epsilon < \mathcal{R}(\chi_A, \mathcal{P}) = \sum_{k \in N_1} \ell([x_{k-1}, x_k]) < \lambda(A) + \epsilon .
\]

**XIV.7.1.5.** Unbounded functions cannot be Riemann integrable, but they can be HK-integrable because gauges can be made to force certain points to be tags. See **XIV.7.2.4.** for examples.

**XIV.7.2.** The Fundamental Theorem of Calculus

The following version of the Fundamental Theorem of Calculus is one of the most attractive features of HK-integration. It fails for Riemann integration \((\quad)\) and even for Lebesgue integration \((\quad)\). The proof is quite simple but uses a clever choice of gauge.

**XIV.7.2.1.** **Theorem.** [Fundamental Theorem of Calculus, HK Version] Let \([a, b]\) be a closed bounded interval, and \( f \) a differentiable function on \([a, b]\). Then \( f' \) is HK-integrable on \([a, b]\) and
\[
\int_a^b f'(t) \, dt = f(b) - f(a) .
\]

**Proof:** Let \( \epsilon > 0 \). For each \( t \in [a, b] \) there is a \( \delta(t) > 0 \) such that
\[
\left| \frac{f(x) - f(t)}{x - t} - f'(t) \right| < \frac{\epsilon}{b - a}
\]
whenever \( x \in [a, b], \; 0 < |x - t| < \delta(t) \). Then \( \delta \) is a gauge on \([a, b]\).

Let \( \mathcal{P} = \{a = x_0, \ldots, x_n = b; t_1, \ldots, t_n\} \) be a \( \delta \)-fine partition of \([a, b]\). For each \( k \), since \( x_{k-1} \leq t_k \leq x_k \) we have
\[
|f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})| = |f(x_k) - f(t_k) - f'(t_k)(x_k - t_k) + f(t_k) - f(x_{k-1}) - f'(t_k)(t_k - x_{k-1})|.
\]

1639
\[ \leq |f(x_k) - f(t_k) - f'(t_k)(x_k - t_k)| + |f(t_k) - f(x_{k-1}) - f'(t_k)(t_k - x_{k-1})| \]

\[ < \frac{\epsilon}{b-a}(x_k - t_k) + \frac{\epsilon}{b-a}(t_k - x_{k-1}) = \frac{\epsilon}{b-a}(x_k - x_{k-1}) \]

since \( x_k - t_k, t_k - x_{k-1} < \delta(t_k) \), and thus

\[ |f(b) - f(a) - \mathcal{R}(f', \hat{P})| = \left| \sum_{k=1}^{n} [f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})] \right| \]

\[ \leq \sum_{k=1}^{n} |f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})| \leq \sum_{k=1}^{n} \frac{\epsilon}{b-a}(x_k - x_{k-1}) = \frac{\epsilon}{b-a}(b-a) = \epsilon . \]

We can extend this result to the case where \( f \) is not quite differentiable everywhere:

**XIV.7.2.2. Theorem.** [Fundamental Theorem of Calculus, General HK Version] Let \([a, b]\) be a closed bounded interval, and \( f \) a continuous function on \([a, b]\) which is differentiable except at countably many points. If \( g \) is a function on \([a, b]\) which equals \( f' \) a.e. on \([a, b]\), then \( g \) is HK-integrable on \([a, b]\) and

\[ \int_{a}^{b} g(t) \, dt = f(b) - f(a) . \]

**XIV.7.2.3.** In XIV.7.2.2., we cannot relax the hypothesis that \( f \) be differentiable except at countably many points to simply require that \( f \) be differentiable almost everywhere. If \( g \) is a nonnegative measurable function on \([a, b]\) which is not Lebesgue integrable on \([a, b]\), hence not HK-integrable on \([a, b]\) (), there is a continuous and a.e. differentiable function \( f \) on \([a, b]\) with \( f' = g \) a.e. (XIV.7.2.10.).

**XIV.7.2.4. Examples.** (i) Let \( f(x) = x^2 \sin \frac{1}{x^2} \) (with \( f(0) = 0 \)). Then \( f \) is differentiable everywhere, and

\[ f'(x) = 2x \sin \frac{1}{x^2} - 2 \cos \frac{1}{x^2} \]

\( (f'(0) = 0) \). But \( f' \) is not Lebesgue integrable on any interval containing 0; however, it is HK-integrable on any (closed bounded) interval. This is an example of an unbounded function which is HK-integrable.

(ii) For a simpler example, \( f(x) = 2\sqrt{x} \) is continuous on \([0, 1]\) and differentiable except at 0. Then \( f'(x) = x^{-1/2} \) is HK-integrable on \([0, 1]\) even though it is unbounded. We have

\[ \int_{0}^{1} t^{-1/2} \, dt = 2\sqrt{1} - 2\sqrt{0} = 2 \]

i.e. the HK-integral equals the improper Riemann integral. This holds generally ().

1640
XIV.8. Convolution

The operation of convolution of functions is valuable in a great many contexts. It is a “multiplication” of functions quite different from ordinary pointwise product, which bears some resemblance to matrix multiplication.

An informal motivation for the operation of convolution comes from multiplying series. We give a completely heuristic version which ignores all questions of convergence. Suppose

\[ \sum_{k=-\infty}^{\infty} a_k x^k \quad \text{and} \quad \sum_{k=-\infty}^{\infty} b_k x^k \]

are two-way infinite summations. We want to “multiply” these to get a new two-way infinite summation

\[ \left[ \sum_{k=-\infty}^{\infty} a_k x^k \right] \left[ \sum_{k=-\infty}^{\infty} b_k x^k \right] = \sum_{k=-\infty}^{\infty} c_k x^k \]

and the “formula” for \( c_k \) is obtained by formally multiplying the two expressions and collecting together like powers of \( x \) to obtain

\[ c_k = \sum_{i+j=k} a_j b_i = \sum_{j=-\infty}^{\infty} a_j b_{k-j} . \]

Such a computation is meaningless as done informally here, but with suitable tweaking and interpretation becomes meaningful and useful.

We will not trace the full convoluted history of the development of this subject here.

XIV.8.1. Convolution of Functions on \( \mathbb{R} \), Continuous Case

In this subsection, we will discuss properties of convolution which do not require a knowledge of Lebesgue integration; all integrals in this subsection will be Riemann integrals or improper Riemann integrals. We will consider only continuous functions, and mostly only functions with compact support, i.e. functions which vanish (are identically zero) outside a bounded interval.

XIV.8.1.1. Definition. Let \( f \) and \( g \) be continuous functions from \( \mathbb{R} \) to \( \mathbb{R} \). Suppose for every \( x \in \mathbb{R} \) the integral

\[ \int_{-\infty}^{+\infty} f(y)g(x-y) \, dy = \int_{\mathbb{R}} f(y)g(x-y) \, dy \]

exists as an improper Riemann integral and is finite. The function \( f * g : \mathbb{R} \to \mathbb{R} \) defined by

\[ (f * g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y) \, dy \]

is called the convolution of \( f \) and \( g \).
XIV.8.1.2. Proposition. Let $f$ and $g$ be continuous functions on $\mathbb{R}$. If $f \ast g$ is defined, then $g \ast f$ is also defined and $g \ast f = f \ast g$.

Proof: Fix $x \in \mathbb{R}$. Then

$$(f \ast g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y) \, dy = \lim_{a \to -\infty, b \to +\infty} \int_a^b f(y)g(x-y) \, dy.$$ 

Make the change of variable $z = x - y$. Then, for $a, b \in \mathbb{R}$,

$$\int_a^b f(y)g(x-y) \, dy = \int_{x-a}^{x-b} f(x-z)g(z)(-1) \, dz = \int_{x-b}^{x-a} f(x-z)g(z) \, dz$$

and since $c = x - b \to -\infty$ and $d = x - a \to +\infty$ as $a \to -\infty$ and $b \to +\infty$, we have that

$$\int_{-\infty}^{+\infty} g(z)f(x-z) \, dz = \lim_{c \to -\infty, d \to +\infty} \int_c^d g(z)f(x-z) \, dz = \lim_{a \to -\infty, b \to +\infty} \int_a^b f(y)g(x-y) \, dy$$

eexists as an improper Riemann integral and equals $(f \ast g)(x)$. $\blacksquare$

XIV.8.1.3. One can be fooled into thinking that $g \ast f = -f \ast g$ since the substitution $z = x - y$ introduces a minus sign into the integrand. But the direction of integration is also reversed, canceling the negation. It is more straightforward to use the change-of-variables formula (\ref{civ1}); cf. (\ref{civ2}). To see that $g \ast f$ must be $+f \ast g$ and not $-f \ast g$, consider the case where $f$ and $g$ are nonnegative; then $f \ast g$ and $g \ast f$ are also nonnegative.

The Compact Support Case

XIV.8.1.4. Suppose $f$ and $g$ are continuous functions, and $f$ has compact support. Then $f \ast g$ is defined. For if $f$ vanishes outside $[a, b]$, then for each $x$ we have

$$(f \ast g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y) \, dy = \int_a^b f(y)g(x-y) \, dy$$

as an ordinary Riemann integral. The convolution $g \ast f$ is also defined by XIV.8.1.2. So for existence of the convolution it suffices that one of the functions have compact support.

Continuity of the Convolution

To discuss continuity of $f \ast g$, we need the following fact:

XIV.8.1.5. Proposition. Let $f$ be a continuous function on $\mathbb{R}$. For $x \in \mathbb{R}$, define $f_x(y) = f(x + y)$. If $x_n \to x$ in $\mathbb{R}$, then $f_{x_n} \to f_x$ u.c.

Proof: Let $\epsilon > 0$, and $[a, b]$ a bounded interval in $\mathbb{R}$. Since $f$ is uniformly continuous on bounded intervals (\ref{hiv1}), there is a $\delta > 0$ such that if $y, z \in [a + x - 1, b + x + 1]$ and $|y - z| < \delta$, then $|f(y) - f(z)| < \epsilon$. Fix $N$ such that $|x_n - x| < \min(\delta, 1)$ for all $n \geq N$. Then, if $y \in [a, b]$ and $n \geq N$, we have

$$|f_{x_n}(y) - f_x(y)| = |f(x_n + y) - f(x + y)| < \epsilon.$$ $\blacksquare$
XIV.8.1.6. PROPOSITION. Let \( f \) and \( g \) be continuous functions, with \( f \) of compact support. Then \( f \ast g \) is continuous.

PROOF: Let \( x_n \to x \) in \( \mathbb{R} \). Set \( h_n(y) = f(y)g(x_n - y) = f(y)g(x_n - y) \) and \( h(y) = f(y)g(x - y) = f(y)g(x - y) \).

Since \( g_{x_n} \to g_x \) u.c. by XIV.8.1.5., and \( f \) has compact support, \( h_n \to h \) uniformly on \( \mathbb{R} \); and the \( h_n \) and \( h \) all vanish outside a finite interval \([a, b]\) (the support of \( f \)). Thus

\[
(f \ast g)(x_n) = \int_{-\infty}^{+\infty} f(y)g(x_n - y) \, dy = \int_{a}^{b} h_n(y) \, dy \to \int_{a}^{b} h(y) \, dy = \int_{-\infty}^{+\infty} f(y)g(x - y) \, dy = (f \ast g)(x)
\]

so \( f \ast g \) is continuous by the Sequential Criterion.

XIV.8.1.7. PROPOSITION. Let \( f \) and \( g \) be continuous functions of compact support contained in \([a, b]\) and \([c, d]\) respectively. Then \( f \ast g \) has compact support contained in \([a + c, b + d]\).

PROOF: If \( x < a + c \), then for every \( y \) either \( y < a \) or \( x - y < c \), so \( f(y)g(x - y) = 0 \) for all \( y \), and similarly for \( x > b + d \).

Differentiability of the Convolution

The convolution of two functions \( f \) and \( g \) has the pleasant property of being least as smooth as the smoother of \( f \) and \( g \). We next show the continuous version of this principle.

XIV.8.1.8. THEOREM. Let \( f \) be a continuous function and \( g \) a \( C^r \) function on \( \mathbb{R} \) (\( 0 \leq r \leq \infty \)). Suppose either \( f \) or \( g \) has compact support. Then \( f \ast g \) is \( C^r \), and \( (f \ast g)^{(k)} = f \ast (g^{(k)}) \) for \( 1 \leq k \leq r \).

PROOF: The case \( r = 0 \) is XIV.8.1.6. Suppose \( g \) is \( C^1 \). Then the support of \( g' \) is contained in the support of \( g \), so \( f \ast (g') \) is defined.

Set \( H(x, y) = f(y)g(x - y) \). Then \( \frac{\partial H}{\partial x}(x, y) = f(y)g'(x - y) \), which is continuous by assumption.

If \( f \) has compact support contained in \([a, b]\), then by (1) we have

\[
(f \ast g)'(x) = \frac{d}{dx} \left( \int_{-\infty}^{+\infty} f(y)g(x - y) \, dy \right)(x) = \frac{d}{dx} \left( \int_{a}^{b} H(x, y) \, dy \right)(x)
\]

\[
= \int_{a}^{b} \frac{\partial H}{\partial x}(x, y) \, dy = \int_{-\infty}^{+\infty} f(y)g'(x - y) \, dy = (f \ast (g'))(x)
\]

The argument if \( g \) has compact support is similar, but slightly more complicated. Suppose the support of \( g \) is contained in \([a, b]\). Fix \( x_0 \), and fix an interval \([c, d]\) with \( x_0 \in (c, d) \). Then, restricting \( x \) to \([c, d]\),

\[
(f \ast g)'(x_0) = \frac{d}{dx} \left( \int_{-\infty}^{+\infty} f(y)g(x - y) \, dy \right)(x_0) = \frac{d}{dx} \left( \int_{c}^{d} H(x, y) \, dy \right)(x_0)
\]

\[
= \int_{c}^{d} \frac{\partial H}{\partial x}(x_0, y) \, dy = \int_{-\infty}^{+\infty} f(y)g'(x_0 - y) \, dy = (f \ast (g'))(x_0)
\]

This proves the result for \( r = 1 \). For larger \( r \) apply the \( r = 1 \) case inductively to \( f \) and \( g^{(k)} \).
XIV.8.1.9. Of course, by XIV.8.1.2., the result also holds if \( f \) is \( C^r \), \( g \) is continuous, and at least one has compact support; in this case \((f * g)(k) = (f^{(k)}) * g\). If \( f \) and \( g \) are both \( C^1 \), and at least one has compact support, then \((f') * g = f * (g')\). If \( f \) and \( g \) are \( C^2 \), then \((f'') * g = (f') * (g'') = f * (g'')\), etc.

In fact, we have:

XIV.8.1.10. Corollary. Let \( f \) be a \( C^r \) function and \( g \) a \( C^s \) function on \( \mathbb{R} \) \((0 \leq r, s < \infty)\). Suppose either \( f \) or \( g \) has compact support. Then \( f * g \) is \( C^{r+s} \), and \((f * g)^{(r+s)} = (f^{(r)}) * (g^{(s)})\).

Proof: The convolution \( f * g \) is \( C^r \) and \((f * g)^{(r)} = (f^{(r)}) * g\); and \((f^{(r)}) * g \) is \( C^s \) with \([f^{(r)}] * g^{(s)} = (f^{(r)}) * (g^{(s)})\).

Convolutions on Intervals

XIV.8.1.11. If \( f \) has compact support contained in \([a, b]\), and \( g \) is any continuous function on \( \mathbb{R} \), then for any interval \([c, d]\) the values of \( f * g \) on \([c, d]\) depend only on the values of \( g \) on the interval \([c - b, d - a]\).
This leads to two important observations:
(1) The function \( g \) can be modified arbitrarily outside the interval \([c - b, d - a]\) without affecting the values of \( f * g \) on \([c, d]\). In particular, there is a continuous function \( h \) of compact support such that \( f * h \) agrees with \( f * g \) on \([c, d]\). This observation is useful in studying local properties of \( f * g \); in particular, it can be used to give an alternate proof of the “harder” half of XIV.8.1.8.
(2) If \( g \) is any continuous function on \([c - b, d - a]\), then we can define the convolution of \( f \) and \( g \) as a function on the interval \([c, d]\).

Algebraic Properties of Convolution

Convolution satisfies all the algebraic properties of a reasonable “multiplication”. We have already established the commutative property \( f * g = g * f \) (XIV.8.1.2.). Another simple property whose proof is left to the reader (Exercise (i)) is the distributive or bilinear property:

XIV.8.1.12. Proposition. Let \( f_1, f_2, g \) be continuous functions on \( \mathbb{R} \), and \( \alpha \in \mathbb{R} \).

(i) If \( f_1 * g \) and \( f_2 * g \) are defined, then \((f_1 + f_2) * g \) is defined and \((f_1 + f_2) * g = f_1 * g + f_2 * g\).

(ii) If \( f_1 * g \) is defined, then \((\alpha f_1) * g \) is defined and \((\alpha f_1) * g = \alpha (f_1 * g)\).

In other words, for fixed \( g \), convolution \( f * g \) is linear in \( f \). Of course, by commutativity, \( f * g \) is also linear in \( g \) for fixed \( f \). Thus convolution is \emph{bilinear}.

The other important algebraic property, associativity, also holds; the proof requires Fubini’s Theorem for vector calculus (i).
XIV.8.1.13. Proposition. Let \( f, g, h \) be continuous functions on \( \mathbb{R} \). Suppose at least two of the functions have compact support. Then \( (f * g) * h = f * (g * h) \).

Proof: If \( x_0 \in \mathbb{R} \), then the function which does not have compact support can be replaced by a function of compact support without changing either \( (f * g) * h \) or \( f * (g * h) \) in an interval around \( x_0 \), by two applications of XIV.8.1.11.(1). Thus to prove the result we may assume all three functions have compact support. We have

\[
[f * g * h](x_0) = \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} f(z)g(y-z) \, dz \right] h(x_0-y) \, dy
\]

\[
[f * (g * h)](x_0) = \int_{-\infty}^{+\infty} f(z) \left[ \int_{-\infty}^{+\infty} g(y)h(x_0-z-y) \, dy \right] \, dz
\]

where the integrals are actually over a sufficiently large bounded region (rectangle) in \( \mathbb{R}^2 \). By Fubini’s Theorem () the integrals are equal to the double integrals

\[
\int_{\mathbb{R}^2} f(z)g(y-z)h(x_0-y) \, dy \, dz
\]

\[
\int_{\mathbb{R}^2} f(z)g(y)h(x_0-z-y) \, dy \, dz
\]

respectively, where the integrals are again really over a sufficiently large bounded region.

Make the change of variables \( \phi(u,v) = (u+v, v) \). Writing \( y = u + v, z = v \), we have \( x_0 - y = x_0 - u - v \) and \( y - z = u \). We have \( \Delta_\phi(u,v) \equiv 1 \), so by the change-of-variables formula for double integrals () we have

\[
\int_{\mathbb{R}^2} f(z)g(y-z)h(x_0-y) \, dy \, dz = \int_{\mathbb{R}^2} f(v)g(u)h(x_0-u-v) \Delta_\phi(u,v) \, du \, dv = \int_{\mathbb{R}^2} f(v)g(u)h(x_0-u-v) \, du \, dv
\]

(where the integrals are over large bounded regions), so the double integrals are equal.

Integrable Functions

We can give various other conditions insuring the existence of the convolution. If neither function has compact support, then in general restrictions on both functions are necessary.

XIV.8.1.14. Definition. Let \( f : \mathbb{R} \to \mathbb{R} \) be a continuous function. Then \( f \) is integrable (over \( \mathbb{R} \)) if

\[
\|f\|_1 := \int_{-\infty}^{+\infty} |f(x)| \, dx < \infty
\]

(i.e. the improper integral converges).

XIV.8.1.15. The set \( CL^1(\mathbb{R}) \) of integrable continuous functions on \( \mathbb{R} \) is a real vector space, and \( \| \cdot \|_1 \) is a norm () on \( CL^1(\mathbb{R}) \). Every continuous function of compact support is in \( CL^1(\mathbb{R}) \), but \( CL^1(\mathbb{R}) \) is much larger than the set \( C_c(\mathbb{R}) \) of continuous functions of compact support: for example, an integrable continuous function need not even be bounded (Example XIV.8.1.19.). The set \( BCL^1(\mathbb{R}) \) of bounded continuous integrable functions on \( \mathbb{R} \) is a vector space properly between \( C_c(\mathbb{R}) \) and \( CL^1(\mathbb{R}) \).

1645
**XIV.8.1.16.** Proposition. Let \( f \) and \( g \) be continuous functions on \( \mathbb{R} \), with \( f \) integrable and \( g \) bounded. Then \( f \ast g \) is defined and bounded, and \( \| f \ast g \|_\infty \leq \| f \|_1 \| g \|_\infty \).

**Proof:** We have that, for each \( x \),
\[
\int_{-\infty}^{+\infty} |f(x-y)g(y)| \, dy \leq \int_{-\infty}^{+\infty} |f(x-y)| \| g \|_\infty \, dy = \| g \|_\infty \int_{-\infty}^{+\infty} |f(x-y)| \, dy = \| f \|_1 \| g \|_\infty
\]
so by (\( \) the improper integral
\[
\int_{-\infty}^{+\infty} f(x-y)g(y) \, dy
\]
converges and
\[
|(f \ast g)(x)| = \left| \int_{-\infty}^{+\infty} f(x-y)g(y) \, dy \right| \leq \int_{-\infty}^{+\infty} |f(x-y)g(y)| \, dy \leq \| f \|_1 \| g \|_\infty.
\]

**XIV.8.1.17.** Proposition. Let \( f \) and \( g \) be continuous functions on \( \mathbb{R} \), with \( f \) integrable and \( g \) bounded (or vice versa). Then \( f \ast g \) is continuous.

**Proof:** Let \( (f_n) \) be a sequence of continuous functions of compact support with \( \| f_n - f \|_1 \to 0 \) (\( \)). Then
\[
\| f_n \ast g - f \ast g \|_\infty = \| (f_n - f) \ast g \|_\infty \leq \| f_n - f \|_1 \| g \|_{\infty} \to 0
\]
so \( f_n \ast g \to f \ast g \) uniformly. By XIV.8.1.16., each \( f_n \ast g \) is continuous; hence \( f \ast g \) is continuous by (\( \)).

**XIV.8.1.18.** Proposition. Let \( f \) and \( g \) be integrable continuous functions on \( \mathbb{R} \), with at least one of \( f \) and \( g \) bounded. Then \( f \ast g \) is defined and in \( BCL^1(\mathbb{R}) \), and
\[
\| f \ast g \|_1 \leq \| f \|_1 \| g \|_1.
\]

**Proof:** By XIV.8.1.16. and XIV.8.1.17., \( f \ast g \) is defined, and is bounded and continuous. To show it is integrable, consider the integral
\[
\int \int_{\mathbb{R}^2} |f(y)||g(x-y)| \, dx \, dy.
\]
We can change variables in this integral using \( \phi : \mathbb{R}^2 \to \mathbb{R}^2 \) defined by \( \phi(z,y) = (z+y,y) \). Then \( x = z+y \), \( z = x-y \), and we have \( \Delta_\phi(z,y) \equiv 1 \), so
\[
\int \int_{\mathbb{R}^2} |f(y)||g(x-y)| \, dx \, dy = \int \int_{\mathbb{R}^2} |f(y)||g(z)||\Delta_\phi(z,y)| \, dz \, dy = \int \int_{\mathbb{R}^2} |f(y)||g(z)| \, dz \, dy.
\]
By the Fubini-Tonelli Theorem (\( \)) this integral can be computed as an iterated integral
\[
\int_{\mathbb{R}} \int_{\mathbb{R}} |f(y)||g(z)| \, dz \, dy = \int_{\mathbb{R}} |f(y)| \left[ \int_{\mathbb{R}} |g(z)| \, dz \right] \, dy = \left[ \int_{\mathbb{R}} |f(y)| \, dy \right] \left[ \int_{\mathbb{R}} |g(z)| \, dz \right] = \| f \|_1 \| g \|_1.
\]

\[1646\]
XIV.8.1.19. Example. Here is an example showing that the boundedness hypothesis in XIV.8.1.18. cannot be deleted. Define a nonnegative continuous function $f$ on $\mathbb{R}$ as follows:

$$f(x) = \begin{cases} 0 & \text{if } x \in [0, \frac{15}{8}] \\ n & \text{if } x \in [n - \frac{1}{2n^2}, n + \frac{1}{2n^2}], n \geq 2 \\ \text{linear} & \text{if } x \in [n - \frac{1}{n^3}, n - \frac{1}{2n^2}], n \geq 2 \\ \text{linear} & \text{if } x \in [n + \frac{1}{2n^2}, n + \frac{1}{n^3}], n \geq 2 \\ 0 & \text{if } x \in [n + \frac{1}{n^3}, n + 1 - \frac{1}{(n+1)^3}], n \geq 2 \\ f(-x) & \text{if } x < 0 \end{cases}$$

The graph of $f$ is identically 0 except for small “blunt spikes” around integers; see Figure (). We have

$$\int_{-\infty}^{+\infty} f(x) \, dx < \infty$$

$$\int_{-\infty}^{+\infty} [f(x)]^2 \, dx = \infty$$

i.e. $f$ is integrable but not square-integrable.

Consider $f * f$. We have

$$\int_{-\infty}^{+\infty} f(y)f(-y) \, dy = \int_{-\infty}^{+\infty} [f(y)]^2 \, dy = \infty$$

so $f * f(0)$ is not defined. In fact, $f * f(x)$ is defined if and only if $x$ is not an integer.

XIV.8.1.20. It turns out that using Lebesgue measure and integration we can still make a reasonable definition of the convolution $f * f$ of XIV.8.1.19. (or of any two integrable functions), but the convolution is not continuous and only defined “almost everywhere.” This is in fact one of the historical motivations for developing a more general theory of measure and integration. See section ().

XIV.8.1.21. Thus convolution gives a binary operation of $BCL^1(\mathbb{R})$. Convolution on $BCL^1(\mathbb{R})$ retains nice algebraic properties: associativity (), commutativity (XIV.8.1.2.), and bilinearity ()

XIV.8.1.22. Using Hölder’s inequality (), XIV.8.1.16. can be extended. If $1 \leq p < \infty$, a continuous function $f$ on $\mathbb{R}$ is $p$-integrable if

$$\|f\|_p^p := \int_{-\infty}^{+\infty} |f(x)|^p \, dx < \infty .$$
**XIV.8.1.23. Theorem.** Let $1 < p < \infty$, and $q = \frac{p}{p-1}$ the conjugate exponent () of $p$. Let $f$ and $g$ be continuous functions which are $p$-integrable and $q$-integrable respectively. Then $f \ast g$ is defined, continuous, and bounded, and $\|f \ast g\|_\infty \leq \|f\|_p \|g\|_q$.

See () for the proof, which actually gives a more general statement not requiring $f$ and $g$ to be continuous. Using this, another version of XIV.8.1.18. can be obtained with an identical proof:

**XIV.8.1.24. Corollary.** Let $f$ and $g$ be integrable continuous functions on $\mathbb{R}$. Suppose there is a $p$, $1 < p < \infty$, such that $f$ is $p$-integrable and $g$ is $q$-integrable, where $q = \frac{p}{p-1}$ is the conjugate exponent to $p$. Then $f \ast g$ is defined and in $BCL^1(\mathbb{R})$, and

$$\|f \ast g\|_1 \leq \|f\|_1 \|g\|_1.$$ 

In particular, if $f$ and $g$ are continuous, integrable, and 2-integrable (square-integrable), then $f \ast g$ is defined, continuous, and bounded.
XIV.9. The Radon-Nikodým Theorem

XIV.9.1. Exercises

1. Let $\nu$ be a complex measure on $(X, \mathcal{A})$.
   (a) Use the complex Radon-Nikodým Theorem to show that there is a $u \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$ such that
       $$\nu(A) = \int_A u \, d|\nu|$$
   for all $A \in \mathcal{A}$.
   (b) Use Exercise XIII.6.4.1.(b) to show that $|u| = 1$ a.e.
   (c) Let $f \in L^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$, and define a complex measure $\sigma$ on $(X, \mathcal{A})$ by $\sigma(A) = \int_A f \, d\nu$. Show that
       $$|\sigma|(A) = \int_A |f| \, d|\nu|$$
   for $A \in \mathcal{A}$. 

XIV.10. Appendix B

XIV.10.1. Space-Filling Curves

In this section, we will show that there is a continuous function from \([0,1]\) onto the square \([0,1]^2\). This remarkable fact was first discovered by Peano in -_. There are actually many ways to construct such a “space-filling curve”; our approach will be somewhat unconventional, exploiting properties of the Cantor set. For other constructions of space-filling curves, see, for example, [1] or [2] (or many other references.) Essentially the same construction will yield a continuous function from \([0,1]\) onto \([0,1]^n\) for any \(n\), and even a continuous function from \([0,1]\) onto the Hilbert cube \([0,1]^\mathbb{N}\) (an “ultimate space-filling curve”).

Recall that the Cantor set \(K\) is the subset of \([0,1]\) consisting of all numbers which have a triadic expansion consisting only of 0’s and 2’s. Every element of \(K\) has a unique such expansion. \(K\) is an uncountable compact subset of \([0,1]\) which is totally disconnected but without isolated points, which is nowhere dense and has Lebesgue measure 0.

The proof of the next proposition is straightforward and is left to the reader.

**Proposition.** The map
\[
(a_1, a_2, \ldots) \mapsto .a_1a_2\ldots
\]
is a homeomorphism from \(\{0,2\}^\mathbb{N}\) onto \(K\).

Thus \(K\) is homeomorphic to a countably infinite product of 2-point spaces. The key fact about \(K\) we will need, which follows immediately from this (since \(n_0 = nN_0 = 8^2\)), is that \(K\) is homeomorphic to \(K^2 = K \times K\), and also to \(K^n\) for any \(n\) and even to \(K^\mathbb{N}\). An explicit homeomorphism from \(K\) to \(K^2\) is given by:
\[
\rho(a_1a_2a_3\cdots) = (.a_1a_3a_5\cdots, a_2a_4a_6\cdots)
\]
where the \(a_i\) are the triadic digits, either 0 or 2. A similar homeomorphism can be given from \(K\) to \(K^n\) for any \(n\). To get a homeomorphism from \(K\) to \(K^\mathbb{N}\), let \(\gamma\) be a bijection from \(\mathbb{N} \times \mathbb{N}\) to \(\mathbb{N}\), for example \(\gamma(m,n) = 2^m-1(2n-1)\), and set
\[
\sigma(a_1a_2a_3\cdots) = (\sigma_1(a_1a_2a_3\cdots), \sigma_2(a_1a_2a_3\cdots), \sigma_3(a_1a_2a_3\cdots), \cdots)
\]
where \(\sigma_m(a_1a_2a_3\cdots) = a_{\gamma(m,1)}a_{\gamma(m,2)}a_{\gamma(m,3)}\cdots\). So, for the above \(\gamma\),
\[
\sigma(a_1a_2a_3\cdots) = (a_1a_3a_5\cdots, a_2a_4a_6\cdots, a_4a_1a_2a_3\cdots, \cdots)
\]

There is a standard continuous function from \(K\) onto \([0,1]\), defined by
\[
\theta(a_1a_2a_3\cdots) = b_1b_2b_3\cdots
\]
where the \(a_n\) are triadic digits, either 0 or 2, \(b_n = a_n/2\) (i.e. \(b_n = 0\) if \(a_n = 0\) and \(b_n = 1\) if \(a_n = 2\)), and \(b_1b_2b_3\cdots\) is the binary expansion of the image number. This function \(\theta\) is just the restriction of the Cantor function () to \(K\).

Applying \(\theta\) coordinatewise, we obtain continuous functions \(\theta^n\) from \(K^n\) onto \([0,1]^n\), and \(\theta^\mathbb{N}\) from \(K^\mathbb{N}\) onto \([0,1]^\mathbb{N}\). Composing these with the homeomorphisms \(\rho\) and \(\sigma\) above, we obtain continuous surjective maps

\[
\phi = \theta^2 \circ \rho : K \to [0,1]^2, \quad \psi = \theta^\mathbb{N} \circ \sigma : K \to [0,1]^\mathbb{N}
\]
Let $\phi_n$ $(n = 1, 2)$ be the coordinate functions of $\phi$, and $\psi_n$ $(n \in \mathbb{N})$ the coordinate functions of $\psi$. By (\_), $\phi_n$ and $\psi_n$ extend to continuous functions on all of $[0, 1] \setminus K$; also call these extensions $\phi_n$ and $\psi_n$ respectively. Using these as coordinate functions, we obtain surjective continuous functions $\phi : [0, 1] \to [0, 1]^2$ and $\psi : [0, 1] \to [0, 1]^\mathbb{N}$, which is what we wanted. Thus:

**Theorem.** There are continuous functions $\phi$ from $[0, 1]$ onto $[0, 1]^2$, and $\psi$ from $[0, 1]$ onto $[0, 1]^\mathbb{N}$.

The functions we have defined have the following properties:

1. The coordinate functions are differentiable on $[0, 1] \setminus K$, hence differentiable a.e.
2. For any $(s_1, s_2) \in [0, 1]^2$, there is a $t \in [0, 1]$ with $\phi_1(t) = s_1$ and $\phi_2(t) = s_2$. Similarly, for any sequence $(s_1, s_2, \cdots)$ in $[0, 1]$, there is a $t \in [0, 1]$ with $\psi_n(t) = s_n$ for all $n$.

The functions $\phi$ and $\psi$ cannot be injective, since if they were they would be homeomorphisms by (\_); but cubes of different dimensions are not homeomorphic (\_). However, $\phi$ is not terribly far from being injective: no point of $[0, 1]^2$ is traversed more than 5 times (the point $(\frac{1}{2}, \frac{1}{2})$ is traversed exactly 5 times, as are countably many other points). Individual coordinate functions are very far from being injective, however. (In fact, see (\_).)

The coordinate functions $\phi_n$ and $\psi_n$, although differentiable a.e., are not of bounded variation. The graphs of $\phi_1$ and $\phi_2$ look as follows:

The graphs have a somewhat fractal character in that the same patterns are repeated in miniature. However, they are not truly fractal since the scaling factors are different on the horizontal and vertical axes ($\frac{1}{3}$ on the horizontal, $\frac{1}{2}$ on the vertical).

**XIV.11. The Borel Sets Are Not Countably Constructible**

In this section, we will show that none of the Baire classes $\mathcal{BC}_\alpha$ for countable $\alpha$ are equal to the class of all Borel measurable functions, and hence the set $\mathcal{BC}_\alpha$ of sets of Baire class $\alpha$ (for $\alpha$ countable) is not the entire class of Borel sets. Thus the classes $\mathcal{BC}_\alpha$ are strictly increasing in $\alpha$ for $\alpha$ countable, i.e. the entire class of Borel sets or Borel measurable functions cannot be constructed in finitely, or even countably, many steps.

This result was proved by Lebesgue in his 1905 paper [\?], and was the beginning of what today is called descriptive set theory (\_). The following slick refinement of Lebesgue’s original argument, due to W. Sierpiński, which uses an “ultimate space-filling curve,” is adapted from van Rooij and Schikhof, *A Second Course on Real Analysis* [vRS82]. For a rather different argument, see Billingsley, *Probability and Measure* [Bil95], p. 30-32.

**Definition.** Let $\mathcal{F}$ be a set of Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$. A *catalog* for $\mathcal{F}$ is a Borel measurable function $F : [0, 1] \times \mathbb{R} \to \mathbb{R}$ such that, for every $f \in \mathcal{F}$, there is an $s \in [0, 1]$ for which $f(t) = F(s, t)$ for all $t \in \mathbb{R}$.

If $\mathcal{F}$ has a catalog, then obviously so does any subset of $\mathcal{F}$.

We will prove:

**Theorem.** For each countable $\alpha$, $\mathcal{BC}_\alpha$ has a catalog. But $\mathcal{B}$ does not have a catalog. So $\mathcal{BC}_\alpha \neq \mathcal{B}$.

The proof will proceed in a number of steps.
PROPOSITION. For each \( n \), let \( \mathcal{F}_n \) be a set of Borel measurable functions, and let \( \mathcal{F} = \bigcup_n \mathcal{F}_n \). If each \( \mathcal{F}_n \) has a catalog, then \( \mathcal{F} \) has a catalog.

PROOF: Let \( F_n \) be a catalog for \( \mathcal{F}_n \). Let \( h_n \) be a linear (affine) function from \( [\frac{1}{2n}, \frac{1}{2n-1}] \) onto \( [0,1] \), and define \( G_n : [0,1] \times \mathbb{R} \to \mathbb{R} \) by setting \( G_n(s,t) = F_n(h_n(s),t) \) if \( s \in [\frac{1}{2n}, \frac{1}{2n-1}] \) and \( G_n(s,t) = 0 \) otherwise. Then \( G_n \) is Borel measurable by (1) and (2). The sequence of partial sums of \( \sum_n G_n \) converges pointwise on \( [0,1] \times \mathbb{R} \) to a function \( G \) which is thus Borel measurable; and \( G \) is a catalog for \( \mathcal{F} \) since, if \( f \in F_n \) and \( f(t) = F_n(s,t) \), then \( f(t) = G(h_n^{-1}(s),t) \).

COROLLARY. Any countable set of Borel measurable functions has a catalog.

PROOF: It is only necessary to observe that if \( f \) is a single Borel measurable function, then \( \mathcal{F} = \{f\} \) has a catalog. Set \( F(s,t) = f(t) \) for all \( s \).

If \( \mathcal{F} \) is a set of Borel measurable functions, write \( \mathcal{F}^* \) for the set of functions (also Borel measurable) which are pointwise limits of sequences of functions in \( \mathcal{F} \). For example, \( \mathcal{B}C^*_n = \mathcal{B}C_{n+1} \). In general, \( \mathcal{F} \subseteq \mathcal{F}^* \) (use constant sequences).

The most difficult part of the proof of the theorem is to show that if \( \mathcal{F} \) has a catalog, so does \( \mathcal{F}^* \). This will be done using the continuous function \( \psi \) of (1). We first need a useful variation of the notion of limit of a sequence:

DEFINITION. Let \( (a_n) \) be a sequence of numbers. Define \( \lim_{n \to \infty} a_n = \lim_{n \to \infty} a_n \) if the limit exists, and \( \lim_{n \to \infty} a_n = 0 \) if \( \lim_{n \to \infty} a_n \) does not exist.

Thus \( \lim_{n \to \infty} a_n \) exists for every sequence \( (a_n) \) of numbers. If \( (a_n) \subseteq [0,1] \), then \( \lim_{n \to \infty} a_n \in [0,1] \).

PROPOSITION. Let \( (X, \mathcal{A}) \) be a measurable space, and \( (f_n) \) a sequence of real-valued \( \mathcal{A} \)-measurable functions on \( X \). Let \( C = \{x \in X : (f_n(x)) \text{ converges}\} \). Then \( C \in \mathcal{A} \).

PROOF: If \( x \in X \), then \( x \in C \) if and only if \( (f_n(x)) \) is a Cauchy sequence. Thus

\[
C = \bigcap_{k \in \mathbb{N}} \bigcup_{r \in \mathbb{Q}} \bigcup_{n \in \mathbb{N}} \bigcap_{m \geq n} f_m^{-1} \left( \left[ r - \frac{1}{k}, r + \frac{1}{k} \right] \right).
\]

COROLLARY. Let \( (X, \mathcal{A}) \) be a measurable space, and \( (f_n) \) a sequence of real-valued \( \mathcal{A} \)-measurable functions on \( X \). For \( x \in X \), set \( f(x) = \lim_{n \to \infty} f_n(x) \). Then \( f \) is \( \mathcal{A} \)-measurable.

PROOF: If \( C \) is as above, then \( C \in \mathcal{A} \), and \( f_n \to f \) on \( C \), so \( f|_C \) is measurable. \( f \) is constant on \( X \setminus C \), hence measurable. Thus \( f \) is measurable by (1).

THEOREM. If \( \mathcal{F} \) is a set of Borel measurable functions from \( \mathbb{R} \) to \( \mathbb{R} \), and \( \mathcal{F} \) has a catalog, then \( \mathcal{F}^* \) also has a catalog.

1653
Proof: Let $F$ be a catalog for $\mathcal{F}$, and let $\psi : [0,1] \to [0,1]^\mathbb{N}$ be an ultimate space-filling curve as in ??, with coordinate functions $\psi_n \,(n \in \mathbb{N})$. For each $s \in [0,1]$, $t \in \mathbb{R}$, set $G_n(s,t) = F(\psi_n(s),t)$, and $G(s,t) = \lim_{n\to\infty} G_n(s,t)$. Since $h_n(s,t) = (\psi_n(s),t)$ is continuous and hence Borel measurable, so is $G_n = F \circ h_n$, and therefore $G$ is Borel measurable by XIV.11.

We claim $G$ is a catalog for $\mathcal{F}^*$. Suppose $(f_n)$ is a sequence in $\mathcal{F}$ with $f_n \to f$ pointwise. For each $n$ there is an $s_n \in [0,1]$ such that $f_n(t) = F(s_n,t)$ for all $t$. There is an $s \in [0,1]$ such that $s_n = \psi_n(s)$ for all $n \,$. Then $G(s,t) = f(t)$ for all $t \in \mathbb{R}$.

We next show that $\mathcal{B}C_\alpha$ has a catalog for $\alpha$ countable.

Definition. If $\mathcal{F}$ and $\mathcal{G}$ are sets of Borel measurable functions, we say $\mathcal{G}$ is locally uniformly dense in $\mathcal{F}$ if, for every $f \in \mathcal{F}$, $\epsilon > 0$, and $n \in \mathbb{N}$, there is a $g \in \mathcal{G}$ with $|f(t) - g(t)| < \epsilon$ for all $t \in [-n,n]$.

We usually apply this definition when $\mathcal{G} \subseteq \mathcal{F}$.

Example. Here is the most important example for us. Let $\mathcal{Q}$ be the set of polynomials with rational coefficients. Then by the Weierstrass Approximation Theorem $(),$ $\mathcal{Q}$ is locally uniformly dense in the set $\mathcal{C}$ of continuous functions from $\mathbb{R}$ to $\mathbb{R}$.

Proposition. If $\mathcal{G}$ is locally uniformly dense in $\mathcal{F}$, then $\mathcal{F}^* \subseteq \mathcal{G}^*$. (If $\mathcal{G} \subseteq \mathcal{F}$, it follows that $\mathcal{G}^* = \mathcal{F}^*$.)

Proof: Suppose $f_n \in \mathcal{F}$ and $f_n \to f$ pointwise. For each $n$, choose $g_n \in \mathcal{G}$ with $|f_n(t) - g_n(t)| < \frac{1}{n}$ for all $t \in [-n,n]$. Then it is easily checked that $g_n \to f$ pointwise.

In particular, we have $\mathcal{Q}^* = \mathcal{C}^* = \mathcal{B}C_1$.

Proposition. $\mathcal{B}C_\alpha$ has a catalog for every countable $\alpha$.

Proof: $\mathcal{Q}$ is countable, so it has a catalog by ??, Thus $\mathcal{Q}^* = \mathcal{B}C_1$ has a catalog by XIV.11.

Now proceed by transfinite induction. If $\mathcal{B}C_\alpha$ has a catalog, then $\mathcal{B}C_{\alpha+1}$ has a catalog by XIV.11. And if $\alpha$ is a countable limit ordinal, then $\mathcal{B}C_\alpha = \cup_{\beta < \alpha} \mathcal{B}C_\beta$, a countable union, so $\mathcal{B}C_\alpha$ has a catalog by ??.

The next result finishes the proof of ??, The proof is strongly reminiscent of Cantor’s diagonalization argument.

Proposition. $\mathcal{B}$ does not have a catalog.

Proof: Suppose $F$ is a catalog for $\mathcal{B}$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = F(t,t) + 1$ if $t \in [0,1]$, and $f(t) = 0$ otherwise. Then $f|_{[0,1]}$ is a composition of two Borel measurable functions, hence Borel measurable; so $f$ is Borel measurable by $(,)$. Since $F$ is a catalog, there is an $s \in [0,1]$ such that $f(t) = F(s,t)$ for all $t$. But $f(s) = F(s,s) + 1 \neq F(s,s)$, a contradiction.

This completes the proof of ??, We have the following consequence:

Corollary. For any countable $\alpha$, $\mathcal{B}C_\alpha$ is a proper subset of $\mathcal{B}$. Thus the $\mathcal{B}C_\alpha$ are strictly increasing for $\alpha$ countable.

Proof: Suppose $\mathcal{B}C_\alpha = \mathcal{B}$ for some countable $\alpha$. Since $\mathcal{B}C_\alpha$ is an algebra of functions $()$, it follows that every simple Borel measurable function is in $\mathcal{B}C_\alpha$. Hence every Borel measurable function is in $\mathcal{B}C_{\alpha+1}$ by $(,)$, a contradiction.
If $\alpha < \beta < \omega_1$ and $\mathcal{B}_\alpha = \mathcal{B}_\beta$, any countable intersection of sets in $\mathcal{B}_\alpha$ is in $\mathcal{B}_\alpha+1 \subseteq \mathcal{B}_\beta = \mathcal{B}_\alpha$ by (1), and so it follows from (1) that $\mathcal{B}_\alpha$ is a $\sigma$-algebra and hence $\mathcal{B}_\alpha = \mathcal{B}$, a contradiction.  

We then have a fortiori that $\mathcal{B}_\alpha \neq \mathcal{B}_\beta$ for $\alpha < \beta < \omega_1$.

**XIV.12. Exercises**

1. Describe the “motion” of $\phi(t)$ through the unit square as $t$ goes from 0 to 1. [First describe $\phi(t)$ for $t \in \left[\frac{1}{3}, \frac{2}{3}\right]$, $t \in \left[\frac{1}{5}, \frac{2}{5}\right]$, $t \in \left[\frac{7}{9}, \frac{8}{9}\right]$, etc., and then find $\phi([0, \frac{1}{3}])$, $\phi([\frac{2}{3}, 1])$, $\phi([0, \frac{1}{9}])$, $\phi([\frac{2}{9}, \frac{1}{3}])$, etc.]

2. An occasional attempt is made to define a space-filling curve $\omega$ from $[0, 1]$ onto $[0, 1]^2$ by the formula

$$\omega(.b_1 b_2 b_3 \cdots) = (.b_1 b_3 b_5 \cdots , b_2 b_4 b_6 \cdots)$$

where the $b_n$ (0 or 1) are binary digits. Show why this approach does not work. Try to modify the formula to give a well-defined continuous surjection.

3. Each of the coordinate functions of $\phi$ or $\psi$ is a continuous function from $[0, 1]$ to $[0, 1]$ taking each value in $[0, 1]$ uncountably many times. [Fix $a \in [0, 1]$. For each $b \in [0, 1]$ there is a $t_b \in [0, 1]$ such that $\phi(t_b) = (a, b)$, so $\phi_1(t_b) = a$ for all $b$.]
XIV.12.1. A Plane Curve With Positive Area

It is easy to give examples of topologically one-dimensional, and even zero-dimensional, compact subsets of $\mathbb{R}^2$ which have positive planar (Lebesgue) measure (e.g. XIV.12.1.4.). In this section, we construct a simple closed curve in $\mathbb{R}^2$ (Jordan curve), i.e. a subset of $\mathbb{R}^2$ homeomorphic to $S^1$, which has positive planar measure.

Such a set, regarded as a curve, must be highly nondifferentiable: the image of a piecewise continuously differentiable curve (or, more generally, a Lipschitz function) has Hausdorff dimension 1, and hence plane measure 0 ().

The first example of a simple closed curve with positive planar measure was given by Osgood in 1902 ([Osg03], complete with color pictures!); another example was constructed by Besicovitch and Schoenberg in [BS61].

Cantor Sets

XIV.12.1.1. Recall one standard construction of Cantor sets in $\mathbb{R}$.

Let $\alpha$ be a real number $\geq 3$. The Cantor set $K_\alpha$ is constructed by beginning with the closed unit interval $[0,1]$ and at the $n$’th stage removing the open subintervals of length $\alpha^{-n}$ centered in each of the $2^{n-1}$ remaining intervals; $K_\alpha$ is the intersection. $K_\alpha$ is a compact totally disconnected set without isolated points, homeomorphic to a countable product of two-point spaces.

Since the total length of the removed subintervals is

$$\sum_{n=1}^{\infty} \frac{2^{n-1}}{\alpha^n} = \frac{1}{\alpha} \sum_{m=0}^{\infty} \left(\frac{2}{\alpha}\right)^m = \frac{1}{\alpha(1 - \frac{2}{\alpha})} = \frac{1}{\alpha - 2},$$

we obtain that the (one-dimensional) Lebesgue measure of $K_\alpha$ is

$$1 - \frac{1}{\alpha - 2} = \frac{\alpha - 3}{\alpha - 2}.$$  

XIV.12.1.2. The usual Cantor set is $K_3$, just denoted $K$. The measure of $K$ is zero. $K$ consists precisely of all elements of $[0,1]$ with a ternary (base 3) decimal expansion consisting only of 0’s and 2’s, and every element of $K$ has a unique such expansion.

XIV.12.1.3. Given any $\alpha$ and $\beta$, there is an evident homeomorphism of $\mathbb{R}$ (strictly increasing function) carrying $K_\alpha$ onto $K_\beta$. (Note that this shows the well-known fact () that a homeomorphism of $\mathbb{R}$ need not preserve sets of measure zero.)

XIV.12.1.4. If $K_\alpha^2 = K_\alpha \times K_\alpha$, then $K_\alpha^2$ is a compact subset of $[0,1]^2$ homeomorphic to $K$ (a specific homeomorphism will be described below.) The planar measure of $K_\alpha^2$ is

$$\left(\frac{\alpha - 3}{\alpha - 2}\right)^2.$$
A Wandering Curve

We will construct a homeomorphic image of $[0, 1]$ passing through every point of $K^2_\alpha$; thus if $\alpha > 3$ the image will have positive planar measure.

For notational simplicity, we will describe the construction for $\alpha = 3$. The construction for $\alpha > 3$ can be done directly in an analogous way, or by simply composing the curve for $\alpha = 3$ with a homeomorphism of $\mathbb{R}^2$ sending $K^2$ onto $K^2_\alpha$.

**XIV.12.1.5.** We first define an auxiliary function

$$
\psi : \{0\} \cup [1/9, 2/9] \cup [1/3, 2/3] \cup [7/9, 8/9] \cup \{1\} \to [0, 1]^2
$$

by setting $\psi(0) = (0, 0)$, $\psi(1/9) = (0, 1/3)$, $\psi(2/9) = (0, 2/3)$, $\psi(1/3) = (1/3, 2/3)$, $\psi(2/3) = (2/3, 2/3)$, $\psi(7/9) = (1, 2/3)$, $\psi(8/9) = (1, 1/3)$, $\psi(1) = (1, 1)$, and letting $\psi$ be linear (affine) on each of the subintervals. See Figure XIV.23.

![Figure XIV.23: The auxiliary function $\psi$](image)

**XIV.12.1.6.** Before defining the curve $\phi : [0, 1] \to [0, 1]^2$, let us make a definition. A square $S$ with sides parallel to the axes, of length $3^{-n}$ for some $n$, will be called *full* if the four corners of $S$ are in the closure of the intersection of $K^2$ with the interior of $S$. The whole unit square is full; and if a full square is divided into nine subsquares, it is the four corner subsquares which are also full.

**XIV.12.1.7.** We will now define $\phi$ in stages. It is convenient to do the first step of the definition somewhat differently from the others, in order to have $\phi(0) = (1/3, 0)$ and $\phi(1) = (2/3, 0)$; then the segment from $(2/3, 0)$ back to $(1/3, 0)$ can be added to give a simple closed curve. (This segment will be included in the pictures.) So in addition to these definitions of $\phi(0)$ and $\phi(1)$, in the first step we define $\phi$ on $[1/9, 2/9]$ to map linearly to the segment from $\phi(1/9) = (1/3, 1/3)$ to $\phi(2/9) = (1/3, 2/3)$; define $\phi$ on $[1/3, 2/3]$ to
map linearly to the segment from $\phi(1/3) = (1/3, 1)$ to $\phi(2/3) = (2/3, 1)$, and similarly $\phi$ maps $[7/9, 8/9]$ linearly to the segment from $\phi(7/9) = (2/3, 2/3)$ to $\phi(8/9) = (2/3, 1/3)$.

**XIV.12.1.8.** We now proceed inductively. After the $n$’th step $\phi$ is defined except on open intervals of the form $(k/9^n, k+1)/9^n$, where $k$ is 0, 2, 6, or 8 mod 9. For each such $k$, there is a unique full square $S$ with adjacent corners $\phi(k/9^n)$ and $\phi((k+1)/9^n)$, and a unique affine map $\theta$ from the unit square onto $S$ with $\theta(0, 0) = \phi(k/9^n)$ and $\theta(1, 0) = \phi((k+1)/9^n)$. Then for $t \in [1/9, 2/9] \cup [1/3, 2/3] \cup [7/9, 8/9]$ define

$$\phi\left(\frac{k + t}{9^n}\right) = \theta \circ \psi(t)$$

After the third step, the curve looks like Figure XIV.24.

![Figure XIV.24: Three steps of the definition of $\phi$](image)

**XIV.12.1.9.** After the inductive construction, $\phi$ is defined on a dense subset of $[0, 1]$. By construction, if $s, t \in [0, 1]$ and $|s - t| < 9^{-n}$, and $\phi(s)$ and $\phi(t)$ are defined, then $\|\phi(s) - \phi(t)\| < 3^{-n}\sqrt{2}$; thus $\phi$ is uniformly continuous and extends to a continuous map from all of $[0, 1]$ to $[0, 1]^2$.

**XIV.12.1.10.** Proposition. The image of $\phi$ contains $K^2$.

Proof: This is obvious, since after the $n$’th step in the construction every point of $K^2$ lies within $3^{-n}\sqrt{2}$ of a point on the curve.

1658
XIV.12.1.11. **Proposition.** The map $\phi$ is a homeomorphism onto its image.

**Proof:** By compactness, it is only necessary to show that $\phi$ is one-to-one. It clearly maps $[0,1]\setminus K$ one-one into $[0,1]^2\setminus K^2$, so we need only show that $\phi$ maps $K$ one-one into (actually onto) $K^2$. This is geometrically clear, but is rather complicated to write down explicitly. In fact, there is a precise (but complicated) formula giving the first $n$ places of the ternary decimal expansion of the two coordinates of $\phi(t)$ ($t \in K$) in terms of the first $2n$ places of the ternary decimal expansion of $t$. For example, if $t = a_1a_2 \cdots$, $\phi(t) = (b_1 \cdots, c_1 \cdots)$, where $b_1$ is always equal to $a_1$ and $c_1 = a_2$ if $a_1 = 0$, $c_1 = 2 - a_2$ if $a_1 = 2$. It can be checked that if $s, t \in K$ differ in the first $2n$ ternary decimal places, then $\phi(s)$ and $\phi(t)$ differ in the first $n$ ternary decimal places in at least one coordinate.  

XIV.12.1.12. The final curve is depicted in Figure XIV.25. The curve has a fractal character in that all of the full squares with sides of length $\leq 1/3$ are similar up to scaling and rotation.

![Figure XIV.25: The closed curve $\phi$](image)
Connection with Space-Filling Curves

XIV.12.13. This construction bears a close resemblance to one version of the Peano space-filling curve (see, for example, [Edg08, p. 64]). We obtain a space-filling curve by composing $\phi$ with $p \times p$, where $p$ is the usual map from $K$ onto $[0, 1]$ given by

$$p(a_1a_2a_3\cdots) = .b_1b_2b_3\cdots$$

where $a_1a_2\cdots$ is the ternary decimal expansion of $t \in K$, and $b_1b_2\cdots$ is the binary decimal expansion of $p(t)$, with $b_n = a_n/2$. The resulting map $(p \times p) \circ \phi$ maps $[0, 1]$ continuously onto $[0, 1]^2$. (It should be thought of as being constant on the intervals of $[0, 1]_K$.)

XIV.12.14. The map $p$ is not one-to-one since the binary decimal expansion of dyadic rational numbers is not unique, and any cross section $q : [0, 1] \rightarrow K$ for $p$ is necessarily discontinuous; however, the map $\sigma = (p \times p) \circ \phi \circ q$ is continuous and independent of $q$, and maps $[0, 1]$ onto $[0, 1]^2$. This map $\sigma$ is a more usual type of space-filling curve.

Summary and Variations

XIV.12.15. If $C_\alpha$ is the image of the curve $\phi$ done for the Cantor set $K_\alpha$, with the last segment filled in between the endpoints, then $C_\alpha$ is a simple closed curve, contained in $[0, 1]^2$, with planar measure at least the measure of $K_\alpha^2$. Actually, $C_\alpha$ is the union of $K_\alpha^2$ and a countable number of straight line segments, so the measure of $C_\alpha$ is exactly

$$\left(\frac{\alpha - 3}{\alpha - 2}\right)^2.$$

This can be any (nonnegative) number less than 1. If $\alpha$ is large enough ($\alpha > 4 + \sqrt{2}$ will do), the measure of this curve is greater than the measure of the region inside the curve!

There cannot be a simple closed curve in $[0, 1]^2$ of planar measure exactly one, since any nonempty open set in the plane has positive measure.

XIV.12.16. The curve $C_\alpha$ is a Jordan curve, but the region inside the curve is an open set in $\mathbb{R}^2$ (homeomorphic to an open disk) which is not a Jordan region () if $\alpha > 3$ since its boundary does not have measure zero.

XIV.12.17. The example in [BS61] has the additional property that for any $s, t, 0 \leq s < t \leq 1$, the arc from $\phi(s)$ to $\phi(t)$ has positive planar measure. Our example can be modified to have this property, by replacing each straight line segment by countably many small versions of the curve.

XIV.12.18. It would appear that a construction similar to the one of this section can be done in $[0, 1]^n$ for any $n$, giving a simple closed curve passing through every point of $K_\alpha^n$ and thus having positive $n$-dimensional Lebesgue measure. However, the details might be complicated to write carefully. Perhaps a similar construction can even be done in the Hilbert cube. See [PS92] for a version of OSGOOD’S example in $\mathbb{R}^n$. 

1660
XIV.13. Borel Measures on $\mathbb{R}$ and Cumulative Distribution Functions

A finite Borel measure on $\mathbb{R}$ is a finite measure defined on $(\mathbb{R}, B)$. We will also study finite Borel measures on intervals in $\mathbb{R}$, and defined in the same way, and $\sigma$-finite Borel measures on $\mathbb{R}$ which are finite on bounded sets (positive Radon measures). Such measures are important in many contexts, and especially in probability theory.

XIV.13.1. Cumulative Distribution Functions

We would like to have a way of describing the finite Borel measures on $\mathbb{R}$ (or an interval in $\mathbb{R}$). If $\mu$ is a finite Borel measure on $\mathbb{R}$, for each $x \in \mathbb{R}$ we can compute $\mu((1; x])$. It turns out that these numbers are sufficient to recover $\mu$ by the Carathéodory extension process (\()\). This data is conveniently described by a function $F$ from $\mathbb{R}$ to $\mathbb{R}$, setting $F(x) = \mu((1; x])$. Then if $a, b \in \mathbb{R}$, $a < b$, we have $\mu((a, b]) = F(b) - F(a)$.

**Definition.** $F$ is called the cumulative distribution function (or cdf) of $\mu$.

**Proposition.** The function $F$ has the following properties:

(i) $F$ is increasing (nondecreasing).

(ii) $F$ is continuous from the right.

(iii) $\lim_{x \to -\infty} F(x) = 0$.

(iv) $F$ is bounded, and $\sup_{x \in \mathbb{R}} F(x) = \lim_{x \to +\infty} F(x) = \mu(\mathbb{R})$.

**Proof:** (i) is obvious, and (iii) and (iv) follow immediately from (i) since $\bigcup_{n \in \mathbb{N}} (-\infty, n] = \mathbb{R}$ and $\bigcap_{n \in \mathbb{N}} (-\infty, -n] = \emptyset$. (ii) also follows from (i) since for any $x \in \mathbb{R}$, $(-\infty, x] = \bigcap_{n \in \mathbb{N}} (-\infty, x + \frac{1}{n}]$.

The choice to make cdf’s continuous from the right is simply an arbitrary convention. We could have equally well made them continuous from the left by defining $F'_\mu(x) = \mu((-\infty, x))$.

Conversely, a function satisfying (i)–(iv) is the cdf of a unique finite Borel measure:

**Theorem.** Let $F : \mathbb{R} \to \mathbb{R}$ be a function with the following properties:

(i) $F$ is increasing (nondecreasing).

(ii) $F$ is continuous from the right.

(iii) $\lim_{x \to -\infty} F(x) = 0$.

(iv) $F$ is bounded.

Then there is a unique finite Borel measure $\mu$ on $\mathbb{R}$ with $F = F_\mu$, and $\mu(\mathbb{R}) = \sup_{x \in \mathbb{R}} F(x) = \lim_{x \to +\infty} F(x)$.

This theorem will be proved in (\()). The finite Borel measure $\mu$ for which $F = F_\mu$ is called the Lebesgue-Stieltjes measure of (or corresponding to) $F$. If $g$ is a $\mu$-integrable function, $\int g \, d\mu$ is sometimes written $\int g \, dF$ and called the Lebesgue-Stieltjes integral of $g$ with respect to $F$. 

1661
If \( F \) is a nondecreasing function from \( \mathbb{R} \) to \( \mathbb{R} \) and \( E \) is a subset of \( \mathbb{R} \), write \( F^{-1}(E) \) for the “image” of \( E \), saturated by adding intervals for jump discontinuities of \( F \) as in (\).

**Theorem.** Let \( \mu \) be a finite Borel measure on \( \mathbb{R} \), with cdf \( F_\mu \). If \( E \) is any Borel subset of \( \mathbb{R} \), then \( F_\mu^{-1}(E) \) is a Borel set, and \( \mu(E) = \lambda(F_\mu^{-1}(E)) \).

Note that this result is false, even meaningless, if \( E \) is merely a Lebesgue measurable subset of \( \mathbb{R} \); \( F_\mu^{-1}(E) \) need not be Lebesgue measurable (\), and \( \mu(E) \) may not be defined.

**XIV.13.2. Density Functions and Absolute Continuity**

For some, but not all, finite Borel measures, there is a corresponding density function. If \( \mu \) is a finite Borel measure on \( \mathbb{R} \), and there is a Borel measurable function \( \phi \) from \( \mathbb{R} \) to \( \mathbb{R} \) (necessarily nonnegative and integrable) such that \( \mu(E) = \int_E \phi \, d\lambda \) for all Borel sets \( E \subseteq \mathbb{R} \), then \( \phi \) is called a density function for \( \mu \). Two density functions for \( \mu \) must agree almost everywhere (with respect to \( \lambda \)), so by slight abuse of terminology we may refer to the density function of \( \mu \). Any nonnegative integrable function \( \phi \) on \( \mathbb{R} \) is the density function of a unique measure \( \mu_\phi \), defined by \( \mu_\phi(E) = \int_E \phi \, d\lambda \) for \( E \subseteq \mathbb{R} \) Borel.

The function \( \phi \) can be thought of as the “derivative” of the function \( F_\mu \) in a certain sense. This is a precise relationship in nice cases:

**Example.** Let \( \phi \) be a nonnegative continuous function on \( \mathbb{R} \) with \( \int_{-\infty}^{\infty} \phi(t) \, dt \) finite. For a Borel set \( E \subseteq \mathbb{R} \), set \( \mu_\phi(E) = \int_E \phi \, d\lambda \) as above. Then, for \( a, b \in \mathbb{R}, a < b \), we have

\[
F_{\mu_\phi}(b) - F_{\mu_\phi}(a) = \mu_\phi((a, b]) = \int_a^b \phi(t) \, dt
\]

and thus by the Fundamental Theorem of Calculus (\), \( F_{\mu_\phi} \) is an antiderivative for \( \phi \), i.e. \( F_{\mu_\phi} \) is differentiable and \( F'_{\mu_\phi} = \phi \).

This example generalizes to arbitrary density functions \( \phi \):

**Theorem.** Let \( \phi \in L^1_+(\mathbb{R}) \), and \( F_{\mu_\phi} \) the cdf of the corresponding measure \( \mu_\phi \) with density function \( \phi \). Then \( F_{\mu_\phi} \) is continuous, differentiable \( \lambda \)-a.e., and \( F'_{\mu_\phi} = \phi \) \( \lambda \)-a.e.

This theorem will be proved in (\).

A finite Borel measure need not have a density function. Its cdf need not even be continuous:

**Example.** Let \( \delta_0 \) be the Dirac measure (unit point mass) at 0. Then

\[
F_{\delta_0}(x) = \begin{cases} 
1 & \text{if } x \geq 0 \\
0 & \text{if } x < 0
\end{cases}
\]

In fact, the cdf of a finite Borel measure \( \mu \) is a continuous function if and only if \( \mu \) is a continuous measure: in general, for \( x \in \mathbb{R} \),

\[
F_\mu(x+) - F_\mu(x-) = F_\mu(x) - F_\mu(x-) = \mu(\{x\})
\]

1662
where, as usual, \( F_\mu(x+) = \lim_{y \to x^+} F_\mu(y) \) and \( F_\mu(x-) = \lim_{y \to x^-} F_\mu(y) \). (This fact was crucial motivation for the choice of the term “continuous” for a measure with no atoms.)

But even a continuous \( \mu \) need not have a density function:

**Example.** Let \( F \) be the Cantor function on \([0,1]\), extended to be 0 on \((-\infty,0]\) and 1 on \([1,\infty)\). \( F \) satisfies the conditions of XIV.13.1, and thus it is \( F_\mu \) for a unique continuous finite Borel measure \( \mu \) on \( \mathbb{R} \) with \( \mu([0,1]) = 1 \). If \( \mu \) had a density function \( \phi \), then \( \phi \) would have to be equal to \( F' \) \( \lambda \)-a.e. But \( F' = 0 \) \( \lambda \)-a.e., so \( \phi \) would have to be 0 \( \lambda \)-a.e. and \( \mu \) would be 0. Thus \( \mu \) has no density function.

This \( \phi \) is not absolutely continuous with respect to \( \lambda \). For example, if \( K \) is the Cantor set, \( \lambda(K) = 0 \), but \( \mu(K) = 1 \): this follows from XIV.13.1, but can also easily be seen directly by noting that \( \mu(I) = 0 \) for any interval \( I \) in the complement of \( K \), since \( F \) is constant on \( I \). This \( \mu \) is actually the push-forward of the measure of \( \phi \) under the natural map from \( \{0,1\}^\mathbb{N} \) to \( K \subseteq \mathbb{R} \).

So which finite Borel measures have density functions? It follows immediately from (i) that for \( \mu \) to have a density function, it is necessary for \( \mu \) to be absolutely continuous with respect to \( \lambda \). This condition also turns out to be sufficient:

**Theorem.** A finite Borel measure has a density function if and only if it is absolutely continuous with respect to \( \lambda \).

The “if” part of this theorem is a special case of the Radon-Nikodým Theorem (i).

**Corollary.** A finite Borel measure \( \mu \) is absolutely continuous with respect to \( \lambda \) if and only if its cdf \( F_\mu \) is absolutely continuous in the sense of (i).

In fact, this result is the reason behind the choice of the term “absolutely continuous” for the functions of (i).

Even if \( \mu \) does not have a density function, it has a density in a generalized sense (i), which is the “derivative” of \( F_\mu \). Thus, for example, we informally think of the “Dirac \( \delta \)-function” as the “derivative” of the function \( \chi_{[0,\infty)} \). We can think of the actual measure \( \mu \) as being identified with \( F_\mu \), in which case the “density” of \( \mu \) is regarded as the “derivative” of \( \mu \), or we can regard \( \mu \) as being identified with its “density”, with \( F_\mu \) regarded as the “antiderivative” of \( \mu \). The second point of view is more common.

**XIV.13.3. Finite Borel Measures on Intervals**

If \( I \) is an interval in \( \mathbb{R} \), then finite Borel measures on \( I \) can be thought of as finite Borel measures on \( \mathbb{R} \) giving measure 0 to \( \mathbb{R} \setminus I \). The cdf of such a measure is constant on \( \mathbb{R} \setminus I \).

If \( I \) is bounded above, and \( b = \sup(I) \), then the cdf \( F_\mu \) of a finite Borel measure \( \mu \) on \( I \) must satisfy \( F_\mu(b) = \mu(I) \) (= \( \mu(\mathbb{R}) \)). If \( b \notin I \), then \( F_\mu \) is continuous at \( b \); but if \( b \in I \), then \( F_\mu \) can have a discontinuity at \( b \) if \( \mu(\{b\}) > 0 \). In any event, \( \mu(I) = \sup_{x \in I} F_\mu(x) \).

If \( I \) is bounded below with \( a = \inf(I) \), then \( F_\mu \) is identically 0 on \( (-\infty,a) \), but if \( a \in I \), then \( F_\mu(a) = \mu(\{a\}) \), so \( \inf_{x \in I} F_\mu(x) \) is not necessarily 0. But if \( a \notin I \), then we must have \( \inf_{x \in I} F_\mu(x) = 0 \).

**XIV.13.4. Unbounded Distribution Functions and Positive Radon Measures**

The requirement that \( \lim_{x \to -\infty} F(x) = 0 \) for a cdf \( F \) is a convenient normalization, but is not really necessary. If \( \mu \) is a finite Borel measure on \( \mathbb{R} \), \( c \) is any constant, and \( G(x) = F_\mu(x) + c \), then \( G \) can be used just as well.
as $F_\mu$ to recover $\mu$, since for any finite interval $(a, b]$, we have

$$\mu((a, b]] = F_\mu(b) - F_\mu(a) = G(b) - G(a).$$

If $F_\mu$ is regarded as an “antiderivative” of $\mu$, then the general “antiderivative” of $\mu$ is a constant translate of $F_\mu$, and one antiderivative is as good as another for most purposes.

Taking this observation one step farther, if $F$ is any nondecreasing, right-continuous function from $\mathbb{R}$ to $\mathbb{R}$, not necessarily bounded above or below, then $F$ has a Lebesgue-Stieltjes measure $\mu$ defined by the Carathéodory extension process using $\mu((a, b]) = F(b) - F(a)$ for $a, b \in \mathbb{R}$, $a < b$, and $F$ can be regarded as a cdf for $\mu$. This $\mu$ is a positive Radon measure (finite on bounded sets) which is always $\sigma$-finite, but not finite if $F$ is unbounded. Two nondecreasing right-continuous functions define the same Lebesgue-Stieltjes measure if and only if they differ by a constant. The function $F(x) = x$ is the cdf of Lebesgue measure itself.

If $F$ is such a cdf, and $a, b \in \mathbb{R}$, then $F$ can be truncated at $a$ and $b$ to a function $G$ by setting

$$G(x) = \begin{cases} F(a-) & \text{if } x < a \\ F(x) & \text{if } a \leq x \leq b \\ F(b) & \text{if } x > b \end{cases}.$$ 

Then $G$ is a cdf for the measure $\mu_{[a, b]}$, which is a finite Borel measure. Thus the study of positive Radon measures on $\mathbb{R}$ and their cdf’s can be effectively reduced to the theory of finite Borel measures.

The theory can be further extended by allowing $F$ to take extended real values. Then $\{x : F(x) \in \mathbb{R}\}$ is an interval $I$, and $F$ may be thought of as a positive Radon measure on $I$, i.e. finite on closed subintervals in the interior of $I$.

**Example.** Let $F(x) = \tan x$ for $-\pi/2 < x < \pi/2$. Set $F = -\infty$ on $(-\infty, -\pi/2]$ and $+\infty$ on $[\pi/2, \infty)$. Then $F$ is the cdf of a $\sigma$-finite Borel measure $\mu$ on $(-\pi/2, \pi/2]$ which is finite on any closed subinterval. This $\mu$ is absolutely continuous with respect to $\lambda|(-\pi/2, \pi/2)$, and has density function $g(x) = \sec^2 x$.

**XIV.13.5. Signed Measures and Nonincreasing Distribution Functions**

Using the theory of signed measures, many of the results of this section generalize to cdf’s which are not increasing. There are considerable complications in this case, but also important applications. Details are found in (1).
XIV.14. Monotone Functions

In this section, we describe the properties of monotone functions.

XIV.14.1. Definitions and Basics

XIV.14.1.1. Definition. Let $I$ be an interval. A function $f : I \to \mathbb{R}$ is nondecreasing on $I$ if $f(x) \leq f(y)$ whenever $x, y \in I$ and $x < y$, and nonincreasing on $I$ if $f(x) \geq f(y)$ whenever $x, y \in I$ and $x < y$. A function is monotone on $I$ if it is either nondecreasing or nonincreasing on $I$.

XIV.14.1.2. A nondecreasing function is sometimes called increasing, but this term is somewhat ambiguous. If $f(x) < f(y)$ whenever $x, y \in I$, $x < y$, we will say that $f$ is strictly increasing on $I$. A strictly increasing function is, of course, nondecreasing. Strictly decreasing functions are defined analogously. A function is strictly monotone on $I$ if it is strictly increasing or strictly decreasing on $I$. “Increasing” could potentially mean either “nondecreasing” or “strictly increasing,” and different authors use both meanings. We will avoid using the terms “increasing” and “decreasing” with regard to functions on intervals.

XIV.14.1.3. We will work with monotone functions on a closed bounded interval $[a, b]$ for simplicity. Analogous results hold for monotone functions on intervals which are not closed and/or not bounded, but there are some technical complications (e.g. such functions need not be bounded). We will also concentrate on the case of nondecreasing functions; the nonincreasing case is essentially identical with only a few notational changes needed.

The first result is almost obvious.

XIV.14.1.4. Proposition. Let $f$ be a monotone function on $[a, b]$. Then $f$ is bounded on $[a, b]$.

Proof: If $f$ is nondecreasing on $[a, b]$, then $f(a) \leq f(x) \leq f(b)$ for all $x \in [a, b]$. The argument for nonincreasing functions is almost identical. 

One of the most important facts about monotone functions is that they are “almost continuous”, and that the only possible discontinuities are of a very simple form.

XIV.14.1.5. Definition. Let $f$ be a function defined in a neighborhood of $c \in \mathbb{R}$, and $h > 0$. Then $f$ has a jump discontinuity of height $h$ at $c$ if $\lim_{x \to c^+} f(x)$ and $\lim_{x \to c^-} f(x)$ both exist and

$$ \left| \lim_{x \to c^+} f(x) - \lim_{x \to c^-} f(x) \right| = h $$

(One could also keep track of the sign of the jump by removing the absolute values and allowing $h < 0$.)

XIV.14.1.6. If $\lim_{x \to c^+} f(x)$ exists, we will denote it by $f(c^+)$. Similarly, if $\lim_{x \to c^-} f(x)$ exists, we will denote it by $f(c^-)$. We will extend this notation and the notion of a jump discontinuity at an endpoint of the domain of a function $f$: if $a$ is the left endpoint of the domain, set $f(a^-) = f(a)$, and at the right endpoint $b$ set $f(b^+) = f(b)$.
XIV.14.1.7. Proposition. Let \( f \) be a monotone function on an interval \( I \), and \( c \in I \). Then \( f(c+) \) and \( f(c-) \) both exist, and \( f(c) \) satisfies

\[
    f(c-) \leq f(c) \leq f(c+) \quad \text{if } f \text{ is nondecreasing on } I.
\]

\[
    f(c-) \geq f(c) \geq f(c+) \quad \text{if } f \text{ is nonincreasing on } I.
\]

If \( d \in I \) and \( c < d \), and \( f \) is nondecreasing [resp. nonincreasing], then \( f(c+) \leq f(d-) \) [resp. \( f(c+) \geq f(d-) \)].

Proof: We will suppose \( c \) is an interior point of \( I \); the case of an endpoint is similar but easier. Suppose \( f \) is nondecreasing on \( I \). Set \( u = \inf_{x>c} f(x) \). Since \( f(c) \leq f(x) \) for all \( x > c \), \( f(c) \) is a lower bound for \( \{ f(x) : x > c \} \), so \( u \) exists and \( f(c) \leq u \). We claim \( u = f(c+) \). Let \( \epsilon > 0 \). There is a \( y > c \) with \( f(y) < u + \epsilon \), by definition of \( u \). Set \( \delta = y - c \); then \( \delta > 0 \) and, if \( c < x < c + \delta \), then \( c < x < y \) and \( u \leq f(x) \leq f(y) < u + \epsilon \), so \( |f(x) - u| < \epsilon \). Similarly, if \( v = \sup_{x<c} f(x) \), then \( v = f(c-) \) and \( v \leq f(c) \). If \( d > c \), choose \( x \) with \( c < x < d \); then \( f(c+) \leq f(x) \leq f(d-) \). The proof for nonincreasing \( f \) is almost identical.

It follows that if \( f \) is monotone on \( I \) and \( c \in I \), then either \( f \) is continuous at \( c \) (if \( f(c-) = f(c+) \)), or \( f \) has a jump discontinuity at \( c \) (if \( f(c-) \neq f(c+) \)).

XIV.14.1.8. Proposition. Let \( f \) be monotone on a closed interval \( [a, b] \), with jump discontinuities at \( x_1, \ldots, x_n \in [a, b] \). Let \( h_j \) be the height of the jump discontinuity of \( f \) at \( x_j \). If \( M = |f(b) - f(a)| \), then \( \sum_{j=1}^{n} h_j \leq M \).

Proof: Suppose \( f \) is nondecreasing. We may assume

\[
    a \leq x_1 < x_2 < \cdots < x_n \leq b
\]

so we have

\[
    f(a) \leq f(x_1-) < f(x_1+) \leq f(x_2-) < f(x_2+) \leq \cdots \leq f(x_n-) < f(x_n+) \leq f(b)
\]

Thus,

\[
    h_1 + \cdots + h_n = [f(x_1+) - f(x_1-)] + \cdots + [f(x_n+) - f(x_n-)]
\]

\[
    \leq [f(x_1+) - f(a)] + [f(x_2+) - f(x_1+)] + \cdots + [f(b) - f(x_{n-1}+)] = M
\]

The case of nonincreasing \( f \) is the same with some inequalities reversed.

XIV.14.1.9. Corollary. Let \( f \) be monotone on \( [a, b] \), and \( \epsilon > 0 \). If \( M = |f(b) - f(a)| \), then \( f \) has no more than \( M/\epsilon \) jump discontinuities on \([a, b]\) of height \( \geq \epsilon \).

We summarize these results in a theorem.
XIV.14.1.10. THEOREM. Let \( f \) be monotone on an interval \( I \). Then \( f \) has only countably many discontinuities on \( I \), each of which is a jump discontinuity.

**Proof:** We have proved everything except the countability of the set of discontinuities. But if \([a, b] \subseteq I\), and \( M = |f(b) - f(a)|\), then for each \( n \in \mathbb{N} \), \( f \) has only finitely many jump discontinuities of height \( \geq \frac{1}{n} \) (no more than \( Mn \)) on \([a, b]\), and hence only countably many jump discontinuities in all. Since \( I \) is a countable union of closed bounded subintervals, \( f \) can only have countably many jump discontinuities on all of \( I \). \( \square \)

XIV.14.1.11. COROLLARY. Let \( f \) be a monotone function on an interval \( I \). Then there is a monotone function \( g \) on \( I \) which is right continuous on \( I \), and which agrees with \( f \) except at countably many points.

**Proof:** For \( x \in I \), set \( g(x) = f(x+) \). Wherever \( f \) is continuous, \( f(x) = f(x+) = g(x) \). \( \square \)

The same result holds if \( g \) is to be left continuous: set \( g(x) = f(x-) \) for all \( x \).

XIV.14.2. Differentiability of Monotone Functions

A much deeper and more remarkable result is true about monotone functions: they are not only continuous almost everywhere, but even differentiable almost everywhere. This result is due to Lebesgue.

XIV.14.2.12. THEOREM. Let \( f \) be monotone on an interval \( I \). Then \( f \) is differentiable almost everywhere (with respect to Lebesgue measure).

However, the set where a monotone function is nondifferentiable is not necessarily countable. For example, the Cantor function \( () \) is not differentiable at the points of \( K \), which is uncountable (but has measure 0).

XIV.14.2.13. THEOREM. Let \( f \) be a nondecreasing function on \([a, b]\). Then \( f' \) is integrable on \([a, b]\) and

\[
\int_a^b f' \, d\lambda \leq f(b) - f(a) .
\]

**Proof:** By XIV.14.2.12., \( f' \) is defined almost everywhere on \([a, b]\). Extend \( f \) to \([a, b + 1]\) by setting \( f(x) = f(b) \) for \( b \leq x \leq b + 1 \). Define

\[
g_n(x) = n \left[ f \left( x + \frac{1}{n} \right) - f(x) \right]
\]

for \( n \in \mathbb{N} \). Then \( g_n \to f' \) pointwise a.e., so \( f' \) is measurable. The \( g_n \) and \( f' \) are nonnegative, so by Fatou’s Lemma we have

\[
\int_a^b f' \, d\lambda \leq \liminf_n \int_a^b g_n \, d\lambda = \liminf_n \int_a^b \left[ f \left( x + \frac{1}{n} \right) - f(x) \right] \, d\lambda(x)
\]

\[
= \liminf_n \left[ n \int_{a + \frac{1}{n}}^{b + \frac{1}{n}} f \, d\lambda - n \int_a^b f \, d\lambda \right] = \liminf_n \left[ n \int_{b + \frac{1}{n}}^{b + \frac{1}{n}} f \, d\lambda - n \int_a^{a + \frac{1}{n}} f \, d\lambda \right] .
\]
Since $f$ is constant on $[b, b + \frac{1}{n}]$, the first integral in the last expression is $\frac{f(b)}{n}$; and $f(x) \geq f(a)$ for $a \leq x \leq a + \frac{1}{n}$, so the second integral is at least $\frac{f(a)}{n}$. Thus we have

$$\int_a^b f' d\lambda \leq f(b) - \limsup_n n \int_a^{a + \frac{1}{n}} f d\lambda \leq f(b) - f(a).$$

\[
\]

XIV.14.2.14. We do not have equality in XIV.14.2.13. in general. If $f$ is the Cantor function, then $f$ is nondecreasing on $[0, 1]$, but $f' = 0$ a.e., so we have

$$0 = \int_0^1 f' d\lambda < f(1) - f(0) = 1.$$

We do have equality for absolutely continuous functions (and only for absolutely continuous functions).

XIV.14.3. Fubini’s Series Theorem

We have the following interesting and useful result of Fubini. We call this “Fubini’s Series Theorem” (not a standard name) since the name “Fubini’s Theorem” is usually reserved for his better-known result on iterated integrals.

XIV.14.3.1. Theorem. [Fubini’s Series Theorem] Let $(f_k)$ be a sequence of nondecreasing functions on a closed bounded interval $[a, b]$. Suppose $\sum_{k=1}^{\infty} f_k$ converges pointwise on $[a, b]$ to a real-valued function $f$. Then

(i) The series $\sum_{k=1}^{\infty} f_k$ converges to $f$ uniformly on $[a, b]$.  

(ii) The series $\sum_{k=1}^{\infty} f'_k$ converges to $f'$ pointwise a.e. on $[a, b]$. In particular, $f'_k \to 0$ pointwise a.e. on $[a, b]$. (Note that by XIV.14.2.12., $f$ and each $f_k$ are differentiable a.e. on $[a, b]$.)

Of course, there is an analogous result for nonincreasing functions.

The statement of the theorem can be rephrased in terms of convergence of sequences of functions (i.e. the partial sums of the series), but it is a little awkward to do so since the hypothesis is that the successive differences of the functions in the sequence, not the functions themselves, must be nondecreasing.

For the proof of XIV.14.3.1., we need two lemmas:

XIV.14.3.2. Lemma. To prove the theorem, it suffices to assume that each $f_k$ is nonnegative and $f_k(a) = 0$.

Proof: Let $(f_k)$ be as in the statement of the theorem. For each $k$, set $g_k(x) = f_k(x) - f_k(a)$. Then $g_k(a) = 0$, and $g_k$ is nonnegative and nondecreasing since $f_k$ is nondecreasing. We have

$$\sum_{k=1}^{n} g_k(x) = \sum_{k=1}^{n} f_k(x) - \sum_{k=1}^{n} f_k(a)$$

1668
for each \( n \). Since the first sum converges to \( f(x) \) and the second sum to \( f(a) \) as \( n \to \infty \), \( \sum_{k=1}^{\infty} g_k(x) \) converges to \( g(x) := f(x) - f(a) \) for every \( x \in [a, b] \). If the conclusion of the theorem holds for the \( g_k \), then \( \sum_{k=1}^{\infty} g_k \) converges uniformly to \( g \), so for any \( \epsilon > 0 \) there is an \( N \) such that

\[
\left| g(x) - \sum_{k=1}^{n} g_k(x) \right| < \frac{\epsilon}{2}
\]

for all \( n \geq N \) and all \( x \in [a, b] \), and

\[
\left| f(a) - \sum_{k=1}^{n} f_k(a) \right| < \frac{\epsilon}{2}
\]

for all \( n \geq N \). Then, for all \( n \geq N \) and \( x \in [a, b] \),

\[
\left| f(x) - \sum_{k=1}^{n} f_k(x) \right| = \left| (g(x) + f(a)) - \sum_{k=1}^{n} (g_k(x) + f_k(a)) \right| \leq \left| f(x) - \sum_{k=1}^{n} f_k(x) \right| + \left| f(a) - \sum_{k=1}^{n} f_k(a) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon
\]

so \( \sum_{k=1}^{\infty} f_k \) converges uniformly to \( f \) on \([a, b]\). Part (ii) for \( f \) obviously follows from part (ii) for \( g \) since \( f' = g' \) for each \( k \) and \( f' = g' \).

**XIV.14.3.3. Lemma.** Under the hypotheses of XIV.14.3.1., we have that \( \sum_{k=1}^{\infty} f_k \) converges uniformly to \( f \) and that \( f'_k \to 0 \) pointwise a.e. on \([a, b]\).

**Proof:** By XIV.14.3.2., we may assume each \( f_k \) is nonnegative and \( f_k(a) = 0 \). For each \( n \), set

\[
s_n(x) = \sum_{k=1}^{n} f_k(x)
\]

If \( \epsilon > 0 \), choose \( N \) such that

\[
0 \leq f(b) - s_n(b) = f(b) - \sum_{k=1}^{n} f_k(b) < \epsilon
\]

for all \( n \geq N \). Then, for any \( n \geq N \) and \( x \in [a, b] \), we have

\[
0 \leq f(x) - s_n(x) \leq f(b) - s_n(b) < \epsilon
\]

since \( f - s_n \) is nonnegative and nondecreasing. Thus \( s_n \to f \) uniformly on \([a, b]\).

Set

\[
E = \{ x \in [a, b] : f \text{ and each } f_k \text{ are differentiable at } x \}
\]

By XIV.14.2.12., \([a, b] \setminus E\) is a countable union of null sets, so is a null set. Each \( s_n \) is differentiable on \( E \), and since \( s_{n+1} - s_n = f_{n+1} \) and \( f - s_n \) are nondecreasing, we have \((s_{n+1} - s_n)'(x) \geq 0\) and \((f - s_n)'(x) \geq 0\) for all \( x \in E \). So, for every \( x \in E \), we have

\[
s'_1(x) \leq s'_2(x) \leq \cdots \leq f'(x)
\]

and hence \((s'_n(x))\) converges for \( x \in E \), which implies that \( f'_n(x) \to 0 \) for all \( x \in E \).

\[\Box\]
XIV.14.3.4. To complete the proof of XIV.14.3.1., we need to show that \( s_n' \to f' \) pointwise a.e. on \([a, b]\). Let \( E \) be the set in the proof of XIV.14.3.3.. Since the sequence \( (s_n'(x)) \) is nondecreasing for each \( x \in E \), it suffices to find a subsequence that converges pointwise to \( f' \) a.e. on \( E \). For each \( j \), choose an \( n_j \) such that \( f(b) - s_{n_j}(b) < 2^{-j} \), and set \( t_j = f - s_{n_j} \). Then

\[
0 \leq t_j(x) = f(x) - s_{n_j}(x) \leq f(b) - s_{n_j}(b) < 2^{-j}
\]

for all \( x \in [a, b] \), and thus the series \( \sum_{j=1}^{\infty} t_j \) satisfies the hypotheses of XIV.14.3.1.. So by XIV.14.3.3. we have \( t_j' \to 0 \) pointwise on \( E' = \{ x \in [a, b] : t \text{ and each } t_j \text{ are differentiable at } x \} \)

(\( \text{where } t = \sum_{j=1}^{\infty} t_j \)), i.e. \( s_n' \to f' \) pointwise on \( E \cap E' \). Thus \( s_n' \to f' \) pointwise on \( E \cap E' \), whose complement is a null set.

XIV.14.3.5. Note that we do not necessarily have \( E \subset E' \) since there is no assurance that \( t \) is differentiable on \( E \), and thus \( \sum_{k=1}^{\infty} f_k' \) does not necessarily converge to \( f' \) on all of \( E \). See Exercise XIV.14.4.2..

Examples

Examples. (i) A nonconstant increasing function \( f \) can have \( f' = 0 \) a.e. The Cantor function () is an example. One can even have a strictly increasing function \( g \) with \( g' = 0 \) a.e.: take \( g = \sum_{n=1}^{\infty} 2^{-n} f_n \), where the \( f_n \) are copies of the Cantor function scaled to each of the subintervals where \( f \) and all the previous rescalings of \( f \) are constant. (Expand)***

(ii) There is a strictly increasing function \( f \) on \([0, 1]\) (or any other interval), which is continuously differentiable on all of \([0, 1]\), and whose derivative vanishes on a set of positive measure. Let \( K \) be a Cantor set of positive measure (), and \( g : [0, 1] \to [0, 1] \) a continuous function with \( g^{-1}([0]) = K \) (e.g. \( g(x) = \rho(x, K) \); cf. ()). Define

\[
f(x) = \int_0^x g(t) \, dt.
\]

Then \( f' = g \) exists and is continuous on all of \([0, 1]\) (), and, for any \( a, b \in [0, 1], a < b \),

\[
f(b) - f(a) = \int_a^b g(t) \, dt > 0
\]

since \( g \) is continuous, nonnegative, and not identically 0 on \([a, b]\) (); recall that \( K \) contains no nontrivial intervals). Thus \( f \) is strictly increasing. But \( f' = g = 0 \) on \( K \). \( K \) can be taken to have measure \( 1 - \epsilon \), for any \( \epsilon > 0 \).

Note that if \( f \) is differentiable everywhere on an interval \( I \) and \( f' = 0 \) a.e. on \( I \), then \( f \) is constant on \( I \) ().

(iii) [POMPEIU] [?] There is a strictly increasing everywhere differentiable function \( f \) on a closed bounded interval (actually any interval), with bounded derivative, for which \( f' = 0 \) on a dense set. Let \( (r_k) \) be an enumeration of the rationals in \([0, 1]\), and set \( g_k(x) = 2^{-k} \sqrt{x - r_k} \),

\[
g(x) = \sum_{k=1}^{\infty} g_k(x) = \sum_{k=1}^{\infty} 2^{-k} \sqrt{x - r_k}
\]

1670
for \( x \in [0, 1] \). The series converges uniformly on \([0, 1]\), and thus \( g \) is a strictly increasing continuous function on \([0, 1]\). Let \([a, b]\) be the range of \( g \), and \( f : [a, b] \rightarrow [0, 1] \) the inverse function.

We claim that \( f \) is differentiable everywhere on \([a, b]\). Each \( g_k \) is differentiable everywhere except at \( r_k \), where its extended derivative is \(+\infty\). Thus, by the Mean Value Theorem, we have, for all \( x, y \in [0, 1], x \neq y \), we have

\[
\frac{g(y) - g(x)}{y - x} \geq \frac{g_1(y) - g_1(x)}{y - x} > \frac{1}{6}
\]

and so all Dini derivatives of \( g \) are \( \geq \frac{1}{6} \) at every point. Thus to show that \( f \) is differentiable everywhere on \([a, b]\), by the Inverse Function Theorem \((x)\) it suffices to show that \( g \) has an extended derivative everywhere on \([0, 1]\). It will then also follow that \( 0 \leq f' (x) \leq 6 \) for all \( x \in [a, b] \).

In fact, we will show that at every \( x \in [0, 1] \), the extended derivative of \( g \) at \( x \) equals \( \sum_{k=1}^{\infty} g'_k (x) \). This is clear if \( x = r_k \in \mathbb{Q} \) \([g \text{ is the sum of } g_k \text{ and another nondecreasing continuous function, so } g' (x) = g'_k (x) = +\infty]\). Suppose \( x \notin \mathbb{Q} \) but \( \sum_{k=1}^{\infty} g'_k (x) = +\infty \). Set \( h_n = \sum_{k=1}^{n} g_k \). Then for each \( M \) there is an \( n \) such that

\[
h'_n (x) = \lim_{y \to x} \frac{h_n (y) - h_n (x)}{y - x} > M
\]

so there is a \( \delta > 0 \) such that for all \( y \in [x - \delta, x + \delta] \) we have

\[
\frac{g(y) - g(x)}{y - x} \geq \frac{h_n (y) - h_n (x)}{y - x} > M
\]

and thus \( g' (x) = +\infty = \sum_{k=1}^{\infty} g'_k (x) \).

Now suppose \( x \in [0, 1] \) and \( \sum_{k=1}^{\infty} g'_k (x) < \infty \). Then \( x \notin \mathbb{Q} \). We have

\[
|s^{1/3} - t^{1/3}| \leq \frac{4}{3} s^{-2/3} |s - t|
\]

for all \( s, t \in \mathbb{R}, s \neq 0 \) (Exercise XIV.14.4.1). Setting \( s = x - r_k \) (recall that \( x \notin \mathbb{Q} \)) and \( t = y - r_k \) we obtain

\[
\left| \frac{g_k (y) - g_k (x)}{y - x} \right| \leq 4 g'_k (x)
\]

for all \( k \) and all \( y \neq x \). Thus we have, for \( y \neq x \) and any \( n \),

\[
\left| \frac{g(y) - g(x)}{y - x} - \sum_{k=1}^{\infty} g'_k (x) \right| = \sum_{k=1}^{\infty} \left( \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right)
\]

\[
\leq \sum_{k=1}^{\infty} \left| \frac{g_k (y) - g_k (x)}{y - x} \right| - \left| g'_k (x) \right| = \sum_{k=1}^{n} \left| \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right| + \sum_{k=n+1}^{\infty} \left| \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right|
\]

for every \( y \neq x \) and every \( n \). Fix \( \varepsilon > 0 \) and choose \( n \) such that \( \sum_{k=1}^{\infty} 5 g'_k (x) < \frac{\varepsilon}{2} \). There is a \( \delta > 0 \) such that if \( 0 < |y - x| < \delta \), we have

\[
\left| \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right| < \frac{\varepsilon}{2n}
\]

for \( 1 \leq k \leq n \). For \( 0 < |y - x| < \delta \), we have

\[
\left| \frac{g(y) - g(x)}{y - x} - \sum_{k=1}^{\infty} g'_k (x) \right| \leq \sum_{k=1}^{n} \left| \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right| + \sum_{k=n+1}^{\infty} \left| \frac{g_k (y) - g_k (x)}{y - x} - g'_k (x) \right|
\]

1671
\[ \leq \sum_{k=1}^{n} \left| g_k(y) - g_k(x) \right| \frac{y-x}{x} + \sum_{k=n+1}^{\infty} n \frac{g_k'(x)}{x} < n \frac{\epsilon}{2n} + \frac{\epsilon}{2} = \epsilon . \]

Thus \( g'(x) \) exists and equals \( \sum_{k=1}^{n} g_k'(x) \).

Since the extended derivative of \( g \) is \( +\infty \) at every point of \( \mathbb{Q} \cap [0,1] \), we have \( f' = 0 \) at all points of \( g(\mathbb{Q} \cap [0,1]) \) (,), which is dense in \([a,b]\). We also have \( f'(x) = 0 \) for many \( x \) not in this countable set: the set

\[ A = \{ x \in [a,b] : f'(x) = 0 \} \]

is dense in \([a,b]\), hence \( A \) must contain the set

\[ B = \{ x \in [a,b] : f' \text{ is continuous at } x \} \]

which is a residual set by (,). (However, the meager set \([a,b] \setminus A \) must be dense in \([a,b]\) and have positive measure in every subinterval by (,).)

**XIV.4. Exercises**

**XIV.4.4.1.** Show that if \( s,t \in \mathbb{R}, s \neq 0 \), then

\[ |s^{1/3} - t^{1/3}| \leq \frac{4}{3} s^{-2/3}|s - t| . \]

[Convert the problem into showing that

\[ |y - z| \leq \frac{4}{3y^2}|y^3 - z^3| \]

for \( y \neq 0 \) and use the factorization

\[ y^3 - z^3 = (y - z)(y^2 + yz + z^2) . \]

**XIV.4.4.2.** \([\text{vRS82, 4.H]}\) For each \( k \), let

\[ f_k(x) = \frac{1}{k} \tan^{-1} kx - \frac{1}{k+1} \tan^{-1}(k+1)x . \]

(a) Show that \( f_k \) is strictly increasing and differentiable on \( \mathbb{R} \).

(b) Show that \( \sum_{k=1}^{\infty} f_k(x) \) converges pointwise (in fact, uniformly) to \( f(x) = \tan^{-1} x \).

(c) Show that \( \sum_{k=1}^{\infty} f_k'(0) \) does not converge to \( f'(0) \).

**XIV.4.4.3.** Let \( f \) be a monotone function on \([a,b]\).

(a) Show that \( f \) is Borel measurable. [Show that \( f^{-1}((c, +\infty)) \) is an interval for any \( c \in \mathbb{R} \).]

(b) Show that \( f \) is Baire class 1 (cf. XII.2.6.5.)
XIV.15. Convex Functions

In this section, we define and derive the most important properties of convex functions.

XIV.15.1. Definitions

XIV.15.1.1. Definition. Let \( f \) be a function from an interval \( I \) to \( \mathbb{R} \). Then \( f \) is convex on \( I \) if for all \( x, y \in I \) and \( 0 < t < 1 \), we have

\[
f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).
\]

Geometrically, this means that the graph of \( f \) between \( x \) and \( y \) lies below the chord between \((x, f(x))\) and \((y, f(y))\), for any \( x, y \in I \) (Figure 1).

The terms concave upward and convex downward are sometimes used as synonyms for convex (although for us concave upward will mean “strictly convex,” cf. V.8.3.1.). The term convex is most common, and is motivated by the following observation:

XIV.15.1.2. Proposition. Let \( f \) be a real-valued function on an interval \( I \). Then \( f \) is a convex function if and only if its supergraph

\[
S_f = \{(x, y) : x \in I, y \geq f(x)\}
\]

is a convex subset () of \( \mathbb{R}^2 \).

XIV.15.1.3. Proposition. [Chordal Slope Criterion] Let \( I \) be an interval and \( f : I \to \mathbb{R} \). The following are equivalent:

(i) \( f \) is convex on \( I \).

(ii) For all \( x, y, z \in I \), \( x < y < z \), we have

\[
\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x}.
\]

(iii) For all \( x, y, z \in I \), \( x < y < z \), we have

\[
\frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}.
\]

(iv) For all \( x, y, z \in I \), \( x < y < z \), we have

\[
\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y}.
\]
(See Figure 2.)

Proof: (i) \(\Rightarrow\) (ii): Suppose \(f\) is convex on \(I\) and \(x, y, z \in I, x < y < z\). Set \(t = \frac{z - y}{z - x}\). Then

\[
tx + (1 - t)z = \frac{z - y}{z - x} x + \frac{y - x}{z - x} z = \frac{xz - xy + yz - xz}{z - x} = \frac{yz - xy}{z - x} = y
\]

so

\[
f(y) \leq tf(x) + (1 - t)f(z) = \frac{z - y}{z - x} f(x) + \frac{y - x}{z - x} f(z)
\]

and therefore

\[
f(y) - f(x) \leq \left[\frac{z - y}{z - x} - 1\right] f(x) + \frac{y - x}{z - x} f(z)
\]

\[
= -\frac{y - x}{z - x} f(x) + \frac{y - x}{z - x} f(z) = \frac{y - x}{z - x} [f(z) - f(x)]
\]

So we have

\[
\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x}.
\]

(ii) \(\Rightarrow\) (i): Let \(x, z \in I, x < z\), and \(0 < t < 1\), and set \(y = tx + (1 - t)z\). Then \(x < y < z\), and \(t = \frac{z - y}{z - x}\). The steps in (i) \(\Rightarrow\) (ii) can be reversed to show that

\[
f(y) \leq \frac{z - y}{z - x} f(x) + \frac{y - x}{z - x} f(z) = tf(x) + (1 - t)f(z)\,.
\]

(ii) \(\Rightarrow\) (iii): Let \(x, y, z \in I, x < y < z\). Suppose

\[
\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x}.
\]

Since

\[
\frac{f(z) - f(x)}{z - x} (z - x) = f(z) - f(x) = \frac{f(y) - f(x)}{y - x} (y - x) + \frac{f(z) - f(y)}{z - y} (z - y)
\]

\[
\leq \frac{f(z) - f(x)}{z - x} (y - x) + \frac{f(z) - f(y)}{z - y} (z - y)
\]

we have

\[
\frac{f(z) - f(x)}{z - x} (z - x) - \frac{f(z) - f(x)}{z - x} (y - x) \leq \frac{f(z) - f(y)}{z - y} (z - y)
\]

\[
\frac{f(z) - f(x)}{z - x} [(z - x) - (y - x)] = \frac{f(z) - f(x)}{z - x} (z - y) \leq \frac{f(z) - f(y)}{z - y} (z - y)
\]

and the result follows from dividing by \(z - y\).

(iii) \(\Rightarrow\) (ii) is almost identical to (ii) \(\Rightarrow\) (iii), and is left to the reader.

(iii) \(\Rightarrow\) (iv): (iv) is an immediate combination of (ii) and (iii).

(iv) \(\Rightarrow\) (ii) and (iii): Let \(x, y, z \in I, x < y < z\). Then

\[
\frac{f(z) - f(x)}{z - x} = \frac{y - x}{z - x} f(y) - f(x) + \frac{z - y}{z - x} f(z) - f(y)
\]

so \(\frac{f(z) - f(x)}{z - x}\) is a weighted average of \(\frac{f(y) - f(x)}{y - x}\) and \(\frac{f(z) - f(y)}{z - y}\) and thus lies in between. \(\diamondsuit\)
XIV.15.1.4. This result can be rephrased (cf. [vRS82, §2]). For a function \( f \) defined on an interval \( I \), define, for \( x, y \in I \),

\[
\Phi_1 f(x, y) = \frac{f(y) - f(x)}{y - x}.
\]

For \( x, y, z \in I \), all distinct, define

\[
\Phi_2 f(x, y, z) = \frac{f(z) - f(y)}{z - y} - \frac{f(y) - f(x)}{y - x}.
\]

Note that the values of \( \Phi_1 f \) and \( \Phi_2 f \) are unchanged if the variables are permuted, i.e. \( \Phi_1 f(x, y) = \Phi_1 f(y, x) \), \( \Phi_2(x, y, z) = \Phi_2(x, z, y) \), etc.

XIV.15.1.5. Corollary. Let \( f \) be a real-valued function on an interval \( I \). The following are equivalent:

(i) \( f \) is convex on \( I \).

(ii) For each \( x \in I \), the function \( \Phi_1 f(x, \cdot) \) is increasing on \( I \setminus \{x\} \).

(iii) \( \Phi_2 f(x, y, z) \geq 0 \) for all distinct \( x, y, z \in I \).

Here are some other consequences of XIV.15.1.3.:

XIV.15.1.6. Corollary. Let \( f \) be convex on \( I \), and \( w, x, y, z \in I \), \( w < x < y < z \). Then

\[
\frac{f(x) - f(w)}{x - w} \leq \frac{f(z) - f(y)}{z - y}.
\]

Proof: Compare both to \( \frac{f(y) - f(x)}{y - x} \).

XIV.15.1.7. Corollary. If \( f \) is differentiable on \( I \), and \( f' \) is nondecreasing on \( I \), then \( f \) is convex on \( I \). In particular, if \( f'' \) exists and is nonnegative on \( I \), then \( f \) is convex on \( I \).

The converse is also true, and the assumption that \( f \) is differentiable is almost automatic (XIV.15.1.6.).

Proof: Suppose \( f' \) is nondecreasing on \( I \). Let \( x, y, z \in I \), \( x < y < z \). By the Mean Value Theorem (), there are \( c, d \in I \) with \( x < c < y < d < z \) such that

\[
f'(c) = \frac{f(y) - f(x)}{y - x} \quad \text{and} \quad f'(d) = \frac{f(z) - f(y)}{z - y}.
\]

Since \( f' \) is nondecreasing, \( f'(c) \leq f'(d) \). Apply XIV.15.1.3.(iv).
XIV.15.1.8. **Examples.** (i) Any linear (affine) function is convex on \((-\infty, \infty)\).
(ii) \(f(x) = x^r\) is convex on \([0, \infty)\) if \(r = 0\) or \(r \geq 1\), convex on \((0, \infty)\) if \(r < 0\), and not convex on any interval if \(0 < r < 1\). If \(n \in \mathbb{N}\) is even, \(f(x) = x^n\) is convex on \((-\infty, \infty)\) and \(g(x) = x^{-n}\) is convex on \((-\infty, 0)\).
(iii) \(f(x) = e^x\) is convex on \((-\infty, \infty)\).
(iv) A convex function need not be differentiable everywhere: \(f(x) = |x|\) is convex on \((-\infty, \infty)\).
(v) A convex function need not even be continuous. The following function is convex on \([0, 1]\):

\[
f(x) = \begin{cases} 
0 & \text{if } 0 \leq x < 1 \\
1 & \text{if } x = 1
\end{cases}
\]

XIV.15.1.9. **Proposition.** (i) If \(f\) and \(g\) are convex on \(I\), then \(f + g\) is convex on \(I\).
(ii) If \(f\) is convex on \(I\) and \(c \geq 0\), then \(cf\) is convex on \(I\).
(iii) If \(I\) and \(J\) are intervals, \(f\) is convex on \(I\), \(f(I) \subseteq J\), and \(g\) is nondecreasing and convex on \(J\), then \(g \circ f\) is convex on \(I\).

**Proof:** (i) and (ii) are straightforward, and are left to the reader. For (iii), if \(x, y \in I\) and \(0 < t < 1\), then

\[
g(f(tx + (1 - t)y)) \leq g(tf(x) + (1 - t)f(y)) \leq tg(f(x)) + (1 - t)g(f(y)).
\]

XIV.15.1.10. **Note,** however, that a product of convex functions is not convex in general, even if both are nonnegative and one is linear. For example \(f(x) = x^2\) and \(g(x) = 1 - x\) are convex and nonnegative on \([0, 1]\), but \(fg\) is not convex on \([0, 1]\). (A product of two nonnegative, nondecreasing, convex functions is convex (XIV.15.1.24.).) Also note that (iii) fails in general if \(g\) is not nondecreasing, e.g. if \(g(x) = -x\).

**Limits and Suprema of Convex Functions**

XIV.15.1.11. **Proposition.** Let \(I\) be an interval, and \((f_n)\) a sequence of convex functions on \(I\) converging pointwise on \(I\) to a function \(f\). Then \(f\) is convex on \(I\).

**Proof:** If \(x, y \in I\) and \(0 < t < 1\), then \(f_n(tx + (1 - t)y) \rightarrow f(tx + (1 - t)y)\) and \(tf_n(x) + (1 - t)f_n(y) \rightarrow tf(x) + (1 - t)f(y)\). It follows immediately that \(f\) is convex. 

In fact, such a convergent sequence of convex functions almost converges uniformly (XIV.15.3.1.).

XIV.15.1.12. **Proposition.** Let \(\{f_i\}\) be a set of convex functions on an interval \(I\). If \(f = \sup_i f_i\) is finite on \(I\), then \(f\) is convex on \(I\). In particular, if \(f_1, \ldots, f_n\) are convex on \(I\), then \(f_1 \lor \cdots \lor f_n = \max(f_1, \ldots, f_n)\) is convex on \(I\).

**Proof:** This follows immediately from XIV.15.1.2. and (.), since \(S_f = \cap_i S_{f_i}\). 

1676
Continuity of Convex Functions

The discontinuity in XIV.15.1.8.(v) can only happen at an endpoint. In fact, we have:

**XIV.15.1.13.** Proposition. Let \( f \) be convex on \( I \), and \([a,b]\) a closed bounded interval contained in the interior of \( I \). Then \( f \) is Lipschitz on \([a,b]\).

**Proof:** Let \( c,d \in I \), \( c < a < b < d \). If \( x,y \in [a,b] \), then by XIV.15.1.6. we have

\[
\frac{f(a) - f(c)}{a - c} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(d) - f(b)}{d - b}
\]

and so \( |f(y) - f(x)| \leq M|y - x| \), where

\[
M = \max \left( \left| \frac{f(a) - f(c)}{a - c} \right|, \left| \frac{f(d) - f(b)}{d - b} \right| \right).
\]

**XIV.15.1.14.** Proposition. Let \( f \) be convex on \( I \). Then \( f \) is continuous on the interior of \( I \). If \( I \) has a left endpoint \( a \) [resp. right endpoint \( b \)], then \( \lim_{x \to a^+} f(x) \) exists [resp. \( \lim_{x \to b^-} f(x) \) exists], and \( \lim_{x \to a^+} f(x) \leq f(a) \) [resp. \( \lim_{x \to b^-} f(x) \leq f(b) \)]. In particular, \( f \) is lower semicontinuous at endpoints of \( I \).

**Proof:** Continuity of \( f \) on the interior of \( I \) follows immediately from XIV.15.1.13., since a Lipschitz function is (uniformly) continuous.

Suppose \( I \) has a left endpoint \( a \). Let \( c \in I \), \( c > a \). Then, if \( a < x < c \), we have by XIV.15.1.3.(iii),

\[
g(x) := \frac{f(c) - f(x)}{c - x} \geq \frac{f(c) - f(a)}{c - a}
\]

and \( g(x) \) decreases as \( x \) decreases; thus

\[
L := \lim_{x \to a^+} \frac{f(c) - f(x)}{c - x} = \inf_{a < x < c} \frac{f(c) - f(x)}{c - x}
\]

exists, and satisfies

\[
L \geq \frac{f(c) - f(a)}{c - a}.
\]

Since \( \lim_{x \to a^+} (c - x) = c - a \) exists and is nonzero, \( m := \lim_{x \to a^+} f(x) \) also exists and satisfies

\[
L = \frac{f(c) - m}{c - a} \geq \frac{f(c) - f(a)}{c - a}
\]

and so \( m = \lim_{x \to a^+} f(x) \leq f(a) \). The argument for a right endpoint \( b \) of \( I \) is virtually identical. ☑
Differentiability of Convex Functions

An almost identical argument shows that the left and right derivatives of a convex function exist at every point:

**XIV.15.1.15. Theorem.** Let $f$ be a convex function on an interval $I$. Then the left and right derivatives $D_Lf$ and $D_Rf$ of $f$ exist at every point of $I$ (except a left or right endpoint respectively), and are finite except possibly at endpoints of $I$. If $x, y \in I$, $x < y$, then

$$D_Lf(x) \leq D_Rf(x) \leq D_Lf(y) \leq D_Rf(y).$$

$D_Lf$ is continuous from the left wherever it is finite (i.e. except possibly at a right endpoint of $I$), and $D_Rf$ is continuous from the right wherever it is finite (i.e. except possibly at a left endpoint of $I$).

In fact, if $I$ has a right endpoint $b$, and $f$ is continuous at $b$, and $D_Lf(b) = +\infty$, then $\lim_{x \to b^-} D_Lf(x) = +\infty$ (cf. ?), i.e. $D_Lf$ is “continuous” at $b$. A similar statement holds for $D_Rf$ at a left endpoint of $I$.

**Proof:** Let $x \in I$, not a right endpoint. Then there is an $\epsilon > 0$ such that $x + \epsilon \in I$. Then

$$g(h) := \frac{f(x + h) - f(x)}{h}$$

exists for $0 < h < \epsilon$. By XIV.15.1.3.(ii), $g(h)$ decreases as $h$ decreases, so

$$D_Rf(x) = \lim_{h \to 0^+} \frac{f(x + h) - f(x)}{h} = \inf_{0 < h < \epsilon} \frac{f(x + h) - f(x)}{h}$$

exists as an extended real number and is not $+\infty$. If $x$ is not a left endpoint, let $a \in I$, $a < x$; then by XIV.15.1.3.(iv) we have

$$\frac{f(x + h) - f(x)}{h} \geq \frac{f(x) - f(a)}{x - a}$$

for $0 < h < \epsilon$, so $D_Rf(x)$ is finite (i.e. not $-\infty$). A similar argument shows that if $x$ is not a left endpoint, then $D_Lf(x)$ exists and is not $-\infty$; and if $x$ is not a right endpoint, $D_Lf(x)$ is finite (not $+\infty$). If $x, y \in I$, $x < y$, then

$$\frac{f(x + h) - f(x)}{h} \leq \frac{f(y) - f(y - h)}{h}$$

for $0 < h < (y - x)/2$ by XIV.15.1.6., and hence $D_Rf(x) \leq D_Lf(y)$.

To prove left-continuity of $D_Lf$, suppose $x$ is not a left endpoint of $I$ and $D_Lf(x)$ is finite. It follows that $f$ is continuous at $x$. Let $\epsilon > 0$. Then there is a $y \in I$, $y < x$, such that

$$D_Lf(x) - \epsilon < \frac{f(y) - f(x)}{y - x} \leq D_Lf(x).$$

If $(x_n)$ is a sequence in $(y, x)$ converging to $x$, we have $f(x_n) \to f(x)$, so for sufficiently large $n$ we have

$$D_Lf(x) - \epsilon < \frac{f(y) - f(x_n)}{y - x_n} \leq D_Lf(x_n) \leq D_Lf(x).$$

Right-continuity of $D_Rf$ is almost identical. ☞

1678
XIV.15.1.16. **Corollary.** Let \( f \) be convex on \( I \). Then \( f \) is differentiable on \( I \) except on a countable set (which may be finite or empty), \( f' \) is continuous on the set on which it is defined, and \( f' \) is nondecreasing on \( I \).

**Proof:** It is immediate from XIV.15.1.15. that \( D_L f \) is nondecreasing on \( I \), and hence is continuous except at countably many points \( \). If \( x \in I \), then \( D_L f(x) \leq D_R f(x) \leq D_L f(y) \) for all \( y > x \), and hence \( D_R f(x) = D_L f(x) \) if \( D_L f \) is continuous from the right at \( x \) (and only at such points). Thus \( f \) is differentiable at any point where \( D_L f \) is continuous, hence at all but countably many points. 

We then get the following “geometrically obvious” consequence:

XIV.15.1.17. **Corollary.** Let \( f \) be convex and continuous on an interval \( I \). Then one of the following holds:

(i) \( D_L f(x) \geq 0 \) for all \( x \in I \).

(ii) \( D_L f(x) \leq 0 \) for all \( x \in I \).

(iii) There are unique \( c, d \in I \), \( c \leq d \), with \( D_L f(x) < 0 \) for all \( x \in I \), \( x < c \), \( D_L f(x) > 0 \) for all \( x \in I \), \( x > d \), and, if \( c < d \), \( D_L f(x) = 0 \) for \( c < x \leq d \).

Thus either \( f \) is nondecreasing on \( I \), \( f \) is nonincreasing on \( I \), or there are unique \( c, d \in I \), \( c \leq d \), such that \( f \) is strictly decreasing on \( I \cap (-\infty, c] \), strictly increasing on \( I \cap [d, +\infty) \), and, if \( c < d \), constant on \( [c, d] \).

In fact, if \( D_L f \) takes both positive and negative values on \( I \), then \( c = \sup\{x \in I : D_L f(x) < 0\} \) and \( d = \inf\{x \in I : D_L f(x) > 0\} \). Note that we often have \( c = d \); this will happen unless \( f \) is constant on a subinterval.

XIV.15.1.18. **Corollary.** Let \( f \) be convex on an interval \( I \). Then for each \( c \) in the interior of \( I \) there is a linear function \( L \) on \( I \) with \( L(c) = f(c) \) and \( L(x) \leq f(x) \) for all \( x \in I \).

**Proof:** We can take for \( L \) the linear function with \( L(c) = f(c) \) and with slope \( m \), for any \( m \) satisfying

\[
D_L f(c) \leq m \leq D_R f(c).
\]

XIV.15.1.19. Conversely, if \( f \) is a continuous function on an interval \( I \) with the property that for each \( c \) in the interior of \( I \), there is a linear function \( L_c \) on \( I \) with \( L_c(c) = f(c) \) and \( L_c(x) \leq f(x) \) for all \( x \in I \), then \( f = \sup_c(L_c) \), and each \( L_c \) is convex on \( I \), so \( f \) is convex on \( I \) by XIV.15.1.12..

Let \( I \) be an interval, and \( g : I \to \mathbb{R} \) a nondecreasing function. Then \( g \) is Riemann integrable over every closed bounded subset of \( I \). Fix \( c \in I \), and for \( x \in I \) define

\[
f(x) = \int_c^x g(t) \, dt.
\]

1679
XIV.15.1.20. Proposition. $f$ is convex on $I$.

Proof: Let $x, y, z \in I$, $x < y < z$. Then

$$\frac{f(y) - f(x)}{y - x} = \frac{1}{y - x} \int_x^y g(t) \, dt$$

is the average value of $g$ on $[x, y]$, and similarly $\frac{f(z) - f(y)}{z - y}$ is the average value of $g$ on $[y, z]$. Since $g$ is nondecreasing,

$$\frac{f(y) - f(x)}{y - x} \leq g(y) \leq \frac{f(z) - f(y)}{z - y}.$$

Apply XIV.15.1.3.

XIV.15.1.21. We will show that the converse is true, i.e. that every convex function on an open interval is the primitive of a nondecreasing function. This follows immediately from XIV.15.1.13. along with () and (), using the theory of Lebesgue integration and absolute continuity. However, it is not necessary to use Lebesgue integration since bounded nondecreasing functions are Riemann integrable on closed bounded intervals:

XIV.15.1.22. Theorem. Let $f$ be convex on an open interval $I$. Then $f'$ is Riemann integrable over every closed bounded subinterval of $I$. If $a \in I$ is fixed, then, for every $x \in I$,

$$f(x) = f(a) + \int_a^x f'(t) \, dt.$$

Proof: This is a special case of (). We have that $D_L f$ is increasing on $I$, hence Riemann integrable over every closed bounded subinterval of $I$, and equal to $f'$ except at countably many points of $I$.

XIV.15.1.23. Corollary. Let $f$ be a function on an open interval $I$. Then $f$ is convex on $I$ if and only if $f$ is absolutely continuous on every closed bounded subinterval of $I$ and $f'$ is increasing on a subset $A$ of $I$ with $\lambda(I \setminus A) = 0$.

XIV.15.1.24. Corollary. Let $f$ and $g$ be nonnegative, nondecreasing, and convex on an open interval $I$. Then $fg$ is convex on $I$.

Proof: Let $[a, b] \subseteq I$. Then $f$ and $g$ are bounded and Lipschitz on $[a, b]$ (XIV.15.1.13.), and hence $fg$ is also Lipschitz and thus absolutely continuous on $[a, b]$. Let $A$ be the set of points where $f$ and $g$ are both differentiable; then $I \setminus A$ is countable. By the product rule, if $x \in A$, then $fg$ is differentiable at $x$ and

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$
Since $f$, $g$, $f'$, and $g'$ are all nonnegative and nondecreasing on $A$, $(fg)'$ is nondecreasing on $A$.

*** Is there an elementary proof of this? ***

Here is a generalization of the last sentence of XIV.15.1.7. (cf. [?, p. 39]; [HS75, 17.37(c)]). Recall (1) that

$$
\mathcal{D}^{(2)}_S(f)(x) = \limsup_{h \to 0} \frac{f(x + h) + f(x - h) - 2f(x)}{h^2}
$$

$$
\mathcal{D}^{(2)}_S(f)(x) = \liminf_{h \to 0} \frac{f(x + h) + f(x - h) - 2f(x)}{h^2}
$$

XIV.15.1.25.\ Prop. Let $f$ be a function on an open interval $I$. Then the following are equivalent:

(i) $f$ is convex on $I$.

(ii) $f$ is continuous on $I$ and $\mathcal{D}^{(2)}_S(f)(x) \geq 0$ for all $x \in I$.

(iii) $f$ is continuous on $I$ and $\mathcal{D}^{(2)}_S(f)(x) \geq 0$ for all $x \in I$.

Proof: (i) $\Rightarrow$ (iii): If $f$ is convex on $I$ and $x \in I$, then there is an $\epsilon > 0$ such that $(x - \epsilon, x + \epsilon) \subseteq I$. Then, if $|h| < \epsilon$,

$$f(x) \leq \frac{1}{2}f(x - h) + \frac{1}{2}f(x + h)$$

and hence $f(x + h) + f(x - h) - 2f(x) \geq 0$. Thus $\mathcal{D}^{(2)}_S(f)(x) \geq 0$. And $f$ is continuous on $I$ by XIV.15.1.14.

(iii) $\Rightarrow$ (ii) is trivial.

(ii) $\Rightarrow$ (i): Suppose $f$ is continuous and $\mathcal{D}^{(2)}_S(f) \geq 0$ on $I$, and let $[a, b] \subseteq I$, and $\epsilon > 0$. Set

$$\phi(x) = f(x) - \left[ f(a) + \frac{f(b) - f(a)}{b - a} (x - a) \right] + \epsilon(x - a)(x - b)$$

for $a \leq x \leq b$. Then $\phi$ is continuous on $[a, b]$, $\phi(a) = \phi(b) = 0$, and $\mathcal{D}^{(2)}_S(\phi) \geq 2\epsilon$ on $[a, b]$. If $\max_{[a, b]} \phi > 0$, then $\phi$ has a local maximum at some $x_0 \in (a, b)$, and we would have $\mathcal{D}^{(2)}_S(\phi)(x_0) \leq 0$, a contradiction. Thus $\phi(x) \leq 0$ on $[a, b],

$$f(x) \leq f(a) + \frac{f(b) - f(a)}{b - a} (x - a) - \epsilon(x - a)(x - b)$$

for all $x \in [a, b]$. Since $\epsilon > 0$ is arbitrary, we have

$$f(x) \leq f(a) + \frac{f(b) - f(a)}{b - a} (x - a)$$

for all $x \in [a, b]$. Since $[a, b]$ is arbitrary, $f$ is convex on $I$. \hfill \Box
XIV.15.2. Midpoint-Convex Functions

There is a weaker condition which almost implies convexity:

XIV.15.2.1. Definition. A real-valued function $f$ on an interval $I$ is midpoint-convex if, for every $x,y \in I$, we have

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}.$$ 

A convex function is midpoint-convex. The converse is false (assuming the Axiom of Choice – Exercise XIV.15.7.2.), but is “almost true” (i.e. true under very mild additional hypotheses), as the results of this subsection show. These results are quite useful since midpoint-convexity is often much easier to check than convexity.

This exposition is adapted from [...].

XIV.15.2.2. Proposition. Let $f$ be a midpoint-convex function on an interval $I$. Then, for every $x,y \in I$ and dyadic rational $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Proof: We prove this for $t$ of the form $\frac{p}{2^n}$ by induction on $n$. The case $n = 1$ holds by definition. Suppose it is true for $n$, and let $t = \frac{p}{2^{n+1}}$. Since the condition is symmetric in $t$ and $1-t$, we may assume $p < 2^n$, so $q := 2^{n+1} - p > 2^n$. Set $r = q - 2^n$, so $p + r = 2^n$. Then

$$z := \frac{p}{2^{n+1}}x + \frac{q}{2^{n+1}}y = \frac{1}{2} \left[ \frac{p}{2^n}x + \frac{r}{2^n}y + y \right]$$

so we have

$$f(z) \leq \frac{1}{2} \left[ f\left( \frac{p}{2^n}x + \frac{r}{2^n}y \right) + f(y) \right] \leq \frac{1}{2} \left[ \frac{p}{2^n}f(x) + \frac{r}{2^n}f(y) + f(y) \right] = \frac{p}{2^{n+1}}f(x) + \frac{q}{2^{n+1}}f(y).$$

XIV.15.2.3. Corollary. If $f$ is continuous and midpoint-convex on an interval $I$, then $f$ is convex on $I$.

Proof: For fixed $x,y \in I$, the inequality

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

for arbitrary $t \in [0,1]$ follows from the inequality for dyadic rational $t$ by continuity and density of the dyadic rationals.

We can say much more. A midpoint-convex function which is not convex must be not only discontinuous, but rather bizarre:
XIV.15.2.4. Theorem. Let \( f \) be a midpoint-convex function on an open interval \( I \). If \( f \) is not convex on \( I \), then \( f \) is not bounded above in any subinterval. If \( c \in I \), then there is a sequence \( (x_n) \) in \( I \) with \( x_n \to c \) and \( \lim_{n \to \infty} f(x_n) = +\infty \). In particular, \( f \) is discontinuous everywhere on \( I \).

Proof: By XIV.15.2.3., \( f \) is discontinuous at some \( x_0 \in I \). By translation and adding a constant, we may assume \( x_0 = 0 \) and that \( f(0) = 0 \) for ease of notation. There is a sequence \( (y_n) \) in \( I \) with \( y_n \to 0 \) and \( f(y_n) \to m \neq 0 \). If \( f(y_n) < 0 \), we have that \( f(-y_n) \geq -f(y_n) > 0 \) since \( 0 = f(0) \leq \frac{1}{2} (f(y_n) + f(-y_n)) \). Thus, by replacing \( (y_n) \) by a subsequence of \( (-y_n) \) if necessary, we may assume \( m > 0 \). If \( m = +\infty \), we are done. Otherwise, we have

\[
2f(y_n) \leq f(0) + f(2y_n) = f(2y_n)
\]

and so \( \liminf_{n \to \infty} f(2y_n) \geq 2m \). Continuing inductively, for each \( k \) we have

\[
f(2^k y_n) \geq 2^k f(y_n)
\]

for all \( n \) large enough that \( 2^k y_n \in I \), and hence \( \liminf_{n \to \infty} f(2^k y_n) \geq 2^k m \). Since for fixed \( k \) we have \( 2^k y_n \to 0 \), we have that \( f \) is unbounded above in every neighborhood of 0, and there is a sequence \( (z_n) \in I \) with \( z_n \to 0 \) and \( f(z_n) \to +\infty \).

Now let \( a \) be the largest number (or \(+\infty\)) for which \((-a, a) \subseteq I \), and let \( c \in (-a, a) \). We have that \( x_n := c + z_n \to c \), so \( x_n \in I \) for sufficiently large \( n \). Also,

\[
f(z_n) = f \left( c + \frac{2z_n - c}{2} \right) \leq \frac{f(c + 2z_n) + f(-c)}{2} = \frac{f(x_n) + f(-c)}{2}
\]

and, since \( f(z_n) \to +\infty \), it follows that \( f(x_n) \to +\infty \). Thus \( f \) is unbounded above in every subinterval of \((-a, a)\) and discontinuous everywhere on \((-a, a)\).

We end up with an open subinterval \( J \) of \( I \), extending to at least one end of \( I \), such that \( f \) is unbounded above on every subinterval of \( J \) and hence discontinuous everywhere on \( J \). If \( J \neq I \), repeat the argument with an \( x_0 \in J \) near the endpoint of \( J \) not exhausting \( I \) to get another such interval \( J_2 \) centered at \( x_0 \). If \( J \cup J_2 \neq I \), then \( J \subseteq J_2 \), and \( J_2 \) is approximately twice the length of \( J \), with the same properties. Repeating the process, in finitely many steps we cover any bounded subinterval of \( I \), which suffices to prove the result. 

XIV.15.2.5. A midpoint-convex function which is not convex can be bounded below: in fact, if \( f \) is midpoint-convex, so is \( \max(f, 0) \).

XIV.15.2.6. Corollary. A midpoint-convex function on an interval \( I \) which is bounded on at least one subinterval of \( I \) is convex.

This result considerably extends XIV.15.2.3.

We can extend XIV.15.2.4. even further. A midpoint-convex function which is not convex cannot even be bounded on a set of positive Lebesgue measure. The next result is due to Ostrowski [?].
XIV.15.2.7. Theorem. Let \( f \) be midpoint-convex on an interval \( I \). If there is a Lebesgue-measurable subset \( E \) of \( I \) with \( \lambda(E) > 0 \) on which \( f \) is bounded, then \( f \) is convex on \( I \).

Proof: Assume \( f \) is not convex on \( I \). By replacing \( I \) by \( I \cap (-n,n) \) and \( E \) by \( E \cap (-n,n) \) for some sufficiently large \( n \), we may assume \( I = (a,b) \) is bounded. There is an open set \( U \subseteq (a,b) \) with \( E \subseteq U \) and \( \lambda(U) < \frac{4}{3}\lambda(E) \). \( U \) is a disjoint union of finitely or countably many open intervals \( J_n \), and we have

\[
\lambda(E) = \sum_n \lambda(E \cap J_n) \leq \sum_n \lambda(J_n) < \frac{4}{3} \sum_n \lambda(E \cap J_n)
\]

so there is an \( n \) such that

\[
0 < \lambda(E \cap J_n) < \frac{4}{3} \lambda(E \cap J_n)
\]

and replacing \( I \) by \( J_n \) and \( E \) by \( E \cap J_n \) we may assume \( \lambda(I) = \frac{4}{3}\lambda(E) \).

Suppose \( f(x) \leq M \) for all \( x \in E \). The set

\[
S = \{ x \in I : f(x) > M \}
\]

is dense in \( I \) by XIV.15.2.4., so by contracting \( I \) slightly we may assume the midpoint of \( I \) is in \( S \). By translating, we may assume \( I = (-c,c) \) and that \( 0 \in S \), to simplify notation. Set

\[
\tilde{S} = \{ -x : x \in S \}.
\]

For any \( x \in I \), we have

\[
M < f(0) \leq \frac{f(x) + f(-x)}{2}
\]

so either \( x \) or \( -x \) is in \( S \), i.e. \( S \cup \tilde{S} = I \). The set \( S \) is not obviously Lebesgue measurable, but in any case we have \( \lambda^*(S) = \lambda^*(\tilde{S}) \), where \( \lambda^* \) denotes Lebesgue outer measure (). We have \( E \cap S = \emptyset \), so \( E \subseteq \tilde{S} \) and \( S \subseteq (I \setminus E) \). Then

\[
\lambda(E) \leq \lambda^*(S) = \lambda^*(\tilde{S}) \leq \lambda(I \setminus E) = \lambda(I) - \lambda(E)
\]

\[
2\lambda(E) \leq \lambda(I) < \frac{4}{3}\lambda(E)
\]

which is a contradiction (to the assumption that \( f \) is not convex on \( I \)).

We then obtain the following corollary, due to Sierpiński:

XIV.15.2.8. Corollary. Let \( f \) be a midpoint-convex function on an interval \( I \). If \( f \) is a Lebesgue measurable function, then \( f \) is convex on \( I \).

Proof: We have that \( E_n = f^{-1}([-n,n]) \) is a Lebesgue measurable subset of \( I \) for all \( n \), and \( \cup_n E_n = I \), so \( \sup_n \lambda(E_n) = \lambda(I) > 0 \), and \( f \) is bounded on each \( E_n \).

Summarizing, and combining with XIV.15.1.13., we have:
XIV.15.2.9. Theorem. Let \( f \) be a midpoint-convex function on an interval \( I \), and \([a, b]\) a closed bounded interval in the interior of \( I \). The following are equivalent:

(i) \( f \) is convex on \([a, b]\).
(ii) \( f \) is Lipschitz on \([a, b]\).
(iii) \( f \) is continuous on \([a, b]\).
(iv) \( f \) is bounded on \([a, b]\).
(v) \( f \) is Lebesgue measurable on \([a, b]\).

XIV.15.3. Convergence of Sequences of Convex Functions

Pointwise convergence of a sequence of convex functions implies several much stronger types of convergence. We describe some of the rather remarkable consequences of pointwise convergence in this subsection.

XIV.15.3.1. Theorem. Let \( I \) be an interval, and \((f_n)\) a sequence of convex functions on \( I \) converging pointwise on \( I \) to a function \( f \). Then \( f \) is convex on \( I \), and the convergence is uniform on any closed bounded interval contained in \( I \) on which \( f \) is continuous (e.g. any closed bounded interval contained in the interior of \( I \)).

Proof: It has been noted in XIV.15.1.11. that \( f \) is convex.

If \([a, b]\) is a closed bounded interval in \( I \), suppose the convergence is not uniform on \([a, b]\). Then there is an \( \eta > 0 \) such that, for infinitely many \( n \), there is a \( c_n \in [a, b] \) with \( |f_n(c_n) - f(c_n)| \geq \eta \). Passing to a subsequence, we may assume we have such a \( c_n \) for all \( n \).

The sequence \((c_n)\) cannot have any eventually constant subsequences, which would contradict the pointwise convergence \( f_n \to f \). Thus \((c_n)\) has a strictly monotone subsequence. Passing to this subsequence, we may assume \((c_n)\) is strictly monotone. In the rest of the argument, we will assume \((c_n)\) is strictly increasing; the strictly decreasing case is essentially identical, interchanging the roles of \( a \) and \( b \).

We have \( c_n \to c \) for some \( c, a < c < b \). There are two cases to consider:

Case 1. \( f_n(c_n) \geq f(c_n) + \eta \) for infinitely many \( n \). Again passing to a subsequence, in this case we may assume \( f_n(c_n) \geq f(c_n) + \eta \) for all \( n \). Fix \( \epsilon > 0 \). Choose \( n_0 \) so that, for all \( n \geq n_0 \), we have \( c_n > c - \epsilon \), \( |f_n(c) - f(c)| < \frac{\eta}{2} \), and \( |f_n(a) - f(a)| < \frac{\eta}{2} \). Then, for \( n \geq n_0 \), we have

\[
\frac{f(c) - f(a) - \eta}{c - a} < \frac{f_n(c) - f_n(a)}{c - a} \leq \frac{f_n(c) - f_n(c_n)}{c - c_n} < -\frac{\eta/2}{\epsilon}.
\]

This is a contradiction if \( f(c) - f(a) - \eta \geq 0 \), or if \( f(c) - f(a) - \eta < 0 \) and \( \epsilon \) is sufficiently small, specifically if

\[
\epsilon < \frac{\eta(c - a)}{2[\eta + f(a) - f(c)]}.
\]
Case 2. \( f_n(c_n) \leq f(c_n) - \eta \) for infinitely many \( n \); we may again assume this holds for all \( n \). Fix \( z \) with \( a < z < c \) and let \( \lambda > 0 \) with

\[
-\lambda < \frac{f(z) - f(a) - \eta}{z - a}.
\]

Fix \( d \) with \( a < d < c \) such that \( d > c - \frac{\eta}{4\lambda} \) and \( |f(x) - f(c)| < \frac{\eta}{4} \) for all \( x \in [d, c] \) (this is possible since \( f \) is continuous at \( c \)). Choose \( n_0 \) such that, for all \( n \geq n_0 \), \( c_n > d \), \( |f_n(a) - f(a)| < \frac{\eta}{4} \), \( |f_n(z) - f(z)| < \frac{\eta}{4} \), and \( |f_n(d) - f(d)| < \frac{\eta}{4} \). Then, for \( n \geq n_0 \), we have \( f_n(d) > f(c) - \frac{\eta}{2} \) and \( f_n(c_n) < f(c_n) - \eta < f(c) - \frac{3\eta}{4} \). Thus \( f_n(c_n) - f_n(d) < -\frac{\eta}{4} \). We also have

\[
\frac{f_n(c_n) - f_n(d)}{c_n - d} \geq \frac{f_n(z) - f_n(a)}{z - a} \geq \frac{f(z) - f(a) - \eta}{z - a} > -\lambda.
\]

Then, since \( 0 < c_n - d < \frac{n}{4\lambda} \), we get

\[
\frac{4\lambda}{\eta} (f_n(c_n) - f_n(d)) > -\lambda
\]

which is a contradiction.

(If \( c < b \) in Case 2, a contradiction can be obtained somewhat more simply using an argument similar to the one in Case 1.)

The first statement of this result which I have found is in [Con74]; however, the proof there is incorrect. A stronger assertion, which is incorrect (see the following example), is made in [? IX, Book IV, Chapter I, §4, p. 52]; the proof outlined there only gives a slightly weaker conclusion than XIV.15.3.1. (cf. XIV.15.7.6.(a)-(b)), but can be adapted to give the full result (XIV.15.7.6.(c)-(d)).

XIV.15.3.2. **Example.** If \( f \) is not continuous on \( I \), we do not have uniform convergence on all of \( I \) in general, even if \( I \) is closed and bounded. If \( f_n(x) = x^n \), then \( f_n \) is convex on \([0, 1]\) and converges pointwise on \([0, 1]\) to the function of XIV.15.1.8.(v). But the convergence is not uniform on \([0, 1]\).

If \( I \) is not closed and bounded, simple examples show that we do not have uniform convergence on all of \( I \) in general, even if \( f \) is continuous on \( I \): if \( I \) is unbounded, consider \( f_n(x) = \frac{x}{n} \), and if \( I = (0, 1] \), consider \( f_n(x) = \frac{1}{x+\pi} \).

Here is a closely related result:

XIV.15.3.3. **Theorem.** Let \( (f_n) \) be a sequence of convex functions on an interval \( I \), converging pointwise on \( I \) to a (convex) function \( f \). Then

(i) If \( x \) is not a left endpoint of \( I \), then

\[
\liminf_{n \to \infty} D_L f_n(x) \geq D_L f(x).
\]

(ii) If \( x \) is not a right endpoint of \( I \), then

\[
\limsup_{n \to \infty} D_R f_n(x) \leq D_R f(x).
\]
In particular, if \( x \) is an interior point of \( I \) and \( f \) is differentiable at \( x \), then

\[
\lim_{n \to \infty} D_L f_n(x) = \lim_{n \to \infty} D_R f_n(x) = f'(x)
\]

(i.e. the limits both exist and equal \( f'(x) \)).

**Proof:** We prove only (i); the proof of (ii) is entirely symmetric and is left to the reader. So suppose \( x \) is not a left endpoint of \( I \), and fix \( m < D_L f(x) \) (note that \( D_L f(x) \neq -\infty \)). Then there is a \( y \in I \), \( y < x \), such that

\[
\frac{f(y) - f(x)}{y - x} > m
\]

and thus there is an \( \epsilon > 0 \) such that

\[
m' := \frac{f(y) - f(x) + 2\epsilon}{y - x} > m
\]

(note that \( y - x < 0 \)). There is an \( N \) such that \( |f_n(x) - f(x)| < \epsilon \) and \( |f_n(y) - f(y)| < \epsilon \) for all \( n \geq N \). Then, for \( n \geq N \), we have \( f_n(y) < f(y) + \epsilon \) and \( f_n(x) > f(x) - \epsilon \), and thus, for any \( z \), \( y \leq z < x \), we have, using convexity of \( f_n \),

\[
\frac{f_n(z) - f_n(x)}{z - x} \geq \frac{f_n(y) - f_n(x)}{y - x} > \frac{[f(y) + \epsilon] - [f(x) - \epsilon]}{y - x} = m'
\]

so \( D_L f_n(x) \geq m' > m \) for \( n \geq N \), and hence \( \lim \inf_{n \to \infty} D_L f_n(x) \geq m' > m \).

For the final statement, since \( D_L f_n(x) \leq D_R f_n(x) \) for all \( n \), we have

\[
f'(x) \leq \lim \inf_{n \to \infty} D_L f_n(x) \leq \lim \sup_{n \to \infty} D_L f_n(x) \leq \lim \sup_{n \to \infty} D_R f_n(x) \leq f'(x)
\]

\[
f'(x) \leq \lim \inf_{n \to \infty} D_L f_n(x) \leq \lim \inf_{n \to \infty} D_R f_n(x) \leq \lim \sup_{n \to \infty} D_R f_n(x) \leq f'(x)
\]

so all the inequalities are equalities.

XIV.15.3.4. **Corollary.** Let \((f_n)\) be a sequence of convex functions on an interval \( I \) converging pointwise on \( I \) to a (convex) function \( f \). Then \( f_n' \to f' \) pointwise almost everywhere.

This corollary can also be proved by combining the next theorem with (\).

XIV.15.3.5. **Theorem.** [Con74] Let \((f_n)\) be a sequence of convex functions on a closed bounded interval \([a, b]\), converging pointwise on \([a, b]\) to a continuous (convex) function \( f \). Then \( f_n' \to f' \) in the \( L^1 \)-norm (\), i.e.

\[
\lim_{n \to \infty} \|f_n' - f'\|_1 = \lim_{n \to \infty} \int_a^b |f_n'(x) - f'(x)| \, dx = 0 .
\]

1687
**XIV.15.3.6.** Corollary. Let \((f_n)\) be a sequence of convex functions on a closed bounded interval \([a, b]\), converging pointwise on \([a, b]\) to a continuous (convex) function \(f\). Then \(f_n \to f\) in the total variation norm (\(\|\cdot\|_{TV}\)).

To see this, we may assume that each \(f_n\) is continuous on \([a, b]\) by redefining it at the endpoints. Then \(f\) and each \(f_n\) is absolutely continuous on \(I\), and by (\(\|\cdot\|_1\)) the total variation norm of an absolutely continuous function \(g\) on \([a, b]\) is

\[
\|g\|_{TV} = |g(a)| + \int_a^b |g'(x)| \, d\lambda(x).
\]

To prove XIV.15.3.5., we need a lemma:

**XIV.15.3.7.** Lemma. Let \(f\) and \(g\) be convex functions on an interval \(I\), and \(P(x) = f(a) + \frac{f(b) - f(a)}{b - a} (x - a)\). Then

\[
\|f' - P'\|_1 = 2\|f - P\|_{\infty}
\]

\[
\|g' - P'\|_1 \leq \|f' - P'\|_1 + 4\|f - g\|_{\infty}
\]

where \(\|\phi\|_{\infty} = \sup\{|\phi(x)| : a \leq x \leq b\}\) is the uniform norm.

**Proof:** For the equality, note that \(f - P\) is convex and is 0 at \(a\) and \(b\), and there is a \(c \in (a, b)\) such that \(f\) is nonincreasing on \([a, c]\) and nondecreasing on \([c, b]\), so the total variation \(\|f' - P'\|_1\) is \(2|m|\), where \(m = f(c)\) is the minimum of \(f - P\) on \([a, b]\); and we also have \(|m| = \|f - P\|_{\infty}\).

For the inequality, \(g - P\) is convex on \([a, b]\), so either \(g - P\) is monotone on \([a, b]\), in which case

\[
\|g' - P'\|_1 = \|[g(b) - P(b)] - [g(a) - P(a)]\| = \|[g(b) - f(b)] - [g(a) - f(a)]\| \leq 2\|f - g\|_{\infty}
\]

or there is a \(d \in (a, b)\) with \(g - P\) nonincreasing on \([a, d]\) and nondecreasing on \([d, b]\), in which case

\[
\|g' - P'\|_1 = ([g(a) - P(a)] - [g(d) - P(d)]) + ([g(b) - P(b)] - [g(d) - P(d)])
\]

\[
= ([g(a) - f(a)] - [g(d) - P(d)]) + ([g(b) - f(b)] - [g(d) - P(d)]) = 2[P(d) - g(d)] + |g(a) - f(a)| + |g(b) - f(b)|
\]

\[
\leq 2|g(d) - P(d)| + 2\|g - f\|_{\infty} \leq 2(|g(d) - f(d)| + |f(d) - P(d)|) + \|g - f\|_{\infty}
\]

\[
\leq 2\|f - P\|_{\infty} + 4\|g - f\|_{\infty} = \|f' - P'\|_1 + 4\|f - g\|_{\infty}.
\]

\(\circ\)

We now give the proof of XIV.15.3.5.

**Proof:** Let \(\epsilon > 0\). Using (\(\|\cdot\|_1\)), fix a partition \(\mathcal{P} = \{a = t_0 < t_1 < \cdots < t_m = b\}\) of \([a, b]\) for which

\[
\|f' - (\mathcal{P}f)'\|_1 < \frac{\epsilon}{4}
\]

where \(\mathcal{P}f\) is the polygonal function whose graph has vertices \((t_k, f(t_k))\) for \(0 \leq k \leq m\). Since \(f_n \to f\) uniformly, choose \(N\) such that

\[
\|f_n - f\|_{\infty} < \frac{\epsilon}{8m}
\]

1688
for all $n \geq N$. For such an $n$, by Lemma XIV.15.3.7, we have, for each $k$,

$$\int_{t_{k-1}}^{t_k} |f'_n - (Pf)'| \leq \int_{t_{k-1}}^{t_k} |f' - (Pf)'| + 4\|f_n - f\|_\infty < \int_{t_{k-1}}^{t_k} |f' - (Pf)'| + \frac{\epsilon}{2m}.$$  

Thus we have

$$\int_a^b |f'_n - f'| \leq \int_a^b |f'_n - (Pf)'| + \int_a^b |(Pf)' - f'|$$

$$\leq \sum_{k=1}^m \int_{t_{k-1}}^{t_k} |f'_n - (Pf)'| + \frac{\epsilon}{4}$$

$$< \sum_{k=1}^m \left( \int_{t_{k-1}}^{t_k} |f' - (Pf)'| + \frac{\epsilon}{2m} \right) + \frac{\epsilon}{4}$$

$$= \int_a^b |f' - (Pf)'| + \frac{\epsilon}{2} + \frac{\epsilon}{4} < \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.$$

If the limit function is differentiable on $I$, we can say something stronger:

**XIV.15.3.8. Proposition.** Let $(f_n)$ be a sequence of convex functions on a closed bounded interval $[a, b]$, converging pointwise on $[a, b]$ to a (convex) function $f$. If $f$ is differentiable on $[a, b]$, then $(D_L f_n)$ and $(D_R f_n)$ converge uniformly to $f'$ on $[a, b]$ (i.e. $(f'_n)$ “converges uniformly” to $f'$ on $[a, b]$).

Indeed, under the hypotheses, $f'$ is continuous on $[a, b]$, so the proposition follows from the next more general result, which is strongly reminiscent of Dini’s Theorem (\():\)

**XIV.15.3.9. Proposition.** Let $(g_n)$ be a sequence of nondecreasing functions on a closed bounded interval $[a, b]$ converging pointwise on $[a, b]$ to a (necessarily nondecreasing) continuous function $g$. Then $g_n \to g$ uniformly on $[a, b]$.

**Proof:** Let $\epsilon > 0$. By continuity of $g$, choose a partition $\{a = t_0 < t_1 < \cdots < t_m = b\}$ of $[a, b]$ such that $g(t_k) - g(t_{k-1}) < \frac{\epsilon}{2}$ for $1 \leq k \leq m$. Choose $N$ such that $|g_n(t_k) - g(t_k)| < \frac{\epsilon}{2}$ for all $n \geq N$ and $0 \leq k \leq m$. Then, if $t_{k-1} \leq x \leq t_k$, we have, for $n \geq N$,

$$g(t_k) - \epsilon < g(t_k) - \frac{\epsilon}{2} \leq g_n(t_{k-1}) \leq g_n(x) \leq g_n(t_k) \leq g(t_k) + \frac{\epsilon}{2}$$

and we also have

$$g(t_k) - \frac{\epsilon}{2} < g(t_{k-1}) \leq g(x) \leq g(t_k)$$

and thus it follows that $|g_n(x) - g(x)| < \epsilon$. This is true for every $k$, i.e. for every $x \in [a, b]$.

If $f$ is not differentiable everywhere, we do not have $f'_n \to f'$ uniformly in general, even on closed bounded intervals contained in the interior of $I$:

1689
XIV.15.3.10. Example. Let \( f(x) = |x| \), and for each \( n \) let \( f_n(x) = |x|^{1+1/n} \). Then \( f \) and each \( f_n \) are convex, \( f_n \) is differentiable everywhere, with \( f'_n \) continuous, and \( f_n \to f \) pointwise on \( \mathbb{R} \) (and uniformly on any bounded subinterval). But \( f'_n \) does not converge uniformly to \( f' \) on any (deleted) interval around 0.

XIV.15.4. Jensen’s Inequality

A simple induction on \( n \), beginning with \( n = 2 \), shows:

XIV.15.4.1. Proposition. If \( f \) is convex on an interval \( I \), \( x_1, \ldots, x_n \in I \), and \( t_1, \ldots, t_n \geq 0 \), \( t_1 + \cdots + t_n = 1 \), then

\[
f(t_1 x_1 + \cdots + t_n x_n) \leq t_1 f(x_1) + \cdots + t_n f(x_n).
\]

The inequality is strict if the \( x_k \) are distinct and the \( t_k \) nonzero, unless \( f \) is linear on an interval containing the \( x_k \).

This can be rephrased:

XIV.15.4.2. Corollary. Let \( f \) be convex on an interval \( I \), \( x_1, \ldots, x_n \in I \), and \( s_1, \ldots, s_n \geq 0 \). Then

\[
\frac{s_1 x_1 + \cdots + s_n x_n}{s_1 + \cdots + s_n} \in I
\]

and

\[
f \left( \frac{s_1 x_1 + \cdots + s_n x_n}{s_1 + \cdots + s_n} \right) \leq \frac{s_1 f(x_1) + \cdots + s_n f(x_n)}{s_1 + \cdots + s_n}.
\]

The inequality is strict if the \( x_k \) are distinct and the \( s_k \) nonzero, unless \( f \) is linear on an interval containing the \( x_k \).

For the proof, just set \( t_k = \frac{s_k}{s_1 + \cdots + s_n} \).

A particularly important sample application is:

XIV.15.4.3. Corollary. Let \( x_1, \ldots, x_n \) be positive numbers, and \( \lambda_1, \ldots, \lambda_n \) positive numbers with \( \lambda_1 + \cdots + \lambda_n = 1 \). Then

\[
x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \cdots + \lambda_n x_n
\]

with strict inequality unless all the \( x_k \) are equal.

Proof: Note that \( f(x) = e^x \) is convex on \( \mathbb{R} \). Set \( y_k = \log x_k \), so \( x_k = e^{y_k} \). We then have

\[
(e^{y_1})^{\lambda_1} \cdots (e^{y_n})^{\lambda_n} = e^{\lambda_1 y_1 + \cdots + \lambda_n y_n} \leq \lambda_1 e^{y_1} + \cdots + \lambda_n e^{y_n}
\]

which is the desired inequality.

The special case where all \( \lambda_k = \frac{1}{n} \) is
Corollary. Let \( x_1, \ldots, x_n \) be positive numbers. Then the geometric mean of \( \{x_1, \ldots, x_n\} \) is less than or equal to the arithmetic mean, i.e.

\[
[x_1 x_2 \cdots x_n]^{1/n} \leq \frac{1}{n} [x_1 + \cdots + x_n]
\]

with strict inequality unless all the \( x_k \) are equal.

Jensen’s Inequality is a generalization of XIV.15.4.2. If \( (X, \mathcal{A}, \mu) \) is a finite measure space and \( f \) is a \( \mu \)-integrable function, recall () that the average value of \( f \) is

\[
av(f) = \frac{1}{\mu(X)} \int_X f \, d\mu.
\]

Theorem. [Jensen’s Inequality] Let \( (X, \mathcal{A}, \mu) \) be a finite measure space, \( f \) a \( \mu \)-integrable function taking values in an interval \( I \), and \( \phi \) a convex function on \( I \). Let \( c = \text{av}(f) \). Then \( c \in I \), and if either \( \phi \) is nonnegative on \( I \) or \( \phi \circ f \) is \( \mu \)-integrable, then

\[
\phi(c) \leq \frac{1}{\mu(X)} \int_X \phi \circ f \, d\mu.
\]

The inequality is strict unless \( f \) is constant a.e. or \( \phi \) is linear on a subinterval containing the essential range of \( f \).

The condition that \( \phi \) is nonnegative or that \( \phi \circ f \) is integrable is simply needed to define the right side of the inequality. (Integrability of \( \phi \circ f \) is not automatic; cf. Exercise XIV.15.7.3.) Actually, in general it turns out that \( (\phi \circ f)_- \) is always integrable (Exercise XIV.15.7.4), so \( \phi \circ f \) is integrable in the extended sense of () and the right side is well defined as \( +\infty \) if \( \phi \circ f \) is not integrable, even if \( \phi \) is not nonnegative.

Proof: If \( \phi \) is nonnegative but \( \phi \circ f \) is not integrable, the inequality is trivial, so it only needs proof in the case where \( \phi \circ f \) is integrable. Let us assume this.

To see that \( c \in I \), it suffices to show that if \( a < f(x) < b \) for all \( x \in X \), then \( a < c < b \). But in this case the function \( g(x) = f(x) - a \) is strictly positive on \( X \), so \( \int_X g \, d\mu > 0 () \), and similarly \( h(x) = b - f(x) \) has \( \int_X h \, d\mu > 0 \). In fact \( c \) is in the interior of \( I \) unless \( f(x) \) is equal almost everywhere to one of the endpoints of \( I \), in which case the inequality is a trivial equality.

Let \( m \) be any number between \( D_L \phi(c) \) and \( D_R \phi(c) () \), and let \( L(t) = \phi(c) + m(t - c) \) for \( x \in I \). Then \( L \) is a linear function, and \( L(t) \leq \phi(t) \) for all \( t \in I \). Thus we have

\[
\phi(c) \mu(X) = \int_X [\phi(c) + m(f(x) - c)] \, d\mu(x) = \int_X L(f(x)) \, d\mu(x) \leq \int_X \phi(f(x)) \, d\mu(x).
\]

We have equality if and only if \( L \circ f = \phi \circ f \) \( \mu \text{-a.e.} \), which is equivalent to the condition in the last line of the statement.

The inequality XIV.15.4.2. can be regarded as a special case of Jensen’s Inequality, where \( X \) is an \( n \)-point space and \( \{s_1, \ldots, s_n\} \) are the masses of the points.

Here are some of the other most commonly stated special cases of Jensen’s Inequality:
**XIV.15.4.7. Corollary.** Let $f$ be integrable on $[a, b]$ with values in an interval $I$, and let $\phi$ be convex on $I$. Let

$$c = \frac{1}{b-a} \int_a^b f \, d\lambda .$$

Then $c \in I$, and if $\phi$ is nonnegative or $\phi \circ f$ is integrable on $[a, b]$, then

$$\phi(c) \leq \frac{1}{b-a} \int_a^b \phi \circ f \, d\lambda .$$

A slightly more general version allows a weight function:

**XIV.15.4.8. Corollary.** Let $f$ be Lebesgue measurable on $[a, b]$ with values in an interval $I$, and let $\phi$ be convex on $I$. Let $\omega$ be a nonnegative Lebesgue integrable function on $[a, b]$, with $f \omega$ integrable on $[a, b]$. Set

$$c = \frac{1}{\int_a^b \omega \, d\lambda} \int_a^b f \omega \, d\lambda .$$

Then $c \in I$, and if $\phi$ is nonnegative or $(\phi \circ f) \omega$ is integrable on $[a, b]$, then

$$\phi(c) \leq \frac{1}{\int_a^b \omega \, d\lambda} \int_a^b (\phi \circ f) \omega \, d\lambda .$$

If $f$ is continuous, then everything can be stated and proved in terms of Riemann integration:

**XIV.15.4.9. Corollary.** Let $f$ be continuous on $[a, b]$ with values in an interval $I$, and let $\phi$ be convex and continuous on $I$. Let

$$c = \frac{1}{b-a} \int_a^b f(x) \, dx .$$

Then $c \in I$, and

$$\phi(c) \leq \frac{1}{b-a} \int_a^b \phi(f(x)) \, dx .$$

An examination of the proof of XIV.15.4.5. shows that only Riemann integrals need be considered in this case. Without the continuity assumption on $\phi$, $\phi \circ f$ can fail to be Riemann-integrable on $[a, b]$ (Exercise XIV.15.7.5.).

**XIV.15.5. Log-Convex Functions**

**XIV.15.5.1. Definition.** If $I$ is an interval, a function $g : I \to (0, \infty)$ is log-convex on $I$ if $\log g$ is convex on $I$. 

1692
XIV.15.5.2. Alternatively, a positive function $g$ on an interval $I$ is log-convex on $I$ if, for all $a, b \in I$, $0 \leq \lambda \leq 1$,

$$g(\lambda a + (1 - \lambda)b) \leq g(a)^\lambda g(b)^{1-\lambda}.$$  

The right-hand side is a sort of weighted geometric mean of $g(a)$ and $g(b)$. If $g$ is continuous, then by XIV.15.2.3. $g$ is log-convex if and only if, for every $a, b \in I$,

$$g \left( \frac{a + b}{2} \right) \leq \sqrt{g(a) g(b)}.$$  

XIV.15.5.3. Proposition. A log-convex function is convex.

Proof: This follows immediately from XIV.15.5.2. and XIV.15.4.3. &

XIV.15.5.4. The converse is false. $g(x) = x^2$ is convex on $(0, \infty)$, but not log-convex on any subinterval. In fact, an increasing log-convex function on an interval $(a, \infty)$ must grow at least exponentially as $x \to +\infty$.

XIV.15.5.5. Since a sum of convex functions is convex, it is obvious that a product of log-convex functions is log-convex. It is also clear that a positive scalar multiple of a log-convex function is log-convex. It is less obvious, but true, that a sum of log-convex functions is log-convex (XIV.15.5.7.).

XIV.15.5.6. Lemma. Let $a_1, a_2, b_1, b_2, c_1, c_2$ be positive real numbers. If $a_1 c_1 - b_1^2 \geq 0$ and $a_2 c_2 - b_2^2 \geq 0$, then

$$(a_1 + a_2)(c_1 + c_2) - (b_1 + b_2)^2 \geq 0.$$  

Proof: If $a, b, c$ are positive real numbers, then the polynomial $f(x) = ax^2 + 2bx + c$ takes only nonnegative real values (for $x$ real) if and only if $ac - b^2 \geq 0$. By assumption, the polynomials $f_1(x) = a_1 x^2 + 2b_1 x + c_1$ and $f_2(x) = a_2 x^2 + 2b_2 x + c_2$ take only nonnegative real values, and hence

$$f_1(x) + f_2(x) = (a_1 + a_2)x^2 + 2(b_1 + b_2)x + (c_1 + c_2)$$

also takes only nonnegative real values. &

XIV.15.5.7. Proposition. A sum of log-convex functions on an interval $I$ is log-convex on $I$.

Proof: Since log-convex functions are continuous, by XIV.15.5.2. it suffices to show that if $f$ and $g$ are log-convex on $I$ and $x, y \in I$, then

$$\left[ f \left( \frac{x + y}{2} \right) + g \left( \frac{x + y}{2} \right) \right]^2 \leq (f(x) + g(x))(f(y) + g(y)).$$

Apply XIV.15.5.6. with $a_1 = f(x)$, $a_2 = g(x)$, $b_1 = f \left( \frac{x+y}{2} \right)$, $b_2 = g \left( \frac{x+y}{2} \right)$, $c_1 = f(y)$, $c_2 = g(y).$ &

1693
XIV.15.5.8. The conclusion of XIV.15.5.7. extends to arbitrary finite sums. Since a positive pointwise limit of log-convex functions is obviously log-convex, it also extends to a positive pointwise limit of finite sums of log-convex functions, e.g. a pointwise-convergent infinite series. By approximating the integrals by Riemann sums, we also obtain:

XIV.15.5.9. PROPOSITION. Let \( f \) be a function on \( I \times [a,b] \) for some interval \( I \) and \( a, b \in \mathbb{R}, a < b \). If \( f(x, t) \) is log-convex on \( I \) for each fixed \( t \), and \( f(x, t) \) is continuous on \( [a,b] \) for each fixed \( x \), then

\[
g(x) = \int_a^b f(x, t) \, dt
\]

is log-convex on \( I \).

PROOF: Let \( n \in \mathbb{N} \) and \( h = \frac{b-a}{n} \). Then \( g_n(x) = h(f(x,a) + f(x,a+h) + \cdots + f(x,a+(n-1)h) \) is log-convex, and \( g_n \to g \) pointwise. Since \( f \) is strictly positive, it is obvious that \( g \) is also strictly positive. 

This result extends to improper integrals:

XIV.15.5.10. COROLLARY. Let \( f \) be a function on \( I \times (a,b) \) for some interval \( I \) and \( a, b \in [-\infty, \infty], a < b \). If \( f(x, t) \) is log-convex on \( I \) for each fixed \( t \), \( f(x, t) \) is continuous on \( (a,b) \) for each fixed \( x \), and the improper Riemann integral \( \int_a^b f(x, t) \, dt \) converges for all \( x \in I \), then

\[
g(x) = \int_a^b f(x, t) \, dt
\]

is log-convex on \( I \).

PROOF: Let \( (a_n) \) be a decreasing sequence with \( a_n \to a \), \( (b_n) \) an increasing sequence with \( b_n \to b \), \( a_n < b_n \) for all \( n \). Set

\[
g_n(x) = \int_{a_n}^{b_n} f(x, t) \, dt .
\]

Then \( g_n \) is log-convex on \( I \) for all \( n \) by XIV.15.5.9., and \( g_n \to g \) pointwise on \( I \). It is also clear that \( g \) is strictly positive.

XIV.15.6. Higher-Dimensional Versions of Convexity

XIV.15.7. Exercises

XIV.15.7.1. Let \( f : \mathbb{R} \to \mathbb{R} \) be an additive function, i.e. \( f(x+y) = f(x) + f(y) \) for all \( x, y \in \mathbb{R} \). Show that if there is any nontrivial closed bounded interval \([a,b]\) on which \( f \) is bounded or Lebesgue measurable, then \( f \) is linear on \( \mathbb{R} \), i.e. there is an \( m \in \mathbb{R} \) with \( f(x) = mx \) for all \( x \in \mathbb{R} \). [Show that \( f \) and \( -f \) are midpoint-convex, and use XIV.15.2.7.]

1694
**XIV.15.7.2.** Let \( \{x_i : i \in I\} \) be a Hamel basis () for \( \mathbb{R} \) over \( \mathbb{Q} \), i.e. for each \( x \in \mathbb{R} \) there is a unique representation
\[
    x = \sum_{i \in I} c_i(x)x_i
\]
where \( c_i(x) \in \mathbb{Q} \) for each \( i \) and only finitely many \( c_i(x) \) are nonzero. (The Axiom of Choice is needed here.)

(a) Show that each function \( c_i \) is \( \mathbb{Q} \)-linear, i.e. \( c_i(x + y) = c_i(x) + c_i(y) \) and \( c_i(ax) = ac_i(x) \) for all \( x, y \in \mathbb{R}, a \in \mathbb{Q} \).

(b) Show that each \( c_i \) is a midpoint-convex function on \( \mathbb{R} \) (and that \(-c_i \) is also midpoint-convex).

(c) Conclude that \( c_i \) is discontinuous everywhere. [If \( j \neq i \), then \( \mathbb{Q}x_j \) and \( x_i + \mathbb{Q}x_j \) are both dense in \( \mathbb{R} \).]

(d) Conclude that \( c_i \) is not convex on any interval. It then follows from XIV.15.2.4. that \( c_i \) is unbounded on every interval (this can be seen directly by noting that if \( j \neq i \), then \( nx_i + \mathbb{Q}x_j \) is dense in \( \mathbb{R} \) for each \( n \in \mathbb{Z} \)). Conclude also that the restriction of \( c_i \) to any interval is not Lebesgue measurable.

**XIV.15.7.3.** Let \( X = (0, 1) \) with Lebesgue measure. Let \( f(x) = x, I = (0, 1), \phi(x) = e^{1/x} \). Show that \( X, f, I, \phi \) satisfy the hypotheses of Jensen’s Inequality, but that \( \phi \circ f \) is not integrable on \( X \).

**XIV.15.7.4.** Let \( (X, A, \mu) \) be a finite measure space, \( f \) a \( \mu \)-integrable function on \( X \) taking values in an integral \( I \), and \( \phi \) a convex function on \( I \). Show that \( \phi \circ f \) is \( \mu \)-integrable. [There is a linear function \( L \) such that \( \phi \geq L \) on \( I \).]

**XIV.15.7.5.** Let \( K \) be a Cantor set of positive measure (), and \( f \) a continuous function from \([0, 1]\) to \([0, 1]\) which takes the value 1 precisely on \( K \), e.g. \( f(x) = 1 - \text{dist}(x, K) \). Let \( \phi \) be the convex function on \([0, 1]\) defined in XIV.15.1.8.(v). Then \( \phi \circ f \) is not Riemann-integrable on \([0, 1]\).

**XIV.15.7.6.** Let \( \mathcal{F} \) be a collection of convex functions on an interval \([a, b]\), and \([c, d]\) a closed bounded subinterval of \((a, b)\).

(a) If the functions in \( \mathcal{F} \) are uniformly bounded above at \( a \) and \( b \), and there is an \( x_0 \in (a, b) \) such that the functions in \( \mathcal{F} \) are uniformly bounded below at \( x_0 \), show that \( \mathcal{F} \) is uniformly bounded on \([a, b]\).

(b) Under the hypotheses of (a), show that there are real numbers \( m \) and \( M \) such that
\[
    m \leq D_L f(x) \leq D_R f(x) \leq M
\]
for all \( f \in \mathcal{F} \) and all \( x \in [c, d] \). Conclude that \( \mathcal{F} \) is uniformly equicontinuous () on \([c, d]\). In particular, \( \mathcal{F} \) is equicontinuous on \((a, b)\).

(c) If \( (f_n) \) is a sequence of convex functions on \([a, b]\) converging pointwise on \([a, b]\) to a (convex) function \( f \), use (a), (b), and the Arzela-Ascoli Theorem () to conclude that \( f_n \to f \) uniformly on \([c, d]\) (cf. XIV.15.3.1.).

(d) If \( (f_n) \) is a sequence of continuous convex functions on \([a, b]\) converging pointwise on \([a, b]\) to a continuous (convex) function \( f \), show that the set
\[
    \mathcal{F} = \{f_n : n \in \mathbb{N}\}
\]
is uniformly bounded and equicontinuous on \([a, b]\).

(e) If \( (f_n) \) is a sequence of convex functions on \([a, b]\) converging pointwise on \([a, b]\) to a function \( f \), use (d) and the Arzela-Ascoli Theorem () to give an alternate proof of XIV.15.3.1..
XIV.15.7.7. Let $f$ be convex on an interval $I$ with right endpoint $b$, and suppose $f$ is continuous at $b$. Show that if $\sup_{x<b} D_L f(x) = m < \infty$, then $D_L f(b) = m$.

XIV.15.7.8. Let $f$ be convex on $[a, +\infty)$ for some $a$. Show that

$$\lim_{x \to +\infty} \frac{f(x)}{x}$$

exists (possibly as $+\infty$, but not $-\infty$) and equals

$$\lim_{x \to +\infty} D_L f(x) = \lim_{x \to +\infty} D_R f(x) = \sup_{x > a} D_L f(x) = \sup_{x \geq a} D_R f(x).$$

Prove a similar statement for $-\infty$.

XIV.15.7.9. Let $f$ be a function on an interval $I$ taking only positive values, with $f''$ defined everywhere on $I$.

(a) Show that $f$ is log-convex on $I$ if and only if

$$f(x)f''(x) \geq [f'(x)]^2$$

for all $x \in I$. [Set $g(x) = \log f(x)$, and compute $g''$.]

(b) What if $f''$ is only defined on the complement of a countable set? What if $f''$ is only defined almost everywhere?

XIV.15.7.10. Let $f$ and $g$ be convex functions on an interval of the form $(a, +\infty)$ for some $a$. Show that if

$$\lim_{n \to \infty} [f(n) - g(n)] = 0$$

(limit for $n \in \mathbb{N}$), then

$$\lim_{x \to +\infty} [f(x) - g(x)] = 0$$

(limit for $x \in \mathbb{R}$).
XIV.16. Functions of Bounded Variation

Functions of bounded variation are ones which vary in a controlled way. Geometrically, a function has bounded variation on an interval if its graph has finite arc length. Functions of bounded variation are well-behaved, with perhaps unexpectedly nice properties which are technically useful.

XIV.16.1. The Total Variation of a Function

XIV.16.1.1. In the definition, we will use the positive and negative part notation: if \( x \) is a real number, write \( x^+ = \max(x, 0) = \frac{|x| + x}{2} \) and \( x^- = \max(-x, 0) = \frac{|x| - x}{2} \), so that \( x = x^+ - x^- \) and \( |x| = x^+ + x^- \). Thus if \( y \) and \( z \) are real numbers, we will write \( (y + z)^+ = \max(y + z, 0) \) and \( (y + z)^- = \max(z - y, 0) \). Thus \( (y - z)^+ = y - z \) if \( y \geq z \) and 0 otherwise, and \( (y - z)^- = |y - z| \) if \( y \leq z \) and 0 otherwise. We always have \( (y - z)^+ - (y - z)^- = y - z \) and \( (y - z)^+ + (y - z)^- = |y - z| \).

XIV.16.1.2. Proposition. If \( x, y \in \mathbb{R} \), then \( (x + y)^+ \leq x^+ + y^+ \) and \( (x + y)^- \leq x^- + y^- \).

This can be proved brute-force on a case-by-case basis, or using the triangle inequality (Exercise (1)).

XIV.16.1.3. Definition. Let \( f \) be a real-valued function on an interval \([a, b]\). If

\[
P = \{ a = x_0, x_1, \ldots, x_n = b \}
\]

is a partition of \([a, b]\), set

\[
V^+(f, P) = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})]^+
\]

\[
V^-(f, P) = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})]^-
\]

\[
V(f, P) = V^+(f, P) + V^-(f, P) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|
\]

\(V^+(f, P), V^-(f, P),\) and \(V(f, P)\) are respectively the positive variation, negative variation, and total variation of \( f \) over the partition \( P \).

These numbers of course depend on the interval \([a, b]\), which is implicitly coded in the partition \( P \).

XIV.16.1.4. Proposition. If \( P \) is any partition of \([a, b]\), then \( V^+(f, P) \) and \( V^-(f, P) \) are nonnegative real numbers, and

\[
V^+(f, P) - V^-(f, P) = f(b) - f(a).
\]

Proof: The sums are obviously nonnegative and finite. We have

\[
V^+(f, P) - V^-(f, P) = \sum_{k=1}^{n} ([f(x_k) - f(x_{k-1})]^+ - [f(x_k) - f(x_{k-1})]^-) \]

1697
The function $f$ has bounded variation on $[a, b]$ if $V_{[a,b]}(f) < +\infty$.

**Examples.** (i) Any monotone function on an interval $[a, b]$ has bounded variation. If $f$ is nondecreasing on $[a, b]$, then $V_{[a,b]}^+(f) = V_{[a,b]}(f) = f(b) - f(a)$ and $V_{[a,b]}^-(f) = 0$; if $f$ is nonincreasing on $[a, b]$, then $V_{[a,b]}^-(f) = V_{[a,b]}(f) = f(a) - f(b)$ and $V_{[a,b]}^+(f) = 0$.

(ii) A Lipschitz function on $[a, b]$ has bounded variation. If $f$ is Lipschitz with Lipschitz constant $M$, i.e., $|f(x) - f(y)| \leq M|x - y|$ for all $x, y \in [a, b]$, then if $\mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\}$ is any partition of $[a, b]$, we have

$$V(f, \mathcal{P}) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \leq \sum_{k=1}^{n} M|x_k - x_{k-1}| = M(b - a)$$

so $V_{[a,b]}(f) \leq M(b - a)$. In particular, if $f$ is differentiable on $[a, b]$ and $f'$ is bounded, then $f$ has bounded variation on $[a, b]$.
(iii) A continuous function need not have bounded variation. Continuous but everywhere non-differentiable functions cannot have bounded variation (XIV.16.2.4(iv)). Even a function which is differentiable everywhere need not have bounded variation. Let

\[ f(x) = x^2 \sin \left( \frac{1}{x^2} \right) \]

for \( x \neq 0 \), \( f(0) = 0 \) (Figure V.4). Then \( f \) is differentiable everywhere, but not of bounded variation on \([0, 1]\).

For each \( n \), let \( x_n = \frac{1}{\sqrt{2} + \pi(n - 1)} \).

Then

\[ f(x_n) = \pm \frac{1}{\sqrt{2} + \pi(n - 1)} \]

where the signs alternate, so

\[ |f(x_{n+1}) - f(x_n)| = \frac{1}{\sqrt{2} + \pi n} + \frac{1}{\sqrt{2} + \pi(n - 1)} \geq \frac{1}{\pi n} \]

so if \( P \) runs over partitions containing large initial finite segments of \( \{x_n : n \in \mathbb{N}\} \), \( V(f, P) \) is arbitrarily large since the harmonic series diverges (IV.2.2.11.). Thus \( V_{[0,1]}(f) = +\infty \). The problem is, roughly speaking (even precisely speaking; cf. XIV.16.2.5.), that \( f' \) is not Lebesgue integrable on \([0, 1]\).

It turns out that if \( f \) is differentiable everywhere on \([a, b]\) and \( f' \) is Lebesgue integrable on \([a, b]\), then \( f \) has bounded variation on \([a, b]\), but this requires considerable work to prove (). This cannot, however, be relaxed to just requiring that \( f \) be continuous, differentiable almost everywhere, and with integrable derivative ()

Example (i) shows that a function of bounded variation need not be continuous. But it must be “almost continuous”: it can have only countably many discontinuities, each of which is a jump discontinuity (XIV.16.2.4.).

XIV.16.1.8. PROPOSITION. Let \( f \) be a real-valued function on an interval \([a, b]\), and let \( \Gamma_f \) be its graph

\[ \Gamma_f = \{(x, f(x)) : a \leq x \leq b\} \]

Then \( f \) has bounded variation on \([a, b]\) if and only if the arc length () \( \ell(\Gamma_f) \) of \( \Gamma_f \) is finite. In fact,

\[ V_{[a,b]}(f) \leq \ell(\Gamma_f) \leq (b - a) + V_{[a,b]}(f) \]

Proof: We have that

\[ \ell(\Gamma_f) = \sup_P \left\{ \sum_{k=1}^{n} ||(x_k, f(x_k)) - (x_{k-1}, f(x_{k-1}))|| \right\} \]

\[ = \sup_P \left\{ \sum_{k=1}^{n} \sqrt{|x_k - x_{k-1}|^2 + |f(x_k) - f(x_{k-1})|^2} \right\} \]

1699
\[
\leq \sup_{\mathcal{P}} \left\{ \sum_{k=1}^{n} (|x_k - x_{k-1}| + |f(x_k) - f(x_{k-1})|) \right\} \\
= \sup_{\mathcal{P}} \left\{ (b-a) + \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \right\} \\
= \sup_{\mathcal{P}} \{ (b-a) + V(f, \mathcal{P}) \} = (b-a) + V_{[a,b]}(f)
\]

where the suprema are over all partitions \( \mathcal{P} = \{a = x_0, \ldots, x_n = b\} \) of \([a,b]\). The first inequality is evident. \( \Box \)

The first inequality is always strict if \( f \) has bounded variation, but the second inequality can be an equality (\( \cdot \)).

**XIV.16.1.9.** Since a function \( f \) of bounded variation need not be continuous, the graph of \( f \) need not be an arc (a homeomorphic image of \([0,1]\)). We normally only consider arc length for arcs. But since a function of bounded variation has only jump discontinuities, and only finitely many jumps of length more than \( \epsilon \) for any \( \epsilon > 0 \), the graph of \( f \) can be expanded to an arc by adding vertical line segments across the jumps. The arc length of this arc is still no more than \( (b-a) + V_{[a,b]}(f) \).

We now explore properties of the variations of a function.

**XIV.16.1.10.** **Proposition.** Let \( f \) be a real-valued function on an interval \([a,b]\). Then

\[
V_{[a,b]}^+(f) + V_{[a,b]}^-(f) = V_{[a,b]}(f)
\]

and, if \( f \) has bounded variation on \([a,b]\), then

\[
V_{[a,b]}^+(f) - V_{[a,b]}^-(f) = f(b) - f(a).
\]

If \( f \) does not have bounded variation, then \( V_{[a,b]}^+(f) = V_{[a,b]}^-(f) = +\infty \).

**Proof:** If \( \mathcal{P} \) is any partition of \([a,b]\), we have

\[
V(f, \mathcal{P}) = V^+(f, \mathcal{P}) + V^-(f, \mathcal{P}) \leq V_{[a,b]}^+(f) + V_{[a,b]}^-(f)
\]

and thus \( V_{[a,b]}(f) \leq V_{[a,b]}^+(f) + V_{[a,b]}^-(f) \). For the opposite inequality, there is nothing to prove if \( V_{[a,b]}(f) = +\infty \), so assume \( V_{[a,b]}(f) \) is finite. Since \( V_{[a,b]}^+(f) \leq V_{[a,b]}(f) \) and \( V_{[a,b]}^-(f) \leq V_{[a,b]}(f) \), they are both also finite. Let \( \epsilon > 0 \). Choose a partition \( \mathcal{P} \) fine enough that \( V^+(f, \mathcal{P}) \geq V_{[a,b]}^+(f) - \epsilon \) and \( V^-(f, \mathcal{P}) \geq V_{[a,b]}^-(f) - \epsilon \). Then

\[
V_{[a,b]}^+(f) + V_{[a,b]}^-(f) - 2\epsilon \leq V^+(f, \mathcal{P}) + V^-(f, \mathcal{P}) = V(f, \mathcal{P}) \leq V_{[a,b]}(f)
\]

and since \( \epsilon > 0 \) is arbitrary, \( V_{[a,b]}^+(f) + V_{[a,b]}^-(f) \leq V_{[a,b]}(f) \).

For the last equation, suppose \( f \) has bounded variation. Let \( \epsilon > 0 \), and fix a partition \( \mathcal{P} \) fine enough that \( |V_{[a,b]}^+(f) - V^+(f, \mathcal{P})| < \frac{\epsilon}{2} \) and \( |V_{[a,b]}^-(f) - V^-(f, \mathcal{P})| < \frac{\epsilon}{2} \). Since by XIV.16.1.4. we have \( V^+(f, \mathcal{P}) - V^-(f, \mathcal{P}) = f(b) - f(a) \), we have

\[
||V_{[a,b]}^+(f) - V_{[a,b]}^-(f)| - (f(b) - f(a))||
\]

1700
This is a contradiction since \( \frac{V}{P} \) over then \( BV \).

XIV.16.1.14. **Proposition.** Let \( f \) be a real-valued function on \([a, b]\), and \( c \in (a, b) \). Then \( V^{+}_{[a, b]}(f) = V^{+}_{[a, b]}(f) + V^{-}_{[a, b]}(f) \), \( V^{-}_{[a, b]}(f) = V^{-}_{[a, b]}(f) + V^{+}_{[a, b]}(f) \), and \( V^{+}_{[a, b]}(f) = V^{+}_{[a, c]}(f) + V^{-}_{[c, b]}(f) \).

**Proof:** Since \( V^{+}(f, P) \) increases as \( P \) becomes finer, in computing \( V^{+}_{[a, b]}(f) \) we may restrict to partitions of \([a, b]\) containing \( c \). Any such \( P \) splits into a partition \( P_1 \) of \([a, c]\) and a partition \( P_2 \) of \([c, b]\). Conversely, any partitions \( P_1 \) of \([a, c]\) and \( P_2 \) of \([c, b]\) combine into a partition \( P \) of \([a, b]\) containing \( c \). For any such \( P \), we have \( V^{+}(f, P) = V^{+}(f, P_1) + V^{+}(f, P_2) \). Taking suprema over \( P_1 \) and \( P_2 \) corresponds to taking supremum over \( P \). The argument for \( V^{-} \) and \( V \) is identical.

The next simple result is used repeatedly.

XIV.16.1.11. **Proposition.** Let \( f \) be a real-valued function on \([a, b]\), and \( c \in (a, b) \). Then \( V^{+}_{[a, b]}(f) = V^{+}_{[a, c]}(f) + V^{-}_{[c, b]}(f) \), \( V^{-}_{[a, b]}(f) = V^{-}_{[a, c]}(f) + V^{+}_{[c, b]}(f) \), and \( V^{+}_{[a, b]}(f) = V^{+}_{[a, c]}(f) + V^{-}_{[c, b]}(f) \).

**Proof:** Since \( V^{+}(f, P) \) increases as \( P \) becomes finer, in computing \( V^{+}_{[a, b]}(f) \) we may restrict to partitions of \([a, b]\) containing \( c \). Any such \( P \) splits into a partition \( P_1 \) of \([a, c]\) and a partition \( P_2 \) of \([c, b]\). Conversely, any partitions \( P_1 \) of \([a, c]\) and \( P_2 \) of \([c, b]\) combine into a partition \( P \) of \([a, b]\) containing \( c \). For any such \( P \), we have \( V^{+}(f, P) = V^{+}(f, P_1) + V^{+}(f, P_2) \). Taking suprema over \( P_1 \) and \( P_2 \) corresponds to taking supremum over \( P \). The argument for \( V^{-} \) and \( V \) is identical.

XIV.16.1.12. **Corollary.** If \( f \) has bounded variation on \([a, b]\), and \([c, d]\) is any (closed) subinterval of \([a, b]\), then \( f \) has bounded variation on \([c, d]\), and \( V^{+}_{[c, d]}(f) \leq V^{+}_{[a, b]}(f) \), \( V^{-}_{[c, d]}(f) \leq V^{-}_{[a, b]}(f) \), \( V^{+}_{[c, d]}(f) \leq V^{-}_{[a, b]}(f) \).

XIV.16.1.13. **Proposition.** Let \( f \) and \( g \) be real-valued functions on \([a, b]\), and \( \alpha \in \mathbb{R} \). Then

(i) \( V^{+}_{[a, b]}(f + g) \leq V^{+}_{[a, b]}(f) + V^{+}_{[a, b]}(g) \), \( V^{-}_{[a, b]}(f + g) \leq V^{-}_{[a, b]}(f) + V^{-}_{[a, b]}(g) \), \( V^{+}_{[a, b]}(f + g) \leq V^{+}_{[a, b]}(f) + V^{+}_{[a, b]}(g) \).

(ii) \( V^{+}_{[a, b]}(\alpha f) = |\alpha| V^{+}_{[a, b]}(f) \). If \( \alpha \geq 0 \), then \( V^{+}_{[a, b]}(\alpha f) = \alpha V^{+}_{[a, b]}(f) \) and \( V^{-}_{[a, b]}(\alpha f) = \alpha V^{-}_{[a, b]}(f) \); if \( \alpha < 0 \), then \( V^{+}_{[a, b]}(\alpha f) = |\alpha| V^{-}_{[a, b]}(f) \) and \( V^{-}_{[a, b]}(\alpha f) = |\alpha| V^{+}_{[a, b]}(f) \).

Part (i) follows immediately from XIV.16.1.2., and the simple proof of (ii) is left as an exercise.

This result shows that if \( BV[a, b] \) is the set of functions of bounded variation on \([a, b]\), then \( BV[a, b] \) is a vector space of functions and \( ||f||_{BV} = |f(a)| + V_{[a, b]}(f) \) is a norm on \( BV[a, b] \). In fact, \( BV[a, b] \) is a Banach space under this norm (.).

1701
XIV.16.2. The Total Variation Functions

We can use the results of the previous section to define positive, negative, and total variation functions for any function of bounded variation, which yield strong structure results about such functions.

XIV.16.2.1. Definition. Let $f$ be a function of bounded variation on $[a,b]$. For $x \in [a,b]$, define

- $T^+ f(x) = V_{[a,x]}^+(f)$
- $T^- f(x) = V_{[a,x]}^-(f)$
- $T_f(x) = V_{[a,x]}(f) = T^+ f(x) + T^- f(x)$

$T^+ f$, $T^- f$, and $T_f$ are called the positive variation function, the negative variation function, and the total variation function of $f$ on $[a,b]$.

These functions depend (up to constants) on the choice of left endpoint of the interval $[a,b]$; when it is necessary to specify the starting point $a$, we will write $T^+_a f$, $T^-_a f$, $T_a f$.

XIV.16.2.2. The functions $T^+ f$, $T^- f$, $T_f$ are nonnegative and nondecreasing by XIV.16.1.12. We define all to be 0 at $a$. By XIV.16.1.4., $T^+ f(x) - T^- f(x) = f(x) - f(a)$ for all $x \in [a,b]$. Thus we obtain the following important consequence:

XIV.16.2.3. Theorem. Every function of bounded variation on an interval $[a,b]$ can be written as a difference of two nondecreasing functions on $[a,b]$.

In fact, if $f$ has bounded variation, then $f = (f(a) + T^+ f) - T^- f$.

Because of the known properties of nondecreasing functions, we obtain:

XIV.16.2.4. Corollary. Let $f$ be a function of bounded variation on an interval $[a,b]$. Then

(i) $f$ is Borel measurable.
(ii) $f$ has left and right limits at every point of $[a,b]$.
(iii) $f$ has only countably many discontinuities on $[a,b]$, each of which is a jump discontinuity.
(iv) $f$ is differentiable a.e. on $[a,b]$. 

1702
XIV.16.2.5. Proposition. Let \( f \) be a function of bounded variation on \([a, b]\). Then \( f' \) (which is defined a.e. on \([a, b]\)) is integrable on \([a, b]\), and

\[
\int_a^b |f'| \, d\lambda \leq V_{[a,b]}(f) .
\]

Proof: Extend \( f \) to \([a, b + 1]\) by setting \( f(x) = f(b) \) for \( x \in [b, b + 1] \). For each \( n \) define

\[
f_n(x) = n \left[ f \left( x + \frac{1}{n} \right) - f(x) \right].
\]

Then each \( f_n \) is (Borel) measurable, and \( f_n(x) \to f'(x) \) for any \( x \in (a, b) \) for which \( f \) is differentiable at \( x \). Thus \( f_n \to f' \) pointwise a.e., so \( f' \) is measurable. Also, since \( T^+f, T^-f, Tf \) are nondecreasing, they are differentiable a.e., with nonnegative derivatives, and since \( f = (f(a) + T^+f) - T^-f \), we have \( f' = (T^+f)' - (T^-f)' \) a.e., and thus

\[
|f'| \leq (T^+f)' + (T^-f)' = (Tf)' \text{ a.e.}
\]

By XIV.14.2.13,

\[
\int_a^b |f'| \, d\lambda \leq \int_a^b (Tf)' \, d\lambda \leq Tf(b) - Tf(a) = V_{[a,b]}(f) .
\]

XIV.16.2.6. Proposition. Let \( f \) be a function of bounded variation on \([a, b]\). Then the variation functions \( T^+f, T^-f, Tf \) are continuous wherever \( f \) is continuous.

XIV.16.2.7. Corollary. If \( f \) is a continuous function of bounded variation on \([a, b]\), then \( T^+f, T^-f, Tf \) are also continuous on \([a, b]\). Thus \( f \) can be written as a difference of two continuous nondecreasing functions on \([a, b]\).

XIV.16.3. Banach’s Multiplicity Theorem

For continuous real-valued functions on a closed bounded interval, there is a close relationship between the notion of bounded variation and the Multiplicity Function (XII.2.5.1.): a function is informally of bounded variation if its graph does not “wiggle up and down” too much, hence the function does not take “too many values too many times.” This theorem is due to Banach ( ).
XIV.16.3.1. **Theorem.** Let $f$ be a continuous function from a closed bounded interval $[a, b]$ to $\mathbb{R}$. Then $N_f$ is finite a.e. and (Lebesgue) integrable on $[c, d] = f([a, b])$ if and only if $f$ has bounded variation on $[a, b]$, and in any case
\[
\int_c^d N_f \, d\lambda = V_{[a,b]}(f) .
\]

**Proof:** Recall (XII.2.5.4.) that $N_f$ is nonnegative and (Borel) measurable, so $\int_c^d N_f \, d\lambda$ makes sense. For each $n$ let $\phi_n$ be the function defined in the proof of XII.2.5.4.. Then we have
\[
\int_c^d \phi_n \, d\lambda = \sum_{k=1}^{2^n} \lambda(f(J_k)) = \sum_{k=1}^{2^n} \lambda(f(J_k)) \leq \sum_{k=1}^{2^n} V_{[x_{k-1}, x_k]}(f) = V_{[a,b]}(f)
\]
(where the $x_k$ are the numbers from the proof of XII.2.5.4., i.e. the endpoints of the intervals $J_k$). We have $\int_c^d \phi_n \, d\lambda \to \int_c^d N_f \, d\lambda$ by the Monotone Convergence Theorem. So if $f$ has bounded variation on $[a, b]$, we have that $N_f$ is integrable on $[c, d]$ and in particular finite a.e. on $[c, d]$, and
\[
\int_c^d N_f \, d\lambda \leq V_{[a,b]}(f) .
\]

Conversely, suppose $N_f$ is finite a.e. and integrable on $[c, d]$. For $a \leq x \leq b$, let $f_x$ be the restriction of $f$ to $[a, x]$, and set $S(x) = \int_c^d N_{f_x} \, d\lambda$. We have $N_{f_x} \leq N_f$, and for $a \leq x_1 \leq x_2 \leq b$ we have
\[
0 = S(a) \leq S(x_1) \leq S(x_2) \leq S(b) = \int_c^d N_f \, d\lambda
\]
and, furthermore, since $f$ takes every value between $f(x_1)$ and $f(x_2)$ on $[x_1, x_2]$ by the Intermediate Value Theorem (), we have $N_{f_{x_2}}(y) \geq N_{f_{x_1}}(y) + 1$ for all $y$ between $f(x_1)$ and $f(x_2)$ and thus we have
\[
S(x_2) - S(x_1) = \int_c^d (N_{f_{x_2}} - N_{f_{x_1}}) \, d\lambda \geq \int_{f(x_1)}^{f(x_2)} 1 \, d\lambda = |f(x_2) - f(x_1)| .
\]

So if $a \leq x_0 < x_1 < \cdots < x_n \leq b$, we have
\[
\int_c^d N_f \, d\lambda = S(b) - S(a) \geq \sum_{k=1}^n (S(x_k) - S(x_{k-1})) \geq \sum_{k=1}^n |f(x_k) - f(x_{k-1})|
\]
and taking the supremum over all such finite subsets of $[a, b]$ we obtain
\[
\int_c^d N_f \, d\lambda \geq V_{[a,b]}(f)
\]
and thus $f$ has bounded variation, and the opposite inequality holds by the first part of the proof, giving equality.

If $f$ does not have bounded variation, equality still holds since both sides are $+\infty$. \(\diamondsuit\)
XIV.16.4. Extensions of Bounded Variation

Noncompact Intervals

We can extend the notion of bounded variation to noncompact intervals:

**XIV.16.4.1. Definition.** Let $I$ be an interval, and $f$ a real-valued function on $I$. Define

$$V^+_I(f) = \sup_{[a,b]} V^+_{[a,b]}(f)$$
$$V^-_I(f) = \sup_{[a,b]} V^-_{[a,b]}(f)$$
$$V_I(f) = \sup_{[a,b]} V_{[a,b]}(f)$$

where the suprema are taken over all closed bounded subintervals $[a,b]$ of $I$.

The function $f$ has **bounded variation on $I$** if $V_I(f) < +\infty$.

Most properties of bounded variation for closed bounded intervals carry over directly. We list some without proof; proofs are similar to the earlier ones and are left to the reader.

**XIV.16.4.2. Proposition.**

(i) $V_I(f) = V^+_I(f) + V^-_I(f)$.

(ii) If $f$ has bounded variation on $I$, then $f$ is bounded on $I$ and

$$\sup_{x \in I} f(x) - \inf_{x \in I} f(x) \leq V_I(f).$$

(iii) If $J$ is a subinterval of $I$, then $V^+_J(f) \leq V^+_I(f)$, $V^-_J(f) \leq V^-_I(f)$, $V_J(f) \leq V_I(f)$. If $f$ has bounded variation on $I$, then $f$ has bounded variation on $J$.

(iv) If $f$ is bounded variation on $I$, and $a$ and $b$ are the left and right endpoints of $I$ (we do not necessarily have $a, b \in I$; we can have $a = -\infty$ and/or $b = +\infty$), then $\lim_{x \to a^+} f(x)$ and $\lim_{x \to b^-} f(x)$ exist.

(v) If $f$ has bounded variation on $I$, then $f$ is Borel measurable, has left and right limits at every point, has only countably many discontinuities each of which is a jump discontinuity, and is differentiable a.e. on $I$.

(vi) If $f$ has bounded variation on $I$, then $f'$ is Lebesgue integrable on $I$ and

$$\int_I |f'| \, d\lambda \leq V_I(f).$$

We can also define variation functions for a function of bounded variation:
**XIV.16.4.3. Definition.** Let \( I \) be an interval, and \( a \) the left endpoint of \( I \) (we do not necessarily have \( a \in I \); \( a \) can be \(-\infty\)). Define

\[
T^+ f(x) = \begin{cases} \mathcal{V}_{[a,x]}(f) & \text{if } a \in I \\ \mathcal{V}^+_{[a,x]}(f) & \text{if } a \notin I \end{cases}
\]

\[
T^- f(x) = \begin{cases} \mathcal{V}_{[a,x]}(f) & \text{if } a \in I \\ \mathcal{V}_{(a,x]}(f) & \text{if } a \notin I \end{cases}
\]

\[
T f(x) = \begin{cases} \mathcal{V}_{[a,x]}(f) & \text{if } a \in I \\ \mathcal{V}_{(a,x]}(f) & \text{if } a \notin I \end{cases}
\]

for \( x \in I \).

\( T^+ f, \ T^- f, \ Tf \) are nonnegative and nondecreasing; \( Tf = T^+ f + T^- f \), \( T^+ f - T^- f = f - C \), where \( C = f(a) \) if \( a \in I \) and \( C = \lim_{x \rightarrow a^+} f(x) \) if \( a \notin I \).

**Complex-Valued Functions**

**XIV.16.4.4.** We can also define bounded variation for complex-valued functions. If \( f \) is a complex-valued function on \( [a,b] \) and \( P = \{a = x_0, \ldots, x_n = b\} \), then \( V^+(f, P) \) and \( V^-(f, P) \) do not make sense, but we can still define

\[
V(f, P) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|
\]

\[
V_{[a,b]}(f) = \sup_P V(f, P)
\]

as before. Then \( f \) has bounded variation on \( [a,b] \) if \( V_{[a,b]}(f) < +\infty \).

**XIV.16.4.5.** If \( g \) and \( h \) are the real and imaginary parts of \( f \), i.e. \( f = g + ih \) with \( g, h \) real-valued, then we have

\[
V_{[a,b]}(g), V_{[a,b]}(h) \leq V_{[a,b]}(f) \leq V_{[a,b]}(g) + V_{[a,b]}(h)
\]

so \( f \) has bounded variation if and only if both \( g \) and \( h \) have bounded variation. (There is, however, no simple formula for \( V_{[a,b]}(f) \) in terms of \( V_{[a,b]}(g) \) and \( V_{[a,b]}(h) \).) We still have:

**XIV.16.4.6. Proposition.** Let \( f \) be a complex-valued function on \( [a,b] \), and \( c \in (a,b) \). Then

\[
V_{[a,b]}(f) = V_{[a,c]}(f) + V_{(c,b]}(f)
\]

**XIV.16.4.7. Proposition.** If \( f \) is a complex-valued function of bounded variation on \( [a,b] \), then \( f \) is differentiable a.e. on \( [a,b] \), \( f' \) is integrable on \( [a,b] \), and

\[
\int_a^b |f'| d\lambda \leq V_{[a,b]}(f)
\]
XIV.16.4.8. We can also define the total variation function \( T_f \) as in the real case, which is nonnegative and nondecreasing, and continuous wherever \( f \) is continuous. If \( g \) and \( h \) are the real and imaginary parts of \( f \), then
\[
T_g, Th \leq T_f \leq Tg + Th
\]
but there is no simple formula for \( T_f \) in terms of \( Tg \) and \( Th \).

XIV.16.5. Exercises

XIV.16.5.1. Let \( f \) be a function of bounded variation on \([a, b]\). Show directly that \( f \) has one-sided limits at every point of \([a, b]\), as follows:
(a) Let \( c \in (a, b) \). Suppose
\[
\liminf_{x \to c-} f(x) < \limsup_{x \to c-} f(x).
\]
Choose \( \alpha, \beta \in \mathbb{R} \) such that
\[
\liminf_{x \to c-} f(x) < \alpha < \beta < \limsup_{x \to c-} f(x).
\]
Choose an increasing sequence \((x_n)\) in \((a, c)\) converging to \( c \) such that \( f(x_n) < \alpha \) for \( n \) odd and \( f(x_n) > \beta \) for \( n \) even.
(b) If \( \mathcal{P} = \{a = x_0, x_1, \ldots, x_{n+1}, b\} \), show that \( V(f, \mathcal{P}) > n(\beta - \alpha) \). Conclude that \( V_{[a,b]}(f) \) must be \(+\infty\).
(c) Give a similar argument for limits from the right.
(d) Conclude from V.8.6.7.(a) that \( f \) has only countably many discontinuities, each of which is a jump discontinuity.

XIV.16.5.2. Let \( f \) be the function of XIV.2.9.10.(i). Show that \( f \) does not have bounded variation on \([0, 1]\) (or on \([0, b]\) for any \( b > 0 \)). Note that \( f \) is differentiable everywhere, and \( f' \) is continuous except at 0.
XIV.17. Absolute Continuity and Lebesgue Density

A function which is differentiable everywhere on an interval \([a, b]\) need not have a (Lebesgue) integrable derivative, hence may not be the integral of its derivative (XIV.2.9.10.(i)); the problem is that such a function need not have bounded variation. But even bounded variation is not sufficient to insure that a continuous function can be recovered from its derivative by integration if it is only differentiable almost everywhere. The Cantor function is the prototype example of a continuous monotone function (hence of bounded variation) which is differentiable almost everywhere, yet is not the integral of its derivative. In this section we discuss conditions which do suffice to recover a function by integration.

Applications of the results to indicator functions demonstrate a remarkable density property of Lebesgue measurable sets: such sets cannot be too uniformly distributed in the real line and must be either thick or sparse almost everywhere.

XIV.17.1. Luzin’s Property \((N)\)

XIV.17.1.1. The principal property of the Cantor function \(f\) which causes difficulties in the context of this section is that there are null sets \(N\) for which \(f(N)\) is not a null set. In fact, the Cantor set \(K\) is a null set and \(f(K)\) is the entire interval \([0, 1]\).

From the point of view of Borel measures, the measure on \([0, 1]\) whose cdf is the Cantor function is a probability measure which is continuous, but concentrated on \(K\), and hence is not absolutely continuous with respect to Lebesgue measure.

We thus make the following definition. This condition was first defined and studied by N. Luzin, who appreciated its importance and relevance to the problem of reconstructing a function by integration.

XIV.17.1.2. Definition. Let \(I\) be an interval, and \(f : I \to \mathbb{R}\). Then \(f\) has Property \((N)\) (or Luzin’s Property \((N)\)) on \(I\) if, whenever \(N\) is a subset of \(I\) with \(\lambda(N) = 0\), we have \(\lambda(f(N)) = 0\) (i.e. \(f\) sends \(\lambda\)-null sets to \(\lambda\)-null sets).

The next result is an immediate rephrasing of the definition.

XIV.17.1.3. Proposition. Let \(\mu\) be a continuous finite signed Borel measure on \([a, b]\), with cdf \(F\). Then \(\mu \ll \lambda\) if and only if \(F\) has property \((N)\).

XIV.17.1.4. Proposition. [?] Let \(I\) be an interval, and \(f : I \to \mathbb{R}\) a continuous function. Then \(f\) has property \((N)\) if and only if, whenever \(A\) is a Lebesgue measurable subset of \(I\), \(f(A)\) is Lebesgue measurable.

Proof: Suppose \(f\) has property \((N)\). If \(A \subseteq I\) is Lebesgue measurable, then \(A = E \cup N\), where \(E\) is an \(F_\sigma\) and \(N\) a null set (XIII.3.4.5.(iv)). Then \(f(E)\) is an \(F_\sigma\) since \(f\) is continuous (XI.11.7.11.), and \(f(N)\) is a null set since \(f\) has Property \((N)\); hence \(f(A) = f(E) \cup f(N)\) is Lebesgue measurable. Conversely, if \(N\) is a null set and \(f(N)\) is not a null set (i.e. \(f(N)\) has positive Lebesgue outer measure), then \(f(N)\) contains a nonmeasurable set \(B\) (XIII.3.9.7.). If \(A = f^{-1}(B)\), then \(A \subseteq N\) and hence \(A\) is Lebesgue measurable, and \(f(A) = B\) is nonmeasurable.

\(\Box\)
XIV.17.1.5. Property \((N)\) is complementary to bounded variation, even for continuous functions: the Cantor function is continuous and monotone, hence has bounded variation, but not Property \((N)\), and there are continuous functions with Property \((N)\) which do not have bounded variation, and are even not differentiable a.e. (XIV.17.5.1).

There is an interesting variation on Property \((N)\), due to Banach:

XIV.17.1.6. Definition. Let \(I\) be an interval, and \(f : I \to \mathbb{R}\). Then \(f\) has Property \((S)\) (or Banach’s Property \((S)\)) on \(I\) if for every \(\epsilon > 0\) there is a \(\delta > 0\) such that, whenever \(A\) is a Lebesgue measurable subset of \(I\) with \(\lambda(A) < \delta\), we have \(\lambda^*(f(A)) < \epsilon\).

XIV.17.1.7. A function with Property \((S)\) obviously has Property \((N)\); hence, if \(f\) is continuous with Property \((S)\), it sends measurable sets to measurable sets, and \(\lambda^*\) can be replaced by \(\lambda\) in the definition. There are continuous functions with Property \((N)\) which do not have Property \((S)\) (XIV.17.5.2).

In fact, we have:

XIV.17.1.8. Theorem. Let \(I\) be an interval, and \(f : I \to \mathbb{R}\) a continuous function. Then \(f\) has property \((S)\) if and only if it has Property \((N)\) and \(N_f()\) is finite a.e.

Combining this with XIV.16.3.1, we obtain:

XIV.17.1.9. Corollary. Let \(I\) be an interval, and \(f : I \to \mathbb{R}\) a continuous function of bounded variation. Then \(f\) has property \((S)\) if and only if it has Property \((N)\).

XIV.17.1.10. There are continuous functions with Property \((S)\) which do not have bounded variation, and are even not differentiable a.e. (XIV.17.5.1).

XIV.17.1.11. Banach said a function \(f\) satisfies Property \((T_1)\) if \(N_f\) is finite a.e., and Property \((T_2)\) if \(N_f\) is countable a.e. (i.e. \(f\) takes almost all values only countably many times). We will avoid this terminology since it could be confused with separation axioms of topology.

The next goal is to show that differentiable functions have Property \((N)\). The next lemma is crucial.

XIV.17.1.12. Lemma. Let \(A\) be a Lebesgue measurable subset of \(\mathbb{R}\), and \(f : A \to \mathbb{R}\) a function. Suppose there is a constant \(M\) such that, for every \(x \in A\), there is a \(\delta > 0\) such that \(|f(y) - f(x)| \leq M(y - x)\) for all \(y \in A \cap (x, x + \delta)\). Then \(\lambda^*(f(A)) \leq M\lambda(A)\).

Proof: We first make a reduction. For each \(n\), let \(B_n = \{x \in A : |f(y) - f(x)| \leq M(y - x)\} \) for all \(y \in A \cap (x, x + 1/n)\).

If \(x \in A \cap B_n\), there is a monotone sequence \((x_k)\) in \(B_n\) with \(x_k \to x\), so for sufficiently large \(k\) we have \(|f(x) - f(x_k)| \leq M|x - x_k|\). Thus \(f(x) = \lim_{k \to \infty} f(x_k)\) and, if \(y \in A \cap (x, x + 1/n)\), then for
sufficiently large \( k \) we have \( y \in (x_k, x_k + 1/n) \), so \( |f(y) - f(x_k)| \leq M(y - x_k) \). Taking \( k \to \infty \), we conclude \( x \in B_n \). Thus \( B_n = A \cap B_n \), and in particular \( B_n \) is measurable. We have that \( A = \cup_n B_n \), so \( \lambda(A) = \sup_n \lambda(B_n) \) and \( \lambda^*(f(A)) = \sup_n \lambda^*(f(B_n)) \). Thus it suffices to show \( \lambda^*(f(B_n)) \leq M\lambda(B_n) \) for each \( n \). Fix \( n \); write \( B_n \) as a disjoint union of measurable sets \( A_{nm} \) of diameter \( < 1/n \). Then \( \lambda(B_n) = \sum_m \lambda(A_{nm}) \) and \( \lambda^*(f(B_n)) \leq \sum_m \lambda^*(f(A_{nm})) \). Thus it suffices to show that \( \lambda^*(f(A_{nm})) \leq M\lambda(A_{nm}) \) for all \( n, m \). But \( A_{nm} \) has the property that if \( x, y \in A_{nm}, x < y \), then \( |f(y) - f(x)| \leq M(y - x) \). Replacing \( A \) by \( A_{nm} \), we may assume \( A \) has the property that \( |f(y) - f(x)| \leq M|y - x| \) for all \( x, y \in A \). We may also assume \( \lambda(A) < \infty \).

Let \( \epsilon > 0 \). Fix a sequence \( (I_n) \) of open intervals covering \( A \) with \( \sum_n \ell(I_n) < \lambda(A) + \epsilon \). If \( x, y \in A \cap I_n \), then \( |f(y) - f(x)| \leq M|y - x| \), so \( f(A \cap I_n) \) is contained in an interval \( J_n \) of length \( \leq M\ell(I_n) \). We then have

\[
\lambda^*(f(A)) \leq \sum_{n=1}^{\infty} \lambda^*(f(A \cap I_n)) \leq \sum_{n=1}^{\infty} \ell(J_n) \leq M \sum_{n=1}^{\infty} \ell(I_n) \leq M(\lambda(A) + \epsilon).
\]

This is true for all \( \epsilon > 0 \), and the result follows. \( \diamond \)

**XIV.17.1.13.** **Corollary.** Let \( A \) be a Lebesgue measurable subset of \( \mathbb{R} \), and \( f : A \to \mathbb{R} \) a function. Suppose there is a constant \( M \) such that, for every \( x \in A \), \( f \) is differentiable at \( x \) and \( |f'(x)| \leq M \). Then \( \lambda^*(f(A)) \leq M\lambda(A) \).

**Proof:** Let \( \epsilon > 0 \). For each \( x \in A \), since

\[
\lim_{y \to x} \frac{|f(y) - f(x)|}{|y - x|} \leq M
\]

there is a \( \delta > 0 \) such that

\[
\frac{|f(y) - f(x)|}{|y - x|} \leq M + \epsilon
\]

whenever \( 0 < |y - x| < \delta \). Thus \( \lambda^*(f(A)) \leq (M + \epsilon)\lambda(A) \). Since \( \epsilon > 0 \) is arbitrary, the result follows. \( \diamond \)

As a special case, we obtain the general one-dimensional version of Sard’s Theorem (.), due to LUZIN:

**XIV.17.1.14.** **Corollary.** Let \( I \) be an interval, and \( f : I \to \mathbb{R} \) a function. Set

\[
A = \{ x \in I : f'(x) \text{ exists and equals } 0 \}.
\]

Then \( f(A) \) is a null set.

**Proof:** \( A \) is Lebesgue measurable by (.). We may write \( A = \cup_n A_n \), where \( \lambda(A_n) < \infty \). Then \( \lambda^*(f(A_n)) = 0 \) for all \( n \), and hence \( \lambda^*(f(A)) \leq \sum_n \lambda^*(f(A_n)) = 0 \). \( \diamond \)

Our main result then is:
XIV.17.15. **Theorem.** Let \( B \) be a Lebesgue measurable set, and \( f : B \to \mathbb{R} \) a function which is differentiable on \( B \) except possibly at countably many points of \( B \). If \( A \) is a subset of \( B \) with \( \lambda(A) = 0 \), then \( \lambda(f(A)) = 0 \).

**Proof:** Let \( C \) be a countable subset of \( B \) such that \( f \) is differentiable on \( B \setminus C \). For each \( n \) let
\[
D_n = \{ x \in B \setminus C : |f'(x)| \leq n \} .
\]
Each \( D_n \) is measurable. If \( A \) is a measurable subset of \( B \) with \( \lambda(A) = 0 \), then \( f(A) \) is the union of \( f(A \cap C) \) and \( f(A \cap D_n) \) for all \( n \). \( f(A \cap C) \) is countable, hence has measure 0; and by XIV.17.13. we have \( \lambda^*(f(A \cap D_n)) \leq n \lambda(A \cap D_n) = 0 \). Thus
\[
\lambda^*(f(A)) \leq \lambda^*(A \cap C) + \sum_{n=1}^{\infty} \lambda(A \cap D_n) = 0 .
\]

XIV.17.16. **Corollary.** Let \( I \) be an interval, and \( f : I \to \mathbb{R} \) a function. If \( f \) is differentiable on \( I \) except possibly at countably many points of \( I \), then \( f \) has Property (N) on \( I \).

Property (N) is relevant in extending consequences of the Mean Value Theorem:

XIV.17.17. **Theorem.** Let \( I \) be an interval, and \( f : I \to \mathbb{R} \) a continuous function with Property (N) on \( I \). If \( f \) is differentiable a.e. and \( f' \geq 0 \) a.e. on \( I \), then \( f \) is nondecreasing on \( I \) (and hence \( f' \geq 0 \) everywhere it is defined). Similarly, if \( f' \leq 0 \) a.e. on \( I \), then \( f \) is nonincreasing on \( I \).

**Proof:** If \( f' > 0 \) a.e., the proof is identical to the proof of V.8.4.1 (i), since if \( f \) is differentiable except on a null set \( A \), then \( f(A) \) is also a null set since \( f \) has Property (N) and thus contains no interval (cf. V.8.4.4.). In the general case, we proceed as follows. Set
\[
A = \{ x \in I : f \text{ is not differentiable at } x \}
\]
\[
B = \{ x \in I : f'(x) < 0 \}
\]
\[
C = \{ x \in I : f'(x) = 0 \}
\]
Then \( A \) and \( B \) are null sets by assumption, so \( f(A) \) and \( f(B) \) are null sets since \( f \) has Property (N). And \( f(C) \) is a null set by XIV.17.14. Thus \( f(A \cup B \cup C) \) is a null set, and contains no interval.

Suppose there are \( a, b \in I \) with \( a < b \) and \( f(a) > f(b) \). Then there is an \( r \notin f(A \cup B \cup C) \) with \( f(a) > r > f(b) \). There is an \( x \in [a, b] \) with \( f(x) = r \) by the Intermediate Value Theorem. Set
\[
c = \sup \{ x \in [a, b] : f(x) = r \} .
\]
Then \( f(c) = r \) by continuity, and \( c < b \). Then \( c \notin A \cup B \cup C \) since \( r = f(c) \notin f(A \cup B \cup C) \), so \( f \) is differentiable at \( c \) and \( f'(c) > 0 \). We have \( c < b \) since \( f(b) < r \), and if \( c < x \leq b \) we have \( f(x) < f(c) = r \), so \( \frac{f(x) - f(c)}{x - c} < 0 \). Thus
\[
f'(c) = \lim_{x \to c^+} \frac{f(x) - f(c)}{x - c} \leq 0
\]

1711
contradicting that $f'(c) > 0$.

Combining this result with XIV.17.16., we obtain:

**XIV.17.18. Corollary.** Let $I$ be an interval, and $f : I \to \mathbb{R}$ a continuous function, which is differentiable on $I$ except possibly at countably many points. If $f' \geq 0$ a.e. on $I$, then $f$ is nondecreasing on $I$ (and hence $f' \geq 0$ everywhere it is defined). Similarly, if $f' \leq 0$ a.e. on $I$, then $f$ is nonincreasing on $I$. If $f' = 0$ a.e. on $I$, then $f$ is constant on $I$.

Here is an interesting consequence of XIV.17.14.:

**XIV.17.19. Corollary.** Let $[a, b]$ be a closed bounded interval, and $f : [a, b] \to \mathbb{R}$ a continuous function. If $f$ is differentiable on $[a, b]$ except possibly at countably many points, then $N_f$ is finite a.e. on $[a, b]$, so $f$ has Property (S) on $[a, b]$.

**Proof:** Let $C$ be a countable set such that $f$ is differentiable on $[a, b] \setminus C$. Let

$$A = \{x \in [a, b] : f'(x) = 0\}.$$ 

By XIV.17.14., $f(A)$ has measure 0. Since $C$ is countable, $f(A \cup C)$ has measure 0.

Suppose $N_f(y) = \infty$. Then $f^{-1}(\{y\})$ is an infinite subset of $[a, b]$, hence has an accumulation point $x$. If $f'(x)$ exists, it must be zero; so $x \in A \cup C$, and hence $y \in f(A \cup C)$. So if $B = \{y : N_f(y) = \infty\}$, then $B \subseteq f(A \cup C)$, $\lambda(B) = 0$. The last statement follows from XIV.17.18.

**XIV.17.2. Absolutely Continuous Functions**

Our basic definition is a refinement of uniform continuity:

**XIV.17.2.1. Definition.** Let $[a, b]$ be a closed bounded interval, and $f : [a, b] \to \mathbb{R}$ a function. Then $f$ is absolutely continuous on $[a, b]$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are finitely many nonoverlapping intervals in $[a, b]$ (nonoverlapping means distinct intervals may have endpoints but not interior points in common) with total length $< \delta$, i.e.

$$\sum_{k=1}^{n} (b_k - a_k) < \delta$$

we have

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \epsilon .$$

Note that the definition of uniform continuity is the same, but with only one subinterval allowed.
**XIV.17.2.2.** Although it will not be necessary, we may assume the intervals \([a_1, b_1], \ldots, [a_n, b_n]\) are in increasing order. The nonoverlapping condition then means
\[
a \leq a_1 < b_1 \leq a_2 < b_2 \leq \cdots < b_n \leq b .
\]
There is also no harm, and some technical advantage, in allowing some intervals to be degenerate, i.e. \(a_k = b_k\), since such degenerate intervals make no contribution to the sums \(\sum_{k=1}^{n} (b_k - a_k)\) and \(\sum_{k=1}^{n} |f(b_k) - f(a_k)|\).
We may also allow infinite collections of subintervals:

**XIV.17.2.3.** Proposition. Let \([a, b]\) be a closed bounded interval, and \(f : [a, b] \to \mathbb{R}\) a function. Then \(f\) is absolutely continuous on \([a, b]\) if and only if for every \(\epsilon > 0\) there is a \(\delta > 0\) such that, whenever \(\{[a_k, b_k] : k \in \mathbb{N}\}\) is a sequence of pairwise nonoverlapping subintervals (not necessarily nondegenerate) of \([a, b]\) with
\[
\sum_{k=1}^{\infty} (b_k - a_k) < \delta
\]
we have
\[
\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \epsilon .
\]

**Proof:** If \(f\) satisfies this condition, it is clearly absolutely continuous (take all but finitely many of the intervals to be degenerate). Conversely, suppose \(f\) is absolutely continuous. Let \(\epsilon > 0\), and choose \(\delta > 0\) corresponding to \(\frac{\epsilon}{2}\) in the definition of absolute continuity. If \(\{[a_k, b_k] : k \in \mathbb{N}\}\) is a sequence of pairwise nonoverlapping subintervals (not necessarily nondegenerate) of \([a, b]\) with
\[
\sum_{k=1}^{\infty} (b_k - a_k) < \delta
\]
we have, for each \(n\),
\[
\sum_{k=1}^{n} (b_k - a_k) < \delta
\]
and therefore
\[
\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \frac{\epsilon}{2} .
\]
Thus
\[
\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| \leq \frac{\epsilon}{2} < \epsilon .
\]
\(\Diamond\)
Our main result, which will be proved in stages, is the following theorem giving various characterizations of absolute continuity, which shows that absolutely continuous functions are precisely the functions which can be recovered from their derivatives by integration.

**XIV.17.2.4. Theorem.** Let \([a, b]\) be a closed bounded interval, and \(F : [a, b] \rightarrow \mathbb{R}\) a function. The following are equivalent:

(i) \(F\) is absolutely continuous on \([a, b]\).

(ii) \(F\) is continuous, has bounded variation, and property \((N)\) on \([a, b]\).

(iii) There is a Lebesgue integrable function \(f : [a, b] \rightarrow \mathbb{R}\) such that \(F\) is the indefinite integral of \(f\) on \([a, b]\), i.e.

\[
F(x) = F(a) + \int_a^x f \, d\lambda
\]

for all \(x \in [a, b]\).

(iv) \(F\) is differentiable a.e. on \([a, b]\), \(F'\) is integrable on \([a, b]\), and \(F\) is the indefinite integral of \(F'\) on \([a, b]\), i.e.

\[
F(x) = F(a) + \int_a^x F' \, d\lambda
\]

for all \(x \in [a, b]\).

(v) \(F\) has bounded variation and is right continuous on \([a, b]\), and the signed Borel measure on \([a, b]\) with cdf \(F()\) is absolutely continuous with respect to Lebesgue measure.

**XIV.17.2.5. Corollary.** [Fundamental Theorem of Calculus, Lebesgue Version I] Let \([a, b]\) be a closed bounded interval, and \(f : [a, b] \rightarrow \mathbb{R}\) a Lebesgue integrable function. Let

\[
F(x) = \int_a^x f \, d\lambda
\]

for all \(x \in [a, b]\). Then \(F\) is differentiable a.e. and \(F' = f\) a.e. on \([a, b]\).

**XIV.17.2.6. Corollary.** Let \([a, b]\) be a closed bounded interval, and \(F : [a, b] \rightarrow \mathbb{R}\) a function. Then \(F\) is Lipschitz on \([a, b]\) if and only if \(F\) is absolutely continuous and \(F'\) is bounded on \([a, b]\). More precisely, \(F\) is Lipschitz on \([a, b]\) with Lipschitz constant \(M\) if and only if \(F\) is absolutely continuous, hence differentiable a.e., on \([a, b]\) and \(|F'(x)| \leq M\) for all \(x \in [a, b]\) for which \(F'(x)\) exists.

In fact, the smallest Lipschitz constant for \(F\) on \([a, b]\) is exactly \(\sup_{a \leq x \leq b} |F'(x)|\) (supremum over those \(x\) where \(F\) is differentiable).
XIV.17.2.7. Corollary. Let $F$ be continuous on $[a, b]$ and differentiable except possibly on a countable subset of $[a, b]$. Then the following are equivalent:

(i) $F$ is absolutely continuous on $[a, b]$.

(ii) $F$ has bounded variation on $[a, b]$.

(iii) $F'$ is Lebesgue integrable on $[a, b]$.

XIV.17.2.8. In XIV.17.2.7., “differentiable except possibly on a countable subset” cannot be weakened to “differentiable a.e.” The Cantor function is a counterexample. The conditions can fail, even if $F$ is differentiable everywhere on $[a, b]$ (XIV.2.9.10.(i)).

Characterizing Functions as Derivatives

Theorem XIV.17.2.4. has another interesting consequence: integrable functions can be characterized without any explicit reference to integration.

XIV.17.2.9. Corollary. Let $f$ be a Lebesgue measurable real-valued function on an interval $[a, b]$. Then $f$ is Lebesgue integrable on $[a, b]$ if and only if there is an absolutely continuous function $F$ on $[a, b]$ with $F' = f$ a.e. on $[a, b]$.

If the word “absolutely” is removed from the statement, the result is false. In fact, it is remarkable that every Lebesgue measurable function is a.e. the derivative of a continuous function. This result is due to Luzin; our proof is adapted from \cite{Sak64}.

XIV.17.2.10. Theorem. Let $f$ be a Lebesgue measurable real-valued function on an interval $[a, b]$. Then there is a continuous function $F$ on $[a, b]$ which is differentiable a.e. and $F' = f$ a.e. In fact, if $G$ is any continuous function on $[a, b]$, for any $\epsilon > 0$ there is a continuous function $F$ on $[a, b]$ which is differentiable a.e., $F' = f$ a.e., and $|F(x) - G(x)| < \epsilon$ for all $x \in [a, b]$.

The proof is a little complicated technically; we begin with an observation and two lemmas.

XIV.17.2.11. Let $\epsilon > 0$. If $G$ is continuous on $[a, b]$, there is a continuous function $H$ on $[a, b]$ which is differentiable a.e. and $H' = 0$ a.e. which is uniformly within \(\frac{\epsilon}{2}\) of $G$ (XV.8.3.6.). If $F_0$ is a continuous function on $[a, b]$ which is differentiable a.e. with $F_0' = f$ a.e., with $|F_0(x)| < \frac{\epsilon}{2}$ for all $x \in [a, b]$, then $F = F_0 + H$ is the desired function. Thus it suffices to show that if $\epsilon > 0$, there is a continuous function $F$ on $[a, b]$ which is differentiable a.e., with $F' = f$ a.e. and $|F(x)| < \epsilon$ for all $x \in [a, b]$.

If $f$ is integrable, this is fairly easily accomplished with judicious use of the Cantor function (this is essentially an explicit version of XV.8.3.6., cf. Exercise XIV.17.5.4.).
**XIV.17.2.12. Lemma.** Let \( \phi \) be an integrable function on an interval \([a,b]\). Then for any \( \epsilon > 0 \) there is a continuous function \( \Phi \) on \([a,b]\) with the following properties:

(i) \( \Phi \) is differentiable a.e. and \( \Phi' = \phi \) a.e. on \([a,b]\).

(ii) \( \Phi(a) = \Phi(b) = 0 \).

(iii) \( |\Phi(x)| < \epsilon \) for all \( x \in [a,b] \).

**Proof:** For \( x \in [a,b] \), set \( \Psi(x) = \int_a^x \phi \, d\lambda \). Then \( \Psi \) is absolutely continuous on \([a,b]\), \( \Psi' = \phi \) a.e. on \([a,b]\), and \( \Psi(a) = 0 \). Fix \( \epsilon > 0 \). Since \( \Psi \) is uniformly continuous on \([a,b]\), there is a \( \delta > 0 \) such that \( |\Psi(x) - \Psi(y)| < \epsilon \) whenever \( x, y \in [a,b] \) and \( |x - y| < \delta \). Let \( P = \{a = x_0, \ldots, x_n = b\} \) be a partition of \([a,b]\) into subintervals of length \( < \delta \). On \([x_{k-1}, x_k]\), let \( H_k \) be a copy of the Cantor Function scaled to \([x_{k-1}, x_k]\) with \( H_k(x_{k-1}) = \Psi(x_{k-1}) \) and \( H_k(x_k) = \Psi_k(x_k) \). Define \( \Phi(x) = \Psi(x) - H_k(x) \) if \( x_{k-1} \leq x \leq x_k \). Then \( \Phi(x_k) = 0 \) for all \( k \), and hence \( \Phi \) is well defined and continuous; and \( \Phi' = \Psi' = \phi \) a.e. since \( H'_k = 0 \) a.e. for each \( k \). If \( x \in [x_{k-1}, x_k] \), we have \( |\Psi(x) - \Psi(x_{k-1})| < \epsilon \) and \( |\Psi(x) - \Psi(x_k)| < \epsilon \); since \( H_k(x) \) is between \( H_k(x_{k-1}) = \Psi(x_{k-1}) \) and \( H_k(x_k) = \Psi(x_k) \), we also have \( |\Phi(x)| = |\Psi(x) - H_k(x)| < \epsilon \).

To handle the more difficult case where \( f \) is not integrable, we need the following technical lemma:

**XIV.17.2.13. Lemma.** Let \( \phi \) be an integrable function on an interval \([a,b]\), and \( C \) a closed subset of \([a,b]\). Then for any \( \epsilon > 0 \) there is a continuous function \( \Phi \) on \([a,b]\) with the following properties:

(i) \( \Phi \) is differentiable a.e. and \( \Phi' = \phi \) a.e. on \([a,b] \setminus C\).

(ii) \( \Phi(x) = \Phi'(x) = 0 \) for all \( x \in C \).

(iii) \( |\Phi(x + h)| = |\Phi(x) + h - \Phi(x)| < \epsilon |h| \) whenever \( x \in C \) and \( x + h \in [a,b] \).

**Proof:** Fix \( \epsilon > 0 \). The open set \((a,b) \setminus C\) is a disjoint union of open intervals \((a_k, b_k)\). For each \( k \), fix a set \( \{x_{k,n} : n \in \mathbb{Z}\} \) with \( x_{k,n} < x_{k,n+1} \) for all \( n \), \( \lim_{n \to -\infty} x_{k,n} = a_k \), and \( \lim_{n \to +\infty} x_{k,n} = b_k \). For each \( k \) and \( n \), let \( \Phi_{k,n} \) be a continuous function on \([x_{k,n}, x_{k,n+1}]\) such that

(i) \( \Phi_{k,n} \) is differentiable a.e. and \( \Phi'_{k,n} = \phi \) a.e. on \([x_{k,n}, x_{k,n+1}]\).

(ii) \( \Phi_{k,n}(x_{k,n}) = \Phi_{k,n}(x_{k,n+1}) = 0 \).

(iii) \( |\Phi_{k,n}(x)| < \min \left( \frac{\epsilon (x_{k,n} - a_k)}{|k| + |n|}, \frac{\epsilon (b_k - x_{k,n+1})}{|k| + |n|} \right) \) for all \( x \in [x_{k,n}, x_{k,n+1}] \).

(Lemma XIV.17.2.12). Define \( \Phi \) on \([a,b]\) by \( \Phi(a) = \Phi(b) = 0 \), \( \Phi(x) = 0 \) for \( x \in C \), and \( \Phi(x) = \Phi_{k,n}(x) \) if \( x \in [x_{k,n}, x_{k,n+1}] \). Then \( \Phi \) is well defined and continuous, and it is routine to check it has the other required properties (Exercise XIV.17.5.3.).

1716
XIV.17.2.14. We now prove the theorem (in the reduced form $G = 0$). Fix $\epsilon > 0$. Inductively for $n \geq 0$ construct continuous functions $\Phi_n$ which are differentiable a.e. and closed subsets $(C_n)$ of $[a, b]$ with $C_n \subseteq C_{n+1}$ for all $n$, such that, with $F_n = \sum_{k=1}^{n} \Phi_k$, the following conditions are satisfied for $n \geq 1$:

(i) $F_n$ is differentiable at all points of $C_n$ and $F_n' = f$ on $C_n$.

(ii) $\Phi_n(x) = \Phi_n'(x) = 0$ for all $x \in C_{n-1}$.

(iii) $|\Phi_n(x + h)| = |\Phi_n(x) - \Phi_n(x)| \leq \frac{\epsilon |h|}{2^n(b-a)}$ whenever $x \in C_{n-1}$ and $x + h \in [a, b]$.

(iv) $\lambda([a, b] \setminus C_n) < \frac{1}{n}$.

Begin with $C_0 = \emptyset$ and $\Phi_0 = 0$. Suppose $\Phi_k$ and $C_k$ have been chosen for $k \leq n$. Since $F_n'$ and $f$ are defined a.e., for some sufficiently large $M$ the measurable set

$$A_n = \{x \in [a, b] \setminus C_n : F_n'(x) \text{ and } f(x) \text{ exist and } |F_n'(x)|, |f(x)| \leq M\}$$

satisfies $\lambda([a, b] \setminus C_n) < \frac{1}{n+1}$.

Define $\phi_n$ on $[a, b]$ by $\phi_n = f - F_n'$ on $A_n$ and $\phi_n = 0$ on $[a, b] \setminus A_n$. Then $\phi_n$ is bounded, hence integrable, so by Lemma XIV.17.2.13. there is a continuous function $\Phi_{n+1}$ on $[a, b]$ which is differentiable a.e., $\Phi_{n+1}'(x) = f(x) - F_n'(x)$ for almost all $x \in A_n$, $\Phi_{n+1}(x) = \Phi_{n+1}'(x) = 0$ for all $x \in C_n$, and

$$|\Phi_{n+1}(x + h)| < \frac{\epsilon |h|}{2^{n+1}(b-a)}$$

if $x \in C_n$ and $x + h \in [a, b]$.

By (), there is a closed subset $B_n$ of $A_n$ such that $\Phi_{n+1}'(x) = f(x) - F_n'(x)$ for all $x \in B_n$ and $\lambda([0, 1] \setminus (C_n \cup B_n)) < \frac{1}{n+1}$. Set $C_{n+1} = C_n \cup B_n$. This completes the inductive construction, and properties (i)–(iv) obviously hold.

From (iii), we have $|\Phi_n(x)| < \frac{\epsilon}{2^n}$ for all $x \in [a, b]$; thus the series

$$\sum_{n=1}^{\infty} \Phi_n$$

converges uniformly to a continuous function $F$ on $[a, b]$ and $|F(x)| < \epsilon$ for all $x \in [a, b]$. Set $C = \cup_{n=1}^{\infty} C_n$. Then $\lambda([a, b] \setminus C) = 0$, and if $x \in C$, then $x \in C_n$ for sufficiently large $n$; thus for any $h$ with $x + h \in [a, b]$ we have

$$\left| \frac{F(x + h) - F(x)}{h} - \frac{F_n(x + h) - F_n(x)}{h} \right| = \left| \sum_{k=n+1}^{\infty} \frac{\Phi_k(x + h) - \Phi_k(x)}{h} \right| < \frac{\epsilon}{2^n(b-a)}$$

for sufficiently large $n$, and hence

$$\limsup_{h \to 0} \left| \frac{F(x + h) - F(x)}{h} - f(x) \right| \leq \frac{\epsilon}{2^n(b-a)}$$

for all sufficiently large $n$, i.e. $F'(x) = f(x)$.

This completes the proof of Theorem XIV.17.2.10. \( \square \)
XIV.17.3. $L^p$ Versions, $p > 1$

There is an $L^p$ analog of absolute continuity for $p > 1$. Surprisingly, the distinction between absolute continuity and bounded variation disappears in this case. This section is adapted from [?].

XIV.17.3.1. **Theorem.** Let $F$ be a function from $[a, b]$ to $\mathbb{R}$, and $p > 1$. The following are equivalent:

(i) There is a measurable function $f$ on $[a, b]$ with $\int_a^b |f|^p \, d\lambda < \infty$ (i.e. $f \in L^p([a, b])$) and

$$F(x) = F(a) + \int_a^x f \, d\lambda$$

for all $x \in [a, b]$.

(ii) There is a constant $M$ such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each $k$, we have

$$\sum_{k=1}^n |F(b_k) - F(a_k)|^p h_k^{1-p} \leq M.$$

(iii) There is a constant $M$ such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each $k$, we have

$$\sum_{k=1}^n |F(b_k) - F(a_k)|^p h_k^{1-p} \leq M$$

and, for every $\epsilon > 0$, there is a $\delta > 0$ such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each $k$ and $\sum_{k=1}^n h_k < \delta$,

$$\sum_{k=1}^n |F(b_k) - F(a_k)|^p h_k^{1-p} < \epsilon.$$

If these conditions are satisfied, the smallest $M$ in (ii) and (iii) is precisely $\int_a^b |f|^p \, d\lambda$.

XIV.17.3.2. The conditions in (iii) are apparently more restrictive than in (ii). The condition in (i) implies that $f$ is integrable on $[a, b]$, and hence that $F$ is absolutely continuous on $[a, b]$.

If $p = 1$, the condition in (ii) is precisely that $F$ have bounded variation on $[a, b]$, and the second condition in (iii) is that $F$ be absolutely continuous on $[a, b]$, which implies that $F$ has bounded variation and hence satisfies the first condition of (iii). Thus Theorem XIV.17.2.4. says that (i) and (iii) are equivalent also if $p = 1$; however, (ii) is not equivalent to (i) and (iii) in general if $p = 1$ (the Cantor function is a counterexample).
XIV.17.3.3. We now give the proof of XIV.17.3.1.

Proof: (iii) ⇒ (ii) is trivial.

(i) ⇒ (iii): By (), for each \( k \) we have

\[ |F(b_k) - F(a_k)| = \left| \int_{a_k}^{b_k} f \, d\lambda \right| \leq \left[ \int_{a_k}^{b_k} |f|^p \, d\lambda \right]^{1/p} \left[ \int_{a_k}^{b_k} 1 \, d\lambda \right]^{1-1/p} = h_k^{1-1/p} \left[ \int_{a_k}^{b_k} |f|^p \, d\lambda \right]^{1/p} \]

so

\[ \sum_{k=1}^{n} |F(b_k) - F(a_k)| h_k^{-p} \leq \sum_{k=1}^{n} h_k^{1-p} \left[ h_k^{1-1/p} \right]^{p} \left[ \int_{a_k}^{b_k} |f|^p \, d\lambda \right] = \sum_{k=1}^{n} \left[ \int_{a_k}^{b_k} |f|^p \, d\lambda \right] \leq \int_{a}^{b} |f|^p \, d\lambda. \]

Also, by (), for any \( \epsilon > 0 \) there is a \( \delta > 0 \) such that if \( \sum_{k=1}^{n} h_k < \delta \),

\[ \sum_{k=1}^{n} \left[ \int_{a_k}^{b_k} |f|^p \, d\lambda \right] < \epsilon \]

so condition (iii) is satisfied with \( M = \int_{a}^{b} |f|^p \, d\lambda \).

(ii) ⇒ (i): Let \([a_1, b_1], \ldots, [a_n, b_n]\) be nonoverlapping subintervals of \([a, b]\). Then, by Hölder’s Inequality (), we have

\[ \sum_{k=1}^{n} |F(b_k) - F(a_k)| = \sum_{k=1}^{n} \left[ |F(b_k) - F(a_k)| h_k^{-p+1} \right]^{1/p} h_k^{1-1/p} = \sum_{k=1}^{n} \left[ |F(b_k) - F(a_k)| h_k^{1-1/p} \right]^{1/p} h_k^{1-1/p} \]

\[ \leq \left( \sum_{k=1}^{n} \left[ |F(b_k) - F(a_k)| h_k^{1-1/p} \right] \right)^{1/p} \left( \sum_{k=1}^{n} h_k \right)^{1-1/p} \leq M^{1/p} \left( \sum_{k=1}^{n} h_k \right)^{1-1/p} \]

and thus, for any \( \epsilon > 0 \), there is a \( \delta > 0 \) such that if \( \sum_{k=1}^{n} h_k < \delta \), \( \sum_{k=1}^{n} |F(b_k) - F(a_k)| < \epsilon \). Thus \( F \) is absolutely continuous. Hence there is an integrable function \( f \) on \([a, b]\) with

\[ F(x) = F(a) + \int_{a}^{x} f \, d\lambda \]

for all \( x \in [a, b] \). It remains to show that \( \int_{a}^{b} |f|^p \, d\lambda \leq M \).

For each \( n \) let \( \mathcal{P}_n = \{ a = x_{n,0}, x_{n,1}, \ldots, x_{n,n} = b \} \) be a partition of \([a, b]\) into \( n \) subintervals of equal length. For \( x_{n,k} \leq x < x_{n,k+1} \), set

\[ f_n(x) = \frac{F(x_{n,k+1}) - F(x_{n,k})}{x_{n,k+1} - x_{n,k}}. \]

Since \( F \) is absolutely continuous, it is differentiable a.e. and \( F' = f \) a.e. Suppose \( x_{n,k} < x < x_{n,k+1} \) and \( F' \) is differentiable at \( x \). Then

\[ f_n(x) = \frac{F(x_{n,k+1}) - F(x)}{x_{n,k+1} - x} \cdot \frac{x_{n,k+1} - x}{x_{n,k+1} - x_{n,k}} + \frac{F(x) - F(x_{n,k})}{x - x_{n,k}} \cdot \frac{x - x_{n,k}}{x_{n,k+1} - x_{n,k}} \]

\[ = (F'(x) + \eta_1) \frac{x_{n,k+1} - x}{x_{n,k+1} - x_{n,k}} + (F'(x) + \eta_2) \frac{x - x_{n,k}}{x_{n,k+1} - x_{n,k}} = F'(x) + \eta_3 \]

1719
where \( \eta_1, \eta_2, \eta_3 \) depend on \( x \) and \( n \); for fixed \( x \) and \( n \) we have \( |\eta_3| \leq |\eta_1| + |\eta_2| \), and for fixed \( x \) we have \( \eta_1, \eta_2 \to 0 \) as \( n \to \infty \). So if \( x \) is not one of the \( x_{n,k} \) for any \( n \) and \( F \) is differentiable at \( x \), \( f_n(x) \to F'(x) \).

Thus \( f_n \to f \) a.e. By Fatou’s Lemma (),

\[
\int_a^b |f|^p \, d\lambda \leq \liminf_{n \to \infty} \int_a^b |f_n|^p \, d\lambda = \liminf_{n \to \infty} \sum_{k=1}^n \frac{|F(x_{n,k}) - F(x_{n,k-1})|^p}{(x_{n,k} - x_{n,k-1})^p} (x_{n,k} - x_{n,k-1}) \leq M.
\]

\[\blacklozenge\]

**XIV.17.4. The Lebesgue Density Theorem**

If \( A \) and \( B \) are Lebesgue measurable subsets of \( \mathbb{R} \), with \( 0 < \lambda(B) < \infty \), it makes sense to interpret \( \frac{\lambda(A \cap B)}{\lambda(B)} \) as the “fraction of the points of \( B \) which are in \( A \).” We will particularly make this interpretation when \( B \) is a finite interval.

If \( A \) is a measurable subset of \( \mathbb{R} \) and \( c \in \mathbb{R} \), we similarly want to carefully define “the fraction of points near \( c \) which are in \( A \).” There is a way to do this, and it leads to a remarkable result (XIV.17.4.3.): for almost all points \( x \) of \( A \), almost all points near \( x \) are in \( A \), and for almost all points \( y \) of \( A^c \), almost all points near \( y \) are in \( A^c \). As a consequence, a Lebesgue measurable subset of \( \mathbb{R} \) is “lumpy” and cannot be distributed very evenly throughout \( \mathbb{R} \), or an interval in \( \mathbb{R} \), unless it is very densely or very sparsely distributed in that interval in a measure-theoretic sense.

**XIV.17.4.1. Definition.** Let \( A \) be a Lebesgue measurable subset of \( \mathbb{R} \). If \( c \in \mathbb{R} \), define the left and right upper and lower densities of \( A \) at \( c \) to be

- The left lower density \( d_- (A, c) = \liminf_{x \to c^-} \frac{\lambda([x, c] \cap A)}{\lambda([x, c])} \).
- The left upper density \( d^- (A, c) = \limsup_{x \to c^-} \frac{\lambda([x, c] \cap A)}{\lambda([x, c])} \).
- The right lower density \( d_+ (A, c) = \liminf_{x \to c^+} \frac{\lambda([c, x] \cap A)}{\lambda([c, x])} \).
- The right upper density \( d^+ (A, c) = \limsup_{x \to c^+} \frac{\lambda([c, x] \cap A)}{\lambda([c, x])} \).

If \( d_- (A, c) = d^- (A, c) \), the left density \( d_l (A, c) \) is defined and equals the common value.

If \( d_+ (A, c) = d^+ (A, c) \), the right density \( d_r (A, c) \) is defined and equals the common value.

If \( d_l (A, c) = d_r (A, c) \), the density \( d(A, c) \) is defined and equals the common value.

Note that the left and right upper and lower densities are always defined and are between 0 and 1. It should not be expected that left or right densities, much less densities, always exist (indeed, see XIV.17.5.5.); in fact, it is remarkable that the density exists for almost all \( c \) (and even more is true).

Comparing definitions, we immediately get the following characterization of densities as Dini derivatives:
**XIV.17.4.2.** Proposition. Let $A \subseteq \mathbb{R}$ be Lebesgue measurable, and $c \in \mathbb{R}$. Fix $a < c$, and define

$$F(x) = \int_a^x \chi_A \, d\lambda$$

for $x \geq a$. Then $d_-(A, c)$, $d^-(A, c)$, $d_+(A, c)$, $d^+(A, c)$ are the Dini derivatives ($D_- F(c)$, $D^- F(c)$, $D_+ F(c)$, $D^+ F(c)$) respectively. Thus $d_l(A, c) [d_r(A, c)]$ is defined if and only if $F$ is left [right] differentiable at $c$, and $d_l(A, c) = D_L F(c)$ [$d_r(A, c) = D_R F(c)$]. The density $d(A, c)$ is defined if and only if $F$ is differentiable at $c$, and $d(A, c) = F'(c)$.

**XIV.17.4.3.** Theorem. (Lebesgue Density Theorem) Let $A$ be a Lebesgue measurable subset of $\mathbb{R}$. Then the density of $A$ exists and is 1 for almost all points of $A$, and exists and is 0 for almost all points of $A^c$.

Proof: This is an almost immediate corollary of XIV.17.2.5. We may assume $A$ is bounded, since for $x \in (a, b)$ the density of $A$ at $x$ is the same as the density of $A \cap (a, b)$ at $x$. So suppose $A \subseteq (a, b)$. Set

$$F(x) = \int_a^x \chi_A \, d\lambda$$

for $x \in [a, b]$. Then $F$ is differentiable a.e. and $F' = \chi_A$ a.e. on $[a, b]$, i.e. $F'(x) = 1$ for almost all $x \in A$ and $F'(x) = 0$ for almost all $x \in [a, b] \setminus A$.

The next Corollary makes precise the statement that Lebesgue measurable sets are “lumpy” and cannot be evenly distributed.

**XIV.17.4.4.** Corollary. Let $A$ be a measurable subset of $\mathbb{R}$, and $I$ a bounded interval in $\mathbb{R}$.

(i) If $\lambda(A \cap I) > 0$, then for every $\epsilon > 0$ there is a subinterval $J$ of $I$ such that

$$\frac{\lambda(A \cap J)}{\lambda(J)} \geq 1 - \epsilon.$$ 

(ii) If $\lambda(A \cap I) < \lambda(I)$, then for every $\epsilon > 0$ there is a subinterval $J$ of $I$ such that

$$\frac{\lambda(A \cap J)}{\lambda(J)} \leq \epsilon.$$ 

**XIV.17.4.5.** Thus, for example, there cannot be a Lebesgue measurable set $A$ such that $\lambda(A \cap I) = \frac{1}{2} \lambda(I)$ for every bounded interval $I$, i.e. such that $A$ contains “half” the points in every interval. In fact, if $A$ is Lebesgue measurable, there is no $\epsilon > 0$ such that

$$\epsilon < \frac{\lambda(A \cap I)}{\lambda(I)} < 1 - \epsilon$$

for every bounded interval $I$.

**XIV.17.4.4.** does not hold for $\epsilon = 0$ in general (XIV.17.5.7.)

1721
XIV.17.5. Exercises

XIV.17.5.1. [Sak64] Let $A$ be a Cantor set of positive measure in $[0, 1]$, containing 0 and 1. Let $\{I_n : n \in \mathbb{N}\}$ be the open intervals in $[0, 1] \setminus A$, in decreasing order of length, and $c_n$ the midpoint of $I_n$. Let $\ell_n$ be the length of $I_n$, and $r_n$ the length of the longest subinterval of $[0, 1]$ disjoint from $I_1, \ldots, I_n$. Define a continuous function $f : [0, 1] \to \mathbb{R}$ by letting $f$ be identically zero on $A$, $f(c_n) = \ell_n + r_n$, and $f$ piecewise-linear on $I_n$.

(a) Show that $r_n \to 0$ (obviously $\ell_n \to 0$), and hence $f$ is indeed continuous.

(b) Show that $f$ is absolutely continuous on each $I_n$.

(c) Show that $f$ has Property (N). [If $N$ is a null set in $[0, 1]$, write $N = \bigcup_{k=0}^{\infty} N_k$, where $N_0 = N \cap A$ and $N_k = N \cap I_k$ for $k \geq 1$. Show that $\lambda(f(N_k)) = 0$ for all $k$.] This $f$ is an ACG function in the language of [Sak64].

(d) If $x \in A$, show that $D_-f(x) \leq 0$ and $D^+f(x) \geq 0$, and either $D_-f(x) \leq -1$ or $D^+f(x) \geq 1$ (or both). Hence $f$ fails to be differentiable anywhere on $A$, i.e. on a set of positive measure. Conclude that $f$ does not have bounded variation.

(e) Show that if $x > 0$, then $N_f(x)$ is finite. Thus $f$ has Property (S) by XIV.17.1.8.

XIV.17.5.2. Let $f$ be as in XIV.17.5.1., and set $g(x) = x + f(x)$.

(a) Show that $g$ has property (N). [See XIV.17.5.1 (c).]

(b) Show that for almost all $x \in A$, $N_g(x) = +\infty$. Thus $N_g$ is not finite a.e.

(c) By Theorem XIV.17.1.8., $g$ also does not satisfy property (S).

XIV.17.5.3. Verify that the function $\Phi$ defined in the proof of Lemma XIV.17.2.13. has all the required properties.

XIV.17.5.4. Show that Lemma XIV.17.2.12. is essentially an explicit version of XV.8.3.6., as follows.

(a) Deduce XIV.17.2.12. from XV.8.3.6. by taking $H$ to be a singular function approximating $\Psi$. [One additional correction with a scaled Cantor function is needed to make $H(a) = \Psi(a)$ and $H(b) = \Psi(b)$.

(b) Slightly modify the proof of XIV.17.2.12. to obtain XV.8.3.6.. [If $G$ is any continuous function, construct $H$ to approximate $\Psi - G$.]

XIV.17.5.5. Fix $c \in \mathbb{R}$ and $m, M$ with $0 \leq m \leq M \leq 1$. Show that a sequence

$$a_1 < b_1 < a_2 < b_2 < \cdots < c$$

of real numbers converging to $c$ can be constructed so that, if $A = \{c\} \cup \bigcup_{n=1}^{\infty} [a_n, b_n]$, then $d_-(A, c) = m$ and $d^-(A, c) = M$. Do the same for right densities. Thus the four densities at a given point can be arbitrary within the obvious constraints, even if $A$ is a closed set (or open set).

XIV.17.5.6. Let $A$ and $B$ be Lebesgue measurable subsets of $\mathbb{R}$. If $\lambda(A \Delta B) = 0$, show that $d_-(A, x) = d_-(B, x)$ for all $x \in \mathbb{R}$, and similarly for $d^-, d_+, d^+$. 

1722
XIV.17.5.7. Construct a Lebesgue measurable subset $A$ of $\mathbb{R}$ such that

$$0 < \frac{\lambda(A \cap I)}{\lambda(I)} < 1$$

for every bounded interval $I$ in $\mathbb{R}$. (Note that these numbers cannot be bounded away from 0 or 1 by XIV.17.4.4.) [Let $f$ be a differentiable function which is not monotone on any interval (V.12.1.7., XI.12.5.7.), and let $A = \{x : f'(x) > 0\}$. Use XIV.17.1.18..]
Chapter XV

Functional Analysis

Functional analysis is, in a nutshell, the study of topological vector spaces and operators on them. The name arose in part because historically (and still today), many of the most important topological vector spaces are function spaces. Another reason for the name is that one of the main underlying philosophies of functional analysis is that spaces can be studied by means of real- or complex-valued functions \((\text{functionals})\) on them.

Functional analysis combines linear algebra with analysis and topology. The interaction makes for a rich and powerful theory that has many applications in both pure and applied mathematics.

If a scientist wants to study a forest, there are two principal ways to examine it. First, one can walk through the forest. In this way, the small-scale properties can be studied in detail: the individual trees and the local ecosystems. But it is not easy to determine the large-scale structure of the forest this way, its overall extent and shape, the relative distribution of tree species and open spaces, etc. The other way is to fly over the forest. If you do this, the small-scale structure is not easily seen: individual trees are almost reduced to points in the overall picture; but the large-scale structure can be easily seen at a glance. A full study of the forest would undoubtedly include both methods.

Calculus and traditional elementary analysis give the basic tools necessary to do small-scale analysis of spaces of functions. Functional analysis provides the way to do large-scale analysis. Individual functions, although they have a rich internal structure of their own, are just points in an overall space from the functional analysis point of view. Flying over the function space using functional analysis is often a powerful way of easily establishing properties or solving problems related to function spaces.

Functional analysis reaches its full development and power in combination with measure theory, and many if not most texts assume familiarity with measure theory. But the essentials of functional analysis can be done without measure theory, so we will go as far as possible without weaving measure theory into the picture.
XV.1. Normed Vector Spaces

XV.1.1. Completeness

XV.1.1.1. Proposition. Let $X$ be a normed vector space and $Y$ a (necessarily closed) subspace of $X$. If $Y$ is complete, and $X/Y$ is complete in the quotient norm, then $X$ is complete.

Proof: Let $\hat{X}$ be the completion of $X$. Then $Y$ is closed in $\hat{X}$ since $Y$ is complete. Thus the quotient map from $X$ to $X/Y$ extends to a (bounded) quotient map $\pi$ from $\hat{X}$ to $\hat{X}/Y$; so $X/Y$ is naturally a subspace of $\hat{X}/Y$. On the one hand, $X/Y$ is dense in $\hat{X}/Y$ since $X$ is dense in $\hat{X}$ and $\pi$ is continuous; on the other hand, $X/Y$ is complete, hence closed in $\hat{X}/Y$. Thus $X/Y = \hat{X}/Y$, and hence $X = \hat{X}$ [if $z \in \hat{X}$, let $x \in X$ with $\pi(x) = \pi(z)$; then $z - x \in Y \subseteq X$, so $z = x + (z - x) \in X$].

Since finite-dimensional normed spaces are complete, we obtain:

XV.1.1.2. Corollary. Let $X$ be a normed vector space and $Y$ a (necessarily closed) subspace of $X$. If $Y$ is complete, and $X/Y$ is finite-dimensional, then $X$ is complete.

XV.1.2. Isometries of Normed Vector Spaces

In this section, we prove the following theorem due to Mazur and Ulam []:

XV.1.2.1. Theorem. Let $V$ and $W$ be normed vector spaces over $\mathbb{R}$. Then any isometry from $V$ onto $W$ is an $\mathbb{R}$-affine function.

XV.1.2.2. This result is false if $\mathbb{R}$ is replaced by $\mathbb{C}$. For example, $\alpha \mapsto \bar{\alpha}$ is an isometry on the complex vector space $\mathbb{C}$ which is not $\mathbb{C}$-affine (it is, of course, $\mathbb{R}$-affine).

We begin with a simple observation:

XV.1.2.3. Proposition. Let $V$ and $W$ be Hausdorff topological vector spaces over $\mathbb{R}$, and $f : V \to W$ a continuous function. Then $f$ is $\mathbb{R}$-affine if and only if, for every $x, y \in V$, we have

$$f \left( \frac{x + y}{2} \right) = \frac{f(x) + f(y)}{2}$$

(i.e. $f$ preserves midpoints).

Proof: It is obvious that an affine function preserves midpoints. Conversely, suppose $f$ preserves midpoints. By composing $f$ with a translation, we may assume $f(0) = 0$. It then follows that

$$f \left( \frac{x}{2} \right) = \frac{f(x)}{2}$$
for every \( x \in \mathcal{V} \) by taking \( y = 0 \) in the midpoint property. Thus \( f(2x) = 2f(x) \) for every \( x \in \mathcal{V} \), and thus, for any \( x,y \in \mathcal{V} \), we have

\[
f(x+y) = f\left(\frac{2x+2y}{2}\right) = \frac{f(2x) + f(2y)}{2} = \frac{2f(x) + 2f(y)}{2} = f(x) + f(y)
\]

so \( f \) is additive. Then by iteration we get that \( f(\alpha x) = \alpha f(x) \) for any dyadic rational \( \alpha \) and any \( x \). If \( \alpha \in \mathbb{R} \) is arbitrary, let \( (\alpha_n) \) be a sequence of dyadic rationals converging to \( \alpha \). Then, for any \( x \),

\[
f(\alpha x) = \lim_{n \to \infty} f(\alpha_n x) = \lim_{n \to \infty} \alpha_n f(x) = \alpha f(x)
\]

since \( \mathcal{W} \) is Hausdorff, i.e. the (translated) function \( f \) is linear.

The special case of XV.1.2.1. for \( \mathcal{V} = \mathcal{W} = \mathbb{R}^n \) follows almost immediately:

**XV.1.2.4. Theorem.** Every isometry of \( \mathbb{R}^n \) (with the usual Euclidean norm) is affine.

**Proof:** Let \( f \) be an isometry of \( \mathbb{R}^n \), and let \( x,y \in \mathbb{R}^n \). Let \( z = \frac{x+y}{2} \) be the midpoint of \( x \) and \( y \). Then \( z \) is the unique point of \( \mathbb{R}^n \) with the property that

\[
2(\rho_2(x,z) - \rho_2(z,y)) = \frac{1}{2} \rho_2(x,y)
\]

where \( \rho_2 \) is the standard metric coming from the norm. It follows that \( f(z) \) must satisfy

\[
\rho_2(f(x), f(z)) = \frac{1}{2} \rho_2(f(x), f(y))
\]

since \( f \) is an isometry. Thus \( f(z) \) must be the midpoint of \( f(x) \) and \( f(y) \). Since \( f \) is continuous, it is affine by XV.1.2.3.

**XV.1.2.5.** This proof works more generally for XV.1.2.1. if \( \mathcal{V} \) and \( \mathcal{W} \) are strictly convex (\( \rho \)). For the general case of XV.1.2.1., we need a subtler argument:

**Proof:** Write \( \rho \) and \( \sigma \) for the norm metrics on \( \mathcal{V} \) and \( \mathcal{W} \) respectively. Fix \( x \) and \( y \) in \( \mathcal{V} \), and set \( z = \frac{x+y}{2} \). By XV.1.2.3. it suffices to show that

\[
f(z) = f\left(\frac{x+y}{2}\right)
\]

Let \( A_1 \) be the set of all \( a \in \mathcal{V} \) such that

\[
\rho(x,a) = \rho(a,y) = \frac{1}{2} \rho(x,y)
\]

Then \( A_1 \) is a subset of \( \mathcal{V} \) containing \( z \). We show that \( A_1 \) is symmetric around \( z \), i.e. if \( a \in A_1 \), then \( b = a + 2(z - a) = 2z - a \) is also in \( A_1 \) (\( b \) is the point for which \( z \) is the midpoint of \( a \) and \( b \)). We have \( 2z = x + y \), and thus \( b - x = y - a \) and \( b - y = x - a \), so

\[
\rho(x, b) = \rho(y, a) = \rho(a, x) = \rho(b, y) = \frac{1}{2} \rho(x, y)
\]
and \( b \in A_1 \).

Let \( \delta_1 \) be the diameter of \( A_1 \). Then \( \delta_1 \) is finite since if \( a, b \in A_1 \), we have

\[
\rho(a, b) \leq \rho(a, x) + \rho(x, b) = \rho(x, y).
\]

For any \( a \in A_1 \), set \( b = 2z - a \in A_1 \). Since \( z \) is the midpoint of \( a \) and \( b \), we have

\[
\rho(a, z) = \frac{1}{2} \rho(a, b) \leq \delta_1.
\]

Now let \( A_2 \) be the set of \( c \in A_1 \) such that \( \rho(a, c) \leq \frac{1}{2} \delta_1 \) for all \( a \in A_1 \). Then \( z \in A_2 \). If \( c \in A_2 \), set \( d = 2z - c \). If \( a \in A_1 \), set \( b = 2z - a \in A_1 \). We then have \( 2z = a + b \), so \( d - a = c - b \) and hence

\[
\rho(a, d) = \rho(b, c) \leq \frac{1}{2} \delta_1,
\]

so \( d \in A_2 \) and \( A_2 \) is also symmetric around \( z \). Also, if \( c, d \in A_2 \), then \( c \in A_1 \) so

\[
\rho(c, d) \leq \frac{1}{2} \delta_1.
\]

Thus the diameter \( \delta_2 \) of \( A_2 \) satisfies \( \delta_2 \leq \frac{1}{2} \delta_1 \).

If \( c \in A_2 \), then \( d = 2z - c \in A_2 \), so we have

\[
\rho(c, z) = \rho(z, d) = \frac{1}{2} \rho(c, d) \leq \frac{1}{2} \delta_2.
\]

Let \( A_3 \) be the set of \( v \in A_2 \) for which \( \rho(c, v) \leq \frac{1}{2} \delta_2 \) for all \( c \in A_2 \). Then \( z \in A_3 \) and, as before, \( A_3 \) is symmetric around \( z \) and the diameter \( \delta_3 \) of \( A_3 \) satisfies \( \delta_3 \leq \frac{1}{2} \delta_2 \).

Iterating the process, we obtain a decreasing sequence \((A_n)\) of sets containing \( z \), symmetric around \( z \), with diameter

\[
\delta_n \leq \frac{1}{2} \delta_{n-1} \leq \frac{1}{2^n} \delta_1
\]

and hence \( \cap_n A_n = \{ z \} \).

Let \( w \) be the midpoint of \( f(x) \) and \( f(y) \) in \( W \). Similarly, using \( \sigma \) construct a decreasing sequence \((B_n)\) of subsets of \( W \) containing \( w \). Since the \( A_n \) and \( B_n \) are constructed using just the metrics \( \rho \) and \( \sigma \), and \( f \) is an isometry, we must have \( f(A_n) = B_n \) for all \( n \). Thus we have that

\[
f(\{ z \}) = f(\cap_n A_n) = \cap_n B_n = \{ w \}.
\]

\( \diamond \)

### XV.1.3. Exercises

#### XV.1.3.1. Use the Baire Category Theorem to give an alternate proof that that no Banach space has vector space dimension \( \aleph_0 \) ([If \( \{ \xi_n : n \in \mathbb{N} \} \) is a Hamel basis for \( X \), consider \( X_n = \text{span}\{\xi_1, \ldots, \xi_n\} \).]
XV.2. Topological Vector Spaces

XV.2.1. Basic Definitions and Properties

XV.2.2. Locally Convex Topological Vector Spaces

As pointed out in [KN76], the main extra feature in topological vector spaces distinguishing them from general topological groups is convexity.

Convexity is hard to categorize mathematically. It is essentially a purely algebraic concept; but it is defined only in vector spaces, or affine spaces, over the real numbers (or over a field containing the real numbers), and the primary applications are in analysis.

XV.2.2.1. Definition. Let $C$ and $D$ be convex sets in vector spaces $X$ and $Y$ respectively. An affine map from $C$ to $D$ is a function $f : C \to D$ such that

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y)$$

for all $x, y \in C$ and all $t, 0 \leq t \leq 1$.

XV.2.2.2. An affine map from $C$ to $D$ is a function which preserves the convex structure. If $f$ is an affine map from $X$ to $Y$ in the usual sense of $()$, and $f(C) \subseteq D$, then $f|_C$ is an affine map from $C$ to $D$ in the sense of XV.2.2.1. It can be shown that every affine map from $C$ to $D$ arises in this way $()$, so the terminology is consistent.

XV.2.2.3. Definition. Let $C$ and $D$ be convex sets in topological vector spaces $X$ and $Y$ respectively. An isomorphism from $C$ to $D$ is an affine homeomorphism from $C$ to $D$.

Thus an isomorphism from $C$ to $D$ is a bijection which preserves both the convex structure and the topology.

XV.2.3. Metrizability and Normability
XV.3. Bounded Operators

When considering normed vector spaces, or more generally topological vector spaces, the most natural and important functions between them should preserve both the algebraic structure, i.e. be linear, and the topological structure, i.e. be continuous. Thus we usually want to work with continuous linear operators. In the case of normed vector spaces, continuity is equivalent to another property, boundedness, which is usually easier to work with directly.

XV.3.1. Bounded Operators on Normed Vector Spaces

XV.3.1.1. Definition. Let $X$ and $Y$ be vector spaces over $F$. A (linear) operator from $X$ to $Y$ is a linear transformation () from $X$ to $Y$. Thus operator and linear transformation are synonymous. The term operator is primarily used in the case $X = Y$, in which case we call it an operator on $X$. If $T : X \to Y$ is an operator, we usually write $Tx$ instead of $T(x)$ for the image of $x \in X$ under $T$.

XV.3.1.2. Definition. Let $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$ be normed vector spaces over $F$, and $T : X \to Y$ an operator (linear transformation). Then $T$ is bounded, or a bounded operator, if there is a constant $M$ such that

$$\|Tx\|_Y \leq M\|x\|_X$$

for all $x \in X$. The smallest such $M$ is called the operator norm of $T$, written $\|T\|_{op}$ or usually just $\|T\|$ when there is no possibility of confusion.

Although the notation does not reflect it, boundedness of $T$ and its operator norm depend on the choices of norms on $X$ and $Y$. Thus, although we often write $T : X \to Y$ by slight abuse of notation, there must be understood norms on $X$ and $Y$.

XV.3.1.3. If $T$ is a bounded operator from $X$ to $Y$, we have

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \|Tx\|_Y$$

since $T$ is linear and any vector is a scalar multiple of a unit vector. The 0 operator is bounded, and $\|0\| = 0$; if $T$ is bounded, then $\|T\| \geq 0$, and $\|T\| = 0$ if and only if $T = 0$.

The most important feature of boundedness is that it is equivalent to continuity:

XV.3.1.4. Proposition. Let $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$ be normed vector spaces over $F$, and $T : X \to Y$ a linear transformation. The following are equivalent:

(i) $T$ is continuous at 0.

(ii) $T$ is continuous.

(iii) $T$ is uniformly continuous.

(iv) $T$ is bounded.
Proof: (iii) ⇒ (ii) ⇒ (i) is trivial.
(iv) ⇒ (iii): Suppose $T$ is bounded. We may assume $T \neq 0$, so $\|T\| > 0$. Let $\epsilon > 0$, and set $\delta = \frac{\epsilon}{\|T\|}$. If $x_1, x_2 \in X$ with $\|x_2 - x_1\|_X < \delta$, then

$$\|Tx_2 - Tx_1\|_Y = \|T(x_2 - x_1)\|_Y \leq \|T\|\|x_2 - x_1\|_X < \|T\|\delta = \epsilon.$$ 

Thus $T$ is uniformly continuous.

(i) ⇒ (iv): Suppose $T$ is not bounded. Then for every $n$ there is a $v_n \in X$, $v_n \neq 0$, with $\|Tv_n\|_Y > n\|v_n\|_X$. Set $x_n = \frac{1}{n\|v_n\|_X}v_n$. Then $\|x_n\|_X = \frac{1}{n}$ and

$$\|Tx_n\|_Y = \left\|T\left(\frac{1}{n\|v_n\|_X}v_n\right)\right\|_Y = \frac{1}{n\|v_n\|_X}\|Tv_n\|_Y > 1$$

for all $n$. Thus $x_n \to 0$ in $X$ but $Tx_n \not\to 0$ in $Y$, so $T$ is not continuous at 0. \qed

**XV.3.1.5.** Let $X$ and $Y$ be vector spaces over $\mathbb{F}$. The set $\mathcal{L}(X,Y)$ of linear transformations from $X$ to $Y$ is a vector space over $\mathbb{F}$ under the usual operations $()$.

**XV.3.1.6.** Proposition. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces over $\mathbb{F}$, and $S$ and $T$ bounded operators from $X$ to $Y$. Then

(i) $S + T$ is a bounded operator from $X$ to $Y$, and $\|S + T\| \leq \|S\| + \|T\|$.

(ii) If $\alpha \in \mathbb{F}$, then $\alpha T$ is bounded and $\|\alpha T\| = |\alpha|\|T\|$.

Proof: (i): If $x \in X$, we have

$$\|(S + T)x\|_Y = \|Sx + Tx\|_Y \leq \|Sx\|_Y + \|Tx\|_Y \leq \|S\|\|x\|_X + \|T\|\|x\|_X = (\|S\| + \|T\|)\|x\|_X.$$ 

(ii): We have

$$\|\alpha T\| = \sup_{\|x\|_X = 1} \|\alpha(Tx)\|_Y = \sup_{\|x\|_X = 1} \|\alpha(Tx)\|_Y = \sup_{\|x\|_X = 1} |\alpha|\|Tx\|_Y = |\alpha| \sup_{\|x\|_X = 1} \|Tx\|_Y = |\alpha|\|T\|.$$ 

\qed

**XV.3.1.7.** Corollary. Let $X$ and $Y$ be normed vector spaces over $\mathbb{F}$. Then the set $\mathcal{B}(X,Y)$ of bounded operators from $X$ to $Y$ is a vector subspace of $\mathcal{L}(X,Y)$, and the operator norm is a norm on $\mathcal{B}(X,Y)$.

The operator norm has one more simple but important property:
**XV.3.1.8.** Proposition. Let \( X, Y, Z \) be normed vector spaces over \( F \), \( T \in \mathcal{B}(X, Y) \), \( S \in \mathcal{B}(Y, Z) \). Then \( ST = S \circ T \in \mathcal{B}(X, Z) \) and \( \| ST \| \leq \| S \| \| T \| \).

Proof: If \( x \in X \), we have
\[
\| (ST)x \|_Z = \| S(Tx) \|_Z \leq \| S \| \| Tx \|_Y \leq \| S \| \| T \| \| x \|_X.
\]

Simple examples with \( 2 \times 2 \) matrices show that the inequality can be strict; if fact, we can have \( ST = 0 \) even if \( S \) and \( T \) are both nonzero, if the range of \( T \) is contained in the null space of \( S \).

Completeness

**XV.3.1.9.** Theorem. Let \( X \) and \( Y \) be normed vector spaces over \( F \). If \( Y \) is complete (a Banach space), then \( \mathcal{B}(X, Y) \) is complete (a Banach space) in the operator norm.

Proof: The proof follows the standard pattern of completeness proofs. Suppose \((T_n)\) is a Cauchy sequence in \( \mathcal{B}(X, Y) \). We first find a candidate for the limit \( T \).

Fix \( x \in X \). We have, for \( n, m \in \mathbb{N} \),
\[
\| T_n x - T_m x \|_Y = \| (T_n - T_m) x \|_Y \leq \| T_n - T_m \| \| x \|_X
\]
so the sequence \((T_n x)\) is a Cauchy sequence in \( Y \). Since \( Y \) is complete, the sequence \((T_n x)\) converges to an element of \( Y \) we may call \( T x \). As \( x \) varies over \( X \), this procedure defines a function \( T : X \to Y \), and \( T_n \to T \) pointwise.

We must show that \( T \) is a bounded operator and that \( T_n \to T \) in operator norm. If \( x_1, x_2 \in X \), then \( T_n x_1 \to T x_1, T_n x_2 \to T x_2 \), and \( T_n (x_1 + x_2) \to T (x_1 + x_2) \). But \( T_n (x_1 + x_2) = T_n x_1 + T_n x_2 \) by linearity, and \( T_n x_1 + T_n x_2 \to T x_1 + T x_2 \) by continuity of addition in \( Y \). Thus \( T(x_1 + x_2) = T x_1 + T x_2 \) by uniqueness of limits in \( Y \). Similarly, if \( \alpha \in F \), we have \( T_n (\alpha x) \to T (\alpha x) \) and \( T_n (\alpha x) = \alpha T_n x \to \alpha (T x) \), so \( T (\alpha x) = \alpha (T x) \). Thus \( T \) is linear.

To show that \( T \) is bounded, note that the sequence \((T_n)\), being a Cauchy sequence, is bounded, say \( \| T_n \| \leq M \) for all \( n \). Fix \( x \) with \( \| x \| = 1 \). Then \( \| T_n x \|_Y \leq M \| x \|_X = M \), and since \( T_n x \to T x \), we also have \( \| T x \|_Y \leq M \). This is true for every \( x \) of norm 1, so \( T \) is bounded and \( \| T \| \leq M \).

Finally, let \( \epsilon > 0 \). Since \((T_n)\) is a Cauchy sequence, there is an \( N \) such that \( \| T_n - T_m \| < \epsilon \) for all \( n, m \geq N \). Fix \( x \in X \), \( \| x \|_X = 1 \). Then
\[
\| T_n x - T_m x \|_Y = \| (T_n - T_m) x \|_Y \leq \| T_n - T_m \| \| x \|_X < \epsilon
\]
for all \( n, m \geq N \). Fix \( n \geq N \); since \( T_n x \to T x \), we have \( \| (T_n - T) x \|_Y = \| T_n x - T x \|_Y \leq \epsilon \). This is true for all such \( x \), so \( \| T_n - T \| \leq \epsilon \). This is true for all \( n \geq N \), so \( T_n \to T \) in operator norm.

**XV.3.1.10.** Note that completeness of \( X \) is not needed for this result. But if \( Y \) is complete, any bounded operator from \( X \) to \( Y \) extends uniquely to a bounded operator from the completion \( \bar{X} \) to \( Y \) with the same norm, so \( \mathcal{B}(X, Y) \cong \mathcal{B}(\bar{X}, Y) \):
**XV.3.1.11.** PROPOSITION. Let $X$ and $Y$ be normed vector spaces, with $Y$ complete, and let $Z$ be a dense subspace of $X$. If $T$ is a bounded operator from $Z$ to $Y$, then $T$ extends uniquely to a bounded operator $\hat{T} : X \to Y$, and $\|\hat{T}\| = \|T\|$.

PROOF: Since $T$ is uniformly continuous (XV.3.1.4.), and $Y$ is complete, $T$ extends uniquely to a continuous function $\hat{T}$ from $X$ to $Y$. If $x_1, x_2 \in X$, there are sequences $(z_{n1})$ and $(z_{n2})$ in $Z$ converging to $x_1$ and $x_2$ respectively. Then $z_{n1} + z_{n2} \to x_1 + x_2$, so $\hat{T}(z_{n1} + z_{n2}) = T(z_{n1} + z_{n2}) \to \hat{T}(x_1 + x_2)$. On the other hand, $Tz_{n1} = \hat{T}z_{n1} \to \hat{T}x_1$ and $Tz_{n2} = \hat{T}z_{n2} \to \hat{T}x_2$, so

$$\hat{T}(z_{n1} + z_{n2}) = T(z_{n1} + z_{n2}) = Tz_{n1} + Tz_{n2} \to \hat{T}x_1 + \hat{T}x_2$$

so $\hat{T}(x_1 + x_2) = \hat{T}x_1 + \hat{T}x_2$ by uniqueness of limits in $Y$. Similarly, if $\alpha \in F$, $\hat{T}(\alpha z_{n1})$ converges to both $T(\alpha x_1)$ and $\alpha(Tx_1)$, so $\hat{T}(\alpha z_{n1}) = \alpha(Tx_1)$ and $\hat{T}$ is linear.

Let $x \in X$, $\|x\| = 1$. There is a sequence $(z_n)$ in $Z$ converging to $x$. Then $\|z_n\|_X \to \|x\|_X = 1$, so if $w_n = \frac{z_n}{\|z_n\|_X}$, then $w_n \in Z$ for all $n$, $\|w_n\|_X = 1$ for all $n$, and $w_n \to x$ by joint continuity of scalar multiplication. We have $\|T w_n\|_Y \leq \|T\|$ for all $n$, and since $\hat{T} w_n = T w_n \to \hat{T} x$, we also have $\|\hat{T} x\|_Y \leq \|T\|$. This is true for all such $x$, so $\hat{T}$ is bounded and $\|\hat{T}\| \leq \|T\|$.

For the opposite inequality, let $\epsilon > 0$. There is a $z \in \hat{Z}$ with $\|z\|_X = 1$ and $\|Tz\|_Y > \|T\| - \epsilon$. We then have

$$\|T\| - \epsilon < \|\hat{T} z\|_Y = \|\hat{T} z\|_Y \leq \|\hat{T}\| \|z\|_X = \|\hat{T}\|.$$

The extreme inequality is true for all $\epsilon > 0$, and thus $\|T\| \leq \|\hat{T}\|$.

** XV.3.2. Bounded Linear Functionals **

**XV.3.2.12.** DEFINITION. Let $X$ be a vector space over $\mathbb{F}$. A linear functional on $X$ is a linear transformation from $X$ to $\mathbb{F}$. If $X$ is a normed vector space, a linear functional $\phi$ on $X$ is bounded if it is bounded as a linear transformation (regarding $\mathbb{F}$ as a normed vector space with norm $|\cdot|$), i.e. if there is a constant $M$ such that $|\phi(x)| \leq M \|x\|$ for all $x \in X$. The smallest such $M$ is the norm of $\phi$, denoted $\|\phi\|$.

**XV.3.2.13.** NOTATION: If $X$ is a vector space over $\mathbb{F}$, we will write $X'$ for the set $\mathcal{L}(X, \mathbb{F})$ of linear functionals on $X$. If $X$ is normed, write $X^*$ for the set $\mathcal{B}(X, \mathbb{F})$ of bounded linear functionals on $X$. Then $X'$ is a vector space over $\mathbb{F}$, and by XV.3.1.7., $X^*$ is a subspace which is normed by $\|\cdot\|_X$; $X^*$ is a Banach space by XV.3.1.9.. $X'$ is called the algebraic dual space of $X$, and $X^*$ the (topological) dual space of $X$.

There is no uniformity of notation for dual spaces. Many authors use $X'$ for the topological dual space, and some of these (e.g. [?]) even use $X^*$ to denote the algebraic dual. We use what seems to be the most common notation. We usually use $\phi$, $\psi$, etc. to denote linear functionals; some authors use $x^*$ to denote a general element of $X^*$, leading to poor and confusing expressions like $x^*(x)$ where $x$ denotes a general element of $X$ and $x^*$ a general element of $X^*$, with no relation between them.

**XV.3.2.14.** By XV.3.1.4., a linear functional on a normed vector space is continuous if and only if it is bounded. The null space of any bounded operator is closed () . The converse does not hold in general, e.g. there are unbounded one-to-one linear transformations. However:
**Proposition.** Let $X$ be a normed vector space over $F$, and $\phi : X \to F$ a linear functional on $X$. If the null space $N_\phi$ is closed in $X$, then $\phi$ is bounded.

**Proof:** Suppose $\phi$ is unbounded. Then for each $n \in \mathbb{N}$ there is a $y_n \in X$ with $\|y_n\| = 1$ but $|\phi(y_n)| > n$. Set $x_n = \frac{1}{\phi(y_n)}y_n$. Then $\|x_n\| < 1/n$ and $\phi(x_n) = 1$ for all $n$. Fix $v \in X$ with $\phi(v) = 1$. We have $x_n \to 0$, so $v - x_n \to v$. But $v - x_n \in N_\phi$ for all $n$, so $v \in \overline{N_\phi} \setminus N_\phi$, and $N_\phi$ is not closed.

In fact, this argument shows that if $\phi$ is unbounded, then $N_\phi$ is dense in $X$. (Actually it would have to be either closed or dense since it has codimension 1.)

**XV.3.3. Operators on Finite-Dimensional Normed Spaces**

**XV.3.4. Bounded Operators on Hilbert Spaces**

**XV.3.5. Algebras of Operators**

**XV.3.5.1. Definition.** A (concrete) operator algebra is a *-subalgebra of $B(\mathcal{H})$ which is topologically closed in a suitable sense (there is also a subject of non-self-adjoint operator algebras, but it is outside the scope of this volume.) A von Neumann algebra is a (necessarily unital) *-subalgebra $M$ of $B(\mathcal{H})$ such that $M = M''$.

A von Neumann algebra is weakly (hence strongly, norm-, ...) closed, and in particular is a concrete C*-algebra (XV.14.1.3.), albeit of a very special kind.

**Commutant and Bicommutant**

One of the first, yet still one of the most crucial basic theorems of operator algebra theory is von Neumann’s Bicommutant Theorem [?]. We say a *-algebra $A$ of operators on a Hilbert space $\mathcal{H}$ acts nondegenerately if $T \xi = 0$ for all $T \in A$ implies $\xi = 0$. Since $A$ is a *-algebra, this is equivalent to the condition that the subspace $A\mathcal{H} = \text{span}\{T \xi : T \in A, \xi \in \mathcal{H}\}$

is dense in $\mathcal{H}$. If $I \in A$, then $A$ obviously acts nondegenerately.

**Theorem. [Bicommutant Theorem]** Let $A$ be a *-subalgebra of $B(\mathcal{H})$ acting nondegenerately. Then $A$ is $\sigma$-strongly dense in $A''$.

**Proof:** Using ?, the proof reduces to showing that if $\xi \in \mathcal{H}$ and $T \in A''$, there is a sequence $(T_n)$ in $A$ with $T_n \xi \to T \xi$. If $\mathcal{X} = \{S \xi : S \in A\}^\perp$, then $P_X \in A'$ (since both $\mathcal{X}$ and $\mathcal{X}^\perp$ are invariant under $A$), so $T$ commutes with $P_X$, i.e. $T$ leaves $\mathcal{X}$ invariant. It remains to show that $\xi \in \mathcal{X}$, so that $T \xi \in \mathcal{X}$. This is trivial if $I \in A$. In the general nondegenerate case, for each $S \in A$,

$$S[(I - P_X)\xi] = (I - P_X)[S\xi] = 0$$

since $S\xi \in \mathcal{X}$, so $(I - P_X)\xi = 0$.

The Bicommutant Theorem relates a topological closure property with the simple, natural, and purely algebraic property of being the bicommutant (or just commutant) of a *-closed set of operators.
XV.3.5.3. Corollary. Let $A$ be a unital (or just nondegenerate) $*$-subalgebra of $\mathcal{B}(\mathcal{H})$. The following are equivalent:

(i) $A$ is strongly closed.

(ii) $A$ is $\sigma$-strongly closed.

(iii) $A$ is weakly closed.

(iv) $A$ is $\sigma$-weakly closed.

(v) $A$ is a von Neumann algebra, i.e. $A''$. 

XV.4. Dual Spaces and Weak Topologies

One of the core tenets of functional analysis is that spaces can be studied via the scalar-valued continuous functions on them. This section describes the linear version of this philosophy.

XV.4.1. Bounded Linear Functionals

XV.4.1.4. Definition. Let $X$ be a vector space over $F$. A linear functional on $X$ is a linear transformation from $X$ to $F$. If $X$ is a normed vector space, a linear functional $\phi$ on $X$ is bounded if it is bounded as a linear transformation, i.e. if there is a constant $M$ such that $|\phi(x)| \leq M\|x\|$ for all $x \in X$. The smallest such $M$ is the norm of $\phi$, denoted $\|\phi\|$.

XV.4.1.5. Notation: If $X$ is a vector space over $F$, we will write $X'$ for the set of linear functionals on $X$. If $X$ is normed, write $B(X;F)$ for the set of bounded linear functionals on $X$. We have that $X'$ is a vector space over $F$ in the obvious way, and by ( ), $X'$ is a subspace which is normed by $\|\cdot\|$; $X^*$ is a Banach space by ( ). $X'$ is called the algebraic dual space of $X$, and $X^*$ the (topological) dual space of $X$.

There is no uniformity of notation for dual spaces. Many authors use $X'$ for the topological dual space, and some of these (e.g. [?]) even use $X^*$ to denote the algebraic dual. We use what seems to be the most common notation. We usually use $\phi$, $\psi$, etc. to denote linear functionals; some authors use $x^*$ to denote a general element of $X^*$, leading to poor and confusing expressions like $x^*(x)$ where $x$ denotes a general element of $X$ and $x^*$ a general element of $X^*$, with no relation between them.

XV.4.1.6. By ( ), a linear functional on a normed vector space is continuous if and only if it is bounded. The null space of any bounded operator is closed ( ). The converse does not hold in general, e.g. there are unbounded one-to-one linear transformations. However:

XV.4.1.7. Proposition. Let $X$ be a normed vector space over $F$, and $\phi : X \to F$ a linear functional on $X$. If the null space $N_\phi$ is closed in $X$, then $\phi$ is bounded.

Proof: Suppose $\phi$ is unbounded. Then for each $n \in \mathbb{N}$ there is a $y_n \in X$ with $\|y_n\| = 1$ but $|\phi(y_n)| > n$. Set $x_n = \frac{1}{\phi(y_n)}y_n$. Then $\|x_n\| < 1/n$ and $\phi(x_n) = 1$ for all $n$. Fix $v \in X$ with $\phi(v) = 1$. We have $x_n \to 0$, so $v - x_n \to v$. But $v - x_n \in N_\phi$ for all $n$, so $v \in \overline{N_\phi} \setminus N_\phi$, and $N_\phi$ is not closed.

In fact, this argument shows that if $\phi$ is unbounded, then $N_\phi$ is dense in $X$. (Actually it would have to be either closed or dense since it has codimension 1.)
XV.5. The Hahn-Banach Theorem

The Hahn-Banach Theorem is one of the most fundamental results of functional analysis. In various forms it asserts the existence of “many” linear functionals on real or complex vector spaces. Indeed, this theorem underlies and justifies the whole philosophy of functional analysis: that a space can be completely studied via the scalar-valued functionals on the space.

There are many slightly different forms in which this theorem is commonly expressed. All the general versions require use of some form of the Axiom of Choice, but versions for vector spaces which are “not too large” can be proved directly without using the AC.

XV.5.1. Sublinear Functionals and Version 1

The first version of the Hahn-Banach Theorem concerns existence of linear functionals on a real vector space which are trapped between a sublinear functional and a superlinear functional, and more generally extensions to the whole vector space of such a linear functional already defined on a subspace.

**XV.5.1.1. Definition.** Let $V$ be a real vector space. A function $p : V \to (-\infty, +\infty]$ is a *sublinear functional* on $V$ if

(i) $p(x + y) \leq p(x) + p(y)$ for all $x, y \in V$.

(ii) $p(\alpha x) = \alpha p(x)$ for all $\alpha > 0$.

A function $q : V \to [-\infty, +\infty)$ is a *superlinear functional* on $V$ if

(i) $q(x + y) \geq q(x) + q(y)$ for all $x, y \in V$.

(ii) $q(\alpha x) = \alpha q(x)$ for all $\alpha > 0$.

A sublinear or superlinear functional taking only real values is called *finite*.

Note that a sublinear functional can take the value $+\infty$ but not $-\infty$, and a superlinear functional can take the value $-\infty$ but not $+\infty$. In most applications, sublinear or superlinear functionals are finite.

**XV.5.1.2. Examples.** (i) Any linear functional is both sublinear and superlinear.

(ii) The function $p(x) = +\infty$ for all $x$ is sublinear; the function $q(x) = -\infty$ for all $x$ is superlinear.

(iii) Any seminorm () is a finite sublinear functional taking only nonnegative values.

(iv) Let $C$ be a convex set in $V$ containing $0$. For $x \in V$, set

$$p(x) = \inf\{\alpha > 0 : \alpha^{-1}x \in C\}$$

(where $\inf(\emptyset) = +\infty$). It is easily checked () that $p$ is a sublinear functional on $V$, called the *Minkowski functional* of $C$. It takes only nonnegative values, and is finite if and only if $C$ is absorbing ().

(v) If $p$ is a sublinear functional, then $-p$ is a superlinear functional, and vice versa.

The first version of the Hahn-Banach Theorem is:
Theorem. [Hahn-Banach Theorem, Version 1] Let \( V \) be a real vector space, \( p \) a sublinear functional on \( V \), and \( q \) a superlinear functional on \( V \) satisfying \( q(x) \leq p(x) \) for all \( x \in V \). Assume at least one of \( p \) and \( q \) is finite. Let \( W \) be a subspace of \( V \), and \( \psi \) a linear functional on \( W \) satisfying \( q(y) \leq \psi(y) \leq p(y) \) for all \( y \in W \). Then there is a linear functional \( \phi \) on \( V \) extending \( \psi \) with
\[
q(x) \leq \phi(x) \leq p(x)
\]
for all \( x \in V \).

Taking \( q(x) = -\infty \) for all \( x \), we get the following commonly-stated version of the theorem:

Corollary. [Hahn-Banach Theorem, Version 1A] Let \( V \) be a real vector space, and \( p \) a finite sublinear functional on \( V \). Let \( W \) be a subspace of \( V \), and \( \psi \) a linear functional on \( W \) satisfying \( \psi(y) \leq p(y) \) for all \( y \in W \). Then there is a linear functional \( \phi \) on \( V \) extending \( \psi \) with
\[
\phi(x) \leq p(x)
\]
for all \( x \in V \).

There is an analogous result with inequalities reversed for a superlinear functional.

If \( p \) is a sublinear functional on \( V \), then \( p(0) = \alpha p(0) \) for all \( \alpha > 0 \). Thus \( p(0) \) is either 0 or \( +\infty \). Similarly, if \( q \) is superlinear, then \( q(0) \) is either 0 or \( -\infty \). Thus, if \( p \) is sublinear and \( q \) superlinear, then \( q(0) \leq 0 \leq p(0) \), so the hypotheses of XV.5.1.3. and XV.5.1.4. are automatically satisfied if \( W \) is the zero subspace and \( \psi \) the zero functional. Thus we obtain:

Corollary. Let \( V \) be a real vector space, \( p \) a sublinear functional on \( V \), and \( q \) a superlinear functional on \( V \) satisfying \( q(x) \leq p(x) \) for all \( x \in V \). Assume at least one of \( p \) and \( q \) is finite. Then there is a linear functional \( \phi \) on \( V \) with
\[
q(x) \leq \phi(x) \leq p(x)
\]
for all \( x \in V \).

There are also one-sided versions of this result.

The proof of Theorem XV.5.1.3. consists of two steps:

1. Show that \( \psi \) can be extended to a subspace spanned by \( W \) and one additional element.
2. “Iterate” step 1 to extend \( \psi \) all the way to \( V \).

The first step is elementary and straightforward, but takes some work. The second step in general requires some form of the Axiom of Choice.

We begin with step 1.
XV.5.1.8. **Lemma.** Let $V$ be a real vector space, $p$ a sublinear functional on $V$, and $q$ a superlinear functional on $V$ satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of $p$ and $q$ is finite. Let $W$ be a subspace of $V$, and $\psi$ a linear functional on $W$ satisfying $q(y) \leq \psi(y) \leq p(y)$ for all $y \in W$. Let $x_0 \in V \setminus W$, and let $X$ be the subspace of $V$ spanned by $W$ and $x_0$. Then there is a linear functional $\theta$ on $X$ extending $\psi$ with

$$q(x) \leq \theta(x) \leq p(x)$$

for all $x \in X$.

**Proof:** Every $x \in X$ can be uniquely written as $y + \alpha x_0$, where $y \in W$ and $\alpha \in \mathbb{R}$. Any linear functional on $X$ extending $\psi$ is of the form

$$\theta(y + \alpha x_0) = \psi(y) + \alpha c$$

for some $c \in \mathbb{R}$, so we must choose $c$ so that $q(x) \leq \theta(x) \leq p(x)$ for all $x \in X$, i.e.

$$q(y + \alpha x_0) \leq \psi(y) + \alpha c \leq p(y + \alpha x_0)$$

for all $y \in W$ and $\alpha \in \mathbb{R}$. To show this is possible, we obtain some inequalities.

First, if $y, z \in W$, we have

$$\psi(y) - \psi(z) = \psi(y - z) \leq p(y - z) = p([y + x_0] + [-x_0 - z]) \leq p(y + x_0) + p(-x_0 - z)$$

and similarly we obtain

$$p(-x_0 - z) - \psi(z) \leq q(-x_0 - z) - \psi(z)$$

Finally, we have

$$\psi(y) - \psi(z) = \psi(y + x_0) - \psi(x_0 + z) \leq p(y + x_0) - q(x_0 + z)$$

and we similarly obtain

$$p(-x_0 - z) - \psi(z) \leq q(-x_0 - y) - \psi(y)$$

Now define

$$\sigma_1 = \sup\{-p(-x_0 - y) - \psi(y) : y \in W\}$$
$$\sigma_2 = \sup\{q(x_0 + y) - \psi(y) : y \in W\}$$
$$\tau_1 = \inf\{p(x_0 + y) - \psi(y) : y \in W\}$$
$$\tau_2 = \inf\{-q(-x_0 - y) - \psi(y) : y \in W\}$$

We have $\sigma_1 \leq \tau_1$ by XV.1, $\sigma_2 \leq \tau_2$ by XV.2, $\sigma_2 \leq \tau_1$ by XV.3, and $\sigma_1 \leq \tau_2$ by XV.4. Thus

$$\max(\sigma_1, \sigma_2) \leq \min(\tau_1, \tau_2)$$

If there is a $y \in W$ with $p(x_0 + y) < +\infty$, then $\tau_1$ is not $+\infty$, so $\sigma_1$ and $\sigma_2$ are not $+\infty$; if there is a $y \in W$ with $p(-x_0 - y) < +\infty$, then $\sigma_1$ is not $-\infty$, so $\tau_1$ and $\tau_2$ are not $-\infty$. A similar statement holds for $q$. Thus, if $p$ or $q$ is finite, we have that $\max(\sigma_1, \sigma_2) \neq +\infty$ and $\min(\tau_1, \tau_2) \neq -\infty$.

Choose any $c \in \mathbb{R}$ such that

$$\max(\sigma_1, \sigma_2) \leq c \leq \min(\tau_1, \tau_2)$$

1738
We will show that this $c$ works, i.e. that $\theta(y + \alpha x_0) = \psi(y) + \alpha c$ satisfies
\[
q(y + \alpha x_0) \leq \psi(y) + \alpha c \leq p(y + \alpha x_0)
\]
for all $y \in W$ and $\alpha \in \mathbb{R}$.

Let $x \in X$. Write $x = y + \alpha x_0$. If $\alpha = 0$, then $x = y \in W$ and $q(x) \leq \theta(x) = \psi(x) \leq p(x)$ by assumption. If $\alpha > 0$, we have $\alpha^{-1}y \in W$, so
\[
c \leq \tau_1 \leq p(x_0 + \alpha^{-1}y) - \psi(\alpha^{-1}y)
\]
\[
\alpha c \leq \alpha p(x_0 + \alpha^{-1}y) - \alpha \psi(\alpha^{-1}y) = p(\alpha x_0 + y) - \psi(y) = p(x) - \psi(y)
\]
\[
\theta(x) = \psi(y) + \alpha c \leq p(x).
\]
Similarly using $\sigma_2$ we obtain $q(x) \leq \theta(x)$.

If $\alpha < 0$, we similarly have
\[
-p(-x_0 - \alpha^{-1}y) - \psi(\alpha^{-1}y) \leq \sigma_1 \leq c
\]
\[
-q(x) + \psi(y) = -p(\alpha x_0 + y) + \psi(y) = (-\alpha)[-p(-x_0 - \alpha^{-1}y)] - (-\alpha)\psi(\alpha^{-1}y) \leq -\alpha c
\]
\[
\theta(x) = \psi(y) + \alpha c \leq p(x).
\]
In the same manner using $\tau_2$ we have $q(x) \leq \theta(x)$.

\[\boxdot\]

XV.5.1.9. In the case of XV.5.1.3, where $V/W$ is countably generated, we can iterate XV.5.1.8 inductively to extend $\psi$ to all of $V$ (the Countable AC is technically needed to do this; if $V$ itself is countably generated, the countable AC can be dispensed with). The same construction can be done transfinitely if a set of generators for $V/W$ can be well ordered; assuming the Well-Ordering Principle (i.e. assuming AC: cf. ()), the proof of the general version of XV.5.1.3. can be completed this way.

XV.5.1.10. We will give an alternate argument using Zorn’s Lemma to finish the proof of XV.5.1.3. in the general case. Let $\mathcal{E}$ be the set of pairs $(X, \theta)$, where $X$ is a subspace of $V$ containing $W$ and $\theta$ is a linear functional on $X$ extending $\psi$, with $q(x) \leq \theta(x) \leq p(x)$ for all $x \in X$. Put a partial ordering $\preceq$ on $\mathcal{E}$ by setting $(X_1, \theta_1) \preceq (X_2, \theta_2)$ if and only if $X_1 \subseteq X_2$ and $\theta_2$ extends $\theta_1$.

If $\mathcal{C} = \{(X_i, \theta_i) : i \in I\}$ is a chain in $\mathcal{E}$, set $X = \bigcup_i X_i$. If $x, y \in X$, then $x \in X_i$ and $y \in X_j$ for some $i, j$; but either $X_i \subseteq X_j$ or $X_j \subseteq X_i$ since $\mathcal{C}$ is a chain, so either $x, y \in X_i$ or $x, y \in X_j$; in either case $x + y \in X$, so $X$ is closed under addition. $X$ is obviously closed under scalar multiplication, so $X$ is a subspace of $V$. Define $\theta : X \to \mathbb{R}$ by $\theta(x) = \theta_i(x)$ if $x \in X_i$: this function is well defined since the $\theta_i$ all agree wherever they are defined. It is routine to check that $\theta$ is linear. Thus $(X, \theta)$ is an upper bound for $\mathcal{C}$ in $\mathcal{E}$.

Thus Zorn’s lemma applies to give a maximal element $(X, \theta)$ in $\mathcal{E}$. But we must have $X = V$, for otherwise if $x_0 \in V \setminus X$ Lemma XV.5.1.8. gives an extension of $\theta$ to the span of $X$ and $x_0$ with the right properties, contradicting maximality of $(X, \theta)$. Then $\phi = \theta$ is the desired extension.

This completes the proof of Theorem XV.5.1.3..

XV.5.1.11. The proof we have given of the Hahn-Banach Theorem uses the AC. But the full AC is not necessary; one can get away with weaker assumptions such as the BPI ( ). Even this assumption is not necessary: the Hahn-Banach Theorem is logically equivalent (over ZF) to the Measure Extension Property ( ), which is strictly weaker than the BPI [ ]. The Hahn-Banach Theorem is not a theorem of ZF, i.e. some form of Choice is necessary.

1739
**XV.5.2. Extension of Bounded Linear Functionals**

One of the most common special cases of XV.5.1.4. is extension of bounded linear functionals from subspaces of normed vector spaces without increasing the norm:

**XV.5.2.1. Theorem.** [Hahn-Banach Theorem, Version 2] Let $V$ be a real vector space, and $\| \cdot \|$ a norm on $V$. Let $W$ be a subspace of $V$, and $\psi$ a bounded linear functional on $W$. Then there is a bounded linear functional $\phi$ on $V$ extending $\psi$ with $\| \phi \| = \| \psi \|$.

**Proof:** Apply XV.5.1.4. with $p(x) = \| \psi \| \| x \|$.

**XV.5.2.2. Note** that no completeness is assumed in this result, and the subspace $W$ need not be closed in $V$. (This is not a significant observation, though, since a bounded linear functional can automatically be extended to the completion or closure without increase in norm by uniform continuity.)

**XV.5.2.3. Corollary.** Let $V \neq \{0\}$ be a normed real vector space, and $x \in V$. Then there is a bounded linear functional $\phi$ of norm 1 on $V$ with $\phi(x) = \| x \|$. In particular,

$$\| x \| = \max\{\| \phi(x) \| : \phi \in V^*, \| \phi \| = 1\} = \sup\{\| \phi(x) \| : \phi \in V^*, \| \phi \| = 1\}.$$  

**Proof:** Suppose $x \neq 0$. Let $W$ be the one-dimensional subspace of $V$ spanned by $x$. Define $\psi : W \to \mathbb{R}$ by $\psi(\alpha x) = \alpha \| x \|$. Then $\psi$ is a bounded linear functional of norm 1 on $W$ with $\psi(x) = \| x \|$. Extend $\psi$ to a bounded linear functional of norm 1 on $V$.

If $x = 0$, let $\phi$ be any bounded linear functional on $V$ of norm 1 (such a linear functional exists by the first part of the proof applied to any nonzero vector in $V$).

**XV.5.2.4. This result shows** that if $V$ is a normed vector space, then $V^*$ is “large” (large enough to determine the norm on $V$). Without this result it is not obvious that a normed vector space has any nonzero bounded linear functionals.

**The Canonical Embedding into the Second Dual**

**XV.5.2.5. If** $V$ is a normed (real) vector space, then $V^*$ is a Banach space ($\cdot$). $V^*$ itself has a dual space $V^{**} = (V^*)^*$. If $x \in V$, then there is a canonically associated $\hat{x} \in V^{**}$ defined by

$$\hat{x}(\phi) = \phi(x)$$

for $\phi \in V^*$. It is easily checked that $\hat{x}$ is a linear functional on $V^*$, and that $x \mapsto \hat{x}$ is linear in $x$. Since $|\hat{x}(\phi)| = |\phi(x)| \leq \| x \| \| \phi \|$ for any $\phi \in V^*$, $\hat{x}$ is bounded and $\| \hat{x} \| \leq \| x \|$. By XV.5.2.3., we actually have $\| \hat{x} \| = \| x \|$. So:
XV.5.2.6. **Corollary.** Let $V$ be a normed real vector space. The map $x \mapsto \hat{x}$ is an isometric linear embedding of $V$ into $V^{**}$, called the canonical embedding.

XV.5.2.7. $V$ is said to be reflexive if the canonical map from $V$ to $V^{**}$ is surjective. Since $V^{**}$ is a Banach space, if $V$ is reflexive it must be complete (a Banach space). Some Banach spaces are reflexive (), (1) while others are not ()

XV.5.3. The Complex Hahn-Banach Theorem

XV.5.4. Separation of Convex Sets

XV.5.4.1. **Example.** Here is an interesting “folklore” example of a pair of convex sets which cannot be separated by a linear functional.

Let $X = c_{00} ()$. Let $A$ be the set of nonzero sequences in $X$ whose last nonzero term is positive. It is obvious that $A$ is convex. Similarly, let $B = -A$ be the set of nonzero sequences whose last nonzero term is negative. Then $B$ is convex, and $X$ is the disjoint union of $A$, $B$, and $\{0\}$.

If $X$ is given the uniform norm () (or any vector space topology), then $A$ and $B$ are dense in $X$: if $x \in X$, fix a $k$ larger than the index of the last nonzero term in $x$, and then $x + \epsilon e_k \in A$ for any $\epsilon > 0$, where $e_k$ is the $k$th standard basis vector (). A similar argument shows that $A$ and $B$ have no interior points.

Since $A$ and $B$ are dense in $X$, it is obvious they cannot be separated by a bounded linear functional. But in fact they cannot be separated by any nonzero linear functional, bounded or not, i.e. there is no nonzero linear functional $\phi$ on $X$ such that $\phi(x) \geq 0$ for all $x \in A$ (a separating linear functional or its negative would have to have this property). For if $\phi$ is such a linear functional, and $x \in X$ is arbitrary, fix $k$ larger than the index of the last nonzero term in $x$. Then $x + \epsilon e_k \in A$ for any $\epsilon > 0$, so

$$0 \leq \phi(x + \epsilon e_k) = \phi(x) + \epsilon \phi(e_k)$$

for any $\epsilon > 0$. Letting $\epsilon \to 0$, we obtain $\phi(x) \geq 0$. Since $x \in X$ is arbitrary, we also have $\phi(-x) \geq 0$, so $\phi(x) = 0$ for any $x$ and $\phi$ is the zero linear functional.

Note that we have referred to a norm topology on $X$, but actually this example is purely algebraic and can be written (except for the density of $A$ and $B$) without any reference to a topology on $X$.

XV.5.5. **Exercises**

XV.5.5.1. Here are two natural separation properties between separation and strict separation.

(a) If $X$ is a topological vector space and $A$ and $B$ disjoint convex sets in $X$, then $A$ and $B$ are completely separated if there is a hyperplane $H$ such that $A$ and $B$ are contained in distinct open half-spaces for $H$.

(b) If $X$ is a topological vector space and $A$ and $B$ disjoint convex sets in $X$, then $A$ and $B$ are fully separated if there is a hyperplane $H$ such that one of $A$ and $B$ is contained in an open half-space for $H$ and the other in the complementary closed half-space.

(c) Observe that strict separation $\Rightarrow$ complete separation $\Rightarrow$ full separation $\Rightarrow$ separation.

(d) Show that $A$ and $B$ are fully separated if and only if there is a continuous linear functional $\phi$ on $X$ such that $\phi(a) < \phi(b)$ for all $a \in A$, $b \in B$, and that $A$ and $B$ are completely separated if and only if there is a continuous linear functional $\phi$ and a real number $m$ such that $f(a) < m < f(b)$ for all $a \in A$, $b \in B$.
(e) Show that if $A$ and $B$ are convex sets in a topological vector space, and if one is open, then $A$ and $B$ are fully separated. If both are open, then they are completely separated. If $A$ and $B$ are not disjoint, then $A$ and $B$ cannot be completely separated, and if $A$ and $B$ are also not disjoint, then $A$ and $B$ cannot be fully separated.

(f) Let $X = \mathbb{R}^2$, $A$ the open upper half-plane along with the positive $x$-axis, and $B$ the lower half-plane along with the negative $x$-axis. Show that $A$ and $B$ are disjoint, convex, separated, but not fully separated.

(g) Let $X = \mathbb{R}^2$, $A$ the $x$-axis, and $B = \{(x, y) : x > 0, y \geq \frac{1}{2}\}$. Show that $A$ and $B$ are disjoint closed convex sets which are fully separated but not completely separated.

(h) Show that if $A$ and $B$ are disjoint closed convex subsets of $\mathbb{R}^2$, then $A$ and $B$ are fully separated.

(i) [KN76, 3G] Let $X = \mathbb{R}^3$,

$$A = \{(x, y, z) : x \geq 0, y \geq 0, z \geq 0, z^2 \leq xy\}$$

$$B = \{(0, y, 1) : y \in \mathbb{R}\}.$$

Show that $A$ and $B$ are disjoint, closed, convex, and separated, but not fully separated.

Thus full or complete separation is not always possible, even in finite-dimensional spaces.
XV.6. The Open Mapping Theorem, Closed Graph Theorem, and Uniform Boundedness Theorem

These are three of the most basic theorems of functional analysis. The first two are essentially one theorem, in the sense that either of them can be derived relatively easily from the other. The Uniform Boundedness Theorem is a consequence of the other two, or can be proved independently (it is apparently a “lower-level theorem” since the other two theorems cannot be obviously deduced from Uniform Boundedness). All of them are a consequence of the Baire Category Theorem, although they are not immediate corollaries and take some considerable additional work to prove.

XV.6.1. Open Mappings

XV.6.1.1. Recall () that if \( f \) is a function from a topological space \( X \) to a topological space \( Y \), then \( f \) is an **open mapping** if \( f(U) \) is open in \( Y \) for every open set \( U \) in \( X \). An open mapping does not have to be continuous. It is unusual for a continuous function, even a surjective one, to be an open mapping; however, we will show that a surjective bounded linear transformation between Banach spaces is an open mapping.

In this subsection, we give alternate characterizations of when linear transformations are open mappings. The arguments make repeated use of the following principle which is an immediate result of linearity:

**XV.6.1.2. [Basic Principle of Open Balls]** Let \( X \) and \( Y \) be normed vector spaces, and \( T : X \to Y \) a linear transformation. Then the images under \( T \) of all open balls in \( X \) look essentially the same, up to scaling and translation.

**XV.6.1.3. Proposition.** Let \( X \) and \( Y \) be normed vector spaces, and \( T : X \to Y \) a linear transformation (not necessarily bounded). The following are equivalent:

(i) \( T \) is an open mapping.

(ii) The image under \( T \) of some open ball in \( X \) contains an open ball in \( Y \).

(iii) The image under \( T \) of every open ball in \( X \) contains an open ball in \( Y \).

(iv) The image under \( T \) of some open ball in \( X \) centered at 0 contains an open ball in \( Y \) centered at 0.

(v) The image under \( T \) of every open ball in \( X \) centered at 0 contains an open ball in \( Y \) centered at 0.

**Proof:** (i) \(\Rightarrow\) (iii) \(\Rightarrow\) (ii) and (i) \(\Rightarrow\) (iv) \(\Rightarrow\) (ii) are trivial (since \( T(0) = 0 \)). (ii) \(\Rightarrow\) (iii) is an instance of the Basic Principle of Open Balls: if \( T(B_r(x_0)) \) contains \( B_r(y_0) \), then for any \( x_1 \) we have that \( T(B_r(x_1)) \) contains \( B_r(y_1) \), where \( y_1 = y_0 + T(x_1 - x_0) \). In particular, \( T(B_r(0)) \) contains \( B_r(y_2) \), where \( y_2 = y_0 - T(x_0) \). Then, for any \( s \), we have \( T(B_s(0)) \) contains \( B_{s\delta}(\tfrac{\delta}{s}y_2) \), where \( \delta = \frac{s}{r} \varepsilon \), and, for any \( x_1 \), \( T(B_s(x_1)) \) contains \( B_{s\delta}(T(x_1) + \tfrac{\delta}{s}y_2) \). The proof of (iv) \(\Rightarrow\) (v) is similar but easier.

To prove (iii) \(\Rightarrow\) (iv), suppose \( T(B_r(0)) \) contains \( B_{\varepsilon}(y_0) \). Then there is an \( x_0 \in B_r(0) \) with \( T(x_0) = y_0 \). Then \( T(B_r(-x_0)) \) contains \( B_r(0) \). If \( s = r + ||x_0|| \), then \( B_{r}(x_0) \subseteq B_s(0) \), so \( T(B_s(0)) \) contains \( B_{\varepsilon}(0) \).

To finish the proof, it suffices to show (v) \(\Rightarrow\) (i). Let \( U \) be an open set in \( X \), and \( y_0 \in T(U) \). Let \( x_0 \in U \) with \( T(x_0) = y_0 \), and \( \delta > 0 \) for which \( B_{\delta}(x_0) \subseteq U \). There is an \( \varepsilon > 0 \) such that \( T(B_{\delta}(0)) \) contains \( B_{\varepsilon}(0) \) by
(v). Then $T(B_{\delta}(x_0))$ contains $B_\epsilon(y_0)$, so $T(U)$ contains $B_\epsilon(y_0)$. Since $y_0 \in T(U)$ is arbitrary, $T(U)$ is open.

A slight modification of a subset of the proof of XV.6.1.3. (adding the words “the closure of” in a few places) gives the following:

**XV.6.1.4. PROPOSITION.** Let $X$ and $Y$ be normed vector spaces, and $T : X \to Y$ a linear transformation (not necessarily bounded). The following are equivalent:

(ii) The closure of the image under $T$ of some open ball in $X$ contains an open ball in $Y$.

(iii) The closure of the image under $T$ of every open ball in $X$ contains an open ball in $Y$.

(iv) The closure of the image under $T$ of some open ball in $X$ centered at 0 contains an open ball in $Y$ centered at 0.

(v) The closure of the image under $T$ of every open ball in $X$ centered at 0 contains an open ball in $Y$ centered at 0.

**XV.6.2. The Open Mapping Theorem**

The main result of this subsection is:

**XV.6.2.1. THEOREM.** [OPEN MAPPING THEOREM] Let $X$ and $Y$ be Banach spaces, and $T : X \to Y$ a surjective bounded operator. Then $T$ is an open mapping.

A crucial special case is when $T$ is injective as well as surjective:

**XV.6.2.2. COROLLARY.** [BOUNDED INVERSE THEOREM] Let $X$ and $Y$ be Banach spaces, and $T : X \to Y$ a bijective bounded operator. Then $T^{-1} : Y \to X$ is also bounded. Equivalently, $T$ is bounded below, i.e. a homeomorphism.

**Proof:** $T^{-1}$ is continuous, hence bounded by (), if and only if $T$ is an open mapping, which follows from the Open Mapping Theorem.
We can derive the Open Mapping Theorem from the Bounded Inverse Theorem, as follows. Let $X$ and $Y$ be Banach spaces and $T : X \rightarrow Y$ a surjective bounded operator. Let $N$ be the null space of $T$; then $N$ is a closed subspace of $X$, so we may form the quotient space $X/N$ which is a Banach space under the quotient norm. Let $P : X \rightarrow X/N$ be the projection map. Then there is an injective bounded operator $\tilde{T} : X/N \rightarrow Y$ with $T = \tilde{T} \circ P$. $\tilde{T}$ is an open mapping by the Bounded Inverse Theorem. $P$ is also an open mapping: this is a general fact about quotient spaces, but in this instance it is also an immediate corollary of XV.6.1.3., since it follows from the definition of the quotient norm that the image of the open unit ball in $X$ is exactly the open unit ball in $X/N$. Thus $T$ is a composition of two open mappings, hence also an open mapping. Hence the Open Mapping Theorem and the Bounded Inverse Theorem are equivalent.

Although the Open Mapping Theorem can be easily deduced from the Bounded Inverse Theorem, there does not appear to be a proof of the Bounded Inverse Theorem which is simpler than the proof of the Open Mapping Theorem.

Examples. All the hypotheses of the Open Mapping Theorem are necessary in general, as the following examples show. These examples illustrate that the statement of the Open Mapping Theorem is a nontrivial assertion which may even be considered somewhat surprising.

(i) Let $X = Y = \ell^2$, and define $T : X \rightarrow Y$ by

$$T((\xi_1, \xi_2, \ldots)) = \left(\xi_1, \frac{1}{2}\xi_2, \ldots, \frac{1}{n}\xi_n, \ldots\right).$$

Then $T$ is injective and has dense range, but is not surjective. $T$ is bounded ($\|T\| = 1$). $T$ is not an open map since no positive scalar multiple of $(1, \frac{1}{2}, \ldots)$ is in the range. Thus the hypothesis that $T$ is surjective is necessary. (In fact, a linear map which is not surjective can never be an open map, since the range is a subspace and a proper subspace can never be an open set.)

(ii) Take the same example as in (i) but let $Y$ be the range of $T$. $Y$ is a dense subspace of $\ell^2$ and inherits a norm from $\ell^2$, but is not complete. $T$ is now a bijection from $X = \ell^2$ to $Y$, but $T^{-1}$ is not bounded since $T$ is not bounded below. Thus the hypothesis that $Y$ be complete is necessary. (In fact, a linear map which is not surjective can never be an open map, since the range is a subspace and a proper subspace can never be an open set.)

(iii) It is trickier to give an example showing that the hypothesis that $X$ be a Banach space is necessary. Let $Z$ be a Banach space and $N$ a closed subspace of $Z$. Set $Y = Z/N$ with the quotient norm, and $P : Z \rightarrow Z/N$ the quotient map. Suppose $X$ is a subspace of $Z$ complementary to $N$ which is not closed (e.g. $X$ is the null space of a discontinuous linear functional on $Z$ and $N$ a complementary one-dimensional subspace; cf. also ). $X$ is a normed vector space with norm inherited from $Z$, and $T = P|_X$ is an injective bounded operator from $X$ onto $Y$. But $T^{-1} : Y \rightarrow X$ cannot be bounded, since if it were $X$ would be complete, hence closed in $Z$.

(iv) The hypothesis that $T$ be bounded is also necessary. Expand the standard orthonormal basis of $\ell^2$ to a Hamel basis, and take a permutation of the Hamel basis interchanging the orthonormal basis with a subset of the remaining Hamel basis vectors. This permutation extends to a bijective but unbounded operator $T$ from $\ell^2$ to $\ell^2$. If $T$ were an open mapping, then $T^{-1}$ would be continuous, hence bounded, so $T$ would be bounded by the Bounded Inverse Theorem, a contradiction.

Note that some form of the Axiom of Choice is needed in (iii) and (iv). (Actually, some form of Choice is even needed for the Open Mapping Theorem, since the Baire Category Theorem depends on DC.)
We now give the proof of the Open Mapping Theorem.

**Proof:** Write $B^X_r$ and $B^Y_r$ for the open balls of radius $r$ around 0 in $X$ and $Y$ respectively. We have that

$$
\bigcup_{n=1}^{\infty} T(B^X_n) = T(X) = Y
$$

since $T$ is surjective. Thus, since $Y$ is a Banach space, by the Baire Category Theorem $T(B^X_n)$ is not nowhere dense in $Y$ for some $n$, i.e. $\overline{T(B^X_n)}$ contains an open ball in $Y$. Thus by XV.6.1.4., $T(B^X_r)$ contains an open ball around 0 in $Y$ for any $r$. In particular, there is an $\epsilon > 0$ such that $T(B^X_1)$ contains $B^Y_\epsilon$. We will show that $T(B^X_n)$ contains $B^Y_\epsilon$; it will then follow from XV.6.1.3. that $T$ is an open mapping.

Let $y \in B^Y_\epsilon$; we will construct an $x \in X$ with $\|x\| < 2$ and $Tx = y$. A key observation is that $\overline{T(B^X_1)}$ contains $B^Y_r$ for every $r > 0$. Since $y$ is in the closure of $T(B^X_1)$, for any $\delta > 0$ we can find an $x_0 \in X$, $\|x_0\| < 1$, with $\|y - Tx_0\| < \delta$. Let $x_1 \in X$, $\|x_1\| < 1$, with

$$
\|y - Tx_1\| < \frac{\epsilon}{2}.
$$

Since $y - Tx_1$ is in the closure of $T(B^X_{1/2})$, there is an $x_2 \in X$, $\|x_2\| < \frac{1}{2}$, with

$$
\| (y - Tx_1) - Tx_2 \| < \frac{\epsilon}{4}.
$$

Continuing inductively, we obtain for each $n$ an $x_n \in X$ with $\|x_n\| < 2^{-n+1}$ and

$$
\| y - Tx_1 - Tx_2 - \cdots - Tx_n \| < 2^{-n} \epsilon.
$$

The infinite series

$$
\sum_{k=1}^{\infty} x_k
$$

converges absolutely, hence converges to some $x \in X$ since $X$ is a Banach space (). We have

$$
\|x\| < \sum_{k=1}^{\infty} 2^{-k+1} = 2.
$$

Since $T$ is continuous, we have

$$
Tx = \sum_{k=1}^{\infty} Tx_k
$$

in $Y ()$. But we have

$$
\left\| y - \sum_{k=1}^{n} Tx_k \right\| < 2^{-n} \epsilon
$$

for all $n$, and thus the infinite series

$$
\sum_{k=1}^{\infty} Tx_k
$$

converges to $y$, i.e. $Tx = y$, completing the proof.

$\blacksquare$
XV.6.3. The Closed Graph Theorem
XV.6.3.1. Recall () that if \( X \) and \( Y \) are sets and \( f : X \to Y \) is a function, the set
\[
\Gamma(f) = \{(x,y) \in X \times Y : y = f(x)\}
\]
is the graph of \( f \). (Technically, using the ordered pair definition of function, a function is its graph.) If \( X \) and \( Y \) are topological spaces, then \( f \) has closed graph if \( \Gamma(f) \) is closed in \( X \times Y \).

XV.6.3.2. Proposition. If \( X \) and \( Y \) are topological spaces and \( f : X \to Y \) is a function, then \( f \) has closed graph if and only if, whenever \((x_i)\) is a net in \( X \) with \( x_i \to x \) and \( f(x_i) \to y \) for some \( x \in X \) and \( y \in Y \), we have \( y = f(x) \).

Proof: By definition, \( f \) has closed graph if and only if, whenever \((x_i, y_i)\) is a net in \( \Gamma(f) \) converging to \((x,y) \in X \times Y \), we have \((x,y) \in \Gamma(f)\). The statement of the proposition is a simple rephrasing.

A continuous function between Hausdorff spaces has closed graph (), but a function between Hausdorff spaces which has closed graph need not be continuous (). However, we have the following, which is the main result of this subsection:

XV.6.3.3. Theorem. [Closed Graph Theorem] Let \( X \) and \( Y \) be Banach spaces, and \( T : X \to Y \) be a linear transformation. Then \( T \) is bounded (continuous) if and only if \( T \) has closed graph.

Proof: If \( T \) is continuous, it has closed graph by (). Conversely, suppose \( T \) has closed graph. We have that \( X \times Y \) is a Banach space under a number of equivalent norms (). Fix one of these norms, say
\[
\|(x,y)\| = \max(\|x\|, \|y\|)
\]
Since \( T \) is linear, \( \Gamma(T) \) is a subspace of \( X \times Y \); by assumption, it is closed, hence a Banach space under the induced norm. The map \( S : \Gamma(T) \to X \) given by \( S((x,Tx)) = x \) is bijective and linear, and bounded (\( \|S\| \leq 1 \)). Thus by the Bounded Inverse Theorem \( S^{-1} : X \to \Gamma(T) \) is bounded. If \( P \) is the projection from \( X \times Y \) to \( Y \), then \( P \) is bounded, and hence \( T = P \circ S^{-1} \) is also bounded.

XV.6.3.4. The proof shows that the Closed Graph Theorem is a simple corollary of the Bounded Inverse Theorem. The converse is true too: if \( X \) and \( Y \) are Banach spaces and \( T : X \to Y \) is a continuous (bounded) bijection, then \( \Gamma(T) \) is closed. Thus \( \Gamma(T^{-1}) \) is also closed since it is essentially the same as \( \Gamma(T) \) with coordinates reversed. So \( T^{-1} \) is bounded by the Closed Graph Theorem.

We thus obtain that the Open Mapping Theorem, the Bounded Inverse Theorem, and the Closed Graph Theorem are equivalent statements by simple arguments.

XV.6.3.5. The importance of the Closed Graph Theorem is that it is often much easier to prove that a map \( T : X \to Y \) has closed graph than to directly prove that it is continuous or bounded. To prove directly that \( T \) is continuous, it must be shown that if \((x_n)\) is a sequence in \( X \) with \( x_n \to x \), then the sequence \((Tx_n)\) converges to \( Tx \). The sequence \((Tx_n)\) cannot be assumed to converge or even to have a convergent subsequence. But to show that \( T \) has closed graph, one only needs to show that if \((x_n)\) is a sequence in \( X \) with \( x_n \to x \), and \( Tx_n \to y \) for some \( y \in Y \), then \( y = Tx \), i.e. the sequence \((Tx_n)\) can be assumed to converge to something. The next result is a good illustration of the use of the Closed Graph Theorem.
XV.6.3.6. **Proposition.** Let $H_1$ and $H_2$ be Hilbert spaces, and $T : H_1 \to H_2$ a linear operator. If $T$ has an adjoint $T^* : H_2 \to H_1$, i.e. $T^*$ is a function (which is automatically a linear operator) satisfying

$$\langle x, T^*y \rangle_{H_1} = \langle Tx, y \rangle_{H_2}$$
for all $x \in H_1$, $y \in H_2$, then $T$ is bounded (and hence $T^*$ is also bounded).

**Proof:** We need only show that $T$ has closed graph. Suppose $(x_n)$ is a sequence in $H_1$ with $x_n \to x$ and $Tx_n \to y$ for some $y \in H_2$. Then, for any $z \in H_2$, we have

$$\langle Tx_n, z \rangle_{H_2} \to \langle y, z \rangle_{H_2}$$
by continuity of the inner product in $H_2$ (iii). On the other hand, we have

$$\langle Tx_n, z \rangle_{H_2} = \langle x_n, T^*z \rangle_{H_1} \to \langle x, T^*z \rangle_{H_1} = \langle Tx, z \rangle_{H_2}$$
by continuity of the inner product in $H_1$. Thus by uniqueness of limits we have

$$\langle y, z \rangle_{H_2} = \langle Tx, z \rangle_{H_2}$$
for every $z \in H_2$, and so $y = Tx$ by (iii).

XV.6.3.7. **Caution:** Our convention is that “$T$ is a linear map from $X$ to $Y$” means that the domain of $T$ is all of $X$. In some references, this phrase means only that the domain of $T$ is a subspace (usually dense) of $X$. The Closed Graph Theorem does not apply in this situation; in fact, such an operator can be unbounded but have closed graph. Densely defined unbounded operators with closed graph, called **closed operators**, are very important; see section (iii). Proposition XV.6.3.6. is also false in general for operators which are just densely defined (with densely defined adjoint); cf. (iv).

XV.6.3.8. Since in the statement of the Closed Graph Theorem $T$ is not required to be surjective, one might think the hypothesis that $Y$ be complete is unnecessary: $Y$ could just be replaced by its completion $\hat{Y}$. But this is incorrect. The graph of $T$ can be a closed subset of $X \times Y$ but not closed in $X \times \hat{Y}$.

For a counterexample, let $X$, $Y$, $T$ be as in XV.6.2.4.(iii). Then the graph of $T$ is closed in $X \times Y$ since $T$ is continuous, so the graph of $T^{-1} : Y \to X$ is closed in $Y \times X$. But $T^{-1}$ is not bounded.

XV.6.4. **The Uniform Boundedness Theorem**

The Uniform Boundedness Theorem is a theorem which states that a family of bounded operators on a Banach space which is “pointwise bounded” is actually uniformly bounded. This result, like the Open Mapping Theorem and Closed Graph Theorem, has many uses. It can be obtained as a simple consequence of the Closed Graph Theorem or proved directly from scratch; we will give three arguments.

1748
XV.6.4.1. **Theorem.** [Uniform Boundedness Theorem] Let \( X \) be a Banach space, \( \{Y_i : i \in I\} \) a collection of normed vector spaces, and \( T_i : X \to Y_i \) a bounded operator for each \( i \). Suppose, for each \( x \in X \), the set
\[
\{\|T_i(x)\| : i \in I\}
\]
is bounded (the bound may depend on \( x \)). Then the set
\[
\{\|T_i\| : i \in I\}
\]
is bounded.

XV.6.4.2. In other words, if for each \( x \in X \) there is a constant \( M_x \) such that \( \|T_i x\| \leq M_x \) for all \( i \in I \), then there is a single constant \( M \) such that \( \|T_i\| \leq M \) for all \( i \in I \).

The theorem is commonly applied in the case where all the \( Y_i \) are the same, and/or when \( I = \mathbb{N} \), but these restrictions do not make the proof any easier.

XV.6.4.3. **Proof:** Our first proof deduces the theorem from the Closed Graph Theorem.

First note that since the \( T_i \) are not assumed to be surjective, there is no harm in replacing \( Y_i \) by its completion. Thus we may assume without loss of generality that each \( Y_i \) is a Banach space.

Let \( Y \) be the \( \ell^\infty \) product of the \( Y_i \). Then \( Y \) is a Banach space. Define \( T : X \to Y \) by
\[
T(x) = (\cdots T_i(x) \cdots)
\]
(the pointwise boundedness hypothesis exactly means that \( T(x) \in Y \) for all \( x \in X \)). Then \( T \) is a linear operator; \( T \) is bounded if and only if the \( T_i \) are uniformly bounded (in fact, \( \|T\| = \sup_i \|T_i\| \)).

To show \( T \) is bounded, it suffices to show that \( T \) has closed graph. But this is easy. Suppose \( (x_n) \) is a sequence in \( X \) and \( x_n \to x \) and \( T x_n \to y \) for some \( x \in X \), \( y \in Y \). Set \( y = (\cdots y_i \cdots) \). Since convergence in \( Y \) implies coordinatewise convergence, we have \( T_i x_n \to y_i \) for all \( i \). But since \( T_i \) is bounded, we have \( T_i x_n \to T_i x \), i.e. \( y_i = T_i x \) for all \( i \), and thus \( y = T x \).

XV.6.4.4. **Proof:** We give a second direct proof from the Baire Category Theorem.

For \( n \in \mathbb{N} \), define
\[
A_n = \{x \in X : \|T_i x\| \leq n \text{ for all } i \in I\}
\]
and note that \( A_n \) is a closed subset of \( X \) (\( A_n = \bigcap_{i \in I} T_i^{-1}(\{y \in Y_i : \|y\| \leq n\}) \)).

By the pointwise boundedness assumption, \( \cup_n A_n = X \), so by the Baire Category Theorem some \( A_n \) has nonempty interior. Thus there is an \( x_0 \in X \) and \( \epsilon > 0 \) such that \( \|T_i x\| \leq n \) whenever \( \|x - x_0\| < \epsilon \). Let \( m \) be a number such that \( \|T_i(x_0)\| \leq m \) for all \( i \in I \). Then, if \( \|x\| < \epsilon \), we have \( \|x + x_0 - x_0\| < \epsilon \), so, for all \( i \),
\[
\|T_i(x + x_0)\| \leq n
\]
\[
\|T_i(x + x_0) - T_i(x_0)\| = \|T_i(x + x_0) - T_i(x_0)\| \leq \|T_i(x + x_0)\| + \|T_i(x_0)\| \leq n + m
\]
which implies that
\[
\|T_i\| \leq \frac{n + m}{\epsilon}
\]
for all \( i \).
**XV.6.4.5.** Proof: Our third proof is the most elementary, using a “gliding hump” argument. This proof is due to Hausdorff [Hau32] (cf. [Hen80]).

Suppose \( \{\|T_i\| : i \in I\} \) is unbounded. Inductively define sequences \((x_n)\) in \( X \) and \( i_n \in I \) as follows. Suppose \( x_k \) and \( i_k \) have been defined for \( k < n \). Set \( M_0 = 0 \), and if \( n > 1 \) set \( M_{n-1} = \sup_i \{T_i(x_1 + \cdots + x_{n-1})\} \).

Now choose \( i_n \) such that
\[
\|T_{i_n}\| > 3 \cdot 4^n |n + M_{n-1}|
\]
and \( x_n \) such that
\[
\|x_n\| = 4^{-n} \quad \text{and} \quad \|T_{i_n}x_n\| > \frac{2}{3} \|T_{i_n}\| \|x_n\|.
\]

Set
\[
x_n = \sum_{n=1}^{\infty} x_n
\]
(note that the series converges absolutely, hence converges since \( X \) is complete). For each \( n \) set
\[
z_n = x_1 + \cdots + x_{n-1}
\]
(set \( z_1 = 0 \))
\[
y_n = \sum_{k=n+1}^{\infty} x_k
\]
and note that \( \|y_n\| \leq \frac{4^{-n}}{3} = \frac{1}{3} \|x_n\| \). We then have
\[
\|T_{i_n}x_n\| > 2 |n + M_{n-1}|
\]
\[
\|T_{i_n}y_n\| \leq \|T_{i_n}\| \|y_n\| \leq \frac{1}{3} \|T_{i_n}\| \|x_n\| < \frac{1}{2} \|T_{i_n}x_n\|
\]
for each \( n \). Thus for each \( n \) we have
\[
\|T_{i_n}x\| = \|T_{i_n}(x_n + y_n + z_n)\| \geq \|T_{i_n}x_n\| - \|T_{i_n}y_n\| - \|T_{i_n}z_n\| > |n + M_n| - M_n = n
\]
which contradicts that \( \{\|T_ix\| : i \in I\} \) is bounded.

**XV.6.5. Exercises**

**XV.6.5.1.** Let \( X \) and \( Y \) be normed vector spaces, with \( X \) complete, and \( T : X \rightarrow Y \) a linear operator.

(a) Show that \( T \) is bounded if and only if \( T^{-1}(B_Y) \) is closed, where \( B_Y \) is the closed unit ball of \( Y \). (One direction is trivial from continuity. Conversely, suppose \( T^{-1}(B_Y) \) is closed. Then \( T^{-1}(B) \) is closed for every closed ball \( B \) in \( Y \). Apply the Baire Category Theorem to conclude that \( T^{-1}(B_Y) \) contains an open ball.)

(b) Let \( T : c_00 \rightarrow c_00 \) be defined by \( T(\xi_1, \xi_2, \ldots) = (\xi_1, 2\xi_2, \ldots, n\xi_n, \ldots) \).

Show that \( T^{-1}(B_{c_00}) \) is closed. Thus the result does not necessarily hold if \( X \) is not complete.

(c) What is the relation between this property of \( T \) and having closed graph?
XV.7. Compact Convex Sets

A fairly detailed analysis can be given of the structure of compact convex sets in locally convex topological vector spaces. We describe the basics of this theory in this section; for more detailed analyses, see [Alf71], [?].

We will work throughout this section with real topological vector spaces. The results will immediately apply also to complex topological vector spaces, since a complex topological vector space has an underlying structure as a real topological vector space and none of the statements or proofs of the results make use of complex scalar multiplication. To avoid unnecessary subscripts, if \( K \) is a compact Hausdorff space, in this section write \( C(K) \) for the set \( C_{\mathbb{R}}(K) \) of real-valued continuous functions on \( K \).

XV.7.1. Faces and Extreme Points

The crucial notion underlying the whole analysis of compact convex sets is that of a face.

XV.7.1.1. Definition. Let \( K \) be a convex set in a vector space. A nonempty convex subset \( F \) of \( K \) is a face of \( K \) if, whenever \( x, y \in K \) and \( tx + (1 - t)y \in F \) for some \( t, 0 < t < 1 \), then both \( x \) and \( y \) are in \( F \). A one-point face is called an extreme point of \( K \). More precisely, \( p \) is an extreme point of \( K \) if \( \{p\} \) is a face of \( K \). Thus an extreme point of \( K \) is a point which cannot be written as a nontrivial convex combination of two distinct points in \( K \).

Note that this definition makes no use of any topology. However, faces and extreme points are most interesting for closed convex sets in topological vector spaces.

In some references, faces are called extreme subsets or supports.

XV.7.1.2. \( K \) itself is a face of \( K \); we are mainly interested in proper faces of \( K \). Roughly (imprecisely, and not quite correctly) speaking, a proper face of \( K \) is a convex subset of the “boundary” of \( K \).

XV.7.1.3. Examples. (i) In \( \mathbb{R} \), convex subsets are intervals. If \( I \) is an interval, the only faces are the extreme points; the extreme points of \( I \) are the endpoints of \( I \) which are contained in \( I \) (by definition, a face of a convex set is a subset of the convex set). Thus a nondegenerate closed bounded interval has two extreme points, an open interval has no extreme points, and other intervals have one extreme point.

(ii) Let \( K \) be a closed convex (solid) polygon in \( \mathbb{R}^2 \), e.g. a (solid) triangle or rectangle. Each closed edge of \( K \) is a face of \( K \), and each vertex is an extreme point, and these are all the proper faces.

(iii) Let \( K \) be a closed convex (solid) polyhedron in \( \mathbb{R}^3 \), e.g. a tetrahedron or cube. There are two-dimensional faces, the closed bounding (solid) polygons (which are usually called the “faces” of the polyhedron); one-dimensional faces, the closed edges; and 0-dimensional faces (extreme points), the vertices. These are all the proper faces. This example is undoubtedly the origin of the term “face” for general convex sets.

(iv) Let \( K \) be the closed unit ball in \( \mathbb{R}^n \). Then the only proper faces of \( K \) are the extreme points, which are precisely all the points on the unit sphere. Thus if \( n > 1 \) there are uncountably many extreme points.

(v) An open ball in \( \mathbb{R}^n \) has no extreme points. A nontrivial subspace of \( \mathbb{R}^n \) has no extreme points.

1751
**XV.7.1.4.** PROPOSITION. Let $K$ be a convex set in a vector space.

(i) Let \( \{ F_i : i \in I \} \) be a collection of faces of $K$. If \( \cap_i F_i \neq \emptyset \), then it is a face of $K$.

(ii) A face of a face of $K$ is a face of $K$.

(iii) If $C$ is a convex subset of $K$, $F$ is a face of $K$, and \( F \cap C \neq \emptyset \), then $F \cap C$ is a face of $C$.

**PROOF:** Straightforward and left to the reader.

**XV.7.1.5.** Thus, if $S$ is any nonempty subset of $K$, there is a smallest face $F(S)$ containing $S$, called the **face generated by** $S$. $F(S)$ is the intersection of all faces containing $S$.

We can define an **interior point** of $K$ to be any point $p$ of $K$ such that $F(p) = K$. Note that such a $p$ is not necessarily an interior point in the topological sense in general (indeed, there may be no topology in the picture!); but if $K$ is a convex set in a topological vector space which has a nonempty interior, then points in the topological interior are interior points in this sense.

**XV.7.1.6.** PROPOSITION. Let $X$ and $Y$ be vector spaces, and $T : X \to Y$ a linear transformation. If $K$ is a convex set in $Y$ and $F$ is a face of $K$, then $T^{-1}(F)$ is a face in $T^{-1}(K)$.

The simple proof is again left to the reader.

**XV.7.1.7.** The same result does not hold for images in general: if $T : X \to Y$ is linear, $K$ is a convex set in $X$, and $F$ is a face in $K$, then $T(F)$ is not necessarily a face in $T(K)$ (consider the case where $Y$ is one-dimensional).

**XV.7.1.8.** A particularly important case of XV.7.1.6. is when $Y = \mathbb{R}$. If $\phi$ is a linear functional on $X$, $K$ is a convex subset of $X$, $I = \phi(K)$ (an interval in $\mathbb{R}$), and $c$ is an endpoint of $I$ contained in $I$, then the intersection of $K$ with the hyperplane $\phi^{-1}(\{c\})$ is a face of $K$ (cf. XV.7.1.4.(iii)). In other words, if $\phi(x) \geq c$ (or $\phi(x) \leq c$) for all $x \in K$, and there is an $x \in K$ with $\phi(x) = c$, then the set

\[
\{x \in K : \phi(x) = c\}
\]

is a face of $K$. In this case, the hyperplane $\phi^{-1}(\{c\})$ is called a **support hyperplane** for $K$, and the intersection with $K$ a **support face**. Such a support face is frequently, although far from always, an extreme point of $K$.

**Closed Faces**

Now let $X$ be a topological vector space, and $K$ a convex subset of $X$. The faces of $K$ which are relatively closed in $K$ are particularly important, and are called the **closed faces** of $K$. This situation is especially important if $K$ is closed in $X$, although we do not make this assumption in the definition. Note that (if $X$ is Hausdorff) every extreme point is a closed face.

**XV.7.1.9.** It looks at first from the definition that a face should be closed; in fact, if $F$ is a face in $K$ and $Y$ is a one-dimensional (more generally, finite-dimensional) affine subspace of $X$, then $F \cap Y$ is always relatively closed in $K \cap Y$. However, a face of a convex set, even a compact convex set (Choquet simplex), need not be closed in general, and the closure of a face is not necessarily a face (XV.7.7.2.).
As in XV.7.1.4 (i), an arbitrary intersection of closed faces of a convex set \( K \), if nonempty, is a closed face (the other parts of XV.7.1.4. have straightforward analogs for closed faces too, with “face” replaced by “closed face” throughout). Thus if \( S \) is an arbitrary nonempty subset of \( K \), there is a smallest closed face \( F(S) \) of \( K \) containing \( S \), the intersection of all closed faces containing \( S \).

Similarly, if \( K \) is a convex set in a topological vector space \( X \), and \( \phi \) is a continuous linear functional on \( X \) (XV.7.1.8.) is a support face for \( K \) from \( \phi \).

### Convex Hulls and Extreme Points

**Proposition.** Let \( X \) be a topological vector space, and \( K_1, \ldots, K_n \) a finite collection of compact convex sets in \( X \). Let \( K \) be the convex hull of \( \bigcup_{j=1}^n K_j \). Then \( K \) is compact.

**Proof:** Since each \( K_j \) is convex, it is easy to check that

\[
\left\{ \sum_{j=1}^n \alpha_j x_j : x_j \in K_j, \alpha_j \geq 0, \sum_{j=1}^n \alpha_j = 1 \right\}
\]

is convex; since it is contained in \( K \) and contains each \( K_j \), it equals \( K \). Set

\[
S = \left\{ (\alpha_1, \ldots, \alpha_n) : \alpha_j \geq 0, \sum_{j=1}^n \alpha_j = 1 \right\}.
\]

Then \( S \) is a compact subset of \( \mathbb{R}^n \) (it is an \((n-1)\)-simplex), so the map

\[
(\alpha_1, \ldots, \alpha_n, x_1, \ldots, x_n) \mapsto \sum_{j=1}^n \alpha_j x_j
\]

is a continuous function from the compact set \( S \times K_1 \times \cdots \times K_n \) onto \( K \). Thus \( K \) is compact by (i).

The next result is known as Milman’s (partial) converse to the Krein-Milman Theorem. Our proof is adapted from [SW99].

**Theorem.** Let \( S \) be a subset of a locally convex (Hausdorff) topological vector space \( X \) such that the closed convex hull \( K \) of \( S \) is compact. Then all extreme points of \( K \) lie in \( S \).

**Proof:** Since \( \bar{S} \) is a subset of \( K \), it is compact. Let \( U \) be an open convex neighborhood of 0 in \( X \). For \( x \in \bar{S} \), set \( U_x = x + U \). The \( U_x \) form an open cover of \( \bar{S} \), so there is a finite subcover \( \{U_{x_1}, \ldots, U_{x_n}\} \). Let \( K_j \) be the closed convex hull of \( S \cap U_{x_j} \). Then \( K_j \subset K \) for all \( j \). The convex hull \( K' \) of \( \bigcup_{j=1}^n K_j \) contains \( S, K' \subset K \), and \( K' \) is compact by XV.7.1.12.; thus \( K' = K \). So if \( p \) is an extreme point of \( K \), we have \( p = \sum_{j=1}^n \alpha_j y_j \) for some \( \alpha_j \)'s in \( S \) and \( y_j \in K_j \); since \( p \) is extreme, we must have \( p = y_j \) for some \( j \), i.e. \( p \in K_j \). Since \( U_{x_j} \) is closed and convex, \( K_j \subset U_{x_j} \subset \bar{S} + U \). Thus \( p \in \bar{S} + U \). This is true for every \( U \), so \( p \in \bar{S} \).
XV.7.2. Compact Convex Sets in Finite-Dimensional Spaces

We now look at compact convex sets in finite-dimensional topological vector spaces. A finite-dimensional real vector space has a unique topology making it a Hausdorff topological vector space, which can be identified with $\mathbb{R}^n$ for some $n$. So without loss of generality we will work in Euclidean space. We can thus use geometric arguments in our analysis.

XV.7.2.1. If $K$ is a convex set in $\mathbb{R}^n$, then $K$ spans an $m$-dimensional affine subspace of $\mathbb{R}^n$ for some $m \leq n$; we say $K$ is $m$-dimensional (this terminology is justified in ??). By replacing $\mathbb{R}^n$ with this subspace and identifying it with $\mathbb{R}^m$, we may if desired for analysis of $K$ assume it spans all of $\mathbb{R}^n$ as an affine space (i.e. is $n$-dimensional), although we will not routinely make this restriction. Dimension in this sense is an isomorphism invariant of $K$.

XV.7.2.2. Proposition. If $K$ is an $n$-dimensional convex set in $\mathbb{R}^n$, then $K$ has nonempty (topological) interior $K^\circ$; in fact $K^\circ$ is an open convex set in $\mathbb{R}^n$ which is dense in $K$.

Proof: $K$ contains $n + 1$ points $x_0, \ldots, x_n$ which span $\mathbb{R}^n$ as an affine space (any set which spans $\mathbb{R}^n$ as an affine space contains $n + 1$ points which also span it). Set $x = \frac{1}{n+1} \sum_{k=0}^n x_k$. Every point $y$ of $\mathbb{R}^n$ can be uniquely written as $\sum_{k=0}^n \alpha_k x_k$ for $\alpha_k \in \mathbb{R}$, $\sum_{k=0}^n \alpha_k = 1$; if $y$ is sufficiently close to $x$, each $\alpha_k$ is close to $\frac{1}{n+1}$ and is hence positive. Thus $y \in K$ if $y$ is sufficiently close to $x$, i.e. $K$ contains an open ball around $x$. Thus $x$ is in $K^\circ$, so $K^\circ \neq \emptyset$. The density and convexity of $K^\circ$ follow from ()

This result fails for infinite-dimensional compact convex sets, since it implies that the spanned affine subspace contains a nonempty open set with compact closure.

Combining this with the geometric version of the Hahn-Banach Theorem in () (which is elementary in a finite-dimensional vector space, not requiring any form of Choice), we obtain:

XV.7.2.3. Proposition. Let $K$ be an $n$-dimensional convex set in $\mathbb{R}^n$. If $x \in K$ is in the topological boundary of $K$, then $x$ is contained in a proper face of $K$, and hence is not an interior point in the sense of XV.7.1.5.

Proof: By () applied to $K^\circ$ and $x$, there is a linear functional $\phi$ on $\mathbb{R}^n$, which is automatically continuous, and a number $r$ such that $\phi(x) \geq r$ and $\phi(y) < r$ for all $y \in K^\circ$. Since $\phi$ is continuous, $\phi(y) \leq r$ for all $y \in K^\circ$, and in particular $\phi(x) = r$ and $\phi(y) \leq r$ for all $y \in K$ since $K^\circ$ is dense in $K$. The set $F = \{ y \in K : \phi(y) = r \}$ is thus a proper face of $K$ containing $x$.

This result fails in the infinite-dimensional case, and even in the finite-dimensional case if the dimension of the convex set is strictly smaller than the dimension of the space.

Extreme Points

Our next goal is to show that a compact convex set $K$ in a finite-dimensional vector space has “many” extreme points, enough that $K$ is the convex hull of its extreme points. This result is a special case of the Krein-Milman Theorem, although we will obtain a sharper result with a more elementary proof in the finite-dimensional case.
**XV.7.2.4.**  Example. It can be shown that if $K$ is a compact convex set in $\mathbb{R}^2$, then the set of extreme points of $K$ is closed in $K$, hence also compact. But this fails in $\mathbb{R}^3$. Let $K$ be the convex hull of the points of $C = \{(x, y, 0) : (x-1)^2 + y^2 = 1\}$ and the points $(0, 0, 1)$ and $(0, 0, -1)$; see Fig. (). The extreme points of $K$ are the points of $C$ except the origin, and the points $(0, 0, 1)$ and $(0, 0, -1)$. The origin is in the closure of the set of extreme points but is not itself an extreme point.

**XV.7.2.5.**  Proposition. An $n$-dimensional compact convex set $K$ has at least $n + 1$ extreme points. The extreme points of $K$ span the same affine subspace as $K$.

**XV.7.2.6.**  Lemma. A nonempty compact convex set in $\mathbb{R}^n$ has an extreme point.

**Proof:** Let $K$ be a compact convex set in $\mathbb{R}^n$, and $x \in K$. There is a point $y \in K$ of maximum distance $d$ from $x$ by compactness ($y$ is not necessarily unique). Then $y$ is an extreme point of the ball $B_d(x)$, which contains $K$, so $y$ is an extreme point of $K$.

This result and proof hold more generally in any strictly convex Banach space. The result is true in much greater generality (XV.7.3.1), but requires a different proof.

**Proof:** Proof of XV.7.2.5: let $K$ be an $n$-dimensional compact convex subset of $\mathbb{R}^n$. Let $A$ be the affine subspace spanned by the extreme points of $K$ (there is at least one extreme point by XV.7.2.6). If $A$ does not contain $K$, then there is a point $x$ of $K$ of maximum distance $d > 0$ from $A$. If $S$ is the set of points of $K$ of distance $d$ from $A$, then the face $F(\{x\})$ of $K$ generated by $x$ is contained in $S$. $F(\{x\})$ is a compact convex set and thus has an extreme point $y$ (XV.7.2.6), which must be an extreme point of $K$ also. But $y \in S$, so $y \notin A$, a contradiction. The first statement then follows since the affine subspace spanned by $m$ points has dimension $\leq m - 1$.

The next result is the main theorem of this subsection, due first to MINKOWSKI without the $n + 1$, and in the sharp form with $n + 1$ by CARATHÉODORY.

**XV.7.2.7.**  Theorem. Let $K$ be a compact convex set in $\mathbb{R}^n$. Then every point of $K$ is a convex combination of at most $n + 1$ extreme points of $K$.

**Proof:** By complete induction on the dimension $m$ of $K$. If $m = 1$, $K$ is isomorphic to a closed bounded interval and is the convex hull of its two endpoints (and the case $m = 0$ is trivial). For a general $K$ of dimension $m$, let $x \in K$. If $x$ is in the topological boundary of $K$, then $x$ lies in a proper face $F$ of $K$ (XV.7.2.3), and thus $x$ is a convex combination of at most $m$ extreme points of $F$ by the inductive hypothesis, all of which are extreme points of $K$. If $x$ is an interior point of $K$, let $y$ be an extreme point of $K$. The ray from $y$ through $x$ intersects the boundary of $K$ in a (unique) point $z$, which is in $K$ since $K$ is compact. Then as before $z$ is a convex combination of at most $m$ extreme points in a proper face of $K$. Since $x$ is a convex combination of $z$ and $y$, $x$ is a convex combination of $y$ and these at most $m$ other extreme points of $K$.

Combining this result with XV.7.1.13, we get:
XV.7.2.8. Corollary. Let $E$ be a compact subset of $\mathbb{R}^n$. Then the convex hull $K$ of $E$ is also compact.

Proof: Let

$$S = \left\{ (\alpha_0, \ldots, \alpha_n) : \alpha_j \geq 0, \sum_{j=0}^n \alpha_j = 1 \right\} \subseteq \mathbb{R}^{n+1}$$

and consider the map $f : S \times \mathbb{R}^{n+1} \to K$ defined by

$$f(\alpha_0, \ldots, \alpha_n, x_0, \ldots, x_n) = \sum_{j=0}^n \alpha_j x_j.$$ 

Then $f$ is continuous. Each extreme point of $K$ is in $E$ by XV.7.1.13., and every element of $K$ is a convex combination of at most $n + 1$ extreme points by XV.7.2.7., so $f$ is surjective. $S \times \mathbb{R}^{n+1}$ is compact, so $K$ is compact by (1).

XV.7.3. The Krein-Milman Theorem

XV.7.3.1. Theorem. [Krein-Milman Theorem] Let $K$ be a nonempty compact convex set in a locally convex (Hausdorff) topological vector space $X$. Then $K$ is the closed convex hull of its extreme points (and in particular has extreme points).

Proof: Recall that we regard $X$ as a real topological vector space by ignoring the complex structure if necessary, since all relevant notions (convexity, compactness, etc.) depend only on the real topological vector space structure. Let $F$ be the collection of all closed faces of $K$. Partially order $F$ by reverse inclusion: if $E, F \in F$, then $E \leq F$ if $F \subseteq E$. If $C = \{ F_i : i \in I \}$ is a chain in $F$, then $C$ has an upper bound, namely $F = \bigcap_i F_i$, which is a closed face (XV.7.1.10.) since it is nonempty because $C$ is a nested family of compact subsets of $K$ (this is the first of several places compactness of $K$ is used).

Thus, by Zorn’s Lemma, $F$ has a maximal element, i.e. $K$ has a minimal closed face $E$. We must have that $E$ is a singleton: if $x, y \in E$, $x \neq y$, there is a continuous linear functional $\phi$ on $X$ with $\phi(x) \neq \phi(y)$ (this is the first place where we need that $X$ is locally convex and Hausdorff). Then $\{ \phi(z) : z \in E \}$ is a compact subset of $\mathbb{R}$ (actually a closed bounded interval) since $E$ is compact. If $\lambda$ is the largest number in this set, then $F = \{ z \in E : \phi(z) = \lambda \}$ is a closed face of $E$ and hence of $K$, and is a proper subset of $E$ since $x$ and $y$ cannot both be in $F$, contradicting that $E$ is a minimal closed face of $K$. Thus $E = \{ p \}$ for an extreme point $p$ of $K$, and $K$ has an extreme point. A stronger conclusion also holds: if $F$ is any closed face of $K$, then $\tilde{F}$ is itself a nonempty compact convex set in $X$, and so has a point which is an extreme point of $F$ and thus an extreme point of $K$. Therefore every closed face of $K$ contains an extreme point of $K$.

Now let $L$ be the closed convex hull of the extreme points of $K$. If $L \neq K$, let $x \in K \setminus L$. Then, since $L$ is compact and convex, there is a continuous linear functional $\psi$ on $X$ with

$$\psi(x) > \sup \{ \psi(y) : y \in L \}$$

by the geometrical Hahn-Banach Theorem (1) (we again use that $X$ is locally convex and Hausdorff). The set $\{ \psi(z) : z \in K \}$ is a compact subset of $\mathbb{R}$ (actually a closed bounded interval), so has a largest number $\mu$. Then $F = \{ z \in K : \psi(z) = \mu \}$ is a closed face in $K$ disjoint from $L$. But $F$ contains an extreme point of $K$ by the previous proof, contradicting that all extreme points of $K$ are in $L$. Thus $L = K$. \( \diamond \)
XV.7.3.2. The local convexity hypothesis in the Krein-Milman Theorem is to insure that there are sufficiently many continuous linear functionals, or equivalently enough closed supporting hyperplanes. There exists a compact convex set in a topological vector space which is complete and metrizable, but not locally convex, which has no extreme points [Rob77]. (In a non-Hausdorff topological vector space, there is a trivial counterexample: the closure of \( \{0\} \).)

XV.7.3.3. The hypotheses of the Krein-Milman Theorem are not necessary. If \( A \) is a unital C*-algebra (), then every element of the open unit ball of \( A \) is a convex combination of unitaries, which are extreme points of the closed unit ball [Bla06]. Thus the closed unit ball of a unital C*-algebra is the closed convex hull of its extreme points, even though there is no (Hausdorff) topological vector space topology on \( A \) in general under which the closed unit ball is compact. (The closed unit ball of a nonunital C*-algebra has no extreme points.) A closed bounded set in a Banach space with the Radon-Nikodym Property () is the closed convex hull of its extreme points ([Phe74], [Edg75]).

XV.7.4. Choquet’s Theorem

There is an important sharpening of the Krein-Milman Theorem for metrizable compact convex sets, due to G. Choquet [Cho56]. Our exposition is adapted from [Phe66].

Representing Measures

XV.7.4.1. Definition. Let \( X \) be a topological vector space, and \( E \) a compact subset of \( X \). Let \( \mu \) be a Borel probability measure on \( E \), and \( x \in X \). Then \( \mu \) represents \( x \), or \( x \) is the barycenter, resultant, or moment of \( \mu \), if for every continuous linear functional \( \phi \) on \( X \) we have

\[
\phi(x) = \mu(\phi) := \int_E \phi \, d\mu
\]

(note that \( \phi \) is Borel measurable and bounded on \( E \), so the integral is defined; the notation \( \mu(\phi) \) comes from regarding \( \mu \) as a bounded linear functional on \( C(E) \) by the Riesz Representation Theorem ()).

The point \( x \), if it exists, is the weak integral \( \int_E \iota \, d\mu \), where \( \iota \) is the identity map on \( E \), regarded as a continuous function from \( E \) to \( X \).

XV.7.4.2. Examples. (i) If \( x \in X \) and \( E \) is any compact subset of \( X \) containing \( x \), then the point mass \( \delta_x \) (regarded as a Borel probability measure on \( E \)) represents \( x \). The measure \( \delta_x \) is called \( \epsilon_x \) in some references.

(ii) Let \( K \) be the convex hull of \( n \) extreme points \( \{p_1, \ldots, p_n\} \) in a Hausdorff topological vector space \( X \). Then \( K \) is compact. If \( x \in K \), then \( x = \sum_{j=1}^{n} \alpha_j p_j \) for some \( \alpha_1, \ldots, \alpha_n \geq 0 \) with \( \sum_{j=1}^{n} \alpha_j = 1 \). If \( \mu \) is the purely atomic Borel probability measure on \( K \) with \( \mu(p_j) = \alpha_j \) (i.e. \( \mu = \sum_{j=1}^{n} \alpha_j \delta_{p_j} \)), then \( \mu \) represents \( x \) (\( x \) is the barycenter of \( \mu \)). Thus every point of \( K \) is the barycenter of a Borel probability measure on \( K \) supported on the extreme points of \( K \). (Since the convex combination giving \( x \) is not necessarily unique, the representing measure for \( x \) is not unique in general.)

Choquet’s Theorem extends (ii) to an arbitrary metrizable compact convex set \( K \), giving a representing measure (not unique in general) for any point of \( K \) which is supported on the extreme points of \( K \).
XV.7.4.3. **Proposition.** If $X$ is locally convex (Hausdorff), and $\mu$ is a Borel probability measure on a compact subset $E$ of $X$, then $\mu$ represents at most one point of $X$.

**Proof:** If $\mu$ represents $x$ and $y$, then $\phi(x) = \mu(\phi) = \phi(y)$ for all $\phi \in X^*$. Thus $x = y$ since $X^*$ separates the points of $X$.

XV.7.4.4. **Proposition.** If $K$ is a compact convex set in a topological vector space $X$, and $\mu$ is a Borel probability measure on $K$, of $X$, then $\mu$ represents at least one point of $X$ (exactly one point $x_\mu$ if $X$ is locally convex), and the barycenter is in $K$. If $X$ is locally convex, the function $\mu \mapsto x_\mu$ is weak-* continuous, i.e. if $\mu_i \to \mu$ weak-* in $C(K)^*$, then $x_{\mu_i} \to x_\mu$ in $K$.

**Proof:** If $x \in X$ is represented by $\mu$, then for any $\phi \in X^*$,

$$\phi(x) = \mu(\phi) \leq \sup \{ \phi(y) : y \in K \}$$

and thus by the geometrical Hahn-Banach Theorem (X.7.4.3) $x \in K$ since it is closed and convex.

For existence, for each $\phi \in X^*$ let $H_\phi = \{ x \in K : \phi(x) = \mu(\phi) \}$. Then $H_\phi$ is a closed hyperplane in $X$, and we wish to show that

$$\bigcap_{\phi} H_\phi \cap K \neq \emptyset.$$ 

Since these sets are compact, it suffices to show that for any $\phi_1, \ldots, \phi_n \in X^*$,

$$(\bigcap_{\phi_i} H_{\phi_i}) \cap K \neq \emptyset.$$ 

Define a bounded linear map $T : X \to \mathbb{R}^n$ by

$$T(x) = (\phi_1(x), \ldots, \phi_n(x)).$$

Then $T(K)$ is a compact convex subset of $\mathbb{R}^n$. Set $y = (\mu(\phi_1), \ldots, \mu(\phi_n)) \in \mathbb{R}^n$. It suffices to show $y \in T(K)$.

If $y \notin T(K)$, there is a linear functional $\psi$ on $\mathbb{R}^n$ with $\psi(y) > \sup \{ \psi(z) : z \in T(K) \}$. There is a vector $a \in \mathbb{R}^n$ such that $\psi(z) = (a, z)$ for all $z \in \mathbb{R}^n$. Set $a = (a_1, \ldots, a_n)$. We have

$$\psi(y) = \langle a, y \rangle > \sup \{ \langle a, T(x) \rangle : x \in K \}.$$ 

If $\theta \in X^*$ is defined by $\theta = \sum_{j=1}^n a_j \phi_j$, we have

$$\int_K \theta d\mu = \psi(y) > \sup \{ \theta(x) : x \in K \}.$$ 

This is impossible since $\mu$ is a probability measure. Thus $y \in T(K)$.

Now suppose $X$ is locally convex (Hausdorff), and $(\mu_i)$ is a net of Borel probability measures on $K$ with $\mu_i \to \mu$ weak-*. Set $x_i = x_{\mu_i}$ and $x = x_\mu$. To show $x_i \to x$, we use the hard-to-state criterion XI.3.2.13. Fix a subnet $(x_j)$ of $(x_i)$. Then $(x_j)$ has a convergent subnet by compactness of $K$. Fix a convergent subnet $(x_r)$, say $x_r \to y \in K$. Then, for any $\phi \in X^*$, $\phi(x_r) \to \phi(y)$ by continuity of $\phi$. On the other hand,

$$\phi(x_i) = \mu_i(\phi) \to \mu(\phi) = \phi(x)$$

and so the subnet $(\phi(x_i))$ also converges to $\phi(x)$. Thus $\phi(y) = \phi(x)$. This is true for all $\phi \in X^*$, so $y = x$ since $X^*$ separates the points of $X$. So $(x_j)$ has a subnet converging to $x$, and so $x_i \to x$ by XI.3.2.13.
XV.7.4.5. Corollary. If $E$ is a compact set in a locally convex topological vector space $X$, and the closed convex hull $K$ of $E$ is also compact (this is automatic if $X$ is topologically complete, e.g. if $X$ is a Banach space in either the norm or weak topology, or the dual of a Banach space in the weak-* topology, cf. ()), then every Borel probability measure $\mu$ on $E$ has a unique barycenter $x_\mu$ in $X$, and $x_\mu \in K$. The map $\mu \mapsto x_\mu$ is weak-* continuous.

XV.7.4.6. Proposition. If $E$ is a compact set in a locally convex topological vector space $X$, and $x \in X$, then $x$ is in the closed convex hull $\overline{co}(E)$ if and only if $x$ is represented by a Borel probability measure on $E$.

Proof: One direction generalizes part of XV.7.4.5. (since it is not assumed that $\overline{co}(E)$ is compact), and has the same proof: if $x$ is represented by $\mu$, then for any $\phi \in X^*$,

$$\phi(x) = \mu(\phi) \leq \sup\{\phi(y) : y \in E\} = \sup\{\phi(y) : y \in \overline{co}(E)\}$$

and thus $x \in \overline{co}(E)$ since it is closed and convex.

For the converse, suppose $x \in \overline{co}(E)$. Then there is a net $(x_i)$ in $co(E)$ with $x_i \to x$. Each $x_i$ can be represented by a Borel probability measure $\mu_i$ on $E$ which is a finite convex combination of point masses as in XV.7.4.2.(ii). Since the set of Borel probability measures on $E$ is a weak-* compact subset of $C(E)^*$, there is a subnet $(\mu_j)$ of the net $(\mu_i)$ which converges weak-* to a Borel probability measure $\mu$. Then $x_j \to x$, and for any $\phi \in X^*$ we have

$$\phi(x_j) \to \phi(x)$$

and thus $\phi(x) = \mu(\phi)$. Since this holds for all $\phi \in X^*$, $\mu$ represents $x$.

The Krein-Milman Theorem Revisited

The Krein-Milman Theorem can thus be rephrased using XV.7.4.6.:

XV.7.4.7. Theorem. [Krein-Milman Theorem, Representing Measure Form] Let $K$ be a compact convex set in a locally compact (Hausdorff) topological vector space $X$, and $E$ the closure of the set $\partial_e(K)$ of extreme points of $K$. Then every point of $K$ is the barycenter of a Borel probability measure on $E$.

For the proof, just note that $\overline{co}(E) = \overline{co}(\partial_e(K))$, and apply XV.7.3.1. and XV.7.4.6.. Conversely, XV.7.3.1. follows from XV.7.4.7. and XV.7.4.6., so XV.7.3.1. and XV.7.4.7. are equivalent via XV.7.4.6..

XV.7.4.8. This form of the Krein-Milman Theorem is not optimally useful, since $E$ can be a rather large subset of $K$; in fact, $E$ is “generically” the whole set $K$ (i.e. the set of extreme points of $K$ is generically dense in $K$)! [Kle59]. It would be far better to have representing measures supported on the set $\partial_e(K)$ of extreme points of $K$ itself, not just on the closure.
**XV.7.4.9.** In order for the desired statement to even make sense, we need the set \( \partial_e(K) \) to be a Borel set. Although it is not closed in general, even if \( K \) is a metrizable Choquet simplex \( (\cdot) \), in the metrizable case we do have:

**XV.7.4.10.** **Proposition.** Let \( K \) be a metrizable compact convex set in a topological vector space \( X \). Then the set \( \partial_e(K) \) of extreme points of \( K \) is a \( G_\delta \), and in particular a Borel set.

**Proof:** Fix a metric \( \rho \) giving the topology of \( K \). For \( n \in \mathbb{N} \), let \( A_n \) be the set of all \( x \in K \) such that there exist \( y, z \in K \) with \( x = \frac{1}{2}(y + z) \) and \( \rho(y, z) \geq \frac{1}{n} \). Then \( \bigcup_{n=1}^\infty A_n \) is the complement of \( \partial_e(K) \), so it suffices to show that each \( A_n \) is a closed set.

Fix \( n \in \mathbb{N} \), and let \( (x_j) \) be a sequence in \( A_n \) with \( x_j \to x \in K \). For each \( j \), there are \( y_j, z_j \in K \) with \( x_j = \frac{1}{2}(y_j + z_j) \) and \( \rho(y_j, z_j) \geq \frac{1}{n} \). By compactness of \( K \), there are convergent subsequences \( (y_{r_j}) \) and \( (z_{r_j}) \) of \( (y_j) \) and \( (z_j) \), which can be taken to have the same indices \( (\cdot) \), say \( y_{r_j} \to y \), \( z_{r_j} \to z \). Then

\[
\frac{1}{2}(y_{r_j} + z_{r_j}) \to \frac{1}{2}(y + z).
\]

But

\[
\frac{1}{2}(y_{r_j} + z_{r_j}) = x_{r_j} \to x
\]

and thus \( x = \frac{1}{2}(y + z) \) since \( K \) is Hausdorff. Since \( \rho(y, z) \geq \frac{1}{n} \) by continuity of \( \rho \), we have \( x \in A_n \) and \( A_n \) is closed.

We then have the following strengthening of **XV.7.4.7.** in the metrizable case:

**XV.7.4.11.** **Theorem.** [Choquet’s Theorem] Let \( K \) be a metrizable compact convex set in a locally convex (Hausdorff) topological vector space \( X \). Then every \( x \in K \) is represented by a measure which is supported on the set \( \partial_e(K) \) of extreme points of \( K \).

**XV.7.4.12.** We give a nice proof due to BONSALL [Bon63] (cf. [Phe66]) based on the Hahn-Banach Theorem. The argument will be a series of lemmas. To begin, fix a point \( x_0 \in K \); we will show that \( x_0 \) is represented by a measure supported on \( \partial_e(K) \). We will work with the subspace \( \text{Aff}(K) \) of \( C(K) \) of real-valued continuous affine functions on \( K \). Since \( K \) is metrizable, \( \text{Aff}(K) \) is separable; fix a sequence \( (h_n) \) which is dense in the unit sphere of \( \text{Aff}(K) \), and set

\[
f = \sum_{n=1}^\infty 2^{-n}h_n^2 \in C(K).
\]

**XV.7.4.13.** **Lemma.** The function \( f \) is strictly convex on \( K \), i.e. \( f(tx + (1-t)y) < tf(x) + (1-t)f(y) \) for \( x, y \in K \), \( x \neq y \), \( 0 < t < 1 \).

**Proof:** Each \( h_n^2 \) is clearly convex, so \( f \) is convex. If \( x \neq y \), there is an \( h_n \) with \( h_n(x) \neq h_n(y) \), so \( h_n \) is affine and not constant on the segment between \( x \) and \( y \), and therefore \( h_n^2 \) is strictly convex on this segment. Hence \( f \) is too. ☐

1760
XV.7.4.14. Definition. If \( g \in C(K) \), write
\[
\tilde{g} = \inf \{ h : h \in \text{Aff}(K), g \leq h \}.
\]
The function \( \tilde{g} \) is called the upper envelope of \( g \).

XV.7.4.15. Lemma. Let \( g, g_1, g_2 \in C(K) \). Then

(i) \( \tilde{g} \) is concave, i.e. \( \tilde{g}(tx + (1-t)y) \geq t\tilde{g}(x) + (1-t)\tilde{g}(y) \) for \( x, y \in K, 0 < t < 1 \).

(ii) \( \tilde{g} \) is bounded and upper semicontinuous, hence Borel measurable.

(iii) \( g \leq \tilde{g} \), and \( g = \tilde{g} \) if and only if \( g \) is concave.

(iv) \( \frac{1}{g_1 + g_2} \leq \frac{1}{g_1 + g_2}, \) and \( \frac{1}{g_1 + g_2} = \frac{1}{g_1 + g_2} = \frac{1}{g_1 + g_2} \) if \( g_2 \in \text{Aff}(K) \).

(v) \( |\tilde{g}_1 - \tilde{g}_2| \leq ||g_1 - g_2|| \).

(vi) If \( \alpha \geq 0 \), then \( \alpha \tilde{g} = \alpha \tilde{g} \).

The simple proof is straightforward and left to the reader.

XV.7.4.16. Let \( Y \) be the subspace of \( C(K) \) spanned by \( \text{Aff}(K) \) and \( f \) (defined above). Define a linear functional \( \psi \) on \( Y \) by \( \psi(h + \alpha f) = h(x_0) + \alpha f(x_0) \).

XV.7.4.17. Lemma. For \( g \in C(K) \), define \( p(g) = \tilde{g}(x_0) \). Then

(i) \( p \) is subadditive and positive-homogeneous.

(ii) If \( h + \alpha f \in Y \), then \( \psi(h + \alpha f) \leq p(h + \alpha f) \).

Proof: (i) follows immediately from XV.7.4.15.(iv) and (vi).

(ii): If \( \alpha \geq 0 \), then \( \frac{1}{h + \alpha f} = h + \alpha f \) by XV.7.4.15.(iv) and (vi), and in particular
\[
\psi(h + \alpha f) = p(h + \alpha f).
\]

If \( \alpha < 0 \), then \( h + \alpha f \) is concave, so \( \frac{1}{h + \alpha f} = h + \alpha f \geq h + \alpha f \) and in particular
\[
\psi(h + \alpha f) = p(h + \alpha f).
\]

XV.7.4.18. Thus by the Hahn-Banach Theorem there is a linear functional \( \phi \) on \( C(K) \) with \( \phi(g) \leq p(g) = \tilde{g}(x_0) \)

for all \( g \in C(K) \), which extends \( \psi \).
**XV.7.4.19.** Lemma. The linear functional $\phi$ is a state (positive linear functional of norm 1) on $C(K)$.

**Proof:** If $g \in C(K)$, $g \leq 0$, then since the constant function 0 is in $\text{Aff}(K)$, $\bar{g} \leq 0$, and in particular $\phi(g) \leq \mu(g) = \bar{g}(x_0) \leq 0$. Taking negatives, $\phi$ is a positive linear functional. Since the constant function 1 is in $\text{Aff}(K)$, we have $\phi(1) = \psi(1) = 1(x_0) = 1$.

**XV.7.4.20.** Thus, by the Riesz Representation Theorem, there is a Borel probability measure $\mu$ on $K$ with

$$\phi(g) = \mu(g) := \int_K g \, d\mu$$

for all $g \in C(K)$. We have $\mu(f) = \bar{f}(x_0)$. Since $f \leq \bar{f}$, $\mu(f) \leq \mu(\bar{f})$ (note that $\mu(\bar{f}) = \int_K \bar{f} \, d\mu$ is defined even though $\bar{f}$ may not be continuous, since it is bounded and Borel measurable); but if $h \in \text{Aff}(K)$, $h \geq f$, we have $h \geq \bar{f}$, so

$$h(x_0) = \psi(h) = \phi(h) = \mu(h) \geq \mu(\bar{f}) .$$

Taking the infimum over all such $h$, we get

$$\bar{f}(x_0) \geq \mu(\bar{f}) \geq \mu(f) = \bar{f}(x_0)$$

and thus equality holds, i.e.

$$\int_K \bar{f} \, d\mu = \mu(\bar{f}) = \mu(f) = \int_K f \, d\mu$$

and so

$$\int_K (\bar{f} - f) \, d\mu = 0 .$$

Thus $\mu$ vanishes on the complement of the set

$$S = \{ x \in K : f(x) = \bar{f}(x) \} .$$

**XV.7.4.21.** Lemma. The set $S$ is contained in $\partial_e(K)$.

**Proof:** If $x, y, z$ are distinct points of $K$ with $x = \frac{1}{2}(y + z)$, then by the strict convexity of $f$ we have

$$f(x) < \frac{1}{2} f(y) + \frac{1}{2} f(z) \leq \frac{1}{2} \bar{f}(y) + \frac{1}{2} \bar{f}(z) \leq \bar{f}(x)$$

since $\bar{f}$ is concave.

This completes the proof of **XV.7.4.11**.

**XV.7.4.22.** In fact, the set $S$ is exactly $\partial_e(K)$; cf. [Phe66, 3.1].

1762
The Nonmetrizable Case

XV.7.4.23. There is a generalization of Choquet’s Theorem to nonmetrizable compact convex sets, the Bishop-de Leeuw Theorem \[BdL59\]. This theorem is considerably more difficult to prove, and is even somewhat tricky to state, because (among other things) the set of extreme points in a nonmetrizable compact convex set, even a nonmetrizable Choquet simplex, need not be a Borel set. See e.g. \[Phe66\] or \[Alf71\] for details.

XV.7.4.24. The Bishop of Bishop-de Leeuw is the Errett Bishop of constructivism fame (I.6.1.20.). Karel de Leeuw is known (besides for his work) for a darker reason: he was murdered in his office by a graduate student.

XV.7.5. Regular Embeddings, Cones, and Affine Functions

XV.7.6. Choquet Simplexes

XV.7.6.1. A compact convex set can be thought of as being defined by “generators” and “relations.” The “generators” are the set of extreme points, and the “relations” are pairs of measures which represent the same point. For example, a square (or parallelogram, cf. XV.7.7.1. (b)) has four generators \(v_1, v_2, v_3, v_4\) and one relation

\[
\frac{1}{2}(v_1 + v_3) = \frac{1}{2}(v_2 + v_4).
\]

From this (somewhat imprecise) point of view, a Choquet simplex is a “free” compact convex set, i.e. one with no relations.

Bauer Simplexes

XV.7.6.2. Definition. A Choquet simplex \(\Delta\) in which the set of extreme points is closed in \(\Delta\) is a Bauer simplex.

XV.7.6.3. Since a Bauer simplex is compact, its set of extreme points is a compact Hausdorff space. Every compact Hausdorff space \(X\) is the set of extreme points of a Bauer simplex \(\Delta_X\), which is unique up to isomorphism. \(\Delta_X\) can be concretely described as the state space of \(C(X)\) (\(\cdot\)), which by the Riesz\(\) Representation Theorem can be identified with the set of Radon probability measures on \(X\). The extreme points are just the point masses, which can be identified with \(X\) (\(\cdot\)). The representing measure of a point is just the point itself; thus uniqueness is essentially tautological, so \(\Delta_X\) is a Choquet simplex. \(\Delta_X\) is metrizable if and only if \(X\) is metrizable, i.e. \(C(X)\) is separable (\(\cdot\)). All finite-dimensional simplexes are, of course, Bauer simplexes.

Other Simplexes

XV.7.6.4. “Most” Choquet simplexes are not Bauer simplexes. The simplest example of a Choquet simplex which is not a Bauer simplex is given in Exercise XV.7.7.2.. The construction can be modified to give far more complicated examples.

In fact, if \(X\) is any Polish space, there is a metrizable Choquet simplex \(\Delta\) such that \(\partial_e(\Delta)\) is homeomorphic to \(X\) (\(\cdot\); the set of extreme points of a metrizable Choquet simplex is always a \(G_\delta\), hence a Polish space in the relative topology). The simplex \(\Delta\) is highly nonunique in general; for example, the set of extreme

1763
points for each of the $\Delta_c$ ($0 < c \leq 1$) of XV.7.7.2. (f) is a countable discrete space (and there are much more complicated simplexes with the same space of extreme points).

**XV.7.6.5.** The most extreme case is the Poulsen simplex. There is a metrizable Choquet simplex $\Delta$, unique up to isomorphism, called the *Poulsen simplex*, for which the set of extreme points of $\Delta$ is dense in $\Delta$ [?]. Every metrizable Choquet simplex is isomorphic to a face of the Poulsen simplex.

The set of extreme points of the Poulsen simplex is thus a Polish space in which every Polish space can be embedded as a closed subset. It can be shown that $\partial_c(\Delta)$ is homeomorphic to $\mathbb{R}^\infty ()$, the underlying topological space of every separable infinite-dimensional Banach space ().

**XV.7.7. Exercises**

**XV.7.7.1.** (a) Show that all (solid) triangles are isomorphic as compact convex sets.

(b) Show that all parallelograms are isomorphic as convex sets. Show that a convex quadrilateral which is not a parallelogram is not isomorphic to a parallelogram. Describe the isomorphism classes of convex quadrilaterals. [Consider how the diagonals intersect.]

**XV.7.7.2.** Give $\ell^1$ the weak-* topology as $c_0^\infty ()$. Then the closed unit ball is compact (), metrizable (), and convex. Let $(e_n)$ be the standard Schauder basis. Let $\Delta$ be the set of sequences $x = (\lambda_1, \lambda_2, \ldots)$ such that $\|x\| \leq 1$, $-1 \leq \lambda_1 \leq 1$, and $\lambda_n \geq 0$ for all $n > 1$.

(a) Show that $\Delta$ is the closed convex hull of $E = \{e_1, -e_1, e_2, e_3, \ldots\}$, and that $E$ is the set of extreme points of $\Delta$. Thus $\Delta$ is compact (and metrizable).

(b) Show that $\Delta$ is a Choquet simplex. [Use ().]

(c) Show that $E$ is not closed in $\Delta$: $\bar{E} = E \cup \{0\}$. Note that $0$ is the midpoint of the segment between $e_1$ and $-e_1$.

(d) Let $F$ be the set of sequences in $\Delta$ for which $\lambda_1 = 0$ and $\sum_{k=2}^\infty \lambda_k = 1$. Show that $F$ is a face of $\Delta$.

(e) Show that $\bar{F}$ is the set of sequences in $\Delta$ with $\lambda_1 = 0$, which is not a face of $\Delta$.

(f) More generally, fix $c$, $0 \leq c \leq 1$, and let $\Delta_c$ be the closed convex hull of $\{e_1, -ce_1, e_2, e_3, \ldots\}$. Then each $\Delta_c$ is a metrizable Choquet simplex with a sequence of extreme points converging to $0$. No two of the $\Delta_c$ are isomorphic. $\Delta_0$ is a Bauer simplex ($0$ is an extreme point).

(g) In all these examples except $\Delta_0$, the relative topology on $\partial_c(\Delta_c)$ is the discrete topology. The facial topology on $\partial_c(\Delta_c)$ is the $T_1$ topology consisting of a sequence of isolated points $(e_n$ for $n \geq 2$) converging simultaneously to $e_1$ and $-ce_1$. The relative and facial topologies on $\partial_c(\Delta_0)$ coincide and consist of a sequence of isolated points (the $e_n$) converging to $0$.

**XV.7.7.3.** Let $\Delta_0$ be the simplex of XV.7.7.2. (f), and let $\pi$ be the quotient map from $\ell^1$ onto $\ell^1 / X$, where $X$ is the one-dimensional subspace of $\ell^1$ spanned by $x = (2^{-1}, 2^{-2}, 2^{-3}, \ldots)$. Set $\Delta = \pi(\Delta_0)$.

(a) Show that $\Delta$ is a Choquet simplex, and that $\partial_c(\Delta)$ consists of $\{f_n : n \in \mathbb{N}\}$, where $f_n = \pi(e_n)$.

(b) The relative topology on $\partial_c(\Delta)$ is the discrete topology, while the facial topology is the finite complement topology.

This $\Delta$ is a *prime simplex* (cf. ()).
XV.7.7.4. ([Bau61]; cf. [Phe66, 1.4], [Alf71, 1.2.4]) Let $K$ be a nonempty compact convex set in a locally convex topological vector space.

(a) If $x \in K$, show that $x$ is an extreme point of $K$ if and only if $\delta_x$ is the only Borel probability measure on $K$ representing $x$.

(b) Use (a) and XV.7.4.6. to give an alternate proof of Milman’s partial converse XV.7.1.13.

XV.7.7.5. Let $K$ be a convex set in a topological vector space $X$. A point $p$ of $K$ is an exposed point of $K$ if there is a closed hyperplane $H$ such that $H \cap K = \{p\}$, i.e. if there is a continuous linear functional $\phi$ on $X$ with $\phi(x) < \phi(p)$ for all $x \in K \setminus \{p\}$.

(a) Show that every exposed point is an extreme point.

(b) Let $K$ be the union in $\mathbb{R}^2$ of the closed unit disk and the square

$$\{(x, y) : -1 \leq x \leq 1, -2 \leq y \leq 0\}.$$ 

Then $K$ is compact and convex. Show that $(1, 0)$ and $(-1, 0)$ are extreme points of $K$ but not exposed points. Find all extreme points and exposed points of $K$.

There are other related notions: strongly exposed points, dentable points, etc. These notions are tied in with the structure theory of Banach spaces.
XV.8. The Stone-Weierstrass Theorem

The Stone-Weierstrass Theorem is one of the important theorems of functional analysis, giving a useful criterion for uniformly approximating general continuous functions by “nice” functions.

XV.8.1. The Weierstrass Approximation Theorem

The original version proved by Weierstrass in 1885 was:

**Theorem.** Let $I$ be a closed bounded interval in $\mathbb{R}$. Then every continuous function from $I$ to $\mathbb{R}$ can be uniformly approximated arbitrarily closely on $I$ by a polynomial.

**Bernstein’s Proof**

In 1912 Sergei Bernstein gave a better proof of the Weierstrass Approximation Theorem in which he gave an explicit sequence of polynomials converging uniformly to a given continuous function on an interval:

**Theorem.** If $f$ is a continuous real-valued function on $[0,1]$, define, for $0 \leq x \leq 1$,

$$B_n(f)(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}$$

(This is the $n$'th Bernstein polynomial of $f$.) Then $(B_n(f))$ converges uniformly to $f$ on $[0,1]$.

Other intervals can be handled via an affine transformation.

**Bernstein’s probabilistic argument** (cf. [?]), which suggests that $B_n(f)$ should at least converge to $f$ pointwise.

Fix $f$ and $x \in [0,1]$. We repeatedly toss a coin with probability $x$ of heads. Suppose we toss the coin a large number $n$ of times, and receive a payoff of $f(r)$ dollars, where $r$ is the fraction of tosses which are heads. In the long run (as $n \to \infty$), $r$ will approach $x$ with probability 1, so the payoff will approach $f(x)$ since $f$ is continuous.

If we toss the coin $n$ times, we will get exactly $k$ heads with probability

$$\binom{n}{k} x^k (1-x)^{n-k}$$

so the expected payoff is

$$\sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} = B_n(f)(x).$$

A direct proof of Bernstein’s result can be found in [?] or almost any book on numerical analysis or approximation theory; a probability version of the proof can be found in [?]. In fact, books on numerical analysis prove the following sharpened form:
**Theorem.** Let $f$ be a bounded real-valued function on $[0, 1]$, and $B_n(f)$ the $n$’th Bernstein polynomial of $f$ (XV.8.1.2). Then

$$|f(x) - B_n(f)(x)| \leq \frac{3}{2} \omega\left(f, \frac{1}{\sqrt{n}}\right)$$

for all $x \in [0, 1]$, where $\omega(f, \cdot)$ is the modulus of continuity ($\cdot$) of $f$ on $[0, 1]$.

Since $\lim_{\delta \to 0} \omega(f, \delta) = 0$ if $f$ is continuous, XV.8.1.2. is a corollary of XV.8.1.4.

**Korovkin’s Theorem**

We give a proof of XV.8.1.2., and hence XV.8.1.1., based on a result of independent interest due to KOROVKIN [Kor53]:

**Theorem.** Let $I = [a, b]$ be a closed bounded interval, and $(L_n)$ a sequence of positive linear operators from $C(I)$ to $C(I)$ (i.e. if $f \geq 0$ then $L_n f \geq 0$ for all $n$). Let $p_0(x) = 1$, $p_1(x) = x$, and $p_2(x) = x^2$. If $L_n p_k \to p_k$ uniformly for $k = 0, 1, 2$, then $L_n f \to f$ uniformly for all $f \in C(I)$.

**Proof:** (cf. [Bri98]) For each $t \in I$, set $g_t(x) = (x - t)^2 = t^2 p_0(x) - 2t p_1(x) + p_2(x)$. Then

$$[L_n g_t](x) = t^2 [L_n p_0](x) - 2t [L_n p_1](x) + [L_n p_2](x)$$

for all $n, t, x$. Since $g_t \geq 0$, we have

$$0 \leq [L_n g_t](t) = t^2 ([L_n p_0](t) - 1) + 2t ([L_n p_1](t) - t) + ([L_n p_2](t) - t^2)$$

$$\leq t^2 [L_n p_0 - p_0] + 2t [L_n p_1 - p_1] + [L_n p_2 - p_2]$$

and thus, since $t^2$ and $|2t|$ are bounded on $I$, for any $\eta > 0$ there is an $N$ such that $[L_n g_t] < \eta$ for all $n \geq N$ and all $t \in I$.

Now fix $f \in C(I)$ and $\epsilon > 0$. Since $f$ is uniformly continuous on $I$, there is a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $x, y \in I$ and $|x - y| < \delta$. Fix $t \in I$. If $x \in I$, $|x - t| \geq \delta$, we have

$$|f(t) - f(x)| \leq 2 \|f\| \leq 2 \|f\| \frac{(x - t)^2}{\delta^2} = \frac{2}{\delta^2} \|f\| g_t(x)$$

so for any $x \in I$ we have

$$|f(t) - f(x)| \leq \frac{2}{\delta^2} \|f\| g_t(x) + \epsilon.$$

Thus

$$-\epsilon p_0 - \frac{2}{\delta^2} \|f\| g_t \leq f(t)p_0 - f \leq \epsilon p_0 + \frac{2}{\delta^2} \|f\| g_t$$

and hence

$$-\epsilon L_n p_0 - \frac{2}{\delta^2} \|f\| L_n g_t \leq f(t) L_n p_0 - L_n f \leq \epsilon L_n p_0 + \frac{2}{\delta^2} \|f\| [L_n g_t]$$

$$|f(t) [L_n p_0](t) - [L_n f](t)| \leq \epsilon \|L_n p_0\| + \frac{2}{\delta^2} \|f\| [L_n g_t](t).$$

1767
Thus
\[ |f(t) - [L_n f](t)| \leq |f(t) - f(t)[L_n p_0](t)| + |f(t)[L_n p_0](t) - [L_n f](t)| \]
\[ \leq |f(t)|1 - |L_n p_0|(t)| + \epsilon |L_n p_0| + \frac{2}{\delta^2} ||f||[L_n g_k](t) \]
\[ \leq ||f|||p_0 - L_n p_0| + \epsilon (||p_0|| + ||p_0 - L_n p_0||) + \frac{2\eta}{\delta^2} ||f||L_n g_k(t) . \]
so if \( \eta > 0 \) and \( n \) is large enough,
\[ |f(t) - [L_n f](t)| \leq ||f|||p_0 - L_n p_0| + \epsilon (||p_0|| + ||p_0 - L_n p_0||) + \frac{2\eta}{\delta^2} ||f|| . \]
The right side is independent of \( t \) and can be made less than \( 2\epsilon \) by choosing \( \eta \) small enough and then \( n \) large enough. Since \( \epsilon > 0 \) is arbitrary, we have \( L_n f \to f \) uniformly on \( I \).

\[ \text{(XV.8.1.6.)} \]

To prove XV.8.1.2., and hence XV.8.1.1., it thus suffices to show that each \( B_n \) is a positive operator (obvious) and that \( B_n p_k \to p_k \) uniformly for \( k = 0, 1, 2 \), which is a simple calculation.

**XV.8.2. The Stone-Weierstrass Theorem**

In 1937 Marshall Stone formulated and proved a general version of Weierstrass’s Theorem, which is now known as the Stone-Weierstrass Theorem. There are several commonly stated versions of the theorem, all of which are simple consequences of the following general version. Stone also proved a lattice version (XV.8.2.14.).

**XV.8.2.1. Theorem.** [Stone-Weierstrass] Let \( X \) be a compact Hausdorff space, and \( C(X) \) the algebra of real-valued continuous functions on \( X \). Let \( A \) be a subalgebra of \( C(X) \) which is uniformly closed () and separates points (i.e. if \( x, y \in X \) with \( x \neq y \) there is an \( f \in A \) with \( f(x) \neq f(y) \)). Then either \( A = C(X) \) or there is a \( p \in X \) such that \( A \) consists of all \( f \in C(X) \) with \( f(p) = 0 \).

Here is the most commonly stated version of the Stone-Weierstrass Theorem (to prove it, consider the uniform closure of \( A \), which is also an algebra):

**XV.8.2.2. Corollary.** Let \( X \) be a compact Hausdorff space, and \( A \) a subalgebra of \( C(X) \) which separates points and contains the constant functions. Then \( A \) is dense in \( C(X) \) for the uniform norm.

Weierstrass’s Theorem is an immediate corollary of this, applied to \( X = I \) and \( A \) the algebra of polynomials.

There is also a complex version:

**XV.8.2.3. Theorem.** [Complex Stone-Weierstrass] Let \( X \) be a compact Hausdorff space, and \( C(X) \) the algebra of complex-valued continuous functions on \( X \). Let \( A \) be a (complex) subalgebra of \( C(X) \) which is closed under complex conjugation, uniformly closed, and separates points. Then either \( A = C(X) \) or there is a \( p \in X \) such that \( A \) consists of all \( f \in C(X) \) with \( f(p) = 0 \).

A complex version of XV.8.2.2. follows immediately:

1708
**XV.8.2.4. Corollary.** Let $X$ be a compact Hausdorff space, and $A$ a subalgebra of $C(X)$ which separates points, contains the constant functions, and is closed under complex conjugation. Then $A$ is dense in $C(X)$ for the uniform norm.

To prove XV.8.2.3. from XV.8.2.1., note that if $A$ is closed under complex conjugation and $f \in A$, then the real and imaginary parts $Re(f) = (f + \bar{f})/2$ and $Im(f) = (f - \bar{f})/2i$ are also in $A$ (the converse is also true). The real-valued functions in $A$ thus satisfy the hypotheses of XV.8.2.1.

**XV.8.2.5.** The hypothesis that $A$ be closed under complex conjugation, or something similar, is necessary; see Exercise XV.8.3.10.

**XV.8.2.6.** Here is an important observation for the proof of XV.8.2.1. If $A$ is an algebra of real-valued functions on a space $X$, $f \in A$, and $p$ is a polynomial with real coefficients and no constant term, then $p \circ f \in A$: if $p(t) = a_1t + a_2t^2 + \cdots + a_nt^n$, then $p \circ f = a_1f + a_2f^2 + \cdots + a_nf^n \in A$. If $A$ contains the constant functions, the same is true if $p$ is a polynomial with constant term. If $(p_n)$ is a sequence of polynomials (without constant term) which converge uniformly on the range of $f$ to a function $g$, then $p_n \circ f \to g \circ f$ uniformly on $X$, so if $A$ is uniformly closed, $g \circ f \in A$. Here is an important application:

**XV.8.2.7. Lemma.** Let $A$ be a uniformly closed algebra of real-valued bounded functions on a set $X$. If $f \in A$, then $|f| \in A$.

Since $|f| = g \circ f$, where $g(t) = |t|$, this follows from the next lemma:

**XV.8.2.8. Lemma.** For any $m \in \mathbb{N}$ there is a sequence of polynomials without constant term converging uniformly to $g(t) = |t|$ on $[-m, m]$.

**Proof:** This can be quickly deduced from XV.8.1.1., but we give a simple direct argument. First consider the interval $[-1, 1]$. By V.17.7.9., the Taylor polynomials $q_n(t)$ around 0 converge uniformly to $h(t) = (1 + t)^{1/2}$ on $[-1, 1]$. Set $r_n(t) = q_n(t^2 - 1)$; then $r_n$ is a polynomial, and $r_n(t)$ converges uniformly to $h(t^2 - 1) = g(t)$ on $[-1, 1]$ (since $t^2 - 1$ maps $[-1, 1]$ into $[-1, 1]$). It is unclear whether $r_n$ has a nonzero constant term; but if $c_n$ is the constant term of $r_n$, then $c_n = r_n(0) \to g(0) = 0$, so if $p_n(t) = r_n(t) - c_n$, then $p_n$ is a polynomial without constant term, and $p_n \to g$ uniformly on $[-1, 1]$.

For $[-m, m]$, set $\phi_n(t) = mp_n(t/m)$. Then $\phi_n$ is a polynomial without constant term, and $\phi_n \to g$ uniformly on $[-m, m]$.

**XV.8.2.9. Lemma.** Let $A$ be a uniformly closed algebra of real-valued bounded functions on a set $X$. If $f, g \in A$, then $f \vee g = \max(f, g)$ and $f \wedge g = \min(f, g)$ are in $A$.

**Proof:** $f \vee g = \frac{1}{2}(f + g + |f - g|)$ and $f \wedge g = \frac{1}{2}(f + g - |f - g|)$.

By induction, if $f_1, \ldots, f_n \in A$, then $f_1 \vee \cdots \vee f_n$ and $f_1 \wedge \cdots \wedge f_n$ are in $A$.  

1769
**XV.8.2.10. Lemma.** Let $A$ be an algebra of real-valued functions on $X$ which separates the points of $X$, and $x, y \in X$. If there are $f, g \in A$ with $f(x) \neq 0$ and $g(y) \neq 0$, then for any $a, b \in \mathbb{R}$ there is an $h \in A$ with $h(x) = a$ and $h(y) = b$.

**Proof:** If $\phi = f^2 + g^2$, then $\phi(x) > 0$ and $\phi(y) > 0$. If $\phi(x) = \phi(y)$, let $\psi$ be a function in $A$ with $\psi(x) \neq \psi(y)$. Then for some (almost any) $r \in \mathbb{R}$ the function $\theta = \phi + r\psi$ has $c = \theta(x)$ and $d = \theta(y)$ distinct nonzero numbers. It is easy to find a quadratic polynomial $p$ with no constant term such that $p(c) = a$ and $p(d) = b$ by solving two linear equations for the coefficients. Then $h = p \circ \theta$ is the desired element of $A$. $\blacksquare$

We are thus in the situation where the following definition applies (for a more general version, see ()):

**XV.8.2.11. Definition.** Let $X$ be a compact Hausdorff space. A subspace $L$ of $C_R(X)$ is a lattice if $f \land g \in L$ for all $f, g \in L$.

A lattice $L$ in $C_R(X)$ has the strong separation property if, whenever $x, y \in X$, $x \neq y$, and $a, b \in \mathbb{R}$, there is an $f \in L$ with $f(x) = a$ and $f(y) = b$.

Since $f \lor g = -[(-f) \land (-g)]$, if $L$ is a lattice and $f, g \in L$, then $f \lor g \in L$.

**XV.8.2.12.** A uniformly closed subalgebra $A$ of $C_R(X)$ is a lattice by XV.8.2.9., and, if $A$ separates points and there is no point of $X$ at which all the functions in $A$ vanish, $A$ has the strong separation property by XV.8.2.10.. Note also:

**XV.8.2.13. Proposition.** Let $X$ be a compact Hausdorff space, and $L$ a lattice in $C_R(X)$. If $L$ separates points and the constant functions are in $L$, then $L$ has the strong separation property.

**Proof:** Let $x, y \in A$, $x \neq y$, and $a \leq b \in \mathbb{R}$. There is an $f \in L$ with $f(x) \neq f(y)$. Replacing $f$ by $-f$ if necessary, we may assume $f(x) < f(y)$, and multiplying $f$ by a scalar we may assume $f(y) - f(x) \geq b - a$. Adding a constant to $f$, we may assume $f(x) = a$, and thus $f(y) \geq b$. If $g = f \land (b1)$, then $g \in L$, $g(x) = a$, and $g(y) = b$. The case $a \geq b$ is obtained by interchanging $x$ and $y$. $\blacksquare$

We then have the lattice version of XV.8.2.2:

**XV.8.2.14. Theorem.** [STONE] Let $X$ be a compact Hausdorff space, and $L$ a lattice in $C_R(X)$ with the strong separation property. Then $L$ is uniformly dense in $C_R(X)$.

**Proof:** Let $f \in C_R(X)$. By the strong separation property, for each $x, y \in X$ there is a $g_{xy} \in L$ with $g_{xy}(x) = f(x)$ and $g_{xy}(y) = f(y)$. Let $\epsilon > 0$. Fix $x$. Since $f$ and $g_{xy}$ are continuous, for each $y$ there is a neighborhood $U_y$ of $y$ such that $g_{xy}(z) < f(z) + \epsilon$ for all $z \in U_y$. Since $X$ is compact, there are finitely many of these sets which cover $X$, say $U_{y_1}, \ldots, U_{y_n}$. Set $h_x = g_{xy_1} \land \cdots \land g_{xy_n}$. Then $h_x \in L$, $h_x(x) = f(x)$, and $h_x(z) < f(z) + \epsilon$ for all $z \in X$. There is such an $h_x$ for every $x \in X$. For each $x$, there is a neighborhood $V_x$ of $x$ such that $h_x(z) > f(z) - \epsilon$ for all $z \in V_x$. Choose a finite subcover $\{V_{x_1}, \ldots, V_{x_m}\}$. If $\phi = h_{x_1} \lor \cdots \lor h_{x_m}$, then $\phi \in L$ and $\|\phi - f\|_{\infty} < \epsilon$. There is such a $\phi$ for every $\epsilon$, so $f \in \bar{L}$. (The proof uses the AC.) $\blacksquare$
**XV.8.2.15.** Corollary. Let $X$ be a compact Hausdorff space, and $L$ a lattice in $C_R(X)$ which separates points and contains the constant functions. Then $L$ is uniformly dense in $C_R(X)$.

**XV.8.2.16.** We now prove XV.8.2.1. Suppose first that for every $x \in X$ there is a $g \in A$ with $g(x) \neq 0$. Then the hypotheses of XV.8.2.14. hold (XV.8.2.12.), so $A$ is uniformly dense in $C_R(X)$, and since $A$ is uniformly closed, $A = C_R(X)$.

Now suppose there is a $p \in X$ such that $g(p) = 0$ for all $g \in A$. (There cannot be more than one such $p$ since $A$ separates points.) Let $B$ be the set of all functions of the form $f = g + c$, where $g \in A$ and $c$ is a constant. Note that for such an $f$, $c = f(p)$. It is clear that $B$ is an algebra of functions. Suppose $(f_n)$ is a sequence in $B$, and $f_n \to f$ uniformly. Write $f_n = g_n + c_n$, with $g_n \in A$ and $c_n = f_n(p)$ a constant. Then $c_n \to c = f(p)$, and hence $g_n \to g = f - c$ uniformly. Thus $g \in A$ since $A$ is uniformly closed, and $f = g + c \in B$. Thus $B$ is uniformly closed. By the first part of the proof, $B = C_R(X)$. Now suppose $f \in C_R(X)$ with $f(p) = 0$. Then $f \in B$, so $f = g + c$ for some $g \in A$ and constant $c$. But $0 = f(p) = g(p) + c = 0 + c = c$, so $f = g \in A$.

This completes the proof of XV.8.2.1. (The proof uses the AC.)

**XV.8.3. Exercises**

**XV.8.3.1.** Prove the following corollary of XV.8.2.1.:

**Theorem.** Let $X$ be a locally compact Hausdorff space. Let $A$ be a uniformly closed subalgebra of $C_R^0(X)$ which separates points of $X$, such that there is no point of $X$ at which all the functions in $A$ vanish. Then $A = C_R^0(X)$.

[Regard $A$ as a subalgebra of $C_R(X)$.]

**XV.8.3.2.** Let $X$ be a locally compact, $\sigma$-compact Hausdorff space. Let $A$ be an algebra of real-valued continuous functions on $X$ which separate points of $X$, such that there is no point of $X$ at which all the functions in $A$ vanish. Show that if $f$ is any real-valued continuous function on $X$, then there is a sequence $(f_n)$ of functions in $A$ converging u.c. to $f$ on $X$.

**XV.8.3.3.** General Weierstrass Approximation Theorem. Let $X$ be a compact subset of $\mathbb{R}^n$ for some $n$.

(a) Show that any continuous function from $X$ to $\mathbb{R}$ can be uniformly approximated arbitrarily closely by a polynomial in the coordinates (in $\mathbb{R}^n$).

(b) Show that any continuous function from $X$ to $\mathbb{R}^m$ for any $m$ can be uniformly approximated arbitrarily closely by a function whose $m$ coordinate functions are all polynomials in the coordinates (in $\mathbb{R}^n$).

**XV.8.3.4.** Let $X$ be a compact subset of $\mathbb{C}$. Let $A$ be the subalgebra of $C(X)$ consisting of all functions of the form $g(z) = p(z,)$, where $p$ is a polynomial of two variables with complex coefficients. Then $A$ separates points of $X$ and contains the constant functions. Show that $A$ is closed under complex conjugation and hence is dense in $C(X)$ by XV.8.2.4.
XV.8.3.5. (a) Let $X$ and $Y$ be compact Hausdorff spaces. Show that every real-valued continuous function on $X \times Y$ can be uniformly approximated arbitrarily closely by (finite) sums of functions of the form $f \otimes g$, where $f \in C_R(X)$, $g \in C_R(Y)$, and

$$(f \otimes g)(x, y) = f(x)g(y).$$

(b) Prove a similar result for complex-valued continuous functions. In fact, every complex-valued continuous function on $X \times Y$ is a uniform limit of finite sums of functions of the form $\lambda(f \otimes g)$, where $\lambda$ is 1 or $i$ and $f$ and $g$ are real-valued continuous functions on $X$ and $Y$ respectively.

(c) Use (a) to give an alternate proof of Fubini’s Theorem for continuous functions (XIV.3.4.7.):

**Theorem.** Let $R = [a, b] \times [c, d]$ be a rectangle in $\mathbb{R}^2$, and $h : R \to \mathbb{R}$ a continuous function. Then

$$\int \int_{R} h \, dA = \int_{a}^{b} \int_{c}^{d} h(x, y) \, dy \, dx = \int_{c}^{d} \int_{a}^{b} h(x, y) \, dx \, dy.$$

[First prove the theorem if $h = f \otimes g$, and take uniform limits.] Generalize to Radon measures on general compact Hausdorff spaces.

XV.8.3.6. Let $a, b \in \mathbb{R}$, $a < b$. Let $C_{sing}([a, b])$ be the set of all continuous functions $f : [a, b] \to \mathbb{R}$ such that $f$ is differentiable a.e. and $f' = 0$ a.e. on $[a, b]$.

(a) Show that $C_{sing}([a, b])$ is a subalgebra of $C_R([a, b])$ containing the constant functions.

(b) By scaling the Cantor function, show that $C_{sing}([a, b])$ separates points of $[a, b]$.

(c) Conclude that $C_{sing}([a, b])$ is uniformly dense in $C_R([a, b])$.

There is an explicit proof of this result not using the Stone-Weierstrass Theorem; cf. XIV.17.5.4..

XV.8.3.7. Let $f$ be a continuous function on an interval $[a, b]$, and suppose that all the moments

$$\int_{a}^{b} f(x)x^n \, dx$$

$(n \in \mathbb{N}_0)$ of $f$ are zero. Prove that $f$ is identically zero on $[a, b]$. [It suffices to show that

$$\int_{a}^{b} [f(x)]^2 \, dx = 0.$$

Show that

$$\int_{a}^{b} f(x)p(x) \, dx = 0$$

for any polynomial $p$, and apply the Weierstrass Approximation Theorem.]
XV.8.3.8. Here is an “integral” version of Weierstrass’s Theorem.
(a) If \( f \) is a polynomial with integer coefficients, regarded as an element of \( C([0, 1]) \), then of course \( f(0) \) and \( f(1) \) are integers. The rest of the problem will show that if \( f \) is an element of \( C([0, 1]) \) such that \( f(0) \) and \( f(1) \) are integers, then \( f \) is a uniform limit on \([0, 1]\) of polynomials with integer coefficients.
(b) Let \( g(x) = x + x(1 - 2x)(1 - x) = 2x - 3x^2 + 2x^3 \). Show that \( g \) is a strictly increasing function from \([0, 1]\) onto \([0, 1]\). The function \( g \) has fixed points 0, 1/2, and 1; if 0 < \( x \) < 1/2, then \( x < g(x) < 1/2 \); if 1/2 < \( x \) < 1, then 1/2 < \( g(x) \) < \( x \).
(c) Show that the iterates \( (g_i) \) of \( g \) converge pointwise on \((0, 1)\) to the constant function 1/2, and that the convergence is uniform on compact subsets (i.e. on subintervals of the form \([\epsilon, 1 - \epsilon]\) for \( \epsilon > 0 \)). Such iterates are strictly increasing.
(d) Show by induction on \( n \) that if \( k \) is any integer, 1 ≤ \( k \) ≤ 2\( n \) − 1, and \( \epsilon > 0 \), then there are integer linear combinations \((\phi_m)\) of powers of the \( g_i \) and \( 1 - g_i \) such that 0 ≤ \( \phi_m \) ≤ 1 on \([0, 1]\) and such that \((\phi_m)\) converges uniformly to the constant function \( \frac{1}{2n} \) on \([\epsilon, 1 - \epsilon]\).
(e) If \( p \) is a polynomial with integer coefficients and \( p(0) = p(1) = 0 \), and \( \lambda \in \mathbb{R} \), show that there is a sequence of polynomials with integer coefficients converging uniformly to \( \lambda p \) on \([-1, 1]\).
(f) If \( f \) is a continuous function on \([0, 1]\) such that \( f(0) \) and \( f(1) \) are integers, then there is a linear polynomial \( q \) with integer coefficients with \( q(x) = f(x) \) for \( x = 0, 1 \). Apply the Stone-Weierstrass Theorem and (e) to conclude that \( f - q \), and hence \( f \), can be approximated uniformly on \([0, 1]\) by polynomials with integer coefficients.

XV.8.3.9. This problem is an extension of Exercise XV.8.3.8.
(a) Let \( f \) be a polynomial with integer coefficients, regarded as an element of \( C([-1, 1]) \). Show that \( f(-1) \), \( f(0) \), and \( f(1) \) are integers, and that \( f(-1) \equiv f(1) \mod 2 \) (i.e. both are even or both odd). The rest of the problem will show that if \( f \) is an element of \( C([-1, 1]) \) such that \( f(-1), f(0) \), and \( f(1) \) are integers, and \( f(-1) \equiv f(1) \mod 2 \), then \( f \) is a uniform limit on \([-1, 1]\) of polynomials with integer coefficients.
(b) If \( p \) is a polynomial with integer coefficients and \( p(-1) = p(0) = p(1) = 0 \), and \( \lambda \in \mathbb{R} \), show that there is a sequence of polynomials with integer coefficients converging uniformly to \( \lambda p \) on \([-1, 1]\). [Consider functions of the form \( \psi(x) = \phi(1 - x^2) \), where \( \phi \) is one of the \( \phi_m \) from (d).]
(c) If \( f \) is a continuous function on \([-1, 1]\) such that \( f(-1), f(0), \) and \( f(1) \) are integers and \( f(-1) \equiv f(1) \mod 2 \), then there is a quadratic polynomial \( q \) with integer coefficients with \( q(x) = f(x) \) for \( x = -1, 0, 1 \). Apply the Stone-Weierstrass Theorem and (b) to conclude that \( f - q \), and hence \( f \), can be approximated uniformly on \([-1, 1]\) by polynomials with integer coefficients.

XV.8.3.10. (a) Let \( T \) be the unit circle in \( \mathbb{C} \). Each polynomial with complex coefficients can be regarded as a continuous function from \( T \) to \( \mathbb{C} \), and this set of polynomials forms a subalgebra \( P \) of \( C(T) \). \( P \) contains the constant functions and separates points of \( T \). But \( P \) is not closed under complex conjugation (in fact, if \( p \) and \( \bar{p} \) are both in \( P \), then \( p \) is a constant function).
(b) Regard \( T \) alternatively as \([0, 2\pi]\) with the endpoints identified, and \( C(T) \) as the algebra of continuous functions \( f : [0, 2\pi] \to \mathbb{C} \) with \( f(2\pi) = f(0) \). \( P \) then becomes an algebra of “trigonometric polynomials” on \([0, 2\pi]\) (but not all trigonometric polynomials), functions of the form

\[
f(t) = \sum_{k=0}^{n} c_k e^{ikt}
\]

1773
for some $n \in \mathbb{N}$ (depending on $f$) and $\{c_k\}$ in $\mathbb{C}$.

(c) If $f \in P$, then $\int_0^{2\pi} f(t) e^{it} \, dt = 0$. The same holds if $f$ is in the uniform closure $\tilde{P}$ of $P$ in $C(\mathbb{T})$.

(d) For $z \in \mathbb{T}$, let $f(z) = \bar{z}$. Alternatively, if $t \in [0, 2\pi]$, let $f(t) = e^{-it}$. Then $f \in C(\mathbb{T})$, and $\int_0^{2\pi} f(t) e^{it} \, dt = 2\pi$. Thus $f$ is not in $\tilde{P}$.

(e) Let $D$ be the open unit disk in $\mathbb{C}$. Each $f \in P$ defines an analytic function on $D$. If $(f_n)$ is a sequence in $P$ converging uniformly on $T$ to a function $f \in \mathbb{P}$, then, by the Maximum Modulus Theorem, $(f_n)$ also converges uniformly on $D$ to a function which is continuous on $\overline{D}$ and analytic on $D$. Conversely, if $f$ is continuous on $D$ and analytic on $D$, the Taylor polynomials (partial sums of the Taylor series) for $f$ around 0 converge uniformly to $f$ on $\overline{D}$ (and, in particular, on $T$), so $f$ is in $\tilde{P}$. Thus $\tilde{P}$ consists precisely of the complex-valued continuous functions on $T$ which extend to a function continuous on $D$ and analytic on $D$.

(f) $\tilde{P}$ also consists precisely of the functions in $C(T)$ whose negative Fourier coefficients vanish.

(g) A trigonometric polynomial on $[0, 2\pi]$ is a function of the form

$$f(t) = \sum_{k=-n}^{n} c_k e^{ikt}$$

for some $n \in \mathbb{N}$ (depending on $f$) and $\{c_k\}$ in $\mathbb{C}$. Then $f \in C(\mathbb{T})$. A trigonometric polynomial can also be written

$$f(t) = \sum_{k=0}^{n} a_k \cos kt + i \sum_{k=1}^{n} b_k \sin kt$$

for $n \in \mathbb{N}$ and $a_k, b_k \in \mathbb{C}$. Show that the trigonometric polynomials are dense in $C(T)$.

See books on function algebras, e.g. [Bro69], for a thorough study of more sophisticated examples of this type.

XV.8.3.11. Let $X$ be a compact Hausdorff space, and let $P$ be a set of nonnegative real-valued continuous functions on $X$ with the following properties:

(i) $f + g \in P$ for all $f, g \in P$.

(ii) $\alpha f \in P$ for all $f \in P$, $\alpha > 0$.

(iii) Either

(1) $f \land g \in P$ for all $f, g \in P$

or

(2) $f \lor g \in P$ for all $f, g \in P$

Let $L = P - P = \{f - g : f, g \in P\}$.

(a) Show that $L$ is a lattice in $C_{\mathbb{R}}(X)$. Use

$$(f_1 - g_1) \land (f_2 - g_2) = [(f_1 + g_2) \land (f_2 + g_1)] - (g_1 + g_2)$$

$$(f_1 - g_1) \lor (f_2 - g_2) = [(f_1 + g_2) \lor (f_2 + g_1)] - (g_1 + g_2)$$

1774
and $f \vee g = -[(-f) \wedge (-g)]$.]

(b) Conclude from XV.8.2.15. that if $P$ separates points and contains a nonzero constant function, then $L$ is uniformly dense in $C_\mathbb{R}(X)$.

Examples of sets $P$ satisfying the hypotheses are the set of nonnegative continuous convex functions on an interval $[a, b]$ and the set of nonnegative continuous concave functions on $[a, b]$. Thus every continuous function on $[a, b]$ can be uniformly approximated arbitrarily closely by the sum of a continuous convex function and a continuous concave function (difference of two continuous convex functions).

**XV.8.3.12.** Use the Stone-Weierstrass Theorem to give another proof of the Tietze Extension Theorem (X) for compact Hausdorff spaces: let $X$ be a compact Hausdorff space, and $Z$ a closed subset of $X$. Let $A$ be the subalgebra of $C_\mathbb{R}(Z)$ consisting of the restrictions to $Z$ of continuous functions from $X$ to $\mathbb{R}$.

(a) Use Urysohn’s Lemma to show that $A$ is uniformly closed in $C_\mathbb{R}(Z)$. (This is a bit tricky.)

(b) Conclude from the Stone-Weierstrass Theorem that $A = C_\mathbb{R}(Z)$, i.e. that every continuous function from $Z$ to $\mathbb{R}$ extends to $X$.

This proof is not easier than the proof in (X), and the result is less general, but it is a different approach. A simpler argument for (a) can be given using C*-algebra theory (XV.14.2.3.).
XV.9. Hilbert Spaces

A Hilbert space is a vector space over \( \mathbb{R} \) or \( \mathbb{C} \) with an inner product or “dot product” which is complete with respect to the norm induced by the inner product. A Hilbert space can be thought of as a special kind of Banach space which has an intrinsic Euclidean geometry; in fact, every finite-dimensional subspace of a Hilbert space is a Euclidean space.

Hilbert spaces are the nicest kinds of topological vector spaces, essentially Euclidean spaces which are allowed to be infinite-dimensional. They share many pleasant properties with Euclidean spaces. In fact, they are so nice they are not themselves very interesting mathematically: there is up to isometric isomorphism exactly one of each dimension, finite or infinite. Like vector spaces, they are mainly important not in and of themselves, but as the setting for operators, which are highly nontrivial and very important.

Hilbert spaces can be either real or complex. The theories of real and complex Hilbert spaces are essentially identical and can be done in parallel. When we get to operators, however, the theories diverge somewhat, and for technical reasons (e.g. the Fundamental Theorem of Algebra) operator theory works better on complex Hilbert spaces. Thus in applications and more advanced work on operators, complex Hilbert spaces are used almost exclusively. Adhering to custom, we will take the default meaning of the term “Hilbert space” to be “complex Hilbert space.”

XV.9.1. Inner Products

Although a unified treatment can be given (XV.9.1.10.), it is clearer to separately define inner products in real and complex vector spaces.

Real Inner Products

XV.9.1.1. Definition. Let \( \mathcal{V} \) be a vector space over \( \mathbb{R} \). A bilinear form on \( \mathcal{V} \) is a function

\[
\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}
\]

(i.e. a function which assigns to each ordered pair of vectors a scalar) which is linear in each entry when the other entry is fixed, i.e.

(i) \( \langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle \) for all \( x_1, x_2, y \in \mathcal{V} \).

(i') \( \langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle \) for all \( x, y_1, y_2 \in \mathcal{V} \).

(ii) \( \langle \alpha x, y \rangle = \alpha \langle x, y \rangle \) for all \( x, y \in \mathcal{V}, \alpha \in \mathbb{R} \).

(ii') \( \langle x, \alpha y \rangle = \alpha \langle x, y \rangle \) for all \( x, y \in \mathcal{V}, \alpha \in \mathbb{R} \).

A bilinear form is symmetric if

(iii) \( \langle y, x \rangle = \langle x, y \rangle \) for all \( x, y \in \mathcal{V} \).

A symmetric bilinear form is positive (or positive semidefinite) if

(iv) \( \langle x, x \rangle \geq 0 \) for all \( x \in \mathcal{V} \).

A positive symmetric bilinear form is definite (or positive definite) if

(v) \( \langle x, x \rangle = 0 \) for \( x \in \mathcal{V} \) if and only if \( x = 0 \).

A positive symmetric bilinear form is called a real pre-inner product.

A positive definite symmetric bilinear form is called a real inner product.
XV.9.1.2. Remarks. (i) Note that the inner product of two vectors is a *scalar*. Thus an inner product is not really a “multiplication” on the vector space, although it has some flavor of a product.

(ii) In a symmetric form, properties (i′) and (ii′) follow automatically from (i) and (ii) respectively. Since most forms we consider are symmetric (or hermitian in the complex case), we will rarely have to separately consider (i′) and (ii′).

XV.9.1.3. Notation. The notation \( \langle x; y \rangle \) is probably the most common notation for an inner product. Other common notations are \((x; y), \langle x|y \rangle, \text{ and } (x|y)\). Additional notations are also occasionally used.

XV.9.1.4. Examples. (i) On \( \mathbb{R}^n \), the usual dot product is a real inner product: if \( x = (x_1, \ldots, x_n) \) and \( y = (y_1, \ldots, y_n) \), then

\[
\langle x, y \rangle = x_1y_1 + \cdots + x_ny_n = \sum_{k=1}^{n} x_ky_k.
\]

This is the *standard inner product* on \( \mathbb{R}^n \).

(ii) On \( C_\mathbb{R}([a,b]) \), the real vector space of continuous real-valued functions on \([a,b]\), define

\[
\langle f, g \rangle = \int_{a}^{b} f(t)g(t) \, dt.
\]

then \( \langle \cdot, \cdot \rangle \) is a real inner product on \( C_\mathbb{R}([a,b]) \).

These two examples obviously have considerable similarity, and can be adapted to a wide range of settings. In fact, these are the “generic” examples; every inner product looks essentially like them (XV.9.5.21.).

(iii) Write elements of \( \mathbb{R}^n \) as column vectors. If \( A \) is an \( n \times n \) matrix over \( \mathbb{R} \), for \( x, y \in \mathbb{R}^n \) define

\[
\langle x, y \rangle = (Ax, y) = y^tAx
\]

where \( \langle \cdot, \cdot \rangle \) is the standard inner product. Then \( \langle \cdot, \cdot \rangle \) is a bilinear form on \( \mathbb{R}^n \). We have

\[
\langle y, x \rangle = x^tAy = [y^tA^tx]^t = y^tA^tx = (A^t x, y)
\]

since transpose is the identity on \( \mathbb{R} \); thus \( \langle \cdot, \cdot \rangle \) is symmetric if and only if \( A \) is symmetric (i.e. \( A^t = A \)). The form is a pre-inner product if and only if \( A \) is positive \( () \), or equivalently if \( A = A^t \) and all eigenvalues of \( A \) are nonnegative; it is an inner product if and only if \( A \) is positive definite, i.e. positive and invertible (i.e. \( A = A^t \) and all eigenvalues of \( A \) are strictly positive).

We can recover \( A \): if \( A_{jk} \) is the \((j,k)\) entry of \( A \), then \( A_{jk} = \langle e_k, e_j \rangle \), where \( e_i \) is the \( i \)'th standard basis vector. Conversely, if \( [\cdot, \cdot] \) is any bilinear form on \( \mathbb{R}^n \), set \( B_{jk} = \langle e_k, e_j \rangle \); then a simple calculation shows that

\[
[x, y] = \langle Bx, y \rangle
\]

for any \( x, y \in \mathbb{R}^n \), where \( B = (B_{jk}) \), so every bilinear form on \( \mathbb{R}^n \) is of this type. The matrix \( B \) is called the *matrix of the bilinear form* with respect to the standard basis of \( \mathbb{R}^n \).
Complex Inner Products

XV.9.1.5. The definitions in XV.9.1. make sense if \( \mathbb{R} \) is merely replaced by \( \mathbb{C} \). However, they are not very interesting. In fact, bilinearity is essentially incompatible with positivity: if \( \alpha \in \mathbb{C} \) with \( \alpha \not\in \mathbb{R} \), and \( x \in V \), if \( \langle \cdot, \cdot \rangle \) is bilinear we have

\[
\langle \alpha x, \alpha x \rangle = \alpha^2 \langle x, x \rangle
\]

and since \( \alpha^2 \) is not a nonnegative real number, \( \langle x, x \rangle \) and \( \langle \alpha x, \alpha x \rangle \) cannot both be nonnegative unless they are zero.

Fortunately, there is a fairly simple way around the difficulty, based on the fact that if \( \alpha \in \mathbb{C} \), then \( \alpha \bar{\alpha} \geq 0 \). We replace bilinearity by “sesquilinearity” (sesqui- is a Latin prefix meaning “one and a half”):

XV.9.1.6. Definition. Let \( V \) be a vector space over \( \mathbb{C} \). A sesquilinear form on \( V \) is a function

\[
\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}
\]

(i.e. a function which assigns to each ordered pair of vectors a scalar) which is linear in the first entry when the second entry is fixed and conjugate-linear in the second entry when the first entry is fixed, i.e.

(i) \( \langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle \) for all \( x_1, x_2, y \in V \).

(i') \( \langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle \) for all \( x, y_1, y_2 \in V \).

(ii) \( \langle \alpha x, y \rangle = \alpha \langle x, y \rangle \) for all \( x, y \in V \), \( \alpha \in \mathbb{C} \).

(ii') \( \langle x, \alpha y \rangle = \bar{\alpha} \langle x, y \rangle \) for all \( x, y \in V \), \( \alpha \in \mathbb{C} \).

A sesquilinear form is hermitian if

(iii) \( \langle y, x \rangle = \overline{\langle x, y \rangle} \) for all \( x, y \in V \).

A hermitian sesquilinear form is positive (or positive semidefinite) if

(iv) \( \langle x, x \rangle \geq 0 \) for all \( x \in V \).

A positive hermitian sesquilinear form is definite (or positive definite) if

(v) \( \langle x, x \rangle = 0 \) for \( x \in V \) if and only if \( x = 0 \).

A positive hermitian sesquilinear form is called a pre-inner product.

A positive definite hermitian sesquilinear form is called a (complex) inner product.

XV.9.1.7. Notation: We have made the convention that sesquilinear forms and inner products are linear in the first variable and conjugate-linear in the second, the usual convention in mathematics. For obscure reasons, the opposite convention is common in mathematical physics, so readers should be aware of the convention used in any reference. The difference will have no practical effect, and a trivial notational change (reversing order in the inner product) will convert to the opposite convention.

1778
XV.9.1.8. Examples. (i) On \(\mathbb{C}^n\), if \(x = (x_1, \ldots, x_n)\) and \(y = (y_1, \ldots, y_n)\), define
\[
\langle x, y \rangle = x_1\overline{y_1} + \cdots + x_n\overline{y_n} = \sum_{k=1}^{n} x_k\overline{y_k}.
\]
This is an inner product, called the standard inner product on \(\mathbb{C}^n\).

(ii) On \(C([a, b])\), the complex vector space of continuous complex-valued functions on \([a, b]\), define
\[
\langle f, g \rangle = \int_{a}^{b} f(t)\overline{g(t)} \, dt.
\]
then \(\langle \cdot, \cdot \rangle\) is an inner product on \(C([a, b])\).

These two examples obviously have considerable similarity, and can be adapted to a wide range of settings. In fact, these are the “generic” examples; every inner product looks essentially like them (XV.9.5.21.).

(iii) Write elements of \(\mathbb{C}^n\) as column vectors. If \(A\) is an \(n \times n\) matrix over \(\mathbb{C}\), for \(x, y \in \mathbb{C}^n\) define
\[
(x, y) = \langle Ax, y \rangle = y^*Ax
\]
where \(\langle \cdot, \cdot \rangle\) is the standard inner product and, for a matrix \(B\),
\[
B^* = \overline{B^t}
\]
is the complex conjugate of the transpose. Then \(\langle \cdot, \cdot \rangle\) is a sesquilinear form on \(\mathbb{C}^n\). We have
\[
(y, x) = x^*Ay = [y^*A^*x]^* = \overline{y}^*A^*x = \overline{\langle Ax, y \rangle}
\]
and thus \(\langle \cdot, \cdot \rangle\) is hermitian if and only if \(A\) is hermitian (i.e. \(A^* = A\)). The form is a pre-inner product if and only if \(A\) is positive (\(\langle x, x \rangle \geq 0\)), or equivalently if \(A = A^*\) and all eigenvalues of \(A\) are nonnegative; it is an inner product if and only if \(A\) is positive definite, i.e. positive and invertible (i.e. \(A = A^*\) and all eigenvalues of \(A\) are strictly positive).

We can recover \(A\): if \(A_{jk}\) is the \((j, k)\) entry of \(A\), then \(A_{jk} = (e_k, e_j)\), where \(e_i\) is the \(i^{th}\) standard basis vector. Conversely, if \([\cdot, \cdot]\) is any sesquilinear form on \(\mathbb{C}^n\), set \(B_{jk} = [e_k, e_j]\); then a simple calculation shows that
\[
[x, y] = \langle Bx, y \rangle
\]
for any \(x, y \in \mathbb{C}^n\), where \(B = (B_{jk})\), so every sesquilinear form on \(\mathbb{C}^n\) is of this type. The matrix \(B\) is called the matrix of the sesquilinear form with respect to the standard basis of \(\mathbb{C}^n\).

XV.9.1.9. If \(V\) is a complex vector space, then \(V\) can be regarded as a real vector space by restricting scalar multiplication. However, if \(\langle \cdot, \cdot \rangle\) is a (complex) inner product on \(V\), it does not become a real inner product by simply restricting scalar multiplication (for one thing, it is complex-valued, not real-valued). But there is a simple conversion: if we define
\[
(x, y) = \text{Re} \langle x, y \rangle
\]
then it is easily checked that \(\langle \cdot, \cdot \rangle\) is a real inner product on \(V\) satisfying \(\langle x, x \rangle = \langle x, x \rangle\) for all \(x \in V\).
A Unified Treatment

A unified treatment of real and complex inner products can be given by making the convention that “complex conjugation” $\alpha \mapsto \bar{\alpha}$ is the identity map on $\mathbb{R}$. With this convention, we may meld the two definitions:

**XV.9.1.10.** Definition. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$, and $V$ a vector space over $\mathbb{F}$. A pre-inner product on $V$ is a function

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$$

satisfying

(i) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$ for all $x_1, x_2, y \in V$.

(ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in V$, $\alpha \in \mathbb{F}$.

(iii) $\langle y, x \rangle = \overline{\langle x, y \rangle}$ for all $x, y \in V$.

(iv) $\langle x, x \rangle \geq 0$ for all $x \in V$.

The pre-inner product is an inner product if it additionally satisfies

(v) $\langle x, x \rangle = 0$ for $x \in V$ if and only if $x = 0$.

It is automatic that such a pre-inner product also satisfies

(i') $\langle x_1 + y_1 + y_2, x \rangle = \langle x_1, x \rangle + \langle y_1, x \rangle + \langle y_2, x \rangle$ for all $x, y_1, y_2 \in V$.

(ii') $\langle x, \alpha y \rangle = \bar{\alpha} \langle x, y \rangle$ for all $x, y \in V$, $\alpha \in \mathbb{F}$.

**XV.9.1.11.** Definition. A (real or complex) inner product space is a pair $(V, \langle \cdot, \cdot \rangle)$, where $V$ is a vector space over $\mathbb{R}$ or $\mathbb{C}$ and $\langle \cdot, \cdot \rangle$ is an inner product on $V$.

An inner product space is a vector space $V$ with specified inner product on $V$. As usual, we will sometimes abuse language by saying “Let $V$ be an inner product space.” There will then implicitly be a specified inner product on $V$. (By convention, we will take “inner product space” to mean “complex inner product space” unless otherwise qualified.)

**XV.9.2.** The CBS Inequality and the Norm

An inner product on a real or complex vector space defines a “norm”

$$\| x \| = \sqrt{\langle x, x \rangle}$$

but to prove that this is truly a norm, specifically to prove the triangle inequality, requires a preliminary inequality which is itself extremely important.
**XV.9.2.1.** **Theorem.** [CBS Inequality] Let $(\cdot,\cdot)$ be a pre-inner product on a real or complex vector space $V$. Then

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

for all $x, y \in V$. If the pre-inner product is an inner product, we have equality if and only if $\{x, y\}$ is linearly dependent.

**Proof:** Fix $x, y \in V$. Suppose first that $\langle y, y \rangle \neq 0$ and set

$$\alpha = -\frac{\langle x, y \rangle}{\langle y, y \rangle}.$$

We then have

$$0 \leq \langle x + \alpha y, x + \alpha y \rangle = \langle x, x \rangle + \alpha \langle y, x \rangle + \alpha \langle x, y \rangle + \alpha^2 \langle y, y \rangle$$

$$= \langle x, x \rangle - \left(\frac{\langle x, y \rangle}{\langle y, y \rangle}\right) \langle y, x \rangle + \frac{\langle x, y \rangle \langle y, x \rangle}{\langle y, y \rangle} + \alpha^2 \langle y, y \rangle$$

$$= \langle x, x \rangle - \frac{|\langle x, y \rangle|^2}{\langle y, y \rangle}.$$

and putting the second term on the left and multiplying by $\langle y, y \rangle$ gives the inequality.

Now suppose $\langle y, y \rangle = 0$. If the pre-inner product is an inner product, then $y = 0$, so $\langle x, y \rangle = 0$ by (real) linearity in the second variable and there is nothing to prove. For the general case, set $\beta = -n \langle x, y \rangle$ for $n \in \mathbb{N}$. Then

$$0 \leq \langle x + \beta y, x + \beta y \rangle = \langle x, x \rangle + \beta \langle y, x \rangle + \beta \langle x, y \rangle + \beta^2 \langle y, y \rangle$$

$$= \langle x, x \rangle + \beta \langle y, x \rangle + \beta \langle x, y \rangle = \langle x, x \rangle - 2n |\langle x, y \rangle|^2.$$

This must be true for all $n \in \mathbb{N}$, so $\langle x, y \rangle = 0$.

If $x$ or $y$ is zero, we clearly have equality. Suppose $x$ and $y$ are nonzero and we have equality, and the pre-inner product is an inner product. Then there are suitable scalar multiples $z$ and $w$ of $x$ and $y$ respectively with

$$\langle z, z \rangle = \langle w, w \rangle = \langle z, w \rangle = 1.$$

We then have

$$\langle z - w, z - w \rangle = \langle z, z \rangle - 2\langle z, w \rangle + \langle w, w \rangle = 0$$

so $z = w$, i.e. $x$ and $y$ are scalar multiples of each other. $\blacktriangle$

**XV.9.2.2.** The name “CBS inequality” is not the one most commonly used, but is the most appropriate. The names CAUCHY, BUNYAKOVSKII, and SCHWARZ are associated with the inequality, and it is commonly called by various subsets of these names (most commonly the “Cauchy-Schwarz inequality.”) Here is a summary of the history:

(i) The first version of the CBS inequality to appear in print was for the standard inner product in $\mathbb{R}^n$, by CAUCHY in [Cau21] in 1821. See Exercise XV.9.10.1. for CAUCHY’s proof.

(ii) In 1859, V. BUNYAKOVSKII stated and indicated a proof of the inequality for the inner product of XV.9.1.4.(ii) [Bun59]. This paper, although written in French and not Russian (BUNYAKOVSKII was actually...
Ukrainian and studied in Paris), was published in Russia and was apparently little noticed in Western Europe. See Exercise XV.9.10.3. for BUNYAKOVSKIĬ’s statement and argument.

(iii) In 1885, H. SCHWARZ published an important and influential paper on minimal surfaces [Sch88] in which he proved, as a relatively minor technical result, a slightly more general integral version of the inequality. He was apparently unaware of BUNYAKOVSKIĬ’s paper. Because SCHWARZ’s paper was widely read and studied, the inequality became known and was often called the “Schwarz inequality.” SCHWARZ did make a substantive contribution: although he only stated the result for integrals, he gave a simple and elegant proof which works verbatim for any real inner product. See Exercise XV.9.10.4. for SCHWARZ’S proof.

(iv) It is unclear who first stated the inequality for complex inner products, or who first discovered the proof given above. The arguments of CAUCHY, BUNYAKOVSKIĬ, and SCHWARZ do not work directly for complex inner products, although the complex case can be easily derived from the real case; see Exercise XV.9.10.5.

The website http://www-stat.wharton.upenn.edu/~steele/Publications/Books/CSMC/CSMC_index. html for the book [Ste04] contains more information and links to electronic copies of the papers by BUNYAKOVSKIĬ and SCHWARZ. The book itself has a thorough study of the CBS inequality and many related inequalities.

The Norm

XV.9.2.3. If $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ is an inner product space over $F = \mathbb{R}$ or $\mathbb{C}$, for $x \in \mathcal{V}$ define

$$\|x\| = \sqrt{\langle x, x \rangle}.$$ 

Whenever we refer to a norm on an inner product space, we mean this induced norm unless otherwise specified.

With this notation, the CBS inequality can be rephrased:

$$|\langle x, y \rangle| \leq \|x\| \|y\|$$

for $x, y \in \mathcal{V}$. A useful consequence is:

XV.9.2.4. Proposition. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be an inner product space, and $x \in \mathcal{V}$. Then

$$\|x\| = \max_{\|y\|=1} |\langle x, y \rangle| = \sup_{\|y\|=1} |\langle x, y \rangle|.$$ 

In particular, if $\langle x, y \rangle = 0$ for all $y$ with $\|y\| = 1$, then $x = 0$.

Proof: We have $\|x\| \geq |\langle x, y \rangle|$ for $\|y\| = 1$ by the CBS inequality, and equality is obtained for $y = \frac{x}{\|x\|}$ if $x \neq 0$ and for any $y$ if $x = 0$.

XV.9.2.5. Proposition. $\| \cdot \|$ is a norm on $\mathcal{V}$.

Proof: It is clear that $\|x\| \geq 0$ for all $x$ and $\|x\| = 0$ if and only if $x = 0$. We have

$$\|\alpha x\| = \sqrt{\langle \alpha x, x \rangle} = \sqrt{\alpha \overline{\alpha} \langle x, x \rangle} = |\alpha| \sqrt{\langle x, x \rangle} = |\alpha| \|x\|$$
for \( x \in \mathcal{V}, \alpha \in \mathbb{F} \) since \(|\alpha| = \sqrt{\alpha \bar{\alpha}}\). For the triangle inequality,
\[
\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2
\]
for \( x, y \in \mathcal{V} \).

### The Parallelogram Law and Polarization Identity

The norm on an inner product space satisfies the following geometric property:

**XV.9.2.6. Proposition.** [Parallelogram Law] Let \((\mathcal{V}, \langle \cdot , \cdot \rangle)\) be an inner product space (real or complex), and \( x, y \in \mathcal{V} \). Then
\[
\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).
\]

**Proof:**
\[
\|x + y\|^2 + \|x - y\|^2 = \langle x + y, x + y \rangle + \langle x - y, x - y \rangle
\]
\[
= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle + \langle x, x \rangle - \langle x, y \rangle - \langle y, x \rangle + \langle y, y \rangle
\]
\[
= 2\langle x, x \rangle + 2\langle y, y \rangle = 2(\|x\|^2 + \|y\|^2).
\]

**XV.9.2.7.** “Most” norms on vector spaces do not satisfy the Parallelogram Law and therefore do not come from inner products. For example, on \( \mathbb{R}^2 \), the norm \( \| \cdot \|_2 \) comes from the standard inner product; but neither \( \| \cdot \|_1 \) or \( \| \cdot \|_\infty \) come from an inner product: since \( \|e_1\|_1 = \|e_2\|_1 = 1 \) but \( \|e_1 + e_2\|_1 = \|e_1 - e_2\|_1 = 2 \), the Parallelogram Law is not satisfied for \( \| \cdot \|_1 \), and similarly for \( \| \cdot \|_\infty \).

In fact, we have a converse to XV.9.2.6, sometimes called the Fréchet-von Neumann-Jordan Theorem:

**XV.9.2.8. Proposition.** [Polarization Identity] (i) Let \( \mathcal{V} \) be a real vector space, and \( \| \cdot \| \) a norm on \( \mathcal{V} \). Then \( \| \cdot \| \) is induced from a (real) inner product if and only if it satisfies the Parallelogram Law, and the inner product giving the norm is unique and satisfies
\[
\langle x, y \rangle = \frac{1}{4} (\|x + y\|^2 - \|x - y\|^2)
\]
for \( x, y \in \mathcal{V} \).

(ii) Let \( \mathcal{V} \) be a complex vector space, and \( \| \cdot \| \) a norm on \( \mathcal{V} \). Then \( \| \cdot \| \) is induced from an inner product if and only if it satisfies the Parallelogram Law, and the inner product giving the norm is unique and satisfies
\[
\langle x, y \rangle = \frac{1}{4} (\|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2)
\]

1783
for $x, y \in \mathcal{V}$.

**Proof:** We write the proof of (ii); the proof of (i) is nearly identical but easier. If the norm comes from an inner product, then a straightforward argument similar to the proof of XV.9.2.6. (Exercise XV.9.10.9.) shows that the inner product satisfies the Polarization Identity and is thus unique. Conversely, if the “inner product” defined by the Polarization Identity is a true inner product, then for $x \in \mathcal{V}$

$$\langle x, x \rangle = \frac{1}{4} \left( \|2x\|^2 - \|0\|^2 + i\|1 + i\|x\|^2 - i\|(1 - i)\|x\|^2 \right) = \|x\|^2$$

since $|1 + i| = |1 - i|$; so the norm it defines is the given one, which therefore satisfies the Parallelogram Law. Thus it suffices to show that if the norm satisfies the Parallelogram Law, the Polarization Identity defines an inner product on $\mathcal{V}$.

So suppose the norm satisfies the Parallelogram Law, and define $\langle \cdot, \cdot \rangle$ by the Polarization Identity. A simple straightforward calculation (not using the Parallelogram Law) shows that $\langle y, x \rangle = \overline{\langle x, y \rangle}$ and $\langle x, iy \rangle = -i \langle x, y \rangle$ for any $x, y \in \mathcal{V}$, and hence $\langle ix, y \rangle = i \langle x, y \rangle$ for $x, y \in \mathcal{V}$. Next, for $x, y, z \in \mathcal{V}$,

$$\|x + y + 2z\|^2 + \|x - y\|^2 = \|(x + z) + (y + z)\|^2 + \|(x + z) - (y + z)\|^2 = 2(\|x + z\|^2 + \|y + z\|^2)$$

by the Parallelogram Law, and similarly

$$\|x + y - 2z\|^2 + \|x - y\|^2 = 2(\|x - z\|^2 + \|y - z\|^2)$$

$$\|x + y + 2iz\|^2 + \|x - y\|^2 = 2(\|x + iz\|^2 + \|y + iz\|^2)$$

$$\|x + y - 2iz\|^2 + \|x - y\|^2 = 2(\|x - iz\|^2 + \|y - iz\|^2)$$

so we have

$$4\langle x + y, 2z \rangle = \|x + y + 2z\|^2 - \|x + y - 2z\|^2 + i\|x + y + 2iz\|^2 - i\|x + y - 2iz\|^2$$

$$= 2\|x + z\|^2 + 2\|y + z\|^2 - 2\|x - z\|^2 - 2\|y - z\|^2 + 2i\|x + iz\|^2 - 2i\||y + iz\|^2 - 2i\|x - iz\|^2 - 2i\|y - iz\|^2$$

$$= 8(\langle x, z \rangle + \langle y, z \rangle)$$

and in the special case $y = 0$,

$$\langle x, 2z \rangle = 2\langle x, z \rangle$$

and therefore, for general $x, y, z$,

$$\langle x + y, z \rangle = \frac{1}{2} \langle x + y, 2z \rangle = \langle x, z \rangle + \langle y, z \rangle$$

and $\langle \cdot, \cdot \rangle$ is additive in each variable. We have also

$$\langle 2x, y \rangle = 2\langle x, y \rangle$$

for any $x, y$ and we get by iteration that

$$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$$

for all $x, y \in \mathcal{V}$ and every dyadic rational number $\alpha$. Since

$$\langle \alpha x, y \rangle = \frac{1}{4} \left( \|\alpha x + y\|^2 - \|\alpha x - y\|^2 + i\|\alpha x + iy\|^2 - i\|\alpha x - iy\|^2 \right)$$

1784
is clearly a continuous function of $\alpha$ for fixed $x, y$, we have
\[ \langle \alpha x, y \rangle = \alpha \langle x, y \rangle \]
for all $\alpha \in \mathbb{R}$. Since it is also true for $\alpha = i$, it holds for all $\alpha \in \mathbb{C}$ and $\langle \cdot, \cdot \rangle$ is an inner product.  

Thus the norm and inner product completely determine each other. One way of phrasing this is:

**XV.9.2.9. Corollary.** Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ and $(\mathcal{W}, \langle \cdot, \cdot \rangle)$ be inner product spaces over the same field, and $T : \mathcal{V} \to \mathcal{W}$ a linear transformation. Then $T$ is an isometry (i.e. $\|Tx\| = \|x\|$ for all $x \in \mathcal{V}$) if and only if $T$ preserves inner products (i.e. $(Tx, Ty) = \langle x, y \rangle$ for all $x, y \in \mathcal{V}$).

**XV.9.2.10. Continuity of the Inner Product**

Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be an inner product space. Then the inner product is **jointly continuous**: if $(x_n)$ and $(y_n)$ are sequences in $\mathcal{V}$, $x, y \in \mathcal{V}$, and $x_n \to x$, $y_n \to y$, then
\[ \langle x_n, y_n \rangle \to \langle x, y \rangle . \]
In particular, the inner product is **separately continuous**:
\[ \langle x_n, y \rangle \to \langle x, y \rangle \text{ and } \langle x, y_n \rangle \to \langle x, y \rangle . \]

**Proof:** We have
\[ |\langle x_n, y_n \rangle - \langle x, y \rangle| \leq |\langle x_n, y_n \rangle - \langle x_n, y \rangle| + |\langle x_n, y \rangle - \langle x, y \rangle| = |(x_n, y_n - y)| + |(x_n - x, y)| \leq \|x_n\|\|y_n - y\| + \|x_n - x\|\|y\| . \]
Since $(x_n)$ converges, it is bounded; hence both terms go to 0 as $n \to \infty$. 

**XV.9.3. Completeness and Hilbert Spaces**

**XV.9.3.1. Definition.** A (complex) **Hilbert space** is a complex inner product space which is complete with respect to the induced norm. A **real Hilbert space** is a real inner product space which is complete with respect to the induced norm. In light of XV.9.2.8., we can also say, by slight abuse of terminology, that a Hilbert space is a Banach space whose norm satisfies the Parallelogram Law.

In this section, the term “Hilbert space” will denote either a real or complex Hilbert space. But in the rest of this volume, the term “Hilbert space,” unless qualified with “real,” will denote a complex Hilbert space. The letter $\mathcal{H}$ will generally be used to denote a Hilbert space.

Note that specification of an inner product (or, equivalently, a norm satisfying the Parallelogram Law) is part of the definition of a Hilbert space. But we will frequently employ the usual abuse of terminology by saying “Let $\mathcal{H}$ be a Hilbert space.”
XV.9.3.2. Examples. (i) \( \mathbb{C}^n \) is a Hilbert space (with respect to the standard inner product) for any \( n \).
\( \mathbb{R}^n \) is a real Hilbert space.

(ii) \( C([a,b]) \) is not a Hilbert space with respect to the inner product of XV.9.1.8.(ii) since it is not complete (Exercise XV.9.10.11.).

(iii) The sequence space \( \ell^2 \) of square-summable sequences of complex numbers is a Hilbert space with respect to the inner product
\[
\langle x, y \rangle = \sum_{k=1}^{\infty} x_k \overline{y_k}
\]
where \( x = (x_1, x_2, \ldots), y = (y_1, y_2, \ldots) \).

(iv) More generally, \( L^2(X, \mathcal{A}, \mu) \), the space of square-integrable functions on a measure space \((X, \mathcal{A}, \mu)\) (or, more precisely, of equivalence classes of square-integrable functions, with functions agreeing almost everywhere identified), with inner product
\[
\langle f, g \rangle = \int_X f \overline{g} \, d\mu
\]
is a Hilbert space. If \( S \) is a set, let \( \mu \) be counting measure on \( S \), and denote \( L^2(S, \mu) \) by \( \ell^2(S) \). Then \( \ell^2 = \ell^2(\mathbb{N}) \).

XV.9.3.3. The definition and basic properties of \( \ell^2 \) were given in 1906 by D. Hilbert, who was the first to describe something approaching the modern notion of a “Hilbert space,” and developed by E. Schmidt, F. Riesz, and others in immediately succeeding years. According to [Jah03, p. 385], the name “Hilbert space” was first used in 1908 by A. Schönflies, apparently to refer to what we today call \( \ell^2 \). The definition of an abstract Hilbert space was, however, not given until 1928 by J. von Neumann [vN30].

Completion of an Inner Product Space

Just as with metric spaces and normed vector spaces, every inner product space can be completed to a Hilbert space:

XV.9.3.4. Theorem. Let \((V, \langle \cdot, \cdot \rangle)\) be a (real or complex) inner product space. Then \( V \) can be isometrically embedded, in a unique way up to isometric isomorphism, as a dense subspace of a (real or complex) Hilbert space, called the completion of \( V \).

Proof: This theorem can be proved directly in the same way as (i) and (ii); see Exercise XV.9.10.13.. But there is a much easier argument using XV.9.2.8.: if \( \mathcal{H} \) is the completion of \( V \) with respect to the induced norm, then the norm on \( \mathcal{H} \) satisfies the Parallelogram Law on a dense subspace, hence by continuity on the whole space. Thus it comes from an inner product which must extend the inner product on \( V \) by Polarization. 

XV.9.3.5. As usual, there is often a concrete realization of the completion of an inner product space: if it can be identified with a subspace of a Hilbert space, the closure of this subspace is the completion.
Pre-Hilbert Spaces

XV.9.3.6. Definition. A pre-Hilbert space is a complex vector space with a pre-inner product.

Real pre-Hilbert spaces can be defined accordingly, but we will stick to the complex case.

Any (complex) inner product space is a pre-Hilbert space. But there are other natural ones. Here is an example of a very common kind:

XV.9.3.7. Example. Let $PC([a; b])$ be the set of piecewise-continuous complex-valued functions on $[a; b]$. Define a sesquilinear form on $PC([a; b])$ by

$$\langle f, g \rangle = \int_a^b f(t) \overline{g(t)} \, dt.$$ 

This is a pre-inner product, but not an inner product: if $f$ is zero except at finitely many points, then $\langle f, f \rangle = 0$, but $f$ is not the zero element of $PC([a; b])$.

XV.9.3.8. Proposition. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be a pre-Hilbert space. Then

(i) If $\mathcal{N} = \{ x \in \mathcal{V} : \langle x, x \rangle = 0 \}$

then $\mathcal{N}$ is a closed subspace of $\mathcal{V}$.

(ii) The pre-inner product on $\mathcal{V}$ induces an inner product on $\mathcal{V}/\mathcal{N}$ by

$$\langle [x], [y] \rangle = \langle x, y \rangle$$

for $x, y \in \mathcal{V}$.

Proof: This is more or less a special case of (i), and is proved in a similar way.

XV.9.3.9. Example. Let $PC([a; b])$ be as in XV.9.3.7. Then $\mathcal{N}$ consists of all functions which are zero except at finitely many points. If $f$ is piecewise-continuous, the equivalence class $[f]$ of $f$ consists precisely of all (necessarily piecewise-continuous) functions which agree with $f$ except at finitely many points. Let $PC([a; b])$ be the set of equivalence classes of piecewise-continuous functions. Then the pre-inner product of XV.9.3.7. gives a well-defined inner product on $PC([a; b])$. We often think of the elements of $PC([a; b])$ as functions on $[a, b]$, but they are technically equivalence classes of functions; but the distinction can often be overlooked without essential difficulty. We can naturally identify $C([a, b])$ with a (dense) subspace of $PC([a; b])$.

XV.9.3.10. Definition. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be a pre-Hilbert space. The completion of $\mathcal{V}$ is the Hilbert space obtained by completing $\mathcal{V}/\mathcal{N}$. 1787
XV.9.4. Best Approximation

Hilbert spaces have a nice property not shared by general Banach spaces which is important in applications. If $A$ is a nonempty closed convex set in a Hilbert space $H$ (e.g. a closed subspace), and $x \in H$, then there is a unique vector $y \in A$ which is closest to $x$. This theorem was proved by BEPPO LEVI in 1906 [?].

We use the notation $\langle x; y \rangle = \|x - y\|$, and

$$\rho(x, A) = \inf_{y \in A} \rho(x, y) = \inf_{y \in A} \|x - y\|.$$

XV.9.4.1. Theorem. Let $H$ be a Hilbert space and $A$ a nonempty closed convex subset of $H$. For any $x \in H$, there is a unique $y \in A$ such that

$$\|x - y\| = \rho(x, y) = \rho(x, A) = \inf_{z \in A} \|x - z\| = \min_{z \in A} \|x - z\|.$$

We need a lemma which is an immediate consequence of the Parallelogram Law:

XV.9.4.2. Lemma. Let $A$ be a nonempty convex set in an inner product space $\mathcal{V}$, $x \in \mathcal{V}$, and $\epsilon > 0$. If $y, z \in A$ with $\|x - y\|^2 < \rho(x, A)^2 + \epsilon$ and $\|x - z\|^2 < \rho(x, A)^2 + \epsilon$, then $\|y - z\|^2 < 4\epsilon$.

Proof: Set $w = \frac{1}{2}(y + z)$. Then $w \in A$ since $A$ is convex, so $\|x - w\| \geq \rho(x, A)$. We have

$$x - w + \frac{1}{2}(y - z) = x - z$$

$$x - w - \frac{1}{2}(y - z) = x - y$$

so by the Parallelogram Law

$$\|x - z\|^2 + \|x - y\|^2 = \left\| x - w + \frac{1}{2}(y - z) \right\|^2 + \left\| x - w - \frac{1}{2}(y - z) \right\|^2 = 2 \left( \|x - w\|^2 + \left\| \frac{1}{2}(y - z) \right\|^2 \right)$$

$$\|y - z\|^2 = 2\|x - z\|^2 + 2\|x - y\|^2 - 4\|x - w\|^2 < 2\rho(x, A)^2 + 2\epsilon + 2\rho(x, A)^2 + 2\epsilon - 4\rho(x, A)^2 = 4\epsilon.$$

We now prove Theorem XV.9.4.1..

Proof: Let $(y_n)$ be a sequence in $A$ with $\|x - y_n\| \to \rho(x, A)$. Then Lemma XV.9.4.2. implies that $(y_n)$ is a Cauchy sequence in $\mathcal{H}$. Let $y = \lim y_n$. Then $y \in A$ since $A$ is closed, and

$$\|x - y\| = \lim_{n \to \infty} \|x - y_n\| = \rho(x, A).$$

If $z \in A$ with $\|x - z\| = \rho(x, A)$, then the hypotheses of Lemma XV.9.4.2. are satisfied for every $\epsilon > 0$, and we conclude $y = z$. 

1788
XV.9.4.3. Lemma XV.9.4.2. really says that an inner product space is uniformly convex (XV.9.10.16.). The argument above shows that the closest approximation theorem holds in any uniformly convex Banach space.

XV.9.4.4. Examples. (i) The closest approximation result does not hold in general Banach spaces. In \((\mathbb{R}^2, \| \cdot \|_\infty)\), if \(A\) is the closed unit ball and \(x = (2, 0)\), then there is a closest vector in \(A\) to \(x\) (there must be since \(A\) is compact), but it is not unique; in fact, any vector \((1, y)\) with \(|y| \leq 1\) is a closest approximation to \(x\) in \(A\).

(ii) The proof of XV.9.4.1. uses completeness. The result fails in general in an inner product space which is not complete. For example, let \(V = C([0, 1])\) with the standard inner product, and let \(A = \{ f \in C([0, 1]) : \int_0^{1/2} f(t) \, dt - \int_{1/2}^1 f(t) \, dt = 1 \} \). Then it is easily checked that \(A\) is a closed convex set in \(V\). But there is no closest vector to 0 (vector of minimum norm) in \(A\): \(\rho(0, A) = 1\), but there is no \(f\) in \(A\) of norm 1. (In the completion, the unique closest vector in \(A\) to 0 is the function which is 1 on \([0, \frac{1}{2})\) and \(-1\) on \((\frac{1}{2}, 1]\).)

XV.9.5. Orthogonality

Hilbert spaces have a geometric structure similar to that of Euclidean space. The most important geometric notion is orthogonality:

XV.9.5.1. Definition. Let \((V, \langle \cdot, \cdot \rangle)\) be an inner product space over \(F = \mathbb{R}\) or \(\mathbb{C}\). If \(x, y \in V\), then \(x\) and \(y\) are orthogonal (with respect to \(\langle \cdot, \cdot \rangle\)), written \(x \perp y\), if \(\langle x, y \rangle = 0\).

XV.9.5.2. The geometric term perpendicular is sometimes used as a synonym for orthogonal. But in functional analysis we prefer to use orthogonal, especially for complex inner product spaces. Note that in a complex inner product space, “perpendicular” and “orthogonal” do not quite have the same meaning: for example, in the one-dimensional (complex) Hilbert space \(\mathbb{C}^1\) with standard inner product, the vectors 1 and \(i\) are “perpendicular” but not orthogonal.

We have \(0 \perp y\) for every \(y\). Conversely, we have a useful rephrasing of the simple result XV.9.2.4.:

XV.9.5.3. Proposition. Let \((V, \langle \cdot, \cdot \rangle)\) be an inner product space, and \(x \in V\). Then \(x \perp y\) for all \(y \in V\) if and only if \(x = 0\).

Orthogonal Complement

XV.9.5.4. Definition. Let \((V, \langle \cdot, \cdot \rangle)\) be an inner product space, and \(S \subseteq V\). The orthogonal complement of \(S\) (in \(V\)) is \(S^\perp = \{ x \in V : \langle x, y \rangle = 0 \text{ for all } y \in S\} \).
\( S^\perp \) is a closed subspace of \( V \) (whether or not \( S \) is a subspace; in fact, \( S^\perp = (\text{span}(S))^\perp \)).

The next norm relation is fundamental.

**XV.9.5.5. Proposition.** [Pythagorean Theorem] Let \( V \) be an inner product space, and \( x, y \in V \) with \( x \perp y \). Then
\[
\|x + y\|^2 = \|x\|^2 + \|y\|^2.
\]

**Proof:** Expand \( \langle x + y, x + y \rangle \).

The converse is true in a real inner product space, but not a complex one (take 1 and \( i \) in \( \mathbb{C} \)). But there is a partial converse:

**XV.9.5.6. Proposition.** Let \( V \) be an inner product space over \( \mathbb{F} \), and \( x, y \in V \). The following are equivalent:

(i) \( x \perp y \).

(ii) \( \|x + \alpha y\|^2 = \|x\|^2 + |\alpha|^2 \|y\|^2 \) for all \( \alpha \in \mathbb{F} \).

(iii) \( \|x + \alpha y\| \geq \|x\| \) for all \( \alpha \in \mathbb{F} \).

**Proof:** (i) \( \Rightarrow \) (ii) is XV.9.5.5., and (ii) \( \Rightarrow \) (iii) is trivial. For (iii) \( \Rightarrow \) (i), we may assume \( y \neq 0 \). Set
\[
\alpha = -\frac{\langle x, y \rangle}{\langle y, y \rangle}
\]
and then a calculation as in the proof of XV.9.2.1. shows that
\[
\|x + \alpha y\|^2 = \|x\|^2 - \frac{|\langle x, y \rangle|^2}{\|y\|^2}
\]
which is only \( \geq \|x\|^2 \) if \( \langle x, y \rangle = 0 \).

If \( \mathcal{H} \) is a Hilbert space, \( \mathcal{Y} \) is a closed subspace of \( \mathcal{H} \), and \( x \in \mathcal{H} \), then there is a closest vector \( y \in \mathcal{Y} \) to \( X \). It turns out that \( z = x - y \in \mathcal{Y}^\perp \):

**XV.9.5.7. Proposition.** Let \( \mathcal{H} \) be a Hilbert space, and \( \mathcal{Y} \) a closed subspace of \( \mathcal{H} \). Then every \( x \in \mathcal{H} \) can be uniquely written as \( y + z \), where \( y \in \mathcal{Y} \) and \( z \in \mathcal{Y}^\perp \). The \( y \) is the closest vector in \( \mathcal{Y} \) to \( x \).

**Proof:** By XV.9.4.1. there is a (unique) closest vector \( y \in \mathcal{Y} \) to \( x \). Set \( z = x - y \). Let \( w \in \mathcal{Y} \). Then, for any \( \alpha \in \mathbb{C} \),
\[
\|z + \alpha w\| = \|x - (y - \alpha w)\| \geq \|x - y\| = \|z\|
\]
since \( y - \alpha w \in \mathcal{Y} \). Thus \( z \perp w \) by XV.9.5.6.. If also \( x = y' + z' \) with \( y' \in \mathcal{Y} \), \( z' \in \mathcal{Y}^\perp \), we have \( y + z - y' - z' = 0 \), so
\[
y - y' = z' - z \in \mathcal{Y} \cap \mathcal{Y}^\perp = \{0\}
\]
and \( y = y' \), \( z = z' \).

1790
XV.9.5.8. **Example.** Completeness is essential in XV.9.5.7. As a slight variation of Example XV.9.4.4., let

\[\mathcal{V} = \left\{ f \in C([0,1]) : \int_0^{1/2} f(t) \, dt - \int_{1/2}^1 f(t) \, dt = 0 \right\}.\]

Then \(\mathcal{V}\) is a closed subspace of \(C([0,1])\) (with respect to the norm induced by the standard inner product). But if \(f\) is the constant function 1, there is no decomposition \(f = g + h\) with \(g \in \mathcal{V}\) and \(h \in \mathcal{V}^\perp\). In fact, \(\mathcal{V}^\perp = \{0\}\) in \(C([0,1])\) even though \(\mathcal{V}\) is a proper subspace of \(C([0,1])\).

XV.9.5.9. If \(\mathcal{V}\) is an inner product space and \(S \subseteq \mathcal{V}\), then \(S^\perp\) is a closed subspace of \(\mathcal{V}\) containing \(S\). If \(\mathcal{V}\) is complete, then \(S^\perp\) is the smallest closed subspace containing \(S\):

**Proposition.** Let \(\mathcal{H}\) be a Hilbert space and \(S \subseteq \mathcal{H}\). Then \(S^\perp\) is the closed linear span of \(S\).

**Proof:** Let \(\mathcal{V}\) be the closed linear span of \(S\), i.e. the closure of \(\text{span}(S)\). Then \(\mathcal{V} \subseteq S^\perp\), and \(\mathcal{V}^\perp = S^\perp\). If \(x \in S^\perp\), then \(x = y + z\) with \(y \in \mathcal{V}\) and \(z \in \mathcal{V}^\perp\). But since \(x, y \in S^\perp\), \(z = x - y \in S^\perp = \mathcal{V}^\perp\). Thus \(z \in \mathcal{V}^\perp \cap \mathcal{V}^\perp = \{0\}\) so \(x = y \in \mathcal{V}\), i.e. \(\mathcal{V} = S^\perp\).

XV.9.5.11. **Corollary.** Let \(\mathcal{H}\) be a Hilbert space and \(\mathcal{V}\) a closed subspace. Then \(\mathcal{V} = \mathcal{V}^\perp\). In particular, \(\mathcal{V} = \mathcal{H}\) if and only if \(\mathcal{V}^\perp = \{0\}\), so if \(\mathcal{S} \subseteq \mathcal{H}\), then the closed linear span of \(S\) is \(\mathcal{S}\) if and only if \(S^\perp = \{0\}\).

This result can fail if the space is not complete (Example XV.9.5.8.).

**Orthonormal Sets**

XV.9.5.12. **Definition.** A set \(\{e_j : j \in J\}\) of vectors in an inner product space is orthonormal if \(\langle e_j, e_k \rangle = \delta_{jk}\) for all \(j, k \in J\), i.e. it is a mutually orthogonal set of unit vectors.

Of course, any subset of an orthonormal set is orthonormal. A set is orthonormal if and only if every finite subset (in fact, every two-element subset) is orthonormal.

XV.9.5.13. **Proposition.** An orthonormal set in an inner product space is linearly independent.

**Proof:** Suppose \(S = \{e_j : j \in J\}\) is an orthonormal set, and suppose

\[\alpha_1 e_{j_1} + \cdots + \alpha_n e_{j_n} = 0\]

for some \(j_1, \ldots, j_n\) and scalars \(\alpha_1, \ldots, \alpha_n\). Then for \(1 \leq m \leq n\)

\[0 = \langle \alpha_1 e_{j_1} + \cdots + \alpha_n e_{j_n}, e_{j_m} \rangle = \sum_{r=1}^n \alpha_r \langle e_{j_r}, e_{j_m} \rangle = \alpha_m\]

1791
so $S$ is linearly independent.

Orthonormal sets are very nice to work with. If $S$ is an orthonormal set, then it is easy to see how to explicitly write any vector in $\text{span}(S)$ as a convex combination of the vectors in $S$:

**XV.9.5.14. Proposition.** Let $S = \{e_1, \ldots, e_n\}$ be a finite orthonormal set in an inner product space, and $x \in \text{span}(S)$. Then the unique representation of $x$ as a linear combination of $\{e_1, \ldots, e_n\}$ is

$$x = \sum_{k=1}^{n} (x, e_k)e_k.$$  

We also have

$$\|x\|^2 = \sum_{k=1}^{n} |(x, e_k)|^2.$$  

**Proof:** Suppose

$$x = \sum_{k=1}^{n} \alpha_k e_k$$

is the unique expression of $x$ as a linear combination of $S$, then, for each $j$,

$$(x, e_j) = \sum_{k=1}^{n} \alpha_k (e_k, e_j) = \alpha_j.$$

The number $(x, e_j)$ is often called the *Fourier coefficient* of $x$ with respect to $e_j$ (cf. (1)).

This finite result has an important generalization to infinite orthonormal sets:

**XV.9.5.15. Theorem.** Let $S = \{e_j : j \in J\}$ be an orthonormal set in an inner product space $\mathcal{V}$, and $\mathcal{Y}$ the closed linear span of $S$ in $\mathcal{V}$. Then, for any $y \in \mathcal{Y}$:

(i) The sum

$$\sum_{j \in J} \langle y, e_j \rangle e_j$$

converges unconditionally to $y$ in $\mathcal{V}$, i.e. the net of finite subsums converges to $y$.

(ii) The unordered sum

$$\sum_{j \in J} |\langle y, e_j \rangle|^2$$

converges (absolutely) to $\|y\|^2$, i.e. the net of finite subsums converges to $\|y\|^2$.  

1792
Proof: Let $F = \{e_{j_1}, \ldots, e_{j_n}\}$ be a finite subset of $S$, and set

$$y_F = \sum_{k=1}^{n} \langle y, e_{j_k} \rangle e_{j_k}$$

and $z_F = y - y_F$. We have that $y_F \in \text{span}(F)$, and that

$$\langle z_F, e_{j_m} \rangle = \langle y, e_{j_m} \rangle - \langle y_F, e_{j_m} \rangle = 0$$

for $1 \leq m \leq n$; thus $z_F \in (\text{span}(F))^\perp$. So we have

$$\|y\|^2 = \|y_F\|^2 + \|z_F\|^2 \geq \|y_F\|^2 = \sum_{k=1}^{n} |\langle y, e_{j_k} \rangle|^2.$$  

Thus the sum in (ii) converges, and is $\leq \|y\|^2$.

On the other hand, if $\epsilon > 0$ there is an $x \in \text{span}(S)$ with $\|x - y\| < \epsilon$. We have $x \in \text{span}(E)$ for some finite subset $E$ of $S$. If $F = \{e_{j_1}, \ldots, e_{j_n}\}$ is a finite subset of of $S$ containing $E$, then

$$\|y - y_F\| \leq \|y - x\|$$

since $y_F$ is the closest vector in $\text{span}(F)$ to $y$. Thus

$$\left\| y - \sum_{k=1}^{n} \langle y, e_{j_k} \rangle e_{j_k} \right\| < \epsilon$$

and the finite subsums in (i) converge to $y$. Also,

$$\sum_{k=1}^{n} |\langle y, e_{j_k} \rangle|^2 = \|y_F\|^2 \geq (\|y\| - \|y - y_F\|)^2 \geq (\|y\| - \epsilon)^2$$

which can be made arbitrarily close to $\|y\|^2$ by choosing $\epsilon$ sufficiently small. Thus the sum in (ii) converges to $\|y\|^2$. 

\[\Box\]

XV.9.5.16. Corollary. Let $\mathcal{H}$ be a Hilbert space, $S = \{e_j : j \in J\}$ an orthonormal set in $\mathcal{H}$, and $x \in \mathcal{H}$. Then

$$\sum_{j \in J} |\langle x, e_j \rangle|^2 \leq \|x\|^2$$

(Bessel's inequality). In particular, $\langle x, e_j \rangle \neq 0$ for only countably many $j$. We have

$$\sum_{j \in J} |\langle x, e_j \rangle|^2 = \|x\|^2$$

(Parseval's relation) if and only if $x$ is in the closed linear span of $S$.

Proof: Let $\mathcal{Y}$ be the closed linear span of $S$, and write $x = y + z$ with $y \in \mathcal{Y}$, $z \in \mathcal{Y}^\perp$. Then

$$\|x\|^2 = \|y\|^2 + \|z\|^2 \geq \|y\|^2 = \sum_{j \in J} |\langle y, e_j \rangle|^2 = \sum_{j \in J} |\langle x, e_j \rangle|^2$$

with equality if and only if $z = 0$. 

\[\Box\]
Orthonormal Bases

The next result is a combination of parts of XV.9.5.11. and XV.9.5.15:

**Corollary.** Let \( \mathcal{H} \) be a Hilbert space, and \( S = \{ e_j : j \in J \} \) an orthonormal set in \( \mathcal{H} \). The following are equivalent:

(i) \( S \) is a maximal orthonormal set (not contained in any strictly larger orthonormal set).

(ii) \( S^\perp = \{ 0 \} \).

(iii) The closed linear span of \( S \) is \( \mathcal{H} \).

(iv) For every \( x \in \mathcal{H} \), there is a unique unordered sum

\[
\sum_{j \in J} \alpha_j e_j
\]

which converges to \( x \). We necessarily have that the unordered sum

\[
\sum_{j \in J} |\alpha_j|^2
\]

converges to \( \|x\|^2 \), and \( \alpha_j = \langle x, e_j \rangle \).

**Proof:** For the equivalence of (i) and (ii), note that if \( S' \) is an orthonormal set containing \( S \), then \( S' \setminus S \subseteq S^\perp \). Conversely, if \( S^\perp \neq \{ 0 \} \), let \( x \) be a unit vector in \( S^\perp \); then \( S \cup \{ x \} \) is an orthonormal set strictly larger than \( S \).

For the uniqueness in (iv), suppose the unordered sum \( \sum_j \alpha_j e_j \) converges to \( x \). Then, by linearity and continuity of the inner product, we have, for each \( k \),

\[
\langle x, e_k \rangle = \lim_{F \subseteq S} \sum_{j \in F} \langle \alpha_j e_j, e_k \rangle = \alpha_k
\]

where the limit is over the net of finite subsets \( F \) of \( S \).

---

**Definition.** If \( \mathcal{H} \) is a Hilbert space, an **orthonormal basis** for \( \mathcal{H} \) is an orthonormal set \( S \) in \( \mathcal{H} \) satisfying the equivalent conditions of XV.9.5.17.

An orthonormal basis is linearly independent (XV.9.5.13.). If \( \mathcal{H} \) is finite-dimensional, then an orthonormal basis for \( \mathcal{H} \) is finite, hence a Hamel basis. But an orthonormal basis for an infinite-dimensional Hilbert space is not a Hamel basis. For example, if \( \{ e_{j_1}, e_{j_2}, \ldots \} \) is a sequence of distinct elements of an orthonormal basis for \( \mathcal{H} \), then the series

\[
\sum_{n=1}^{\infty} \frac{1}{n} e_{j_n}
\]

converges in \( \mathcal{H} \) to a vector which cannot be written as a finite linear combination of the \( e_j \) by the uniqueness in XV.9.5.17.. A countable orthonormal basis for a Hilbert space is an unconditional Schauder basis ( ).
As with Hamel bases in vector spaces, there is nothing remotely unique about orthonormal bases for Hilbert spaces. For example, any two perpendicular unit vectors in $\mathbb{R}^2$ form an orthonormal basis for $\mathbb{R}^2$ or $\mathbb{C}^2$ (with respect to the usual inner product). $\mathbb{C}^2$ has many other orthonormal bases not contained in the subset $\mathbb{R}^2$.

If $\mathcal{H}$ is a complex Hilbert space, whether an orthonormal set is an orthonormal basis depends whether $\mathcal{H}$ is regarded as a complex Hilbert space or as a real Hilbert space with inner product $(\cdot,\cdot) = \text{Re}(\langle \cdot , \cdot \rangle)$. If $B$ is an orthonormal basis for $\mathcal{H}$ as a complex Hilbert space, it is an orthonormal set, but not an orthonormal basis, for $\mathcal{H}$ as a real Hilbert space; the set

$$B \cup iB = \{ e, ie : e \in B \}$$

is an orthonormal basis for $\mathcal{H}$ as a real Hilbert space.

We have a companion result slightly generalizing XV.9.5.17., which is an immediate corollary of XV.9.5.17., the continuity of the inner product, and the CBS inequality:

**Corollary.** Let $\mathcal{H}$ be a Hilbert space, and $\{ e_j : j \in J \}$ an orthonormal basis for $\mathcal{H}$. Let $x, y \in \mathcal{H}$. Write

$$x = \sum_{j \in J} \alpha_j e_j, \quad y = \sum_{j \in J} \beta_j e_j$$

as in XV.9.5.17.. Then the unordered sum

$$\sum_{j \in J} \alpha_j \overline{\beta_j}$$

converges absolutely to $\langle x, y \rangle$.

The next theorem is the main theoretical result about orthonormal bases:

**Theorem.** Let $\mathcal{H}$ be a Hilbert space. Then

(i) $\mathcal{H}$ has an orthonormal basis.

(ii) Every orthonormal set in $\mathcal{H}$ is contained in an orthonormal basis.

(iii) Any two orthonormal bases for $\mathcal{H}$ have the same cardinality.

**Proof:** The proof of (i) is a prototype Zorn’s Lemma argument (using the characterization of an orthonormal basis as a maximal orthonormal set), similar to the proof in III.11.2.32. (but easier), and is left to the reader. Part (ii) can be proved in a similar way: if $S$ is an orthonormal set in $\mathcal{H}$, consider only orthonormal sets containing $S$. Alternately, let $\mathcal{V}$ be the closed linear span of $S$. Then $\mathcal{V}^\perp$ is a closed subspace of $\mathcal{H}$ and thus itself a Hilbert space, so has an orthonormal basis $B'$. Then $B = S \cup B'$ is an orthonormal basis for $\mathcal{H}$ containing $S$.

To prove (iii), first note that if $\mathcal{H}$ has a finite orthonormal basis $B$ with $n$ elements, then $B$ is a Hamel basis for $\mathcal{H}$ and thus $\mathcal{H}$ is an $n$-dimensional vector space. Thus any other orthonormal basis for $\mathcal{H}$ also has exactly $n$ elements. Now assume that one (hence every) orthonormal basis for $\mathcal{H}$ is infinite. Let $\{ e_j : j \in J \}$ be an orthonormal basis for $\mathcal{H}$ of cardinality $\lambda$, and $\{ f_k : k \in K \}$ another orthonormal basis for $\mathcal{H}$ of cardinality $\kappa$. Then each $e_j$ is in the closed linear span of countably many $f_k$ ( ). Every $f_k$ occurs in one of
these, since otherwise $e_j \in \{f_k\}$ for all $j$. Thus $K$ is a union of $\lambda$ countable sets, so $\kappa \leq \aleph_0 \cdot \lambda$. Similarly, $\lambda \leq \aleph_0 \cdot \kappa$. Since $\lambda$ and $\kappa$ are infinite, $\lambda = \kappa$ (II.9.7.4.).

Note that the AC is needed for this proof; in fact, (i) and (ii) may be equivalent to the AC. (The full AC is not needed for (iii), but some weaker form is (II.9.7.4.).)

**XV.9.5.23. Definition.** Let $\mathcal{H}$ be a Hilbert space over $\mathbb{F}$ ($\mathbb{R}$ or $\mathbb{C}$). The **orthogonal dimension** (or **Hilbert dimension**) of $\mathcal{H}$ (over $\mathbb{F}$) is the cardinality of some (hence every) orthonormal basis for $\mathcal{H}$.

**XV.9.5.24.** The orthogonal dimension of $\mathcal{H}$ is finite if and only if $\mathcal{H}$ is a finite-dimensional vector space, in which case the orthogonal dimension is the same as the vector space dimension. The two dimensions do not coincide for infinite-dimensional Hilbert spaces in general ('). We usually only consider the orthogonal dimension for a Hilbert space $\mathcal{H}$; it is often just called the **dimension** of $\mathcal{H}$ and denoted $\dim(\mathcal{H})$.

The orthogonal dimension of a finite-dimensional complex Hilbert space changes (is multiplied by 2) if it is regarded as a real Hilbert space with inner product $(\cdot, \cdot) = \text{Re}(\langle \cdot, \cdot \rangle)$. The orthogonal dimension of an infinite-dimensional complex Hilbert space does not change if it is regarded as a real Hilbert space (since $2\kappa = \kappa$ for any infinite cardinal $\kappa$).

The orthogonal dimension classifies Hilbert spaces up to isometric isomorphism; thus the orthogonal dimension is the only isomorphism invariant for Hilbert spaces:

**XV.9.5.25. Theorem.** Let $\mathcal{H}$ and $\mathcal{H}'$ be Hilbert spaces over $\mathbb{F}$. Then $\mathcal{H}$ and $\mathcal{H}'$ are isometrically isomorphic (over $\mathbb{F}$) if and only if they have the same orthogonal dimension. More precisely, if $B = \{e_j : j \in J\}$ is an orthonormal basis for $\mathcal{H}$ and $B' = \{f_j : j \in J\}$ is an orthonormal basis for $\mathcal{H}'$, the map $e_j \mapsto f_j$ extends to a unique isometric isomorphism $T : \mathcal{H} \to \mathcal{H}'$.

**Proof:** It is obvious from the uniqueness of orthogonal dimension that Hilbert spaces of different orthogonal dimensions cannot be isometrically isomorphic. So we need only prove the last statement. The map $T_0 : \text{span}(B) \to \text{span}(B')$ defined by

$$T_0 \left( \sum_{j \in F} \alpha_j e_j \right) = \sum_{j \in F} \alpha_j f_j$$

(where $F$ is an arbitrary finite subset of $J$) is a well-defined linear map since $B$ is linearly independent. $T_0$ is an isometry by XV.9.5.14.. Thus $T_0$ extends uniquely to an isometric linear isomorphism $T$ from $\mathcal{H}$ onto $\mathcal{H}'$. The extension satisfies

$$T \left( \sum_{j \in J} \alpha_j e_j \right) = \sum_{j \in J} \alpha_j f_j$$

by continuity (cf. XV.9.5.17.).

Although the AC is not used in this proof, this theorem depends on the AC since we must know that $\mathcal{H}$ and $\mathcal{H}'$ have orthonormal bases. But if they are particular Hilbert spaces for which we know the existence of orthonormal bases of the same cardinality without using the AC (e.g. XV.9.5.29.), then the isomorphism also does not depend on the AC.
XV.9.5.26. For each cardinal \( \kappa \), there is a Hilbert space whose orthogonal dimension is \( \kappa \), namely \( \ell^2(X) \) for a set \( X \) of cardinality \( \kappa \) (indicator functions of singleton subsets of \( X \) form an orthonormal basis of cardinality \( \kappa \)). Thus (assuming AC) there is a precise classification of Hilbert spaces up to isometric isomorphism: there is exactly one for each cardinal. (There is, of course, a separate classification for real Hilbert spaces and complex Hilbert spaces.) This classification is analogous to, but not exactly the same as, the classification up to isomorphism of vector spaces over a fixed field by their dimension.

Separable Hilbert spaces

XV.9.5.27. Proposition. Let \( \mathcal{H} \) be a Hilbert space. Then \( \mathcal{H} \) is separable if and only if its orthogonal dimension is countable.

Proof: If \( \mathcal{H} \) has a countable orthonormal basis, then it is the closed linear span of a countable set and thus separable. Conversely, any orthonormal basis for a Hilbert space is discrete in the relative topology (the distance between any two orthogonal unit vectors is \( \sqrt{2} \)), and a separable metric space cannot have an uncountable discrete subset.

This proof depends on the existence of orthonormal bases and thus on the AC. But we will give a constructive proof that every separable Hilbert space has a (countable) orthonormal basis without using even the countable AC; thus this result does not depend on any form of Choice.

XV.9.5.28. Suppose \( \mathcal{H} \) is a separable Hilbert space. Fix a countable dense (or just total) set in \( \mathcal{H} \) and arrange it into a sequence \( (x_n) \). We will describe an explicit inductive procedure to manufacture an orthonormal basis out of the \( x_n \), called the Gram-Schmidt orthogonalization procedure.

To avoid annoying notational problems, we will first go through the sequence \( (x_n) \) to discard any \( x_n \) which is a linear combination of the previous ones. Reindexing, we may thus assume the \( x_n \) are linearly independent.

In particular, \( x_1 \neq 0 \). Set \( e_1 = \frac{x_1}{\|x_1\|} \). Then \( e_1 \) is a unit vector spanning the same one-dimensional subspace as \( x_1 \).

Now set \( y_2 = x_2 - \langle x_2, e_1 \rangle e_1 \). Then \( y_2 \neq 0 \) since \( \{x_1, x_2\} \) is linearly independent and hence \( \{e_1, x_2\} \) is linearly independent, and
\[
\langle y_2, e_1 \rangle = \langle x_2, e_1 \rangle - \langle x_2, e_1 \rangle \langle e_1, e_1 \rangle = 0
\]
So \( y_2 \perp e_1 \). Set \( e_2 = \frac{y_2}{\|y_2\|} \). Then \( e_2 \) is a unit vector orthogonal to \( e_1 \), and span\( (\{e_1, e_2\}) = \text{span}(\{x_1, x_2\}) \).

Continue in this manner, i.e.
\[
y_3 = x_3 - \langle x_3, e_1 \rangle e_1 - \langle x_3, e_2 \rangle e_2
\]
is a nonzero vector orthogonal to \( e_1 \) and \( e_2 \) (it is just the component of \( x_3 \) in the orthogonal complement of \( \text{span}(\{x_1, x_2\}) \) in \( \text{span}(\{x_1, x_2, x_3\}) \)), and if \( e_3 = \frac{y_3}{\|y_3\|} \), then \( \{e_1, e_2, e_3\} \) is an orthonormal set with the same span as \( \{x_1, x_2, x_3\} \). In this way we generate an orthonormal sequence \( (e_n) \), with
\[
\text{span}(\{e_1, \ldots, e_n\}) = \text{span}(\{x_1, \ldots, x_n\})
\]
for all \( n \). Since the sequence \( (x_n) \) is total in \( \mathcal{H} \), so is the sequence \( (e_n) \), i.e. \( \{e_n : n \in \mathbb{N}\} \) is an orthonormal basis for \( \mathcal{H} \).

The result is important enough that we state it as a theorem:

1797
XV.9.5.29. **Theorem.** [Gram-Schmidt] Let $H$ be a separable Hilbert space, and $S$ a countable total subset of $H$. Then there is a countable orthonormal basis $B$ for $H$ with $\text{span}(B) = \text{span}(S)$. The orthonormal basis $B$ is constructed from $S$ by an explicit algorithm.

XV.9.5.30. Several things to note about this process:

(i) The orthonormal basis obtained depends not only on the $x_n$, but on the order in which they are taken. Even in a two-dimensional Hilbert space the orthonormal basis is quite different in general if the order of the original vectors is changed. The reader unfamiliar with the Gram-Schmidt procedure should try some two- and three-dimensional examples to understand the process.

(ii) Nowhere in the procedure did we use that $H$ is complete. Thus the process works equally well in any separable inner product space, i.e. any dense subspace of a separable Hilbert space. The conclusion is that every dense subspace of a separable Hilbert space $H$ contains an orthonormal basis for $H$. This is useful when $H$ contains a dense subspace of nice concrete elements; cf. XV.9.5.33.

(iii) No form of Choice is used in the Gram-Schmidt process.

We thus obtain the following result without any form of AC:

XV.9.5.31. **Corollary.** Two separable Hilbert spaces over $\mathbb{F}$ are isometrically isomorphic (over $\mathbb{F}$) if and only if they have the same orthogonal dimension (over $\mathbb{F}$), which is in $\mathbb{N}_0 \cup \{ \aleph_0 \}$. Any separable infinite-dimensional Hilbert space (over $\mathbb{F}$) is isometrically isomorphic (over $\mathbb{F}$) to $\ell^2$; hence any two separable infinite-dimensional Hilbert spaces (over $\mathbb{F}$) are isometrically isomorphic (over $\mathbb{F}$).

XV.9.5.32. Topologists generally use the term “Hilbert space” to denote the unique separable infinite-dimensional Hilbert space (i.e. $\ell^2$), regarded either as a metric space or as a topological space (where it is homeomorphic to $\mathbb{R}^\infty$ with the product topology ($\_\_\_\_\_$), although there is no linear homeomorphism). Thus to a topologist there is only one “Hilbert space” (up to isometry), but to an analyst there are many.

XV.9.5.33. **Example.** [Legendre Polynomials] Let $H = L^2([-1, 1])$ (we can take this simply to be the completion of $C([-1, 1])$, $PC([-1, 1])$, or $R([-1, 1])$; cf. ($\_\_\_\_$), and we can work over either $\mathbb{R}$ or $\mathbb{C}$). The algebra $P$ of polynomials is uniformly dense in $C([-1, 1])$ by the Stone-Weierstrass Theorem ($\_\_\_\_$); since the uniform topology is stronger than the topology defined by the standard inner product

$$
\langle f, g \rangle = \int_{-1}^{1} f(t)g(t) \, dt
$$

we have that $P$ is dense in $L^2([-1, 1])$. Thus, if we take $f_n(t) = t^n$, then the sequence $(f_n)$ ($n \geq 0$) is total in $L^2([-1, 1])$.

We make an orthonormal basis for $L^2([-1, 1])$ by the Gram-Schmidt procedure. Since $\|f_0\| = \sqrt{2}$, $e_0$ is the constant function $\frac{1}{\sqrt{2}}$. We have that

$$
g_1 := f_1 - \langle f_1, e_0 \rangle e_0 = t - \frac{1}{\sqrt{2}} \int_{-1}^{1} t \frac{1}{\sqrt{2}} \, dt = f_1
$$

1798
(actually $f_1 \perp e_0$, so $e_1$ is just a unit vector in the direction of $f_1$). Since

$$
\|g_1\| = \|f_1\| = \left[ \int_{-1}^{1} |t|^2 \, dt \right]^{1/2} = \sqrt{\frac{2}{3}}
$$

we have that $e_1(t) = \sqrt{\frac{2}{3}} t$.

Continuing in this manner, we obtain

$$
e_n(t) = \sqrt{\frac{2n+1}{2}} p_n(t)
$$

where

$$
p_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} [(t^2 - 1)^n]
$$

is the $n$'th Legendre polynomial (this expression is called Rodrigues’s formula). The constants are chosen so that $p_n(1) = 1$ for all $n$. The polynomial $p_n$ is the solution to the Legendre differential equation

$$(1-t^2)y'' - 2ty' + n(n+1)y = 0$$

with $y(1) = 1$. See e.g. [?, 3.7] for details, and for more examples of orthonormal bases of function spaces constructed by the Gram-Schmidt procedure.

Note that, as expected, we obtain an orthonormal basis for $L^2([-1, 1])$ consisting of polynomials.

**XV.9.5.34.** The statement (XV.9.5.30.(ii)) that a dense subspace of a Hilbert space $H$ contains an orthonormal basis for $H$ can be false if $H$ is nonseparable. In fact, a Hilbert space of orthogonal dimension $2^{\aleph_0}$ has a dense subspace which does not contain any uncountable orthonormal set [Dix53].

**A Crinkled Curve**

“There’s a lot of room in Hilbert space.”

*Traditional*

**XV.9.5.35.** In an infinite-dimensional Hilbert space, even a separable one, there are continuous arcs which are “crinkled” in a way impossible in a finite-dimensional Euclidean space: curves which make a right-angled turn at every point. Such a curve cannot be differentiable anywhere in any reasonable sense.

**XV.9.5.36.** Example. In $L^2([0, 1])$ (we can avoid involving measure theory by working in the dense subspace $PC([0, 1])$, cf. ()), for $0 \leq t \leq 1$ let $f_t = \chi_{[0,t]}$, the indicator function of the subinterval $[0, t]$. It is easy to check that $t \mapsto f_t$ is a one-to-one continuous function from $[0, 1]$ into $L^2([0, 1])$. If $0 \leq a < b < c \leq 1$, then the chord $f_b - f_a$ from $f_a$ to $f_b$ is orthogonal to the chord $f_c - f_b$ from $f_b$ to $f_c$. Thus any two consecutive chords (or indeed any two nonoverlapping chords) are orthogonal!
XV.9.6. Projections

The theory of subspaces and orthogonal complements in a Hilbert space has a nice alternate description in terms of orthogonal projection operators (projections).

XV.9.6.1. Definition. Let $H$ be a Hilbert space, and $Y$ a closed subspace. For any vector $x \in H$, write $x = y + z$ with $y \in Y$ and $z \in Y^\perp$, and define $P_Y(x)$ to be $y$. Then $P_Y$ is a function from $H$ to $H$, with range $Y$. $P_Y$ is called the orthogonal projection operator, or just projection, of $H$ onto $Y$.

Note that $P_Y(x)$ is well defined by the uniqueness of the decomposition $x = y + z$ (XV.9.5.7.). We have $P_{\{0\}} = 0$ and $P_H = I$.

XV.9.6.2. Proposition. Let $H$ be a Hilbert space, and $Y$ a closed subspace. Then $P_Y$ is a bounded linear operator on $H$ with operator norm $\|P_Y\| = 1$ (unless $Y = \{0\}$), and $N(P_Y) = Y^\perp$.

Proof: If $x_1, x_2 \in H$, and $x_1 = y_1 + z_1$, $x_2 = y_2 + z_2$ with $y_j \in Y$, $z_j \in Y^\perp$, then $x_1 + x_2 = (y_1 + y_2) + (z_1 + z_2)$ must be the unique decomposition for $x_1 + x_2$, and thus

$$P_Y(x_1 + x_2) = y_1 + y_2 = P_Y(x_1) + P_Y(x_2).$$

Similarly, if $x \in H$ and $\alpha \in \mathbb{F}$, and $x = y + z$ is the decomposition of $x$, then $\alpha x = \alpha y + \alpha z$ is the decomposition of $\alpha x$, and thus

$$P_Y(\alpha x) = \alpha y = \alpha P_Y(x).$$

Thus $P_Y$ is a linear operator. If $x \in H$, $x = y + z$, then by XV.9.5.5.

$$\|P_Y(x)\|^2 = \|y\|^2 \leq \|y\|^2 + \|z\|^2 = \|x\|^2$$

so $P_Y$ is a bounded operator and $\|P_Y\| \leq 1$. But if $Y \neq \{0\}$, and $y$ is a nonzero vector in $Y$, then $P_Y(y) = y$, so $\|P_Y\| = 1$.

It is obvious that $P_Y(x) = 0$ if and only if $x \in Y^\perp$. \(\diamondsuit\)

XV.9.6.3. Orthogonal projection operators are the simplest kinds of non-scalar bounded operators on a Hilbert space. They are the “building blocks” for general self-adjoint operators via the Spectral Theorem ().

XV.9.6.4. Proposition. Let $H$ be a Hilbert space, and $Y$ a closed subspace. Then $P_Y$ is an idempotent operator on $H$, i.e. $P_Y^2 = P_Y$. Every idempotent bounded operator $P$ on $H$ with $\|P\| \leq 1$ is an orthogonal projection operator, i.e. $P = P_X$ for some closed subspace $X$ of $H$.

Proof: If $x \in H$, $x = y + z$, then

$$P_Y^2(x) = P_Y(P_Y(x)) = P_Y(y) = y = P_Y(x).$$

1800
Conversely, if $P^2 = P$, let $\mathcal{X}$ be the range of $P$. If $x \in \mathcal{X}$, then $x = P(v)$ for some $v \in \mathcal{H}$, so $P(x) = P^2(v) = P(v) = x$. If $w = v - x$, we then have $w \in \mathcal{N}(P)$ and $v = x + w$; thus, if $v$ is any vector in $\mathcal{H}$, we can write $v = P(v) + w$ with $w \in \mathcal{N}(P)$. If $x \in \mathcal{X}$ and $w \in \mathcal{N}(P)$, then, for any $v \in \mathbb{F}$,

$$\|x\| = \|P(x + \alpha w)\| \leq \|x + \alpha w\|$$

if $\|P\| \leq 1$; thus $x \perp w$ by XV.9.5.6, i.e. $\mathcal{N}(P) \subseteq \mathcal{X}^\perp$. But if $v \in \mathcal{X}^\perp$, we can write $v = P(v) + w$ with $w \in \mathcal{N}(P)$. Since $P(v) \in \mathcal{X}$ and $w \in \mathcal{X}^\perp$, we must have $v = w$ by XV.9.5.7, i.e. $v \in \mathcal{N}(P)$, $\mathcal{N}(P) = \mathcal{X}^\perp$. Thus, if $v \in \mathcal{H}$, and $v = x + w$ for $x \in \mathcal{X}$, $w \in \mathcal{X}^\perp$, we have $P(v) = x = P_X(v)$.

Thus the projections have a simple algebraic characterization: they are exactly the idempotent bounded operators on $\mathcal{H}$ of norm $\leq 1$.

Here are some other elementary properties of projections:

**XV.9.6.5.** PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace. Then $P_{\mathcal{Y}^\perp} = I - P_{\mathcal{Y}}$.

**XV.9.6.6.** The projection $I - P$ is often written $P^\perp$. Thus $P_{\mathcal{Y}^\perp} = P_Y^\perp$. We say two projections $P$ and $Q$ are orthogonal, written $P \perp Q$, if $PQ = 0$. We have $P_X \perp P_Y$ if and only if $\mathcal{Y} \subseteq \mathcal{X}^\perp$. Also, if $P \perp Q$, then $P + Q$ is a projection (and conversely). We have $P \perp P^\perp$ for any $P$.

**XV.9.6.7.** PROPOSITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be closed subspaces of $\mathcal{H}$. The following are equivalent:

(i) $\mathcal{X} \subseteq \mathcal{Y}$.

(ii) $P_X P_Y = P_Y P_X = P_X$.

(iii) $P_Y - P_X$ is a projection (it is $P_{Y \cap \mathcal{X}^\perp}$).

**XV.9.6.8.** This result suggests that we can put a reasonable partial order on the projections on $\mathcal{H}$ by setting $P \leq Q$ if $PQ = QP = P$. Thus $P_X \leq P_Y$ if and only if $\mathcal{X} \subseteq \mathcal{Y}$, i.e. $P \leq Q$ if and only if $\mathcal{R}(P) \subseteq \mathcal{R}(Q)$. There is a “largest” projection $I$ and a “smallest” projection $0$ under this partial order. Any collection $\{P_j\}$ of projections has an infimum $P = \bigwedge P_j$ and a supremum $Q = \bigvee P_j$: if $P_j = P_{X_j}$, then $P = P_X$, where $\mathcal{X} = \bigcap \mathcal{X}_j$, and $Q = P_Y$, where $\mathcal{Y}$ is the closed linear span of $\bigcup \mathcal{X}_j$. In particular, if $P$ and $Q$ are projections, with $P = P_X$ and $Q = P_Y$, then $P \wedge Q$ is $P_{\mathcal{X} \cap \mathcal{Y}}$ and $P \vee Q$ is $P_{(\mathcal{X} + \mathcal{Y})^\perp}$. Thus the projections on $\mathcal{H}$ form a complete lattice $\bigvee$.

Using the complement $P \mapsto P^\perp = I - P$, the projections on $\mathcal{H}$ form a complete Boolean algebra.

**XV.9.6.9.** Caution: Two projections $P$ and $Q$ do not commute (i.e. $PQ \neq QP$) in general: for example, consider projections from $\mathbb{R}^2$ onto two one-dimensional subspaces which are distinct but not orthogonal. $P_X$ and $P_Y$ commute if and only if the subspaces $\mathcal{X}$ and $\mathcal{Y}$ are “nicely situated” with respect to each other. If $PQ = QP$, we have $P \wedge Q = PQ$ and $P \vee Q = P + Q - PQ$. In general, there are no simple algebraic formulas for $P \wedge Q$ or $P \vee Q$. If $P \perp Q$, i.e. if $PQ = 0$, then also $QP = 0$ and $P \wedge Q = 0$, $P \vee Q = P + Q$.

1801
XV.9.7. Dual Spaces and Weak Topology

Bounded linear functionals on a Hilbert space are easy to describe:

XV.9.7.1. Theorem. [Riesz-Fréchet] Let $H$ be a Hilbert space, and $\phi$ a bounded linear functional on $H$. Then there is a (unique) vector $\xi \in H$ with $\phi(\eta) = \langle \eta, \xi \rangle$ for all $\eta \in H$; and $\|\xi\| = \|\phi\|$.

So the dual space $H^*$ of $H$ may be identified with $H$ (the identification is conjugate-linear), and the weak (= weak-* topology is given by the inner product: $\xi_i \to \xi$ weakly if $\langle \xi_i, \eta \rangle \to \langle \xi, \eta \rangle$ for all $\eta$. In particular, a Hilbert space is a reflexive Banach space and the unit ball is therefore weakly compact. A useful way of stating this is:

XV.9.7.2. Corollary. Let $H$ be a Hilbert space. If $(\xi_i)$ is a bounded weak Cauchy net in $H$ (i.e. $(\langle \xi_i, \eta \rangle)$ is Cauchy for all $\eta \in H$), then there is a (unique) vector $\xi \in H$ such that $\xi_i \to \xi$ weakly, i.e. $\langle \xi_i, \eta \rangle \to \langle \xi, \eta \rangle$ for all $\eta \in H$.

The weak topology is, of course, distinct from the norm topology (strictly weaker) if $H$ is infinite-dimensional: for example, an orthonormal sequence of vectors converges weakly to 0. The next proposition gives a connection.

XV.9.7.3. Proposition. Let $H$ be a Hilbert space, $\xi_i, \xi \in H$, with $\xi_i \to \xi$ weakly. Then $\|\xi\| \leq \lim \inf \|\xi_i\|$, and $\xi_i \to \xi$ in norm if and only if $\|\xi_i\| \to \|\xi\|$.

Proof: $\|\xi\|^2 = \lim \langle \xi_i, \xi \rangle \leq \|\xi\| \lim \inf \|\xi_i\|$ by the CBS inequality.

$$\|\xi_i - \xi\|^2 = \langle \xi_i, \xi_i \rangle - \langle \xi_i, \xi \rangle - \langle \xi, \xi_i \rangle + \langle \xi, \xi \rangle$$

which goes to zero if and only if $\|\xi_i\| \to \|\xi\|$.

XV.9.7.4. If $H$ is infinite-dimensional, then the weak topology on $H$ is not first countable, and a weakly convergent net need not be (norm-)bounded. It is an easy consequence of Uniform Boundedness (XV.6.4.1.) that a weakly convergent sequence is bounded. For example, if $\{\xi_n\}$ is an orthonormal sequence in $H$, then 0 is in the weak closure of $\{\sqrt{n}\xi_n\}$, but no subsequence converges weakly to 0 (cf. [Hal82, Problem 28], [?]). If $H$ is separable, then the restriction of the weak topology to the unit ball of $H$, and hence to any bounded set, is metrizable [Hal82, Problem 24].

XV.9.8. Standard Constructions

There are two standard constructions on Hilbert spaces which are used repeatedly, direct sum and tensor product.
XV.9.8.1. If \( \{ \mathcal{H}_i : i \in \Omega \} \) is a set of Hilbert spaces, we can form the Hilbert space direct sum, denoted \( \bigoplus_{\Omega} \mathcal{H}_i \), as the set of “sequences” (indexed by \( \Omega \)) \( (\cdots \xi_i, \cdots) \), where \( \xi_i \in \mathcal{H}_i \) and \( \sum_i \| \xi_i \|^2 < \infty \) (so \( \xi_i \neq 0 \) for only countably many \( i \)). The inner product is given by

\[
\langle (\cdots \xi_i, \cdots), (\cdots \eta_i, \cdots) \rangle = \sum_i \langle \xi_i, \eta_i \rangle.
\]

This is the completion of the algebraic direct sum with respect to this inner product. If all \( \mathcal{H}_i \) are the same \( \mathcal{H} \), the direct sum is called the amplification of \( \mathcal{H} \) by \( card(\Omega) \). The amplification of \( \mathcal{H} \) by \( n \) (the direct sum of \( n \) copies of \( \mathcal{H} \) is denoted by \( \mathcal{H}^n \); the direct sum of countably many copies of \( \mathcal{H} \) is denoted \( \mathcal{H}^\infty \) (or sometimes \( l^2(\mathcal{H}) \)).

XV.9.8.2. If \( \mathcal{H}_1, \mathcal{H}_2 \) are Hilbert spaces, let \( \mathcal{H}_1 \odot \mathcal{H}_2 \) be the algebraic tensor product over \( \mathbb{C} \). Put an inner product on \( \mathcal{H}_1 \odot \mathcal{H}_2 \) by

\[
\langle (\xi_1 \otimes \xi_2, \eta_1 \otimes \eta_2) = \langle \xi_1, \eta_1 \rangle \langle \xi_2, \eta_2 \rangle
\]

extended by linearity, where \( \langle \cdot, \cdot \rangle \) is the inner product on \( \mathcal{H}_i \). (It is easily checked that this is an inner product, in particular that it is positive definite.) Then the Hilbert space tensor product \( \mathcal{H}_1 \otimes \mathcal{H}_2 \) is the completion of \( \mathcal{H}_1 \odot \mathcal{H}_2 \). Tensor products of finitely many Hilbert spaces can be defined similarly. (Infinite tensor products are trickier and will be discussed in ??.)

If \( \mathcal{H} \) and \( \mathcal{H}' \) are Hilbert spaces and \( \{ \eta_i : i \in \Omega \} \) is an orthonormal basis for \( \mathcal{H}' \), then there is an isometric isomorphism from \( \mathcal{H} \otimes \mathcal{H}' \) to the amplification of \( \mathcal{H} \) by \( card(\Omega) \), given by \( \sum_i \xi_i \otimes \eta_i \mapsto (\cdots \xi_i, \cdots) \). In particular, \( \mathcal{H}^n \) is naturally isomorphic to \( \mathcal{H} \otimes \mathbb{C}^n \) and \( \mathcal{H}^\infty \cong \mathcal{H} \otimes l^2 \).

If \( (X, \mu) \) is a measure space and \( \mathcal{H} \) is a separable Hilbert space, then the Hilbert space \( L^2(X, \mu) \otimes \mathcal{H} \) is naturally isomorphic to \( L^2(X, \mu, \mathcal{H}) \), the set of weakly measurable functions \( f : X \to \mathcal{H} \) (i.e. \( x \mapsto \langle f(x), \eta \rangle \) is a complex-valued measurable function for all \( \eta \in \mathcal{H} \)) such that \( \int_X \| f(x) \|^2 \, d\mu(x) < \infty \), with inner product

\[
\langle f, g \rangle = \int_X (\langle f(x), g(x) \rangle) \, d\mu(x).
\]

The isomorphism sends \( f \otimes \eta \) to \( (x \mapsto \langle f(x), \eta \rangle) \).

Unless otherwise qualified, a “direct sum” or “tensor product” of Hilbert spaces will always mean the Hilbert space direct sum or tensor product.

XV.9.9. Real Hilbert Spaces

One can also consider real inner product spaces, real vector spaces with a (bilinear, real-valued) inner product \( \langle \cdot, \cdot \rangle \) satisfying the properties of ?? for \( \alpha \in \mathbb{R} \). A real Hilbert space is a complete real inner product space. A real Hilbert space \( \mathcal{H}_R \) can be complexified to \( \mathbb{C} \otimes_{\mathbb{R}} \mathcal{H}_R \); conversely, a (complex) Hilbert space can be regarded as a real Hilbert space by restricting scalar multiplication and using the real inner product \( \langle \xi, \eta \rangle = \text{Re} \langle \xi, \eta \rangle \). (Note, however, that these two processes are not quite inverse to each other.) All of the results of this section, and many (but by no means all) results about operators, hold verbatim or have obvious exact analogs in the case of real Hilbert spaces.

Throughout this volume, the term “Hilbert space,” unless qualified with “real,” will always denote a complex Hilbert space; and “linear” will mean “complex-linear.”

1803
XV.9.10. Exercises

XV.9.10.1. Prove the CBS inequality for $\mathbb{R}^n$ as follows. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$.

(a) Show that
\[
\left( \sum_{k=1}^{n} x_k y_k \right)^2 + \sum_{j<k} (x_j y_k - x_k y_j)^2 = \left( \sum_{k=1}^{n} x_k^2 \right) \left( \sum_{k=1}^{n} y_k^2 \right).
\]

(b) Use that all the terms in the second sum are nonnegative to deduce the CBS inequality.

(c) Equality holds in the CBS inequality if and only if all terms in the second sum are zero. Show that this implies that $\{x, y\}$ is linearly dependent.

This was Cauchy’s proof of the CBS inequality from [Can21].

XV.9.10.2. Give another proof of the CBS inequality for $\mathbb{R}^n$ as follows.

(a) Show that for $x, y \in \mathbb{R}$, $|xy| \leq \frac{1}{2}(x^2 + y^2)$, with equality if and only if $y = \pm x$. [Use that $(x + y)^2$ and $(x - y)^2$ are nonnegative.]

(b) Show that it suffices to prove the CBS inequality for $x, y$ of norm 1.

(c) if $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ have norm 1, then
\[
\left| \sum_{k=1}^{n} x_k y_k \right| \leq \sum_{k=1}^{n} |x_k y_k| \leq \frac{1}{2} \left( \sum_{k=1}^{n} x_k^2 + \sum_{k=1}^{n} y_k^2 \right) = 1
\]
with equality if and only if $y = \pm x$.

XV.9.10.3. Let $f$ and $g$ be real-valued continuous functions on $[a, b]$. Use Riemann sums for $\int_{a}^{b} f(t)g(t) \, dt$, $\int_{a}^{b} [f(t)]^2 \, dt$, and $\int_{a}^{b} [g(t)]^2 \, dt$, and the CBS inequality for $\mathbb{R}^n$, to prove
\[
\left( \int_{a}^{b} f(t)g(t) \, dt \right)^2 \leq \left( \int_{a}^{b} [f(t)]^2 \, dt \right) \left( \int_{a}^{b} [g(t)]^2 \, dt \right)
\]
with equality if and only if $\{f, g\}$ is linearly dependent. This was Bunyakovskii’s inequality and proof from [Bun59].

XV.9.10.4. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be a real inner product space, and $x, y \in \mathcal{V}$. 

(a) Write $\langle x + \alpha y, x + \alpha y \rangle$ as a quadratic function of $\alpha$.

(b) Use the quadratic formula and the fact that $\langle x + \alpha y, x + \alpha y \rangle \geq 0$ for all $\alpha$ to deduce the CBS inequality.

This was essentially Schwarz’s proof from [Sch88] (he only explicitly considered the case where the inner product was given by a (multiple) integral as in XV.9.1.4.(ii).)

XV.9.10.5. Deduce the CBS inequality for complex inner products from the inequality for real inner products. [Consider the real inner product $\langle x, y \rangle = \Re(\langle x, y \rangle)$ and suitable scalar multiples.]

XV.9.10.6. [Ste04] Prove the CBS inequality for the standard inner product on $\mathbb{R}^n$ by induction, starting with the case $n = 2$ (it is trivial for $n = 1$).
XV.9.10.7. Justify the name “Parallelogram Law” by giving a geometric interpretation in terms of the lengths of the diagonals of a parallelogram. If \( x, y \in \mathbb{R}^2 \), consider the parallelogram with vertices \( 0, x, y, x + y \). See VI.3.1.22.

XV.9.10.8. Let \( \mathcal{V} \) be a normed vector space over \( \mathbb{F} = \mathbb{R} \) or \( \mathbb{C} \). Show that the norm on \( \mathcal{V} \) comes from an inner product if and only if every two-dimensional subspace of \( \mathcal{V} \) is isometrically isomorphic to \( \mathbb{F}^2 \) with the induced norm from the standard inner product.

XV.9.10.9. Suppose \((\mathcal{V}, \langle \cdot, \cdot \rangle)\) is an inner product space. Expand the right side of the Polarization Identity (XV.9.2.8) to show that the identity is satisfied.

XV.9.10.10. (a) Show that the Parallelogram Law is equivalent to the statement that
\[
\|x + y\|^2 - \|x - y\|^2 = 2(\|x + y\|^2 - \|x\|^2 - \|y\|^2)
\]
for all \( x, y \).
(b) Show that in an inner product space,
\[
\langle x, y \rangle = \frac{1}{2}[(\|x + y\|^2 - \|x\|^2 - \|y\|^2) + i(\|x + iy\|^2 - \|x\|^2 - \|y\|^2)]
\]
for all \( x, y \). Give a similar formula for real inner products.

XV.9.10.11. (a) In \( C([0, 1]) \), let \( f_n (n \geq 2) \) be defined by
\[
f_n(t) = \begin{cases} 
0 & \text{if} \ 0 \leq t \leq \frac{1}{2} \\
\text{linear} & \text{if} \ \frac{1}{2} \leq t \leq \frac{1}{2} + \frac{1}{n} \\
1 & \text{if} \ \frac{1}{2} \frac{1}{n} \leq t \leq 1
\end{cases}
\]
Show that \((f_n)\) is a uniformly bounded sequence in \( C([0, 1]) \) which is a Cauchy sequence with respect to the norm induced by the inner product, but which does not converge in \( C([0, 1]) \) in the induced norm. (Note that we have referred here to the uniform norm \( \| \cdot \|_\infty \) as well as the norm induced by the inner product; in this problem and succeeding problems we discuss Cauchy sequences and convergence only with respect to the norm induced by the inner product. The sequence \((f_n)\) is bounded but not Cauchy in the uniform norm.)
(b) It is slightly harder to show that \( PC([0, 1]) \) is not complete. Let \( g_n (n \geq 2) \) be defined by
\[
g_n(t) = \begin{cases} 
0 & \text{if} \ 0 \leq t \leq \frac{1}{n} \\
\sin \left( \frac{\pi}{t} \right) & \text{if} \ \frac{1}{n} \leq t \leq 1
\end{cases}
\]
Show that \((g_n)\) defines a uniformly bounded Cauchy sequence in \( C([0, 1]) \) which does not converge in \( PC([0, 1]) \).
(c) It is easier to give an example of a nonconvergent Cauchy sequence which is not uniformly bounded. Let
\[
h_n(t) = \min \left( n, \frac{1}{\sqrt{n}} \right)
\]
for \( 0 \leq t \leq 1 \). Then \((h_n)\) is a Cauchy sequence in \( C([0, 1]) \) which does not converge even in \( PC([0, 1]) \).
XV.9.10.12. Let \( \mathcal{R}([a, b]) \) be the set of complex-valued Riemann-integrable functions on \([a, b]\).

(a) Show that the formula

\[
\langle f, g \rangle = \int_a^b f(t)g(t) \, dt
\]

(where the integral is a Riemann integral) defines a pre-inner product on \( \mathcal{R}([a, b]) \).

(b) Explicitly describe the space \( \mathcal{N} \).

(c) Let \( R([a, b]) = \mathcal{R}([a, b]) / \mathcal{N} \). Show that \( PC([a, b]) \) can be naturally identified with a dense subspace of \( R([a, b]) \).

(d) The Cauchy sequence \((h_n)\) of XV.9.10.11.(c) does not converge in \( R([0, 1]) \).

(e) Find a uniformly bounded Cauchy sequence in \( \mathcal{R}([a, b]) \) that does not converge. [A Cantor set of positive measure () is helpful.]

(f) Try to describe the completion of \( R([a, b]) \) (which will also be the completion of \( C([a, b]) \) and \( PC([a, b]) \)) since these are dense in \( R([a, b]) \). The example in (d) shows that we must at least add all nonnegative improperly Riemann square-integrable functions on \([a, b]\), but (e) shows that even linear combinations of these do not give enough. The answer involves Lebesgue measure and integration (); this example gives good motivation for development of the Lebesgue theory.

XV.9.10.13. Let \((V, \langle \cdot, \cdot \rangle)\) be a (real or complex) inner product space. Let \( \mathcal{H} \) be the set of equivalence classes of Cauchy sequences in \( V \) as in the proof of (). Show that

\[
\langle [(x_n)], [(y_n)] \rangle = \lim_{n \to \infty} \langle x_n, y_n \rangle
\]

gives a well-defined inner product on \( \mathcal{H} \) inducing the norm defined in ()

XV.9.10.14. Let \( \mathcal{H} \) be an infinite-dimensional Hilbert space.

(a) Show that for every \( r > 0 \), an open ball of radius \( 3r \) in \( \mathcal{H} \) contains infinitely many pairwise disjoint open balls of radius \( r \). If \( \mathcal{H} \) is nonseparable, an open ball of radius \( 3r \) contains uncountably many pairwise disjoint open balls of radius \( r \). [Consider open balls centered at points of an orthonormal basis.] This is one justification for the statement “There’s a lot of room in Hilbert space.”

(b) Use (a) to show that there is no translation-invariant Borel measure on \( \mathcal{H} \) for which open balls have finite positive measure.

(c) Show that an open ball of radius \( 2r \) in a Hilbert space (finite- or infinite-dimensional) cannot contain more than two disjoint open balls of radius \( r \). [If there are three such balls, consider the two-dimensional real affine subspace spanned by the centers.]

In (a), 3 may be replaced by \( 1 + \sqrt{2} \approx 2.414 \) and in (c), 2 may be replaced by any number less than \( 1 + \frac{2}{\sqrt{3}} \approx 2.1547 \). What about numbers in between?

XV.9.10.15. Let \( \{X_i : i \in I\} \) be an indexed collection of Banach spaces. Define the \( \ell^2 \)-direct sum \( \bigoplus_{i \in I} X_i \) to be the subset of \( \prod_{i \in I} X_i \) consisting of all elements \(( \cdots, x_i, \cdots)\) for which \( \sum_{i \in I} \|x_i\|^2 < \infty \). (Note that \( \bigoplus_{i \in I} X_i = \prod_{i \in I} X_i \) if \( I \) is finite.)
(a) Let $\mathcal{X} = \bigoplus_{i \in I} \mathcal{X}_i$. Show that $\mathcal{X}$ is a vector subspace of $\prod_{i \in I} \mathcal{X}_i$ and that $\| \cdot \|$ defined on $\mathcal{X}$ by

$$
\|(\cdots x_i \cdots)\| = \left[ \sum_{i \in I} |x_i|^2 \right]^{1/2}
$$

is a norm on $\mathcal{X}$.

(b) Show that $(\mathcal{X}, \| \cdot \|)$ is complete, i.e. a Banach space.

(c) Show that if each $\mathcal{X}_i$ is a Hilbert space (i.e. the norm satisfies the Parallelogram Law), then $\mathcal{X}$ is also a Hilbert space.

**XV.9.10.16. Uniformly Convex Banach Spaces.** Let $\mathcal{X}$ be a Banach space. $\mathcal{X}$ is **strictly convex** if, whenever $x, y \in \mathcal{X}$, $x \neq y$, $\|x\| = \|y\| = 1$, we have $\|1/2 (x + y)\| < 1$. $\mathcal{X}$ is **uniformly convex** if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $x, y \in \mathcal{X}$, $\|x\| = \|y\| = 1$, $\|x - y\| > \epsilon$, we have $\|1/2 (x + y)\| < 1 - \delta$. Note that a uniformly convex Banach space is strictly convex.

(a) Show that $\mathcal{X}$ is strictly convex if and only if every element of $\mathcal{X}$ of norm 1 is an extreme point of the closed unit ball of $\mathcal{X}$.

(b) Show that any (closed) subspace of a strictly [resp. uniformly] convex Banach space is strictly [resp. uniformly] convex.

(c) Use the Parallelogram Law to show that any Hilbert space is uniformly convex.

(d) Show that $(\mathbb{R}^n, \| \cdot \|_p)$ is uniformly convex for $1 < p < \infty$, but that $(\mathbb{R}^n, \| \cdot \|_1)$ and $(\mathbb{R}^n, \| \cdot \|_\infty)$ are not strictly convex.

(e) Show that a strictly convex finite-dimensional Banach space is uniformly convex.

(f) Show that any $\ell^2$-direct sum (XV.9.10.15.) of strictly convex Banach spaces is strictly convex.

(g) Show that $\bigoplus_{p \in \mathbb{N}} (\mathbb{R}^2, \| \cdot \|_p)$ for $p \in \mathbb{N}, p \geq 2$, is strictly convex but not uniformly convex. Thus (f) is not true if “strictly” is replaced by “uniformly.” (But it is if the index set is finite.)

(h) Formulate and prove a precise statement to the effect that $\bigoplus_{p \in \mathbb{N}} \mathcal{X}_i$ is uniformly convex if and only if the $\mathcal{X}_i$ are “uniformly uniformly convex.” Deduce the last statement of (g).

It is known that a uniformly convex Banach space is always reflexive (Milman-Pettis Theorem). A strictly convex Banach space need not be reflexive; in fact, every separable Banach space has an equivalent strictly convex norm (cf. [https://math.stackexchange.com/questions/488790/is-there-a-non-reflexive-banach-space-whi](https://math.stackexchange.com/questions/488790/is-there-a-non-reflexive-banach-space-whi)).
XV.9.11. Sesquilinear Forms and Operators

XV.9.11.1. If $\mathcal{X}$ is a normed complex vector space, $\mathcal{H}$ a Hilbert space, and $T : \mathcal{X} \to \mathcal{H}$ a bounded operator, set

$$B(x, y) = \langle Tx, y \rangle.$$ 

Then it is easily checked that $B$ is a sesquilinear form on $\mathcal{X} \times \mathcal{H}$, which is bounded, i.e. there is a $K$ such that

$$|B(x, y)| \leq K \|x\| \|y\|$$

for all $x \in \mathcal{X}$, $y \in \mathcal{H}$. The smallest such $K$, the norm $\|B\|$ of $B$, is equal to $\|T\|$.

The following converse establishes a one-one correspondence between bounded operators and bounded sesquilinear forms. We state it for complex spaces; there is an essentially identical analogous version for real spaces, replacing “sesquilinear” with “bilinear.”

XV.9.11.2. Proposition. Let $\mathcal{X}$ be a normed vector space, $\mathcal{H}$ a Hilbert space, and $B$ a bounded sesquilinear form on $\mathcal{X} \times \mathcal{H}$. Then there is a unique bounded operator $T : \mathcal{X} \to \mathcal{H}$ such that

$$B(x, y) = \langle Tx, y \rangle$$

for all $x \in \mathcal{X}$, $y \in \mathcal{H}$. $\|T\| = \|B\|$.

Proof: Fix $x \in \mathcal{X}$. Define $\phi : \mathcal{H} \to \mathbb{C}$ by $\phi(y) = B(x, y)$. Then $\phi$ is a conjugate-linear functional on $\mathcal{H}$, and $\phi$ is bounded ($\|\phi\| \leq \|B\| \|x\|$). Thus by () there is a unique $z \in \mathcal{H}$ with

$$\phi(y) = B(x, y) = \langle z, y \rangle.$$ 

Set $Tx = z$. In this way a function $T : \mathcal{X} \to \mathcal{H}$ is defined. It is an easy consequence of the linearity of $B$ in the first variable that $T$ is linear. If $x \in \mathcal{X}$, we have

$$\|Tx\| = \sup_{\|y\|=1} |\langle Tx, y \rangle| = \sup_{\|y\|=1} |B(x, y)| \leq \|B\| \|x\|$$

and thus $T$ is bounded and $\|T\| \leq \|B\|$. By taking the supremum over all $x$ with $\|x\| = 1$, we obtain $\|T\| = \|B\|$. \(

XV.9.11.3. There is thus a one-one correspondence between bounded operators and bounded sesquilinear forms. Each point of view has its advantages. Some early authors such as HILBERT used the form approach exclusively instead of working with operators, but some parts of the theory become almost impossibly difficult and cumbersome (e.g. composition of operators) from this point of view. I. FREDHOLM and F. RIESZ emphasized the operator point of view, which is superior for most purposes.

The following special case of XV.9.11.2. is especially useful in the theory of partial differential equations:
**XV.9.11.4. Proposition.** Let $\mathcal{H}$ be a Hilbert space, and $B$ a bounded sesquilinear form on $\mathcal{H} \times \mathcal{H}$ for which there is an $\epsilon > 0$ such that

$$|B(x, x)| \geq \epsilon \|x\|^2$$

for all $x \in \mathcal{H}$. Then there is a unique $T \in \mathcal{B}(\mathcal{H})$ with

$$B(x, y) = \langle Tx, y \rangle$$

for all $x, y \in \mathcal{H}$. $T$ is invertible and $\|T\| = \|B\|$. Thus there is a unique $S \in \mathcal{B}(\mathcal{H})$, namely $S = T^{-1}$, such that

$$B(Sx, y) = \langle x, y \rangle$$

for all $x, y \in \mathcal{H}$. $S$ is invertible and $\|S\| \leq \frac{1}{\epsilon}$.

There is an analogous result for real Hilbert spaces, with “sesquilinear” replaced by “bilinear.”

**XV.9.11.5. Corollary.** Let $\mathcal{H}$ be a Hilbert space, and $B$ a bounded sesquilinear form on $\mathcal{H} \times \mathcal{H}$ for which there is an $\epsilon > 0$ such that

$$|B(x, x)| \geq \epsilon \|x\|^2$$

for all $x \in \mathcal{H}$. If $\phi$ is any bounded linear functional on $\mathcal{H}$, then there is a unique $z \in \mathcal{H}$ such that

$$\phi(x) = B(x, z)$$

for all $x \in \mathcal{H}$. Conversely, if $z$ is such a vector, then $\phi(x) = \langle x, T^* z \rangle$ for all $x$, so $T^* z = y, z = S' y$.

**Proof:** Let $T$ and $S = T^{-1}$ be as in XV.9.11.4. (so $(S^*)^{-1} = T^*$). There is a unique $y \in \mathcal{H}$ with $\phi(x) = \langle x, y \rangle$ for all $x \in \mathcal{H}$. Then $z = S' y$ is such a $z$:

$$B(x, z) = \langle Tx, z \rangle = \langle x, T^* z \rangle = \langle x, y \rangle = \phi(x)$$

for all $x$. Conversely, if $z$ is such a vector, then $\phi(x) = \langle x, T^* z \rangle$ for all $x$, so $T^* z = y, z = S' y$.

**XV.9.11.6.** For the proof of XV.9.11.4., everything follows immediately from XV.9.11.2. except invertibility of $T$, which is a consequence of the next elementary result, and uniqueness of $S$, which is immediate if $S'$ is another such operator, then for each $x \in \mathcal{H}$ we have $B(Sx, y) = B(S'x, y)$ and hence $B((S - S')x, y) = 0$ for all $y \in \mathcal{H}$, and in particular for $y = (S - S')x$, so

$$0 = |B((S - S')x, (S - S')x)| \geq \epsilon \|(S - S')x\|^2$$

and $(S - S')x = 0$.

**XV.9.11.7. Proposition.** Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$ with the property that there is an $\epsilon > 0$ such that $\langle Tx, x \rangle \geq \epsilon \|x\|^2$ for all $x \in \mathcal{H}$. Then $T$ is invertible. $\|T^{-1}\| \leq \frac{1}{\epsilon}$.

**Proof:** By the CBS inequality, $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{H}$, i.e. $T$ is bounded below. Thus $T$ is one-to-one with closed range ($\mathcal{R}(T)$). If $x \perp \mathcal{R}(T)$, then $\langle Tx, x \rangle = 0$, so $x = 0$, and $\mathcal{R}(T)$ is dense in $\mathcal{H}$. Thus $\mathcal{R}(T) = \mathcal{H}$. We have, for any $x \in \mathcal{H}$, $\|x\| = \|T(T^{-1}x)\| \geq \epsilon \|T^{-1}x\|$.

1809
XV.9.11.8. In many references, Proposition XV.9.11.4., or sometimes XV.9.11.5., is called the “Lax-Milgram Theorem” (or sometimes “the famous Lax-Milgram Theorem”) since it appeared in [LM54]. Although it is useful, it does not seem to merit named status, let alone fame, since

(1) It is such an easy and obvious consequence of () (although the proofs in [LM54] and in some other references such as [Yos80] are unnecessarily complicated).

(2) The more general result XV.9.11.2. was well known to Hilbert, Riesz, and other early workers in Hilbert space, long before Lax and Milgram came along.

Not everything published by prominent mathematicians deserves hagiography.
XV.10. Banach Algebras and Spectral Theory

Most of the advanced study of operators is connected with the spectrum, which is a generalization of the set of eigenvalues. While the most immediate applications of spectral theory are to operators, the theory itself is done most efficiently in the more general context of Banach algebras, and spectral theory has numerous applications beyond the case of operators. We will thus first set up the theory of Banach algebras and develop spectral theory there.

XV.10.1. Banach Algebras

A Banach algebra is a Banach space which has a multiplication with nice compatibility properties:

XV.10.1.1. Definition. Let \( F \) be \( \mathbb{R} \) or \( \mathbb{C} \). A normed algebra over \( F \) is an algebra \( \mathcal{A} \) over \( F \) with a norm satisfying
\[
\|xy\| \leq \|x\|\|y\|
\]
for all \( x, y \in \mathcal{A} \) (submultiplicative norm). If \( \mathcal{A} \) is a normed algebra which is a Banach space (i.e. is complete), \( \mathcal{A} \) is a Banach algebra.

We thus have real Banach algebras and complex Banach algebras. For technical reasons, we will usually consider complex Banach algebras, especially for spectral theory; thus we will make the convention that the term “Banach algebra” without the qualification “real” will denote a complex Banach algebra.

There are a great many important Banach algebras with varied properties. For example:

XV.10.1.2. Examples. (i) Let \( X \) be a Banach space (over \( F \)). Then \( B(X) \), with the usual operations and operator norm \( \| \cdot \| \), is a Banach algebra (over \( F \)).

(ii) Let \( X \) be a compact Hausdorff space, and let \( C(X) \) be the set of complex-valued continuous functions on \( X \). Give \( C(X) \) the usual operations of pointwise addition, pointwise multiplication, and pointwise scalar multiplication, and the supremum norm \( \| \cdot \| \). Then \( C(X) \) is a (complex) Banach algebra. Similarly, \( C_0(X) \) is a real Banach algebra. More generally, if \( X \) is locally compact, the set \( C_0(X) \) of complex-valued continuous functions vanishing at infinity \( \| \cdot \| \) is a Banach algebra with the same operations and norm.

(iii) As a special case of (i), \( M_n(F) = B(F^n) \) becomes a Banach algebra when given the operator norm from any norm on \( F^n \). Thus there are many norms on \( M_n(F) \) making it into a Banach algebra. As a special case, or as a special case of (ii) with \( X \) a one-point space, \( F \) is a Banach algebra over \( F \) with norm \( \| \alpha \| = | \alpha | \).

(iv) If \( X \) is any Banach space, define \( xy = 0 \) for all \( x, y \in X \). Then \( X \) is a Banach algebra. (This example is not very interesting, to be sure.)

These examples only scratch the surface of the possibilities, even the interesting possibilities.

Joint Continuity of Multiplication

In a Banach space (or just normed vector space), addition and scalar multiplication are jointly continuous. In a normed algebra, multiplication is also jointly continuous:
XV.10.1.3. Proposition. Let $A$ be a normed algebra. If $x_n, y_n, x, y \in A$ and $x_n \to x, y_n \to y$, then $x_n y_n \to xy$.

Proof: By the usual argument,

$$\|x_n y_n - xy\| = \|x_n y_n - x_n y + x_n y - xy\| \leq \|x_n y_n - x_n y\| + \|x_n y - xy\| \leq \|x_n\| \|y_n - y\| + \|y\| \|x_n - x\| .$$

Since $(x_n)$ converges, it is bounded.

Completion of a Normed Algebra

XV.10.1.4. Similarly, if $A$ is a normed algebra, then its completion $\overline{A}$ has a natural multiplication making it a Banach algebra. For if $(x_n)$ and $(y_n)$ are Cauchy sequences in $A$, then, for any $n$ and $m$,

$$\|x_n y_n - x_m y_m\| = \|x_n y_n - x_n y_m + x_n y_m - x_m y_m\| \leq \|x_n y_n - x_n y_m\| + \|x_n y_m - x_m y_m\| \leq \|x_n\| \|y_n - y_m\| + \|y_m\| \|x_n - x_m\| .$$

Since Cauchy sequences are bounded, it follows that $(x_n y_n)$ is a Cauchy sequence.

Unital Banach Algebras and Unitization

XV.10.1.5. Examples XV.10.1.2 (ii) and (iv) show that a Banach algebra need not have a multiplicative identity. A Banach algebra with a multiplicative identity is a unital Banach algebra. The multiplicative identity, if it exists, is generally denoted $1$. We usually assume $\|1\| = 1$; while this is not necessarily true in general, the norm can be replaced by an equivalent Banach algebra norm with this property (Exercise XV.10.6.1).

If a Banach algebra is not unital, a unit can be added in a systematic way:

XV.10.1.6. Proposition. Let $A$ be a nonunital Banach algebra over $\mathbb{F}$. Let $A^\dagger = A \oplus \mathbb{F}$. Define addition and scalar multiplication coordinatewise, and multiplication and a norm by

$$(x, \lambda) \cdot (y, \mu) = (xy + \lambda y + \mu x, \lambda \mu)$$

$$\|(x, \lambda)\| = \|x\| + |\lambda| .$$

Then $A^\dagger$ is a unital Banach algebra containing a copy of $A$ as an ideal of codimension 1, called the unitization of $A$.

The rule for multiplication can be remembered by thinking of $(x, \lambda)$ as $x + \lambda 1$. (The multiplicative identity is $(0,1)$.)

Proof: Checking the algebraic rules for an algebra is routine but tedious, and is left to the reader. It is clear that $\| \cdot \|$ is a norm on $A^\dagger$. For submultiplicativity, if $x, y \in A, \lambda, \mu \in \mathbb{F}$,

$$\|xy + \lambda y + \mu x\| + |\lambda \mu| \leq \|xy\| + \|\lambda y\| + \|\mu x\| + |\lambda \mu|$$

$$\leq \|x\| \|y\| + |\lambda| \|y\| + |\mu| \|x\| + |\lambda \mu| = (\|x\| + |\lambda|)(\|y\| + |\mu|) .$$

The map $x \mapsto (x, 0)$ identifies $A$ with a subalgebra (ideal) of $A^\dagger$. To see that $A^\dagger$ is complete, if $(a_n, \lambda_n)$ is a Cauchy sequence in $A^\dagger$, then $(a_n)$ is a Cauchy sequence in $A$ and $(\lambda_n)$ a Cauchy sequence in $\mathbb{F}$. These converge to some $a$ and $\lambda$, and $(a_n, \lambda_n) \to (a, \lambda)$.
Many aspects of Banach algebra theory are technically simpler for unital Banach algebras; allows the nonunital case to be largely reduced to the unital case. We do not want to just restrict to unital Banach algebras, however, since many naturally occurring Banach algebras are nonunital. Many, but not all, nonunital Banach algebras have an approximate unit (which can be used as a substitute).

Invertibility, Ideals, and Quotients

One of the remarkable and appealing aspects of the theory of Banach algebras is the intricate interplay between the topology and the algebraic structure. One of the most basic, and most important, examples of this is the use of power series in Banach algebras to construct inverses (and, later, other elements).

Recall that an element $x$ in a unital algebra $A$ is invertible if there is a $y \in A$ with $xy = yx = 1$. Caution: Since a Banach algebra is not necessarily commutative, for $x$ to be invertible it is not generally enough that there is a $y$ with $yx = 1$ (such an $x$ is left invertible), or that there is a $z$ with $xz = 1$ (such an $x$ is right invertible); cf. (1). However, if $x$ is both left and right invertible, the left and right inverses coincide and $x$ is invertible (and conversely): if $yx = xz = 1$, then $y = y(xz) = (yx)z = z$.

Theorem. Let $A$ be a unital Banach algebra, and $x \in A$. If the power series

$$\sum_{k=0}^{\infty} x^k$$

(where $x^0 = 1$ by convention) converges in $A$, the sum is a multiplicative inverse for $1 - x$. If $\|x\| < 1$, then the power series converges to a multiplicative inverse for $1 - x$, and

$$\|(1 - x)^{-1}\| \leq \frac{1}{1 - \|x\|}.$$ 

Proof: Suppose the series converges, and let $y$ be the sum. If

$$s_n = \sum_{k=0}^{n} x^k$$

is the partial sum to $n$, so that $s_n \to y$, we have by a simple calculation that

$$s_n(1 - x) = (1 - x)s_n = 1 - x^{n+1}.$$ 

The left side converges to $y(1 - x) = (1 - x)y$, and the right side to 1 since $x^{n+1} \to 0$. For the second statement, since $\|x^k\| \leq \|x\|^k$ for every $k$ by submultiplicativity, we have that

$$\sum_{k=1}^{\infty} \|x^k\|$$
converges by comparison to the geometric series with ratio \(|x|\), and hence \(\sum_{k=0}^{\infty} x^k\) converges. By the triangle inequality,
\[
\left\| \sum_{k=0}^{\infty} x^k \right\| \leq \sum_{k=0}^{\infty} \|x^k\| \leq \sum_{k=0}^{\infty} \|x\|^k = \frac{1}{1 - \|x\|}.
\]

\[\vspace{1cm}\]

**XV.10.2.10. Corollary.** Let \(A\) be a unital Banach algebra, and \(u \in A\). If \(\|u - 1\| < 1\), then \(u\) is invertible in \(A\), and
\[
\|u^{-1}\| \leq \frac{1}{1 - \|u - 1\|}.
\]

**Proof:** Apply the theorem with \(x = 1 - u\).

\[\vspace{1cm}\]

**XV.10.2.11. Corollary.** Let \(A\) be a unital Banach algebra, and \(u, v \in A\). If \(u\) is invertible in \(A\) and
\[
\|u - v\| < \frac{1}{\|u^{-1}\|}
\]
then \(v\) is also invertible in \(A\), and
\[
\|v^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}\||u - v|}.
\]
Thus the open ball around \(u\) of radius \(\frac{1}{\|u^{-1}\|}\) consists entirely of invertible elements. In particular, the set of invertible elements in \(A\) is an open set.

**Proof:** If \(\|u - v\| < \frac{1}{\|u^{-1}\|}\), then
\[
\|1 - u^{-1}v\| = \|u^{-1}(u - v)\| \leq \|u^{-1}\||u - v| < 1
\]
so \(u^{-1}v\) is invertible and \(v\) is left invertible. Similarly, \(\|1 - vu^{-1}\| < 1\) so \(v\) is right invertible. We have
\[
\|v^{-1}u\| = \|(u^{-1}v)^{-1}\| \leq \frac{1}{1 - \|u^{-1}v - 1\|}
\]
so
\[
\|v^{-1}\| = \|(v^{-1}u)^{-1}\| \leq \|v^{-1}u\||u^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}v - 1\|} = \frac{\|u^{-1}\|}{1 - \|u^{-1}(v - u)\|} \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}\||v - u|}.
\]

\[\vspace{1cm}\]

1814
XV.10.2.12. Note that the bound
\[ \|v^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}v - 1\|} \]
in the proof is generally better than the bound in the statement of the result; but the bound in the statement is usually easier to compute.

Continuity of Inversion

XV.10.2.13. If \( A \) is a unital algebra, write \( A^{-1} \) for the set of invertible elements of \( A \). Note that \( A^{-1} \) is a group under multiplication, since a product of invertible elements is invertible.

XV.10.2.14. Corollary. Let \( A \) be a unital Banach algebra. Then the map \( x \mapsto x^{-1} \) is a continuous function from \( A^{-1} \) to \( A^{-1} \).

Proof: By the sequential criterion, it suffices to show that if \((u_n)\) is a sequence in \( A^{-1} \) converging to \( u \in A^{-1} \), then \( u_n^{-1} \to u^{-1} \). There is an \( N \) such that
\[ \|u_n - u\| < \frac{1}{2\|u^{-1}\|} \]
for all \( n \geq N \). Thus, for \( n \geq N \), we have
\[ \|u_n^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}\||u_n - u\|} < 2\|u^{-1}\| . \]
So, for \( n \geq N \),
\[ \|u_n^{-1} - u^{-1}\| = \|u_n^{-1}(u - u_n)u^{-1}\| \leq \|u_n^{-1}\||u_n - u\||u^{-1}\| < 2\|u^{-1}\|^2\|u_n - u\| \]
which goes to 0 as \( n \to \infty \).

XV.10.2.15. Corollary. If \( A \) is a unital Banach algebra, then \( A^{-1} \) (with the norm topology) is a topological group under multiplication.

XV.10.2.16. If \( I \) is an ideal in a Banach algebra \( A \), it follows from continuity of multiplication that the closure \( \bar{I} \) is also an ideal. If \( I \) is a proper ideal, it is not immediately obvious that \( \bar{I} \) is a proper ideal (i.e. \( I \) could be dense in \( A \)); in fact, this can happen if \( A \) is nonunital. But it cannot happen if \( A \) is a unital Banach algebra:

XV.10.2.17. Corollary. Let \( A \) be a unital Banach algebra, and \( I \) proper ideal of \( A \). Then \( \bar{I} \) is also a proper ideal of \( A \).

Proof: An ideal in a unital algebra is a proper ideal if and only if it contains no invertible elements. Thus if \( I \) is a proper ideal, it is contained in the set \( S \) of noninvertible elements of \( A \). But \( S \) is closed in \( A \), so \( \bar{I} \subseteq S \) and \( I \) is a proper ideal.
XV.10.3. Spectrum

We now define the crucial notion of the spectrum of an element of a Banach algebra. For reasons partially explained in XV.10.3.5., we will consider only complex Banach algebras. For simplicity, we will also only develop the theory for unital Banach algebras. Thus, in this section (except (XV.10.3.2.)), all Banach algebras will be unital and complex; subalgebras will be assumed to contain the multiplicative identity. All homomorphisms will be assumed unital.

XV.10.3.1. Definition. Let $A$ be a unital Banach algebra, $x \in A$. The spectrum of $x$ in $A$ is

$$\sigma_A(x) = \{ \lambda \in \mathbb{C} : \lambda I \text{ is not invertible} \}.$$  

The complement of $\sigma_A(x)$ in $\mathbb{C}$ is called the resolvent set of $x$ in $A$, denoted $\rho_A(x)$.

Thus $\sigma_A(x)$ is a subset of $\mathbb{C}$, which depends on $x$, and in general also on $A$ (XV.10.3.11.). In many instances the $A$ is understood (or, more importantly, the spectrum is independent of the choice of $A$, e.g. ()), and we just write $\sigma(x)$ for $\sigma_A(x)$.

XV.10.3.2. If $A$ is a nonunital Banach algebra and $x \in A$, we can define $\sigma_A(x) = \sigma_{B(X)}(x)$. We then have $0 \in \sigma_A(x)$ for every $x \in A$. Although much of the theory we develop works in the nonunital case with occasional technical complications, for simplicity we will not consider spectrum in a nonunital Banach algebra from now on.

Here are the two most important examples, which together underly and motivate spectral theory:

XV.10.3.3. Examples. (i) Let $\mathcal{X}$ be a complex Banach space, and $T \in \mathcal{B}(\mathcal{X})$. Then $\sigma(T) = \sigma_{\mathcal{B}(\mathcal{X})}(T)$ is the set of $\lambda \in \mathbb{C}$ such that $T - \lambda I$ is not invertible, i.e. is either not injective or not surjective (the Bounded Inverse Theorem implies that if $T - \lambda I$ is injective and surjective, then $(T - \lambda I)^{-1} \in \mathcal{B}(\mathcal{X})$). If $T - \lambda I$ is not injective, then $\lambda$ is by definition an eigenvalue of $T$. Thus $\sigma(T)$ can be thought of as the set of generalized eigenvalues of $T$.

If $\mathcal{X}$ is finite-dimensional, then by () $T - \lambda I$ is surjective if and only if it is injective. Thus in this case $\sigma(T)$ is exactly the set of eigenvalues of $T$, a finite nonempty subset of $\mathbb{C}$.

(ii) Let $X$ be a compact Hausdorff space, and $A = C(X)$. If $f \in A$, then $f$ is invertible in $A$ if and only if $f$ never takes the value 0. Thus $f - \lambda I$ is invertible if and only if $f$ never takes the value $\lambda$, i.e. $\sigma_A(f)$ is precisely the range of $f$, a nonempty compact subset of $\mathbb{C}$.

XV.10.3.4. Note that the spectrum of an element of a Banach algebra is a purely algebraic concept, and the norm plays no role in its definition. We may more generally define the spectrum of an element in any unital algebra over $\mathbb{C}$ in an identical way. (Thus, if $V$ is a finite-dimensional complex vector space and $T \in \mathcal{L}(V)$, then $\sigma(T)$ is the set of eigenvalues of $T$.) However, the interplay between the topological and algebraic structure in a Banach algebra has important consequences for the properties of the spectrum.

XV.10.3.5. Example XV.10.3.3.(i) shows one basic reason why we restrict to complex algebras: operators on real vector spaces can fail to have (real) eigenvalues since $\mathbb{R}$ is not algebraically closed, for example a rotation in $\mathbb{R}^2$. There are deeper reasons why we restrict to complex algebras, involving applications of Complex Analysis (e.g. XV.10.3.25.).
Topological Properties of the Spectrum and Resolvent

XV.10.3.6. Proposition. Let $A$ be a unital Banach algebra, and $x \in A$. Then $\sigma_A(x)$ is closed in $\mathbb{C}$, i.e. $\rho_A(x)$ is open in $\mathbb{C}$.

Proof: This follows immediately from the fact that $A^{-1}$ is open in $A$, since $\lambda \mapsto x - \lambda 1$ is a continuous function from $\mathbb{C}$ to $A$.

XV.10.3.7. Proposition. Let $A$ be a unital Banach algebra, and $x \in A$. If $\lambda \in A(x)$, then $|\lambda| \leq \|x\|$. Thus $\sigma_A(x)$ is bounded in $\mathbb{C}$, hence compact.

Proof: The case $x = 0$ is trivial ($A(0) = \{0\}$), so suppose $x \neq 0$. Let $\alpha \in \mathbb{C}$. If $|\alpha| < \frac{1}{\|x\|}$, then the series $\sum_{k=0}^{\infty} (\alpha x)^k$ converges to an inverse for $1 - \alpha x$ by XV.10.2.9. So if $\lambda \in \mathbb{C}$, $|\lambda| > \|x\|$, we have that $1 - \frac{1}{\lambda} x$ is invertible, and hence $x - \lambda 1 = -\lambda(1 - \frac{1}{\lambda} x)$ is also invertible and $\lambda \notin \sigma_A(x)$.

XV.10.3.8. We will see later (XV.10.3.25.) that $\sigma_A(x)$ is always nonempty. A more careful analysis (XV.10.3.30.) along the lines of the argument in the proof of XV.10.3.7. will give a more precise description of the size of the set $\sigma_A(x)$.

XV.10.3.9. Let $A$ be a unital Banach algebra, and $x \in A$. Then continuity of inversion implies that the function $R : \rho_A(x) \to A^{-1}$ given by $R(\lambda) = (\lambda 1 - x)^{-1} = -(x - \lambda 1)^{-1}$ is continuous. $R$ is called the resolvent function of $x$, with $R(\lambda)$ usually denoted $R_A(x)$; it is actually far more than just continuous in $\lambda$ (XV.10.3.23.). (The sign convention for $R_A(x)$ is not universal.)

Dependence of Spectrum on the Algebra

XV.10.3.10. The spectrum of an element can depend on which containing Banach algebra is considered: if $A$ is a (unital) Banach algebra and $B$ a (unital) closed subalgebra of $A$, hence itself a Banach algebra, and $x \in B$, then $x$ can be invertible in $A$ but not in $B$, i.e. we can have $x^{-1} \notin B$. Obviously $\sigma_A(x) \subseteq \sigma_B(x)$, but the containment can be proper:

XV.10.3.11. Example. Let $\mathbb{T}$ be the unit circle in $\mathbb{C}$, and $A = C(\mathbb{T})$. Let $f$ be the “identity function” $f(z) = z$ from $\mathbb{T}$ to $\mathbb{C}$ (note that the term “identity function” can be confusing here since $f$ is not the multiplicative identity of $A$). Then $\sigma_A(f)$ is the range of $f$ (XV.10.3.3.(ii)), i.e. $\sigma_A(f) = \mathbb{T}$.

Let $B$ be the subset of $A$ consisting of functions which extend to functions continuous on the closed unit disk $\overline{\mathbb{D}}$ and holomorphic on $\mathbb{D}$ (B is exactly the algebra $\mathcal{P}$ of Exercise XV.8.3.10.). By the Maximum Modulus Theorem (...), such an extension, if it exists, is unique. By the Maximum Modulus Theorem and X.4.1.2., $B$ is a closed unital subalgebra of $A$, and $f \in B$. However, if $|\lambda| \leq 1$, the unique extension $g_{\lambda}$ of
$f - \lambda 1$ to $\overline{D}$ is $g_\lambda(z) = z - \lambda$, which is not invertible in $B$. If $|\lambda| > 1$, then $\frac{1}{\overline{f}}$ is holomorphic on $\mathbb{C} \setminus \{\lambda\}$, hence on $\overline{D}$, so $f - \lambda 1$ is invertible in $B$. Thus $\sigma_B(f) = \overline{D}$.

However, $\sigma_B(x)$ cannot be too much larger than $\sigma_A(x)$:

**XV.10.3.12.** Proposition. Let $A$ be a unital Banach algebra and $B$ a closed unital subalgebra of $A$, and let $x \in B$. Then $\sigma_A(x) \subseteq \sigma_B(x)$ and $\partial(\sigma_B(x)) \subseteq \partial(\sigma_A(x)) \subseteq \sigma_A(x)$, where $\partial$ denotes topological boundary $()$.

**Proof:** We have already made the trivial observation that $\sigma_A(x) \subseteq \sigma_B(x)$. Suppose $\lambda \in \partial(\sigma_B(x))$. Then $\lambda \in \sigma_B(x)$ since $\sigma_B(x)$ is closed, and there is a sequence $(\lambda_n)$ in $\rho_B(x)$ with $\lambda_n \to \lambda$. Since $\rho_B(x) \subseteq \rho_A(x)$, $\lambda_n \in \rho_A(x)$ for all $n$, so either $\lambda \in \partial(\sigma_A(x))$ (if $\lambda \in \sigma_A(x)$) or $\lambda \in \rho_A(x)$. But if $\lambda \in \rho_A(x)$ (i.e. $x - \lambda 1$ is invertible in $A$), then $(x - \lambda_n 1)^{-1} \to (x - \lambda 1)^{-1}$ by XV.10.3.9., and $(x - \lambda_n 1)^{-1} \in B$ for all $n$, so $(x - \lambda 1)^{-1} \in B$ since $B$ is closed in $A$, contradicting that $\lambda \in \sigma_B(x)$. Thus $\lambda \in \partial(\sigma_A(x))$.  

**XV.10.3.13.** Thus, if $A$ is a unital Banach algebra and $B$ a closed unital subalgebra of $A$, and $x \in B$, then $\sigma_B(x)$ is the union of $\sigma_A(x)$ and possibly some bounded connected components of the complement of $\sigma_A(x)$. If $\sigma_A(x)$ does not divide $\mathbb{C}$ (i.e. if the complement is connected), then necessarily $\sigma_B(x) = \sigma_A(x)$. This is also true if $\sigma_B(x)$ has empty interior.

**Spectral Mapping Properties**

**XV.10.3.14.** If $f$ is a polynomial with complex coefficients, and $x$ is an element of a unital algebra $A$, then there is an obvious way to apply $f$ to $x$ to obtain an element $f(x) \in A$: if

$$f(X) = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$$

then

$$f(x) = a_0 1 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$ 

If $f$ and $g$ are polynomials with complex coefficients, and $x \in A$, then $(f + g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$, i.e. the map $f \mapsto f(x)$ is a homomorphism $()$ from the polynomial ring $\mathbb{C}[X]$ to $A$.

**XV.10.3.15.** Proposition. [Spectral Mapping Theorem for Polynomials] Let $A$ be a unital algebra over $\mathbb{C}$, $x \in A$, and $f$ a polynomial with complex coefficients. Then

$$\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\}.$$ 

**Proof:** Let $\mu \in \mathbb{C}$. Since $\mathbb{C}$ is algebraically closed, the polynomial $f(X) - \mu$ factors into linear factors in $\mathbb{C}[X]$:

$$f(X) - \mu = (X - \lambda_1)(X - \lambda_2) \cdots (X - \lambda_n)$$

for some $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ (not necessarily distinct). We then have

$$f(x) - \mu 1 = (x - \lambda_1 1) \cdots (x - \lambda_n 1)$$

1818
in $A$. Then by the next lemma $f(x) - \mu 1$ is invertible in $A$ if and only if $x - \lambda_k 1$ is invertible in $A$ for all $k$, i.e. $\mu \in \sigma_A(f(x))$ if and only if $\lambda_k \in \sigma_A(x)$ for some $k$. Since the numbers $\lambda_1, \ldots, \lambda_n$ are precisely the complex numbers $\lambda$ such that $f(\lambda) = \mu$, the result follows.

**XV.10.3.16. Lemma.** Let $A$ be a unital ring (e.g. a unital algebra over a field $F$), and $a_1, \ldots, a_n \in A$. If the $a_k$ commute, then $a = a_1 a_2 \cdots a_n$ is invertible in $A$ if and only if each $a_k$ is invertible in $A$, i.e. $a$ is invertible if and only if each $a_k$ is invertible in $A$.

**Proof:** If the $a_k$ are invertible, then $a$ is invertible, and $a^{-1} = a_1^{-1} \cdots a_n^{-1}$ (this direction holds even if the $a_k$ do not commute). Conversely, if the $a_k$ commute and $a$ is invertible, we have for each $k$

$$a_k(a_1 a_2 \cdots a_{k-1} a_{k+1} \cdots a_n a^{-1}) = 1$$

so $a_k$ is both left and right invertible in $A$.

Note that the conclusion can be false if the $a_k$ do not commute ()

**XV.10.3.17. Proposition.** [Spectral Mapping Theorem for Inverses] Let $A$ be a unital algebra over $\mathbb{C}$, and $x \in A$. If $x$ is invertible in $A$ (i.e. $0 \notin \sigma_A(x)$), then

$$\sigma_A(x^{-1}) = \left\{ \frac{1}{\lambda} : \lambda \in \sigma_A(x) \right\}.$$

**Proof:** Since $x^{-1}$ is invertible, $0 \notin \sigma_A(x^{-1})$. If $\lambda \neq 0$, we have

$$x^{-1} - \frac{1}{\lambda} = -\frac{1}{\lambda} x^{-1}(x - \lambda 1)$$

and $x^{-1}$ and $x - \lambda 1$ commute, so $x^{-1} - \frac{1}{\lambda} 1$ is invertible if and only if $x - \lambda 1$ is invertible.

Here is a cute fact related in spirit:

**XV.10.3.18. Proposition.** Let $A$ be a unital algebra over $\mathbb{C}$, $x, y \in A$. Then

$$\sigma_A(xy) \cup \{0\} = \sigma_A(yx) \cup \{0\}.$$

**Proof:** Suppose $\lambda \neq 0$ and $xy - \lambda 1$ is invertible (i.e. $\lambda \notin \sigma_A(xy)$). Then, if $r = (xy - \lambda 1)^{-1}$, we have

$$(xy - \lambda 1)r = r(xy - \lambda 1) = 1$$

$$xyr = rxy = 1 - \lambda r$$

So we have

$$-\frac{1}{\lambda} (1 - yr)(xy - \lambda 1) = -\frac{1}{\lambda} [yx - yrxy + \lambda yrx - \lambda 1] = -\frac{1}{\lambda} [yx - y(1 - \lambda r)x + \lambda yrx - \lambda 1] = 1$$

and similarly $-\frac{1}{\lambda} (yx - \lambda 1)(1 - yr) = 1$, so $yx - \lambda 1$ is invertible and $\lambda \notin \sigma_A(yx)$. The $x$ and $y$ can be interchanged in the argument.
**XV.10.3.19.** So, while $0 \in \sigma_A(xy)$ (i.e. $xy$ not invertible) does not imply $0 \in \sigma_A(yx)$ (i.e. $yx$ not invertible), the rest of the two spectra are identical. If $xy$ is invertible but $yx$ is not invertible, then 0 is an isolated point of the spectrum of $yx$ (since $\sigma_A(xy) = \sigma_A(yx) \setminus \{0\}$ is closed in $\mathbb{C}$).

**Holomorphic Properties of the Resolvent**

Recall (XV.10.3.9.) that if $x$ is an element of a unital Banach algebra $A$, the resolvent function of $x$ is defined to be

$$R_\lambda(x) = (\lambda I - x)^{-1}$$

for $\lambda \in \rho_A(x) \subseteq \mathbb{C}$. In what follows, we keep $A$ and $x$ fixed, and just write $R_\lambda$ for $R_\lambda(x)$.

**XV.10.3.20.** **Proposition.**

$$\lim_{|\lambda| \to \infty} R_\lambda = 0 .$$

**Proof:** Note that since $\sigma_A(x)$ is bounded (XV.10.3.7.), $R_\lambda$ is defined for $|\lambda|$ sufficiently large, so the limit potentially exists. If $|\lambda|$ is large, we have

$$R_\lambda = (\lambda 1 - x)^{-1} = \alpha(1 - \alpha x)^{-1}$$

where $\alpha = \frac{1}{\lambda}$. We have

$$\lim_{\alpha \to 0}(1 - \alpha x) = 1$$

and by XV.10.2.10. and XV.10.2.14. we also have

$$\lim_{\alpha \to 0}(1 - \alpha x)^{-1} = 1 .$$

Thus

$$\lim_{\alpha \to 0} \alpha(1 - \alpha x)^{-1} = 0 .$$

\[ \checkmark \]

**XV.10.3.21.** **Proposition.** Let $\lambda, \mu \in \rho_A(x)$. Then

$$R_\mu = (1 - y)^{-1}R_\lambda$$

where $y = (\lambda - \mu)R_\lambda$.

**Proof:** We have

$$\mu 1 - x = (\lambda 1 - x) - (\lambda - \mu) = (\lambda 1 - x)(1 - y)$$

so the formula follows by taking inverses.\[ \checkmark \]

Putting this together with XV.10.2.9., we obtain:
**XV.10.3.22.** COROLLARY. Let \( \lambda_0 \in \rho_A(x) \), and \( \lambda \in \mathbb{C} \). If the series

\[
\sum_{k=0}^{\infty} (\lambda_0 - \lambda)^k R_{\lambda_0}^{k+1}
\]

converges in \( A \), then \( \lambda \in \rho_A(x) \) and the sum is \( R_\lambda \). In particular, if \( |\lambda - \lambda_0| < \frac{1}{\|R_{\lambda_0}\|} \), \( \lambda \in \rho_A(x) \) and the series converges to \( R_\lambda \).

**Proof:** Apply XV.10.3.21. with \((\lambda, \mu)\) replaced by \((\lambda_0, \lambda)\), and apply XV.10.2.9. to \( y = (\lambda_0 - \lambda)R_{\lambda_0} \). ☑️

**XV.10.3.23.** Thus \( R_\lambda \) is locally representable as a power series, and thus is an “analytic function” of \( \lambda \). There is a theory of Banach-space-valued analytic functions () which applies. However, we will need only the following weaker form which reduces consideration to complex-valued functions:

**XV.10.3.24.** COROLLARY. Let \( \phi \in A^* \), and define \( f_\phi(\lambda) = \phi(R_\lambda) \) for \( \lambda \in \rho_A(x) \). Then \( f_\phi \) is a holomorphic (analytic) complex-valued function on \( A(x) \).

**Proof:** Fix \( \lambda_0 \in \rho_A(x) \). If \( |\lambda - \lambda_0| < \frac{1}{\|R_{\lambda_0}\|} \), then the infinite series

\[
\sum_{k=0}^{\infty} \phi(R_{\lambda_0}^{k+1})(\lambda_0 - \lambda)^k = \sum_{k=0}^{\infty} (-1)^k \phi(R_{\lambda_0}^{k+1})(\lambda - \lambda_0)^k
\]

converges to \( f_\phi(\lambda) \) by XV.10.3.22. and continuity of \( \phi \). Thus \( f_\phi \) is locally representable by a convergent power series. ☑️

**The Spectrum is Nonempty!**

The first major consequence is:

**XV.10.3.25.** THEOREM. Let \( A \) be a unital Banach algebra, and \( x \in A \). Then \( \sigma_A(x) \) is nonempty.

**Proof:** Suppose \( \sigma_A(x) = \emptyset \), so \( \rho_A(x) = \mathbb{C} \). Let \( \phi \in A^* \). Then \( f_\phi \) is an entire function (). We have

\[
\lim_{|\lambda| \to \infty} f_\phi(\lambda) = 0
\]

by XV.10.3.20. and continuity of \( \phi \). Thus \( f_\phi \) is bounded, hence constant by Liouville’s Theorem (X.3.3.10.). Since the limit at infinity is 0, \( f_\phi \) is identically 0, i.e. \( \phi(R_\lambda) = 0 \) for all \( \lambda \in \mathbb{C} \) and all \( \phi \in A^* \). By () this implies that \( R_\lambda = 0 \) for all \( \lambda \), a contradiction. ☑️

A simple corollary is:
XV.10.3.26. **Corollary.** [Gelfand-Mazur Theorem] The only (complex) Banach division algebra is $\mathbb{C}$.

**Proof:** Recall that $A$ is a division algebra if every nonzero element is invertible. If $x \in A$, there is a $\lambda \in \sigma_A(x)$. For this $\lambda$, $x - \lambda 1$ is not invertible; hence $x - \lambda 1 = 0$, $x = \lambda 1$. Thus every element of $A$ is a scalar multiple of 1. ☐

### Spectral Radius

Let $x$ be an element of a unital Banach algebra $A$. Since $\sigma_A(x)$ is a nonempty compact subset of the plane, it makes sense to make the following definition:

**XV.10.3.27. Definition.** The *spectral radius* of $x$ is

$$
 r_\sigma(x) = \sup_{\lambda \in \sigma_A(x)} |\lambda| = \max_{\lambda \in \sigma_A(x)} |\lambda| .
$$

**XV.10.3.28.** The notation $r_\sigma(x)$ does not reflect the choice of the Banach algebra $A$ containing $x$, on which the spectral radius potentially depends. But it is an immediate consequence of XV.10.3.12. that if $B$ is a unital Banach subalgebra of $A$ and $x \in B$, then the same spectral radius for $x$ is obtained whether $\sigma_A(x)$ or $\sigma_B(x)$ is used in the definition, so $r_\sigma(x)$ in fact does not depend on which Banach algebra containing $x$ is used. We could indeed just use the *smallest* Banach subalgebra of $A$ containing $x$, i.e. the closed linear span of $\{x^n : n \geq 0\}$.

**XV.10.3.29.** Note also that the spectral radius of $x$ depends only on the *algebraic* structure of $A$ and not on the norm. However, we needed to know that $A$ was a Banach algebra with respect to *some* norm in order to conclude that $\sigma_A(x)$ is nonempty and bounded so that the spectral radius is well defined.

There is nonetheless a tight relationship between the spectral radius and norm:

**XV.10.3.30. Theorem.** Let $A$ be a Banach algebra and $x \in A$. Then

$$
 r_\sigma(x) = \lim_{n \to \infty} \|x^n\|^{1/n} = \inf_n \|x^n\|^{1/n} .
$$

The theorem will be proved in several steps.

**XV.10.3.31.** Part of the content of the theorem is that

$$
 \lim_{n \to \infty} \|x^n\|^{1/n}
$$

exists and equals

$$
 \inf_n \|x^n\|^{1/n} .
$$
This can be proved directly using IV.11.3., setting \( n = \| x \|^n \), since \( \| x^n + m \| = \| x^n \| \leq \| x \|^n \| x^m \| \) for any \( n, m \in \mathbb{N} \) by submultiplicativity. It will not be necessary for us to prove this since it will follow automatically from our other arguments. The sequence \( \| x^n \|^{1/n} \) is not necessarily nonincreasing.

The first observation for the proof of XV.10.3.30. is that \( \mathcal{r}(x) \leq \| x \| \) by XV.10.3.7. We can do better by noting the following consequence of the Spectral Mapping Theorem for polynomials (XV.10.3.15):

**XV.10.3.32. Proposition.** Let \( x \) be an element of a unital Banach algebra \( A \), and \( f \) a polynomial with complex coefficients. Then
\[
r_\sigma(f(x)) = \max_{\lambda \in \sigma(A)} |f(\lambda)|.
\]
In particular, \( r_\sigma(x^n) = [r_\sigma(x)]^n \) for any \( n \in \mathbb{N} \).

**XV.10.3.33. Corollary.** Let \( x \) be an element of a unital Banach algebra \( A \). Then
\[
r_\sigma(x) \leq \inf_n \| x^n \|^{1/n}.
\]

**Proof:** By XV.10.3.32. and XV.10.3.7. we have, for any \( n \),
\[
[r_\sigma(x)]^n = r_\sigma(x^n) \leq \| x^n \|.
\]

**XV.10.3.34.** So to finish the proof of Theorem XV.10.3.30., it suffices to show that
\[
r_\sigma(x) \geq \limsup_{n \to \infty} \| x^n \|^{1/n}.
\]
This will take some work using Complex Analysis. Our argument follows [Ped89]; roughly speaking, it is based on X.4.4.6.

**XV.10.3.35.** Fix \( \phi \in \mathcal{A}^* \) and let \( f_\phi(\lambda) = \phi(R_\lambda) \). Then \( f_\phi \) is analytic on \( \rho_A(x) \), and in particular on \( \Omega = \{ \lambda \in \mathbb{C} : |\lambda| > r_\sigma(x) \} \). We have that
\[
f_\phi(\lambda) = \frac{1}{\lambda} \phi \left( \left[ 1 - \frac{1}{\lambda} x \right]^{-1} \right) = \frac{1}{\lambda} \phi \left( \sum_{k=0}^{\infty} \left[ \frac{1}{\lambda} x \right]^k \right) = \frac{1}{\lambda} \sum_{k=0}^{\infty} \phi \left( \left[ \frac{1}{\lambda} x \right]^k \right) = \sum_{k=0}^{\infty} \lambda^{-k-1} \phi(x^k)
\]
for \( |\lambda| > \| x \| \); thus this is the Laurent series for \( f_\phi \) on \( \Omega \) and converges to \( f_\phi \) u.c. on \( \Omega \) (X.4.4.11.).
Now fix $s > r_\sigma(x)$. For any $n \in \mathbb{N}$, the series

$$\sum_{k=0}^{\infty} \lambda^{n-k} \phi(x^k)$$

converges uniformly to $g_n(\lambda) = \lambda^{n+1} f_\phi(\lambda)$ on the circle $\gamma$ of radius $s$ around $0$, so we can compute the integral of $g_n$ around $\gamma$ by integrating term-by-term:

$$\int_0^{2\pi} s^{n+1} e^{i(n+1)\theta} \phi(se^{i\theta}) d\theta = \sum_{k=0}^{\infty} \left[ \int_0^{2\pi} s^{-k} e^{i(n-k)\theta} \phi(x^k) d\theta \right] = \int_0^{2\pi} \phi(x^n) d\theta = 2\pi \phi(x^n)$$

since the integrals for $k \neq n$ are zero. Thus we obtain that

$$|\phi(x^n)| \leq s^{n+1} M(s) \|\phi\|$$

where $M(s) = \max_{|\lambda|=s} \|R_\lambda\|$. Since $\phi \in A^*$ is arbitrary, by the Hahn-Banach Theorem we conclude that

$$\|x^n\| \leq s^{n+1} M(s)$$

$$\|x^n\|^{1/n} \leq s^{\frac{n+1}{n}} M(s)^{1/n}$$

for every $n$. Taking lim sup as $n \to \infty$, we obtain

$$\limsup_{n \to \infty} \|x^n\|^{1/n} \leq s.$$ 

This holds for all $s > r_\sigma(x)$, so we conclude

$$\limsup_{n \to \infty} \|x^n\|^{1/n} \leq r_\sigma(x).$$

This completes the proof of Theorem XV.10.3.30.

XV.10.4. Holomorphic Functional Calculus

XV.10.4.1. If $f(t) = \alpha_1 t + \cdots + \alpha_n t^n$ is a polynomial with complex coefficients, without constant term (i.e. $f(0) = 0$), and $x$ is an element of an algebra $A$ over $\mathbb{C}$, then there is an obvious way to apply $f$ to $x$ to obtain an element $f(x) = \alpha_1 x + \cdots + \alpha_n x^n \in A$. If $A$ is a Banach algebra, there is a very important way of extending this procedure to holomorphic functions, called functional calculus.

XV.10.4.2. If $X$ is a compact subset of $\mathbb{C}$, denote by $H(X)$ the algebra of functions holomorphic in a neighborhood of $X$ and vanishing at 0 if $0 \in X$, with functions identified if they agree on a neighborhood of $X$. Functional calculus gives a homomorphism from $H(\sigma_A(x))$ to the Banach subalgebra of $A$ generated by $x$ extending the map for polynomials. The image of $f$ is denoted $f(x)$. The element $f(x)$ can be defined using the Cauchy Integral Formula, but in some cases (e.g. if $f$ is entire) it is also given by a power series. If $A$ is unital, functional calculus is also defined for holomorphic functions not vanishing at zero. Functional calculus has the following properties, which (along with the elementary definition for polynomials) determine it uniquely:
XV.10.4.3. **Proposition.** Let $A$ be a Banach algebra and $x \in A$. Then

(i) For any $f \in H(\sigma_A(x))$, $\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\}$.

(ii) If $f \in H(\sigma_A(x))$ and $g \in H(f(\sigma_A(x))) = H(\sigma_A(f(x)))$, so $g \circ f \in H(\sigma_A(x))$, then $(g \circ f)(x) = g(f(x))$.

(iii) If $f_n, f \in H(\sigma_A(x))$ and $f_n \to f$ uniformly on a neighborhood of $\sigma_A(x)$, then $f_n(x) \to f(x)$.

(iv) If $B$ is a Banach algebra and $\phi : A \to B$ a continuous (bounded) homomorphism, then $\phi(f(x)) = f(\phi(x))$ for any $f \in H(\sigma_A(x))$.

**Proof:** (i): Suppose $A$ is unital. Let $f$ be analytic on $U$ containing $\sigma_A(x)$. If $\lambda \in \sigma_A(x)$, then $f(z) - f(\lambda) = (z - \lambda)g(z)$ with $g$ analytic on $U$; then $f(x) - f(\lambda)1 = (x - \lambda 1)g(x)$, and since $x - \lambda 1$ and $g(x)$ commute and $x - \lambda 1$ is not invertible, $f(\lambda) \in \sigma_A(f(x))$. Conversely, if $\mu \notin \{f(\lambda) : \lambda \in \sigma_A(x)\}$, then $h(z) = (f(z) - \mu)^{-1}$ is analytic on $\{z \in U : f(z) \neq \mu\}$, which contains $\sigma_A(x)$, and $h(x) = (f(x) - \mu 1)^{-1}$.

(ii)-(iv) are straightforward.

XV.10.4.4. An especially important case of functional calculus uses the exponential function $f(z) = e^z$. If $x$ is any element of a unital Banach algebra $A$, then $f(x)$ is defined and denoted $e^x$. The element $e^x$ is given by the power series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$. For any $x, e^x$ is invertible, with inverse $e^{-x}$; $e^{x+y} = e^x e^y$ if $x$ and $y$ commute. If $A$ is a Banach *-algebra, then $(e^x)^* = e^{(x^*)}$. Conversely, if $\sigma_A(x)$ is contained in a simply connected open set not containing 0, let $f(z)$ be a branch of the logarithm holomorphic in a neighborhood of $\sigma_A(x)$; then $y = f(x)$ satisfies $e^y = x$. In particular, if $\|1 - x\| < 1$, then there is a $y \in A$ with $\|y\| < \pi/2$ with $e^y = x$ (use the principal branch of log).

XV.10.4.5. If $A$ is a unital Banach algebra, write $A^{-1}$ (also often written $GL_1(A)$) for the (open) set of invertible elements in $A$, and $A_0^{-1}$ (or $GL_1(A)_0$) for the connected component of the identity in $A^{-1}$. Then

$$exp(A) = \{e^y : y \in A\} \subseteq A^{-1}$$

is path-connected, and by the above the subgroup of $A^{-1}$ generated algebraically by $exp(A)$ is a connected open subgroup, hence equal to $A_0^{-1}$. In particular, $A_0^{-1}$ is an open subgroup of $A^{-1}$, and every element of $A_0^{-1}$ is a finite product of exponentials.

If $A$ and $B$ are unital Banach algebras and $\phi$ is a bounded homomorphism from $A$ onto $B$, and $y \in B^{-1}$, then there is not necessarily an $x \in A^{-1}$ with $\phi(x) = y$.

XV.10.4.6. **Example.** Let $\mathbb{D}$ be the closed unit disk in $\mathbb{C}$, and $\mathbb{T}$ its boundary. There is a *-homomorphism $\phi : C(\mathbb{D}) \to C(\mathbb{T})$ given by restriction. But if $g \in C(\mathbb{T})$ with $g(z) = z$, then $g \in C(\mathbb{T})^{-1}$, but there is no $f \in C(\mathbb{D})^{-1}$ with $\phi(f) = g$.

The image in the Calkin algebra of a Fredholm operator of nonzero index is another example. However:
XV.10.4.7. PROPOSITION. If $\phi : A \to B$ is a surjective bounded homomorphism of unital Banach algebras, then $\phi(A_o^{-1}) = B_o^{-1}$.

This can be proved either by an application of the Open Mapping Theorem or by noting that if $x \in A$, then $\phi(e^x) = e^{\phi(x)}$.

Homomorphisms

We have a generalization of XV.10.3.10:

XV.10.4.8. PROPOSITION. Let $A$ and $B$ be unital Banach algebras, and $\phi : A \to B$ a unital homomorphism. If $x \in A$, then $\sigma_B(\phi(x)) \subseteq \sigma_A(x)$.

PROOF: If $y \in A$ is invertible in $A$, then $\phi(y)$ is invertible in $B \ [\phi(y)^{-1} = \phi(y^{-1})]$. ♦
XV.10.5. Spectral Theory for Bounded Operators

In this section, we examine the spectrum and its properties for bounded operators on Banach spaces, concentrating almost exclusively on the case of operators on Hilbert spaces.

We will essentially follow the first comprehensive development and exposition of this theory, given by F. Riesz [7] in 1913, abstracted to Banach spaces and Hilbert spaces (which did not exist in 1913!) and with some modern technical simplifications.

XV.10.5.1. If \( T \in B(\mathcal{X}, \mathcal{Y}) \) (\( \mathcal{X}, \mathcal{Y} \) Banach spaces) is invertible in the algebraic sense, i.e. \( T \) is one-to-one and onto, then \( T^{-1} \) is automatically bounded by the Open Mapping Theorem. So there are two ways a bounded operator \( T \) can fail to be invertible:

- \( T \) is not bounded below (recall that \( T \) is bounded below if \( \exists \epsilon > 0 \) such that \( \|Tx\| \geq \epsilon \|x\| \) for all \( x \in \mathcal{X} \)

- \( \mathcal{R}(T) \) is not dense.

These possibilities are not mutually exclusive. For example, if \( T \) is a normal operator on a Hilbert space, then \( \mathcal{N}(T) = \mathcal{R}(T)^\perp \) (\( \perp \)), so if \( \mathcal{R}(T) \) is not dense, then \( T \) is not one-to-one and hence not bounded below. Thus a normal operator on a Hilbert space is invertible if and only if it is bounded below.

XV.10.5.2. Proposition. Let \( T \) be a bounded operator on a Hilbert space \( \mathcal{H} \). Then \( T \) is invertible if and only if both \( T \) and \( T^* \) are bounded below.

Proof: If \( T \) is invertible, then for any \( x \) we have \( \|x\| = \|T^{-1}(Tx)\| \leq \|T^{-1}\|\|Tx\| \), so \( T \) is bounded below. Since \( T^* \) is also invertible, \( T^* \) is bounded below too.

Conversely, suppose \( T \) and \( T^* \) are bounded below. In particular, \( T^* \) is one-to-one, so \( \mathcal{N}(T^*) = \mathcal{R}(T)^\perp = \{0\} \), and thus \( \mathcal{R}(T) \) is dense in \( \mathcal{H} \). If \( y \in \mathcal{H} \), then there is a sequence \( (y_n) \) in \( \mathcal{R}(T) \) with \( y_n \to y \). For each \( n \) there is an \( x_n \in \mathcal{H} \) (unique, since \( T \) is one-to-one) with \( y_n = Tx_n \). Fix \( \epsilon > 0 \) such that \( \|Tx\| \geq \epsilon \|x\| \) for all \( x \in \mathcal{H} \). Then, for any \( n \) and \( m \),

\[
\|x_n - x_m\| \leq \epsilon^{-1}\|Tx_n - Tx_m\| = \epsilon^{-1}\|y_n - y_m\|
\]

and since \( (y_n) \) is a Cauchy sequence, so is \( (x_n) \). Let \( x_n \to x \). We then have \( Tx_n \to Tx \) since \( T \) is continuous (bounded), so \( y = Tx \) and \( T \) is surjective. \( T^{-1} \) is automatically bounded by the Bounded Inverse Theorem.

Recall the following definition from XV.10.3.1., rewritten in the context of operators and expanded:

XV.10.5.3. Definition. Let \( \mathcal{X} \) be a normed vector space, \( T \in B(\mathcal{X}) \). The spectrum of \( T \) is

\[
\sigma(T) = \{ \lambda \in \mathbb{C} : T - \lambda I \text{ is not invertible} \}
\]

The point spectrum of \( T \) is

\[
\sigma_p(T) = \{ \lambda \in \mathbb{C} : T - \lambda I \text{ is not one-to-one} \}
\]

(thus \( \sigma_p(T) \) is exactly the set of eigenvalues of \( T \)). The continuous spectrum \( \sigma_c(T) \) of \( T \) is the set of \( \lambda \in \mathbb{C} \) such that \( T - \lambda I \) is one-to-one with dense range, but not bounded below. The residual spectrum \( \sigma_r(T) \) is the set of \( \lambda \in \mathbb{C} \) such that \( T - \lambda I \) is one-to-one but does not have dense range.

Note that \( \sigma(T) \) is the disjoint union of \( \sigma_p(T) \), \( \sigma_c(T) \), and \( \sigma_r(T) \). One or more of these sets can be empty (XV.10.5.7.). But we have the fundamental fact from XV.10.3.6., XV.10.3.25., and XV.10.3.30.:
**XV.10.5.4. Theorem.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{B}(\mathcal{X})$. Then

(i) $\sigma(T)$ is a nonempty compact subset of $\mathbb{C}$.

(ii) $r_\sigma(T) = \max_{\lambda \in \sigma(T)} |\lambda| = \lim_{n \to \infty} \|T^n\|^{1/n} = \inf_n \|T^n\|^{1/n}$.

if $\mathcal{X}$ is finite-dimensional, then $\sigma(T) = \sigma_p(T)$, and if $\mathcal{X}$ is a Hilbert space and $T$ is normal, then $\sigma_*(T) = \emptyset$.

The spectrum of $T$ can be thought of as the set of “generalized eigenvalues” of $T$ (note that every actual eigenvalue of $T$ is in the spectrum). The first definition (not the same as the one here – essentially the reciprocal) and name “spectrum” were given by HILBERT; the modern definition is due to F. RIESZ.

**XV.10.5.5. Proposition.** Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. Then $\sigma(T^*) = \{\bar{\lambda} : \lambda \in \sigma(T)\}$.

**Proof:** This follows immediately from the fact that if $S \in \mathcal{B}(\mathcal{H})$, then $S$ is invertible if and only if $S^*$ is invertible, and $(T + \lambda I)^* = T^* + \bar{\lambda}I$.

The relation between $\sigma_p(T)$, $\sigma_c(T)$, $\sigma_r(T)$ and $\sigma_p(T^*)$, $\sigma_c(T^*)$, $\sigma_r(T^*)$ is slightly more complicated:

**XV.10.5.6. Proposition.** Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. If $\lambda \in \mathbb{C}$, then

(i) If $\lambda \in \sigma_p(T)$, then $\bar{\lambda} \in \sigma_p(T^*) \cup \sigma_r(T^*)$.

(ii) If $\lambda \in \sigma_r(T)$, then $\bar{\lambda} \in \sigma_p(T^*)$.

(iii) $\lambda \in \sigma_p(T) \cup \sigma_r(T)$ if and only if $\bar{\lambda} \in \sigma_p(T^*) \cup \sigma_r(T^*)$.

(iv) $\lambda \in \sigma_c(T)$ if and only if $\bar{\lambda} \in \sigma_c(T^*)$.

**Proof:** Parts (i) and (ii) follow immediately from the relation $\mathcal{N}(S) = \mathcal{R}(S^*)^\perp$; (iii) follows trivially from (i) and (ii), and (iv) follows from (iii) and XV.10.5.5. by process of elimination.

**XV.10.5.7. Examples.** (i) If $\mathcal{X}$ is finite-dimensional, then $\sigma(T) = \sigma_p(T)$ is the set of eigenvalues of $T$.

(ii) If $T$ is a normal operator on a Hilbert space, then $N(T) = R(T)^\perp$ (i), so if $R(T)$ is not dense, then $T$ is not one-to-one and hence not bounded below. Thus a normal operator on a Hilbert space is invertible if and only if it is bounded below, and $\sigma_r(T) = \emptyset$.

(iii) Let $(\alpha_n)$ be a bounded sequence in $\mathbb{C}$. Define $T : \ell^2 \to \ell^2$ by

$$T(\xi_1, \xi_2, \ldots) = (\alpha_1 \xi_1, \alpha_2 \xi_2, \ldots).$$

Then $T \in \mathcal{B}(\ell^2)$, $\|T\| = \sup_n |\alpha_n|$. We have $\sigma_p(T) = \{\alpha_n : n \in \mathbb{N}\}$, $\sigma_r(T) = \emptyset$ since $T$ is normal, and $\sigma(T)$ is the closure of $\sigma_p(T)$, so if $\sigma_p(T)$ is not closed we have $\sigma_c(T) = \sigma(T) \setminus \sigma_p(T) \neq \emptyset$. 

1828
(iv) Let $T$ be the unilateral shift on $\ell^2$: 

$$T(\xi_1, \xi_2, \ldots) = (0, \xi_1, \xi_2, \ldots).$$

Then $T$ is an isometry, hence $\|T\| = 1$. $T$ is one-to-one, but $\mathcal{R}(T)$ is not dense in $\ell^2$; thus $0 \in \sigma_r(T)$. It is easy to see that $\sigma_p(T) = \emptyset$. By XV.10.5.5., XV.10.5.6., and (v) below, $\sigma(T)$ is the closed unit disk $\mathbb{D}$ in $\mathbb{C}$, $\sigma_r(T)$ is the open unit disk, and $\sigma_e(T)$ is the unit circle.

(v) Let $T^*$ be the adjoint of the unilateral shift of (iv). Then $T^*$ is the left shift:

$$T^*(\xi_1, \xi_2, \ldots) = (\xi_2, \xi_3, \ldots).$$

If $\lambda \in \mathbb{C}$, $|\lambda| < 1$, then the vector $x_\lambda = (1, \lambda, \lambda^2, \ldots)$ is in $\ell^2$ and $T^*(x_\lambda) = \lambda x_\lambda$; thus $\lambda \in \sigma_p(T^*)$. It is easy to see that no complex number of absolute value 1 can be an eigenvalue of $T^*$, so $\sigma_p(T^*)$ is the open unit disk $\mathbb{D}$. Thus $\sigma(T^*)$ contains $\mathbb{D}$ since it is closed. But $\|T^*\| = 1$, so $r(\sigma(T^*)) \leq 1$ and $\sigma(T^*) \subseteq \mathbb{D}$. Thus $\sigma(T^*) = \mathbb{D}$. By XV.10.5.5., XV.10.5.6., and (iv), $\sigma_e(T^*)$ is the unit circle and $\sigma_r(T^*) = \emptyset$.

### Numerical Range

**XV.10.5.8.** The numerical range of an operator $T$ on a Hilbert space $\mathcal{H}$ is 

$$W(T) = \{\langle Tx, x \rangle : \|x\| = 1\}.$$ 

Set 

$$w(T) = \sup\{\|x\| : \lambda \in W(T)\}.$$ 

It is obvious that $W(T^*) = \{\lambda : \lambda \in W(T)\}$, so $w(T^*) = w(T)$; if $T = T^*$, then $W(T) \subseteq \mathbb{R}$. (Conversely, if $W(T) \subseteq \mathbb{R}$, then $T = T^*$: for $\xi, \eta \in \mathcal{H}$, expand $\langle T(\xi + \eta), \xi + \eta \rangle \in \mathbb{R}$ and $\langle T(\xi + i\eta), \xi + i\eta \rangle \in \mathbb{R}$.)

**XV.10.5.9.** **Proposition.** If $T \in \mathcal{B}(\mathcal{H})$, then $\sigma(T) \subseteq W(T)$ and $r(\sigma(T)) \leq w(T) \leq \|T\|$.

**Proof:** To see that $\sigma(T) \subseteq W(T)$, note that if $\lambda \in \sigma(T)$, then by XV.10.5.2. either $T - \lambda I$ or $(T - \lambda I)^*$ is not bounded below, so there is a sequence $(x_n)$ of unit vectors such that either $(T - \lambda I)x_n \to 0$ or $(T - \lambda I)^*x_n \to 0$. In either case, we obtain that $(T - \lambda I)x_n, x_n) \to 0$ by the CBS inequality. The inequality $r(\sigma(T)) \leq w(T)$ follows immediately, and $w(T) \leq \|T\|$ is immediate from the CBS inequality.

**XV.10.5.10.** **Examples.** (i) Note that $r(\sigma(T)) = w(T) = \|T\|$ does not hold for general operators: for example, let $T$ be the (nilpotent) operator on $\mathbb{C}^2$ given by the matrix $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. Then $\sigma(T) = \{0\}$, so $r(\sigma(T)) = 0$; but $\|T\| = 1$. It can be shown by an easy calculation that $W(T) = \{\lambda : |\lambda| \leq 1/2\}$, so $w(T) = 1/2$.

(ii) If $\mathcal{H}$ is finite-dimensional and $T \in \mathcal{B}(\mathcal{H})$, then $W(T)$ is closed since the unit sphere in $\mathcal{H}$ is compact, and thus $\sigma(T) \subseteq W(T)$. But if $\mathcal{H}$ is infinite-dimensional, neither statement is true in general. Let $T \in \mathcal{B}(\ell^2)$ be defined by 

$$T(\xi_1, \xi_2, \ldots) = (\xi_1, \xi_2/2, \ldots, \xi_n/n, \ldots).$$

Then $\sigma(T) = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}$. But $\langle Tx, x \rangle \neq 0$ if $x \neq 0$, so $0 \notin W(T)$. It is not hard to compute that $W(T) = (0, 1]$. 

1829
**XV.10.5.11.** The expression $\langle Tx, x \rangle$ is not linear in $x$, and anyway the set of unit vectors is not convex. However, perhaps remarkably, the numerical range $W(T)$ is always convex (Hausdorff-Toeplitz Theorem). See XV.10.6.2. for a proof. It is also true that $w(T)$ is a norm on $B(H)$ for a Hilbert space $H$ (XV.10.6.3).

**Self-Adjoint and Positive Operators**

**XV.10.5.12.** Definition. A (necessarily self-adjoint; cf. XV.10.6.4.) bounded operator $T$ on a Hilbert space $H$ is positive if $\langle Tx, x \rangle \geq 0$ for all $x \in H$ (i.e. $W(T) \subseteq [0, +\infty)$). Write $T \geq 0$ if $T$ is positive, and write $B(H)_+$ for the set of positive operators on $H$.

**XV.10.5.13.** If $T \in B(H)$, then $T^*T \geq 0$ (this is also true if $T$ is conjugate-linear); if $T = T^*$, then $\lambda I - T \geq 0$ for $\lambda \geq \|T\|$. Sums and nonnegative scalar multiples of positive operators are positive.

If $S,T$ are self-adjoint operators, write $S \leq T$ if $T - S \geq 0$, i.e. if $\langle Sx, x \rangle \leq \langle Tx, x \rangle$ for all $x$. If $T = T^*$, we have $-w(T)I \leq T \leq w(T)I$, and in particular $-\|T\|I \leq T \leq \|T\|I$.

**XV.10.5.14.** Let $S$ be a positive operator, and let $m$ and $M$ be the inf and sup of $W(S)$ respectively. The form $(x, y) = \langle Sx, y \rangle$ is positive semidefinite, and thus satisfies the CBS inequality, yielding, for all $x,y$, $$|\langle Sx, y \rangle|^2 \leq \langle Sx, x \rangle \langle Sy, y \rangle \leq \langle Sx, x \rangle M \|y\|^2 \leq M^2 \|x\|^2 \|y\|^2.$$ Setting $y = Sx$ and dividing through by $\|Sx\|^2$, we obtain, for all $x$, $$\langle S^2x, x \rangle = \langle Sx, Sx \rangle = \|Sx\|^2 \leq M \langle Sx, x \rangle \leq M^2 \|x\|^2.$$ Since $0 \leq M \leq \|S\|$, we conclude:

**XV.10.5.15.** Proposition. Let $S$ be a positive operator, and let $m$ and $M$ be the inf and sup of $W(S)$ respectively. Then

(i) $\|S\| = M$ and $S^2 \leq MS$.

(ii) If $S$ is bounded below, then $m > 0$.

[For (ii), if $\|Sx\| \geq \epsilon \|x\|$ for all $x$, then $M > 0$, and for $\|x\| = 1$ we have $$\epsilon^2 = \epsilon^2 \|x\|^2 \leq \langle Sx, Sx \rangle \leq M \langle Sx, x \rangle$$ so $m \geq \frac{\epsilon^2}{M}$. In fact, $\epsilon = m$ works and this is the largest possible $\epsilon$.]

**XV.10.5.16.** Proposition. If $T$ is a self-adjoint operator on a Hilbert space $H$, and $m$ and $M$ are the inf and sup of $W(T)$ respectively, then $\sigma(T) \subseteq [m, M] \subseteq \mathbb{R}$ and $m, M \in \sigma(T)$. Thus $r_\sigma(T) = w(T)$.

Proof: $\sigma(T) \subseteq [m, M]$ by XV.10.5.9. Note that $mI \leq T \leq MI$. Applying XV.10.5.15.(ii) to $T - mI$ and $MI - T$, we obtain $m, M \in \sigma(T)$.

1830
 XV.10.5.17.  **Proposition.** Let $T$ be a self-adjoint element of $\mathcal{B}(\mathcal{H})$. Then $r_\sigma(T) = \|T\| (= w(T))$.

**Proof:** By scaling, we may assume $r_\sigma(T) = w(T) = 1$, so $-I \leq T \leq I$. Set $S = (T + I)/2$; then $0 \leq S \leq I$. By XV.10.5.15., $\|S\|^2 \leq S$, so $I - T^2 = 4(S - S^2) \geq 0$, $0 \leq T^2 \leq I$, $\|T^2\| = \|T\|^2 \leq 1$ by () and XV.10.5.15. \hfill \Box

 XV.10.5.18.  In particular, if $T \geq 0$, then $\sigma(T) \subseteq [0, \|T\|]$ and $\|T\| \in \sigma(T)$.

 XV.10.5.19.  The result of XV.10.5.17. also follows easily from the spectral radius formula, since if $T$ is self-adjoint we have $\|T^2\| = \|T\|^2$ (), so by iteration $\|T^{2n}\| = \|T\|^{2^n}$ and thus

$$r_\sigma(T) = \lim_{m \to \infty} \|T^m\|^{1/m} = \lim_{n \to \infty} \|T^{2^n}\|^{1/2^n} = \|T\|.$$  

However, the proof in XV.10.5.17. is more elementary, not requiring ()

We can by a similar argument extend XV.10.5.17.:  

 XV.10.5.20.  **Proposition.** Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. If $T$ is normal, then $r_\sigma(T) = w(T) = \|T\|$.

**Proof:** We have $\|T\|^2 = \|T^*T\|$ (), and for each $n$, $\|T^n\|^2 = \|(T^n)^*T^n\| = \|(T^*T)^n\|$ since $T$ is normal. But $T^*T$ is self-adjoint, so $\|(T^*T)^{2n}\| = \|T^*T\|^{2^n} = \|T\|^{2^{n+1}}$. Thus we have

$$r_\sigma(T) = \lim_{n \to \infty} \|T^{2^n}\|^{1/2^n} = \lim_{n \to \infty} \|(T^*T)^{2^n}\|^{1/2^n} = \lim_{n \to \infty} \|T^{2^n}\|^{1/2^{n+1}} = \|T\|.$$  

 XV.10.5.21.  **Proposition.** Let $T$ be a positive operator on a Hilbert space $\mathcal{H}$, and $x, y \in \mathcal{H}$. Then

$$\langle T(x + y), x + y \rangle \leq \|\langle Tx, x \rangle + \langle Ty, y \rangle \|^{1/2}.$$  

**Proof:** We have $\langle T(x + y), x + y \rangle = \langle Tx, x \rangle + 2Re\langle Tx, y \rangle + \langle Ty, y \rangle$, and by the CBS inequality $Re\langle Tx, y \rangle \leq \|\langle Tx, x \rangle + \langle Ty, y \rangle \|^{1/2}$. \hfill \Box

 XV.10.5.22.  **Corollary.** Let $T$ be a positive operator on $\mathcal{H}$, and $\{x_i\}$ a total set of vectors in $\mathcal{H}$. If $\langle Tx_i, x_i \rangle = 0$ for all $i$, then $T = 0$.

 XV.10.5.23.  If $X$ is a compact subset of $\mathbb{C}$ not containing 0, and $k \in \mathbb{N}$, there is in general no bound on the norm of $T^{-1}$ as $T$ ranges over all operators with $\|T\| \leq k$ and $\sigma(T) \subseteq X$. For example, let $S_n \in \mathcal{B}(l^2)$ be the truncated shift:

$$S_n(\alpha_1, \alpha_2, \ldots) = (0, \alpha_1, \alpha_2, \ldots, \alpha_n, 0, 0, \ldots)$$

and let $T_n = I - S_n$, $\|S_n\| = 1$, so $\|T_n\| \leq 2$ for all $n$. Since $S_n$ is nilpotent, $\sigma(S_n) = \{0\}$, so $\sigma(T_n) = \{1\}$ for all $n$. $T_n$ is invertible, with $T_n^{-1} = I + S_n + S_n^2 + \cdots + S_n^n$, and $\|T_n^{-1}\| = \sqrt{n + 1}$, so $\|T_n^{-1}\| \geq \sqrt{n + 1}$.

If $T$ is restricted to normal operators, then there is a bound: if $T$ is an invertible normal operator, then $\|T^{-1}\| = d^{-1}$, where $d$ is the distance from 0 to $\sigma(T)$. This follows immediately from II.1.6.3 and the elementary fact that $\sigma(T^{-1}) = \{\lambda^{-1} : \lambda \in \sigma(T)\}$ (cf. II.1.5.2(i)).

1831
XV.10.6. Exercises

(a) If $A$ is a unital normed algebra, show that $\|1\| \geq 1$. [1 = 1 · 1.]

(1) Show that $\|x\| = \sup_{\|y\|=1} \langle x, y \rangle$ is a convex subset of $XV.10.6.3.$

(b) Give an example of a Banach algebra norm $\|\cdot\|$ on $M_2(\mathbb{C})$ for which $\|I\| > 1$. [Try $\|a_{ij}\| = \sum_{i,j} |a_{ij}|$.

(c) Conclude from (a) and (b) that to show that $\|\cdot\|$ is a Banach algebra norm on $A$ equivalent to $\|\cdot\|$, and that $\|1\| = 1$.

(d) [Kat04, VIII.1.4] Show, more generally, that if $A$ is a Banach space with norm $\|\cdot\|$ which is a unital algebra under a multiplication which is separately continuous, then the norm $\|\cdot\|$ defined in (c) is a Banach algebra norm on $A$ equivalent to $\|\cdot\|$, with $\|1\| = 1$. [First observe that $\|a\| \geq \|a\|_{\text{op}}$ for any $a$. Then show the map $a \mapsto L_a$ from $A$ to $B(A)$ has closed graph. Conclude that $(A, \|\cdot\|)$ is complete and (using the Closed Graph Theorem) that $\|\cdot\|$ is equivalent to $\|\cdot\|.$]

(e) What if $A$ is nonunital? See XV.14.1.2. for a nonunital version.

XV.10.6.2. This problem gives a proof of the Hausdorff-Toeplitz Theorem:

THEOREM. Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not necessarily bounded). Then $W(T)$ is a convex subset of $\mathbb{C}$.

(a) Show that $W(\alpha T + \beta I) = \{\alpha \lambda + \beta : \lambda \in W(T)\}$ for $\alpha, \beta \in \mathbb{C}$, and $\langle Tx, x \rangle = \langle T(\lambda x), \lambda x \rangle$ for $|\lambda| = 1$.

(b) If $\alpha$ and $\beta$ are complex numbers, show that there is a $\lambda \in \mathbb{C}$, $|\lambda| = 1$, with $\lambda \alpha + \lambda \beta$ a nonnegative real number. [Consider $\frac{\alpha - \beta}{|\alpha - \beta|}$ if $\beta \neq \bar{\alpha}$.] Conclude that if $x, y \in \mathcal{V}$, there is a $\lambda \in \mathbb{C}$, $|\lambda| = 1$, such that $\langle T(\lambda x), y \rangle + \langle T y, \lambda x \rangle$ is a nonnegative real number.

(c) Conclude from (a) and (b) that to show that $W(T)$ is convex for all $T$, it suffices to show that if $\langle Tx, x \rangle = 0$ and $\langle Ty, y \rangle = 1$ for $x, y \in \mathcal{V}$, $\|x\| = \|y\| = 1$, $(Tx, x) + (Ty, y) \geq 0$, then for any $\lambda \in [0, 1]$ there is a $z \in \mathcal{V}$, $\|z\| = 1$, with $\langle Tx, z \rangle = \lambda$.

(d) Let $x, y \in \mathcal{V}$ be as in (c). Show that if $0 \leq \alpha \leq 1$, then $(1 - \alpha)x + \alpha y \neq 0$.

(e) For $0 \leq \alpha \leq 1$, set

$$z_{\alpha} = \frac{(1 - \alpha)x + \alpha y}{\| (1 - \alpha)x + \alpha y \|}$$

and $f(\alpha) = \langle Tz_{\alpha}, z_{\alpha} \rangle$. Show that $f$ is a real-valued continuous function on $[0, 1]$.

(f) Since $f(0) = 0$ and $f(1) = 1$, $f$ takes all values in $[0, 1]$ by the Intermediate Value Theorem.

XV.10.6.3. (a) Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not assumed bounded). If $w(T) = 0$, i.e. if $\langle Tx, x \rangle = 0$ for all $x \in \mathcal{V}$, show that $T = 0$. [If $x, y \in \mathcal{V}$, expand $\langle T(x + y), x + y \rangle$ and $T(x + iy), x + iy \rangle$ and conclude that $\langle Tx, y \rangle = 0$.

(b) Show that the result in (a) is not necessarily true in a real inner product space. [Consider a 90° rotation in $\mathbb{R}^2$.]

(c) Let $\mathcal{H}$ be a Hilbert space. If $S, T \in \mathcal{B}(\mathcal{H})$, show that $w(S + T) \leq w(S) + w(T)$. Thus $w$ is a norm on $\mathcal{B}(\mathcal{H})$. Is it equivalent to the operator norm? Is $\mathcal{B}(\mathcal{H})$ complete with respect to the norm $w$?
XV.10.6.4. Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not necessarily bounded). Suppose $W(T) \subseteq \mathbb{R}$, i.e. $\langle Tx, x \rangle \in \mathbb{R}$ for all $x \in \mathcal{V}$. Show that $\langle Tx, y \rangle = \langle x, Ty \rangle$ for all $x, y \in \mathcal{V}$. [As in XV.10.6.3., expand $\langle T(x+y), x+y \rangle$ and $\langle T(x+iy), x+iy \rangle$.]

XV.10.6.5. Prove the following nonlinear version of XV.10.2.9.:

Theorem. Let $X$ be a Banach space, and $f : X \to X$ a strict contraction (a Lipschitz function with Lipschitz constant $< 1$). Let $i : X \to X$ be the identity function. Then $g = i - f$ is a homeomorphism from $X$ onto $X$.

[Reduce to the case $f(0) = 0$. Set $f^0 = i$ and $f^k = f \circ f \circ \cdots \circ f$ ($k$ times) for $k > 0$, and $h_n = \sum_{k=0}^{n} f^k$.

Show that $(h_n)$ converges pointwise on $X$ to an inverse for $g$.]

XV.10.6.6. (a) Let $T \in \mathcal{B}(\mathbb{R}^2)$ have matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

with respect to the standard basis. Show that

$$\|T\| = \|A\| = \sqrt{\frac{a^2 + b^2 + c^2 + d^2 + \sqrt{a^4 + 6a^2b^2 - 2a^2c^2 - 2a^2d^2 + 8abcd + b^4 - 2b^2c^2 - 2b^2d^2 + c^4 + 6c^2d^2 + d^4}}{2}}.$$

[Use that $\|A\|$ is the square root of the largest eigenvalue of $A^tA$.]

(b) Find a similar formula if $T \in \mathcal{B}(\mathbb{C}^2)$, i.e. $a, b, c, d \in \mathbb{C}$.

(c) (For masochists only, or for someone adept at using computer algebra programs) Find a similar formula for $3 \times 3$ and $4 \times 4$ matrices.

(d) Show that if $n \geq 5$, there is no algebraic formula for the norm of an $n \times n$ matrix $A$ in terms of the entries of $A$. Specifically, show that the norm of the matrix

$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which in this case is just the largest eigenvalue, approximately 3.917, cannot be written in terms of radicals and rational numbers, i.e. its characteristic polynomial is not solvable by radicals over $\mathbb{Q}$ (D. Zare; cf. https://mathoverflow.net/questions/268453/is-there-an-algebraic-formula-for-the-eigenvalues-of-a-symmetric-n).

For larger $n$ just add rows and columns of 0’s.

The formula in (a) is not for serious use; it is only to indicate that even in the $2 \times 2$ case, the norm of a matrix or linear transformation does not have a simple algebraic formula.
**XV.11. Fredholm Operators and Fredholm Index**

Fredholm operators are a very important class of “almost invertible” operators. To each Fredholm operator there is associated an integer “index” which roughly speaking measures the extent to which the operator fails to be invertible (or, more precisely, to have a “small” invertible perturbation). Arguably the most important fact about Fredholm operators is that a compact perturbation of a Fredholm operator is also Fredholm, with the same index; in particular, any compact perturbation of an invertible operator is Fredholm, with index 0.

The name “Fredholm operator” is in honor of Swedish mathematician I. Fredholm. It is argued in [Lax02] that there is “no historical justification for this,” and that the notions of Fredholm operator and Fredholm index should be attributed to Fritz Noether (son of Max Noether and brother of Emmy Noether). However, Fredholm, who worked around the turn of the 20th century, before Banach spaces were defined, developed some of the crucial basic properties of such operators in his work on integral equations (nicely described in [Lax02]), although he did indeed only consider operators of index 0. The names “Fredholm operator” and “Fredholm index” are also well established so we will use them instead of the names “Noether operator” and “Noether index” used in [Lax02].

The theory of Fredholm operators and Fredholm index is not only important for applications, but (in retrospect) was a pioneering example of bringing algebraic topology into the theory of operator algebras. In fact, it was one of the first examples of the process called “noncommutative topology,” which underlies much of the modern theory of operator algebras. See the introduction to [Con94] for an excellent survey of noncommutative topology and geometry.

**XV.11.1. Banach Space Preliminaries**

The first elementary result uses the quotient norm: if \( \mathcal{Y} \) is a closed subspace of a normed vector space \( \mathcal{X} \), the **quotient norm** on \( \mathcal{X}/\mathcal{Y} \) is the norm \( \|x + \mathcal{Y}\| = \inf \{\|x - y\| : y \in \mathcal{Y}\} = \inf \{\|z\| : z + \mathcal{Y} = x + \mathcal{Y}\} \).

**Proposition.** Let \( \mathcal{Y} \) be a proper closed subspace of a normed vector space \( \mathcal{X} \), and \( \epsilon > 0 \). Then there is a \( x \in \mathcal{X} \) with \( \|x\| \leq 1 + \epsilon \) and \( \|x - y\| \geq 1 \) for all \( y \in \mathcal{Y} \).

**Proof:** Let \( x \) be a suitable preimage of a unit vector in \( \mathcal{X}/\mathcal{Y} \).

Next we obtain a useful fact which is a simple consequence of the Hahn-Banach Theorem:

**Proposition.** Let \( \mathcal{X} \) be a normed vector space, and \( \mathcal{Y} \) a finite-dimensional subspace (which is automatically closed.) Then \( \mathcal{Y} \) has a closed complementary subspace \( \mathcal{Z} \) (i.e. \( \mathcal{Y} + \mathcal{Z} = \mathcal{X} \) and \( \mathcal{Y} \cap \mathcal{Z} = \{0\} \).) If \( \mathcal{R} \) is a closed subspace of \( \mathcal{X} \) with \( \mathcal{Y} \cap \mathcal{R} = \{0\} \), then \( \mathcal{Z} \) may be chosen to contain \( \mathcal{R} \).

**Proof:** Let \( \{x_1, \ldots, x_n\} \) be a basis for \( \mathcal{Y} \), and define linear functionals \( \phi_i \) (1 \( \leq i \leq n \)) on \( \mathcal{Y} \) by \( \phi_i(\sum_{j=1}^n \alpha_j x_j) = \alpha_i \) (i.e. \( \phi_i(x_j) = \delta_{ij} \)). Then each \( \phi_i \) is bounded since \( \mathcal{Y} \) is finite-dimensional. If \( \mathcal{W} \) is the subspace of \( \mathcal{X} \) spanned by \( \mathcal{Y} \) and \( \mathcal{R} \), then \( \mathcal{W} \) is closed (the image of \( \mathcal{W} \) in \( \mathcal{X}/\mathcal{R} \) is the same as the image of \( \mathcal{Y} \), hence is finite-dimensional and therefore closed), so \( \mathcal{W} = \mathcal{Y} \oplus \mathcal{R} \). Extend \( \phi_i \) to \( \mathcal{W} \) by setting it equal to 0 on \( \mathcal{R} \). Then extend \( \phi_i \) to a bounded linear functional \( \psi_i \) on \( \mathcal{X} \) by the Hahn-Banach Theorem. Define \( P \in B(\mathcal{X}) \) by \( Py = \sum_{i=1}^n \psi_i(y) x_i \). Then \( \mathcal{R}(P) = \mathcal{Y} \), and \( P|_\mathcal{Y} \) is the identity (\( P \) is a projection onto \( \mathcal{Y} \).) Then \( \mathcal{Z} = N(P) \) is a closed subspace of \( \mathcal{X} \) complementary to \( \mathcal{Y} \), containing \( \mathcal{R} \).

We recall the definitions of some standard terms from operator theory which will be used in this section.
XV.11.1.3. Definition. Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces, and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. Then $T$ is bounded below if there is an $\epsilon > 0$ such that $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{X}$.

An operator which is bounded below is one-to-one, and its range is closed if $\mathcal{X}$ is a Banach space.

XV.11.1.4. Definition. An operator $T \in \mathcal{B}(\mathcal{X})$ has finite ascent $n$ if the chain $\mathcal{N}(T) \subseteq \mathcal{N}(T^2) \subseteq \cdots$ stabilizes at $n$, i.e. $\mathcal{N}(T^n) = \mathcal{N}(T^{n+1}) = \cdots$.

$T$ has finite descent $m$ if the chain $\mathcal{R}(T) \supseteq \mathcal{R}(T^2) \supseteq \cdots$ stabilizes at $m$.

XV.11.1.5. Note that if $\mathcal{N}(T^n) = \mathcal{N}(T^{n+1})$, then $\mathcal{N}(T^n) = \mathcal{N}(T^{n+m})$ for all $m$; if $x \in \mathcal{N}(T^{r+1}) \setminus \mathcal{N}(T^r)$ for some $r > n$, then $T^{r-n}x \in \mathcal{N}(T^{n+1}) \setminus \mathcal{N}(T^n)$. Similarly, if $\mathcal{R}(T^n) = \mathcal{R}(T^{n+1})$, then $\mathcal{R}(T^n) = \mathcal{R}(T^{n+m})$ for all $m$.

XV.11.1.6. It is fairly exceptional for an operator to have finite ascent or descent, but note that an injective operator has finite ascent 1, and a surjective operator finite descent 1. Every operator on a finite-dimensional space has finite ascent and descent. A normal operator $T$ on a Hilbert space has finite ascent 1, but not necessarily finite descent, even if it is compact (Exercise XV.11.6.1.).

XV.11.1.7. Definition. If $T : \mathcal{X} \to \mathcal{Y}$ is a bounded operator, the adjoint of $T$ is the operator $T^* : \mathcal{Y}^* \to \mathcal{X}^*$ defined by $T^*(\phi) = \phi \circ T$. ($\mathcal{X}^*$ is the usual Banach space dual of $\mathcal{X}$, the space of bounded linear functionals on $\mathcal{X}$.)

XV.11.1.8. $T^* \in \mathcal{B}(\mathcal{Y}^*, \mathcal{X}^*)$ and $\|T^*\| = \|T\|$ by the Hahn-Banach Theorem. If $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces, then $\mathcal{X}^*$ and $\mathcal{Y}^*$ can be identified with $\mathcal{X}$ and $\mathcal{Y}$ in the standard way, and then $T^*$ agrees with the usual adjoint operator defined by the inner product: $\langle Tx, y \rangle = \langle x, T^*y \rangle$ for $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

XV.11.1.9. Proposition. If $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ is finite-rank, then $T^*$ is finite-rank.

Proof: $T$ factors through $\mathcal{R}(T)$, so $T^*$ factors through $\mathcal{R}(T)^*$, which is finite-dimensional. ☑

XV.11.2. Compact Operators

XV.11.2.1. Definition. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. An operator $T : \mathcal{X} \to \mathcal{Y}$ is compact if $T$ sends bounded subsets of $\mathcal{X}$ to precompact subsets of $\mathcal{Y}$.

XV.11.2.2. In other words, if $B$ is a bounded subset of $\mathcal{X}$, then $\overline{T(B)}$ is a compact subset of $\mathcal{Y}$. Equivalently, whenever $(x_n)$ is a bounded sequence of vectors in $\mathcal{X}$, then $(Tx_n)$ has a convergent subsequence. $T$ is compact if and only if $T(\mathcal{B}_X)$ is totally bounded. Since a compact metric space is separable, if $T : \mathcal{X} \to \mathcal{Y}$ is compact, $\overline{\mathcal{R}(T)}$ is a separable subspace of $\mathcal{Y}$.

XV.11.2.3. In many references, particularly older ones, compact operators are called completely continuous operators.
**XV.11.2.4.** It is obvious that a compact operator is bounded, and the composition (in either order) of a compact operator with a bounded operator is compact. It is easily seen using the total boundedness criterion that a (norm-)limit of compact operators is compact. The set of compact operators from $\mathcal{X}$ to $\mathcal{Y}$, denoted $\mathcal{K}(\mathcal{X}, \mathcal{Y})$, is a closed subspace of $\mathcal{B}(\mathcal{X}, \mathcal{Y})$: $\mathcal{K}(\mathcal{X}, \mathcal{Y}) = \mathcal{K}(\mathcal{X}, \mathcal{X})$ is a closed two-sided ideal in $\mathcal{B}(\mathcal{X})$. Every finite-rank bounded operator is compact; for many, but not all, Banach spaces, every compact operator is a limit of finite-rank operators (see the discussion below.)

**XV.11.2.5.** It is easily seen that a bounded operator is weakly continuous (cf. [?]). So if $\mathcal{X}$ is a Hilbert space (or, more generally, a reflexive Banach space), $\mathcal{B}_\mathcal{X}$ is weakly compact, so if $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, $T(\mathcal{B}_\mathcal{X})$ is weakly compact and thus (norm-)closed. In particular, if $T \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$, then $T(\mathcal{B}_\mathcal{X})$ is already closed and therefore compact.

**XV.11.2.6.** Proposition. If $T \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$, then $T^* \in \mathcal{K}(\mathcal{Y}^*, \mathcal{X}^*)$.

**Proof:** Let $\epsilon > 0$. Since $T(\mathcal{B}_\mathcal{X})$ is totally bounded, there are $x_1, \ldots, x_n \in \mathcal{B}_\mathcal{X}$ such that, for each $x \in \mathcal{B}_\mathcal{X}$, there is an $i$ with $\|Tx - Tx_i\| < \epsilon/3$. Give $\mathbb{F}^n$ the infinity norm, i.e. $\| (\alpha_1, \ldots, \alpha_n) \| = \max |\alpha_i|$, and define $S : \mathcal{Y}^* \to \mathbb{F}^n$ by $S\phi = (\phi(Tx_1), \ldots, \phi(Tx_n))$. $S$ is finite-rank and bounded and hence compact, so $S(\mathcal{B}_{\mathcal{Y}^*})$ is totally bounded; let $\phi_1, \ldots, \phi_m \in \mathcal{B}_{\mathcal{Y}^*}$ be such that, for every $\phi \in \mathcal{B}_{\mathcal{Y}^*}$, there is a $j$ such that $\| S\phi - S\phi_j \| < \epsilon/3$. If $\phi \in \mathcal{B}_{\mathcal{Y}^*}$, choose a corresponding $\phi_j$; then, for any $x \in \mathcal{B}_\mathcal{X}$, with corresponding $x_i$,

$$\| T^*\phi(x) - T^*\phi_j(x) \| = \| \phi(Tx) - \phi_j(Tx) \|$$

$$\leq \| \phi(Tx) - \phi(Tx_i) \| + \| \phi_j(Tx_i) - \phi_j(Tx) \|$$

$$\leq \| Tx - Tx_i \| = \| S\phi - S\phi_j \| + \| Tx_i - Tx \| < \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon .$$

Thus $\| T^*\phi - T^*\phi_j \| \leq \epsilon$, so $T^*(\mathcal{B}_{\mathcal{Y}^*})$ is totally bounded. ✷

While the proof of XV.11.2.6. is not difficult, there is an alternate approach (of comparable difficulty) on a Hilbert space which is worth noting:

**XV.11.2.7.** Theorem. If $\mathcal{Y}$ is a Hilbert space, then $\mathcal{K}(\mathcal{X}, \mathcal{Y})$ is the closure of the finite-rank bounded operators from $\mathcal{X}$ to $\mathcal{Y}$.

**Proof:** We again use a total boundedness argument. Let $T \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$, and $\epsilon > 0$. Choose $x_1, \ldots, x_n \in \mathcal{B}_\mathcal{X}$ such that, for any $x \in \mathcal{B}_\mathcal{X}$, there is an $i$ with $\| Tx - Tx_i \| < \epsilon/2$. Let $P \in \mathcal{B}(\mathcal{Y})$ be the orthogonal projection onto the subspace spanned by $\{Tx_1, \ldots, Tx_n\}$. Then $P$ is finite-rank, $PTx_i = Tx_i$ for all $i$, and $\|P\| = 1$. If $x \in \mathcal{B}_\mathcal{X}$, and a corresponding $x_i$ is chosen, then

$$\| PTx - Tx \| \leq \| PTx - PTx_i \| + \| Tx_i - Tx \| < \epsilon/2 + \epsilon/2 = \epsilon .$$

Thus $\| PT - T \| \leq \epsilon$, and $PT$ has finite rank. ✷

**XV.11.2.8.** If $T \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{Y}$ is a Hilbert space, an alternate proof of XV.11.2.6. is then obtained by taking $(T_n)$ of finite rank and $T_n \to T$; then $T_n^* \to T^*$, and $T_n^*$ is finite-rank by XV.11.1.9. Similarly, if $T \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$ and $\mathcal{X}$ is a Hilbert space, apply XV.11.2.7. to $T^* \in \mathcal{K}(\mathcal{Y}^*, \mathcal{X})$ (using XV.11.2.6.) to conclude that $\mathcal{K}(\mathcal{X}, \mathcal{Y})$ is also the closure of the finite-rank bounded operators from $\mathcal{X}$ to $\mathcal{Y}$ (there is also a simple direct proof.)

1836
XV.11.2.9. Definition. A Banach space $\mathcal{X}$ has the approximation property if every compact operator on $\mathcal{X}$ is a (norm-)limit of finite-rank (bounded) operators on $\mathcal{X}$.

XV.11.2.10. XV.11.2.7. shows that Hilbert spaces have the approximation property. All standard Banach spaces have the approximation property. P. Enflo [Enf73] constructed the first example of a Banach space without the approximation property; see [Jam82] for a discussion of this and related matters, and [Gow95] for more recent developments.

Strictly Singular Operators
There is a slightly larger class of operators which to some extent behave like compact operators:

XV.11.2.11. Definition. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, with $\mathcal{X}$ infinite-dimensional. An operator $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ is strictly singular if $T$ is not bounded below on any infinite-dimensional subspace of $\mathcal{X}$.

XV.11.2.12. It follows from () that every compact operator is strictly singular. Every strictly singular operator on a Hilbert space, or many standard Banach spaces, is compact; but there are operators on some Banach spaces which are strictly singular but not compact (XV.11.6.2.).

XV.11.3. Fredholm Operators

XV.11.3.1. Definition. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. $T$ is a Fredholm operator if $\mathcal{N}(T)$ is finite-dimensional and $\mathcal{R}(T)$ has finite codimension. The index of $T$ is $\dim \mathcal{N}(T) - \text{codim} \mathcal{R}(T)$. Denote the set of Fredholm operators from $\mathcal{X}$ to $\mathcal{Y}$ by $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, and write $\mathcal{F}(\mathcal{X})$ for $\mathcal{F}(\mathcal{X}, \mathcal{X})$.

XV.11.3.2. Examples.

(i) If $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional, then every operator from $\mathcal{X}$ to $\mathcal{Y}$ is Fredholm with index $\dim \mathcal{X} - \dim \mathcal{Y}$.

(ii) If $T$ is invertible, then $T$ is Fredholm with index 0.

(iii) Let $S$ be the unilateral shift on the Hilbert space $l^2$ of square-summable sequences, defined by $S(\alpha_1, \alpha_2, \ldots) = (0, \alpha_1, \alpha_2, \ldots)$. Then $S$ is Fredholm with index $-1$. $S^n$ is Fredholm with index $-n$, and $(S^n)^n$ is Fredholm with index $n$.

XV.11.3.3. Definition XV.11.3.1. is not the most common one, which in addition requires $\mathcal{R}(T)$ to be closed. However, the two definitions are equivalent, since a Fredholm operator in the sense of XV.11.3.1. always has closed range. Although strictly speaking we will not need this fact for our development of the theory, we give a proof for completeness (and to show that we are not redefining the notion of Fredholm operator), and also because the argument gives some important structural properties of Fredholm operators.
**XV.11.3.4. Proposition.** Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then $\mathcal{R}(T)$ is closed.

**Proof:** Let $W$ be a complementary subspace for $\mathcal{R}(T)$ in $\mathcal{Y}$; then $W$ is finite-dimensional and hence closed. Let $Z$ be a closed subspace of $\mathcal{X}$ complementary to $\mathcal{N}(T)$ (XV.11.1.2.). Then $\mathcal{R}(T|_Z) = \mathcal{R}(T)$. Define a bounded operator $S$ from the Banach space $W \oplus Z$ to $\mathcal{Y}$ by $S(w, z) = w + Tz$. Then $S$ is a bijection, hence a homeomorphism by the Open Mapping Theorem, and $S(\{0\} \oplus Z) = \mathcal{R}(T)$. 

A corollary of the proof gives a simple picture of a Fredholm operator:

**XV.11.3.5. Corollary.** Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces, and $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there are closed subspaces $Z, W$ of $\mathcal{X}$ and $\mathcal{Y}$ complementary to $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively, such that $\mathcal{X} \cong \mathcal{N}(T) \oplus Z$, $\mathcal{Y} \cong W \oplus \mathcal{R}(T)$, and $T \cong 0 \oplus U$, where $U$ is a linear isomorphism from $Z$ onto $\mathcal{R}(T)$ (i.e. $U$ is bounded below.)

**XV.11.3.6. Corollary.** Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there is an $S \in \mathcal{F}(\mathcal{Y}, \mathcal{X})$, with $\text{index}(S) = -\text{index}(T)$, such that $ST = I_\mathcal{X} - P$ and $TS = I_\mathcal{Y} - Q$, where $P$ and $Q$ are finite-rank bounded idempotent operators.

**Proof:** Under the above identifications, $S$ may be taken as $0 \oplus U^{-1} : W \oplus \mathcal{R}(T) \to \mathcal{N}(T) \oplus Z$. 

**XV.11.3.7.** The operator $S$ is a quasi-inverse for $T$: $T = TST$. (If $S$ is as defined in the proof, $T$ is also a quasi-inverse for $S$.)

We obtain another characterization of Fredholm operators from XV.11.3.5.:

**XV.11.3.8. Corollary.** Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces, and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. Then $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ if and only if there is a closed subspace $Z$ of $\mathcal{X}$, of finite codimension, with $T(Z)$ of finite codimension in $\mathcal{Y}$, such that $T$ is bounded below on $Z$. If there is such a $Z$, then $\text{index}(T) = \text{codim}(Z) - \text{codim}T(Z))$.

**XV.11.3.9. Corollary.** Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces, and $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there is an $\epsilon > 0$ such that $T + S \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ and $\text{index}(T + S) = \text{index}(T)$ for all $S \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, $\|S\| < \epsilon$.

**Proof:** Let $Z$ be a closed subspace of finite codimension in $\mathcal{X}$ such that $T$ is bounded below on $Z$. Then, if $\|S\|$ is sufficiently small, $T + S$ is bounded below on $Z$ and $(T + S)(Z)$ is a small enough perturbation of $T(Z)$ to have the same codimension ( ).

Another important consequence of XV.11.3.5. is:
**XV.11.3.10.** Corollary. Let \( T \in \mathcal{F}(\mathcal{X}, \mathcal{Y}) \). Then \( T^* \in \mathcal{F}(\mathcal{Y}^*, \mathcal{X}^*) \), \( \dim(\mathcal{N}(T^*)) = \text{codim}(\mathcal{R}(T)) \), \( \text{codim}(\mathcal{R}(T^*)) = \dim(\mathcal{N}(T)) \), and \( \text{index}(T^*) = -\text{index}(T) \).

**Proof:** If \( T = 0 \oplus U : \mathcal{N}(T) \oplus \mathcal{Z} \to \mathcal{W} \oplus \mathcal{R}(T) \), then \( T^* = 0 \oplus U^* : \mathcal{W}^* \oplus \mathcal{R}(T)^* \to \mathcal{N}(T)^* \oplus \mathcal{Z}^* \), and \( U^* \) is invertible; thus \( \mathcal{N}(T^*) = \mathcal{W}^* \oplus \{0\} \) and \( \mathcal{R}(T^*) = \{0\} \oplus \mathcal{Z}^* \). We have \( \mathcal{W}^* \cong \mathcal{W} \) and \( \mathcal{N}(T)^* \cong \mathcal{N}(T) \) since they are finite-dimensional. \( \blacksquare \)

**XV.11.3.11.** Actually, if \( T \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \), then \( \mathcal{N}(T^*) = \mathcal{R}(T)^\perp = \{\phi \in \mathcal{Y}^* : \phi|_{\mathcal{R}(T)} = 0\} \cong (\mathcal{Y}/\mathcal{R}(T))^* \).

Thus \( T \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \) is Fredholm if and only if \( \mathcal{R}(T) \) is closed and \( \mathcal{N}(T) \) and \( \mathcal{N}(T^*) \) are finite-dimensional, and \( \text{index}(T) = \dim \mathcal{N}(T) - \dim \mathcal{N}(T^*) \).

It is useful not to require that the range of a Fredholm operator be closed as part of the definition. The next proof is an example.

**XV.11.3.12.** Proposition. Let \( \mathcal{X} \) and \( \mathcal{Y} \) be Banach spaces, \( T \in \mathcal{B}(\mathcal{X}, \mathcal{Y}) \), \( S \in \mathcal{B}(\mathcal{Y}, \mathcal{X}) \). If \( ST \) and \( TS \) are Fredholm, then \( S \) and \( T \) are Fredholm.

**Proof:** We show \( T \) is Fredholm. \( \mathcal{N}(T) \subseteq \mathcal{N}(ST) \), which is finite-dimensional, and \( \mathcal{R}(T) \supseteq \mathcal{R}(TS) \), which has finite codimension. \( \blacksquare \)

Conversely:

**XV.11.3.13.** Theorem. Let \( \mathcal{X}, \mathcal{Y}, \mathcal{Z} \) be Banach spaces, and \( T \in \mathcal{F}(\mathcal{X}, \mathcal{Y}) \) and \( S \in \mathcal{F}(\mathcal{Y}, \mathcal{Z}) \). Then \( ST \in \mathcal{F}(\mathcal{X}, \mathcal{Z}) \) and \( \text{index}(ST) = \text{index}(S) + \text{index}(T) \).

**Proof:** Let \( \mathcal{U} \) be a closed subspace of \( \mathcal{X} \) complementary to \( \mathcal{N}(T) \) on which \( T \) is bounded below. Then \( U = T|_{\mathcal{U}} \) is an isomorphism from \( \mathcal{U} \) onto \( \mathcal{R}(T) \). Set \( \mathcal{X}_2 = \mathcal{R}(T) \cap \mathcal{N}(S) \) and \( \mathcal{X}_2 = U^{-1}(\mathcal{X}_2) \). Then \( \mathcal{X}_2 \) is a finite-dimensional subspace of \( \mathcal{U} \). Let \( \mathcal{X}_1 \) be a closed subspace of \( \mathcal{U} \) complementary to \( \mathcal{X}_2 \), and let \( \mathcal{Y}_1 = U(\mathcal{X}_1) = T(\mathcal{X}_1) \). Then \( \mathcal{Y}_1 \) is a closed subspace of \( \mathcal{Y} \) with zero intersection with \( \mathcal{N}(S) \), and \( \mathcal{Y}_1 \) and \( \mathcal{Y}_2 \) span \( \mathcal{R}(T) \), and since \( \mathcal{Y}_2 \) is finite-dimensional and \( \mathcal{R}(T) \) has finite codimension in \( \mathcal{Y} \), \( \mathcal{Y}_1 \) has finite codimension in \( \mathcal{Y} \). Let \( \mathcal{V} \) be a closed subspace of \( \mathcal{Y} \) containing \( \mathcal{Y}_1 \) which is complementary to \( \mathcal{N}(S) \) (XV.11.1.2.), and \( \mathcal{Y}_3 \) a subspace of \( \mathcal{V} \) complementary to \( \mathcal{Y}_1 \). Then \( \mathcal{Y}_3 \) is finite-dimensional, and we have

\[
\mathcal{Y} \cong \mathcal{Y}_1 \oplus \mathcal{Y}_2 \oplus \mathcal{Y}_3 \oplus \mathcal{Y}_4
\]

where \( \mathcal{Y}_4 \) is a subspace of \( \mathcal{N}(S) \) complementary to \( \mathcal{Y}_2 \). Then \( \mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4 \) are finite-dimensional, \( \mathcal{Y}_2 \) and \( \mathcal{Y}_4 \) span \( \mathcal{N}(S) \), and \( \mathcal{Y}_1 \) and \( \mathcal{Y}_2 \) span \( \mathcal{R}(T) \). Then \( S \) is bounded below on \( \mathcal{Y} \) (cf. the proof of XV.11.3.4.), and \( V = S|_{\mathcal{Y}} \) maps \( \mathcal{V} \) isomorphically onto \( \mathcal{R}(S) \). Also, \( UV = ST|_{\mathcal{X}_1} \) maps \( \mathcal{X}_1 \) isomorphically onto \( \mathcal{Z}_1 = V(\mathcal{Y}_1) = S(\mathcal{Y}_1) \). Since \( \mathcal{Y}_1 \) has finite codimension in \( \mathcal{Y} \) and \( \mathcal{R}(S) \) has finite codimension in \( \mathcal{Z} \), \( \mathcal{Z}_1 \) has finite codimension in \( \mathcal{Z} \) and \( ST \) is Fredholm.

We have

\[
\mathcal{X} \cong \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \mathcal{X}_3
\]

\[
\mathcal{Z} \cong \mathcal{Z}_1 \oplus \mathcal{Z}_2 \oplus \mathcal{Z}_3
\]
where \( X_3 = N(T), Z_3 = V(Y_3) = S(Y_3), \) and \( Z_2 \) is a subspace of \( Z \) complementary to \( R(S) \). Set \( a_k = \text{dim}(X_k), b_k = \text{dim}(Y_k), \) and \( c_k = \text{dim}(Z_k) \). We then have
\[
\text{index}(T) = a_3 - (b_3 + b_4) \\
\text{index}(S) = (b_2 + b_4) - c_2 \\
\text{index}(ST) = (a_2 + a_3) - (c_2 + c_3).
\]
But \( b_2 = a_2 \) and \( c_3 = b_3 \) since \( U \) and \( V \) are isomorphisms, so we have
\[
\text{index}(ST) = b_2 + a_3 - b_3 - c_2 = \text{index}(S) + \text{index}(T).
\]

**XV.11.4. Fredholm Operators of Index 0 and the Fredholm Alternative**

We first note an immediate consequence of **XV.11.3.5.**:

**XV.11.4.1. Proposition.** Let \( T \in \mathcal{F}(X, Y) \) have index 0. Then there are \( V, F \in \mathcal{B}(X, Y) \) with \( V \) invertible and \( F \) finite-rank, with \( T = V + F \).

The converse is also true (**XV.11.5.12.**).

**Proof:** Under the identifications \( X \cong N(T) \oplus Z, Y \cong W \oplus R(T), T \cong 0 \oplus U, \) we have that \( \text{dim}(N(T)) = \text{dim}(W) < \infty, \) so there is a (necessarily bounded) linear isomorphism \( S : N(T) \to W. \) Set \( V = S \oplus U, F = -S \oplus 0. \)

The next observation is called the **Fredholm Alternative.** The last statement follows from **XV.11.3.10.**

**XV.11.4.2. Proposition.** Let \( T \in \mathcal{F}(X, Y) \) have index 0. Then \( T \) is injective if and only if \( T \) is surjective. So either the equation \( Tx = y \) has a unique solution for each \( y \in Y, \) or the homogeneous equation \( Tx = 0 \) has a nontrivial (but finite-dimensional) space of solutions.

Furthermore, the equation \( T^* \phi = \psi \) has a unique solution for each \( \psi \) if and only if \( Tx = y \) has a unique solution for each \( y, \) and the solution set of \( T^* \phi = 0 \) has the same (finite) dimension as the solution set of \( Tx = 0. \)

**XV.11.4.3.** The result is sometimes paraphrased: “If the solution to \( Tx = y \) is unique, then it exists.”

**XV.11.5. Relation between Compact and Fredholm Operators**

The next theorem is the main result of this section.

**XV.11.5.1. Theorem.** Let \( X \) be a Banach space, \( T \in \mathcal{K}(X). \) Then \( I + T \) is Fredholm, of index 0.

The proof that \( I + T \) is Fredholm consists of the next three lemmas.
**XV.11.5.2. Lemma.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{N}(I+T)$ is finite-dimensional.

Proof: $T = -I$ on $\mathcal{N}(I+T)$, and thus the unit ball of $\mathcal{N}(I+T)$ is compact. 

**XV.11.5.3. Lemma.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{R}(I+T)$ is closed.

Proof: Let $Z$ be a closed subspace of $\mathcal{X}$ complementary to $\mathcal{N}(I+T)$ (XV.11.1.2). It suffices to show that $I+T$ is bounded below on $Z$; then $\mathcal{R}(I+T) = \mathcal{R}((I+T)|_{Z})$ is closed. Suppose $(x_{n})$ is a sequence of unit vectors in $Z$ with $\|(I+T)x_{n}\| \to 0$. Passing to a subsequence we may assume that $Tx_{n} \to y$ for some $y$, since $T$ is compact. But $(I+T)x_{n} \to 0$, so $x_{n} \to -y$, and thus $\|y\| = 1$ and $Ty = -y$, i.e. $(I+T)y = 0$, $y \in \mathcal{N}(I+T)$. But $y \in Z$ since $-x_{n} \to y$ and $Z$ is closed; hence $y \in Z \cap \mathcal{N}(I+T) = \{0\}$, a contradiction. 

**XV.11.5.4. Lemma.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{R}(I+T)$ has finite codimension.

Proof: Consider the Banach space $\mathcal{Y} = \mathcal{X}/\mathcal{R}(I+T)$. We show that $\mathcal{Y}$ is finite-dimensional by showing that the dual space $\mathcal{Y}^{*}$ is finite-dimensional. $\mathcal{Y}^{*}$ can be identified with the subspace of $\mathcal{X}^{*}$ consisting of functionals vanishing on $\mathcal{R}(I+T)$. But $\mathcal{Y}^{*}$ is precisely $\mathcal{N}((I+T)^{*}) = \mathcal{N}(I+T^{*})$. $T^{*} \in \mathcal{K}(\mathcal{X}^{*})$ by XV.11.2.6, so $\mathcal{N}(I+T^{*})$ is finite-dimensional by XV.11.5.2.

It remains to be shown that $\text{index}(I+T) = 0$. Our proof is partly based in spirit on [RSN90, §70] (repeated without attribution in the rather dreadful paper [Ram01]). We first obtain another important property of $I+T$ which is a simple (but nonobvious) consequence of the compactness of $T$.

**XV.11.5.5. Proposition.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $I+T$ has finite ascent and finite descent.

Proof: If $I+T$ does not have finite ascent, by XV.11.1, choose for each $n$ a vector $x_{n} \in \mathcal{N}((I+T)^{n})$ with $\|x_{n}\| \leq 2$ and $\|x_{n} - y\| \geq 1$ for all $y \in \mathcal{N}((I+T)^{n-1})$. Then, for $m < n$,

$$\|Tx_{m} - Tx_{n}\| = \|x_{n} - (I+T)x_{n} + (I+T)x_{m} - x_{m}\| \geq 1$$

since $x_{m} - (I+T)x_{m} + (I+T)x_{n} \in \mathcal{N}((I+T)^{n-1})$. Thus $(Tx_{n})$ has no convergent subsequence, contradicting compactness of $T$.

The proof that $I+T$ has finite descent is virtually identical. 

1841
**XV.11.5.6.** **Lemma.** If $\mathcal{X}$ is a Banach space and $T \in \mathcal{K}(\mathcal{X})$, then $\text{index}(I + T) = 0$.

**Proof:** Let $n$ be the maximum of the ascent and descent of $I + T$, so $\mathcal{N}((I + T)^n) = \mathcal{N}((I + T)^{n+k})$ and $\mathcal{R}((I + T)^n) = \mathcal{R}((I + T)^{n+k})$ for all $k$. Note that $(I + T)^n = I + S$ for $S \in \mathcal{K}(\mathcal{X})$, so $(I + T)^n$ is Fredholm and hence $\mathcal{R} = \mathcal{R}((I + T)^n)$ is closed and $\mathcal{N} = \mathcal{N}((I + T)^n)$ is finite-dimensional. Then $\mathcal{N} \cap \mathcal{R} = \{0\}$, for if $0 \neq y = (I + T)^nx \in \mathcal{N} \cap \mathcal{R}$, then $x \in \mathcal{N}((I + T)^{2n}) \setminus \mathcal{N}((I + T)^n)$, a contradiction. Thus $(I + T)^n$ maps $\mathcal{R}$ isomorphically onto $\mathcal{R}$ (since $\mathcal{R} = \mathcal{R}((I + T)^{2n})$.) If $\mathcal{Z}$ is a closed subspace of $\mathcal{X}$ containing $\mathcal{R}$ which is complementary to $\mathcal{N}$, then $(I + T)^n$ also maps $\mathcal{Z}$ isomorphically onto $\mathcal{R}$, and it follows that $\mathcal{Z} = \mathcal{R}$, i.e. $\mathcal{N}$ and $\mathcal{R}$ are complementary subspaces of $\mathcal{X}$. Since both $\mathcal{N}$ and $\mathcal{R}$ are invariant under $I + T$ and $(I + T)|_{\mathcal{R}}$ is an isomorphism, $\text{index}(I + T) = \text{index}((I + T)|_{\mathcal{N}})$. But $\mathcal{N}$ is finite-dimensional, so $\text{index}((I + T)|_{\mathcal{N}}) = 0$. \hfill $\Box$

This completes the proof of XV.11.5.1..

The proof of XV.11.5.6. actually shows that $I + T$ has a simple form:

**XV.11.5.7.** **Corollary.** Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then there are complementary closed subspaces $\mathcal{N}$ and $\mathcal{R}$ of $\mathcal{X}$, each invariant under $T$ (and hence under $I + T$), with $\mathcal{N}$ finite-dimensional, such that $(I + T)|_{\mathcal{R}}$ is invertible and $(I + T)|_{\mathcal{N}}$ is nilpotent.

It is well known (and useful) that every operator on a finite-dimensional space has this property.

The next theorem [Atk53] gives a useful alternate characterization of Fredholm operators.

**XV.11.5.8.** **Theorem.** [Atkinson’s Theorem] Let $\mathcal{X}$ be an infinite-dimensional Banach space and $\pi : \mathcal{B}(\mathcal{X}) \to \mathcal{Q}(\mathcal{X})$ the quotient map onto the Calkin algebra $\mathcal{Q}(\mathcal{X}) = \mathcal{B}(\mathcal{X})/\mathcal{K}(\mathcal{X})$ of $\mathcal{X}$. Then $T \in \mathcal{B}(\mathcal{X})$ is Fredholm if and only if $\pi(T)$ is invertible in $\mathcal{Q}(\mathcal{X})$.

**Proof:** If $T$ is Fredholm, let $S$ be a quasi-inverse for $T$ as in XV.11.3.6.; then $\pi(S)$ is an inverse for $\pi(T)$.

Conversely, suppose $\pi(T)$ is invertible in $\mathcal{Q}(\mathcal{X})$, and let $S \in \mathcal{B}(\mathcal{X})$ such that $\pi(S) = \pi(T)^{-1}$. Then $ST = I + K$, $TS = I + L$ for $K, L \in \mathcal{K}(\mathcal{X})$. $I + K$ and $I + L$ are Fredholm by XV.11.5.1., so $S$ and $T$ are Fredholm by XV.11.3.12. \hfill $\Box$

**XV.11.5.9.** **Corollary.**

(i) If $\mathcal{X}$ is a Banach space, $S, T \in \mathcal{F}(\mathcal{X})$, and $K \in \mathcal{K}(\mathcal{X})$, then $ST$ and $T + K$ are Fredholm.

(ii) The set $\mathcal{F}(\mathcal{X})$ of Fredholm operators on $\mathcal{X}$ is open in $\mathcal{B}(\mathcal{X})$.

Part (ii) is an immediate consequence of XV.11.3.9., or follows directly from (i). It also follows from XV.11.3.9. that index is locally constant on $\mathcal{F}(\mathcal{X})$, hence constant on connected components. In particular, we obtain:
**XV.11.5.10.** Corollary. Let $\mathcal{X}$ be a Banach space, and $T \in \mathcal{F}(\mathcal{X})$. Then $\text{index}(T) = \text{index}(T + K)$ for any $K \in \mathcal{K}(\mathcal{X})$.

Proof: $T$ and $T + K$ are connected in $\mathcal{F}(\mathcal{X})$ by the path $T + tK$, $0 \leq t \leq 1$.

**XV.11.5.11.** The index map from the connected components of $\mathcal{F}(\mathcal{X})$ to $\mathbb{Z}$ is thus well defined, but it is not either injective or surjective in general: it can fail to be injective since the invertible elements of $\mathcal{B}(\mathcal{X})$ are not necessarily connected [?], and we can even have $\text{index}(T) = 0$ for every $T \in \mathcal{F}(\mathcal{X})$ (e.g. for the examples of [?]; cf. [?]). If $\mathcal{H}$ is a Hilbert space, the index map is a bijection from the connected components of $\mathcal{F}(\mathcal{H})$ onto $\mathbb{Z}$.

**XV.11.5.12.** Theorem. Let $T \in \mathcal{F}(\mathcal{X})$. Then the following are equivalent:

(i) $\text{index}(T) = 0$.

(ii) There are $U, K \in \mathcal{B}(\mathcal{X})$ with $U$ invertible, $K$ compact, and $T = U + K$.

(iii) There are $V, F \in \mathcal{B}(\mathcal{X})$ with $V$ invertible, $F$ finite-rank, and $T = V + F$.

Proof: (i) $\Rightarrow$ (iii) is XV.11.4.1., and (iii) $\Rightarrow$ (ii) is trivial. For (ii) $\Rightarrow$ (i), if $T = U + K$, then $U^{-1}T = I + U^{-1}K$ is Fredholm with index 0 by XV.11.5.11. It is obvious that if $S$ is Fredholm and $W$ is invertible, then $WS$ is Fredholm and $\text{index}(WS) = \text{index}(S)$; thus $T = U(U^{-1}T)$ is Fredholm and $\text{index}(T) = 0$.

The argument of [RSN90, §70] gives a direct proof of (iii) $\Rightarrow$ (i).

**XV.11.5.13.** Note that if $\mathcal{X}$ has the approximation property, then there is a simple proof of (ii) $\Rightarrow$ (iii): if $T = U + K$ with $U$ invertible and $K$ compact, approximate $K$ closely enough by a finite rank $F$ that $V = U + (K - F)$ is invertible (using the fact that the set of invertible elements of $\mathcal{B}(\mathcal{X})$ is open); then $T = V + F$.

Finally, we state the Fredholm Alternative in its usual form:

**XV.11.5.14.** Corollary. [Fredholm Alternative] If $\mathcal{X}$ is a Banach space and $T \in \mathcal{K}(\mathcal{X})$, then $I + T$ is injective if and only if it is surjective. So either the equation $(I + T)x = y$ has a unique solution for every $y$, or the homogeneous equation $(I + T)x = 0$ has a nontrivial (but finite-dimensional) space of solutions.

Furthermore, the equation $(I + T^*)\phi = \psi$ has a unique solution for each $\psi$ if and only if $(I + T)x = y$ has a unique solution for each $y$, and the solution set of $(I + T^*)\phi = 0$ has the same (finite) dimension as the solution set of $(I + T)x = 0$.

**XV.11.5.15.** Note also that if $T \in \mathcal{K}(\mathcal{X})$, then XV.11.5.1. and its consequences, notably XV.11.5.14., are valid not only for $I + T$ but also for $T - \lambda I = -\lambda(I + (-\lambda^{-1}T))$ for any $\lambda \neq 0$. This is useful in describing the eigenvalues of $T$. The results of Fredholm theory, especially XV.11.5.5., are used to prove that the spectrum of a compact operator on an infinite-dimensional Banach space consists of 0 and at most countably many isolated nonzero eigenvalues of finite multiplicity.

1843
XV.11.5.16. The results of this section can be extended: if $T$ is Fredholm and $S$ is strictly singular, then $T + S$ is Fredholm and $\text{index}(T + S) = \text{index}(T)$ (cf. ()). Thus the Fredholm Alternative holds more generally for strictly singular operators.

XV.11.6. Exercises

XV.11.6.1. (a) Use the Spectral Theorem () to show that every normal operator on a Hilbert space has finite ascent 1.

(b) Consider the multiplication operator on $\ell^2$ given by

$$T(x_1, x_2, \ldots) = \left( x_1, \frac{x_2}{2}, \ldots, \frac{x_n}{n}, \ldots \right).$$

Show that $T$ is self-adjoint, hence normal, and compact, but does not have finite descent.

The difference between ascent and descent (at least in this case) is that the null space of an operator is closed but the range is not necessarily closed. In fact, if $T$ is a normal operator on a Hilbert space, then the closure of the range of $T^n$ is the same as the closure of $\mathcal{R}(T)$ for all $n$ (it is the orthogonal complement of $\mathcal{N}(T) = \mathcal{N}(T^n)$).

XV.11.6.2. (a) Show that every strictly singular operator on a Hilbert space is compact.

(b) Show that the inclusion map from $\ell^1$ to $\ell^2$ is not compact. [Consider the standard basis.]

(c) Show that the inclusion map from $\ell^1$ to $\ell^2$ is strictly singular. [Use that the weak topology on $\ell^1$ has the same convergent sequences as as the norm topology ()].

(d) Let $X = \ell^1 \oplus \ell^2$. Define $T \in \mathcal{B}(X)$ by $T(x, y) = (0, x)$. Show that $T$ is strictly singular but not compact.

(e) Show that if $p < q$, the inclusion of $\ell^p$ into $\ell^q$ is strictly singular but not compact. So if $1 < p < q < \infty$, there is an operator on the reflexive (even uniformly convex, if the 2-norm is used on the direct sum) Banach space $\ell^p \oplus \ell^q$ which is strictly singular but not compact.

XV.11.6.3. Let $X$ be a Banach space. Show that the set of strictly singular operators on $X$ is a closed two-sided ideal in $\mathcal{B}(X)$.

XV.11.6.4. Let $X$ be a Banach space, and $S$ a strictly singular operator on $X$. Modify the proofs for compact operators to show the following:

(a) If $T \in \mathcal{F}(X)$, then $T + S \in \mathcal{F}(X)$ and $\text{index}(T + S) = \text{index}(T)$. In particular, if $T$ is invertible, then $T + S$ is a Fredholm operator of index 0.

(b) $\sigma(S)$ consists of 0 and at most countably many nonzero numbers, each an eigenvalue of finite multiplicity, with no nonzero accumulation point.
XV.12. The Spectral Theorem

The Spectral Theorem, in its various forms, gives a precise description of the structure of self-adjoint and, more generally, normal operators on a Hilbert space. It roughly says that such an operator is, up to unitary equivalence, a “multiplication operator with multiplicity” with a simple explicit description. The name “Spectral Theorem” comes from the fact that the description is expressed in terms of the spectrum of the operator.

XV.12.1. The Finite-Dimensional Spectral Theorem

We first discuss the Spectral Theorem on a finite-dimensional Hilbert space. As might be expected, this situation is pure Linear Algebra with no essential use of functional analysis.

XV.12.1.1. Theorem. [Finite-Dimensional Spectral Theorem] Let \( \mathcal{H} \) be a finite-dimensional Hilbert space, and \( T \) a normal operator on \( \mathcal{H} \). Then there is an orthonormal basis for \( \mathcal{H} \) consisting of eigenvectors of \( T \).

XV.12.1.2. Corollary. Let \( A \in M_n = M_n(\mathbb{C}) \) be a normal matrix. Then \( A \) is unitarily equivalent to a diagonal matrix.

XV.12.1.3. The converse is true too: if \( A \in M_n \) is unitarily equivalent to a diagonal matrix, then \( A \) is normal.

We now give the proof of Theorem XV.12.1.1.

Proof: The proof of XV.12.1.1. is a simple complete induction (infinite regress), based on () and the elementary facts that if \( T \) is normal, then so is \( T - \lambda I \) for any \( \lambda \in \mathbb{C} \) and that any operator on a finite-dimensional complex vector space has at least one eigenvalue ()

Suppose the result holds for Hilbert spaces of dimension \( < n \), and that \( \mathcal{H} \) is \( n \)-dimensional. Let \( \lambda \) be an eigenvalue of \( T \), and \( E_\lambda \) the corresponding eigenspace, i.e.

\[
E_\lambda = \{ x \in \mathcal{H} : Tx = \lambda x \}.
\]

Then \( E_\lambda = \mathcal{N}(T - \lambda I) \) is a nontrivial subspace of \( \mathcal{H} \), so \( E_\lambda^\perp = \mathcal{R}(T - \lambda I) \) is a subspace of \( \mathcal{H} \) of dimension \( < n \) (recall that every subspace of \( \mathcal{H} \) is closed). If \( x \in E_\lambda^\perp \), then \( x = (T - \lambda I)y \) for some \( y \), and hence

\[
Tx = T(T - \lambda I)y = (T - \lambda I)(Ty) \in \mathcal{R}(T - \lambda I) = E_\lambda^\perp.
\]

So \( T \) maps \( E_\lambda^\perp \) into itself, and so \( S = T|_{E_\lambda^\perp} \) is an operator on \( E_\lambda^\perp \). It is easily checked that the adjoint of \( S \) is \( T^*|_{E_\lambda^\perp} \), since \( E_\lambda = \mathcal{N}((T - \lambda I)^*) \) and \( E_\lambda^\perp = \mathcal{R}((T - \lambda I)^*) \) because \( T - \lambda I \) is normal, and therefore \( E_\lambda \) and \( E_\lambda^\perp \) are invariant under \( T^* \) by the same argument as for \( T \). Thus \( S \) is normal, so by the inductive assumption there is an orthonormal basis for \( E_\lambda^\perp \) consisting of eigenvectors of \( S \), and putting this basis together with any orthonormal basis of \( E_\lambda \) gives an orthonormal basis for \( \mathcal{H} \) consisting of eigenvectors of \( T \). ☑
XV.12.1.4. The orthonormal basis in XV.12.1.1. is not unique. However, the decomposition of \( \mathcal{H} \) into eigenspaces of \( T \) is unique, and is a more natural way to phrase the Spectral Theorem:

\[ \text{XV.12.1.5. Corollary.} \]
Let \( \mathcal{H} \) be a finite-dimensional Hilbert space, and \( T \) a normal operator on \( \mathcal{H} \). Let \( \lambda_1, \ldots, \lambda_m \) be the distinct eigenvalues of \( T \), and \( \mathcal{E}_{\lambda_1}, \ldots, \mathcal{E}_{\lambda_m} \) the corresponding eigenspaces. Then

(i) \( \mathcal{E}_{\lambda_j} \perp \mathcal{E}_{\lambda_k} \) for \( j \neq k \).

(ii) The subspaces \( \{\mathcal{E}_{\lambda_1}, \ldots, \mathcal{E}_{\lambda_m}\} \) span \( \mathcal{H} \).

Thus \( \mathcal{H} \cong \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_m} \).

There is an alternate phrasing in terms of projection operators. For each \( k \), let \( E_k \) be the orthogonal projection of \( \mathcal{H} \) onto \( \mathcal{E}_{\lambda_k} \). Then:

\[ \text{XV.12.1.6. Corollary.} \]
Let \( \mathcal{H} \) be a finite-dimensional Hilbert space, and \( T \) a normal operator on \( \mathcal{H} \). Let \( \lambda_1, \ldots, \lambda_m \) be the distinct eigenvalues of \( T \). Then there are unique projections \( E_1, \ldots, E_m \in \mathcal{B}(\mathcal{H}) \) such that

(i) \( E_j \perp E_k \) for \( j \neq k \).

(ii) \( E_1 + \cdots + E_m = I \).

(iii) \( E_k T = TE_k = E_k T E_k = \lambda_k E_k \) for each \( k \).

(iv) \( \lambda_1 E_1 + \cdots + \lambda_m E_m = T \).

The expression in (iv) is called the spectral resolution of \( T \).

\[ \text{XV.12.1.7.} \]
This representation can be extended. If \( f \) is a polynomial with complex coefficients, and \( f(T) \) is defined in the standard way (), then an easy calculation shows that

\[ f(T) = f(\lambda_1)E_1 + \cdots + f(\lambda_m)E_m \]

where \( E_1, \ldots, E_m \) are as in XV.12.1.6. Since \( f(T) \) is also normal, and the eigenvalues of \( f(T) \) are \( f(\lambda_1), \ldots, f(\lambda_m) \), then this expression is almost the spectral resolution of \( f(T) \), except that the \( f(\lambda_k) \) need not be distinct and if not some terms in this sum could be combined.

\[ \text{XV.12.1.8.} \]
In particular, if \( f_k \) is a polynomial with \( f_k(\lambda_k) = 1 \) and \( f_k(\lambda_j) = 0 \) for \( j \neq k \) (cf. V.11.1.3.), then \( f_k(T) = E_k \). Thus the \( E_k \) are expressible as polynomials in \( T \).
XV.13. \textit{L}^p\textit{-Spaces}

In this section, we define and prove the basic properties of the \textit{L}^p\textit{-spaces of a measure space. These are very important examples of Banach spaces, especially \textit{L}^2\textit{-spaces, which are Hilbert spaces.}

XV.13.1. Conjugate Exponents

This subsection simply sets up some simple notation and terminology which will prove very useful.

XV.13.1.1. Definition. Let \( p \) be a real number greater than 1. The number \( q = \frac{p}{p-1} \) is called the \textit{conjugate exponent} to \( p \). The pair \((p,q)\) is a \textit{conjugate pair of exponents}.

If \( p > 1 \), then \( q > 1 \) also. This definition is symmetric in \( p \) and \( q \), as is immediate from the next proposition, which gives several equivalent versions of the defining equality. The proof is a simple exercise in high-school algebra.

XV.13.1.2. Proposition. Let \( p, q > 1 \). The following are equivalent:

(i) \( q \) is the conjugate exponent to \( p \), i.e. \( q = \frac{p}{p-1} \).

(ii) \( p \) is the conjugate exponent to \( q \), i.e. \( p = \frac{q}{q-1} \).

(iii) \( \frac{1}{p} + \frac{1}{q} = 1 \).

(iv) \( pq - q = p \).

(v) \( p - \frac{p}{q} = 1 \).

(vi) \( \frac{p}{q} = p - 1 \).

(vii) \( (p - 1)(q - 1) = 1 \).

Note that if \( 1 < p < 2 \), then \( q > 2 \), and if \( p = 2 \) then \( q = 2 \) (i.e. \( 2 \) is “self-conjugate”). As \( p \) increases, \( q \) decreases; as \( p \to 1 \), \( q \to \infty \) (and vice versa). We extend the definition of a conjugate pair to include the pairs \((1, \infty)\) and \((\infty, 1)\), i.e. the conjugate exponent to 1 is \( \infty \) and vice versa. These pairs behave somewhat differently than the pairs for \( 1 < p < \infty \), however, and some results do not apply to them.

XV.13.2. \textit{L}^p\textit{-Spaces, Norms, and Inequalities}

XV.13.2.1. Definition. Let \((X, \mathcal{A}, \mu)\) be a measure space and \( 0 < p < \infty \). If \( f \) is an \( \mathcal{A}\)-measurable function from \( X \) to \( \mathbb{C} \), set

\[
J_p(f) = \int_X |f|^p \, d\mu 
\]

\[
||f||_p = [J_p(f)]^{1/p} = \left[ \int_X |f|^p \, d\mu \right]^{1/p}
\]

(where \( \infty^{1/p} = \infty \)).
We let \( L^p_c(X,A,\mu) \) (or \( L^p(\mu) \)) if \( X \) and \( A \) are understood) be the set of measurable functions from \( X \) to \( \mathbb{C} \) such that \( J_\mu(f) < \infty \). \( L^p_c(X,A,\mu) \) is defined similarly. We let \( L^p(X,A,\mu) \) stand for either \( L^p_c(X,A,\mu) \) or \( L^p_c(X,A,\mu) \) when the base field is unimportant (it must of course always be specified and kept in mind).

The space \( L^p(\mu) \) is really only interesting if \( \mu \) is \( \sigma \)-finite or at least semifinite, since it is rather degenerate otherwise (XV.13.5.7., Exercise XV.13.7.4.–XV.13.7.5.). However, the definition and basic properties need no assumptions about the measure space.

The next proposition records some obvious properties of the \( L^p \)-spaces.

**XV.13.2.2.** Proposition. Let \( (X,A,\mu) \) be a measure space and \( 0 < p < \infty \). Then

(i) \( L^p(X,A,\mu) \subseteq L^p_c(X,A,\mu) \).

(ii) If \( f \in L^p(X,A,\mu) \), \( g \) is \( A \)-measurable, and \( |g| \leq |f| \) a.e., then \( g \in L^p(X,A,\mu) \). In particular, if \( g \) is \( A \)-measurable, then \( g \in L^p(X,A,\mu) \) if and only if \( |g| \in L^p(X,A,\mu) \).

(iii) If \( r > 0 \) and \( f \in L^p(X,A,\mu) \), then \( |f|^{p/r} \in L^r(X,A,\mu) \) and \( \| |f|^{p/r} \|_r = \| f \|_{p/r} \). In particular, if \( p > 1 \) and \( q \) is the conjugate exponent to \( p \), and \( f \in L^p(X,A,\mu) \), then \( |f|^{p^{-1}} \in L^q(X,A,\mu) \) and \( \| |f|^{p^{-1}} \|_q = \| f \|_{p^{-1}} \) (XV.13.1.2.(vi)).

(iv) If \( f : X \to \mathbb{C} \) is \( A \)-measurable, then \( f \in L^p_c(X,A,\mu) \) if and only if the real and imaginary parts of \( f \) are in \( L^p_c(X,A,\mu) \).

We will eventually show that \( \| \cdot \|_p \) is a seminorm on \( L^p(\mu) \) if \( p \geq 1 \). The triangle inequality (which only holds for \( p \geq 1 \)) will take some work. We first get a weak form of the triangle inequality which will be enough to show that \( L^p(\mu) \) is a vector space, even for \( 0 < p < 1 \):

**XV.13.2.3.** Proposition. Let \( (X,A,\mu) \) be a measure space, \( 0 < p < \infty \), and \( f \in L^p_c(X,A,\mu) \), \( \alpha \in \mathbb{C} \). Then

\[
J_p(f + g) \leq 2^p[J_p(f) + J_p(g)] \\
J_p(\alpha f) = |\alpha|^p J_p(f)
\]

In particular, \( f + g \) and \( \alpha f \) are in \( L^p_c(X,A,\mu) \), and \( L^p_c(X,A,\mu) \) is a complex vector space (hence \( L^p(X,A,\mu) \) is a real vector space).

**Proof:** The statement for \( \alpha f \) is obvious. For the first statement, we have

\[
J_p(f + g) = \int_X |f + g|^p d\mu \leq \int_X (|f| + |g|)^p d\mu \leq \int_X (2\max(|f|,|g|))^p d\mu \\
= 2^p \int_X (\max(|f|^p,|g|^p)) d\mu \leq 2^p \int_X (|f|^p + |g|^p) d\mu \\
= 2^p \left[ \int_X |f|^p d\mu + \int_X |g|^p d\mu \right] = 2^p[J_p(f) + J_p(g)].
\]

\( \blacksquare \)

In order to prove the triangle inequality for \( \| \cdot \|_p \) \((1 < p < \infty)\), we need to establish some inequalities: Young’s Inequality for products, Hölder’s Inequality, and finally Minkowski’s Inequality. We begin with Young’s Inequality for products after observing a simple preliminary lemma:
**XV.13.2.4.** **Lemma.** Let \( x, y \in (0, \infty) \) and \( 0 \leq t \leq 1 \). Then
\[
t \log x + (1-t) \log y \leq \log(tx + (1-t)y) .
\]
We have equality if and only if \( x = y \) or \( t = 0 \) or 1.

The proof is essentially just the observation that the graph of \( \log x \) is (strictly) concave downward.

**XV.13.2.5.** **Proposition.** [Young’s Inequality for Products] Let \( \alpha, \beta \in [0, \infty) \) and \( (p,q) \) a conjugate pair of exponents with \( 1 < p < \infty \). Then
\[
\alpha \beta \leq \frac{1}{p} \alpha^p + \frac{1}{q} \beta^q
\]
with equality if and only if \( \alpha^p = \beta^q \).

**Proof:** If \( \alpha \) or \( \beta \) is 0, the result is trivial. Otherwise, we have
\[
\log(\alpha \beta) = \log \alpha + \log \beta = \frac{1}{p} \log(\alpha^p) + \frac{1}{q} \log(\beta^q) \leq \log \left( \frac{1}{p} \alpha^p + \frac{1}{q} \beta^q \right)
\]
where the inequality comes from Lemma XV.13.2.4. with \( x = \alpha^p, \ y = \beta^q, \ t = \frac{1}{p} \). Since the log function is strictly increasing, the result follows.

See Exercise XV.13.7.1. for an equivalent statement and another proof of Young’s Inequality for products, and Exercise XV.13.7.2. for a more general version. There is another inequality due to Young for convolutions ()

We now obtain Hölder’s Inequality, which is a generalization of the CBS Inequality (XV.9.2.1.). This inequality will not only be used in this section, but will be important in other applications.

**XV.13.2.6.** **Proposition.** [Hölder’s Inequality] Let \( (X, \mathcal{A}, \mu) \) be a measure space, \( (p,q) \) a conjugate pair with \( 1 < p < \infty \), \( f \in L^p(\mu) \), and \( g \in L^q(\mu) \). Then \( fg \in L^1(\mu) \) and
\[
\int_X |fg| \, d\mu \leq \|f\|_p \|g\|_q .
\]
We have equality if and only if \( |g|^q = \alpha |f|^p \) a.e. for some \( \alpha > 0 \), or if \( f = 0 \) a.e. or \( g = 0 \) a.e.

**Proof:** If \( f = 0 \) a.e. or \( g = 0 \) a.e. both sides are 0, so equality holds. Otherwise, let \( F = \frac{1}{\|f\|_p} f \), \( G = \frac{1}{\|g\|_q} g \). Then \( \|F\|_p = \|G\|_q = 1 \), and we have, using Young’s Inequality for products,
\[
\frac{1}{\|f\|_p \|g\|_q} \int_X |fg| \, d\mu = \int_X |FG| \, d\mu \leq \frac{1}{p} \int_X |F|^p \, d\mu + \frac{1}{q} \int_X |G|^q \, d\mu = \frac{1}{p} + \frac{1}{q} = 1 .
\]
We have equality if and only if \( |F|^p = |G|^q \) a.e. Note that multiplying \( f \) or \( g \) by a nonzero scalar does not change \( |F| \) and \( |G| \).

We can now prove the triangle inequality for \( \| \cdot \|_p \) for \( 1 \leq p < \infty \). This inequality is usually called Minkowski’s Inequality.
**XV.13.2.7.** Theorem. [Minkowski’s Inequality] Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(1 \leq p < \infty\). If \(f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu)\), then \(f + g \in \mathcal{L}^p(X, \mathcal{A}, \mu)\) and
\[
\|f + g\|_p \leq \|f\|_p + \|g\|_p.
\]

Proof: The case \(p = 1\) is essentially just the usual triangle inequality for numbers, and is left to the reader. So we assume \(p > 1\).

We have already noted that \(f + g \in \mathcal{L}^p(\mu)\) (XV.13.2.3.). Let \(q\) be the conjugate exponent to \(p\). Then by XV.13.2.2.(iii), \(|f + g|^{p-1} \in \mathcal{L}^q(\mu)\) and
\[
\|\|f + g|^{p-1}\|_q = \|f + g\|_p^{p/q} = \|f + g\|_p^{p-1}.
\]

Next note that
\[
|f + g|^p = |f + g| \cdot |f + g|^{p-1} \leq (|f| + |g|) \cdot |f + g|^{p-1} = |f| |f + g|^{p-1} + |g| |f + g|^{p-1}.
\]

Thus, by Hölder’s Inequality,
\[
\|f + g\|_p^p = \int_X |f + g|^p d\mu \leq \int_X |f| |f + g|^{p-1} d\mu + \int_X |g| |f + g|^{p-1} d\mu
\leq \|f\|_p \|f + g|^{p-1}\|_q + \|g\|_p \|f + g|^{p-1}\|_q = (\|f\|_p + \|g\|_p) \|f + g\|_p^{p-1}.
\]

If \(\|f + g\|_p = 0\), the result is trivial; otherwise, divide by \(\|f + g\|_p^{p-1}\).

**XV.13.2.8.** Corollary. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(1 \leq p < \infty\). Then \(\| \cdot \|_p\) is a seminorm on \(\mathcal{L}^p(X, \mathcal{A}, \mu)\).

Proof: It is obvious that \(\|f\|_p \in [0, \infty)\) if \(f \in \mathcal{L}^p(\mu)\), and it follows from XV.13.2.3. that if \(f \in \mathcal{L}^p(\mu)\) and \(\alpha\) is a scalar, then \(\|\alpha f\|_p = |\alpha| \|f\|_p\) (these statements are also true if \(0 < p < 1\)). Minkowski’s Inequality gives the triangle inequality (for \(1 \leq p < \infty\)).

This seminorm is usually not a norm: \(\|f\|_p = 0\) if (and only if) \(f = 0\) a.e. We will remedy this in the next section.

The function \(\| \cdot \|_p\) does not satisfy the triangle inequality for \(0 < p < 1\) in general:

**XV.13.2.9.** Example. Let \(X = \{a, b\}\) be a two-point space, and \(\mu\) counting measure on \((X, \mathcal{P}(X))\). Let \(f(a) = 1\), \(f(b) = 0\), \(g(a) = 0\), \(g(b) = 1\). For any \(p\), \(0 < p < \infty\), we have \(\|f\|_p = \|g\|_p = 1\), \(\|f + g\|_p = 2^{1/p}\).

So if \(p < 1\), \(\|f + g\|_p > 2\) and \(\| \cdot \|_p\) does not satisfy the triangle inequality.
XV.13.3. \textit{L}^{p}\text{-Spaces}

Definition as a quotient, completeness, density of simple functions, density of continuous functions for Borel measures

Let \((X, \mathcal{A}, \mu)\) be a measure space. Since \(\|\cdot\|_p\) is usually only a seminorm and not a norm on \(L^p(X, \mathcal{A}, \mu)\), by the usual process () we reduce to the quotient by the subspace of vectors of seminorm 0, which in this case is the set of functions which are 0 a.e. This reduction effectively means that we identify functions which are equal a.e.

XV.13.3.1. Definition. Let \(0 < p < \infty\). \(L^p(X, \mathcal{A}, \mu)\) (denoted just \(L^p\) when \(X\) and \(\mathcal{A}\) are understood) is the quotient of \(L^p(X, \mathcal{A}, \mu)\) by the subspace

\[\{ f \in L^p(X, \mathcal{A}, \mu) : \| f \|_p = 0 \} = \{ f : f \text{ measurable, } f = 0 \text{ a.e.} \} .\]

(As with \(L^p\), \(L^p\) can stand for either \(L^p_{\mathbb{R}}\) or \(L^p_{\mathbb{C}}\); the base field should always be specified in a specific context.)

If \(f \in L^p(\mu)\), we denote by \([f]\) the image of \(f\) in \(L^p(\mu)\):

\([f] = \{ g \in L^p(\mu) : f = g \text{ a.e.} \} .\]

As in (), the seminorm \(\| \cdot \|_p\) on \(L^p(\mu)\) drops to a norm on \(L^p(\mu)\), also denoted \(\| \cdot \|_p\), by \(\|[f]\|_p = \|f\|_p\).

Many authors are rather sloppy in distinguishing between \(L^p(\mu)\) and \(L^p(\nu)\), and functional analysts tend to be rather casual about the difference; we will eventually follow this convention of thinking of elements of \(L^p(\mu)\) as functions rather than equivalence classes of functions, but in this section we will make a careful distinction in establishing the basic properties of these spaces (which can be used to carefully justify the casual convention).

The most important fact about \(L^p\)-spaces is that they are Banach spaces:

XV.13.3.2. Theorem. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(1 \leq p < \infty\). Then \(L^p(X, \mathcal{A}, \mu)\) is complete.

Proof: We will use (). Suppose \((f_k)\) is a sequence in \(L^p(X, \mathcal{A}, \mu)\) with

\[\sum_{k=1}^{\infty} \| f_k \|_p = M < \infty .\]

We will show that there is an \(f \in L^p(X, \mathcal{A}, \mu)\) with

\[\lim_{n \to \infty} \left\| f - \sum_{k=1}^{n} f_k \right\|_p = 0\]

which will imply that the series \(\sum_{k=1}^{\infty} |f_k|\) converges to \([f]\) in \(L^p(X, \mathcal{A}, \mu)\).

For each \(n\), let \(g_n = \sum_{k=1}^{n} |f_k|\). Then \((g_n)\) is an increasing sequence; let \(g = \lim g_n = \sup g_n\). Then \(g\) is an extended nonnegative real-valued measurable function on \((X, \mathcal{A})\). By Minkowski’s Inequality, for each \(n\) we have

\[\int_X g_n^p \, d\mu = \| g_n \|_p^p \leq \left[ \sum_{k=1}^{n} \| f_k \|_p \right]^p \leq M^p .\]
and so, by the Monotone Convergence Theorem,
\[
\int_X g^p \, d\mu = \sup_n \int_X g_n^p \, d\mu \leq M^p < \infty .
\]
Thus, although \( g \) may technically not be in \( L^p(X, \mathcal{A}, \mu) \) since it may take the value \(+\infty\), it is finite a.e. and equal a.e. to a function \( h \in L^p(X, \mathcal{A}, \mu) \). For any \( x \in X \) with \( g(x) < \infty \) the series \( \sum_{k=1}^\infty f_k(x) \) of scalars converges absolutely, hence converges; call the sum \( f(x) \). Extend \( f \) to be 0 on the \( x \)'s for which \( g(x) = +\infty \). Then \( f \) is an \( \mathcal{A} \)-measurable function from \( X \) to \( \mathbb{C} \), and \( |f| \leq |h| \) a.e., so \( f \in L^p(X, \mathcal{A}, \mu) \). We also have
\[
\left| f - \sum_{k=1}^n f_k \right|^p \leq |h|^p \text{ a.e.}
\]
for all \( n \), and \( |h|^p \in L^1(X, \mathcal{A}, \mu) \); since \(|f - \sum_{k=1}^n f_k|^p \to 0 \) pointwise a.e., by the Dominated Convergence Theorem we have
\[
\lim_{n \to \infty} \left\| f - \sum_{k=1}^n f_k \right\|_p^p = \lim_{n \to \infty} \left[ \int_X \left| f - \sum_{k=1}^n f_k \right|^p \right] = 0 .
\]
It is interesting to note that, although it is somewhat easier to prove that \( \| \cdot \|_p \) is a seminorm (i.e. satisfies the triangle inequality) for \( p = 2 \) than for general \( p \) (cf. ()), and considerably easier for \( p = 1 \), it is no easier to prove completeness of \( L^p \) for \( p = 1 \) or \( p = 2 \) than for general \( p \).

It is usually important to identify some dense subsets of “nice” elements in any Banach space, and this is particularly important in the \( L^p \)-spaces. For example, we will show (that in \( L^p \)-spaces arising from Borel measures on good topological spaces, the continuous functions are dense in \( L^p \) for any \( p \). One general and very useful fact is true in any \( L^p \)-space:

**Theorem.** Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(1 \leq p < \infty\). Then the (classes of) simple functions in \( L^p(X, \mathcal{A}, \mu) \) are dense in \( L^p(X, \mathcal{A}, \mu) \).

**Proof:** By working with real and imaginary parts, it suffices to prove the result for \( L^p_{\mathbb{R}}(\mu) \). If \( f \in L^p_{\mathbb{R}}(\mu) \), then by (1) there is a sequence of real-valued simple functions \((s_n)\) converging pointwise to \( f \), with \( |s_n| \leq |f| \) for all \( n \). Thus each \( s_n \in L^p_{\mathbb{R}}(\mu) \). We have \( |s_n - f|^p \leq 2|f|^p \) for all \( n \), and \( 2|f|^p \in L^1(\mu) \). Since \( |s_n - f|^p \to 0 \) pointwise, by the Dominated Convergence Theorem we have
\[
\lim_{n \to \infty} \|s_n - f\|_p^p = \lim_{n \to \infty} \left[ \int_X |s_n - f|^p \, d\mu \right] = 0 .
\]
XV.13.4. $L^\infty$- and $L^\infty$-Spaces

XV.13.5. Comparison Between $L^p$-Spaces and $L^p$-Spaces

Let $(X, \mathcal{A}, \mu)$ be a measure space. There is no containment between the $L^p$-spaces or the $L^p$-spaces of $(X, \mathcal{A}, \mu)$ as $p$ varies, except in certain special cases. There are two competing constraints which can cause a (measurable) function $f$ to fail to be in $L^p(X, \mathcal{A}, \mu)$:

(i) $f$ can go to zero “too slowly” (or not at all) at “infinity.”

(ii) $f$ can go to infinity “too rapidly” on a small portion of $X$.

The first problem gets worse as $p$ decreases, and the second problem gets worse as $p$ increases. We will make these problems precise in this subsection, and give examples.

Consider the first constraint.

XV.13.5.1. Example. Suppose $(X, \mathcal{A}, \mu)$ contains measurable subsets of arbitrarily large finite measure (e.g. $\mu$ is $\sigma$-finite but not finite). Choose an increasing sequence $(B_n)$ in $\mathcal{A}$ with $\mu(B_1) \geq 1$ and $\mu(B_n) + 1 < \mu(B_{n+1}) < \infty$ for all $n > 1$. Then, by taking successive differences, there is a sequence $(A_n)$ of pairwise disjoint sets in $\mathcal{A}$ with $1 \leq \mu(A_n) < \infty$ for all $n$. Fix $p, 0 < p < 1$, and define a function $f : X \to [0, 1]$ by setting

\[ f(x) = \frac{1}{[n\mu(A_n)]^{1/p}} \text{ if } x \in A_n \]

\[ f(x) = 0 \text{ if } x \notin \cup_n A_n . \]

Then $f$ is measurable, and we have

\[ J_p(f) = \sum_{n=1}^{\infty} \frac{1}{n\mu(A_n)} \mu(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty \]

so $f \notin L^p(X, \mathcal{A}, \mu)$. However, if $p < r < \infty$, we have

\[ J_r(f) = \sum_{n=1}^{\infty} \frac{1}{n^{r/p}\mu(A_n)^{r/p}} \mu(A_n) = \sum_{n=1}^{\infty} \frac{1}{n^{r/p}\mu(A_n)^{r/p}} \leq \sum_{n=1}^{\infty} \frac{1}{n^{r/p}} < \infty \]

so $f \in \cap_{r>p} L^r(X, \mathcal{A}, \mu)$. (Note that $f$ is bounded, so $f \in L^\infty(X, \mathcal{A}, \mu)$ also.) In particular, $L^r(X, \mathcal{A}, \mu)$ is not contained in $L^p(X, \mathcal{A}, \mu)$ for $p < r \leq \infty$.

In the absence of such subsets, we have a containment:

XV.13.5.2. Proposition. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and let $0 < p \leq r \leq \infty$. Then $L^r(X, \mathcal{A}, \mu)$ is contained in $L^p(X, \mathcal{A}, \mu)$.

Proof: If $r = \infty$, the result holds since any bounded measurable function on a finite measure space is integrable. So assume $r < \infty$ and let $f \in L^r(X, \mathcal{A}, \mu)$. Set

\[ A = \{ x \in X : |f(x)| \geq 1 \} \]
\[ B = \{ x \in X : |f(x)| < 1 \} . \]

Then \(|f|^p \leq |f|^r\) on \(A\), so we have
\[
J_p(f) = \int_X |f|^p d\mu = \int_A |f|^p d\mu + \int_B |f|^p d\mu \leq \int_A |f|^r d\mu + \int_B |f|^p d\mu
\]
\[
\leq J_r(f) + \mu(B) < \infty
\]
so \(f \in L^p(X, \mathcal{A}, \mu)\).  

Now we turn to the second constraint.

\textbf{XV.13.5.3. Example.} Let \((X, \mathcal{A}, \mu)\) be a measure space which contains measurable sets of arbitrarily small positive measure. Let \(B_1 \in \mathcal{A}\) with \(0 < \mu(B_1) \leq 1\), and inductively let \(B_n \in \mathcal{A}\) with \(0 < \mu(B_n) \leq \frac{1}{2} \mu(B_{n-1})\) for \(n > 1\). Set \(A_n = B_n \setminus (\bigcup_{k=n+1}^{\infty} B_k)\). Then the \(A_n\) are pairwise disjoint, and since
\[
\mu(\bigcup_{k=n+1}^{\infty} B_k) \leq \sum_{k=n+1}^{\infty} \mu(B_k) \leq \sum_{m=1}^{\infty} 3^{-m} \mu(B_n) = \frac{\mu(B_n)}{2}
\]
we have \(0 < \frac{1}{2} \mu(B_n) \leq \mu(A_n)\). Also, \(\mu(A_n) \leq \mu(B_n) \leq 3^{-n+1}\).

Fix \(p, 0 < p < \infty\). Define a function \(f : X \to [0, \infty)\) by
\[
f(x) = [\mu(A_n)]^{-1/p} \text{ if } x \in A_n
\]
\[
f(x) = 0 \text{ if } x \notin \bigcup_n A_n .
\]
Then we have
\[
J_p(f) = \sum_{n=1}^{\infty} [\mu(A_n)]^{-1} \mu(A_n) = \infty
\]
so \(f \notin L^p(X, \mathcal{A}, \mu)\). But if \(0 < r < p\), we have
\[
J_r(f) = \sum_{n=1}^{\infty} [\mu(A_n)]^{-r/p} \mu(A_n) = \sum_{n=1}^{\infty} [\mu(A_n)]^{1-r/p} \leq \sum_{n=1}^{\infty} [3^{1-r/p}]^{-n+1} < \infty
\]
so \(f \in \cap_{r<p} L^r(X, \mathcal{A}, \mu)\). In particular, \(L^r(X, \mathcal{A}, \mu)\) is not contained in \(L^p(X, \mathcal{A}, \mu)\) for any \(r < p\). Also, \(f\) is not essentially bounded, so \(f \notin L^\infty(X, \mathcal{A}, \mu)\), i.e. \(L^p(X, \mathcal{A}, \mu)\) is not contained in \(L^\infty(X, \mathcal{A}, \mu)\).

We again have a containment in the absence of this situation:
**XV.13.5.4.** Proposition. Let \((X, \mathcal{A}, \mu)\) be a measure space. Suppose there is an \(\epsilon > 0\) such that, for every \(A \in \mathcal{A}\), either \(\mu(A) = 0\) or \(\mu(A) \geq \epsilon\). Let \(0 < r \leq p \leq \infty\). Then \(L^r(X, \mathcal{A}, \mu)\) is contained in \(L^p(X, \mathcal{A}, \mu)\).

**Proof:** Let \(f \in L^r(\mu)\). We first show that \(f\) is essentially bounded. We may assume \(\|f\|_r > 0\), since otherwise \(f = 0\) a.e. Set \(M = \frac{\|f\|_r}{\epsilon}\) and let \(C = \{x \in X : |f(x)| > M\}\). Since

\[
M^r \mu(C) = \frac{\|f\|_r^r}{\epsilon} \mu(C) < \int_C |f|^r \, d\mu \leq \int_X |f|^r \, d\mu = \|f\|_r^r
\]

we have \(\mu(C) < \epsilon\), so \(\mu(C) = 0\), i.e. \(|f| \leq M\) a.e. and \(f \in L^\infty(X, \mathcal{A}, \mu)\). If \(p < \infty\), let \(A = \{x : |f(x)| \geq 1\}\) and \(B = \{x : |f(x)| < 1\}\). Then \(|f|^p < |f|^r\) on \(B\), and

\[
\mu(A) \leq \int_A |f|^r \, d\mu \leq \int_X |f|^r \, d\mu = J_r(f) < \infty
\]

and we have

\[
J_p(f) = \int_X |f|^p \, d\mu = \int_A |f|^p \, d\mu + \int_B |f|^p \, d\mu < \int_A |f|^p \, d\mu + \int_B |f|^r \, d\mu \leq M^p \mu(A) + J_r(f) < \infty
\]

so \(f \in L^p(X, \mathcal{A}, \mu)\).

**XV.13.5.5.** Examples. (a) A set with counting measure satisfies the hypotheses of XV.13.5.4. (with \(\epsilon = 1\)). Thus \(\ell^r(X) \subseteq \ell^p(X)\) for any \(X\) if \(r < p\).

(b) If \((X, \mathcal{A}, \mu)\) is a measure space containing both sets of arbitrarily large and arbitrarily small finite positive measure (e.g. \((\mathbb{R}, \mathcal{M}, \lambda))\), then there is no containment either way between \(L^p(X, \mathcal{A}, \mu)\) and \(L^r(X, \mathcal{A}, \mu)\) for \(p \neq r\).

There is, however, a large subspace which is common to \(L^p(\mu)\) for all \(p\). A measurable function \(f\) has support of finite measure if \(\{x : f(x) \neq 0\}\) has finite measure. (Some authors say that such a function has finite support, but this term, while more efficient, is ambiguous.) We can similarly define functions with support of \(\sigma\)-finite measure.

If \((X, \mathcal{A}, \mu)\) is a measure space, let \(L^{oo}(X, \mathcal{A}, \mu)\) (distinguish between oo and \(\infty\!\!)\) be the set of essentially bounded measurable functions with support of finite measure. The characteristic function of a set of finite measure is in \(L^{oo}(X, \mathcal{A}, \mu)\), hence so is any linear combination of such functions (simple functions with support of finite measure).

**XV.13.5.6.** Proposition. \(L^{oo}(X, \mathcal{A}, \mu) \subseteq L^p(X, \mathcal{A}, \mu)\) for \(0 < p \leq \infty\), and \(L^{oo}(X, \mathcal{A}, \mu)\) is dense in \(L^p(X, \mathcal{A}, \mu)\) if \(p < \infty\).

The density for \(p < \infty\) is immediate from the proof of XV.13.3.3., since a simple function which is in \(L^p\) for some \(p\) must be in \(L^{oo}\) (Exercise ())).
\( L^{\infty}(X, \mathcal{A}, \mu) \) is not dense in \( L^{\infty}(X, \mathcal{A}, \mu) \) unless \( \mu \) is finite, in which case \( L^{\infty}(X, \mathcal{A}, \mu) = L^{\infty}(X, \mathcal{A}, \mu) \).

We say a function \( f \) has support of almost finite measure if \( \{ x : |f(x)| > \epsilon \} \) has finite measure for all \( \epsilon > 0 \). This implies, but is not equivalent to, that \( f \) has support of \( \sigma \)-finite measure (consider the constant function 1 on \((\mathbb{R}, \mathcal{M}, \lambda)\)). It is easily shown (Exercise (\ref{ex})) that the closure of \( L^{\infty}(X, \mathcal{A}, \mu) \) in \( L^{\infty}(X, \mathcal{A}, \mu) \) is the set of essentially bounded functions with support of almost finite measure.

It is worth noting the next result:

\textbf{XV.13.5.7.} \textsc{Proposition.} Let \((X, \mathcal{A}, \mu)\) be a measure space and \(0 < p < \infty\). Then any function in \( L^p(X, \mathcal{A}, \mu) \) has support of almost finite measure (in particular, of \( \sigma \)-finite measure).

This proposition is false for \( L^{\infty}(X, \mathcal{A}, \mu) \) if \( \mu \) is not finite.

\textbf{XV.13.5.8.} \textsc{Remark.} For a fixed measure space \((X, \mathcal{A}, \mu)\), \( L^p(X, \mathcal{A}, \mu) \) is constructed from \( L^p(X, \mathcal{A}, \mu) \) by dividing out by the same subspace of functions for every \( p \). Thus if \( L^r(X, \mathcal{A}, \mu) \) is contained in \( L^p(X, \mathcal{A}, \mu) \) for some \( p \) and \( r \), it is technically strictly correct also that \( L^r(X, \mathcal{A}, \mu) \) is contained in \( L^p(X, \mathcal{A}, \mu) \). The converse also holds, as is easy to check. Thus, in all the above containment results of this section, \( L^p \) and \( L^r \) can be replaced by \( L^p \) and \( L^r \) to give strictly correct containment statements.

\textbf{XV.13.5.9.} Now suppose \((X, \mathcal{A}, \mu)\) is a measure space and \( \mathcal{B} \) is a sub-\( \sigma \)-algebra of \( \mathcal{A} \). Set \( \nu = \mu|_{\mathcal{B}} \). It is then obvious that \( L^p(X, \mathcal{B}, \nu) \) is a subspace of \( L^p(X, \mathcal{A}, \mu) \) for any \( p \), and the \( p \)-seminorms agree.

However, it is technically not strictly correct in most cases to say that \( L^p(X, \mathcal{B}, \nu) \) is a subspace of \( L^p(X, \mathcal{A}, \mu) \), since the elements of these spaces are different: if \( f \in L^p(X, \mathcal{B}, \nu) \), the corresponding elements are

\[
[f]_\mathcal{A} = \{ g : g \text{-measurable}, g = f \text{ a.e.} \}
\]

\[
[f]_\mathcal{B} = \{ g : g \text{-measurable}, g = f \text{ a.e.} \}
\]

which are not the same sets in general. However, the map \( [f]_\mathcal{B} \mapsto [f]_\mathcal{A} \) is a linear isometry from \( L^p(X, \mathcal{B}, \nu) \) onto a (complete, hence closed) subspace of \( L^p(X, \mathcal{A}, \mu) \), usually called the inclusion map, and by making this identification we usually, if rather casually, think of \( L^p(X, \mathcal{B}, \nu) \) as a closed subspace of \( L^p(X, \mathcal{A}, \mu) \) for any \( p \).

Normally, the inclusion map sends \( L^p(X, \mathcal{B}, \nu) \) onto a proper subspace of \( L^p(X, \mathcal{A}, \mu) \). But there is one important case where it should be noted that the inclusion map is surjective, hence an isometric isomorphism:

\textbf{XV.13.5.10.} \textsc{Proposition.} Let \((X, \mathcal{A}, \mu)\) be a measure space, and \((X, \mathcal{M}, \bar{\mu})\) its completion. Then, for any \( p \), the inclusion map from \( L^p(X, \mathcal{A}, \mu) \) to \( L^p(X, \mathcal{M}, \bar{\mu}) \) is an isomorphism.

The proof is immediate from the fact that every \( \mathcal{M} \)-measurable function is equal a.e. to an \( \mathcal{A} \)-measurable function (\ref{eq}). Note that \( L^p(X, \mathcal{M}, \bar{\mu}) \) is considerably larger than \( L^p(X, \mathcal{A}, \mu) \) in general.

In particular, \( L^p(\mathbb{R}, \mathcal{M}, \lambda) \) is naturally isomorphic to \( L^p(\mathbb{R}, \mathcal{B}, \lambda) \) for any \( p \).

\textbf{XV.13.6. \textit{L}^p\text{-spaces for Borel Measures}}

Density of continuous functions, separability, Egorov’s Theorem

1856
XV.13.7. Exercises

XV.13.7.1. Let $0 < r < 1$.
(a) Let $f(x) = x^r - rx$ for $x \geq 0$. Show that $f$ is increasing on $[0,1]$ and decreasing on $[1,\infty)$, and has a unique maximum $1 - r$ at $x = 1$.
(b) If $x \in [0,\infty)$, show that $x^r \leq rx + (1 - r)$, with equality if and only if $x = 1$.
(c) If $a, b > 0$, show that $a^r b^{1-r} \leq ra + (1 - r)b$, with equality if and only if $a = b$. [Apply (b) with $x = \frac{a}{b}$.
(d) Show that the inequality in (c) is equivalent to Young’s Inequality for products. [Let $r = \frac{1}{p}$, $\alpha = a^r$, $\beta = b^{1-r}$.

XV.13.7.2. Here is the general version of Young’s Inequality for products:

**Theorem.** [Young’s Inequality for Products, General Version] Let $\phi$ be a strictly increasing continuous function from $[0,\infty)$ onto $[0,\infty)$ (that is, $\phi(0) = 0$ and $\lim_{x \to \infty} \phi(x) = +\infty$), and let $\psi$ be the inverse function of $\phi$. For $x \geq 0$, set

$$\Phi(x) = \int_0^x \phi(t) \, dt$$
$$\Psi(x) = \int_0^x \psi(t) \, dt$$

If $\alpha, \beta \in [0,\infty)$, then

$$\alpha \beta \leq \Phi(\alpha) + \Psi(\beta)$$

with equality if and only if $\beta = \phi(\alpha)$.

See Figure XV.1 for a geometric interpretation in terms of areas. The red area is $\Phi(\alpha)$ and the blue area is $\Psi(\beta)$, and $\alpha \beta$ is the area of the rectangle bounded by the dotted lines and the axes.

(a) Use (a) to show that, for $\beta \geq 0$,

$$\Psi(\beta) = \int_0^\beta \int_0^{\psi(y)} dx \, dy = \int_0^{\psi(\beta)} \int_0^\beta dy \, dx$$

(b) If $\beta \geq \phi(\alpha)$, then

$$\alpha \beta = \int_0^\alpha \int_0^\beta dy \, dx = \int_0^\alpha \int_0^{\phi(x)} dy \, dx + \int_0^\beta \int_0^{\beta} dy \, dx = \Phi(\alpha) + \int_0^\alpha \int_{\phi(x)}^\beta dy \, dx$$

(c) Use (a) and (b) to prove the Theorem if $\beta \geq \phi(\alpha)$. Give a similar proof if $\beta \leq \phi(\alpha)$.

(d) Show that the Young’s Inequality for products of XV.13.2.5. is the special case $\phi(x) = x^{p-1}$.
XV.13.7.3. Prove the following generalized versions of Hölder’s Inequality:

**Theorem. [Hölder’s Inequality, General Version]** Let \((X, \mathcal{A}, \mu)\) be a measure space.

(i) Let \(p, q > 1, r \geq 1\), with

\[
\frac{1}{p} + \frac{1}{q} = \frac{1}{r}
\]

and let \(f \in L^p(\mu)\) and \(g \in L^q(\mu)\). Then \(fg \in L^r(\mu)\) and

\[
\|fg\|_r \leq \|f\|_p \|g\|_q.
\]

(ii) Let \(p_1, \ldots, p_n > 1, r \geq 1\), with

\[
\frac{1}{p_1} + \cdots + \frac{1}{p_n} = \frac{1}{r}
\]

and let \(f_k \in L^{p_k}(\mu)\) for \(1 \leq k \leq n\). Then \(f_1 f_2 \cdots f_n \in L^r(\mu)\) and

\[
\|f_1 f_2 \cdots f_n\|_r \leq \|f_1\|_{p_1} \|f_2\|_{p_2} \cdots \|f_n\|_{p_n}.
\]

[To prove (i), note that]

\[
\frac{1}{p/r} + \frac{1}{q/r} = 1
\]
and apply Hölder’s inequality to $|f|^r$ and $|g|^r$. Prove (ii) by induction on $n$. For the inductive step set

$$
\frac{1}{p} = \frac{1}{p_1} + \cdots + \frac{1}{p_n}
$$

and $q = p_{n+1}$ to obtain $f_1 f_2 \cdots f_n \in L^p$ and $\|f_1 \cdots f_{n+1}\|_r \leq \|f_1 \cdots f_n\|_p \|f_{n+1}\|_{p_{n+1}}$.

**XV.13.7.4.** Let $X = \{x\}$ be a one-point space, and let $\mu(\{x\}) = \infty$. Show that $L^p(X, \mathcal{P}(X), \mu) = \{0\}$ for $1 \leq p < \infty$, and that $L^\infty(X, \mathcal{P}(X), \mu)$ is one-dimensional (the set of all scalar-valued functions).

**XV.13.7.5.** Let $X$ be an uncountable set, and for $A \subseteq X$ let $\mu(A) = 0$ if $A$ is countable and $\mu(A) = \infty$ if $A$ is uncountable.

(a) If $1 \leq p < \infty$, show that $L^p(X, \mathcal{P}(X), \mu)$ is the set of functions vanishing except on a countable set, and $L^p(X, \mathcal{P}(X), \mu) = \{0\}$.

(b) Show that $L^\infty(X, \mathcal{P}(X), \mu)$ is the set of all functions which are bounded on a cocountable set, and $L^\infty(X, \mathcal{P}(X), \mu)$ is infinite-dimensional and even nonseparable.

**XV.13.7.6.** (a) Even in a finite continuous measure space the $L^p$-spaces can be small simply because of the lack of sufficiently many measurable functions. Let $X$ be an uncountable set, $\mathcal{A}$ the $\sigma$-algebra of countable-cocountable sets, and $\mu$ the measure on $(X, \mathcal{A})$ with $\mu(A) = 0$ if $A$ is countable and $\mu(A) = 1$ if $A$ is cocountable. A scalar-valued measurable function on $A$ must be constant except on a countable set (1). Show that $L^p(X, \mathcal{A}, \mu)$ is one-dimensional for all $p$, $1 \leq p \leq \infty$ (every measurable function is equal a.e. to a constant function).

(b) Show that the result of (a) holds true in any measure space in which every measurable set has measure either 0 or 1. Thus, if $X$ is a set of measurable cardinality () and $\mu$ is a full $\{0,1\}$-valued measure on $(X, \mathcal{P}(X))$, then $L^p(X, \mathcal{P}(X), \mu)$ is one-dimensional for $1 \leq p \leq \infty$ even though there are plenty of measurable functions; every scalar-valued function is still equal a.e. to a constant function.

**XV.13.7.7.** (a) Show that, in the situation of Example XV.13.5.1., for any $p$ there is an $f \in L^p(X, \mathcal{A}, \mu)$ which is not in $L^r(X, \mathcal{A}, \mu)$ for any $r < p$. [For each $n$, construct a function $f_n$ as in Example XV.13.5.1. which is in $L^p(\mu)$ but not in $L^{p-1/n}(\mu)$, and consider $\sum_{n=1}^{\infty} \frac{2^{-n}}{\|f_n\|_p} f_n$. Modify the argument slightly for $p = \infty$.]

(b) In the situation of XV.13.5.3., use a similar argument to show that if $0 < p < \infty$, there is an $f \in L^p(X, \mathcal{A}, \mu)$ which is not in $L^r(X, \mathcal{A}, \mu)$ for any $r > p$.

(c) If $(X, \mathcal{A}, \mu)$ contains sets of both arbitrarily large and arbitrarily small finite positive measure, and $0 < s \leq r \leq \infty$, show that

(i) There is an $f$ such that $f \in L^p(X, \mathcal{A}, \mu)$ if and only if $s \leq p \leq r$ (in particular, if $s = r$, there is an $f$ such that $f \in L^p(X, \mathcal{A}, \mu)$ if and only if $p = r$).

(ii) If $s < r$, there is an $f$ such that $f \in L^p(X, \mathcal{A}, \mu)$ if and only if $s < p \leq r$.

(iii) If $s < r$, there is an $f$ such that $f \in L^p(X, \mathcal{A}, \mu)$ if and only if $s \leq p < r$.

(iv) If $s < r$, there is an $f$ such that $f \in L^p(X, \mathcal{A}, \mu)$ if and only if $s < p < r$.  

1859
XV.13.7.8. Let \((X, \mathcal{A}, \mu)\) be a measure space, and \(0 < s < p < r \leq \infty\).

(a) Using the arguments of the proofs of XV.13.5.2. and XV.13.5.4., show that every \(f \in L^p(X, \mathcal{A}, \mu)\) can be written as \(f = g + h\) with \(g \in L^s(X, \mathcal{A}, \mu), h \in L^r(X, \mathcal{A}, \mu), gh = 0\), and \(J_p(f) \leq J_s(g) + J_r(h)\).

(b) Show that the function

\[
\|f\| = \inf \{\|g\|_s + \|h\|_r : g \in L^s(\mu), h \in L^r(\mu), f = g + h\}
\]

is a seminorm on the subspace \(L^s(\mu) + L^r(\mu)\) of the space of \(\mathcal{A}\)-measurable functions, that \(\|f\| = 0\) if and only if \(f \in Z\), and that the quotient, which we denote \(L^s(\mu) + L^r(\mu)\), is complete under the corresponding norm. Show that the inclusion map from \(L^p(\mu)\) into \(L^s(\mu) + L^r(\mu)\) is bounded.

(c) Show that \(L^s(\mu) \cap L^r(\mu) \subseteq L^p(\mu)\), that \(\|f\| = \|f\|_s + \|f\|_r\) is a seminorm on \(L^s(\mu) \cap L^r(\mu)\), \(\|f\| = 0\) if and only if \(f \in Z\), and that the quotient, which we denote \(L^s(\mu) \cap L^r(\mu)\), is complete under the corresponding norm. Show that the inclusion map from \(L^s(\mu) \cap L^r(\mu)\) into \(L^p(\mu)\) is bounded.

XV.13.7.9. (a) Prove directly that if \(p > 1\) and \(q\) is conjugate to \(p\), the map \(a \mapsto \phi_a\) is an isometric isomorphism from \(\ell^q\) to \((\ell^p)^*\), where

\[
\phi_a(x) = \sum_{k=1}^{\infty} \alpha_k \xi_k
\]

for \(a = (\alpha_1, \alpha_2, \ldots)\) and \(x = (\xi_1, \xi_2, \ldots)\). [Use Hölder’s inequality and the fact that \(\{e_k\}\) is a Schauder basis for \(\ell^p\).]

(b) Give a similar proof that \((c_0)^* \cong \ell^1\) and that \((\ell^1)^* \cong \ell^\infty\).

This gives an alternate proof that \(\ell^p\) is complete for \(1 \leq p \leq \infty\).
XV.14. C*-Algebras

C*-algebras are the nicest kinds of Banach algebras, and have a beautiful theory with profound applications. We will touch on only some of the basics of this extensive theory here; see [Bla06] for a much more comprehensive treatment.

XV.14.1. Banach *-Algebras and C*-Algebras

**XV.14.1.1. Definition.** An involution on a Banach algebra $A$ is a conjugate-linear isometric antiautomorphism of order two, usually denoted $x \mapsto x^*$. In other words, $(x + y)^* = x^* + y^*$, $(xy)^* = y^*x^*$, $(\lambda x)^* = \lambda x^*$, $(x^*)^* = x$, $\|x^*\| = \|x\|$ for all $x, y \in A$, $\lambda \in \mathbb{C}$. A Banach *-algebra is a Banach algebra with an involution.

An (abstract) C*-algebra is a Banach *-algebra $A$ satisfying the C*-axiom:

$$\|x^*x\| = \|x\|^2$$

for all $x \in A$.

The deceptively simple and innocuous C*-axiom turns out to be extremely powerful, forcing rigid structure on a C*-algebra. For example, it follows that the norm is completely determined by the algebraic structure and is thus unique (XV.14.2.34.), that *-homomorphisms of C*-algebras are automatically contractive (XV.14.1.11.), and that every C*-algebra can be isometrically represented as a concrete C*-algebra of operators (XV.14.4.35.). One obvious, but useful, consequence is that in a C*-algebra, $x^*x = 0$ implies that each $x_j = 0$ (XV.14.3.2. (i), XV.14.3.3.).

**XV.14.1.2.** In many older references, abstract C*-algebras were called $B^*$-algebras, with the name “C*-algebra” reserved for concrete C*-algebras. The term “C*-algebra,” first introduced in [Seg47] (for concrete C*-algebras, but viewed in a somewhat abstract manner), did not become universal until well after the publication of [Dix69] in 1964 (there were occasional references to “$B^*$-algebras” in the literature at least as late as 1980; see even [Lax02]). According to [DB86, p. 6], the “C” in “C*-algebra” originally meant “closed”, and not, as commonly believed, “continuous”, although the interpretation as standing for “continuous” is nicely in line with the modern point of view of C*-algebra theory as “noncommutative topology.”

The issue of terminology is clouded by the fact that several different (but ultimately equivalent) axiom schemes have been used for C*-algebras over the years. For example, it is easily seen that the C*-axiom, along with submultiplicativity of the norm, implies that the involution is isometric, so it is unnecessary to include isometry of the involution as an axiom. The C*-axiom has sometimes been replaced by the apparently weaker axiom that $\|x^*x\| = \|x\|\|x\|$ for all $x$. It turns out that this weaker axiom, along with the Banach algebra axioms, also implies isometry of the involution (a much harder result), so the weakened axiom is equivalent to the C*-axiom. See also XV.14.3.4.. See [DB86] for details about the C*-algebra axioms.

**XV.14.1.3. Examples.**

(i) $B(\mathcal{H})$ with the usual algebraic operations, involution, and operator norm, is a C*-algebra for any Hilbert space $\mathcal{H}$. If $\mathcal{H}$ is $n$-dimensional, we obtain that the $n \times n$ matrices $M_n = B(\mathbb{C}^n)$ form a C*-algebra with the usual involution (conjugate transpose) and operator norm. More generally, any norm-closed *-subalgebra of $B(\mathcal{H})$ is a C*-algebra; such C*-algebras are called concrete C*-algebras. For example, the algebra $K(\mathcal{H})$ of compact operators on a Hilbert space $\mathcal{H}$ is a concrete C*-algebra. If $\mathcal{H}$ is infinite-dimensional, $K(\mathcal{H})$ is nonunital; thus a C*-algebra need not be unital. We denote by $\mathbb{K}$ the C*-algebra of compact operators on a separable, infinite-dimensional Hilbert space.
(ii) Let $X$ be a locally compact Hausdorff space, and $C_0(X)$ the complex-valued continuous functions on $X$ vanishing at infinity. Give $C_0(X)$ its usual pointwise operations and supremum norm. Define an involution by $f^*(x) = \overline{f(x)}$. Then $C_0(X)$ is a commutative C*-algebra. In fact, every commutative C*-algebra is of this form (XV.14.2.5). $C_0(X)$ has a unit (identity) if and only if $X$ is compact; in this case, we usually write $C(X)$. More generally, if $B$ is a C*-algebra, then the set $C_0(X,B)$ of (norm-)continuous functions from $X$ to $B$ vanishing at infinity, with pointwise operations and supremum norm, is a C*-algebra. In particular, $C_0(X,M_n) \cong M_n(C_0(X))$ is a C*-algebra. (In fact, a matrix algebra over any C*-algebra is a C*-algebra (??); this is a special case of the tensor product for C*-algebras (??)).

(iii) Let $G$ be a locally compact topological group with (left) Haar measure $\mu$. Then $L^1(G,\mu)$ becomes a Banach *-algebra under convolution. It is not a C*-algebra unless $G$ is trivial. This example will be treated in more detail in ??.

XV.14.1.4. If $A$ is a Banach algebra and $J$ is a closed ideal ("ideal" will always mean "two-sided ideal" unless otherwise specified) in $A$, then the quotient norm makes $A/J$ into a Banach algebra. If $A$ is a Banach *-algebra and $J$ is a *-ideal (i.e. closed under *), then $A/J$ is a Banach *-algebra. It turns out that if $A$ is a C*-algebra and $J$ is a closed ideal in $A$, then $J$ is automatically a *-ideal and $A/J$ is a C*-algebra in the quotient norm (XV.14.3.14).

XV.14.1.5. A Banach algebra is separable if and only if it is countably generated as a Banach algebra (a countably generated algebra over $\mathbb{Q}(i)$ is countable). A C*-algebra is separable if and only if it is countably generated as a C*-algebra.

XV.14.1.1 Unitization

XV.14.1.1. A Banach algebra, even a C*-algebra, need not be unital (e.g. $K(H)$, $C_0(X)$ for $X$ non-compact). However, every nonunital Banach algebra $A$ can be embedded in a unital Banach algebra $A^\dagger$ (XV.10.1.6.). If $A$ is a Banach *-algebra, for $(a, \lambda) \in A^\dagger$ define $(a, \lambda)^* = (a^*, \overline{\lambda})$; this is an involution making $A^\dagger$ a (unital) Banach *-algebra. With the norm of XV.10.1.6. $A^\dagger$ is not a C*-algebra, but:

XV.14.1.2. Proposition. If $A$ is a C*-algebra, the operator norm on $A^\dagger$ as left multiplication operators on $A$, i.e.

$$\| (a, \lambda) \| = \sup \{ \| ab + \lambda b \| : \| b \| = 1 \}$$

is a C*-norm on $A^\dagger$ extending the norm of $A$.

Proof: $A^\dagger$ is a subalgebra of $B(A)$ with the operator norm; thus it is a normed algebra, and in particular the norm is submultiplicative. If $(a, \lambda) \in A^\dagger$ and $b \in A$, $\| b \| = 1$, then $(a, \lambda)^*(a, \lambda) = (a^* a + \lambda a^* + \overline{\lambda} a, \overline{\lambda})$, so

$$\| ab + \lambda b \|^2 = \| (ab + \lambda b)^*(ab + \lambda b) \| = \| (b^* a^* + \overline{\lambda} b^*)(ab + \lambda b) \|$$

$$= \| b^* (a^* a b + \lambda a b + \overline{\lambda} a b + \overline{\lambda} \lambda b) \| \leq \| a^* a b + \lambda a b + \overline{\lambda} a b + \overline{\lambda} \lambda b \| \leq \| (a, \lambda)^*(a, \lambda) \|$$

and taking the supremum over all $b$ we obtain $\| (a, \lambda) \|^2 \leq \| (a, \lambda)^*(a, \lambda) \|$, and the opposite inequality follows from submultiplicativity (first use this inequality and submultiplicativity to conclude that $\| (a, \lambda)^* = \ldots$
The algebra \( A^\dagger \) contains \( A \) as a (closed) ideal, and \( A^\dagger/A \cong \mathbb{C} \). If \( A \) is unital, \( A^\dagger \cong A \oplus \mathbb{C} \) as \( C^* \)-algebras (under the map \((a,\lambda) \mapsto (a - \lambda 1, \lambda))\).

**XV.14.1.3.** As an example, if \( X \) is a locally compact noncompact Hausdorff space, it is easy to see that \( C_0(X)^\dagger \cong C(X^\dagger) \), where \( X^\dagger \) is the one-point compactification of \( X \).

**XV.14.1.4.** A bounded homomorphism between the Banach algebras \( A \) and \( B \) extends uniquely to a bounded unital homomorphism from \( A^\dagger \) to \( B^\dagger \). A \( * \)-homomorphism between Banach \( * \)-algebras extends to a unital \( * \)-homomorphism between the unitizations.

### Special Elements in a \( C^* \)-Algebra

In a \( C^* \)-algebra, certain types of elements have standard names arising from operator theory, reflecting the types of operators they become when the \( C^* \)-algebra is represented as a concrete \( C^* \)-algebra of operators:

**XV.14.1.5.** **Definition.** Let \( A \) be a \( C^* \)-algebra and \( x \in A \). Then \( x \) is

- **self-adjoint** if \( x = x^* \).
- **normal** if \( x^*x = xx^* \).
- **a projection** if \( x = x^* = x^2 \).
- **a partial isometry** if \( x^*x \) is a projection.

If \( A \) is unital, then \( x \) is

- **an isometry** if \( x^*x = 1 \).
- **a coisometry** if \( xx^* = 1 \).
- **unitary** if \( x^*x = xx^* = 1 \).

**XV.14.1.6.** The self-adjoint elements of \( A \) form a closed real vector subspace \( A_{sa} \) of \( A \). (Note, however, that \( A_{sa} \) is not closed under multiplication unless \( A \) is commutative.) If \( A \) is unital, then the set of unitaries of \( A \) forms a group \( U(A) \). Every self-adjoint or unitary element is normal. If \( x \) is self-adjoint, then \( e^{ix} \) is unitary. 0 and 1 are projections; isometries (and, in particular, unitaries) are partial isometries. Coisometries are too; in fact, \( x \) is a partial isometry if and only if \( x = xx^*x \) [one direction is obvious; for the other, if \( x^*x \) is a projection and \( y = x - xx^*x \), then \( y^*y = 0 \), so \( y = 0 \)]. Thus if \( x \) is a partial isometry, then so is \( x^* \).

**XV.14.1.7.** If \( x \) is any element, then \( a = (x + x^*)/2 \) and \( b = (x - x^*)/2i \) are self-adjoint, and \( x = a + ib \); thus \( A_{sa} + iA_{sa} = A \). The elements \( a \) and \( b \) are called the **real** and **imaginary** parts of \( x \). It is obvious from their definitions that \( \|a\|, \|b\| \leq \|x\| \).
Norm and Spectrum

XV.14.1.8. It follows from the C*-axiom that every nonzero projection, and hence every nonzero partial isometry (in particular, every unitary), has norm 1. Thus, if \( x \) is unitary, then \( \sigma_A(x) \subseteq \{ \lambda : |\lambda| \leq 1 \} \). Since \( x^{-1} = x^* \) is also unitary and hence \( \sigma_A(x^{-1}) = \{ \lambda^{-1} : \lambda \in \sigma_A(x) \} \) is also contained in the unit disk, \( \sigma_A(x) \) is actually contained in the unit circle.

The next two results give important properties of the spectrum and norm for self-adjoint and, more generally, normal elements of a C*-algebra.

XV.14.1.9. Proposition. If \( x \) is self-adjoint, then \( \sigma_A(x) \subseteq \mathbb{R} \).

Proof: If \( x \) is self-adjoint, then \( e^{ix} \) is unitary, so
\[
\sigma_A(e^{ix}) = \{ e^{i\lambda} : \lambda \in \sigma_A(x) \}
\]
(XV.10.4.3. (i)) is contained in the unit circle.

XV.14.1.10. Proposition. If \( A \) is a C*-algebra and \( x \in A \) is normal (e.g. self-adjoint), then \( r(x) = \|x\| \), i.e. there is a \( \lambda \in \sigma_A(x) \) with \( |\lambda| = \|x\| \).

Proof: If \( x \) is self-adjoint, then from the C*-axiom \( \|x^2\| = \|x\|^2 \), and by iteration \( \|x^{2^n}\| = \|x\|^{2^n} \) for all \( n \). Thus \( r(x) = \lim_{n \to \infty} \|x^{2^n}\|^{2^{-n}} = \|x\| \). More generally, if \( y \) is normal, then
\[
r(y) = \lim_{n \to \infty} \|y^{2^n}\|^{2^{-n}} = \lim_{n \to \infty} \|(y^*)^{2^n}y^{2^n}\|^{2^{-n-1}}
\]
\[
= \lim_{n \to \infty} \|(y^*)^{2^n}y\|^{2^{-n-1}} = [r(y^*y)]^{1/2} = \|y^*y\|^{1/2} = \|y\| \).
\]

As a consequence, *-homomorphisms between C*-algebras are automatically contractive:

XV.14.1.11. Corollary. If \( A \) is a Banach *-algebra, \( B \) a C*-algebra, and \( \phi : A \to B \) a *-homomorphism, then \( \|\phi\| \leq 1 \).

Proof: If \( x \in A \), then \( \sigma_B(\phi(x)) \cup \{0\} \subseteq \sigma_A(x) \cup \{0\} \). Thus
\[
\|\phi(x)\|^2 = \|\phi(x^*x)\| = r(x^*x) \leq r(x^*x) \leq \|x^*x\| \leq \|x\|^2 .
\]

If \( A \) is a unital Banach algebra and \( B \) is a unital Banach subalgebra, and \( x \in B \), then it is not necessarily true that \( \sigma_B(x) = \sigma_A(x) \) (XV.10.3.11.). But if \( A \) is a C*-algebra, the spectra are the same:

1864
**XV.14.1.12.** Corollary. If $B$ is a unital $C^*$-subalgebra of a unital $C^*$-algebra $A$ and $x \in B$, then $\sigma_B(x) = \sigma_A(x)$. If $B$ is a general $C^*$-subalgebra of a general $C^*$-algebra $A$, and $x \in B$, then $\sigma_B(x) \cup \{0\} = \sigma_A(x) \cup \{0\}$.

Proof: This is true if $x = x^*$ by XV.10.3.12. and XV.14.1.9. For general $x$, if $x - \lambda 1$ is invertible in $A$, then so are $(x - \lambda 1)^*(x - \lambda 1)$ and $(x - \lambda 1)(x - \lambda 1)^*$. Thus they are invertible in $B$, so $x - \lambda 1$ is left and right invertible in $B$. \hfill \Diamond

**XV.14.2.** Commutative $C^*$-Algebras and Continuous Functional Calculus

The main result of this section is Gelfand’s Theorem (sometimes called the Gelfand-Naimark Theorem, although this name is more commonly used for a different theorem, cf. XV.14.4.35.):

**XV.14.2.1.** Theorem. (i) Let $A$ be a commutative unital $C^*$-algebra. Then there is a compact Hausdorff space $X$ such that $A$ is isometrically $*$-isomorphic to $C(X)$.

(ii) $X$ can be recovered from $C(X)$ either as $\text{Prim}(C(X)) = \text{Maxspec}(C(X))$ (XI.7.8.16., XV.14.4.39.) with the hull-kernel topology, or as the set $\overline{C(X)}$ of unital $*$-homomorphisms from $C(X)$ to $\mathbb{C}$ with the topology of elementwise convergence (weak-* topology). In particular, if $C(X) \cong C(Y)$, then $X$ is homeomorphic to $Y$.

(iii) If $X$ and $Y$ are compact Hausdorff spaces and $\phi : X \to Y$ is a continuous function, then $\hat{\phi} : C(Y) \to C(X)$ defined by $\hat{\phi}(f) = f \circ \phi$ is a unital $*$-homomorphism, and every unital $*$-homomorphism arises in this manner.

(iv) The correspondences $X \leftrightarrow C(X)$ and $\phi \leftrightarrow \hat{\phi}$ give a contravariant category equivalence between the category of compact Hausdorff spaces and continuous functions and the category of commutative unital $C^*$-algebras and unital $*$-homomorphisms.

**XV.14.2.2.** As a result, all the topology of compact Hausdorff spaces is encoded in the corresponding $C^*$-algebras, and in principle all of topology (of compact Hausdorff spaces) can be done on the $C^*$-algebra level. In practice, this is certainly not the way to do most topology! However, the point of view that a general $C^*$-algebra is a “noncommutative topological space” is very powerful and underlies much of the modern theory and applications of $C^*$-algebras. Interaction with topology has gone both ways: ideas and techniques from topology, particularly $K$-theory, have proved fundamentally useful in operator algebra theory, and on the other hand operator algebra methods have been of significant use in some parts of topology.

**XV.14.2.3.** There is much more fine structure associated with the category equivalence. For example, if $X$ and $Y$ are compact Hausdorff spaces and $\phi : X \to Y$ is continuous, the $*$-homomorphism $\hat{\phi} : C(Y) \to C(X)$ is surjective (i.e. a quotient map, cf. XV.14.3.15.) if and only if $\phi$ is injective, and $\hat{\phi}$ is injective if and only if $\phi$ is surjective (i.e. a quotient map, cf. XI.8.3.13.). The closed ideals of $C(X)$ are in natural one-one correspondence with the closed subsets of $X$: the ideal of $C(X)$ corresponding to a closed subset $Z$ of $X$ is the set $J_Z$ of continuous functions from $X$ to $\mathbb{C}$ vanishing on $Z$. The quotient $C^*$-algebra $C(X)/J_Z$ is naturally isomorphic to $C(Z)$ via restriction of functions (the Tietze Extension Theorem () for compact Hausdorff spaces is included in this fact, since it implies that every element of $C(Z)$ can be extended to an element of $C(X)$, cf. XV.8.3.12.).
The Nonunital Case

XV.14.2.4. There is a nonunital version of XV.14.2.1., but it cannot be stated quite so cleanly. The commutative C*-algebras are easy to describe: they are all of the form \( C_0(X) \), the complex-valued continuous functions vanishing at infinity on a locally compact Hausdorff space \( X \). But the *-homomorphisms from \( C_0(X) \) to \( C_0(Y) \) do not generally correspond to continuous functions from \( Y \) to \( X \). For example, a constant function from \( \mathbb{R} \) to \( \mathbb{R} \) does not give a *-homomorphism from \( C_0(\mathbb{R}) \) to \( C_0(\mathbb{R}) \), and conversely the zero homomorphism from \( C_0(\mathbb{R}) \) to \( C_0(\mathbb{R}) \) does not come from a continuous function from \( \mathbb{R} \) to \( \mathbb{R} \) (there are many less trivial examples too). Instead, the *-homomorphisms from \( C_0(X) \) to \( C_0(Y) \) can be described as corresponding to proper () continuous functions from open subsets of \( Y \) to \( X \).

Alternatively, it is much cleaner to consider pointed compact Hausdorff spaces \((Z, *)\), where \( Z \) is a compact Hausdorff space and \(*\) is a fixed basepoint in \( Z \). There is a natural correspondence with locally compact Hausdorff spaces: if \( X \) is locally compact, associate \( X \) with \((X^\dagger, \infty)\), where \( X^\dagger \) is the one-point compactification of \( X \). Conversely, if \((Z, *)\) is a pointed compact Hausdorff space, associate the locally compact Hausdorff space \( Z \setminus \{ * \} \). The pointed space \((Z, *)\) corresponds to a compact Hausdorff space if and only if \(*\) is an isolated point in \( Z \). Then, if \( X \) is locally compact, we have that \( C_0(X) \) is naturally isomorphic to the ideal of \( C(X^\dagger) \) consisting of functions which are zero at \( \infty \), i.e. if \((Z, *)\) is a pointed compact Hausdorff space, the associated commutative C*-algebra is the algebra of complex-valued continuous functions on \( Z \) vanishing at \(*\). It is then easy to see that the *-homomorphisms from \( C_0(X) \) to \( C_0(Y) \) exactly correspond to continuous functions from \( Y^\dagger \) to \( X^\dagger \) sending \( \infty \) to \( \infty \), i.e. basepoint-preserving continuous maps. (For example, the extreme case of the zero homomorphism corresponds to the constant function from \( Y^\dagger \) to \( X^\dagger \) sending everything to \( \infty \).) Thus, as a slight extension of XV.14.2.1., we obtain:

XV.14.2.5. **Theorem.**

(i) Let \( A \) be a commutative C*-algebra. Then there is a compact Hausdorff space \( Z \) and point \(* \in Z \) such that \( A \) is isometrically *-isomorphic to \( C^*(Z, *) := C_0(X) \), where \( X = Z \setminus \{ * \} \) is locally compact.

(ii) \( X \) can be recovered from \( C_0(X) \) as \( \text{Prim}(C_0(X)) = \text{Maxspec}(C_0(X)) \) with the hull-kernel topology. \( Z \) can be recovered from \( C^*(Z, *) \) as the set \( C^*(Z, *) \) of *-homomorphisms from \( C^*(Z, *) \) to \( \mathbb{C} \) with the topology of elementwise convergence (weak-* topology), with * corresponding to the zero homomorphism. In particular, if \( C_0(X) \cong C_0(Y) \) for locally compact Hausdorff spaces \( X \) and \( Y \), then \( X \) is homeomorphic to \( Y \).

(iii) If \((Z, *_Z)\) and \((W, *_W)\) are pointed compact Hausdorff spaces and \( \phi : Z \to W \) is a pointed continuous function (i.e. \( \phi \) is continuous and \( \phi(*_Z) = *_W \)), then \( \hat{\phi} : C^*(W, *_W) \to C^*(Z, *_Z) \) defined by \( \hat{\phi}(f) = f \circ \phi \) is a *-homomorphism, and every *-homomorphism arises in this manner.

(iv) The correspondences \((Z, *) \leftrightarrow C^*(Z, *)\) and \( \phi \leftrightarrow \hat{\phi} \) give a contravariant category equivalence between the category of pointed compact Hausdorff spaces and pointed continuous functions and the category of commutative C*-algebras and *-homomorphisms.

XV.14.2.6. The fine structure described in XV.14.2.3. persists in this setting, but care must be exercised in phrasing it. Details are left to the interested reader.

**Proof of the Theorem**

We now prove Gelfand’s Theorem. We will only prove the unital version XV.14.2.1.; the proof of the nonunital version XV.14.2.5. is a straightforward variation, but the additional technicalities detract from
an intuitive appreciation of the argument. We first prove (ii) and (iii), which are fairly elementary (given Urysohn’s Lemma), and then return to (i) which requires some nontrivial Banach algebra theory. The proof will actually give some fine structure beyond the statement of the theorem.

We begin with a characterization of the maximal ideals of a C(X).

**XV.14.2.7. Lemma.** Let X be a compact Hausdorff space. For each p ∈ X, the set Mp of all functions in C(X) which vanish at p is a maximal ideal in C(X), and every maximal ideal of C(X) is of this form. If p, q ∈ X, p ≠ q, then Mp ≠ Mq.

**Proof:** Let p ∈ X. It is obvious that C(X)/Mp is isomorphic to C via the correspondence f → f(p). Since the quotient is a field, Mp is a maximal ideal in C(X). Conversely, suppose M is a maximal ideal in C(X), and suppose that there is no point of X for which all functions in M vanish. For each p ∈ X, there is an fp ∈ M with fp(p) ≠ 0. The set Up = {x ∈ X : fp(x) ≠ 0} is an open set in X containing p, and {Up : p ∈ X} is an open cover of X. By compactness, there is a finite subcover {Up1, ..., Upn}. Set

\[ g = f_p1 f_{p1} + \cdots + f_{pn} f_{pn} ∈ M \]

and note that g(x) > 0 for every x ∈ X. Thus g is invertible in C(X), a contradiction since a proper ideal cannot contain an invertible element. Thus there is some p ∈ X at which all functions in M vanish, so M ⊆ Mp, and M = Mp by maximality. Since functions in C(X) separate points of X by Urysohn’s Lemma, and constant functions are in C(X), we have Mp ≠ Mq if p ≠ q.

**XV.14.2.8.** Thus p ↔ Mp is a one-one correspondence between X and Maxspec(C(X)). If φ is a unital *-homomorphism from C(X) to C, then the kernel of φ is a maximal ideal of C(X), hence of the form Mp for some p ∈ X, and φ must just be evaluation at p, i.e. φ(f) = ̂f(p) = f(p). Thus there is also a natural one-one correspondence between X and Maxspec(C(X)), so as sets X, Maxspec(C(X)), and C(X) can be naturally identified.

**XV.14.2.9.** We now show how the topology of X can be recovered as the hull-kernel topology on Maxspec(C(X)). If E is a subset of X ≃ Maxspec(C(X)), recall that the closure E of E in the hull-kernel topology consists of all Mp such that J := ∩q∈E Mq ⊆ Mp. The ideal J clearly consists of all f such that f(q) = 0 for all q ∈ E. If E is the closure in X of E regarded as a subset of X, and p ∈ E, then every f ∈ J vanishes at p by continuity, so Mp ∈ E. But if p ∈ E, by Urysohn’s Lemma there is an f ∈ C(X) such that f = 0 on E and f(p) = 1; this f is in J but not Mp, so Mp /∈ E. Thus E = E and the correspondence p ↔ Mp is a homeomorphism.

**XV.14.2.10.** Now consider the topology of pointwise convergence on C(X): ̂p_i → ̂p if and only if ̂p_i(f) → ̂f(f) for every f ∈ C(X), i.e. f(p_i) → f(p) for all f ∈ C(X). If p_i → p in X, this condition is clearly satisfied. If p_i ∉ p in X, there is an open neighborhood U of p such that p_i /∈ U for a cofinal set of i. By Urysohn’s Lemma there is a function f ∈ C(X) with f(p) = 1 and f = 0 on Uc, and f(p_i) /→ f(p) for this f, i.e. ̂p_i ∉ ̂p. Thus p ↔ ̂p is a homeomorphism.

1867
XV.14.2.11. Finally, we observe that every primitive ideal (XV.14.4.39.) of \( C(X) \) is of the form \( M_p \). By definition, primitive ideals are the kernels of irreducible representations. Every irreducible representation of a commutative C*-algebra is one-dimensional, hence a unital *-homomorphism to \( \mathbb{C} \), hence \( \hat{p} \) for some \( p \in X \), and its kernel is \( M_p \). Alternatively, every primitive ideal is a closed prime ideal (XV.14.4.11.), so it suffices to prove the following fact:

XV.14.2.12. Proposition. Let \( X \) be a compact Hausdorff space, and \( J \) a prime ideal of \( C(X) \). Then there is a unique \( p \in X \) such that \( O_p \subseteq J \subseteq M_p \), where

\[
O_p = \{ f \in C(X) : f = 0 \text{ in a neighborhood of } p \}.
\]

The straightforward proof is left to the reader (cf. [GJ76, 7.15]). Since \( O_p \) is dense in \( M_p \), it follows that every closed prime ideal of \( C(X) \) is maximal.

XV.14.2.13. This completes the proof of (ii). We now turn to (iii). Let \( X \) and \( Y \) be compact Hausdorff spaces. If \( \phi : X \to Y \) is continuous, it is obvious that \( \phi : C(Y) \to C(X) \), defined by \( \phi(f) = f \circ \phi \), is well defined and is a unital *-homomorphism. Conversely, let \( \alpha \) be a unital *-homomorphism from \( C(Y) \) to \( C(X) \). For any \( p \in X \), \( \hat{p} \circ \alpha \) is a unital *-homomorphism from \( C(Y) \) to \( \mathbb{C} \), hence \( \hat{p} \circ \alpha = \hat{q} \) for a unique \( q \in Y \). Set \( \phi(p) = q \). This defines a function \( \phi : X \to Y \), and \( \alpha(f) = f \circ \phi \) for all \( f \in C(X) \), and in particular \( f \circ \phi \in C(X) \) for all \( f \in C(Y) \). It remains only to show that \( \phi \) is continuous. The argument is similar to the one in XV.14.2.10.. If \( \phi \) is discontinuous at \( p \in X \), then there is a net \( (p_i) \) in \( X \) with \( p_i \to p \) in \( X \) but \( \phi(p_i) \not\to \phi(p) \) in \( Y \). There is an open neighborhood \( U \) of \( \phi(p) \) in \( Y \) with \( \phi(p_i) \notin U \) for a cofinal set of \( i \). By Urysohn’s Lemma there is an \( f \in C(Y) \) with \( f(\phi(p)) = 1 \) and \( f = 0 \) on \( U^c \). Then \( f(\phi(p_i)) \not\to f(\phi(p)) \), i.e. \( f \circ \phi \) is not continuous at \( p \), a contradiction. Thus \( \phi \) is continuous.

Now we tackle (i). We will need to bring in some Banach algebra machinery. If \( A \) is a unital commutative Banach algebra, write \( \hat{A} \) for the set of unital homomorphisms from \( A \) to \( \mathbb{C} \). Recall the definitions of spectrum and spectral radius ()

XV.14.2.14. Proposition. Let \( A \) be a unital commutative Banach algebra, and \( \phi \in \hat{A} \). If \( x \in A \), then \( \phi(x) \in \sigma_A(x) \). In particular, \( |\phi(x)| \leq r_\sigma(x) \leq ||x|| \). Thus \( \phi \) is bounded and \( ||\phi|| = 1 \).

Proof: Since \( \phi(x - \phi(x))1 = 0 \), \( x - \phi(x) \) is not invertible in \( A \). The inequality in the second statement is then obvious, so \( ||\phi|| \leq 1 \). But \( \phi(1) = 1 \), so \( ||\phi|| = 1 \).

XV.14.2.15. Corollary. Let \( A \) be a unital commutative C*-algebra, and \( \phi \in \hat{A} \). Then \( \phi \) is a *-homomorphism.

Proof: It suffices to show that if \( x = x^* \in A \), then \( \phi(x) \in \mathbb{R} \). But \( \sigma_A(x) \subseteq \mathbb{R} \) by XV.14.1.9..
**Proposition.** Let $A$ be a unital commutative Banach algebra and $x \in A$. If $\lambda \in \sigma_A(x)$, then there is a $\phi \in \hat{A}$ with $\phi(x) = \lambda$.

**Proof:** Since $x - \lambda 1$ is not invertible in $A$, it generates a proper ideal of $A$ which is contained in a maximal ideal $M$ of $A$. Then $M$ is closed (XV.10.2.17.), so $A/M$ is a Banach algebra which is a field. Thus $A/M \cong \mathbb{C}$ (XV.10.3.26.), and the quotient map $\phi : A \to A/M$ can be regarded as an element of $\hat{A}$ with $\phi(x - \lambda 1) = 0$. 

**Proposition.** Let $A$ be a unital commutative Banach algebra. By XV.14.2.14., $\hat{A}$ can be regarded as a subset of the closed unit ball $B$ of $A^*$. It is easily checked that $\hat{A}$ is closed in $B$ in the weak-* topology (topology of pointwise convergence on $A$). Since $B$ is a compact Hausdorff space in the weak-* topology (XV.8.2.3.), $\hat{A}$ is a compact Hausdorff space in the topology of pointwise convergence on $A$. If $x \in A$, define $\hat{x} : \hat{A} \to \mathbb{C}$ by $\hat{x}(\phi) = \phi(x)$ for $\phi \in \hat{A}$. Then $\hat{x} \in C(\hat{A})$.

**Definition.** The map $\gamma : A \to C(\hat{A})$ defined by $\gamma(x) = \hat{x}$ is a unital homomorphism, called the Gelfand transform of $A$.

We have that $\gamma$ is bounded and $\|\gamma\| = 1$ by XV.14.2.14.. The Gelfand transform is not injective or surjective for general unital commutative Banach algebras. However:

**Lemma.** Let $A$ be a unital commutative C*-algebra. Then the Gelfand transform $\gamma : A \to C(\hat{A})$ is an isometric *-isomorphism.

**Proof:** It follows from XV.14.2.15. that $\gamma$ is a *-homomorphism. Let $x \in A$. Then by (XV.14.2.15.) there is a $\lambda \in \sigma_A(x^*x)$ with $|\lambda| = \|x^*x\| = \|x\|^2$, and hence by XV.14.2.16. there is a $\phi \in \hat{A}$ with $|\phi(x)|^2 = |\phi(x^*x)| = \|x^*x\| = \|x\|^2$. Since $\|\gamma\| = 1$, $\gamma$ is isometric, and hence $A$ can be regarded as the *-subalgebra $\gamma(A)$ of $C(\hat{A})$, which is complete and thus closed. The subalgebra $\gamma(A)$ clearly separates points of $A$ and contains the constant functions, so $\gamma(A) = C(\hat{A})$ by the Stone-Weierstrass Theorem (XV.8.2.3.).

**Applications**

Probably the most important application of Gelfand’s Theorem is continuous functional calculus, discussed in XV.14.2.25.. But here are a couple of other applications of some interest.
XV.14.2.21. Let $X$ be a topological space. Let $BC(X)$ be the set of bounded continuous complex-valued functions on $X$. With pointwise addition and multiplication, pointwise complex conjugation as adjoint, and supremum norm, $BC(X)$ becomes a unital commutative $C^*$-algebra. Thus $BC(X) \cong C(Y)$ for a compact Hausdorff space $Y$. Each point of $X$ gives, by point evaluation, an element of $BC(X)$ and hence a point of $Y$, and it is easy to check that the induced map $\phi$ from $X$ to $Y$ is continuous. If $X$ is completely regular, a simple argument yields that $\phi$ is a homeomorphism onto its image, so $X$ can be identified with a subspace of $Y$. In this case, it is straightforward to identify $\beta X$ as $BC(X)$ or $\text{Prim}(BC(X))$ is perhaps the most elegant and efficient construction of the Stone-Čech compactification.

XV.14.2.22. Here is another interesting and modestly useful application. Give $A = L^\infty(\mathbb{R})$ pointwise addition, multiplication, and complex conjugation, and the usual essential supremum norm. Then $A$ becomes a unital commutative $C^*$-algebra, so $A \cong C(Y)$ for some compact Hausdorff space $Y$. The space $Y$ is rather horrendous as a topological space: it is extremally disconnected and not first countable. But it is occasionally useful to realize $A$ as an algebra of continuous functions on a bad space instead of as measurable functions on a nicer space. The same construction can be done for $L^\infty(X, A, \mu)$ for any measure space $(X, A, \mu)$. If done for $\mathbb{N}$ with counting measure one obtains $Y = \beta \mathbb{N}$ as in XV.14.2.21. These $L^\infty(X, A, \mu)$ are commutative von Neumann algebras (XV.3.5.1.) when realized as algebras of multiplication operators on $L^2(X, A, \mu)$.

Continuous Functional Calculus

The classification of commutative $C^*$-algebras leads to a far-reaching and powerful extension of holomorphic functional calculus (XV.10.4.).

XV.14.2.23. Let $A$ be a unital $C^*$-algebra, and $x$ a normal (e.g. self-adjoint or unitary) element of $A$, i.e. $x^*x = xx^*$. Let $B$ be the $C^*$-subalgebra of $A$ generated by $x$ and $1$: $B$ is the closure of the set of polynomials in $x$ and $x^*$. Then $B$ is a unital commutative $C^*$-algebra, and thus $B \cong C(X)$ for a compact Hausdorff space $X = \hat{B}$.

XV.14.2.24. The space $\hat{B}$ can be easily described. If $\phi \in \hat{B}$, then $\phi(x) \in \sigma_B(x) = \sigma_A(x)$ (XV.14.2.14., XV.14.1.12.). Conversely, for each $\lambda \in \sigma_A(x)$ there is a unique $\phi_\lambda \in \hat{B}$ with $\phi_\lambda(x) = \lambda$ by XV.14.2.16. [for uniqueness, if $\phi, \psi \in \hat{B}$ with $\phi(x) = \psi(x) = \lambda$, then $\phi(x^*) = \psi(x^*) = \bar{\lambda}$ by XV.14.2.15., so $\phi(p(x, x^*)) = \psi(p(x, x^*)) = p(\lambda, \bar{\lambda})$ for any polynomial $p$ of two variables with complex coefficients; such elements are dense in $B$, so $\phi = \psi$.] Thus

$$\hat{B} = \{\phi_\lambda : \lambda \in \sigma_A(x)\}. $$

More precisely, the function $\hat{x} : \hat{B} \to \mathbb{C}$ defined by $\hat{x}(\phi_\lambda) = \lambda$ is a bijection from $\hat{B}$ to $\sigma_A(x)$. Since $\hat{x}$ is continuous and $\hat{B}$ and $\sigma_A(x)$ are compact Hausdorff spaces, it is a homeomorphism. We thus have:

XV.14.2.25. Theorem. [Continuous Functional Calculus, Unital Version] Let $A$ be a unital $C^*$-algebra, $x$ a normal element of $A$, and $B$ the $C^*$-subalgebra of $A$ generated by $x$ and $1$. Then there is an isometric *-isomorphism $\theta : C(\sigma_A(x)) \to B$ with $\theta(1) = x$, where $\theta(\lambda) = \lambda$ for $\lambda \in \sigma_A(x) \subseteq \mathbb{C}$.
 XV.14.2.26. We have that \( \theta \) is just \((x)^{-}\). If \( f \in C(\sigma_A(x)) \), we normally write \( f(x) \) instead of \( \theta(f) \); thus \( \iota(x) = x \). This notation is consistent with previous notation: if \( p \) is a polynomial with complex coefficients, then \( p(x) \) is the element of \( B \) defined in XV.10.4.1. More generally, if \( f \) is holomorphic on an open disk containing \( \sigma_A(x) \), then \( f(x) \) is the element defined by a power series as in XV.10.4.3. Thus “functional calculus” can be thought of as a way to apply suitable functions to elements of a C*-algebra. Note that holomorphic functional calculus is defined for arbitrary elements of a C*-algebra (or even a general Banach algebra), but continuous functional calculus only works for normal elements of a C*-algebra.

There is a nonunital version of continuous functional calculus, which can be easily proved from the unital version by simply adding a unit (XV.14.1.2.). If \( X \) is a compact subset of \( \mathbb{C} \), write \( C_0(X) \) for the set of continuous functions \( f : X \cup \{0\} \to \mathbb{C} \) with \( f(0) = 0 \).

 XV.14.2.27. Theorem. [Continuous Functional Calculus, General Version] Let \( A \) be a C*-algebra, \( x \) a normal element of \( A \), and \( B \) the C*-subalgebra of \( A \) generated by \( x \). Then there is an isometric *-isomorphism \( \theta : C_0(\sigma_A(x)) \to B \) with \( \theta(\iota(x)) = x \), where \( \iota(\lambda) = \lambda \) for \( \lambda \in \sigma_A(x) \cup \{0\} \subseteq \mathbb{C} \).

Note that if \( 0 \notin \sigma_A(x) \), then \( A \) is automatically unital and \( 1 \) is in the C*-subalgebra of \( A \) generated by \( x \).

The following properties of functional calculus are easy consequences:

 XV.14.2.28. Proposition. Let \( A \) be a C*-algebra and \( x \in A \) a normal element. Then

(i) For any \( f \in C_0(\sigma_A(x)) \), \( \sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\} \).

(ii) If \( f \in C_0(\sigma_A(x)) \) and \( g \in C_0(f(\sigma_A(x))) = C_0(\sigma_A(f(x))) \), so \( g \circ f \in C_0(\sigma_A(x)) \), then \( (g \circ f)(x) = g(f(x)) \).

(iii) If \( f_n, f \in C_0(\sigma_A(x)) \) and \( f_n \to f \) uniformly on \( \sigma_A(x) \), then \( f_n(x) \to f(x) \).

(iv) If \( B \) is a C*-algebra and \( \phi : A \to B \) a *-homomorphism, then \( \phi(f(x)) = f(\phi(x)) \) for any \( f \in C_0(\sigma_A(x)) \).

There is a useful joint continuity generalization of (iii):

 XV.14.2.29. Proposition. Let \( Y \) be a compact subset of \( \mathbb{C} \), and \( (f_n) \) a sequence of elements of \( C_0(Y) \) converging uniformly on \( Y \) to \( f \). Let \( A \) be a C*-algebra, \( (x_n) \) a sequence of normal elements of \( A \) with \( x_n \to x \in A \) and \( \sigma_A(x_n) \subseteq Y \) (so \( \sigma_A(x) \subseteq Y \) ). Then \( f_n(x_n) \to f(x) \).

Proof: Let \( \epsilon > 0 \), and approximate \( f \) uniformly on \( Y \) within \( \epsilon/4 \) by a function of the form \( g(\lambda) = p(\lambda, \bar{\lambda}) \), where \( p \) is a polynomial of two variables (XV.8.3.4.); then \( g(x) = p(x, x^*) \), so

\[
\|f_n(x_n) - f(x)\| \leq \|f_n(x_n) - f(x_n)\| + \|f(x_n) - g(x_n)\| + \|p(x_n, x_n^*) - p(x, x^*)\| + \|g(x) - f(x)\|.
\]

If \( n \) is large enough that \( |f_n - f| \) is uniformly less than \( \epsilon/4 \) on \( Y \), the first term is less than \( \epsilon/4 \); the second and fourth terms are \( < \epsilon/4 \) by choice of \( g \), and the third term goes to 0 as \( n \to \infty \) by continuity of addition, multiplication, and involution. Thus \( \|f_n(x_n) - f(x)\| < \epsilon \) for all sufficiently large \( n \).

This result is often used when \( (f_n) \) is a constant sequence.

Another simple consequence of functional calculus is:
**XV.14.2.30. Corollary.** Let $A$ be a C*-algebra, and $x$ a normal element of $A$. Then

- $x$ is self-adjoint if and only if $\sigma_A(x) \subseteq \mathbb{R}$.
- $x$ is unitary if and only if $\sigma_A(x) \subseteq \{\lambda : |\lambda| = 1\}$.
- $x$ is a projection if and only if $\sigma_A(x) \subseteq \{0, 1\}$.

**Proof:** (i): Let $f(\lambda) = \text{Re}(\lambda) = \frac{1}{2}(\lambda + \bar{\lambda})$. Then $f(x) = \frac{1}{2}(x + x^*)$ is the real part of $x$. We have $f(x) = x$ if and only if $f = \iota$ on $\sigma_A(x)$, i.e. if and only if $\sigma_A(x) \subseteq \mathbb{R}$. The other parts are similar.

**Uniqueness of Norm on a C*-Algebra**

Another important consequence of Gelfand’s Theorem is that the norm on a C*-algebra is unique in a strong sense. Thus C*-algebras are quite rigid.

We first prove one of the fine structure properties of commutative C*-algebras (XV.14.2.3.).

**XV.14.2.31. Lemma.** Let $A$ and $B$ be commutative C*-algebras, and let $\alpha : A \rightarrow B$ be a *-homomorphism. If $\alpha$ is injective, it is isometric.

**Proof:** Write $A = C^*(Z, \ast)$ and $B = C^*(W, \ast)$. We have $\alpha = \hat{\phi}$ for a pointed continuous function $\phi : (W, \ast) \rightarrow (Z, \ast)$. If $\alpha$ is injective, then $\phi$ is surjective: set $Y = \phi(W)$ and suppose there is a $p \in Z \setminus Y$. By Urysohn’s Lemma there is an $f \in C(Z)$ with $f = 0$ on $Y$ and $f(p) = 1$. Then $f \in C^*(Z, \ast)$ since $\ast \in Y$, and $\alpha(f) = 0$, contradicting that $\alpha$ is injective.

Thus, if $f \in C^*(Z, \ast)$, we have

$$\|f\| = \sup_{p \in Z} |f(p)| = \sup_{q \in W} |f(\phi(q))| = \|\alpha(f)\|.$$  

**XV.14.2.32. Corollary.** Let $A$ and $B$ be C*-algebras, $\alpha : A \rightarrow B$ an injective *-homomorphism. Then $\alpha$ is isometric.

**Proof:** If $x \in A$, then $\|x\|^2 = \|x^*x\|$ and $\|\alpha(x)\|^2 = \|\alpha(x^*x)\|$, so replacing $A$ and $B$ by $C^*(x^*x)$ and $C^*(\alpha(x^*x))$ respectively, it suffices to assume $A$ and $B$ are commutative, where the result follows from the lemma.

**XV.14.2.33. Corollary.** If $A$ is a C*-algebra with respect to a norm $\| \cdot \|$, then there is no other C*-norm $\| \cdot \|'$ on $A$ (with the same algebraic structure and involution), even incomplete.

**Proof:** Let $B$ be the completion of $A$ with respect to $\| \cdot \|'$. Then $B$ is a C*-algebra, and the natural injective *-homomorphism from $(A, \| \cdot \|)$ to $B$ is isometric.
XV.14.2.34. A \(\ast\)-algebra thus has at most one norm making it a (complete) \(C^*\)-algebra; if it has one, it has no other (even incomplete) norm satisfying the \(C^*\)-axiom (XV.14.2.33). [A \(C^*\)-algebra can have other \(C^*\)-seminorms; and a \(\ast\)-algebra (e.g. a polynomial ring) can have many different (incomplete) norms satisfying the \(C^*\)-axiom.] Note that by XV.14.1.10, the norm in a \(C^*\)-algebra is completely determined by the algebraic structure: if \(x \in A\), then
\[
\|x\| = \|x^*x\|^{1/2} = [r_\sigma(x^*x)]^{1/2}
\]
and \(r_\sigma(x^*x)\) is a purely algebraic quantity.

XV.14.3. Algebraic and Order Theory of \(C^*\)-Algebras

Positive Elements in a \(C^*\)-Algebra

XV.14.3.1. Definition. Let \(A\) be a \(C^*\)-algebra. An element \(x \in A\) is positive if \(x = x^*\) and \(\sigma_A(x) \subseteq [0, \infty)\). The set of positive elements in \(A\) is denoted \(A_+\). If \(x \in A_+\), we write \(x \geq 0\).

Note that by XV.14.1.12, the property of being positive is independent of the containing \(C^*\)-algebra, i.e. if \(B\) is a \(C^*\)-subalgebra of \(A\), then \(B_+ = B \cap A_+\). A positive element of \(C_0(X)\) is just a function taking only nonnegative real values.

The following facts are obvious from functional calculus:

XV.14.3.2. Proposition. Let \(A\) be a \(C^*\)-algebra and \(x, y \in A\). Then

(i) If \(x \geq 0\) and \(-x \geq 0\), then \(x = 0\).

(ii) If \(x\) is normal, then \(x^*x \geq 0\). In particular, if \(x = x^*\), then \(x^2 \geq 0\).

(iii) If \(x \geq 0\), then \(\|x\| = \max\{\lambda : \lambda \in \sigma(x)\}\).

(iv) If \(x = x^*\) and \(\|x\| \leq 2\), then \(x \geq 0\) if and only if \(\|1 - x\| \leq 1\) (in \(A\)).

(v) If \(x, y \geq 0\) and \(xy = yx\), then \(x + y \) and \(xy\) are positive.

(vi) If \(x = x^*\), then there is a unique decomposition \(x = x_+ - x_-\), where \(x_+, x_- \geq 0\), and \(x_+x_- = 0\). We have \(x_+, x_- \in C^*(x)\). \([x_+ = f(x)\) and \(x_- = g(x)\), where \(f(t) = \max(t, 0)\) and \(g(t) = -\min(t, 0)\).] Thus every element of \(A\) is a linear combination of four positive elements.

(vii) Every positive element of a \(C^*\)-algebra has a unique positive square root. More generally, if \(x \geq 0\) and \(\alpha\) is a positive real number, there is a positive element \(x^\alpha \in C^*(x)_+\); these elements satisfy \(x^\alpha x^\beta = x^{\alpha+\beta}\), \(x^1 = x\), and \(\alpha \mapsto x^\alpha\) is continuous. If \(x\) is invertible \(x^\alpha\) is also defined for \(\alpha \leq 0\).

(viii) \((x, \lambda) \geq 0\) in \(A^1\) if and only if \(x = x^*\) and \(\lambda \geq \|x_\|\).

For (vii), set \(x^\alpha = g_\alpha(t)\), where \(g_\alpha(t) = t^\alpha\). Then \(g_{\alpha-1}(x^\alpha) = x\). If \(b \in A_+\) with \(g_{\alpha-1}(b) = x\), then \(x\) commutes with \(b\), hence \(x^\alpha\) commutes with \(b\) since it is a limit of polynomials in \(x\); thus \(b = x^\alpha\) in \(C^*(x^\alpha, b) \subseteq A\) (we clearly have uniqueness in a commutative \(C^*\)-algebra).

Two crucial facts about positive elements are significant generalizations of (ii) and (v):
**Proposition.** Let $A$ be a C*-algebra. Then

(i) $A_+$ is a closed cone (.) in $A$; in particular, if $x, y \geq 0$ then $x + y \geq 0$.

(ii) If $x \in A$, then $x^*x \geq 0$.

**Proof:** To prove (i), since $\mathbb{R}_+A_+ \subseteq A_+$ it suffices to show that $A_+ \cap B$ is a closed convex set, where $B = B_1(A)$ is the closed unit ball in $A$. But $A_+ \cap B$ is the intersection of the closed convex sets $A_{sa}, B$, and \( \{ x : \|1 - x\| \leq 1 \} \) (XV.14.3.2 (iv)).

The proof of property (ii) uses the fact that if $x = a + ib \in A$ with $a, b \in A_{sa}$, then $x^*x \geq 0$ if and only if $xx^* \geq 0$ by XV.10.3.18., and

\[ x^*x + xx^* = 2a^2 + 2b^2 \geq 0 \]

by (i) and XV.14.3.2.(ii). If $x^*x = c = c_+ - c_-$ as in XV.14.3.2.(vi) and $y = xc_-$, then $-y^*y = c_-^2 \geq 0$, so

\[ yy^* = (y^*y + yy^*) + (-y^*y) \geq 0 \]

and thus $y^*y \geq 0$. So $y^*y = 0$ by XV.14.3.2.(i), $c_- = 0$, $x^*x \geq 0$. \( \square \)

**Corollary.** If $A$ is a C*-algebra and $a \in A_+$, then $x^*ax = (a^{1/2}x)^*(a^{1/2}x) \geq 0$ for any $x \in A$.

By XV.14.3.2.(vi), there are “many” positive elements of a C*-algebra. Unital C*-algebras also have “many” unitaries:

**Proposition.** Every element of a unital C*-algebra $A$ is a linear combination of four unitaries. In fact, if $x = x^* \in A$ and $\|x\| \leq 2$, then $x$ is a sum of two unitaries in $U(A)$ (where $U(A)$ is the connected component of 1 in $U(A)$).

**Proof:** Set $a = x/2$. Then $1 - a^2 \geq 0$ and $a \pm i(1 - a^2)^{1/2}$ are unitaries. If $0 \leq t \leq 1$, then $(ta \pm i(1 - t^2a^2)^{1/2})$ gives a path of unitaries from $\pm i$ to $a \pm i(1 - a^2)^{1/2}$, so $a \pm i(1 - a^2)^{1/2} \in U(A)$.

\( \square \)

Invertible elements have a polar decomposition analogous to ():
\textbf{XV.14.3.7. Proposition.} Let \( A \) be a unital C*-algebra, and \( x \in A \) an invertible element. Then there are a unique unitary \( u \in A \) and \( a \in A_+ \) with \( x = ua \). In fact, \( a = (x^*x)^{1/2} \) and \( u = x(x^*x)^{-1/2} \).

\textbf{Proof:} If \( u = x(x^*x)^{-1/2} \) (note that \( x^*x \) is invertible), then \( u^*u = 1 \), and in particular \( u \) is left invertible. We have \( uu^* = x(x^*x)^{-1}x^* \) is invertible, hence \( u \) is also right invertible, i.e. \( u \) is invertible and \( u^{-1} = u^* \), so \( u \) is unitary. Obviously \( u \) and \( a = (x^*x)^{1/2} \geq 0 \) give such a decomposition. For uniqueness, if \( x = ua = vb \), then \( x^*x = a^2 = b^2 \), so \( a = b \) (XV.14.3.9), and then \( u = v = xa^{-1} \).

Noninvertible elements of a C*-algebra do not have an analogous polar decomposition in general; but if \( x \) is any element of a C*-algebra \( A \) and \( 0 < \alpha < 1/2 \), then there is a \( u \in A \) (not unique in general) such that \( x = u(x^*x)\alpha \), so \( x \) has an approximate polar decomposition.

\textbf{XV.14.3.8.} Because of XV.14.3.3(i), it makes sense in a C*-algebra \( A \) to write \( x \leq y \) if \( y - x \geq 0 \). This defines a translation-invariant partial order on \( A \) (it is usually only used on \( A_{sa} \)). Note that \( x \leq y \) is well defined independent of the containing C*-algebra. If \( x = x^* \) and \( \sigma_A(x) \subseteq [\alpha, \beta] \), then \( \alpha 1 \leq x \leq \beta 1 \) (in \( A^1 \)). If \( a \leq b \), then \( x^*ax \leq x^*bx \) for any \( x \). If \( x \in A \), \( a \in A_+ \), then \( x^*ax \leq x^*(\|a\|1)x = \|a\|x^*x \). In \( C_0(X) \), if \( f \) and \( g \) are self-adjoint (i.e. real-valued), then \( f \leq g \) in \( C_0(X) \) if and only if \( f(x) \leq g(x) \) for all \( x \in X \).

\textbf{XV.14.3.9. Proposition.} Let \( A \) be a unital C*-algebra, \( x, y \in A \). If \( 0 \leq x \leq y \) and \( x \) is invertible, then \( y \) is invertible and \( 0 \leq y^{-1} \leq x^{-1} \).

\textbf{Proof:} If \( x \) is invertible, then \( \epsilon 1 \leq x \) for some \( \epsilon > 0 \), so \( \epsilon 1 \leq y \) and \( y \) is invertible by XV.14.2.1., and \( x^{-1}, y^{-1} \geq 0 \). The inequality \( y^{-1} \leq x^{-1} \) is obvious from XV.14.2.1. if \( x \) and \( y \) commute, and in particular if \( y = 1 \). For the general case, if \( x \leq y \), then
\[
y^{-1/2}xy^{-1/2} \leq y^{-1/2}yy^{-1/2} = 1 \]
\[
1 \leq (y^{-1/2}xy^{-1/2})^{-1} = y^{1/2}x^{-1}y^{1/2} \]
\[
y^{-1} = y^{-1/2}1y^{-1/2} \leq y^{-1/2}(y^{1/2}x^{-1}y^{1/2})y^{-1/2} = x^{-1}.
\]

\textbf{Approximate Units}

While a C*-algebra need not be unital, it always contains an approximate unit (sometimes called an approximate identity) which can often be used as a substitute. Some, but not all, other Banach algebras have approximate units in an analogous sense.

\textbf{XV.14.3.10. Definition.} Let \( A \) be a C*-algebra. An \textit{approximate unit} for \( A \) is a net \((h_\lambda)\) of positive elements of \( A \) of norm \( \leq 1 \), indexed by a directed set \( \Lambda \), such that \( h_\lambda x \to x \) for all \( x \in A \). If \( h_\lambda \leq h_\mu \) for \( \lambda \leq \mu \), the approximate unit \((h_\lambda)\) is \textit{increasing}.

An approximate unit \((h_\lambda)\) is \textit{idempotent} if each \( h_\lambda \) is a projection; it is \textit{almost idempotent} if \( h_\lambda \ll h_\mu \) for \( \lambda < \mu \).

An approximate unit \((h_\lambda)\) is \textit{sequential} if \( \Lambda = \mathbb{N} \); it is \textit{continuous} if \( \Lambda = (0, \infty) \) (or a cofinal subinterval) and \( \lambda \mapsto h_\lambda \) is continuous.
In many references, approximate units are assumed to be increasing, but we will not do so since there are situations where this assumption is not natural. At the other extreme, the assumption that \( \|h\| \leq 1 \) for all \( \lambda \), or even that \( \|h\| \) is bounded, is not universal.

An approximate unit in a unital C*-algebra is just a net of (eventually) invertible positive elements in the unit ball converging to 1.

If \( A \) is a C*-algebra and \( I \) is a dense (two-sided) ideal in \( A \) (the case \( I = A \) is the most important one!), let \( A_I \) be the set of all positive elements of \( I \) of norm strictly less than 1. Give \( A_I \) its ordering as a subset of \( A^+ \).

XV.14.3.11. **Proposition.**

(i) \( A_I \) is a directed set.

(ii) \( A_I \) is an increasing approximate unit for \( A \).

**Proof:** (i): By XV.14.3.9,(iii), \( h \mapsto (1 - h)^{-1} - 1 \) is an order-isomorphism of \( A_I \) onto \( I_+ \), which is a directed set. (The computation is done in \( A \), but it is easy to check that the map and its inverse \( h \mapsto 1 - (1 + h)^{-1} \) map \( I \) into \( I \) by considering the quotient map \( A \rightarrow A/I \).)

(ii): It suffices to show that for any \( x \in A \), the decreasing net

\[
\{x^*(1 - h)x : h \in A_I \}
\]

converges to 0, for then

\[
\|(1 - h)x\|^2 = \|x^*(1 - h)^2x\| \leq \|x^*(1 - h)x\| \rightarrow 0.
\]

Since \( A_I \) is dense in \( A \) [if \( x \in A_+ \), choose \( y \in I \) close to \( x^{1/2} \); then \( y^*y \in I_+ \) is close to \( x \)], we need only find for each \( x \) and \( \epsilon \) an \( h \in A \) with \( \|x^*(1 - h)x\| \leq \epsilon \). By XV.14.3.2,(vi), we may assume \( x \geq 0 \) and \( \|x\| < 1 \). But then \( \|x(1 - x^{1/n})x\| < \epsilon \) for sufficiently large \( n \).

XV.14.3.12. **Corollary.** If \( A \) is a C*-algebra and \( I \) is a dense (two-sided) ideal in \( A \), then there is an increasing approximate unit for \( A \) contained in \( I \). If \( A \) is separable, then the increasing approximate unit may be chosen to be sequential or continuous.

To construct an increasing sequential approximate unit for a separable \( A \), let \( \{x_n\} \) be a dense set in \( A \), and inductively find \( h_n \) in \( A_I \) with \( h_n \geq h_{n-1} \) and \( \|h_n x_k - x_k\| < 1/n \) for \( 1 \leq k \leq n \). For a continuous approximate unit, set \( h_t = (n + 1 - t)h_n + (t - n)h_{n+1} \) for \( n \leq t \leq n + 1 \).

**Closed Ideals and Quotients**

In this section, “ideal” means “two-sided ideal” unless otherwise specified.

XV.14.3.13. **Proposition.** Let \( A \) be a C*-algebra, and \( J \) a closed ring-theoretic ideal in \( A \). Then \( J \) is closed under scalar multiplication.

**Proof:** This is obvious if \( A \) is unital. For general \( A \), let \( \{h_\lambda\} \) be an approximate unit for \( A \). If \( x \in J \) and \( \alpha \in \mathbb{C} \), then \( (\alpha h_\lambda)x \in J \) for all \( \lambda \) and \( (\alpha h_\lambda)x = \alpha(h_\lambda x) \rightarrow \alpha x \). Since \( J \) is closed, \( \alpha x \in J \).
XV.14.3.14. **Proposition.** Let $J$ be a closed ideal in a C*-algebra $A$, and $\pi$ the quotient map from $A$ to the Banach algebra $A/J$. Then

(i) $J$ is self-adjoint (so $J$ is a C*-algebra and there is an induced involution on $A/J$).

(ii) If $(h_\lambda)$ is an approximate unit for $J$, then for all $x \in A$,

$$\|\pi(x)\| = \lim_\lambda \|x(1 - h_\lambda)\| = \inf \|x(1 - h_\lambda)\|.$$

(iii) $A/J$ is a C*-algebra in the quotient norm.

**Proof:** (i): If $x \in J$, then $(x^* x)^{1/n} \in J$ for all $n$.

$$\|x^* - (x^* x)^{1/n} x^*\|^2 = \|x - x^* x\|^2$$

$$= \|(1 - (x^* x)^{1/n}) x^* x (1 - (x^* x)^{1/n})\| \to 0.$$  

(ii): $\|\pi(x)\| = \inf \{\|x + y\| : y \in I\} \leq \inf \|x - x h_\lambda\|$. But if $y \in J$, then

$$\|x - x h_\lambda\| \leq \|(x + y)(1 - h_\lambda)\| + \|y - y h_\lambda\| \leq \|x + y\| + \|y - y h_\lambda\|,$$

so $\limsup \|x - x h_\lambda\| \leq \|\pi(x)\|$.

(iii): For any $x \in A$,

$$\|\pi(x)\|^2 = \inf \|x(1 - h_\lambda)\|^2 = \inf \|(1 - h_\lambda)x^* x (1 - h_\lambda)\|$$

$$\leq \inf \|x^* x (1 - h_\lambda)\| = \|\pi(x^* x)\| = \|\pi(x)\|^2.$$

XV.14.3.15. **Corollary.** Let $A$ and $B$ be C*-algebras, $\phi : A \to B$ a *-homomorphism. Then $\phi(A)$ is closed in $B$, and therefore a C*-subalgebra which is isometrically isomorphic to $A/J$, where $J = \text{Ker}(\phi)$.

**Proof:** $J$ is closed since $\phi$ is continuous (XV.14.1.11.), and $\phi$ induces an injective *-homomorphism from the C*-algebra $A/J$ to $B$, which is an isometry by XV.14.2.32.

XV.14.3.16. **Proposition.** Let $A$ be a C*-algebra.

(i) If $I$ and $J$ are closed ideals in $A$, then $I \cap J = IJ$, where $IJ$ is the set of finite sums of elements of the form $xy$ with $x \in I$, $y \in J$.

(ii) If $I$ is a closed ideal in $A$ and $J$ is a closed ideal of $I$, then $J$ is an ideal of $A$.

(iii) If $I$, $J$, $K$ are closed ideals of $A$, then $(I + K) \cap (J + K) = (I \cap J) + K$. 

1877
Proof: (i) Clearly $IJ \subseteq I \cap J$, and any product of two elements of $I \cap J$ is in $IJ$. But it is an easy consequence of functional calculus that any self-adjoint element of a C*-algebra is a product of two elements (in fact, any element of a C*-algebra is the product of two elements, so it is even unnecessary to allow finite sums in the definition of $IJ$).

(ii): $I$ is itself a C*-algebra, hence has an approximate unit $(h_\lambda)$. If $x \in A$ and $y \in J$, then $y \in I$, so $h_\lambda y \to y$, and thus $x h_\lambda y \to xy$. We have $x h_\lambda \in I$ and thus $(x h_\lambda) y \in J$ for all $\lambda$ since $J$ is an ideal of $I$, and so $xy \in J$ since $J$ is closed. A similar argument shows that $yx \in J$.

(iii) follows immediately from (i) and the fact that $(I + K)(J + K) \subseteq IJ + K \subseteq (I \cap J) + K \subseteq (I + K) \cap (J + K)$ (true in any ring).

 XV.14.3.17. A C*-algebra can have many nonclosed ideals, and their structure can be enormously complicated. (See, for example, [GJ76] for a description of the ideal structure of $C(X)$.) Several of the properties of closed ideals can fail for nonclosed ideals: they need not be self-adjoint, a self-adjoint ideal need not be hereditary or even positively generated, and an ideal of an ideal of $A$ need not be an ideal of $A$. See [Bla06] for examples.

 XV.14.4. Representations of Banach *-Algebras and C*-Algebras

Before about 1980, representation theory was the core of the theory of C*-algebras. It has since become less central, but it is still an extremely important part of the theory. Many applications of the theory involve representations. And there are even basic structural facts about C*-algebras which can only be reasonably obtained via representation theory.

 XV.14.4.1. Definition. A representation of a Banach *-algebra $A$ is a *-homomorphism from $A$ to $B(H)$ for some Hilbert space $H$. (Technically this should be called a *-representation; we will not consider representations which are not *-representations.) We will mostly only consider representations of C*-algebras.

Two representations $\pi$ and $\rho$ of $A$ on $\mathcal{X}$ and $\mathcal{Y}$ respectively are (unitarily) equivalent if there is a unitary operator $U \in B(\mathcal{X}, \mathcal{Y})$ with $U \pi(x) U^* = \rho(x)$ for all $x \in A$.

A subrepresentation of a representation $\pi$ on $\mathcal{H}$ is the restriction of $\pi$ to a closed invariant subspace of $\mathcal{H}$ (a subspace $\mathcal{Y}$ of $\mathcal{H}$ is invariant if $\pi(x) \xi \in \mathcal{Y}$ for all $x \in A$, $\xi \in \mathcal{Y}$).

A representation is irreducible if it has no nontrivial closed invariant subspaces.

If $\pi_i$ ($i \in \Lambda$) is a representation of $A$ on $\mathcal{H}_i$, then the sum $\oplus_i \pi_i$ of the $\pi_i$ is the diagonal sum acting on $\oplus_i \mathcal{H}_i$.

If each $\pi_i$ is equivalent to a fixed representation $\rho$, then $\oplus_i \pi_i$ is a multiple or amplification of $\rho$ by $\text{card}(\Lambda)$.

 XV.14.4.2. A representation is always norm-decreasing (XV.14.1.11.) and hence continuous. Thus the kernel of a representation is a closed ideal. A representation with kernel 0 is called faithful. By XV.14.2.32., a faithful representation of a C*-algebra is isometric.

 XV.14.4.3. If $J$ is a closed ideal of $A$, then any representation of $A/J$ gives a representation of $A$ by composition with the quotient map, whose kernel contains $J$. Conversely, if $\pi$ is a representation of $A$ whose kernel contains $J$, then $\pi$ drops to a representation of $A/J$.

1878
**XV.14.4.4.** PROPOSITION. If $\pi$ is a representation of $A$ on $\mathcal{H}$ and $\mathcal{X}$ is a closed subspace of $\mathcal{H}$, then the following are equivalent:

(i) $\mathcal{X}$ is invariant under $\pi(A)$.

(ii) $\mathcal{X}^\perp$ is invariant under $\pi(A)$.

(iii) $P_\mathcal{X} \in \pi(A)'$.

**XV.14.4.5.** If $\mathcal{N}$ is the largest subspace of $\mathcal{H}$ on which $\pi(x) = 0$ for all $x \in A$, then $\mathcal{N}$ is closed, and $\mathcal{X} = \mathcal{N}^\perp$ is an invariant subspace, called the essential subspace of $\pi$. $\mathcal{X}$ is the closed span of $\{\pi(x)\xi : x \in A, \xi \in \mathcal{H}\}$. If $A$ is unital, then $P_\mathcal{X} = \pi(1_A)$. The representation $\pi$ is nondegenerate (or essential) if $\mathcal{X} = \mathcal{H}$. An irreducible representation is always nondegenerate.

**XV.14.4.6.** PROPOSITION. Let $\pi$ be a representation of a C*-algebra $A$ on $\mathcal{H}$, and $\mathcal{X}$ the essential subspace of $\pi$. If $(h_\lambda)$ is an approximate unit for $A$, then $\pi(h_\lambda) \to P_\mathcal{X}$ strongly.

**Proof:** Fix $x \in A$ and $\zeta \in \mathcal{X}$. Since $h_\lambda x \to x$ in norm, $\pi(h_\lambda x) \to \pi(x)$ in norm, so $\pi(h_\lambda x)\zeta \to \pi(x)\zeta$. We have

$$\pi(h_\lambda)[\pi(x)\zeta] = \pi(h_\lambda x)\zeta \to \pi(x)\zeta.$$ 

Since vectors of the form $\pi(x)\zeta$ ($x \in A$, $\zeta \in \mathcal{X}$) are dense in $\mathcal{X}$, and $(h_\lambda)$ is uniformly bounded, for any $\zeta \in \mathcal{X}$ we have $\pi(h_\lambda)\zeta \to \zeta$. Since $\pi(h_\lambda)\eta = 0$ for $\eta \in \mathcal{X}^\perp$, for any $\xi \in \mathcal{H}$ write $\xi = \zeta + \eta$ with $\zeta \in \mathcal{X}$, $\eta \in \mathcal{X}^\perp$; then

$$\pi(h_\lambda)\xi = \pi(h_\lambda)\zeta + \pi(h_\lambda)\eta = \pi(h_\lambda)\zeta \to \zeta = P_\mathcal{X}\xi.$$ 

**XV.14.4.7.** PROPOSITION. If $J$ is a closed ideal in $A$, let $\mathcal{E}$ be the essential subspace of $\pi|_J$. Then $\mathcal{E}$ is invariant under $\pi(A)$, so $P_\mathcal{E} \in \pi(A)'$.

**Proof:** Fix an approximate unit $(h_\lambda)$ for $J$. If $\xi \in \mathcal{E}$, then $\pi(h_\lambda)\xi \to \xi$ (XV.14.4.6.). If $x \in A$, then $\pi(xh_\lambda)\xi \in \mathcal{E}$ since $xh_\lambda \in J$, and $\pi(xh_\lambda)\xi = \pi(x)[\pi(h_\lambda)\xi] \to \pi(x)\xi$ since $\pi(x)$ is continuous; thus $\pi(x)\xi \in \mathcal{E}$ since $\mathcal{E}$ is closed.

**XV.14.4.8.** If $J$ is a closed ideal in $A$, let $\mathcal{E}$ be the essential subspace of $\pi|_J$. Then $\mathcal{E}$ is invariant under $\pi(A)$, so $\pi$ decomposes into a sum of a representation essential on $J$ and a representation which is zero on $J$ (a representation of $A/J$). Conversely, if $\rho$ is a nondegenerate representation of $J$ on an $\mathcal{H}$, and $(h_\lambda)$ is an approximate unit for $J$, then for any $a \in A$ the net $(\rho(ah_\lambda))$ converges strongly in $B(\mathcal{H})$ to an operator we call $\rho_A(a)$ (cf. ??), defining a representation of $A$ extending $\rho$, and we have $\rho_A(A)'' = \rho(J)''$. (This $\rho_A$ is the unique extension of $\rho$ to a representation of $A$ on $\mathcal{H}$.)
XV.14.4.9. Note that for a nondegenerate representation $\pi$, the strong (or weak) closure of $\pi(A)$ is $\pi(A)''$ by the Bicommutant Theorem (XV.3.5.2); for a general $\pi$, the strong closure of $\pi(A)$ is $P_\mathcal{X}\pi(A)''$, where $\mathcal{X}$ is the essential subspace of $\pi$. (Note that $P_\mathcal{X}$ is a central projection in $\pi(A)''$.)

We get the following fact from an application of the Bicommutant Theorem:

XV.14.4.10. **Proposition.** Let $\pi$ be a representation of a C*-algebra $A$ on a Hilbert space $\mathcal{H}$. Then the following are equivalent:

(i) $\pi$ is irreducible.

(ii) $\pi(A)' = \mathbb{C}I$.

(iii) $\pi(A)'' = \mathcal{B}(\mathcal{H})$.

(iv) $\pi(A)$ is strongly dense in $\mathcal{B}(\mathcal{H})$.

It follows from XV.14.4.8. that if $\pi$ is an irreducible representation of $A$, and $J$ is a closed ideal of $A$, then $\pi|_J$ is either zero or irreducible.

XV.14.4.11. **Proposition.** The kernel of an irreducible representation of a C*-algebra is a (closed) prime ideal.

**Proof:** The kernel of any representation is closed. Let $\pi$ be an irreducible representation of a C*-algebra $A$ on $\mathcal{H}$. Replacing $A$ by $A/\ker(\pi)$ we may assume $\pi$ is faithful. Suppose there are nonzero ideals $J$ and $K$ of $A$ with $J \cap K = \{0\}$. If $(h_\lambda)$ and $(k_\mu)$ are increasing approximate units for $J$ and $K$ respectively, then $(\pi(h_\lambda))$ and $(\pi(k_\mu))$ converge strongly to the projections $P$ and $Q$ onto the closures of $\pi(J)\mathcal{H}$ and $\pi(K)\mathcal{H}$. Then $P$ and $Q$ are nonzero projections in $\mathcal{B}(\mathcal{H})$, and $P \perp Q$. Since $\pi(J)\mathcal{H}$ and $\pi(K)\mathcal{H}$ are invariant under $\pi(A)$, we have $P, Q \in \pi(A)'$, a contradiction. \hfill \$\checkmark$

XV.14.4.12. If $\pi$ is a representation of a C*-algebra $A$ on a Hilbert space $\mathcal{H}$, and $0 \neq \xi \in \mathcal{H}$, then the set $\{\pi(x)\xi : x \in A\}$ is an invariant subspace of $\mathcal{H}$, and thus its closure is also invariant. If the closure is all of $\mathcal{H}$, $\xi$ is called a cyclic vector for $\pi$. A representation of $A$ with a cyclic vector is called a cyclic representation of $A$. A cyclic representation is nondegenerate. A simple Zorn’s Lemma argument shows that every nondegenerate representation is a direct sum of cyclic representations (in a highly nonunique way in general). If $\pi$ is an irreducible representation of $A$ on $\mathcal{H}$, then every nonzero vector in $\mathcal{H}$ is a cyclic vector for $\pi$.

**Positive Linear Functionals and States**

The connection between C*-algebras and Hilbert spaces is made via the notion of a state:

XV.14.4.13. **Definition.** Let $A$ be a C*-algebra. A linear functional $\phi$ on $A$ is positive, written $\phi \geq 0$, if $\phi(x) \geq 0$ whenever $x \geq 0$. A state on $A$ is a positive linear functional of norm 1. Denote by $\mathcal{S}(A)$ the set of all states on $A$, called the state space of $A$. 1880
**XV.14.4.14.** PROPOSITION. A positive linear functional is bounded.

**Proof:** Let $\phi$ be a positive linear functional on a C*-algebra $A$. It suffices to show $\phi$ is bounded on $A_+$ by XV.14.3.2.(vi): if it is not, for each $n$ choose $x_n \geq 0$ with $\|x_n\| = 1$ and $\phi(x_n) \geq 4^n$, and then $x = \sum_{n=1}^{\infty} 2^{-n}x_n$ satisfies $\phi(x) \geq 2^{-n}\phi(x_n) \geq 2^n$ for all $n$, a contradiction. $\blacksquare$

**XV.14.4.15.** EXAMPLES.

(i) If $A$ is a concrete C*-algebra of operators acting nondegenerately on $\mathcal{H}$ and $\xi \in \mathcal{H}$, and $\phi_\xi(x) = \langle x\xi, \xi \rangle$ for $x \in A$, then $\phi_\xi$ is a positive linear functional on $A$ of norm $\|\xi\|^2$, so $\phi_\xi$ is a state if $\|\xi\| = 1$. Such a state is called a vector state of $A$. This example is the origin of the term “state”: in the mathematical formulation of quantum mechanics, the states of a physical system are given by probability distributions (unit vectors in an $L^2$-space), and observables are self-adjoint operators; the value of the observable $T$ on the state $\xi$ is $\langle T\xi, \xi \rangle$ (cf. [BR87]).

(ii) By the Riesz Representation Theorem, there is a one-one correspondence between bounded linear functionals on $C_0(X)$ and finite regular complex Baire measures (complex Radon measures) on $X$. A bounded linear functional is positive if and only if the corresponding complex measure takes only nonnegative real values (i.e. is an ordinary measure). Thus the states on $C_0(X)$ are precisely given by the regular Baire probability measures on $X$. Any nonzero homomorphism from $C_0(X)$ to $\mathbb{C}$ is a state by XV.14.2.14., XV.14.2.15., and XV.14.3.2.(ii) and (vii).

A positive linear functional takes real values on self-adjoint elements, and thus $\phi(x^*) = \overline{\phi(x)}$ for any $x$ if $\phi \geq 0$. So if $\phi$ is a positive linear functional on $A$, then $\phi$ defines a pre-inner product on $A$ by $\langle x, y \rangle_\phi = \phi(y^*x)$. Thus we get the CBS inequality:

**XV.14.4.16.** PROPOSITION. Let $A$ be a C*-algebra, $\phi$ a positive linear functional on $A$, $x, y \in A$. Then $|\phi(y^*x)|^2 \leq \phi(x^*x)\phi(y^*y)$. More symmetrically, $|\phi(xy)|^2 \leq \phi(xx^*)\phi(y^*y)$.

**XV.14.4.17.** PROPOSITION. Let $A$ be a C*-algebra and $\phi$ a bounded linear functional on $A$. Then

(i) If $\phi \geq 0$, then $\|\phi\| = \sup\{\phi(x) : x \geq 0, \|x\| = 1\} = \lim \phi(h_\lambda)$ for any approximate unit $(h_\lambda)$ for $A$ (in particular, $\|\phi\| = \phi(1_A)$ if $A$ is unital), and $\phi$ extends to a positive linear functional on $A^1$ by setting $\phi(1) = \|\phi\|$ (or $\phi(1) = t$ for any $t \geq \|\phi\|$). In particular, any state on $A$ extends uniquely to a state on $A^1$.

(ii) If $A$ is unital and $\phi(1_A) = \|\phi\|$, then $\phi \geq 0$.

**Proof:** For the first part of (i), if $(h_\lambda)$ is an approximate unit for $A$, we have

$$\|\phi\| \geq \sup\{\phi(x) : x \geq 0, \|x\| = 1\} \geq \sup \phi(h_\lambda);$$
but if \( y \in A \), \( \|y\| = 1 \), \( |\phi(y)|^2 > \|\phi\|^2 - \epsilon \), then, using XV.14.4.16.,

\[
\|\phi\|^2 - \epsilon < |\phi(y)|^2 = \text{lim} |\phi(h_{1/2}^x y)|^2 \leq \text{lim inf} \phi(h_{1/2}^x) \phi(y^* y) \leq \|\phi\| \text{lim inf} \phi(h_{1/2}^x).
\]

The second statement of (i) follows from the first and XV.14.3.2.(viii).

To prove (ii), if \( x \geq 0 \), restrict \( \phi \) to \( C^*(x,1) \cong C(X) \); then \( \phi \) corresponds to a complex measure \( \mu \) on \( X \), and \( \mu(X) = |\mu|(X) \), so \( \mu \geq 0 \). A more elementary alternate argument can be based on XV.14.1.10.: if \( \phi(1_A) = \|\phi\| = 1 \), and \( x \geq 0 \), suppose \( \phi(\sigma) \notin [0,\infty) \). Then there is a \( \lambda \in \mathbb{C} \) and \( \rho \geq 0 \) with \( \phi(\sigma) \subseteq D = \{ z \in \mathbb{C} : |z - \lambda| \leq \rho \} \), but \( \phi(\sigma) \notin D \). If \( y = x - \lambda 1_A \), then \( \phi \) is normal and \( \|y\| = r_{\sigma}(y) \leq \rho \), so \( |\phi(x) - \lambda| = |\phi(y)| \leq \|y\| \leq \rho \), a contradiction.

XV.14.4.18. Any sum or nonnegative scalar multiple of positive linear functionals on a C*-algebra \( A \) is positive. A weak-* limit of a net of positive linear functionals on \( A \) is positive. XV.14.4.17.(i) shows that if \( \phi \) and \( \psi \) are positive linear functionals on \( A \), then \( \|\phi + \psi\| = \|\phi\| + \|\psi\| \). In particular, a convex combination of states is a state, so if \( A \) is unital, \( S(A) \) is a compact convex subset of the dual of \( A \). If \( A \) is nonunital, then \( S(A) \), though convex, is not compact or even locally compact, but the set of positive linear functionals on \( A \) of norm \( \leq 1 \) is in natural one-one correspondence with \( S(A^1) \), which is a compact convex set.

XV.14.4.19. Definition. A state of a C*-algebra \( A \) is pure if it is an extreme point of \( S(A) \). Denote the set of pure states of \( A \) by \( P(A) \).

XV.14.4.20. If \( A \) is unital, then \( S(A) \) is the closed convex hull of the pure states; in general, the set of positive linear functionals of norm \( \leq 1 \) is the closed convex hull of the pure states and the zero functional. The pure states of \( C_0(X) \) are precisely the homomorphisms to \( \mathbb{C} \), i.e. the states corresponding to point masses (XV.14.4.15.(ii)).

Extension and Existence of States

XV.14.4.21. Proposition. If \( A \) is a C*-algebra and \( B \) a C*-subalgebra, then every state on \( B \) extends to a state on \( A \).

Proof: If \( A \) is a C*-algebra and \( \phi \) is a state on \( A \), then by XV.14.4.17.(i) \( \phi \) extends to a state on \( A^1 \). If \( A \) is a unital C*-algebra and \( B \) a unital C*-subalgebra, and \( \phi \) is a state on \( B \), by the Hahn-Banach Theorem () \( \phi \) extends to a linear functional \( \psi \) on \( A \) of norm \( 1 \). Since \( \psi(1) = \phi(1) = 1 \), \( \psi \) is a state on \( A \). If \( B \) is a (not necessarily unital) C*-subalgebra of a (not necessarily unital) C*-algebra \( A \), and \( \phi \) is a state on \( B \), first extend \( \phi \) to \( B^1 \), regard \( B^1 \) as a unital C*-subalgebra of \( A^1 \) and extend to \( A^1 \), and then restrict to \( A \) to get an extension of \( \phi \) to a state on \( A \).

XV.14.4.22. If \( \phi \) is a pure state on a C*-subalgebra of a (not necessarily unital) C*-algebra \( A \), then the set of extensions of \( \phi \) to \( A \) is a weak-* compact convex subset of \( S(A) \), and any extreme point of this set (extreme points exist by the Krein-Milman Theorem ()) is a pure state of \( A \). Thus we get:
Proposition. Let $A$ be a C*-algebra, $x \in A_{sa}$. Then there is a pure state $\phi$ on $A$ with $|\phi(x)| = \|x\|$.

Proof: Note that there is a pure state (homomorphism) $\phi$ on $C^*(x) \cong C_0(\sigma_A(x))$ with $|\phi(x)| = r_\sigma(x) = \|x\|$, and extend $\phi$ to a pure state on $A$.

There is also a noncommutative version of the Jordan Decomposition Theorem for signed measures:

Theorem. Let $A$ be a C*-algebra, and $\phi$ a bounded self-adjoint linear functional on $A$. Then there are unique positive linear functionals $\phi_+$ and $\phi_-$ on $A$ with $\phi = \phi_+ - \phi_-$ and $\|\phi\| = \|\phi_+\| + \|\phi_-\|$.

Since every linear functional $\phi$ can be canonically written as $\phi_{re} + i\phi_{im}$ with $\phi_{re}, \phi_{im}$ self-adjoint [define $\phi^* = \overline{\phi(x)}$, and let $\phi_{re} = (\phi + \phi^*)/2, \phi_{im} = (\phi - \phi^*)/2i$], every bounded linear functional can be written canonically as a linear combination of four states. A simple consequence of this and the Hahn-Banach Theorem is:

Corollary. Let $A$ be a C*-algebra, $x \in A$.

(i) If $\phi(x) = 0$ for all $\phi \in S(A)$, then $x = 0$.

(ii) $x = x^*$ if and only if $\phi(x) \in \mathbb{R}$ for all $\phi \in S(A)$.

(iii) $x \geq 0$ if and only if $\phi(x) \geq 0$ for all $\phi \in S(A)$.

In fact, $S(A)$ can be replaced by $\mathcal{P}(A)$ in this result.

The GNS Construction

The GNS construction, discovered independently by Gelfand and Naimark [GN43] and I. Segal [Seg47] (the essential idea appeared earlier in [Ge41]), while quite simple, is an “ingenious construction” (E. Hille [Hil43]) and one of the most fundamental ideas of the theory of operator algebras. It provides a method for manufacturing representations of C*-algebras.

Let $A$ be a C*-algebra, $\phi$ a positive linear functional on $A$. Put a pre-inner product on $A$ by $\langle x, y \rangle_\phi = \phi(y^*x)$. If

$$N_\phi = \{x \in A : \phi(x^*x) = 0\},$$

then $N_\phi$ is a closed left ideal of $A$, and $\langle \cdot, \cdot \rangle_\phi$ drops to an inner product on $A/N_\phi$. If $a \in A$, let $\pi_\phi(a)$ be the left multiplication operator by $a$ on $A/N_\phi$, i.e. $\pi_\phi(a)(x + N_\phi) = ax + N_\phi$. Since

$$x^*a^*ax \leq \|a\|^2x^*x$$

(XV.14.3.8.), $\pi_\phi(a)$ is a bounded operator and $\|\pi_\phi(a)\| \leq \|a\|$; so $\pi_\phi(a)$ extends to a bounded operator, also denoted $\pi_\phi(a)$, on the completion (Hilbert space) $\mathcal{H}_\phi = L^2(A, \phi)$ of $A/N_\phi$. The representation $\pi_\phi$ is called the GNS representation of $A$ associated with $\phi$. 1883
XV.14.4.27. There is also a (unique) distinguished vector $\xi_\phi \in \mathcal{H}_\phi$ such that $\phi(a) = \langle \pi_\phi(a)\xi_\phi, \xi_\phi \rangle_\phi$ (and $\|\xi_\phi\|^2 = \|\phi\|$). If $A$ is unital, then $\xi_\phi$ is just the image of $1_A$ in $\mathcal{H}_\phi$. In general, let $\tilde{\phi}$ be the unique extension of $\phi$ to $A^\dagger$ with $\|\tilde{\phi}\| = \|\phi\|$; then from $\tilde{\phi}(1) = \lim \phi(h_\lambda) = \lim \phi(h_\lambda^2)$, where $(h_\lambda)$ is an approximate unit for $A$, it follows that $\phi((1 - h_\lambda)^2) \to 0$, so $A/N_{\phi}$ is dense in $A^\dagger/N_{\tilde{\phi}}$, and thus $\mathcal{H}_\phi$ can be identified with $\mathcal{H}_{\tilde{\phi}}$, and $\xi_\phi$ may be taken to be $\xi_{\tilde{\phi}}$. This $\xi_\phi$ is a cyclic vector for $\pi_\phi$, i.e. $\pi_\phi(A)\xi_\phi$ is dense in $\mathcal{H}_{\tilde{\phi}}$.

XV.14.4.28. This construction is the inverse of XV.14.4.15(i): if $\pi$ is a representation of $A$ on a Hilbert space $\mathcal{H}$ and $\xi$ is a cyclic vector for $\pi$, and $\phi(a) = \langle \pi(a)\xi, \xi \rangle$, then $(\mathcal{H}_\phi, \pi_\phi, \xi_\phi)$ is unitarily equivalent to $(\mathcal{H}, \pi, \xi)$. Thus there is a precise one-one correspondence between positive linear functionals on $A$ and (cyclic) representations of $A$ with a specified cyclic vector.

XV.14.4.29. Example. Let $A = n$, with Tr the (ordinary) trace on $A$. Then $\mathcal{H}_{\text{Tr}}$ can be identified with $\mathbb{M}_n$ with the Hilbert-Schmidt norm, with $\xi_{\text{Tr}} = I$. The action $\pi_{\text{Tr}}$ is the ordinary one by left multiplication. $\mathbb{M}_n$ acting on itself in this way is called the standard form representation of $\mathbb{M}_n$. Its commutant is also isomorphic (or, more naturally, anti-isomorphic) to $\mathbb{M}_n$, acting by right multiplication.

XV.14.4.30. If $\phi$ is a positive linear functional on $A$, then $\phi$ has a canonical extension to a positive linear functional on $\mathcal{B}(\mathcal{H}_\phi)$ of the same norm, also denoted $\phi$, by $\phi(T) = \langle T\xi_\phi, \xi_\phi \rangle_\phi$.

The next simple result is a version of the Radon-Nikodym Theorem (the classical Radon-Nikodym theorem is essentially the special case where $A = L^\infty(X, \mu)$ for a finite measure space $(X, \mu)$, and $\phi(f) = \int f \, d\mu$).

XV.14.4.31. Proposition. Let $\phi$ and $\psi$ be positive linear functionals on $A$, with $\psi \leq \phi$ (i.e. $\psi(x) \leq \phi(x)$ for all $x \in A_+$). Then there is a unique operator $T \in \pi_\phi(A)' \subseteq \mathcal{B}(\mathcal{H}_\phi)$, with $0 \leq T \leq I$, such that $\psi(x) = \phi(T\pi_\phi(x)) = \langle T\pi_\phi(x)\xi_\phi, \xi_\phi \rangle_\phi$ for all $x \in A$.

Proof: For $x, y \in A$, define $(\pi_\phi(x)\xi_\phi, \pi_\phi(y)\xi_\phi) = \psi(y^*x)$. Then $(\cdot, \cdot)$ extends to a bounded sesquilinear form (pre-inner product) on $\mathcal{H}_\phi$, and thus there is a $T \in (\mathcal{H}_\phi)^\prime$, $0 \leq T \leq I$, such that $(\eta, \zeta) = \langle T\eta, \zeta \rangle_\phi$ for all $\eta, \zeta \in \mathcal{H}_\phi$ (??). If $x, y, z \in A$, then

\[
\langle T\pi_\phi(x)\xi_\phi, \pi_\phi(y)\xi_\phi \rangle_\phi = \psi(y^*(xz)) = \psi((x^*y)^*z)
\]

so, fixing $x$ and letting $y, z$ range over $A$, we conclude that $T\pi_\phi(x) = \pi_\phi(x)T$.

XV.14.4.32. In fact, if $\psi \leq \phi$ and $\omega = \phi - \psi$, so that $\phi = \psi + \omega$, then $\pi_\phi$ is unitarily equivalent to the subrepresentation of $\pi_\psi \oplus \pi_\omega$ with cyclic vector $\xi_\psi \oplus \xi_\omega$, and $T$ is projection onto the first coordinate, compressed to this subspace.

Because of the abundance of states on a C*-algebra described in (XV.14.4.23.), every C*-algebra has many representations. We can say even more:
XV.14.4.33. Proposition. If \( \phi \) is a state of \( A \), then \( \pi_\phi \) is irreducible if and only if \( \phi \) is a pure state.

Proof: It follows immediately from XV.14.4.31. and XV.14.4.10. that if \( \pi_\phi \) is irreducible, then any positive linear functional \( \psi \leq \phi \) is a multiple of \( \phi \), and hence \( \phi \) is pure. Conversely, if \( \phi \) is pure, suppose there is a projection \( P \) in \( \pi_\phi(A)' \), \( P \neq 0 \), \( I \). Then \( P\xi_\phi \neq 0 \), for otherwise

\[
0 = \pi_\phi(x) P\xi_\phi = P[\pi_\phi(x)\xi_\phi]
\]

for all \( x \in A \), so \( P = 0 \). Similarly, \( (I - P)\xi_\phi \neq 0 \). For \( x \in A \), write

\[
\phi_1(x) = \langle \pi_\phi(x) P\xi_\phi, P\xi_\phi \rangle = \phi(P\pi_\phi(x))
\]

\[
\phi_2(x) = \langle \pi_\phi(x)(I - P)\xi_\phi, (I - P)\xi_\phi \rangle = \phi((I - P)\pi_\phi(x)).
\]

Then \( \phi_1, \phi_2 \) are positive linear functionals on \( A \), \( \phi = \phi_1 + \phi_2 \), and \( \phi_1 = \lambda \phi \), \( \phi_2 = (1 - \lambda)\phi \) for some \( \lambda, 0 < \lambda < 1 \), because \( \phi \) is pure. For any \( \epsilon > 0 \), there is an \( x \in A \) with \( \|\pi_\phi(x)\xi_\phi - P\xi_\phi\|_\phi < \epsilon \) and \( \|\pi_\phi(x)\xi_\phi\|^2_\phi = \phi(x^*x) = \|P\xi_\phi\|^2_\phi \). Then \( \|(I - P)\pi_\phi(x)\xi_\phi\| = \|(I - P)(\pi_\phi(x)\xi_\phi - P\xi_\phi)\| < \epsilon \), and

\[
(1 - \lambda)\|P\xi_\phi\|^2_\phi = (1 - \lambda)\phi(x^*x) = \phi_2(x^*x) = \|(I - P)\pi_\phi(x)\xi_\phi\|^2_\phi < \epsilon^2
\]

contradicting \( \lambda \neq 1 \) and therefore the existence of \( P \). Thus \( \pi_\phi(A)' = CI \).

XV.14.4.34. Corollary. If \( A \) is a C*-algebra and \( x \in A \), then there is an irreducible representation \( \pi \) of \( A \) with \( \|\pi(x)\| = \|x\| \).

Proof: Let \( \pi = \pi_\phi \), where \( \phi \) is a pure state of \( A \) with \( \phi((xx^*)^2) = \|(xx^*)^2\| = \|x\|^4 \) (XV.14.4.23.). Then

\[
\|x\|^2 = \langle xx^*, xx^* \rangle^{1/2}_\phi = \|xx^*\|_\phi = \|\pi(x)x^*\|_\phi \leq \|\pi(x)\|\|x^*\|_\phi \leq \|\pi(x)\|\|x\|.
\]

By considering sums of irreducible representations, we obtain

XV.14.4.35. Corollary. [Gelfand-Naimark Theorem] If \( A \) is a C*-algebra, then \( A \) has a faithful representation, i.e. \( A \) is isometrically isomorphic to a concrete C*-algebra of operators on a Hilbert space \( H \). If \( A \) is separable, then \( H \) may be chosen to be separable.

XV.14.4.36. Proposition. Let \( A \) be a C*-algebra, \( B \) a C*-subalgebra of \( A \), and \( \rho \) a representation of \( B \) on a Hilbert space \( H_0 \). Then \( \rho \) can be extended to a representation \( \pi \) of \( A \) on a (possibly) larger Hilbert space \( H \), i.e. there is a Hilbert space \( H \) containing \( H_0 \) and a representation \( \pi \) of \( A \) on \( H \) such that \( \pi_\rho \) is invariant under \( \pi|_B \) and \( \pi|_B \) agrees with \( \rho \) on \( H_0 \). If \( \rho \) is irreducible, then \( \pi \) may be chosen irreducible.

Proof: Since every representation is a direct sum of cyclic representations, it suffices to show the result for \( \rho \) cyclic, i.e. \( \rho \) is the GNS representation corresponding to a state \( \psi \) of \( B \). Extend \( \psi \) to a state \( \phi \) of \( A \) (with \( \phi \) pure if \( \psi \) is pure; see XV.14.4.22.), and let \( \pi \) be its GNS representation.

1885
**XV.14.4.37.** Thus every C*-algebra can be faithfully represented on a Hilbert space, but the representation is highly nonunique in general. The major part of representation theory is concerned with how to describe the possibilities. There is a relatively small class of C*-algebras (V) for which a concrete description of the representations is even possible.

Using the GNS construction, we can obtain an extension of XV.14.4.17. (i) and a sharpening of XV.14.4.21. in the case that the subalgebra is an ideal. This result is also closely related to XV.14.4.8.

**XV.14.4.38.** **PROPOSITION.** Let $J$ be a closed ideal in a C*-algebra $A$, and $\psi$ a positive linear functional on $J$. Then there is a unique extension of $\psi$ to a positive linear functional $\phi$ on $A$ with $\|\phi\| = \|\psi\|$. If $(h_\lambda)$ is an increasing approximate unit for $J$, then $\phi(x) = \lim \psi(h_\lambda x) = \lim \psi(x h_\lambda)$ for all $x \in A$; if $x \in A_+$, then

$$\phi(x) = \lim \psi(x^{1/2}h_\lambda x^{1/2}) = \sup \psi(x^{1/2}h_\lambda x^{1/2}).$$

Any other positive linear functional $\phi'$ on $A$ extending $\psi$ satisfies $\phi' \geq \phi$.

**PROOF:** By XV.14.4.21. there is a $\phi$ extending $\psi$ with $\|\phi\| = \|\psi\|$, but we will not use this result. Instead we define $\phi$ directly as follows.

Let $\pi_\psi$ be the GNS representation of $J$ on $H_\psi$ with cyclic vector $\xi_\psi$. Extend $\pi_\psi$ to a representation $\pi$ of $A$ on $H_\psi$ as in XV.14.4.8. Then $\xi_\psi$ defines a positive linear functional $\phi$ on $A$ by $\phi(x) = \langle \pi(x)\xi_\psi, \xi_\psi \rangle$. We have $\|\phi\| = \|\xi_\psi\|^2 = \|\psi\|$, and $(H_\psi, \pi_\phi, \xi_\phi)$ can be naturally identified with $(H_\psi, \pi, \xi_\psi)$. Since $\pi(h_\lambda) \to I$ strongly, we have, for any $x \in A$,

$$\psi(h_\lambda x) = \langle \pi(h_\lambda)\pi(x)\xi_\psi, \xi_\psi \rangle \to \langle \pi(x)\xi_\psi, \xi_\psi \rangle = \phi(x)$$

and similarly $\psi(x h_\lambda) \to \phi(x)$ and, if $x \geq 0$, $\psi(x^{1/2}h_\lambda x^{1/2}) \to \phi(x)$.

If $\phi'$ is any extension of $\psi$ to a positive linear functional on $A$, and $x \in A_+$, then $\phi'(x) \geq \phi'(x^{1/2}h_\lambda x^{1/2}) = \psi(x^{1/2}h_\lambda x^{1/2})$ for all $\lambda$, so $\phi'(x) \geq \phi(x)$. Thus $\omega = \phi' - \phi$ is a positive linear functional, and we have $\|\omega\| = \|\phi\| + \|\omega\| = \|\psi\|$, so if $\|\phi'\| = \|\psi\|$, then $\omega = 0$, $\phi' = \phi$.

**Primitive Ideal Space and Spectrum**

It is common in mathematics (e.g. in algebraic geometry) to associate to a ring $A$ a topological space of ideals of $A$, generally either the set of prime ideals (denoted Spec($A$)) or the set of maximal ideals (denoted Maxspec($A$)), with the hull-kernel topology (XI.7.8.16.). For C*-algebras, the most useful variant is to consider the primitive ideal space:

**XV.14.4.39.** **DEFINITION.** Let $A$ be a C*-algebra. A **primitive ideal of $A$** is an ideal which is the kernel of an irreducible representation of $A$. Denote by Prim($A$) the set of primitive ideals of $A$. The topology on Prim($A$) is the hull-kernel topology, i.e. $\{J_i\}^{-} = \{J : J \supseteq \cap J_i\}$.

**XV.14.4.40.** A primitive ideal is closed by XV.14.4.2. and prime by XV.14.4.11.. By XV.14.4.34., the intersection of the primitive ideals of a C*-algebra $A$ is $\{0\}$. By considering irreducible representations of $A/J$, it follows that every closed ideal $J$ of $A$ is an intersection of primitive ideals. In particular, every maximal ideal is primitive. There is thus a natural one-one correspondence between the closed sets in Prim($A$) and the closed ideals of $A$. If $A$ is commutative, then every closed prime ideal is maximal and hence primitive, and Prim($A$) agrees with the set Maxspec($A$) and may be identified with $A$ as in XV.14.2.1..
If $J$ is a closed ideal in $A$, then we can identify $\text{Prim}(A/J)$ with

\[
\text{Prim}_J(A) = \{ K \in \text{Prim}(A) : J \subseteq K \},
\]

the corresponding closed set in $\text{Prim}(A)$. If $K \in \text{Prim}(A)$, $J \not\subseteq K$, and $\pi$ is an irreducible representation of $A$ with kernel $K$, then $\pi|J$ is an irreducible representation of $J$ (XV.14.4.8.) with kernel $K \cap J$. So there is a map $\rho_J$ from

\[
\text{Prim}'(A) = \{ K \in \text{Prim}(A) : J \not\subseteq K \}
\]
to $\text{Prim}(J)$, defined by $\rho_J(K) = K \cap J$, which is surjective by XV.14.4.8. Also, $\rho_J$ is injective: if $K_1, K_2 \in \text{Prim}(A)$ and $K_1 \cap J = K_2 \cap J \neq J$, then by XV.14.3.16 (iii) we have

\[
(K_1 + K_2) \cap (J + K_2) = (K_1 \cap J) + K_2 = (K_2 \cap J) + K_2 = K_2
\]
and, since $K_2$ is prime and $J + K_2 \neq K_2$, we have $K_1 + K_2 = K_2, K_1 \subseteq K_2$. Symmetrically, $K_2 \subseteq K_1$, so $K_1 = K_2$. Thus $\text{Prim}(J)$ may be identified with $\text{Prim}'(A) = \text{Prim}(A) \setminus \text{Prim}_J(A)$, an open set in $\text{Prim}(A)$.

If $J$ is a closed ideal in a C*-algebra $A$ and $x \in A$, write $\|x\|_J$ for the norm of $x$ mod $J$. If \{\text{J}_i\} is any collection of closed ideals in $A$ and $J = \cap \text{J}_i$, then $\| : \|_J = \sup_i \| : \|_{\text{J}_i}$ (by XV.14.2.33. it suffices to note that this formula defines a C*-norm on $A/J$); and if \{\text{J}_i\} is an increasing net of closed ideals with $J = [\cup \text{J}_i]^{-}$, then $\| : \|_J = \inf_i \| : \|_{\text{J}_i}$ (this follows immediately from the definition of the quotient norm). A simple consequence is:

**XV.14.4.43.** PROPOSITION. Let $A$ be a C*-algebra.

(i) If $x \in A$, define $\hat{x} : \text{Prim}(A) \rightarrow \mathbb{R}_+$ by $\hat{x}(J) = \|x\|_J$. Then $\hat{x}$ is lower semicontinuous.

(ii) If $\{x_i\}$ is a dense set in the unit ball of $A$, and $U_i = \{J \in \text{Prim}(A) : \hat{x}_i(J) > 1/2\}$, then $\{U_i\}$ forms a base for the topology of $\text{Prim}(A)$.

(iii) If $x \in A$ and $\lambda > 0$, then $\{J \in \text{Prim}(A) : \hat{x}(J) \geq \lambda\}$ is compact (but not necessarily closed) in $\text{Prim}(A)$.

**PROOF:** Part (i) is obvious. For (ii), let $J \in \text{Prim}(A)$. A neighborhood of $J$ is a set of the form $V = \{ I \in \text{Prim}(A) : K \not\subseteq I \}$ for a closed ideal $K$ not contained in $J$. Choose $y \in K \setminus J$, and $0 < \epsilon \leq \frac{1}{2}\|y\|^2_2$, and let $x = f_\epsilon(y^*y)$. Then $x \in K$, and $\|x\| = \|x\|_J = 1$. Choose $x_i$ with $\|x_i - x\| < 1/2$; then $J \subseteq U_i$ and $U_i \subseteq V$. To show (iii), let $\{K_i : i \in \Omega\}$ be a set of closed ideals in $A$, $K$ the closed ideal generated by $\cup K_i$ (so $K = [\bigcup K_i]^{-}$), and

\[
U_i = \{J \in \text{Prim}(A) : K_i \not\subseteq J\}.
\]

Then $\{U_i\}$ is an open cover of $F = \{J \in \text{Prim}(A) : \hat{x}(J) \geq \lambda\}$ if and only if $\|x\|_K < \lambda$. In this case, there are $i_1, \ldots, i_n$ such that if $K_0 = K_{i_1} + \cdots + K_{i_n}$, then $\|x\|_{K_0} < \lambda$, so $\{U_{i_1}, \ldots, U_{i_n}\}$ is a cover of $F$. \(\square\)

**XV.14.4.44.** COROLLARY. $\text{Prim}(A)$ is a locally compact space $T_0$-space which is compact if $A$ is unital. If $A$ is separable, then $\text{Prim}(A)$ is second countable.

Actually, there are several inequivalent definitions of local compactness used for non-Hausdorff spaces (XI.11.8.20.), so the statement that $\text{Prim}(A)$ is “locally compact” is ambiguous in the non-Hausdorff case. To be precise, if $A$ is a C*-algebra, then XV.14.4.43. shows that every point of $\text{Prim}(A)$ has a neighborhood base of compact sets, i.e. $\text{Prim}(A)$ satisfies XI.11.8.20.(ii). But a point need not have any closed compact neighborhood, i.e. $\text{Prim}(A)$ does not necessarily satisfy XI.11.8.20.(iii).

1887
The function $\hat{x}$ of XV.14.4.43. is not continuous in general; in fact, if Prim($A$) is not Hausdorff, then the $\hat{x}$ cannot all be continuous since $\{\hat{x} : x \in A\}$ separates the points of Prim($A$). But if Prim($A$) is Hausdorff, XV.14.4.43.(iii) shows that $\hat{x}$ is continuous for all $x \in A$. If $A$ is commutative, or more generally if $x$ is in the center of $A$, then $\hat{x}$ is continuous.

XV.14.4.46. Examples.

(i) Let $A = \mathbb{K}^\dagger$. Then Prim($A$) = $\{0, \mathbb{K}\}$; $\{\mathbb{K}\}$ is a closed point, and $\{0\}$ is a dense open point. Thus Prim($A$) homeomorphic to the Sierpiński space and is not $T_1$.

(ii) Let $A$ be the C*-algebra of all sequences of elements of $\mathbb{M}_2$ which converge to a diagonal matrix. Then Prim($A$) consists of a sequence of points simultaneously converging to two closed points at infinity. Thus Prim($A$) is $T_1$ but not Hausdorff.

(iii) Let $A$ be the C*-algebra of all sequences in $\mathbb{K}^\dagger$ converging to a scalar multiple of 1. Then Prim($A$) is the one-point compactification of a disjoint union of two-point spaces as in (i).

If $A$ is commutative, then Gelfand’s Theorem (XV.14.2.1.) identifies elements of $A$ with continuous functions on $\hat{A} \cong$ Prim($A$).

XV.14.4.47. For a general C*-algebra $A$ we define $\hat{A}$ to be the set of unitary equivalence classes of irreducible representations of $A$. There is a natural map from $\hat{A}$ onto Prim($A$) sending $\pi$ to $\text{ker}(\pi)$. This map is not injective in general, i.e. there can be inequivalent representations with the same kernel. There is also an obvious map from $\mathcal{P}(A)$ onto $\hat{A}$, sending a pure state to its GNS representation; the composite map from $\mathcal{P}(A)$ to Prim($A$) is continuous and open when $\mathcal{P}(A)$ is given the relative weak-* topology. The quotient topology (XI.8.1.1.) on $\hat{A}$ from $\mathcal{P}(A)$, which coincides with the weak topology (XI.5.1.8.) for the map of $\hat{A}$ onto Prim($A$), is the unique topology on $\hat{A}$ making the maps $\mathcal{P}(A) \to \hat{A}$ and $\hat{A} \to$ Prim($A$) continuous and open (XI.8.5.1.), but $\hat{A}$ is not $T_0$ unless the map to Prim($A$) is injective (and hence a homeomorphism).

By XV.14.4.8., if $J$ is a closed ideal in $A$, we may identify $J$ and Prim($J$) with open subsets of $\hat{A}$ and Prim($A$) respectively; the complements are naturally identified with $\hat{A}/J$ and Prim($A/J$). We may do the same with $\mathcal{P}(J)$ since if $\phi \in \mathcal{P}(A)$, then either $|\phi|_J = 0$ or $|\phi|_J$ is a pure state of $J$.

XV.14.4.48. Another way of describing the topology on $\hat{A}$ is to fix a Hilbert space $\mathcal{H}$ and let Irr$_\mathcal{H}(A)$ be the set of irreducible representations of $A$ on $\mathcal{H}$. Irr$_{\mathcal{H}}(A)$ has a natural topology, the topology of elementwise convergence, and the map from Irr$_{\mathcal{H}}(A)$ to $\hat{A}$ is continuous and open, i.e. $\hat{A}$ has the quotient topology from Irr$_{\mathcal{H}}(A)$ (more care must be taken in describing the topology if $A$ has irreducible representations of different dimensions. The difficulty can be avoided by stabilization (XV.14.4.55.) if $A$ is separable, since there are natural homeomorphisms $[A \otimes \mathbb{K}]^\sim \cong \hat{A}$ and Prim($A \otimes \mathbb{K}$) $\cong$ Prim($A$).) This procedure also yields a Borel structure (the Mackey Borel structure) on $\hat{A}$, the quotient of the natural Borel structure on Irr$_{\mathcal{H}}(A)$ induced by its topology, which is finer than the Borel structure on $\hat{A}$ induced by its topology, and which is more useful than the topology if $\hat{A}$ is not $T_0$ (e.g. the Mackey Borel structure separates points of $\hat{A}$).

Matrix Algebras and Stable Algebras

As an example of a basic fact which can be proved easily by representation theory but for which there seems to be no decent space-free proof, we observe that a matrix algebra over a C*-algebra is also a C*-algebra.
**XV.14.4.49.** If $\mathcal{H}$ is a Hilbert space, then $\mathcal{B}(\mathcal{H}^n)$ is naturally isomorphic to the matrix algebra $M_n(\mathcal{B}(\mathcal{H}))$. (More generally, $\mathcal{B}(\mathcal{H}^n, \mathcal{H}^m)$ can be identified with the $m \times n$ matrices over $\mathcal{B}(\mathcal{H})$.) So if $A$ is a concrete C*-algebra of operators on $\mathcal{H}$, then the matrix algebra $M_n(A)$ acts naturally as a concrete C*-algebra of operators on $\mathcal{H}^n$ (it is easily seen to be complete). The adjoint of a matrix $(a_{ij})$ is the matrix $(b_{ij})$, where $b_{ij} = a^*_{ji}$.

**XV.14.4.50.** Thus, if $A$ is a C*-algebra, we need only to take a faithful representation $\pi$ of $A$ on a Hilbert space $\mathcal{H}$; then $\pi$ defines entrywise a faithful representation $\pi^{(n)}$ of $M_n(A)$ (with involution defined as above) on $\mathcal{H}^n$, and thus $M_n(A)$ is a C*-algebra with the induced operator norm.

**XV.14.4.51.** By XV.14.2.34., this operator norm is the unique C*-norm on $M_n(A)$ with this involution. When we use matrix algebras over C*-algebras, we will always implicitly use this involution and norm. Note that there is no explicit formula for the norm of a matrix in terms of the entries in general (e.g. if $A = \mathbb{M}_3$ and $n = 2$, cf. XV.10.6.6.). There is, however, a simple estimate: if $a = (a_{ij})$ is a matrix, then

$$\max_{i,j} \|a_{ij}\| \leq \|a\| \leq \sum_{i,j} \|a_{ij}\|$$

since it is easily seen that if $b$ is a matrix with exactly one nonzero entry $x$, then $\|b\| = \|x\|$.

**XV.14.4.52.** $M_n(A)$ is isomorphic to $A \otimes \mathbb{M}_n = A \otimes_{\mathbb{C}} \mathbb{M}_n$ (it), and it is convenient to use tensor product notation in matrix algebras. In $\mathbb{M}_n$, let $e_{ij}$ be the matrix with 1 in the $(i,j)$'th entry and zeroes elsewhere, and write $a \otimes e_{ij}$ for the element of $M_n(A)$ with $a$ in the $(i,j)$'th entry and zeroes elsewhere. The $e_{ij}$ are called the standard matrix units in $\mathbb{M}_n$. (There is a similar set of standard matrix units in $\mathbb{K}$.)

**XV.14.4.53.** As in XV.14.4.49., $\mathcal{B}(\mathcal{H}^\infty)$ can be identified with an algebra of infinite matrices over $\mathcal{B}(\mathcal{H})$ (although it is impossible to give an explicit description of which matrices give bounded operators). If $A$ is a concrete C*-algebra of operators on $\mathcal{H}$, let $M_\infty(A)$ denote the infinite matrices over $A$ with only finitely many nonzero entries. Then $M_\infty(A)$ acts naturally as a *-algebra of bounded operators on $\mathcal{H}^\infty$, and its closure in $\mathcal{B}(\mathcal{H}^\infty)$ is denoted $A \otimes \mathbb{K}$.

**XV.14.4.54.** If $A$ is an (abstract) C*-algebra, we can form $M_\infty(A)$ in the same manner, and by choosing a faithful representation $\pi$ of $A$, $M_\infty(A)$ may be identified with $M_\infty(\pi(A))$ and thus given a norm. Using the uniqueness of norm on $M_n(A)$ for each $n$, it is easily seen that the norm on $M_\infty(A)$ does not depend on the choice of $\pi$.

**XV.14.4.55.** Definition. The completion of $M_\infty(A)$ is called the stable algebra of $A$, denoted $A \otimes \mathbb{K}$.

This is also a special case of a tensor product of C*-algebras, hence the notation.

**XV.14.4.56.** Definition. A C*-algebra $A$ is stable if $A \cong A \otimes \mathbb{K}$. Two C*-algebras $A$ and $B$ are stably isomorphic if $A \otimes \mathbb{K} \cong B \otimes \mathbb{K}$.

Note that a stable C*-algebra is always nonunital.
**XV.14.4.57.** Proposition. $\mathcal{K}(\mathcal{H}) \otimes M_n = \mathcal{K}(\mathcal{H}^n)$ and $\mathcal{K}(\mathcal{H}) \otimes K = \mathcal{K}(\mathcal{H}^\infty)$. In particular,

$$K \cong M_n(K) \cong M_n \otimes K \cong K \otimes M_n$$

for all $n$, and $K$ is isomorphic to $K \otimes K$.

**XV.14.4.58.** Similarly, it is easily verified that $(A \otimes K) \otimes K \cong A \otimes K$ for any $A$, so the stable algebra $A \otimes K$ of a C*-algebra $A$ is a stable C*-algebra, and $A$ is stably isomorphic to $A \otimes K$. Two stable C*-algebras are stably isomorphic if and only if they are isomorphic. In the same way, $A$ and $M_n(A)$ are stably isomorphic for any $A$ and $n$. Stably isomorphic C*-algebras are “the same up to ‘size’;” a stable C*-algebra has uniformly infinite “size.” (This must be taken with a grain of salt in the non-$\sigma$-unital case, but can be made precise in a nice way for $\sigma$-unital C*-algebras). If $X$ is a locally compact Hausdorff space, then $C_0(X) \otimes K \cong C_0(X,K)$ (XV.14.1.3).

**XV.14.5. Universal and Enveloping C*-Algebras**

**XV.14.6. Exercises**

**XV.14.6.1.** Use XV.14.2.14. and XV.14.2.16. to give an alternate proof of XV.14.1.9., as follows. Let $B$ be a commutative C*-algebra, which may be assumed unital by XV.14.1.2., and $x = x^* \in B$.

(a) Suppose $\lambda = \alpha + i\beta \in \sigma_B(x)$ with $\alpha, \beta \in \mathbb{R}$. By XV.14.2.16. there is a $\phi \in \hat{B}$ with $\phi(x) = \lambda$.

(b) For $t \in \mathbb{R}$, set $x_t = x + i\lambda$. Then $\|x_t\|^2 = \|x^*_tx_t\| = \|x^2 + t^2\| \leq \|x^2\| + t^2 = \|x\|^2 + t^2$.

(c) Since $|\phi(x_t)| \leq \|x_t\|$ by XV.14.2.14., we have

$$\alpha^2 + (\beta + t)^2 \leq \|x\|^2 + t^2$$

for all $t \in \mathbb{R}$. Show that this implies that $\beta = 0$.

(e) If $A$ is a general not necessarily commutative C*-algebra (which may be assumed unital) and $x = x^* \in A$, the above argument can be applied to the C*-subalgebra $B$ of $A$ generated by $x$ and 1. Why does this give an alternate proof of the full result XV.14.1.9. (what is the relation between $\sigma_B(x)$ and $\sigma_A(x)$)?

**XV.14.6.2.** (a) Let $I$ be a closed bounded interval in $\mathbb{R}$, and $A$ the set of bounded functions from $I$ to $\mathbb{C}$ whose real and imaginary parts are in $2\mathcal{C}_c(I)$ (XII.2.3.2.). Give $A$ pointwise operations and supremum norm. Show that $A$ is a (unital) commutative C*-algebra. [Use XII.2.6.7.]

(b) Let $B$ be the set of regulated functions on $I$ (V.8.6.7.). Show that $B$ is a (unital) C*-subalgebra of $A$ containing $C(I)$.

(c) Show that $B$ is homeomorphic to $I \times \{0,1\}$ with the order topology from the lexicographic ordering (). The quotient map to $I \cong C(I)$ is projection onto the first coordinate. Note that $B$ is not metrizable, so $B$ is not separable (this has an easy direct proof too).

(d) Try to describe $A$. (Note that $\hat{B}$ is a quotient of $\hat{A}$.)

**XV.14.6.3.** Let $A$ be the set of bounded uniformly continuous functions from $\mathbb{R}$ to $\mathbb{C}$. Show that $A$ is a proper unital C*-subalgebra of $BC(\mathbb{R})$. Thus $A$ is a quotient of $BC(\mathbb{R}) \cong \beta \mathbb{R}$. Describe the quotient map.
Chapter XVI

Topological Groups and Abstract Harmonic Analysis

XVI.1. Topological Groups

XVI.1.1. Groups and Homomorphisms

XVI.1.2. Topological Groups

XVI.1.2.1. Definition. A topological group is a group $G$ with a topology such that the group operation $(x, y) \mapsto xy$ is a continuous function from $G \times G$ to $G$ (i.e. multiplication is jointly continuous on $G$), and inversion $x \mapsto x^{-1}$ is a continuous function from $G$ to $G$.

XVI.1.2.2. Note that both the group operation and the topology are parts of the definition of a topological group; thus it is technically incorrect to say “Let $G$ be a topological group” (although we do it all the time). In fact, a group can have many different topologies making it into a topological group.

XVI.1.2.3. Examples. (i) Let $G$ be any group. Then $G$ is a topological group with the discrete topology or the indiscrete topology.

A topological group with the indiscrete topology is not very interesting. However, a group with the discrete topology is much more interesting than might be at first anticipated. In fact, we can regard the theory of groups (with no topology) as a special case of the theory of topological groups by using the discrete topology (although this is logically circular if not done properly).

(ii) The additive group $\mathbb{R}$ (or $\mathbb{R}^n$) with its ordinary topology is a topological group. More generally, the additive group of any normed vector space is a topological group with the topology induced by the norm.

(iii) Let $GL_n(\mathbb{R})$ be the multiplicative group of invertible $n \times n$ matrices with real entries, with the relative topology as a subset of $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$. Then $GL_n(\mathbb{R})$ is a topological group.

(iv) Any subgroup of a topological group is a topological group with the relative topology.

(v) Groups of homeomorphisms of topological spaces can often be given natural topologies making them into topological groups, although the details can be subtle (cf. XVI.1.7.8–XVI.1.7.14.).
These examples barely scratch the surface of the possibilities.

The condition defining a topological group can be streamlined:

**XVI.1.2.4. Proposition.** Let $G$ be a group with a topology. Then $G$ is a topological group under the given topology if and only if the map $\phi : G \times G \to G$ given by $\phi(x, y) = x^{-1}y$ is continuous.

**Proof:** If $G$ is a topological group, then $\phi$ is a composition of multiplication and inversion, hence continuous. Conversely, if $\phi$ is continuous, then

$$x \mapsto (x, e) \mapsto x^{-1}e = x^{-1}$$

is continuous, and hence

$$(x, y) \mapsto (x^{-1}, y) \mapsto (x^{-1})^{-1}y = xy$$

is continuous, so $G$ is a topological group in the given topology.

**XVI.1.2.5.** Joint continuity of multiplication in a topological group implies separate continuity (). But separate continuity of multiplication does not imply joint continuity (Exercise XVI.1.7.6.). There are situations where separate continuity does imply joint continuity ()

**Translations and Homogeneity**

**XVI.1.2.6.** **Definition.** Let $G$ be a topological group, and $g \in G$. The functions $\lambda_g, \rho_g$ from $G$ to $G$ defined by

$$\lambda_g(x) = gx$$

$$\rho_g(x) = xg$$

are called *left translation* and *right translation* by $g$ respectively.

**XVI.1.2.7.** **Proposition.** Let $G$ be a topological group, and $g \in G$. Then the left and right translations $\lambda_g$ and $\rho_g$ are homeomorphisms from $G$ to $G$. If $g \neq e$, these homeomorphisms have no fixed points.

**Proof:** Translations are continuous by (separate) continuity of the group operation. Left translation $\lambda_g$ has inverse $\lambda_{g^{-1}}$, hence is a bijection and a homeomorphism; similarly $\rho_g$ has inverse $\rho_{g^{-1}}$. If $\lambda_g(x) = x$ for some $x \in G$, i.e. $gx = x$, then $g = (gx)x^{-1} = xx^{-1} = e$, and similarly for $\rho_g$.

**XVI.1.2.8.** **Corollary.** If $G$ is a topological group, then the topology of $G$ is *homogeneous*, i.e. if $x, y \in G$, then there is a homeomorphism from $G$ to $G$ sending $x$ to $y$.

**Proof:** $\lambda_{yx^{-1}}$ is such a homeomorphism.

Thus a topological group “looks the same” around every point.
XVI.1.2.9. Similarly, inversion \( \iota \), where \( \iota(x) = x^{-1} \), is a continuous map from \( G \) to \( G \) which is its own inverse, i.e. \( \iota \) is a homeomorphism.

The next result, although simple, appears quite bizarre and surprising at first glance, and is a dramatic illustration of the maxim that being open is not the opposite of being closed for subsets of a topological space. See Exercise XVI.1.7.2. for a generalization.

XVI.1.2.10. **Proposition.** Let \( G \) be a topological group. Any open subgroup of \( G \) is closed.

**Proof:** Let \( H \) be an open subgroup of \( G \). Then, for any \( g \in G \), the left coset \( gH \) is the image of \( H \) under the homeomorphism \( \lambda_g \), and hence \( gH \) is also open. The complement of \( H \) in \( G \) is the union of the nonidentity left cosets of \( H \) in \( G \), hence open. \( \Box \)

See Exercise XVI.1.7.2. for an alternate proof.

**Fundamental Systems of Neighborhoods**

XVI.1.2.11. If \( G \) is a topological group and \( U \) is a neighborhood of \( e \), then for any \( g \in G \), \( gU = \lambda_g(U) \) and \( Ug = \rho_g(U) \) are neighborhoods of \( g \), and conversely if \( V \) is a neighborhood of \( g \) then \( g^{-1}V \) and \( Vg^{-1} \) are neighborhoods of \( e \). Thus to describe a topological group topology on a group \( G \), it suffices to specify the neighborhoods of \( e \).

XVI.1.2.12. If \( A \) and \( B \) are subsets of \( G \), write \( AB = \{ xy : x \in A, y \in B \} \), and \( A^2 = AA \). If \( A, B, C \) are subsets of \( G \), then \( (AB)C = A(BC) \) by associativity; call this subset \( ABC \). Products of more than three subsets can be defined similarly. \( A^n \) has an unambiguous meaning for any \( n \in \mathbb{N} \), and \( A^nA^m = A^{n+m} \).

If \( A \subseteq G \), write \( A^{-1} = \iota(A) = \{ x^{-1} : x \in A \} \).

XVI.1.2.13. **Proposition.** If \( G \) is a topological group, \( A, B \subseteq G \), and one of \( A, B \) is open, then \( AB \) is open. If \( A \) is open, then \( A^{-1} \) is also open.

**Proof:** If \( B \) is open, then \( AB = \lambda_a(B) \) is open for each \( a \in A \), so \( AB = \bigcup_{a \in A} aB \) is open. The proof if \( A \) is open is similar. \( A^{-1} = \iota(A) \) is open if \( A \) is. \( \Box \)

XVI.1.2.14. **Proposition.** Let \( G \) be a topological group. The collection \( \mathcal{N}_e \) of neighborhoods of \( e \) has the following properties:

(i) If \( U \in \mathcal{N}_e \), then there is a \( V \in \mathcal{N}_e \) with \( V^2 \subseteq U^o \).

(ii) There is a \( W \in \mathcal{N}_e \) with \( W = W^{-1} \) and \( W \subseteq U^o \).

(iii) For every \( g \in G \), there is a \( V \in \mathcal{N}_e \) with \( gVg^{-1} \subseteq U^o \).

**Proof:** (i): Let \( \mu : G \times G \to G \) be multiplication, i.e. \( \mu(x, y) = xy \). Then \( \mu^{-1}(U^o) \) is an open neighborhood of \( (e, e) \) in \( G \times G \), hence contains a neighborhood of the form \( V_1 \times V_2 \) where each \( V_j \) is in \( \mathcal{N}_e \). Set \( V = V_1 \cap V_2 \).

(ii): Set \( W = U^o \cap (U^o)^{-1} \).

(iii): Set \( V = g^{-1}U^o g = \rho_g(\lambda_g^{-1}(U^o)) = \lambda_g^{-1}(\rho_g(U^o)) \), which is in \( \mathcal{N}_e \). \( \Box \)
XVI.1.2.15. Corollary. If $G$ is a topological group, and $\mathcal{F}$ is a local base at $e$ for the topology (not necessarily consisting of open sets), then $\mathcal{F}$ has the following properties:

(i) If $U, V \in \mathcal{F}$, then there is a $W \in \mathcal{F}$ with $W \subseteq U \cap V$.
(ii) If $U \in \mathcal{F}$ and $g \in U$, then there is a $V \in \mathcal{F}$ with $Vg \subseteq U$.
(iii) If $U \in \mathcal{F}$, then there is a $V \in \mathcal{F}$ with $V^{-1}V \subseteq U$.
(iv) If $U \in \mathcal{F}$ and $g \in G$, then there is a $V \in \mathcal{F}$ with $gV^{-1} \subseteq U$.

Such an $\mathcal{F}$ has additional properties such as:

(iiiia) If $U \in \mathcal{F}$, there is a $V \in \mathcal{F}$ with $V^2 \subseteq U$.
(iiib) If $U \in \mathcal{F}$, there is a $V \in \mathcal{F}$ with $V^{-1} \subseteq U$.

In fact, (iiiia) and (iiib) together are equivalent to (iii).

XVI.1.2.16. Definition. Let $G$ be a group. A fundamental system of neighborhoods of $e$ is a nonempty collection $\mathcal{F}$ of subsets of $G$, each containing $e$, satisfying (i)–(iv) of XVI.1.2.15. (or equivalently (i), (ii), (iiiia), (iiib), (iv)).

Note that in this definition, there is no specified topology on $G$, so the sets in $\mathcal{F}$ are not “neighborhoods” of $e$ in $G$ in the usual sense of the term. But the terminology is justified by:

XVI.1.2.17. Theorem. Let $G$ be a group. If $\mathcal{F}$ is a fundamental system of neighborhoods of $e$ in $G$, then there is a unique topology $T$ on $G$ making $G$ a topological group, for which $\mathcal{F}$ is a local base at $e$.

XVI.1.2.18. So a topological group topology on a group $G$ can be specified by giving a fundamental system of neighborhoods of $e$, and every topological group topology on $G$ arises in this manner. Specifying a topology via a fundamental system of neighborhoods of $e$ is one of the most convenient ways of defining topologies on groups.

Homogeneous Spaces and Quotients

XVI.1.2.19. If $G$ is a topological group and $H$ a subgroup of $G$, not necessarily normal, we can form the left coset space $G/H$. Since $G/H$ is a set-theoretic quotient of $G$, i.e. there is a natural quotient map $\pi : G \to G/H$, we may give $G/H$ the quotient topology. $G/H$ with this topology is called the homogeneous space of $G$ by $H$ (the terminology is justified by XVI.1.2.20.(ii)).
**XVI.1.2.20.** **Proposition.** Let $G$ be a topological group and $H$ a subgroup of $G$. Let $G/H$ be the homogeneous space. Then

(i) The quotient map $\pi : G \to G/H$ is continuous and open.

(ii) If $g \in G$, the map $\tilde{\lambda}_g : G/H \to G/H$ defined by $\tilde{\lambda}_g(xH) = gxH$ is a well-defined homeomorphism, and in particular the topology of $G/H$ is homogeneous in the sense of XVI.1.2.8. The map $(g, xH) \mapsto \lambda_g(xH) = gxH$ is continuous from $G \times G/H$ to $G/H$.

**Proof:** (i) The quotient map $\pi$ is continuous by definition of the quotient topology. If $U$ is open in $G$, then $\pi^{-1}(\pi(U)) = UH$, which is open in $G$ by XVI.1.2.13.; hence $\pi(U)$ is open in $G/H$ by definition of the quotient topology.

(ii): If $xH = yH$, i.e. $y = xz$ for some $z \in H$, then $gy = gxz$, so $gy \in gxH$, i.e. $gyH = gxH$ and $\tilde{\lambda}_g$ is well defined, and is a bijection since $\lambda_g^{-1} = \tilde{\lambda}_g^{-1}$. If $V \subseteq G/H$ is open, then $\pi^{-1}(V)$ is open in $G$, so $\lambda_g^{-1}(\pi^{-1}(V))$ is open in $G$ since $\lambda_g^{-1}$ is a homeomorphism on $G$; thus $\lambda_g^{-1}(V) = \pi(\lambda_g^{-1}(\pi^{-1}(V)))$ is open by (i). So $\lambda_g$ is continuous. Since $\lambda_g^{-1} = \tilde{\lambda}_g^{-1}$ is also continuous, $\tilde{\lambda}_g$ is a homeomorphism. For the last statement, if $x, y \in G$, then $\tilde{\lambda}_{xy^{-1}}$ sends $xH$ to $yH$, so $G/H$ is homogeneous.

**XVI.1.2.21.** **Proposition.** Let $G$ be a topological group, and $H$ a closed normal subgroup. Then

(i) $G/H$ is a topological group in the quotient topology.

(ii) If $G'$ is another topological group, and $\phi : G \to G'$ a continuous homomorphism whose kernel contains $H$, then the induced homomorphism $\psi : G/H \to G'$ () is continuous.

**Proof:** (i): Let $\tilde{V}$ be an open set in $G/H$, and let $V = \pi^{-1}(\tilde{V})$. Then $V$ is open in $G$. Since $\mu$ is continuous, $U = \mu^{-1}(V)$ is open in $G \times G$. The quotient map $\pi \times \pi$ from $G \times G$ to $(G/H) \times (G/H)$ is open, so $U = (\pi \times \pi)(U)$ is open in $(G/H) \times (G/H)$, and $U = \tilde{\mu}^{-1}(\tilde{V})$. Thus $\tilde{\mu}$ is continuous. A similar argument shows that $\tilde{\psi}$ is continuous on $G/H$.

The homeomorphisms $\tilde{\lambda}_g$ are an example of a group action:

**XVI.1.2.22.** **Definition.** Let $G$ be a topological group and $X$ a topological space. A group action of $G$ on $X$ is a continuous function $(g, x) \mapsto \alpha_g(x)$ from $G \times X$ to $X$ such that $\alpha_e(x) = x$ and $\alpha_{gh}(x) = \alpha_g(\alpha_h(x))$ for all $x \in X$, $g, h \in G$.

**XVI.1.2.23.** The point $\alpha_g(x)$ is often written $g \cdot x$. For fixed $g$, $\alpha_g$ is a continuous function from $X$ to $X$ with inverse $\alpha_{g^{-1}}$, i.e. $\alpha_g$ is a homeomorphism from $X$ to $X$. The map $\alpha : G \to \text{Homeo}(X)$ is a group-homomorphism, which is continuous if $\text{Homeo}(X)$ is given the topology of pointwise convergence.

**Separation Axioms**
Topological groups automatically have nice separation properties:

1895
XVI.1.2.24. **Proposition.** Let $G$ be a topological group. The following are equivalent:

(i) $G$ is $T_0$.

(ii) $G$ is Hausdorff.

(iii) $G$ is completely regular.

**Proof:** (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) is trivial. Since the topology on $G$ comes from a uniform structure (XVI.1.3.4.), if $G$ is $T_0$, it is completely regular, so (i) $\Rightarrow$ (iii).

XVI.1.2.25. There is a simple direct proof that a topological group is $T_3$, and hence that a $T_0$ topological group is Hausdorff; cf. Exercise XVI.1.7.3.

XVI.1.2.26. **Corollary.** A second countable $T_0$ topological group is metrizable.

One can do much better (XVI.1.3.7.).

**Closed Subgroups**

XVI.1.2.27. **Proposition.** Let $G$ be a topological group, and $H$ a subgroup. Then $G/H$ is Hausdorff (completely regular) if and only if $H$ is closed in $G$.

XVI.1.2.28. **Proposition.** Let $G$ be a topological group, and $H$ a subgroup of $G$. Then $\bar{H}$ is also a subgroup of $G$. If $H$ is a normal subgroup, so is $\bar{H}$.

XVI.1.2.29. Thus $\langle e \rangle$ is the smallest closed subgroup of $G$. It is normal, and $G/\langle e \rangle$ is the largest Hausdorff quotient group of $G$. Any continuous homomorphism from $G$ into a Hausdorff topological group factors through $G/\langle e \rangle$. The space $G/\langle e \rangle$ is the maximal $T_0$ quotient of $G$ in the sense of ()

**Products of Topological Groups**

XVI.1.2.30. **Proposition.** Any product of topological groups is a topological group with the product topology.

**Inverse Limits**

**Connectedness**

XVI.1.2.31. **Proposition.** Let $G$ be a topological group. If $U$ is any open neighborhood of $e$ in $G$, then the subgroup of $G$ generated by $U$ is open, and coincides with the subgroup of $G$ generated by $\bar{U}$.

**Proof:** The subset $[U \cup (U^{-1})]^n$ of $G$ is open for each $n$ by XVI.1.2.13. The subgroup of $G$ generated by $U$ is $\bigcup_{n=1}^{\infty}[U \cup (U^{-1})]^n$. Since an open subgroup is closed (XVI.1.2.10.), this subgroup contains $\bar{U}$.

1896
XVI.1.2.32.  COROLLARY. Let $G$ be a connected topological group. Then every neighborhood of $e$ generates $G$ as a group.

Proof: Every open subgroup of a topological group is clopen, so a connected topological group cannot contain any proper open subgroups. Every neighborhood of $e$ in $G$ contains an open neighborhood of $e$.  

XVI.1.2.33.  PROPOSITION. Let $G$ be a topological group and $H$ a subgroup. If $H$ and $G/H$ are connected, then $G$ is connected.

Proof: Suppose $H$ is connected and $G$ is not connected. Let $U$ and $V$ be nonempty disjoint clopen sets in $G$ with $U \cup V = G$. Since each left coset of $H$ is connected, it is entirely contained in either $U$ or $V$, i.e. $U$ and $V$ are each unions of left cosets of $H$. But then $\pi(U)$ and $\pi(V)$ are disjoint clopen subsets of $G/H$, so $G/H$ is not connected.

XVI.1.2.34.  PROPOSITION. Let $G$ be a topological group. Then the connected component of the identity $e$ is a closed normal subgroup $H$ of $G$, and $G/H$ is totally disconnected.

Proof: Let $H$ be the connected component of $e$ in $G$. Then $H$ is a closed subset of $G$. $H \times H$ is a connected subset of $G \times G$, and hence its image $H^2$ in $G$ under multiplication is connected, and contains $e$; thus $H^2 \subseteq H$, i.e. $H$ is closed under multiplication. Similarly, $H^{-1}$ is connected and contains $e$, thus $H^{-1} \subseteq H$ and $H$ is closed under inversion. So $H$ is a subgroup of $G$. If $g \in G$, then $gHg^{-1}$ is a connected subset of $G$, hence $gHg^{-1} \subseteq H$. Then $H = g^{-1}(gHg^{-1})g \subseteq g^{-1}Hg \subseteq H$, so $g^{-1}Hg = H$ and $H$ is normal.

To see that $G/H$ is totally disconnected, it suffices to show that the connected component of $\bar{e}$ in $G/H$ is $\{\bar{e}\}$. Let $K$ be the connected component of $\bar{e}$ in $G/H$, and $G' = \pi^{-1}(K)$. Then $G'$ is a subgroup of $G$ and $G'/H \cong K$ is connected, and since $H$ is connected, $G'$ is connected by XVI.1.2.33. Since $H \subseteq G'$, we have $G' = H$ so $K = \{\bar{e}\}$.

XVI.1.3. Uniformities on Topological Groups

Every topological group has a natural uniform structure (actually three natural uniform structures which may or may not coincide). In fact, these uniform structures are one of the models for general uniform structures ()

XVI.1.3.1.  DEFINITION. Let $(G, T)$ be a topological group.

(i) The left uniform structure $\mathcal{L}$ on $G$ is the uniform structure generated by the sets

$$L_U = \{(x, y) \in G \times G : x^{-1}y \in U\}$$

as $U$ ranges over all open neighborhoods of $e$ in $G$.

(ii) The right uniform structure $\mathcal{R}$ on $G$ is the uniform structure generated by the sets

$$R_U = \{(x, y) \in G \times G : xy^{-1} \in U\}$$

as $U$ ranges over all open neighborhoods of $e$ in $G$.

(iii) The two-sided uniform structure $\mathcal{U}$ on $G$ is the uniform structure generated by the sets $L_U$ and $R_U$ as $U$ ranges over all open neighborhoods of $e$ in $G$.  

1897
XVI.1.3.2. The uniform structures $\mathcal{L}$ and $\mathcal{R}$ coincide in many cases, e.g. if $G$ is compact () or abelian, but not in general (). The uniform structure $\mathcal{U}$ coincides with $\mathcal{L}$ or $\mathcal{R}$ if and only if $\mathcal{L} = \mathcal{R}$.

XVI.1.3.3. **Proposition.** Let $G$ be a topological group. The following are equivalent:

(i) $\mathcal{L} = \mathcal{R}$.

(ii) $\mathcal{L} = \mathcal{U}$.

(iii) $\mathcal{R} = \mathcal{U}$.

(iv) $G$ has a fundamental system of neighborhoods of $e$ invariant under all inner automorphisms, i.e. neighborhoods $U$ such that $gUg^{-1} = U$ for all $g \in G$.

However, even though the uniform structures do not coincide in general, they all define the given topology on $G$:

XVI.1.3.4. **Proposition.** Let $(G, T)$ be a topological group. Then the topology on $G$ defined by any of the three uniform structures $\mathcal{L}, \mathcal{R}, \mathcal{U}$ is $T$.

XVI.1.3.5. **Proposition.** Let $G$ be a topological group. Then

(i) The translations $\lambda_g$ are uniformly continuous for $\mathcal{L}$ and $\mathcal{U}$, for every $g \in G$.

(ii) The translations $\rho_g$ are uniformly continuous for $\mathcal{R}$ and $\mathcal{U}$, for every $g \in G$.

(iii) Inversion $\iota$ is uniformly continuous for $\mathcal{U}$.

Inversion is not uniformly continuous for $\mathcal{L}$ or $\mathcal{R}$ in general: inversion interchanges $\mathcal{L}$ and $\mathcal{R}$. In fact:

XVI.1.3.6. **Proposition.** Let $G$ be a topological group. The following are equivalent:

(i) The translations $\lambda_g$ are uniformly continuous for $\mathcal{R}$, for every $g \in G$.

(ii) The translations $\rho_g$ are uniformly continuous for $\mathcal{L}$, for every $g \in G$.

(iii) Inversion $\iota$ is uniformly continuous for $\mathcal{L}$.

(iv) Inversion $\iota$ is uniformly continuous for $\mathcal{R}$.

(v) $\mathcal{L} = \mathcal{R}(= \mathcal{U})$. 

1898
Metrization

It is an immediate corollary of XVI.1.3.4. and () that a first countable topological group is metrizable, since \( \mathcal{L} \), \( \mathcal{R} \), and \( \mathcal{U} \) are all countably generated if \( G \) is first countable. There is a more precise result:

XVI.1.3.7. **Corollary.** Every first countable Hausdorff topological group is metrizable under a left-invariant metric which generates the uniformity \( \mathcal{L} \).

XVI.1.3.8. If \( d \) is a left-invariant metric for \( G \), then the metric \( d'(g,h) = d(g^{-1}, h^{-1}) \) is right-invariant and defines the topology of \( G \) (and the uniformity \( \mathcal{R} \)), so “left-invariant” can be replaced by “right-invariant” in XVI.1.3.7.

However, \( G \) does not have a two-sided invariant metric in general; in fact, it has a two-sided invariant metric if and only if it satisfies the conditions of XVI.1.3.6. There is a metric giving the uniformity \( \mathcal{U} \) if \( G \) is first countable, for example

\[
d'(g,h) = d(g,h) + d(g^{-1}, h^{-1})
\]

if \( d \) is left-invariant; but \( d' \) is neither left- or right-invariant unless \( \mathcal{L} = \mathcal{R} = \mathcal{U} \).

Completion

XVI.1.3.9. **Definition.** A topological group \( G \) is complete if it is complete with respect to the uniformity \( \mathcal{U} \).

A group which is complete in this sense is often called *Raikov-complete*.

XVI.1.3.10. **Proposition.** Let \( G \) be a complete topological group, and \( H \) a subgroup. Then

(i) \( H \) is complete if and only if it is closed in \( G \).

(ii) If \( H \) is normal, then \( G/H \) is complete.

XVI.1.3.11. **Theorem.** Let \( G \) be a topological group. Then there is a complete topological group \( \hat{G} \) containing \( G \) as a dense subgroup. \( \hat{G} \) is unique up to topological isomorphism which is the identity on \( G \). \( \hat{G} \) is called the completion, or Raikov completion, of \( G \).

XVI.1.3.12. A topological group \( G \) does not generally have a completion with respect to \( \mathcal{L} \) or \( \mathcal{R} \) (more precisely, the completion of \( G \) with respect to \( \mathcal{L} \) or \( \mathcal{R} \) as a uniform space, called the Weil completion of \( G \), does not have a compatible structure as a topological group in general); in fact, it has such a completion (as a topological group) if and only if \( \mathcal{L} = \mathcal{R} \).

XVI.1.4. **Locally Compact Topological Groups**

Topological groups which are locally compact and Hausdorff are especially important, and have a lot of nice associated structure which allows analysis to be done on them.
**XVI.1.4.1. Definition.** A *locally compact group* is a locally compact Hausdorff topological group, i.e. a topological group whose topology is locally compact and Hausdorff. A *compact group* is a compact Hausdorff topological group.

Thus the words “Hausdorff” and “topological” will be understood in the names “locally compact group” and “compact group.”

**XVI.1.4.2. Examples.**

(i) The additive group $\mathbb{R}$, or $\mathbb{R}^n$, with its usual topology, is a locally compact group.

(ii) The group $GL_n(\mathbb{R})$ (XVI.1.2.3. (iii)) is locally compact (in its usual topology).

(iii) $\mathbb{T}$ is a compact group (in its usual topology). Any product of copies of $\mathbb{T}$ is a compact group.

(iv) Any group with the discrete topology is a locally compact group.

(v) Any closed subgroup of a locally compact group is locally compact.

(vi) The additive group of an infinite-dimensional normed vector space is not a locally compact group in the norm topology.

(vii) The additive group $\mathbb{Q}$ is not a locally compact group in its usual topology (the relative topology from $\mathbb{R}$). But it is in the discrete topology.

By homogeneity, we obtain:

**XVI.1.4.3. Proposition.** A Hausdorff topological group $G$ is locally compact if and only if there is a compact neighborhood of $e$ in $G$.

**XVI.1.4.4. Proposition.** Let $G$ be a locally compact group. Then $G$ has an open subgroup which is $\sigma$-compact. If $G$ is connected, it is $\sigma$-compact.

**Proof:** Let $U$ be a neighborhood of $e$ in $G$ with $\bar{U}$ compact. Let $H$ be the subgroup of $G$ generated by $U$. Then $H$ is open and is the subgroup of $G$ generated by $\bar{U}$ (XVI.1.2.31.). If $V = U \cup U^{-1}$, then $\bar{V}$ is compact and generates $H$. We have that $\bar{V}^n$ is compact since it is the image of $\bar{V} \times \bar{V}$ under multiplication; by induction, $\bar{V}^n$ is compact for all $n$. Thus $H = \cup_{n=1}^{\infty} \bar{V}^n$ is $\sigma$-compact.

**XVI.1.4.5. Corollary.** Let $G$ be a locally compact group. Then $G$ partitions into clopen subsets, each of which is $\sigma$-compact and thus Lindelöf. So $G$ is paracompact and hence normal.

**Proof:** The cosets of a $\sigma$-compact open subgroup give such a partition.
**Corollary.** Let $G$ be a locally compact group. The following are equivalent:

(i) $G$ is second countable.

(ii) $G$ is first countable and $\sigma$-compact.

(iii) $G$ is first countable and compactly generated.

**Proposition.** Let $G$ be a topological group, $K$ a compact subset of $G$, and $U$ an open neighborhood of $K$. Then there is an open neighborhood $V$ of $e$ in $G$ such that $KV \subseteq U$.

**Proof:** For each $g \in K$, $ge \in U$, so by joint continuity there are neighborhoods $W_g$ of $g$ and $V_g$ of $e$ with $W_gV_g \subseteq U$. The $W_g$ cover $K$, so there is a finite subcover $\{W_{g_1}, \ldots, W_{g_n}\}$. Set $V = V_{g_1} \cap \cdots \cap V_{g_n}$. If $g \in K$, then $g \in W_{g_k}$ for some $k$, so $gV \subseteq W_{g_k}V \subseteq W_{g_k}V_{g_k} \subseteq U$. Thus $KV \subseteq U$.

Similarly, we have:

**Proposition.** Let $G$ be a topological group, $K$ a compact subset of $G$, and $U$ an open neighborhood of $e$ in $G$ such that $g^{-1}Vg \subseteq U$ for all $g \in K$.

**Proof:** For each $g \in K$, $g^{-1}eg = e$, so by joint continuity there are neighborhoods $W_g$ of $g$ and $V_g$ of $e$ with $h^{-1}Vgh \subseteq U$ for all $h \in W_g$. The $W_g$ cover $K$, so there is a finite subcover $\{W_{g_1}, \ldots, W_{g_n}\}$. Set $V = V_{g_1} \cap \cdots \cap V_{g_n}$. If $g \in K$, then $g \in W_{a_k}$ for some $k$, so $g^{-1}Vg \subseteq g^{-1}V_{g_k}g \subseteq U$.

**Proposition.** Let $G$ be a (necessarily locally compact) group. If $U$ is a compact open neighborhood of $e$ in $G$, then $U$ contains a compact open subgroup of $G$.

**Proof:** There is a neighborhood $V$ of $G$ with $V = V^{-1}$ and $UV \subseteq U$ (XVI.1.4.7 with $K = U$). Then $V^2 \subseteq U$, and by induction $V^n \subseteq U$. Thus the subgroup of $G$ generated by $V$, which is open, is contained in $U$.

**Theorem.** Let $G$ be a locally compact group. Then every neighborhood of $e$ in $G$ contains a compact subgroup $H$ such that $G/H$ is metrizable. If $G$ is compactly generated, the subgroup $H$ may be chosen normal.

**Proof:** First let $U$ be a compact neighborhood of $e$ in $G$, and let $G_0$ be the subgroup of $G$ generated by $U$. Then $G_0$ is an open subgroup. If $H$ is a subgroup of $G_0$, then $G/H$ is a separated union of cosets of $G_0/H$; hence $G/H$ is metrizable if $G_0/H$ is metrizable. Thus it suffices to show the last statement, i.e. that if $G$ is generated by a compact symmetric neighborhood $U$ of $e$, then for any neighborhood $V$ of $e$, $G$ has a compact normal subgroup $H$ contained in $V$ with $G/H$ metrizable (first countable).

Fix a neighborhood $V$, which we may take to be compact and symmetric. Set $V_1 = V$. Generate a decreasing sequence $(V_n)$ of symmetric compact neighborhoods of $e$ such that $g^{-1}V_{n+1}g \subseteq V_n$ for all $g \in U$.
(XVI.1.4.8.) and such that $V_{n+1}^2 \subseteq V_n$. Set $H = \cap_n V_n$. Then $H$ is a compact subgroup of $G$, and $g^{-1}Hg \subseteq H$ for all $g \in U$; since $U$ generates $G$, $H$ is normal in $G$. If $\pi : G \to G/H$ is the quotient map, the sets $(\pi(V_n))$ are compact neighborhoods of the identity which form a local neighborhood base by XI.11.1.13. Thus $G/H$ is first countable.

**XVI.1.4.11.** PROPOSITION. Every locally compact group is complete.

**XVI.1.4.12.** PROPOSITION. Let $G$ be a locally compact group and $H$ a closed subgroup. Then $G/H$ is locally compact in the quotient topology. So if $H$ is normal, $G/H$ is a locally compact group.

The next result is closely related to the Open Mapping Theorem for Banach spaces (), with a similar proof based on the Baire Category Theorem. This result is a special case of XVI.1.6.2.

**XVI.1.4.13.** THEOREM. [Open Mapping Theorem for Locally Compact Groups] Let $G$ and $G'$ be locally compact groups, and $\phi$ a continuous group homomorphism from $G$ onto $G'$. If $G$ is $\sigma$-compact (so $G'$ is also $\sigma$-compact), then $\phi$ is an open mapping, and thus $G'$ is topologically isomorphic to $G/\ker(\phi)$.

As in the Banach space case (), we also get a Closed Graph Theorem for locally compact groups:

**XVI.1.4.14.** THEOREM. [Closed Graph Theorem for Locally Compact Groups] Let $G$ and $G'$ be $\sigma$-compact locally compact groups, and $\phi : G \to G'$ a homomorphism. If the graph $\Gamma(\phi)$ is a closed subset of $G \times G'$, then $\phi$ is continuous.

**Proof:** $G \times G'$ is locally compact and $\sigma$-compact, and $\Gamma(\phi)$ is a closed subgroup, hence also locally compact and $\sigma$-compact. The map $\pi_G : (x, \phi(x)) \mapsto x$ is a continuous bijective homomorphism from $\Gamma(\phi)$ onto $G$, hence a topological isomorphism by XVI.1.4.13.; thus the inverse map $\psi$ is continuous, and hence $\phi = \pi_{G'} \circ \psi$ is continuous.

**XVI.1.4.15.** COROLLARY. Let $G$ be a group, and let $\mathcal{T}$ be a Hausdorff topology on $G$, not necessarily a topological group topology. Then there is at most one topology on $G$ stronger than $\mathcal{T}$ making it into a $\sigma$-compact locally compact group.

**Proof:** Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be topologies stronger than $\mathcal{T}$ making $G$ into a locally compact $\sigma$-compact group, and let $\phi$ be the identity map from $(G, \mathcal{T}_1)$ to $(G, \mathcal{T}_2)$. The graph of $\phi$ is the diagonal in $G \times G$. The product topology $\mathcal{T}_1 \times \mathcal{T}_2$ on $G \times G$ is stronger than the topology $\mathcal{T} \times \mathcal{T}$. Since $\mathcal{T}$ is Hausdorff, the diagonal is closed in $G \times G$ for the topology $\mathcal{T} \times \mathcal{T}$, and hence also closed in the topology $\mathcal{T}_1 \times \mathcal{T}_2$. Thus $\phi$ has closed graph and is therefore continuous, i.e. $\mathcal{T}_1$ is stronger than $\mathcal{T}_2$. The same argument with $\mathcal{T}_1$ and $\mathcal{T}_2$ interchanged shows that $\mathcal{T}_2$ is stronger than $\mathcal{T}_1$. □

1902
XVI.1.4.16. XVI.1.4.13–XVI.1.4.15. are false in general without “σ-compact”: the additive group \( \mathbb{R} \) is a locally compact group in either the ordinary topology or the discrete topology. The \( T \) in XVI.1.4.15. is also necessary in general if \( G \) is uncountable: \( \mathbb{R} \) has discontinuous group-automorphisms (using AC).

XVI.1.4.17. XVI.1.4.15. does not say that a group has at most one σ-compact Hausdorff topology making it into a topological group. In fact, there is an abelian group \( G \) which has \( 2^{\aleph_0} \) mutually nonequivalent topological group topologies, each of which is compact and metrizable [Fuc59] (where two topological group topologies \( T_1 \) and \( T_2 \) on a group \( G \) are equivalent if \( (G, T_1) \) and \( (G, T_2) \) are isomorphic as topological groups). However, we do have:

XVI.1.4.18. Corollary. The only locally compact group topology on a countable group is the discrete topology.

Any closed subgroup of a locally compact group is locally compact (). Conversely:

XVI.1.4.19. Theorem. Let \( G \) be a Hausdorff topological group, and \( H \) a subgroup which is locally compact in the relative topology. Then \( H \) is a closed subgroup of \( G \).

XVI.1.5. Polish Groups and Baire Groups

XVI.1.5.1. Definition. A Polish group is a topological group which is a Polish space (), i.e. homeomorphic to a separable complete metric space.

XVI.1.5.2. There is a large overlap between the class of locally compact groups and the class of Polish groups: the overlap consists precisely of the second countable locally compact groups (XVI.1.4.6.). The theory of Polish groups is to a large extent similar and parallel to the theory of locally compact groups. Except for one little thing: Polish groups which are not locally compact do not have Haar measure, so the most crucial tool for analysis is lacking. Thus the theory of Polish groups is not as satisfactory overall as that of locally compact groups.

There are many important topological groups which are Polish but not locally compact, for example:

XVI.1.5.3. Examples. (i) The additive group of any separable Banach space is a Polish group.
(ii) The multiplicative group of invertible elements in any separable Banach algebra is a Polish group in the relative norm topology (),().
(iii) The group of unitary operators on a separable Hilbert space is a Polish group under the strong or weak operator topologies (which coincide on this group; cf. ()).
(iv) The group of homeomorphisms of a compact metric space is a Polish group in the topology of uniform convergence ()
(v) The group of isometries of a separable complete metric space is a Polish group in the topology of pointwise convergence (which coincides with the topology of u.c. convergence; cf. ()).
(vi) Any closed subgroup of a Polish group is a Polish group.
(vii) A countable product of Polish groups is a Polish group.

XVI.1.5.4. There is one sense in which Polish groups are nicer than locally compact groups: there is a universal Polish group $\tilde{G}$, a Polish group with the property that every Polish group is topologically isomorphic to a closed subgroup of $\tilde{G}$. In fact, the group of homeomorphisms of the Hilbert cube with the topology of uniform convergence is such a universal group ([Usp86]; cf. [Kec95, 9.18]).

XVI.1.6. Open Mappings and Automatic Continuity

In this section, we collect together some miscellaneous results about when surjective homomorphisms of topological groups are automatically open mappings, and when group operations and group actions are automatically continuous.

We first give simple criteria for when homomorphisms are continuous or open:

XVI.1.6.1. Proposition. Let $G$ and $G'$ be topological groups, and $\phi : G \to G'$ a group-homomorphism. Then:

(i) If $\phi$ is continuous at one point, then it is continuous everywhere.
(ii) $\phi$ is an open mapping if and only if, for every open neighborhood $U$ of $e$, $\phi(U)$ has nonempty interior (i.e. contains a nonempty open set).

Proof: (i): Suppose $\phi$ is continuous at $g$, i.e. if $g_i \to g$, then $\phi(g_i) \to \phi(g)$. If $x_i, x \in G$ and $x_i \to x$, then $gx^{-1}x_i \to g$, so

$$\phi(g)\phi(x)^{-1}\phi(x_i) = \phi(gx^{-1}x_i) \to \phi(g)$$

and thus

$$\phi(x_i) = \phi(x)\phi(g)^{-1}\phi(g)\phi(x)^{-1}\phi(x_i) \to \phi(x)\phi(g)^{-1}\phi(g) = \phi(x).$$

The first main result is a general version of XVI.1.4.13.: 

XVI.1.6.2. Theorem. [Open Mapping Theorem for $\sigma$-Compact Groups] Let $G$ and $G'$ be topological groups, with $G'$ a Baire group, and $\phi$ a continuous group homomorphism from $G$ onto $G'$. If $G$ is $\sigma$-compact (so $G'$ is also $\sigma$-compact), then $\phi$ is an open mapping, and thus $G'$ is topologically isomorphic to $G/\ker(\phi)$.

Proof: The homomorphism $\phi$ factors topologically through $G/\ker(\phi)$, so we may assume without loss of generality that $\phi$ is a bijection. Let $G = \bigcup_{n=1}^{\infty} K_n$, with $K_n$ compact. Then $G' = \bigcup_{n=1}^{\infty} \phi(K_n)$, and $\phi(K_n)$ is compact and hence closed in $G'$ since $G'$ is Hausdorff. By the Baire Category Theorem, some $\phi(K_n)$ has nonempty interior $V$. The restriction of $\phi$ to $K_n$ is a homeomorphism, so $\phi^{-1}$ is continuous on $V$ since $V$ is open. Since $V$ is nonempty, $\phi^{-1}$ is continuous (XVI.1.6.1.(ii)), i.e. $\phi$ is an open map.
XVI.1.6.3. Note that no hypothesis on \( G \) beyond \( \sigma \)-compactness is needed, not even an assumption that \( G \) is Hausdorff. This proof is attributed to H. GLÖCKNER, who observed that the local compactness hypothesis on \( G \) in many references is not necessary.

XVI.1.6.4. Corollary. Let \( G \) be a countable group. Then the only Polish topological group topology on \( G \) is the discrete topology.

XVI.1.7. Exercises

XVI.1.7.1. Let \( G \) be the additive group of real numbers. Give \( G \) the Sorgenfrey topology (XI.7.8.10.). Show that the group operation (addition) is jointly continuous (i.e. continuous from \( G \times G \) to \( G \)), but that inversion (negation) is not continuous. Thus \((\mathbb{R},+)\) with this topology is not a topological group.

XVI.1.7.2. Let \( G \) be a topological group, and \( A \) a subset of \( G \). Show that the closure of \( A \) is contained in \( AV \) for any neighborhood \( V \) of \( e \), and in particular \( U^2 \) if \( U \) is a neighborhood of \( e \) in \( G \). [If \( g \in A \), then \( gV^{-1} \cap A \neq \emptyset \).]

XVI.1.7.3. (a) Use XVI.1.7.2. and XVI.1.2.14.(i) to give a direct proof that a topological group is \( T_3 \) (cf. XI.7.5.1.).
(b) Give a direct proof that a \( T_0 \) topological group is Hausdorff.

XVI.1.7.4. (a) Let \( G \) be a topological group and \( H \) a subgroup. If \( H \) contains a neighborhood of \( e \), show that \( H \) is open. Thus the hypothesis in XVI.1.2.31. that \( U \) be an open neighborhood can be relaxed to just requiring that \( U \) be a neighborhood.
(b) Show that a closed subgroup of a topological group is either open or nowhere dense.

XVI.1.7.5. Let \( G \) be a topological group, and \( U \) an open neighborhood of \( e \) in \( G \). Set \( W = \cup_{n=1}^{\infty} U^n \).
(a) Show that \( W \) is a clopen subset of \( G \).
(b) \( W \) need not be a subgroup of \( G \). [Consider \( \{0,1\} \subseteq \mathbb{Z} \) with the discrete topology.]
(c) If \( G \) is connected, then \( W = G \).

XVI.1.7.6. How much of the theory of topological groups requires joint continuity of multiplication? That is, if we change the definition of a topological group \( G \) to only require continuity of inversion and separate continuity of multiplication (i.e. that \( \iota \) and \( \lambda_g \), \( \rho_g \) are continuous for all \( g \in G \)), how much of the theory persists? [Consider the case where \( G \) is an infinite group with the finite complement topology.]

XVI.1.7.7. For a topological space \( X \) to have a group structure making it a topological group, i.e. for \( X \) to be homeomorphic to a topological group, it is necessary that \( X \) be homogeneous. But this is far from sufficient, even if \( X \) is compact and metrizable.
(a) The Hilbert cube \( \mathbb{H} \) is homogeneous (.), but not homeomorphic to a topological group. [Every continuous function from \( \mathbb{H} \) to \( \mathbb{H} \) has a fixed point (.)].
(b) \( S^n \) for even \( n > 0 \) is not homeomorphic to a topological group. [Use (.)]. In fact, \( S^n \) is homeomorphic to a topological group if and only if \( n \) is 0, 1, or 3 (.), (.)

1905
Groups of Homeomorphisms

If $X$ is a topological space, the group $\text{Homeo}(X)$ of homeomorphisms of $X$ can be given various topologies, e.g. the topology of pointwise convergence or the compact-open topology (which is the topology of u.c. convergence if $X$ is metrizable ())). $\text{Homeo}(X)$ is sometimes, but not always, a topological group under these topologies. The next several problems explore when $\text{Homeo}(X)$, or various natural subgroups, is or is not a topological group.

XVI.1.7.8. Let $X$ be a Hausdorff space. Give $\text{Homeo}(X)$ the compact-open topology. Show that multiplication is jointly continuous.

XVI.1.7.9. Let $X$ be a compact Hausdorff space. Show that $\text{Homeo}(X)$ is a topological group under the compact-open topology (which is the topology of uniform convergence).

XVI.1.7.10. [Are46] Let $X$ be the space $\mathbb{N}_0 \cup \{ \frac{1}{n} : n \in \mathbb{N} \}$, with its relative topology from $\mathbb{R}$. Then $X$ is a locally compact metrizable space. Show that inversion is not continuous in $\text{Homeo}(X)$. [Consider $\phi_n$, where $\phi_n(n) = \frac{1}{n}$, $\phi_n(m) = m - 1$ if $m > n$, $\phi_n(\frac{1}{m}) = \frac{1}{m+1}$ if $m \geq n$, and $\phi_n(x) = x$ otherwise.]

It is shown in [Are46] that if $X$ is locally compact and locally connected, then inversion in $\text{Homeo}(X)$ is continuous in the compact-open topology, and hence $\text{Homeo}(X)$ is a topological group in this topology.

XVI.1.7.11. Let $X$ be a topological space. Give $\text{Homeo}(X)$ the topology of pointwise convergence. Then multiplication is separately continuous.

XVI.1.7.12. Give $\text{Homeo}(\mathbb{R}^2)$ the topology of pointwise convergence. Show that multiplication is not jointly continuous, and that inversion is not continuous.

XVI.1.7.13. Let $X = \mathbb{R}$ or $[0,1]$. Show that the topology of pointwise convergence on $\text{Homeo}(X)$ coincides with the compact-open topology (topology of u.c. convergence), and that $\text{Homeo}(X)$ is a topological group under this topology.

XVI.1.7.14. Let $X$ be a metric space. Show that the group $\text{Isom}(X)$ of isometries (isometric bijections) of $X$ is a topological group in the topology of pointwise convergence.

XVI.1.7.15. Group Topologies on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. If $G$ is a group and $H$ a (Hausdorff) topological group, and $\phi : G \to H$ is an injective homomorphism, $G$ can be identified with $\phi(G)$ and given the relative topology, which makes $G$ into a (Hausdorff) topological group. In particular, if $H$ is a (Hausdorff) topological group and $x$ is an element of $H$ of infinite order, the map $n \mapsto x^n$ from $\mathbb{Z}$ to $H$ can be used to give a (Hausdorff) topological group topology on $\mathbb{Z}$. This gives an enormous variety of nonequivalent (XVI.1.4.17.) group topologies.

(a) Let $S$ be a nonempty subset of $\mathbb{R}$. If $G = \mathbb{Z}, \mathbb{Q}$, or $\mathbb{R}$, map $G$ into $\prod_{s \in S} T$ by sending $x \in G$ to the point $\phi_S(x)$ with $s$-coordinate $e^{2\pi i sx}$. Show that if $G = \mathbb{Z}$ or $\mathbb{Q}$, then $\phi_S$ is injective unless $S$ is contained in a cyclic subgroup of $(\mathbb{Q}, +)$, and if $G = \mathbb{R}$, then $\phi_S$ is injective unless $S$ is contained in a cyclic subgroup of $(\mathbb{R}, +)$. In these cases, give $G$ the relative topology $\mathcal{T}_S$. The topology $\mathcal{T}_S$ is metrizable (first countable) if and only if $S$ is countable.
(b) Let $B$ be a Hamel basis for $\mathbb{R}$ over $\mathbb{Q}$ consisting of irrational numbers. If $S_1, S_2$ are distinct nonempty subsets of $B$ (not singletons if $G = \mathbb{R}$), show that the topologies $T_{S_1}$ and $T_{S_2}$ are nonequivalent. Conclude that there are $2^{\aleph_0}$ mutually nonequivalent Hausdorff topologies on $G$ making $(G, +)$ a topological group. These topologies are all totally bounded, i.e. the completion is compact. These topologies are all weaker than the ordinary topology on $G$.

(c) The topologies $T_S$ are by no means all the Hausdorff topological group topologies on $(\mathbb{Q}, +)$: for example, the discrete topology, the relative topology from $\mathbb{R}$, and the $n$-adic topologies (III.5.3.9.) are not among them. (The completions with respect to these topologies are locally compact but noncompact.) But there are exactly $2^{\aleph_0}$ distinct topologies on $\mathbb{Q}$. The only locally compact or Polish topological group topology on $(\mathbb{Q}, +)$ is the discrete topology. Similar statements are true for $(\mathbb{Z}, +)$.

(d) If $G$ is any infinite abelian group of cardinality $\kappa$, then $G$ has $2^{2^\kappa}$ mutually nonequivalent Hausdorff topological group topologies [Pod77]. In particular, $(\mathbb{R}, +)$ has $2^{2^{\aleph_0}}$ mutually nonequivalent Hausdorff topological group topologies. Using that $(\mathbb{R}, +)$ is isomorphic to the direct sum of $2^{\aleph_0}$ copies of $(\mathbb{Q}, +)$, construct from (b) for $\mathbb{Q}$ a family of $2^{2^{\aleph_0}}$ mutually nonequivalent totally bounded Hausdorff topological group topologies on $(\mathbb{R}, +)$. The additive group of any separable Banach space is also isomorphic to $\mathbb{R}$.

XVI.2. Haar Measure

The most important tool for analysis on locally compact groups is Haar measure. Every locally compact group has an essentially unique translation-invariant measure called Haar measure. Haar measure generalizes Lebesgue measure on $\mathbb{R}$ or $\mathbb{R}^n$ and is of similar use and importance. And the fact that topological groups which are not locally compact do not have a similar translation-invariant measure is a serious obstacle to their study.

XVI.2.1. Definition and Existence

XVI.2.1.1. DEFINITION. Let $G$ be a Hausdorff topological group. A left Haar measure on $G$ is a measure $\mu$ defined on the Baire sets in $G$, with the following properties:

(i) $\mu(U) > 0$ for every nonempty open Baire set $U$ in $G$.

(ii) $\mu(K) < \infty$ for every compact $G$ in $G$.

(iii) $\mu$ is left invariant: $\mu(\lambda_g(A)) = \mu(A)$ for every $g \in G$ and every Baire set $A$ in $G$.

A right Haar measure on $G$ is a measure $\mu$ on the Baire sets in $G$ satisfying (i), (ii), and

(iii′) $\mu$ is right invariant: $\mu(\rho_g(A)) = \mu(A)$ for every $g \in G$ and every Baire set $A$ in $G$.

If $G$ is abelian, there is no difference between a left Haar measure and a right Haar measure on $G$. However, if $G$ is not abelian a left Haar measure need not be a right Haar measure, and conversely (XVI.2.2.6.(c)). If $G$ is second countable, the Baire sets are the same as the Borel sets in $G$, so (i) holds for all open sets and (ii) holds for all compact sets in $G$.

Conditions (i) and (ii) are nondegeneracy conditions: without (i) the zero measure would be a left and right Haar measure, and without (ii) the measure in which every nonempty set has infinite measure would be a left and right Haar measure.
XVI.2.1.2. Examples. (i) Lebesgue measure is a (left and right) Haar measure on \( \mathbb{R} \) or \( \mathbb{R}^n \) with addition.
(ii) A (left or right) Haar measure on \( \mathbb{T} \) can be described in several equivalent ways. One simple way is to identify it with \([0,1]\) by \( t \leftrightarrow e^{2\pi it} \) and transfer Lebesgue measure on \([0,1]\) to \( \mathbb{T} \). One could also take arc length measure on \( \mathbb{T} \) (or 1-dimensional Hausdorff measure) to obtain the same measure up to a factor of \( 2\pi \).
(iii) Counting measure is a (left or right) Haar measure on any discrete group.
(iv) Any positive scalar multiple of a left Haar measure on \( G \) is a left Haar measure, and similarly for right Haar measures.

The main theorem is:

XVI.2.1.3. Theorem. Let \( G \) be a locally compact group. Then there is a left Haar measure on \( G \), which is unique up to positive scalar multiple. There is also a right Haar measure on \( G \), unique up to a positive scalar multiple.

XVI.2.2. Haar Integrals

By the results of (), there is an entirely equivalent formulation of Haar measure as an integral or as a positive linear functional.

XVI.2.2.1. Definition. Let \( G \) be a locally compact group. A left Haar integral on \( G \) is a nonzero positive linear functional \( I \) on \( C_c(G) \) which is left invariant: \( I(g \cdot \phi) = I(\phi) \) for all \( g \in G, \phi \in C_c(G) \), where \((g \cdot \phi)(x) = \phi(gx)\), i.e. \( g \cdot \phi = \phi \circ \lambda_g \). A right Haar integral is a nonzero positive linear functional \( I \) on \( C_c(G) \) which is right invariant: \( I(\cdot g) = I(\cdot) \) for all \( g \in G, \phi \in C_c(G) \), where \((\phi \cdot g)(x) = \phi(xg)\), i.e. \( \phi \cdot g = \phi \circ \rho_g \).

XVI.2.2.2. By the Riesz Representation Theorem (), a positive linear functional on \( C_c(G) \) is given by integration with respect to a Baire measure. So if \( I \) is a left Haar integral on \( G \), then
\[
I(\phi) = \int_G \phi \, d\mu
\]
for a Baire measure \( \mu \).

XVI.2.2.3. Proposition. If \( I \) is a left Haar integral on \( G \), then \( \mu \) is a left Haar measure on \( G \), and similarly for a right Haar integral.

Conversely:

XVI.2.2.4. Proposition. If \( \mu \) is a left Haar measure on \( G \), then \( I \) defined by \( I(\phi) = \int_G \phi \, d\mu \) is a left Haar integral on \( G \), and similarly for a right Haar measure.

Thus Theorem XVI.2.1.3. is equivalent to:
XVI.2.2.5. Theorem. Let $G$ be a locally compact group. Then there is a left Haar integral on $G$, which is unique up to positive scalar multiple. There is also a right Haar integral on $G$, unique up to a positive scalar multiple.

The Haar integral is easier and more elementary to describe and find than the Haar measure in many cases:

XVI.2.2.6. Examples. (a) For $\mathbb{R}$ with addition, the Haar integral (left and right) is simply given by Riemann integration: if $\phi \in C_c(\mathbb{R})$, 
\[
I(\phi) = \int_{-\infty}^{+\infty} \phi(t) \, dt .
\]
This is not really an improper Riemann integral since $\phi$ has compact support: it is an ordinary Riemann (or even Cauchy) integral over a finite interval, which varies from function to function.  

(b) The (normalized) Haar integral on $\mathbb{T}$ is also given by Riemann integration: if $\phi \in C_c(\mathbb{T}) = C(\mathbb{T})$, then 
\[
I(\phi) = \int_{0}^{1} \phi(e^{2\pi it}) \, dt .
\]
(c) Here is perhaps the simplest example of a locally compact group for which left and right Haar measure (or integral) are different: the real “$ax + b$ group.” Let 
\[
G = \left\{ \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} : a, b \in \mathbb{R}, a > 0 \right\} .
\]
$G$ is a subgroup of $GL(2, \mathbb{R})$, i.e. is a group under matrix multiplication. $G$ can be identified as a set with the open right half-plane $\mathbb{R}_+ \times \mathbb{R}$ in $\mathbb{R}^2$ by 
\[
\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \leftrightarrow (a, b)
\]
and given the topology from this identification (which agrees with the subspace topology from $GL(2, \mathbb{R})$ regarded as an open subset of $M_2(\mathbb{R}) \cong \mathbb{R}^4$). From this point of view, the group operations are given by 
\[
(a, b)(x, y) = (ax, ay + b)
\]
\[
(a, b)^{-1} = \begin{bmatrix} 1/a & -b/a \\ a & 0 \end{bmatrix}
\]
(cf. ()); the identity is $(1, 0)$. Since these operations are continuous, $G$ is a locally compact group (it is even a Lie group). 

If $g = (a, b) \in G$, left multiplication by $g$ is an affine map $(x, y) \mapsto (ax, ay + b)$ which is differentiable with constant derivative $\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ and Jacobian $a^2$. 

Consider the linear functional $I$ on $G$ given for $\phi \in C_c(G)$ by 
\[
I(\phi) = \int_{G} \phi(x, y) \frac{dx \, dy}{x^2} = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \phi(x, y) \frac{dy \, dx}{x^2}
\]
1909
(since the support of \( \phi \) is a compact subset of the open right half-plane, the integral is only really over a sufficiently large closed rectangle on which \( \frac{1}{x^2} \) is bounded, hence is a proper Riemann integral). Then \( I \) is a nonzero positive linear functional on \( C_c(G) \). If \( g = (a, b) \in G \), then by the change-of-variables formula, writing \( u = ax, v = ay + b \), we have

\[
I(g \cdot \phi) = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(ax, ay + b)}{x^2} \, dy \, dx = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(ax, ay + b)}{a^2 x^2} a^2 \, dy \, dx
\]

\[
= \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(u, v)}{u^2} \, dv \, du = I(\phi)
\]

(the integrals are actually over suitably large compact regions in \( G \); cf. XVI.2.9.3.). Thus \( I \) is a left Haar integral on \( G \). The corresponding left Haar measure is given by

\[
\mu(A) = \int_{A} \frac{1}{x^2} \, d\lambda(x, y)
\]

for \( A \) a Borel set in \( G \), where \( \lambda \) is Lebesgue measure on \( \mathbb{R}^2 \).

Similarly, right multiplication by \( g = (a, b) \) is given by \((x, y) \mapsto (xa, xb + y)\), with derivative

\[
\begin{bmatrix}
    a & 0 \\
    b & 1
\end{bmatrix}
\]

and Jacobian \( a \). Thus the functional

\[
I_r(\phi) = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(x, y)}{x} \, dy \, dx
\]

is a right Haar integral on \( G \), with right Haar measure

\[
\mu(A) = \int_{A} \frac{1}{x} \, d\lambda(x, y)
\]

for \( A \) Borel. Thus left and right Haar measures on \( G \) are different (XVI.2.9.3.).

\( G \) is called the “\( ax + b \) group” since it is isomorphic to the group of all order-preserving invertible affine transformations of \( \mathbb{R} \), i.e. transformations of the form \( f(x) = ax + b, a > 0 \). There is a similar “\( ax + b \) group” for any vector space over any field, or even modules over rings; some of these have topologies making them interesting topological groups (see e.g. [Bla83]).

XVI.2.3. Invariant Measures for Group Actions

XVI.2.4. Properties of Haar Measure

XVI.2.5. The Modular Function

XVI.2.6. Measures on Homogeneous Spaces

XVI.2.7. Extending Haar Measure

XVI.2.8. Measurable Groups

XVI.2.9. Exercises
XVI.2.9.1. Let $G$ be a locally compact group.

(a) Show that there is a neighborhood of $e$ in $G$ which is a compact $G_δ$, which contains a neighborhood of $e$ which is an open Baire set. [Use that $G$ is completely regular.]

(b) Show that if $μ$ is a left Haar measure on $G$, then there is a nonempty open Baire set $V$ in $G$ with $μ(V) < ∞$.

(c) Conversely, if $μ$ is a left invariant Baire measure on $G$ for which there is a nonempty open Baire set $V$ with $μ(V) < ∞$, then $μ(K) < ∞$ for every compact $G_δ$ $K$. [$K$ can be covered by finitely many left translates of $V$.]

Thus condition (ii) in the definition of a left Haar measure can be replaced by

(ii’) There is a nonempty open Baire set $V$ in $G$ with $μ(V) < ∞$.

The same change can be made in the definition of a right Haar measure.

XVI.2.9.2. Here is a warmup for Example XVI.2.2.6.(c). Let $G = \mathbb{R}_+$ be the positive real numbers with multiplication.

(a) Show that the linear functional on $C_c(G)$ given for $ϕ \in C_c(G)$ by

$$I(ϕ) = \int_0^∞ \frac{ϕ(x)}{x} dx$$

is (left and right) invariant, hence is a Haar integral. Write down the corresponding Haar measure.

(b) Show the result of (a) alternately by noting that $x \mapsto e^x$ is a topological isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}_+, \cdot)$ and transferring the Haar integral on $\mathbb{R}$ to $\mathbb{R}_+$ using the change-of-variables formula.

XVI.2.9.3. (a) In example XVI.2.2.6.(c), give a more careful argument about regions of integration in the application of the change-of-variables formula in showing $I$ is left invariant.

(b) Show by example that $I$ is not right invariant, i.e. that the left and right Haar integrals for $G$ are truly different. (Caution: this involves more than just showing that $I \neq I_r$.)

XVI.2.9.4. (a) Let $G = GL(2, \mathbb{R})$. Identify $G$ with an open set in $\mathbb{R}^4$ by

$$\begin{pmatrix} x & y \\ z & w \end{pmatrix} \mapsto (x, y, z, w).$$

As in XVI.2.2.6.(c), show that the linear functional defined for $ϕ \in C_c(G)$ by

$$I(ϕ) = \int_G \frac{ϕ(x, y, z, w)}{(xw - yz)^2} d\lambda(x, y, z, w)$$

(the integral is actually a Riemann integral over a suitable compact region) is both a left and right Haar integral for $G$. Note that the denominator is the square of the determinant.

(b) Show similarly that if $G = GL(n, \mathbb{R})$, identified with an open set in $\mathbb{R}^{n^2}$, the left and right Haar integral is given by

$$I(ϕ) = \int_G \frac{ϕ(x_{ij})}{|\text{Det}(x_{ij})|^n} d\lambda(x_{ij})$$

where $(x_{ij})$ symbolically denotes the $n^2$-tupple $(x_{11}, \ldots, x_{nn})$, or the corresponding matrix. (Note that the case $n = 1$ is XVI.2.9.2.)
XVI.3. Lie Groups

A Lie group is a group which is a smooth manifold and whose multiplication and inversion are smooth maps. Since a smooth manifold has an underlying structure as a topological space and smooth functions are continuous, a Lie group has an underlying structure as a topological group (although it is technically not quite correct to say that a Lie group “is” a topological group), which is locally compact. Many of the most familiar and important topological groups are Lie groups. It turns out that every topological group whose topology is locally Euclidean has a unique structure as a Lie group. The theory of Lie groups bridges geometry, topology, linear and abstract algebra, and analysis, and impacts much of modern mathematics.

Lie groups are named in honor of Sophus Lie; note that “Lie” is pronounced “Lee.”

XVI.3.1. The Classical Groups

XVI.3.2. Lie Groups and Lie Algebras

XVI.3.3. Locally Euclidean Groups

XVI.3.3.1. Definition. A locally Euclidean group is a topological group whose topology is locally Euclidean (XI.19.1.1.).

XVI.3.3.2. A locally Euclidean space is $T_1$ (XI.19.1.2. (viii)), and thus a locally Euclidean group $G$ is Hausdorff, hence locally compact. By XI.19.1.2. (vi), XVI.1.2.34., and XVI.1.4.4., the connected component of the identity is an open normal subgroup of $G$ which is $\sigma$-compact. Thus, as a topological space, a locally Euclidean group is a topological manifold (and in particular metrizable; cf. XVI.1.3.7.).

XVI.3.3.3. Definition. Let $G$ be a topological group. Then $G$ has no small subgroups, or is an NSS group, if there is a neighborhood of the identity which contains no nontrivial subgroup of $G$.

An NSS group is Hausdorff, since $\{e\}$ is contained in every neighborhood of $e$ (Exercise XVI.1.7.2.). Any Lie group is an NSS group (). By XVI.1.4.10., a locally compact NSS group is metrizable.

XVI.3.3.4. Hilbert asked in his 1900 problem list (#5) whether every locally Euclidean group is a Lie group (i.e. has a Lie group structure; Hilbert did not state the problem quite so cleanly). This proved to be a difficult problem and was solved in 1952 by A. Gleason, D. Montgomery, and L. Zippin ([Gle52], [MZ52]); special cases had been established by von Neumann (compact groups [vN33]), Pontryagin (abelian groups []), and Chevalley (solvable groups [Che41]). The proof of the general theorem was simplified by H. Yamabe ([Yam53a], [Yam53b]) (although it is still deep and difficult). It is one of the remarkable results of mathematics, often described as “getting something for nothing.”

XVI.3.3.5. Theorem. Let $G$ be a locally compact group. The following are equivalent:

(i) $G$ is locally Euclidean.

(ii) $G$ is an NSS group.

(iii) $G$ has a unique structure as a Lie group.
The proof is far beyond the scope of this book. See e.g. [MZ74] or [Kap71].

“[This result] shows that you can’t have a little bit of orderliness without a lot of orderliness.”

A. Gleason

The implication (ii) ⇒ (i) of XVI.3.3.5. has a generalization due to Gleason and Yamabe:

**XVI.3.3.6. Theorem.** Let $G$ be a locally compact group. Then $G$ has an open subgroup $H$ such that, if $U$ is any neighborhood of $e$ in $G$, there is a compact normal subgroup $K$ of $H$, contained in $U$, such that $H/K$ is locally Euclidean (and thus has a unique structure as a Lie group).

The $K$ cannot be chosen normal in $G$ in general (). But since $H/K$ is a clopen subset of the homogeneous space $G/K$, it follows that $G/K$ is locally Euclidean as a topological space, hence a manifold.

XVI.4. Group Algebras and Unitary Representations

One of the most important applications of the existence of Haar measure on locally compact groups is that certain Banach algebras can be associated to any locally compact group which have the “same” representation theory as the group. These Banach algebras are a crucial tool in showing that locally compact groups have a nice representation theory, and in developing a theory of analysis on locally compact groups (abstract harmonic analysis).

XVI.4.1. The Group Algebra

**XVI.4.2. The $L^1$ Algebra and the C*-Algebra of a Locally Compact Group**

**Caution:** It is common in the literature on representation theory to say that a locally compact group $G$ is “separable” if $L^1(G)$ (or, equivalently, $C^*(G)$) is separable, i.e. if $G$ is second countable. However, this use of the term “separable” can be misleading, since a locally compact group, even a compact abelian group, which is separable in the topological sense of having a countable dense subset is not necessarily “separable” in this sense (a product of $2^\mathbb{N}$ circle groups or nontrivial finite groups is a counterexample; cf. XI.6.1.2.). Thus we will avoid using the term “separable” in any sense for topological groups.

---

1[Yan02, p. 144].
XVI.5. Locally Compact Abelian Groups and Pontryagin Duality

Locally compact abelian groups have special properties which make them an especially attractive class of objects to study, with important applications. There is a beautiful duality theory which leads to, among other things, a general theory of Fourier transform. Analysis on locally compact abelian groups, including the general Fourier transform, is called harmonic analysis.

XVI.5.1. The Dual Group

To each locally compact abelian group $G$ there is associated a dual group $\hat{G}$ which is also a locally compact abelian group. The major result, the Pontryagin Duality Theorem ($\ast$), will be that the dual group $\hat{\hat{G}}$ of $\hat{G}$ is topologically isomorphic to $G$, so there is a true duality between $G$ and $\hat{G}$. In this subsection, we define the dual group and examine some of its properties.

XVI.5.1.1. Definition. Let $G$ be a locally compact abelian group, and $\mathbb{T}$ the circle group ($\ast$). A character of $G$ is a continuous homomorphism $\chi : G \to \mathbb{T}$. If $\chi_1, \chi_2$ are characters, define the product $\chi_1 \chi_2$ by $(\chi_1 \chi_2)(g) = \chi_1(g)\chi_2(g)$ for $g \in G$, where the multiplication is in $\mathbb{T}$ (i.e. the pointwise product as complex-valued functions). Let $\hat{G}$ be the set of all characters of $G$, with the topology of uniform convergence on compact subsets of $G$. $\hat{G}$ is called the dual group, or Pontryagin dual group, of $G$.

The terminology is justified by the following result:

XVI.5.1.2. Theorem. Let $G$ be a locally compact abelian group. Then $\hat{G}$ is a locally compact abelian group under the product and topology of XVI.5.1.1.

Proof: First note that $\hat{G}$ is an abelian group under the product: the product is obviously associative and commutative, the constant function 1 is an identity, and if $\chi \in \hat{G}$, then $\chi^{-1} \in \hat{G}$ defined by $\chi^{-1}(g) = \chi(g)^{-1} = \overline{\chi(g)}$ is an inverse for $\chi$.

To show $\hat{G}$ is a topological group,

It remains to show that $\hat{G}$ is locally compact. 

XVI.5.1.3. Examples. (i) Let $G = \mathbb{Z}$ with the discrete topology. Any homomorphism $\chi$ from $\mathbb{Z}$ to $\mathbb{T}$ is continuous, and completely determined by $\chi(1)$, which can be any element of $\mathbb{T}$; thus $\hat{\mathbb{Z}}$ can be identified as a set with $\mathbb{T}$. The product of two characters is just the product in $\mathbb{T}$ under this identification, so $\hat{\mathbb{Z}} \cong \mathbb{T}$ as a group. The topology on $\hat{\mathbb{Z}}$ is the topology of pointwise convergence since $\mathbb{Z}$ has the discrete topology. If $(\chi_j)$ is a net in $\hat{\mathbb{Z}}$, then $\chi_j \to \chi$ pointwise if and only if $\chi_j(1) \to \chi(1)$. Thus the topology on $\hat{\mathbb{Z}}$ is just the usual topology on $\mathbb{T}$ under the identification, i.e. the identification $\chi \mapsto \chi(1)$ is a topological isomorphism from $\hat{\mathbb{Z}}$ onto $\mathbb{T}$.

(ii) Let $G = \mathbb{R}$ with the usual topology. If $\chi$ is a continuous homomorphism from $\mathbb{R}$ to $\mathbb{T}$, then $\chi$ lifts to a continuous function $\tilde{\chi} : \mathbb{R} \to \mathbb{R}$, and it is easily checked that $\tilde{\chi}$ is a homomorphism. But the only continuous homomorphisms from $\mathbb{R}$ to $\mathbb{R}$ are multiplication by a constant: if $\phi : \mathbb{R} \to \mathbb{R}$ is a continuous homomorphism, then $\phi$ is uniquely determined on $\mathbb{Q}$ by $\phi(1) = \lambda$, and by continuity we must have $\phi(x) = \lambda x$ for all $x \in \mathbb{R}$. Thus, if $\tilde{\chi}(x) = \lambda x$ for fixed $\lambda \in \mathbb{R}$, we must have $\chi(x) = e^{2\pi i \lambda x}$, and we can identify $\chi$ with $\lambda$, i.e. call this character $\chi_\lambda$. The map $\lambda \mapsto \chi_\lambda$ is a group-isomorphism from $\mathbb{R}$ onto $\hat{\mathbb{R}}$. 1914
To show that this is a topological isomorphism, suppose $\chi_\lambda \to \chi_\lambda$ uniformly on compact sets. Then in particular

(iii) Let $G = T$. Suppose $\chi$ is a continuous homomorphism from $T$ to $T$. Compose $\chi$ with the homomorphism $\phi : \mathbb{R} \to T$ with $\phi(x) = e^{2\pi i x}$ to get a character $\tilde{\chi}$ of $\mathbb{R}$, which must be of the form $\tilde{\chi}(x) = e^{2\pi i \lambda x}$ for some $\lambda \in \mathbb{R}$, so we have $\chi(e^{2\pi i x}) = e^{2\pi i \lambda x}$. This function is well defined only if $\lambda \in \mathbb{Z}$. So a continuous homomorphism from $T$ to $T$ can only be of the form $\chi_n(x) = \xi^n$ for some fixed $n \in \mathbb{Z}$, so $T \cong \mathbb{Z}$ as groups. Since $T$ is compact, the topology on $\hat{T}$ is the topology of uniform convergence. But no $\chi_n$ for $n \neq 0$ can be uniformly close to the constant function 1, so the topology on $\hat{T}$ is the discrete topology, i.e. $\hat{T} \cong \mathbb{Z}$ with the discrete topology.

(iv) Let $G = \mathbb{Z}_n$ with the discrete topology. Then any $\chi \in \hat{G}$ is again completely determined by $\chi(1)$. But since 1 has order $n$ in $G$, $\chi(1)$ must have order $n$ in $T$, i.e. $\chi(1)$ must be an $n$’th root of unity. Any $n$’th root of unity defines a character, and the characters multiply like the roots of unity, so we have $\hat{\mathbb{Z}_n} \cong \mathbb{Z}_n$, and the topology on $\hat{\mathbb{Z}}_n$ is obviously also the discrete topology. But there is no natural isomorphism between $\hat{\mathbb{Z}}_n$ and $\mathbb{Z}_n$: any isomorphism requires making a choice of generators for the groups.

XVI.5.1.4. The characters of a locally compact abelian group $G$ are just the one-dimensional unitary representations () of $G$. Every one-dimensional representation is obviously irreducible, and every irreducible unitary representation of an abelian group is one-dimensional (), so $\hat{G}$ is just the set of irreducible representations of $G$. Two one-dimensional representations are equivalent () if and only if they are identical as homomorphisms to $T$, so $\hat{G}$ is just the set of equivalence classes of irreducible representations of $G$, i.e. the notation is consistent with the notation of () . But if $G$ is abelian, $\hat{G}$ has the additional natural structure of a (topological) group; $\hat{G}$ for nonabelian $G$ has no such structure.

XVI.5.1.5. By (), every unitary representation of $G$ defines representations of $L^1(G)$ and $C^*(G)$, and conversely. Thus, if $G$ is locally compact abelian, each character of $G$ defines a (bounded) *-homomorphism from $L^1(G)$ or $C^*(G)$ to $\mathbb{C}$, and these are precisely all the irreducible representations of these group algebras. Thus $\hat{G} \cong L^1(G) = C^*(G)$. The topology on $\hat{G}$ is precisely the hull-kernel topology on $C^*(G) \cong \text{Prim}(C^*(G))$.

Since irreducible representations are separating (), we obtain:

XVI.5.1.6. Proposition. Let $G$ be a locally compact abelian group, and $g \in G$, $g \neq e$. Then there is a character $\chi$ of $G$ with $\chi(g) \neq 1$.

XVI.5.1.7. Definition. If $G$ is a locally compact abelian group and $H$ is a closed subgroup, the annihilator of $H$ in $\hat{G}$ is

$$H^\perp = \{ \chi \in \hat{G} : \chi(h) = 1 \text{ for all } h \in H \} .$$

An element of $H^\perp$ defines a character of $G/H$ in the obvious way. Conversely, every character of $G/H$ can be regarded as a character of $G$ whose kernel contains $H$, i.e. an element of $H^\perp$.

1915
XVI.5.1.8. **Proposition.** Let $G$ be a locally compact group and $H$ a closed subgroup. Then the above correspondence is a topological isomorphism between $H^\perp$ and $G/H$.

XVI.5.1.9. **Theorem.** Let $G$ be a locally compact abelian group. Then

(i) $\hat{G}$ is second countable if and only if $G$ is second countable.

(ii) $\hat{G}$ is compact [resp. discrete] if and only if $G$ is discrete [resp. compact].

If $G$ is discrete (so $\hat{G}$ is compact), there are further topological properties of $\hat{G}$ which reflect algebraic properties of $G$:

XVI.5.1.10. **Proposition.** Let $G$ be a discrete abelian group, and $\hat{G}$ the compact dual group. Then

(i) $\hat{G}$ is connected if and only if $G$ is torsion-free.

(ii) $\hat{G}$ is totally disconnected (zero-dimensional) if and only if $G$ is a torsion group.

XVI.5.2. **Exercises**

XVI.5.2.1. (a) Let $G$ be an abelian group with the discrete topology. Show that the topological dimension (covering dimension $dim$) of the compact group $\hat{G}$ equals the rank () of $G$.

(b) A torsion-free abelian group of finite rank is countable. Hence a connected compact abelian group of finite topological dimension is metrizable.

(c) If $H$ is a connected compact abelian group of topological dimension 1, then $H$ is either a circle or a solenoid (). [Apply (a) to $\hat{H}$.]

(d) The rank 2 torsion-free abelian groups, i.e. the subgroups of $\mathbb{Q}^2$ containing $\mathbb{Z}^2$, are “unclassifiable” (in a technically precise sense, cf. [Tho03]). Thus the connected compact abelian groups of topological dimension 2 (all of which are metrizable) are “unclassifiable” in the same sense. In particular, they can be far more complicated than products of two circles and/or solenoids. The classification problem for rank $n$ becomes strictly more complex as $n$ increases.
Chapter XVII

Addenda
XVII.1. Mathematical Names

One problem not often discussed, which every student and mathematician faces, is learning how to pronounce names of mathematicians when citing or describing their work. I know I have struggled with this on occasion, beginning with my student days, and judging from mispronunciations I have heard from other mathematicians I am hardly alone. Thus I have included this guide as a public service. The information here is correct to the best of my knowledge, but I make no guarantees, and I welcome corrections. I have generally omitted names whose pronunciation should be obvious to English speakers. There are a number of websites such as http://pronouncemath.blogspot.com which may be helpful, but none should be regarded as authoritative.

Since neither I nor most potential readers are professional linguists, I have written the pronunciations using English syllables instead of linguistic symbols. The pronunciations are only approximations, and reflect commonly accepted pronunciations by English-speaking mathematicians rather than exact pronunciations in the original language. Of course, languages vary significantly, particularly in the pronunciation of vowels and r’s, and also l’s, n’s, and other consonants; stress patterns also vary. It is probably better for a mathematician speaking in English to use an English pronunciation instead of an exact original pronunciation anyway.

The English syllables and character combinations are slightly ambiguous in places, but should be pronounced in the usual English way; thus “ch” is always pronounced as in “chair”. One sound not present in English is the German ch or Russian x, pronounced like an English “hi” but farther back in the throat. This sound is written “kh”.

The stressed syllable is written in capital letters.

Spellings of names not originally written in the Latin alphabet are also an issue. This arises most critically with Russian names, but can also occur with Arabic, Chinese, or other names. (It is potentially an issue with Indian and Japanese names, but these seem to have standard and universally-used transliterations into the Latin alphabet.)

Russian names are not always transliterated consistently, and the standard English, French, and German transliteration schemes are rather different; and since much mathematical literature has historically been written in German or French, French- and German-style transliterations have come into common use even in English texts. Thus, for example, the name of the well-known Russian mathematician Тихонов, who originally published his papers in German, is frequently transliterated Tychonoff instead of the proper English transliteration TIKHONOV (this difference is exacerbated by the fact that a final consonant like n in Russian is normally pronounced unvoiced). An initial E in Russian is usually pronounced “yeh”, so should be properly transliterated Ye in this case (the name of the former Russian president Ельциным is transliterated YELTSYN), but Eropo is usually transliterated EGOROV or EGOROFF, not YEGOROV. We have given common spelling variations in the list. Another difficulty is that some Russian mathematicians, especially Jewish ones, have names which were originally German which were transliterated into Russian; there is a question whether these should be retransliterated into English or restored to their original German spellings. For example, Бернштейн (Sergei) is usually written Bernstein, and Урьсон is usually URSOHN. Наимарк is sometimes spelled NAIMARK and sometimes NEUMARK. Yet another gray area concerns Russian mathematicians who emigrate and begin spelling their own names a certain way in the Latin alphabet. The prime example for this book is BESICOVITCH (his spelling). We will regard this example as a voluntary name change and use his spelling.

One last note of caution, more relevant for contemporary names than historical ones: many Americans (and Canadians) are of European or Asian ancestry, and their names are not always pronounced or spelled in the ancestral way (this is even true of BLACKADAR, which was originally BLACKADDER, although not of BBC fame!) There are particularly many German-Americans, Italian-Americans, Polish-Americans, and Chinese-Americans, and Jewish Americans whose names are of German, Polish, or Russian origin. (There are also a
great many Americans with Hispanic names, but unfortunately this group is seriously underrepresented in
the world of mathematics.)

Here are some general pronunciation conventions for some common languages:

**German:** a is normally pronounced “ah”, never “eh” unless it has an umlaut (ä). Other vowels which can
have umlauts are ö and ü, which are hard to pronounce for an English speaker; ö sounds something like “er”,
and ü is something like a deep “ee” with rounded lips. The vowels ä, ö, ü are sometimes written ae, oe, ue
respectively.

Double vowels are just pronounced as longer versions of the corresponding single vowel: aa is pronounced
“ah”, ee is pronounced roughly like the name of the English letter a, and oo is pronounced “oh”, not English
“oo”. (The combination eh is pronounced the same as ee, leading to pun jokes among German students
concerning the word root lehr, “teach”, and leer, “empty”. ah and oh are similarly pronounced like aa and
oo respectively.) A final e is always pronounced, normally as “uh” and not “ee”.

Diphthongs are pronounced as follows: au as “ow”, not “aw”; eu as “oy”, not “oo” or “you”; ie as the
name of the English letter i; ie as “ee”.

The letter v is usually pronounced “f”; the letter w is usually pronounced roughly like English v, but
a little softer with rounded lips. z is pronounced “ts”. The combination th is pronounced “t” (the sound
“th”, voiced or unvoiced, does not occur in German). These letters also sometimes occur consecutively at
the break between parts of a compound word, in which case they are pronounced separately; BEETHOVEN
is an example (although this name is of Flemish origin). st is often, but not always, pronounced separately; the sound “sh” is spelled sch. Thus Monatsheft is pronounced “MOE-nahts-heft”, not “MOE-naht-sheft”. The combination sh usually
only occurs, as here, in compound words where the first part ends in s and the second part begins with h.
This applies to names such as Pringsheim.

**Russian:** a (а) is pronounced “ah” after a consonant, some trace of the “y” usually remains, but it mainly disappears into the preceding consonant (which is technically pronounced soft). There is another letter e (e) which is pronounced without the “y”, but this letter mostly appears in foreign words and rarely if ever in Russian names. Sometimes in a stressed position e is pronounced “yaw” (e pronounced this way should be written è, but the “umlaut” is typically omitted). For example, the names of the former Soviet leaders Хрущев and Горбачев are properly pronounced “khroosch-OFF” and “gar-ba-CHOFF” respectively; the most notable instance among mathematical names is in Chebyshев. When this occurs in the list, it is noted in the pronunciation guide. The letter o (о) is pronounced “ow”, or “uh” (sometimes “ah”) in an unstressed position. The letter u (у) is pronounced “oo”, not “uh”; the soft form io is pronounced “you”. The letter i (и) is pronounced “ee”. There
is another vowel i, the hard version of и, transliterated y, which is pronounced similarly to the German ü
but farther back in the throat. The vowel я, “ya”, is the soft form of a.

There are no diphthongs in Russian: consecutive vowels are pronounced separately. For example,
Воеводский is pronounced “vaw-yeh-VAWD-skee”, and the chemist Менделеев is pronounced “men-del-
YAY-yeff”.

Most Russian consonants have a hard and soft form. The hard form is the “default”, and consonants
become soft only when followed by a soft vowel (я, е, и, ю) or a soft sign (е), which is not pronounced. The
soft sign is sometimes transliterated with an apostrophe (as in OL’SHANSKII), but we have avoided using
this. There is also a “hard sign” or “separative sign” ь, which occurs only rarely in modern Russian and
which now usually has the same pronunciation effect as ь. The letter ь, which occurs only after vowels, is
pronounced with a slight “y” sound and usually transliterated i.

1919
As mentioned earlier, some consonants at the end of words, or before other unvoiced consonants, are pronounced unvoiced.

**Polish:** Polish names are frequently mispronounced badly in English, and some have acquired widely accepted English pronunciations quite different from the Polish ones. This is mainly due to the fact that the Polish spelling conventions for many sounds are different than the English ones. As a result, some Polish names look “unpronounceable” to English speakers not familiar with the Polish conventions. But a little familiarity makes them relatively easy to pronounce, at least approximately.

In Polish, \( w \) is pronounced “v”; \( ł \) is pronounced “w”, \( c \) is pronounced “ts” (except when followed by \( i \) or \( z \)), and \( j \) is pronounced “y”.

The combination \( sz \) is pronounced “sh”, \( cz \) is pronounced “ch”, and \( rz \) is pronounced “zh”. These are hard sounds. The accented letters \( ś \), \( ć \), \( ź \) (si, ci, zi before vowels) are soft forms of \( sz \), \( cz \), and \( rz \) respectively; \( ż \) is pronounced like \( rz \).

The vowel \( ó \) is pronounced “oo”, the same as \( u \). There are also nasal vowels and a nasal \( n \) (\( ň \)), but these are perhaps too arcane to worry about here.

For example, the common first name *Jerzy* is pronounced “YEH-zhee”, not “Jersey”.

In Polish words, including names, the stress is almost always on the next-to-last syllable.

Some Poles (mainly Jews, but not exclusively) have names which are originally German but which are written with Polish spelling conventions. Notable examples are the names *Szulc* and *Sznajder*, which are the same (and pronounced the same) as the German *Schulz* and *Schneider* respectively. Among mathematicians, the names *Aronszajn* and *Szpilrajn* come to mind.

**Hungarian:** Hungarian names also have the reputation of being hard for English speakers to pronounce, and the reason is again differences in spelling conventions. The Hungarian spelling conventions are in some instances the direct opposites of the Polish ones.

The vowels \( a \), \( e \), \( i \), \( o \), \( ö \), \( u \), \( ü \) are pronounced similarly to the corresponding German vowels. An accent denotes a long vowel; \( ť \) and \( ţ \) are the long forms of \( ŏ \) and \( ŭ \) respectively.

Hungarian consonants and combinations are as follows: \( c \) is pronounced “ts”, \( cs \) is “ch”, \( gy \) is a soft “j” (actually similar to the British pronunciation of “due”), \( j \) is like “y”; \( ly \), \( ny \), \( ty \) are soft versions of \( l \), \( n \), \( t \); \( s \) is pronounced “sh” and \( sz \) is pronounced “s”; and \( zs \) is “zh” (hence \( dzs \) is a hard “j”).

The stress in Hungarian words is on the first syllable. Note that an accent on a vowel does *not* indicate stress.

**French:** French names truly are hard for many English speakers to pronounce correctly, although most English speakers know how to make a decent approximation since they are relatively familiar.

**Italian:**

**Japanese:** The syllables of Japanese words are all pronounced with roughly the same stress. However, some syllables are partially swallowed. For example, the name of the city *Fukushima*, the site of the recent nuclear disasters, is pronounced something like “FKOO-shi-muh” (at least it sounds like this to me).

**Chinese:** I will not attempt to describe the proper pronunciation of Chinese names, since it is far beyond my expertise. However, I will recall a conversation I had with a Chinese mathematician, about the Chinese name *Ng* (which is usually pronounced “ing” in English, which is at least better than “nig”). I asked the proper pronunciation, and he replied, “Mm?” I thought he had not understood what I said, so I repeated the question, and he again replied, “Mm?”

Of course, there are many other languages, each with its own peculiarities. For example, consider the name of the eminent, and recently tragically deceased, Danish mathematician *Uffe Haagerup*. Danes pronounce
his last name something like “OR-up”, but somewhat less distinctly. English-speaking mathematicians usually say “HOG-er-up”, which Uffe (“OOF-uh”) didn’t seem to mind too much.

I must say I feel fortunate to be a native English speaker, not just because English has become the international language of mathematics (and many other subjects), but also because I am very happy I didn’t have to learn the language as an adult. It seems to me like a difficult and irregular language, particularly because of the use of articles (very hard for nonnative speakers to get right) and the tenuous relationship between spelling and pronunciation for many words, and I admire the ability of so many of my foreign colleagues to master it quite well. Although it is sometimes said that “the international language of mathematics is bad English,” I find most of it surprisingly good.

“English is weird. It can be understood through tough, thorough thought though.”

Unknown

We could say the same thing about analysis!
<table>
<thead>
<tr>
<th>Name</th>
<th>Pronunciation</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abel</td>
<td>AH-bull</td>
<td>abelian pronounced “uh-BEEL-yan”</td>
</tr>
<tr>
<td>Alaoglu</td>
<td>AL-uh-loo</td>
<td>Approximate; “g” very soft (Turkish)</td>
</tr>
<tr>
<td>Aleksandrov</td>
<td>ah-leks-AHN-druff</td>
<td></td>
</tr>
<tr>
<td>Alexandrov</td>
<td>See Aleksandrov</td>
<td></td>
</tr>
<tr>
<td>Aronszajn</td>
<td>AR-un-shine</td>
<td></td>
</tr>
<tr>
<td>Arzela</td>
<td>ahr-ZAY-la</td>
<td></td>
</tr>
<tr>
<td>Ascoli</td>
<td>as-COAL-ee</td>
<td>s is unvoiced</td>
</tr>
<tr>
<td>Baire</td>
<td>BEAR</td>
<td></td>
</tr>
<tr>
<td>Banach</td>
<td>BAHN-akh</td>
<td>Often pronounced BAHN-ock</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>bear-NEW-lee</td>
<td></td>
</tr>
<tr>
<td>Bernstein</td>
<td>BEARN-shtine</td>
<td></td>
</tr>
<tr>
<td>Blackadar</td>
<td>BLACK-uh-dahr</td>
<td></td>
</tr>
<tr>
<td>Bolzano</td>
<td>bol-TSAHN-o</td>
<td>Italian name, but Bohemian</td>
</tr>
<tr>
<td>Borel</td>
<td>burr-ELL</td>
<td></td>
</tr>
<tr>
<td>Bourbaki</td>
<td>bour-bah-KEE</td>
<td></td>
</tr>
<tr>
<td>Bożejko</td>
<td>bo-ZHAY-ko</td>
<td></td>
</tr>
<tr>
<td>Brouwer</td>
<td>BROW-er</td>
<td></td>
</tr>
<tr>
<td>Carathéodory</td>
<td>car-uh-thay-uh-DOOR-ee</td>
<td>“th” is unvoiced</td>
</tr>
<tr>
<td>Cartan</td>
<td>car-TAHN</td>
<td>n is nasal</td>
</tr>
<tr>
<td>Cauchy</td>
<td>KOH-shee</td>
<td></td>
</tr>
<tr>
<td>Cavalieri</td>
<td>cav-ul-YAIR-ee</td>
<td></td>
</tr>
<tr>
<td>Čech</td>
<td>CHEKH</td>
<td>As in “Czech”</td>
</tr>
<tr>
<td>Cesáro</td>
<td>chuh-SAH-row</td>
<td>Italian</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>cheb-uh-SHOFF</td>
<td>often cheb-uh-SHEFF</td>
</tr>
<tr>
<td>d’Alembert</td>
<td>dahl-em-BEAR</td>
<td></td>
</tr>
<tr>
<td>Daniell</td>
<td>dan-YELL</td>
<td></td>
</tr>
<tr>
<td>Darboux</td>
<td>dahr-BOO</td>
<td></td>
</tr>
<tr>
<td>Dedekind</td>
<td>DAY-duh-kind</td>
<td>“kind” as in “kin”</td>
</tr>
<tr>
<td>DeMoivre</td>
<td>duh-MWAVR</td>
<td></td>
</tr>
<tr>
<td>Denjoy</td>
<td>don-ZHWAH</td>
<td>n is nasal</td>
</tr>
<tr>
<td>Desargues</td>
<td>days-ARG</td>
<td></td>
</tr>
<tr>
<td>Descartes</td>
<td>day-CART</td>
<td></td>
</tr>
<tr>
<td>Dirac</td>
<td>durr-ACK</td>
<td></td>
</tr>
<tr>
<td>Dirichlet</td>
<td>DEER-ikh-lay</td>
<td>often DEER-ish-lay. See note</td>
</tr>
<tr>
<td>Egorov</td>
<td>ye-GOR-uff</td>
<td>Not EGG-o-ff!</td>
</tr>
<tr>
<td>Erdős</td>
<td>AIR-dosh</td>
<td>AIR-dush is more correct</td>
</tr>
<tr>
<td>Euclid</td>
<td>YOU-klid</td>
<td></td>
</tr>
<tr>
<td>Euler</td>
<td>OY-ler</td>
<td></td>
</tr>
<tr>
<td>Fatou</td>
<td>fat-TOO</td>
<td></td>
</tr>
<tr>
<td>Fejér</td>
<td>FAY-er</td>
<td>FAY-air is closer to correct</td>
</tr>
<tr>
<td>Fermat</td>
<td>fair-MAH</td>
<td></td>
</tr>
<tr>
<td>Fibonacci</td>
<td>fee-bun-AH-kee</td>
<td>Often fib-un-AH-kee</td>
</tr>
<tr>
<td>Fourier</td>
<td>FOUR-yay</td>
<td></td>
</tr>
<tr>
<td>Fraenkel</td>
<td>FRENK-ul</td>
<td></td>
</tr>
<tr>
<td>Fréchet</td>
<td>fresh-A</td>
<td>A as the name of the letter a</td>
</tr>
<tr>
<td>Frege</td>
<td>FRAY-guh</td>
<td></td>
</tr>
<tr>
<td>Frobenius</td>
<td>fro-BEAN-ee-us</td>
<td></td>
</tr>
<tr>
<td>Fubini</td>
<td>foo-BEE-nee</td>
<td>1922</td>
</tr>
<tr>
<td>Name</td>
<td>Pronunciation</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>Galois</td>
<td>gal-WAH</td>
<td>Hermitian pronounced her-MEE-shun</td>
</tr>
<tr>
<td>Gauss</td>
<td>GOWSS</td>
<td></td>
</tr>
<tr>
<td>Goursat</td>
<td>gor-SAH</td>
<td></td>
</tr>
<tr>
<td>Haar</td>
<td>HAHR</td>
<td></td>
</tr>
<tr>
<td>Hadamard</td>
<td>ad-uh-MAHR</td>
<td></td>
</tr>
<tr>
<td>Hausdorff</td>
<td>HOUSE-dorf</td>
<td></td>
</tr>
<tr>
<td>Heine</td>
<td>HIGH-nuh</td>
<td></td>
</tr>
<tr>
<td>Hermite</td>
<td>air-MEET</td>
<td></td>
</tr>
<tr>
<td>Hilbert</td>
<td>HILL-burt</td>
<td></td>
</tr>
<tr>
<td>Hölder</td>
<td>HEL-dur</td>
<td>Approximate; sometimes HULL-der</td>
</tr>
<tr>
<td>Jacobi</td>
<td>ya-KOE-bee</td>
<td>Actually German, of Jewish ancestry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sometimes ja-KOE-bee; Jacobian pronounced ja-KOE-bee-un</td>
</tr>
<tr>
<td>Jordan</td>
<td>zhor-DAHN</td>
<td>N is nasal</td>
</tr>
<tr>
<td>Khinchin</td>
<td>KHEEN-cheen</td>
<td>Often KHIN-chin</td>
</tr>
<tr>
<td>Kolmogorov</td>
<td>kul-muh-GOR-uff</td>
<td></td>
</tr>
<tr>
<td>Kovalevskaya</td>
<td>ko-va-LYEV-skah-ya</td>
<td></td>
</tr>
<tr>
<td>Krein</td>
<td>KRINE</td>
<td></td>
</tr>
<tr>
<td>Kronecker</td>
<td>KRON-eck-ur</td>
<td></td>
</tr>
<tr>
<td>Kuratowski</td>
<td>kur-a-TOV-skee</td>
<td></td>
</tr>
<tr>
<td>Lagrange</td>
<td>la-GRAHNGE</td>
<td>Not la-GREHNGE!</td>
</tr>
<tr>
<td>Laplace</td>
<td>la-PLAHS</td>
<td>Final “s” unvoiced</td>
</tr>
<tr>
<td>Laurent</td>
<td>lor-ON</td>
<td>N is nasal</td>
</tr>
<tr>
<td>Lebesgue</td>
<td>le-BEG</td>
<td>Second syllable is long</td>
</tr>
<tr>
<td>Legendre</td>
<td>le-ZHAHN-der</td>
<td></td>
</tr>
<tr>
<td>Leibniz</td>
<td>LIBE-nits</td>
<td></td>
</tr>
<tr>
<td>Levi-Civita</td>
<td>LEH-vee-CHIV-ee-tah</td>
<td></td>
</tr>
<tr>
<td>Lévy</td>
<td>LAY-vee</td>
<td></td>
</tr>
<tr>
<td>l’Hôpital</td>
<td>low pee-CAHL</td>
<td>Often pronounced LOW-pit-ahl</td>
</tr>
<tr>
<td>l’Hospital</td>
<td>See l’Hôpital</td>
<td></td>
</tr>
<tr>
<td>Lie</td>
<td>LEE</td>
<td></td>
</tr>
<tr>
<td>Lindelöf</td>
<td>LIN-del-off</td>
<td>Finnish name</td>
</tr>
<tr>
<td>Lindemann</td>
<td>LIN-duh-mahn</td>
<td></td>
</tr>
<tr>
<td>Liouville</td>
<td>LEE-oo-vill</td>
<td></td>
</tr>
<tr>
<td>Lipschitz</td>
<td>LIP-shits</td>
<td>No delicate way to write it!</td>
</tr>
<tr>
<td>Łoś</td>
<td>WAUSH</td>
<td></td>
</tr>
<tr>
<td>Lusin</td>
<td>See Luzin</td>
<td>French transliteration?</td>
</tr>
<tr>
<td>Luzin</td>
<td>LOO-zeen</td>
<td>Often LOO-zin</td>
</tr>
<tr>
<td>Marczewski</td>
<td>mar-CHEV-skee</td>
<td>Same person as Szpilrajn</td>
</tr>
<tr>
<td>Markov</td>
<td>MARK-uff</td>
<td></td>
</tr>
<tr>
<td>Milman</td>
<td>MIL-mun</td>
<td></td>
</tr>
<tr>
<td>Minkowski</td>
<td>min-KOV-skee</td>
<td></td>
</tr>
<tr>
<td>Mycielski</td>
<td>mi-CHEL-skee</td>
<td></td>
</tr>
<tr>
<td>Nagy</td>
<td>See Sz.-Nagy</td>
<td></td>
</tr>
<tr>
<td>Neumann</td>
<td>NOY-mahn</td>
<td></td>
</tr>
<tr>
<td>Nevanlinna</td>
<td>NEV-ann-lin-uh</td>
<td></td>
</tr>
<tr>
<td>Nikodym</td>
<td>NICK-uh-deem</td>
<td>Polish mathematician, Czech name</td>
</tr>
<tr>
<td>Noether</td>
<td>NER-ter</td>
<td>Approximate; first “r” not really there</td>
</tr>
<tr>
<td>Ostrogradskii</td>
<td>ah-strah-GRAHD-skee</td>
<td></td>
</tr>
</tbody>
</table>

1923
Pascal pass-KAL
Peano pay-AH-no not “Piano”!
Perron pair-RONE See note
Picard pec-CAR
Plancherel plain-sher-EL n is nasal
Poincaré pwan-car-AY
Poisson pwah-SAWN N is nasal
Pontryagin pon-tree-AH-geen g is hard (actually three syllables pon-TRYAH-geen)
Pringsheim PRINGZ-hime “hime” as in “time”
Prüfer PREE-fer approximate; often mispronounced PROOF-er
Pythagoras pith-AG-or-us
Raabe RAH-buh
Radon RA-don Not RAY-don
Ramanujan ra-MAHN-oo-jun
Riemann REE-mahn
Riesz REESE
Rolle RAHL
Schwartz (J., L.) SHWARTS
Schwarz (Hermann) SHVARTS Usually pronounced SHWARTS
Shilov SHWARTS
Sierpiński share-PEEN-skee n is technically nasal
Sinai SEEN-eye
Smulian SHMOOL-cee-m S should be written Š
Sobolev SO-buh-jeff SO-buh-lyeff is more accurate
Souslin See Suslin French transliteration
Stietjes STEET-yes Dutch
Suslin SOOS-leen Often SOOS-lin; not SUSS-lin!
Sz.-Nagy NAHDZH DZH is very soft
Szegő SAY-go SAY-ger is closer to correct
Szpilrajn SHPEEL-rine Same person as Marczewski
Tchebycheff See Chebyshev
Tietze TEET-suh not TEET-see or TEET-zee!
Tikhonov TEEKH-uh-noff Often pronounced TICK-uh-noff
Toeplitz TOE-plits Correct pronunciation closer to TER-plits
Tonelli ton-EL-lee
Tychonoff See Tikhonov
Ulam OO-lahm
Urysohn oor-uh-SAWN not YOUR-i-sun
van Rooij van ROY
Vitali vee-TAHL-ee
Volterra voal-TAIR-uh
von Neumann fun NOY-mun
Weierstrass VYE-er-shtrahss

1924
Notes:

Dirichlet These pronunciations are almost universally used, but there is some doubt about their correctness. Dirichlet was German, but his family originally came from the town of Richelette in the French-speaking part of Belgium. His full surname was Lejeune Dirichlet, originally Lejeune de Richelette. Even with a French pronunciation the t would be pronounced in this name. The logical pronunciation of Dirichlet should be something like deer-ish-LET. I have been unable to determine how he himself pronounced his name.

Perron Perron was German, but the origin of his surname is obscure.

Zermelo The origin of this name is obscure. Zermelo, who was German (and not Jewish), told people who asked that it was a contraction of Walzermelodie (cf. [Rei70, p. 98]). More serious speculations on an East Prussian origin of the name, apparently also originating with Zermelo himself, can be found in [Ebb07, p. 1].
XVII.1.1. The Greek Alphabet

One thing every mathematics student must learn is the Greek alphabet, or at least most of it. Greek letters are frequently used as mathematical symbols. Many upper-case Greek letters are essentially identical to Latin letters (although not necessarily the equivalent Latin letter!), and these upper-case Greek letters are rarely if ever used as mathematical symbols. The other upper-case letters are commonly used, but the lower-case letters are more common in mathematics. The only lower-case Greek letter I have never seen used as a mathematical symbol is omicron, although upsilon is rarely used since it is hard to distinguish from nu.

<table>
<thead>
<tr>
<th>Upper Case</th>
<th>Lower Case</th>
<th>Name</th>
<th>Latin Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>α</td>
<td>alpha</td>
<td>a</td>
</tr>
<tr>
<td>B</td>
<td>β</td>
<td>beta</td>
<td>b</td>
</tr>
<tr>
<td>Γ</td>
<td>γ</td>
<td>gamma</td>
<td>g</td>
</tr>
<tr>
<td>Δ</td>
<td>δ</td>
<td>delta</td>
<td>d</td>
</tr>
<tr>
<td>E</td>
<td>ε</td>
<td>epsilon</td>
<td>e (short)</td>
</tr>
<tr>
<td>Z</td>
<td>ζ</td>
<td>zeta</td>
<td>z</td>
</tr>
<tr>
<td>H</td>
<td>η</td>
<td>eta</td>
<td>e (long)</td>
</tr>
<tr>
<td>Θ</td>
<td>θ</td>
<td>theta</td>
<td>th</td>
</tr>
<tr>
<td>I</td>
<td>i</td>
<td>iota</td>
<td>i</td>
</tr>
<tr>
<td>K</td>
<td>κ</td>
<td>kappa</td>
<td>k</td>
</tr>
<tr>
<td>Λ</td>
<td>λ</td>
<td>lambda</td>
<td>l</td>
</tr>
<tr>
<td>M</td>
<td>μ</td>
<td>mu</td>
<td>m</td>
</tr>
<tr>
<td>Ν</td>
<td>ν</td>
<td>nu</td>
<td>n</td>
</tr>
<tr>
<td>Ξ</td>
<td>ξ</td>
<td>xi</td>
<td>ks</td>
</tr>
<tr>
<td>O</td>
<td>o</td>
<td>omicron</td>
<td>o (short)</td>
</tr>
<tr>
<td>Π</td>
<td>π</td>
<td>pi</td>
<td>p</td>
</tr>
<tr>
<td>P</td>
<td>ρ</td>
<td>rho</td>
<td>r</td>
</tr>
<tr>
<td>Σ</td>
<td>σ</td>
<td>sigma</td>
<td>s</td>
</tr>
<tr>
<td>T</td>
<td>τ</td>
<td>tau</td>
<td>t</td>
</tr>
<tr>
<td>U</td>
<td>υ</td>
<td>upsilon</td>
<td>u</td>
</tr>
<tr>
<td>Φ</td>
<td>ϕ or ϕ</td>
<td>phi</td>
<td>ph (f)</td>
</tr>
<tr>
<td>Χ</td>
<td>χ</td>
<td>chi</td>
<td>kh</td>
</tr>
<tr>
<td>Ψ</td>
<td>ψ</td>
<td>psi</td>
<td>ps</td>
</tr>
<tr>
<td>Ω</td>
<td>ω</td>
<td>omega</td>
<td>o (long)</td>
</tr>
</tbody>
</table>

1926
XVII.2. Origin of Terms and Notation

I have collected here my best information of when various standard mathematical terms were first introduced. Corrections and additions are solicited. Information is from [Caj93], http://jeff560.tripod.com/mathword.html, and http://jeff560.tripod.com/mathsym.html unless otherwise specified.

Terms

Many of these terms were introduced in languages other than English, in appropriate form. We have given the English form or translation.

algebra: originally *al-jabr* (according to [Str13], Arabic for “restoring”), Al-Khwarizmi, 825. Not used in the modern sense until much later.

algebraic topology: Lefschetz, 1936 [Jam99, p. 570]. Previously usually called “combinatorial topology” (not exactly the same thing; Emmy Noether was instrumental in the transition from “combinatorial topology” to algebraic topology in the late 1920s [Jam99, p. 564-565]).

algorithm: from Al-Khwarizmi, whose book (825) introduced “Arabic” numbers (actually originating in India) and modern calculation methods to Europe. “The method of Al-Khwarizmi” referred to solving a problem by calculation.

Banach space: Fréchet, 1928. Banach spaces were defined by Banach in 1920; he called them espace du type (*B*).

calculus: The term, from the Latin meaning “small pebble,” was used to refer to calculation (word of same origin) since Roman times. The modern usage of *calculus* (*differential calculus*) was introduced by Leibniz around 1680. Newton first used the term in 1691. The original meaning survives in some other mathematical terms such as *functional calculus* and *calculus of residues*.

compact (topology): Fréchet, 1906 (not quite in the modern sense).

complex number: Gauss. He used the term to resolve the ambiguity in the use of *imaginary number* to refer both to nonreal complex numbers and to pure imaginary numbers (an ambiguity which persists to a degree even today).

corollary: from Latin *corolla*, a small wreath; the English word dates from the 14th century and has been used in the modern mathematical sense (a theorem which is an immediate or easy consequence of another theorem) at least since 1669.


derivative: Lagrange, *Théorie des fonctions analytiques*, 1797 [Gra10]. Lagrange’s term (which he used as early as 1772) was *fonction derivée*. Derivatives were previously described as differentials, fluxions, etc.; the general notion of a function was just appearing in Lagrange’s time.

eigenvalue: A hybrid German-English word, apparently first used by A. S. Eddington in 1927. Many mathematicians have rebelled against the term and used alternates, but *eigenvalue* is now firmly established.


ergodic: coined by Boltzmann in 1884.

field (algebra): Dedekind used *Zahlenkörper* in 1858. The English *field* was first used by E. H. Moore in 1893.

geometry: From Greek “measuring the earth.” Used in ancient Greece before Plato.

harmonic analysis: W. Thompson (Lord Kelvin), 1867.

Hilbert space: Erhard Schmidt, *Über die Auflösung linearer Gleichungen mit unendlich vielen Unbekannten*, 1908 (for $\ell^2$). Abstract Hilbert spaces were defined by von Neumann in 1928 [vN30].

homeomorphism: Poincaré, *Analysis Situs*, 1892 (not quite in the modern sense). The modern usage was standard by 1930.

homology: Used by Poincaré in *Analysis Situs*, 1895, but not clearly in the modern sense [Jam99, p. 85]. The word was used in a different sense by James Stirling in 1749.

homotopy: Dehn and Heegaard, *Analysis Situs*, 1907 [Jam99, p. 88]. The concept of homotopy, or continuous deformation, had been around a long time.

imaginary number: Descartes, *La Geometrie*, 1637. Such numbers had long been considered.

integral: Jakob Bernoulli, *Acta eruditorium*, 1690. Johann Bernoulli disputed this (he disputed many things) and claimed originality for the term. Johann Bernoulli first used the term *integral calculus*, and convinced Leibniz to use it too.


linear algebra: the term was used by Bombelli in about 1550, but the modern usage as the subject of vector spaces and linear transformations only dates from about 1950.

logarithm: John Napier, 1614.

manifold: The term was used to denote affine subspaces of $\mathbb{R}^n$ long before the modern definition of manifold was given. First used in something approaching the modern sense by Riemann. The “correct” modern definition was nailed down by Whitney in the 1930s.

mathematics: from Greek *manthanein*, “to learn.” The term *mathematike tekhne*, “mathematical science,” already existed in ancient Greece. The word *mathematics* gradually shifted from a plural noun to a singular one (at least in American English; it is somewhat ambiguous in British English, where it is abbreviated “maths,” and still plural in some other languages).

matrix: Sylvester, 1848. Matrix notation had been recently introduced by Cayley. Determinants (of arrays) were introduced by Cauchy. The older meaning of *matrix* was “the place of origin of something else,” and Sylvester apparently regarded mathematical matrices as the place of origin of determinants.

parallelepiped: Euclid. The only controversy is about pronunciation: the accent is most commonly on the next-to-last syllable, although it is sometimes put on the previous syllable.

pathological: Used in medicine since the 17th century. The first documented use in mathematics is in Murray and von Neumann, *On Rings of Operators*, 1936 (referring to non-type-I factors).


real number: Descartes, *La Geometrie*, 1637.

ring: Dedekind.


set: Apparently first used (German *Menge*) by Bolzano, *Wissenschaftslehre*, 1837 [Fra66, p. 2].

topology: J. Listing, 1847 [Jam99]. The term did not come into general use until the twentieth century; *analysis situs* (introduced by Leibniz, used by Poincaré) predominated.
transcendental number: Leibniz, 1704 [Rib00b]. Transcendental numbers were not proved to exist until 1844 (by Liouville).

trigonometry: Bartholomaeus Pitiscus, 1595.

vector: Used in astronomy to denote a radius vector. First use to denote a point in $\mathbb{R}^3$ by William Rowan Hamilton in 1844. Only gradually became used to denote a point in a general vector space.

vector space: Peano, Formulario mathematico, 1895.

Notation

$a^n$ (exponent): Descartes, 1637 (natural number exponents); Newton, 1676 (negative and fractional exponents). Exponents were first used (with different notation) by Nicole Oresme about 1360.

$|\cdot|$ (absolute value): evidently introduced by Weierstrass.

$\binom{}{}$ (binomial coefficient): Andreas von Ettingshausen, 1826. Euler used this symbol, but with a fraction bar in the middle.

$\mathbb{C}$ (complex numbers): Nathan Jacobson, 1939.

.(decimal point): Unclear. Used by Bartholomaeus Pitiscus in 1612. Popularized by Napier. Decimals had long been used in the Middle East.

$\div$ (division): Johann Rahn, Teutsche Algebra, 1659.

$>$ (greater than): Thomas Harriot, Artis Analyticae Praxis ad Aequationes Algebraicas Resolvendas, 1631. Wallis used the first version of $\geq$ in 1670.

$\delta$ (limit): Cauchy (for différence). Johann Bernoulli used $\delta$ for a difference of functions in 1706.

$e$ (2.718⋅⋅⋅): Euler, 1727. The reason for $e$ is obscure, but it probably was not because of his name.

$<$ (less than): Thomas Harriot, Artis Analyticae Praxis ad Aequationes Algebraicas Resolvendas, 1631. Wallis used the first version of $\leq$ in 1670.

$\epsilon$ (limit): Cauchy (for erreur).


$f(x)$ (function): Euler, Commentarii Academiae Scientarium Petropolitanae, 1734.

$f: X \to Y$ (function): Hurewicz, 1940.

$f'$ (derivative): Lagrange, Théorie des fonctions analytiques, 1797 [Gra10]. The notations $\frac{dy}{dx}$ (Leibniz) and $\dot{x}$ (Newton) previously predominated.

$\emptyset$ (empty set): Bourbaki, 1939. Introduced by Weil from the Norwegian alphabet.

! (factorial): Christian Kramp, Elements d’arithmétique universelle, 1808. The alternate notation $\vdash$, still occasionally seen, is actually newer (Thomas Jarrett, 1827).

$i$ ($\sqrt{-1}$): Euler, 1777. Complex numbers, especially square roots of negative real numbers, had long been considered, although they were still “imaginary.”

$\infty$ (infinity): Wallis, De sectionibus conicis, 1655.

$\lfloor \cdot \rfloor$ (integer part): Kenneth Iverson, 1962. Gauss defined and used this function and denoted it $\lfloor \cdot \rfloor$.

$f$ (integral): Leibniz, 1675. The definite integral notation $\int_a^b$ was first used by Fourier in 1822.

1929
\cap (intersection): Peano, *Formulario mathematico*, 1895.

\lim_{x \to a} \text{ (limit): John Leathem, *Volume and Surface Integrals Used in Physics*, 1905. Weierstrass used } \lim_{x \to a}.

\in (member): Peano, *Formulario mathematico*, 1895.

\neg (minus): Johannes Widmann, 1489.

\mathbb{N} (natural numbers): Dedekind used \( \mathbb{N} \) in *Was sind und was sollen die Zahlen*, 1888.

\mathcal{O} (big-\( \mathcal{O} \)): Paul Bachmann, *Analytische Zahlentheorie*, 1894; but see \( o \).

\( o \) (little-\( o \)): first apparently used in the modern sense by Landau in 1909 (\( O \) and \( o \) are sometimes called Landau symbols). But \( o \) was used by Newton (cf. [?, p. 204] and V.7.).

\( \partial \) (partial): Condorcet, 1768. \( \partial \) was first used by Legendre, 1786. This symbol is an italic Cyrillic \( d \), and this may be the origin of the mathematical symbol (possibly through Euler?)

\( \pi \) (3.14159\ldots): William Jones, 1706 (for *perimeter*). Usage popularized by Euler. The number itself has been known and studied since antiquity.

\( + \) (plus): Nicole Oresme, 1360.

\( \mathbb{Q} \) (rational numbers): Bourbaki introduced \( \mathbb{Q} \) in the 1930s, evidently for *Quotient*.

\( \mathbb{R} \) (real numbers): Dedekind used \( \mathbb{R} \) in *Stetigkeit und irrationale Zahlen*, 1872 (he used \( R \) for the rational numbers). The real numbers as a named set were rarely referred to before this. Use of “blackboard bold” for \( \mathbb{R} \) and other sets of numbers, which is still not universal, is of uncertain origin.

\( / \) (division): Thomas Twinin, 1718.


\( \sqrt{} \) (radical): Christoff Rudolff, *Die Coss*, 1525. Descartes (1637) added the vinculum. The notation \( \sqrt{} \) was suggested by Albert Girard in 1629 and may have been first used by Rolle in *Traité d’Algèbre* in 1690.

\( \sum \) (sum): Euler, *Institutiones calculi differentialis*, 1755.

\( \times \) (multiplication): William Oughtred, 1631.

\( \cup \) (union): Peano, *Formulario mathematico*, 1895.

\( \mathbb{Z} \) (integers): Edmund Landau introduced \( \mathbb{Z} \) (evidently for *Zahl*) in *Grundlagen der Analysis*, 1930. Bourbaki used \( \mathbb{Z} \).
XVII.3. Miscellany

XVII.3.1. Age and Mathematics

“Very few people do anything creative after the age of thirty-five. The reason is that very few people do anything creative before the age of thirty-five.”

Joel Hildebrand

Weierstrass is a notable counterexample to the widely held belief that mathematics is exclusively an enterprise for the young. He never earned a doctorate (he was eventually awarded an honorary one), did not have a university position until age 41, and did most of his work on the foundations of analysis over the next twenty years after that. He did significant mathematical research well into his sixties and perhaps early seventies (he published the Weierstrass Approximation Theorem at age 70), and lectured regularly about his work into his seventies. (He did do important mathematical work in his twenties and thirties while he was employed as a high-school [gymnasium] teacher, but this work was not published or known in the mathematical community until much later.)

Hausdorff did not do much mathematics until he was in his forties (he previously wrote two books on philosophy under the name Paul Mongré).

John Wallis (1616-1703) is perhaps an even more notable example. He continued to produce and write significant mathematics into his eighties, and retained his chair at Oxford until his death at age (almost) 87. (He apparently did not learn or do any serious mathematics until after his Oxford appointment at age 33!) Although he was 27 years older than Newton, he remained active and productive for several years after Newton had ceased to do creative mathematics. When Wallis was 81, Leibniz invited him to Hannover to instruct the Germans on cryptography (he had been England’s Royal cryptographer for decades); Wallis declined for strategic reasons, not because he was physically incapable. (Wallis’s contemporary Thomas Hobbes (1588-1679), a formidable philosopher but a lightweight mathematician, with whom Wallis had a longstanding polemical debate, lived to age 91. Newton lived to age 84 and remained intellectually sharp almost to the end, but produced little new mathematics after a serious illness at age 50, although he did finally publish part of the work he had done decades earlier.)

“. . . no old Men (excepting Dr. Wallis) love Mathematicks.”

Isaac Newton

Leopold Vietoris died in 2002, two months short of his 111th birthday. He was the oldest person in Austria at the time of his death (and the oldest documented Austrian man ever). He published his last research paper at the age of 103. Sergei Nikolskii (1905-2012) also published research papers until age 101. I am not aware of any other mathematician who published a research work after his/her 100th birthday.

“If you live to a hundred you have it made because very few people die past the age of a hundred.”

George Burns

---

1 R. Byrne, The Other 637 Best Things Anybody Ever Said, 616.
2 William Whiston, Memoirs, p. 315-316; cf. Westfall, Never at Rest, p. 139.
3 R. Byrne, The Other 637 Best Things Anybody Ever Said, 585.
Other notably long-lived mathematicians:

- **Dirk Struik** (1894-2000) 106 years
- **Henri Cartan** (1904-2008) 104 years
- **Boris Bukreev** (1859-1962) 103 years

Notably short-lived mathematicians:

- **Évariste Galois** (1811-1832) 20 years
- **Jacques Herbrand** (1908-1931) 23 years
- **Mikhail Suslin** (1894-1919) 24 years
- **Pavel Urysohn** (1898-1924) 26 years
- **Niels Henrik Abel** (1802-1829) 26 years
- **Frank Ramsey** (1903-1930) 26 years
- **Rufus Bowen** (1947-1978) 31 years

1932
XVII.3.2. The Mathematical Organ

It was fashionable at one time for phrenologists to make studies of the skulls of various classes of people, including mathematicians. The following is adapted from *The Mathematician as a Type*, by William Story.

It has been claimed that a very large number of mathematicians show a striking physiognomic peculiarity called the “mathematical organ.” Franz Joseph Gall, the originator of phrenology, describes it thus:

“The outer part of the roof of the orbita is depressed in such a manner that the upper edge of the hollow of the eye retains its natural arch only on the inner half, and that the outer half becomes a straight line, which runs sloping downward to its outer end. In consequence of this, the outer part of the upper lid falls and covers the eye, more than usually. Still more decided is the character when the lateral part of the orbita is pushed outward, so that the angular process of the frontal bone (the outer upper angle of the edge of the roof) projects sidewise over the front part of the temple.”

Story continues:

“I know a great part of the living mathematicians personally [and] I have studied the busts, and painted and engraved portraits of many others. In all, without exception, I have found the organ described. Consider the portrait of young Colburn; in him the outer part of the orbita is depressed and pushed out in such a way that this peculiarity did not escape the authors of the first notices about the young man in the American newspapers. Consider the portraits of Kepler, Newton, Leibnitz, Huyghens, Descartes, Euler, Roberval, Bernoulli, Lagrange, de la Place, Lalande, Herschell, Olbers, Bessel, Monge, Carnot, etc.”

P. J. Möbius, a psychiatrist and disciple of Gall (and descendent of August Ferdinand Möbius), who published a book on *The Talent for Mathematics*, says:

“Gall’s statements about the mathematical organ are entirely correct, but I have three remarks to make: 1. Nature varies the forms more than might appear from Gall’s description. 2. The mathematical organ is not equally developed on both sides, but as a rule more strongly on the left. 3. The mathematical organ consists in part of a thickening of the soft parts. ... In fact, we find these modifications in almost all great mathematicians. However, sometimes the ridge running downward and backward from the lateral end of the eyebrows is more prominent and sometimes the slope of the outer edge of the orbita. Often the eye stands normally open, often the outer half of the lid hangs down so that the eye is half closed and seems to slant. If we observe the face from in front, it also strikes us that the distance between the outer angle of the eye and the contour of the face is strikingly great. Sometimes the space between the outer angle of the eye and the end of the eyebrow is particularly broad, the distance between the eye and the edge of the forehead magnified. In such cases there is usually neither a distinct ridge nor a slope; the origin is then, so to say, built upward. In individual cases the angular process is uniformly broadened with indefinite outlines, in others it projects with a sharp edge. The description is inaccurate in respect to the variety of nature. A definition that should be in accord with all variations is hard to be found. It would be, perhaps, best to say the mathematical organ consists in an abnormal formation of the angular process that amounts to magnification of the space inclosed by the process. We can say that on no two heads is the organ formed in exactly the same way. ... For the important fact that the mathematical organ is to be found prevalently
on the left, I need only refer to portraits. In the case of only moderate mathematical ability a
distinct modification is usually to be found only on the left. But at times there is scarcely a
perceptible difference between the left and the right, for example, in Leibnitz. At times, again,
the right as well as the left is abnormally formed, but the one is entirely different from the
other, for example, in Gauss. Here, also, the variations are many and it is neither practicable
nor advisable to describe all particular cases in words. . . . Even in considering portraits one
comes to the thought that the prominences are to be referred in part to an abnormally strong
development of the skin. If we examine the living, we can convince ourselves by feeling that a
considerable hyperplasia of the soft part exists. We feel distinctly that the angular process is
indeed uncommonly strongly developed, but also that the skin, more definitely the hypodermal
tissue, is thickened. The skin with an abundant cushion of fat forms a loose sack laid around the
angular process. Very frequently we see also strikingly heavy eyebrows."

With this information, you will undoubtedly be able to recognize the next mathematician you meet on
the street, and even judge the quality of this mathematician at a glance. (Also note that the author has
"strikingly heavy eyebrows," although he lacks an "abundant cushion of fat forming a loose sack laid around
the angular process," making him a good but not great mathematician.)
XVII.3.3. Mathematics and Mental Illness

This is a sensitive subject that I discuss with some hesitation. Mathematicians are often portrayed stereotypically in films and television as social misfits of borderline sanity. There are well-known examples of mathematicians who have suffered from mental illness, such as CANTOR, GÖDEL, and NASH, and other mathematicians whose behavior is at least quite odd by conventional standards (e.g. GROTHENDIECK, PERELMAN). And then there is TED KACZYŃSKI. Popular books, particularly about logic, sometimes suggest that sustained intense thought about mathematics and logic can, and typically will, drive a person insane.

An extreme (and fortunately not so well known) example was ANDRÉ BLOCH, who spent most of his life in an asylum for the criminally insane after killing his brother and two other family members out of fear that they were showing symptoms of hereditary mental illness present in his family history. While in the asylum he did a substantial amount of solid mathematics, particularly in Complex Analysis, and regularly corresponded with leading mathematicians, most of whom apparently did not know he was institutionalized. (He did do odd things such as date all his letters April 1 no matter when they were written.)

But mathematicians in fact appear to be no more, or less, susceptible to mental illness than any other group of people; in any category of people of substantial size, some percentage can be expected to have mental illness. For example, one can easily identify at least as many professional athletes as mathematicians who have had mental health problems, not even counting athletes who suffer brain injuries. Studies have shown that schizophrenics typically have below-average aptitude in mathematics.

Autism is possibly a different matter. There is evidence of positive correlation between autism and some types of mathematical ability. It is likely that some prominent mathematicians (possibly including NEWTON) have been mildly autistic. But the vast majority of mathematicians are in no way autistic.

It is true that many mathematicians are free thinkers who have an unconventional view of the world; one might even expect that the ability to transcend conventional ways of thinking is an important, even necessary, part of being successful in mathematics. And many mathematicians have rather low regard for social customs and norms, although many more are quite conventional socially. Part of the reason many mathematicians are somewhat unconventional may be that they can feel rather free to be so, due to the extreme degree of tolerance in the world of mathematics: mathematicians, probably more than almost any other social group, judge each other to a great extent simply on the quality of their work and ideas and not on social conformity.

“No one ever made a difference by being like everyone else.”

P. T. Barnum

It is also not accurate to characterize mathematicians as misfits who are loners or uncomfortable around others and in social situations. There are some such mathematicians, just as there are such people in most other walks of life, and mathematics may indeed be a comfortable refuge for such a person with sufficient mathematical talent; but most mathematicians are pretty adept and comfortable in social situations (VON NEUMANN was reputedly the life of every party he attended, and he is far from unique among mathematicians in this respect). And mathematicians tend to have broad interests and knowledge, and are not at all single-minded (except when they get working intensively on a mathematical problem, or some other project – mathematicians are very good at focusing their attention when necessary); you will find that mathematicians are often very interesting to talk to about almost any subject.

4Line from the 2017 film The Greatest Showman. There is no documentation that BARNUM actually said it.
Standard Notation in this Book

We will use the following standard (and sometimes not so standard) symbols to denote commonly used mathematical structures. In a few cases, the same letter is used to represent more than one structure; hopefully no confusion will result.

\[B^n\] Closed unit ball in \(\mathbb{R}^n\)

\(\mathbb{C}\) Field of complex numbers

\(\mathbb{D}\)
(1) Ring of dyadic rational numbers ()
(2) Open unit disk in \(\mathbb{R}^2\)

\(\mathbb{F}\) A field, usually \(\mathbb{R}\) or \(\mathbb{C}\)

\(\mathbb{H}\)
(1) Division ring of quaternions ()
(2) Hilbert cube \(\mathbb{I}^N\)

\(\mathbb{I}\) Closed unit interval \([0, 1]\) in \(\mathbb{R}\)

\(\mathbb{J}\) Set of irrational numbers in \(\mathbb{R}\)

\(\mathbb{K}\) C*-Algebra of compact operators on \(\ell^2\)

\(\mathbb{M}_n\) C*-algebra of \(n \times n\) matrices over \(\mathbb{C}\)

\(\mathbb{N}\) Natural numbers (positive integers)

\(\mathbb{Q}\) Field of rational numbers

\(\mathbb{R}\) Field of real numbers

\(\mathbb{R}^n\) Euclidean space of dimension \(n\) (set of \(n\)-tuples from \(\mathbb{R}\))

\(\mathbb{S}^n\) Unit sphere in \(\mathbb{R}^{n+1}\)

\(\mathbb{T}\) Topological group of complex numbers of modulus 1

\(\mathbb{Z}\) Ring of integers

\(\mathbb{Z}_n\) Additive group or ring of integers mod \(n\)

\(\mathbb{Z}_p\)
(1) Additive group or field of integers mod \(p\) (\(p\) prime)
(2) Ring of \(p\)-adic integers ()

The symbol \(\checkmark\) denotes the end of a proof, and \(\mathbb{Q}\) denotes “caution” or, at the end of an argument, “this is not really a proof as it stands.” Other hand symbols occasionally used should have obvious meaning.
XVII.4. Literature Review

“People who like this sort of thing will find this the sort of thing they like.”

Abraham Lincoln

We will give brief reviews here of some of the important books on real analysis and measure theory, and some other topics discussed in this book. In spite of any negative or qualifying comments I make here, almost all of these books are worthwhile to read and study. I have not yet included nearly all the books I am aware of that should be on the list, and I have not included any of the books that should be on the list that I am not aware of.

I generally try to write my own review before looking at any other reviews, and in most cases I have not read any other reviews; in cases where I have and other reviews differ substantially from mine or make additional points, I sometimes add comments to that effect (with attribution).

Undergraduate Analysis

Browder, Mathematical Analysis: An Introduction [Bro96]

This nice book covers undergraduate analysis, both one-variable and multivariable, as well as basic measure theory, and ends with some elementary differential geometry (manifolds and differential forms). Textbooks written by first-rate mathematicians often have a certain quality not found in some other books; the author’s overall approach, insights, and perspectives are quite illuminating. This book covers a lot of ground in a readable way, although some topics covered are not treated in depth.

The author is one of three brothers who are or were all eminent mathematicians. Their father was Earl Browder, a well-known American communist.

Euler, Foundations of Differential Calculus [Eul00]

English translation of the first part of Euler’s 1755 treatise on calculus. Primarily of (considerable) historical interest today, although almost all of the mechanical parts of calculus covered today in first-semester courses are here and treated in a nicely understandable way, if not always too rigorously (this was the first text about which this can be reasonably said), and as in most writings of great mathematicians there is much wisdom and insight in the exposition. And the man certainly could calculate! But, on the other hand, it is remarkable and revealing that even one of the greatest mathematicians had such a degree of misunderstanding about things we today regard as fairly basic, both in mathematics and physics (e.g. Euler seemed to consider it self-evident that matter is infinitely divisible). Throughout this book there is much beating of gums about the nature of infinity and the “infinitely small,” which he repeatedly states is actually zero or nothing while at the same time can be used as a divisor of another infinitely small quantity. This book graphically illustrates how far our understanding of the foundations of analysis has progressed since then.

Fichtenholz, Differential and Integral Calculus [Fic89]

In Russian. This classic three-volume treatise on undergraduate Real Analysis was used as a text by generations of students in the Soviet Union. Far more than just a calculus text. A German translation is available, but not an English one.

5 Book review; cf. R. Byrne, The Third and Possibly the Best 637 Best Things Anybody Ever Said, 514.
Fichtenholz, Foundations of Mathematical Analysis [Fe57]

In Russian. This two-volume set is a reworking of parts of Differential and Integral Calculus, and is aimed at a slightly higher (post-calculus) audience. A generally careful and readable exposition, although it suffers from excessive hand-waving in places (a criticism which can probably be made of most books on this list, including my own). Contains some material and viewpoints not readily found in Western texts. Unfortunately, no translation is available.

Körner, A Companion to Analysis [Kör04]

An insightful and entertainingly written account of undergraduate analysis. Treatment of the topics covered is mostly excellent. Although the mathematics is “honest” and carefully done, it is unsystematic and somewhat informal in places, and not all essential topics are covered, which make the book rather unsuitable as a text. But for its intended purpose as a “companion”, it is excellent.

Krantz, Real Analysis and Foundations [Kra05]

Like “little Rudin,” bridges undergraduate- and graduate-level analysis. I am not terribly impressed with this book, although it has some good features: it is generally pretty solid and complete, and gives a good introduction and motivation for interesting and important applications. But the exposition and organization are of uneven quality, and there are more errors than there should be in a book like this, especially a second edition. (Krantz is very prolific as a writer, but this comes at a price: his writing tends to suffer from a lack of polish and attention to detail, although much of what he writes is pretty good.) He claims that the writing style of this book, unlike older books, speaks to modern-day students in their own language, but I fail to see much difference in style between this book and older books; maybe he simply means the substantial number of statements in the text specifically addressed to the reader, which is indeed rather unusual writing style for a mathematics text but which I don’t imagine makes much practical difference to most readers.

The “preface” is primarily a masterpiece of self-promotion. It also contains the ridiculous assertion (written in 2005) that “real analysis has not appreciably changed in 150 years,” overlooking such fundamental work as that of CANTOR, DEDEKIND, LEBESGUE, . . . , as well as most of the foundational work of WEIERSTRASS (he did not yet even have a university position in 1855) and almost all of topology and functional analysis (and much more). Even the Riemann integral had just barely been developed and was not yet published or widely known. At least half of the present book, and a substantial portion of Krantz’s book, cover real analysis unknown, or at least not properly understood, in 1855.

If he had said “30 years” instead of “150 years,” the statement would have been more defensible, and more in line with what he was claiming in the preface to the first edition: that some other authors of recent analysis books have said that the subject has changed in the last 30 years. But this is a straw man; I am not aware of any author who has made such an assertion.

All in all, this book does not live up to the author’s hype.

The above comments were written about the second edition. A third edition has recently appeared, which has been revised into more of a traditional undergraduate real analysis text: the chapters on Lebesgue measure, wavelets, and differential forms have been removed. A chapter on functional analysis has been added (complete with introductory text directly contradicting the “150 year” statement). The remaining text is a little more refined, but the general comments above still apply. A quick and cursory perusal of the third edition turned up some sample howlers: every metric space is separable (section 13.2 problem 17; he says that this is “fairly tricky to prove” – I’ll say!); every bounded operator on $L^2$ is “essentially” a multiplication operator by the Spectral Theorem (section 14.4 problem 9); and in section 14.6 problem 6, a “norm” is defined which is not a norm (this topological vector space $C^\infty([a,b])$ is in fact not normable,
although it is completely metrizable with a metric defined by a formula similar to the “norm” formula, cf. ()).

There are also a number of inaccurate internal references, which is strange since accurate internal references are so easy in \TeX. Judging by the new preface, the author still has a pretty high opinion of himself and his book. (I should note that the prefaces to some of Krantz’s other books, such as [Kra15], have a far different tone and create a different picture of the author, who I must emphasize I do not know personally.)

Now there is a fourth edition. The slimming down of topics that began in the third edition has continued: for example, the chapter on functional analysis which appeared (with all its difficulties) in the third edition has disappeared just as suddenly. The author says he doesn’t want people to think this is a “high-level” text (I don’t think there is much danger of that; he also says he wants the text “to be the generational touchstone for the subject and the go-to text for developing young scientists” – a laudable but pretentious goal which is not likely either), but he still claims that this book is more rigorous and at the same time more readable than anything else since Little Rudin. But I don’t see either more rigor or more readability than, say, in [Pug15] (or even as much in this case).

For example, the discussion on p. 1 of the inadequacy of \( \mathbb{Q} \) for analysis is logical mush: the claim is that \( \mathbb{Q} \) is “not closed under limits” (apparently meaning limits of sequences) and must be expanded to a number system which is. But this argument presupposes that we already have a (perhaps still ill-defined) larger number system in which limits can be taken: if \( \mathbb{Q} \) is the only number system we have, it is tautologically closed under limits of convergent sequences. And if we want limits of all sequences, \( \mathbb{R} \) is also not closed. He tries to assert that every sequence of real numbers which “should” converge actually does. This is evidently an informal reference to the idea of a Cauchy sequence, a somewhat subtle notion which would certainly not leap to mind for someone (other than a Cauchy) who hadn’t seen it before. Many people would say the sequence \( (1, 2, 3, \ldots) \) “should” converge to \(+\infty\), but \(+\infty \notin \mathbb{R} \), so not every sequence in \( \mathbb{R} \) which “should” converge does, i.e. \( \mathbb{R} \) is not closed under limits either (e.g. in the extended real line). I’m not saying it is easy to give an elementary yet mathematically and logically honest introduction to the notion of completeness, but one can certainly do a lot better than the discussion in the book.

On p. 25, he proves that a sequence in \( \mathbb{C} \) converges if and only if it is Cauchy (after proving this for real sequences) by making the “observation” that a sequence in \( \mathbb{C} \) converges (resp. is Cauchy) if and only if the sequences of real and imaginary parts converge (resp. are Cauchy). There is no previous discussion of anything like this. In fact, both “observations” are propositions which need to be proved. Of course, the proofs are very easy, but they are not trivial, especially for someone who has just seen the definitions of the concepts; at the very least, these should be expressly designated as exercises for the reader, not just “observations.”

On p. 290, it is asserted that “a moment’s thought” shows that an open spherical shell cannot be written as a disjoint union of open balls. Maybe so. But another moment’s thought reveals that this is actually quite nontrivial to prove (and I’m not sure why the question is even relevant to the discussion on that page).

In fact, almost every page I looked at contained something I could quibble with, although many of the criticisms are petty ones concerning the author’s writing style. For example, on p. 19, he says, “If a sequence does not converge then we frequently say that it diverges.” Frequently? When do we not? (OK, I guess we don’t always actually say it diverges, but we could.) Why not just say: “A sequence diverges if it does not converge.”?

I guess someone (other than the author) must like this book, since it has been through four editions.

Pugh, Real Mathematical Analysis [Pug15]

This is overall an excellent book, and is one of the best references for learning and understanding real analysis. The author’s explanations and insights on most topics are outstanding. The book is mathematically honest and generally rigorous. However, many arguments are written a little informally, with some routine
details glossed over; thus a student using this as a textbook might have trouble developing the skill of writing complete rigorous analysis proofs if the book’s proofs are taken as a model (to be sure, there are plenty of proofs in the book written out fully enough to meet anyone’s standards).

Rudin, Principles of Mathematical Analysis [Rud76]

“Little Rudin” or “Baby Rudin.” One of the most widely-used analysis books for a long time, with good reason. Covers pre-measure-theory analysis with a last chapter on Lebesgue measure.

Schumacher, Closer and Closer [?]

A leisurely and non-stressful introduction to undergraduate real analysis. This is a pretty good book, despite its problematic title (cf. IV.1.3.11.); it is nicely written, the mathematics is generally carefully done and is not “dumbed down,” and a surprising amount of material is covered. The discussion of the Implicit Function Theorem is especially good. The treatment is mostly mathematically “honest” (I.7.), although some subtleties, e.g. the Axiom of Choice, are not mentioned. One mistake I noted is the wrong definition of countable compactness (p. 141), confusing countably compact with Lindelöf. The approach is an “active learning” one, based on the Socratic method: some arguments are given in great detail, but many are left as exercises for the reader, often with little hint as to how to proceed. Such an approach is certainly defensible pedagogically. This is a good supplementary text for a reader desiring a gentle but reasonably thorough treatment of the subject.

Stromberg, An Introduction to Classical Real Analysis [Str81]

This is a terrific book (although I would not describe the writing as a literary masterpiece). It is thorough and detailed, but also has much helpful explanatory material. The first few chapters cover standard undergraduate analysis at a similar level to most undergraduate texts, but the pace and depth increase when integration is reached. Riemann integration is not treated, except in passing after the fact; instead, the Riesz-Daniell approach to Lebesgue integration is followed. Some complex analysis is interwoven and integrated into the development at appropriate points. The exercises in this book are especially notable. This is not only a textbook (probably most suited to an honors-level analysis course), but also a great reference work even for mature analysts.

Wade, An Introduction to Analysis [?]

This is a decent book on undergraduate real analysis, covering both the one-variable and multivariable theories, but there are many small things I would wish to do differently. The fourth edition has a number of improvements over the third edition (although the deletion of the chapter on differential forms is not one of them), but the fourth edition seems “cheaper” than the third in both typesetting and binding. The author also did not do a very good job proofreading this edition before publication. There is a long list of typographical errors from the first printing on his website which have been corrected in later printings, but a significant number of typographical and mathematical errors remain.

Real Analysis, Measure Theory, and Integration

These textbooks cover the basic theory of measure and integration, and most also treat the basics of functional analysis.

Many general measure theory texts follow what I call the “reinventing the wheel” plan: first develop Lebesgue measure and integration, then redo everything in general, in many cases just referring back to the
Lebesgue case for proofs of results. This approach is certainly defensible pedagogically, and is preferred by many people, but it is not to my personal taste; I much prefer to start out with general measure theory, spending considerable time discussing and emphasizing the critical special case of Lebesgue measure as I go, as well as discussing other special cases such as probability spaces.

“One of the many fundamental questions which any author on the subject must decide, is whether to begin with ‘general’ measure theory or with ‘Lebesgue’ measure and integration. The point is that Lebesgue measure is rather more than just the most important example of a measure space. It is so close to the heart of the subject that the great majority of the ideas of the elementary theory can be fully realised in theorems about Lebesgue measure. ... If you take the view, as I certainly do when it suits my argument, that the business of pure mathematics is to express and extend the logical capacity of the human mind, and that the actual theorems we work through are merely vehicles for these ideas, then you can correctly point out that all the really important things in [elementary measure theory] can be done without going to the trouble of formulating a general theory of abstract measure spaces; and that by studying the relatively concrete example of Lebesgue measure on \(\mathbb{R}^n\)-dimensional Euclidean space you can avoid a variety of irrelevant distractions.

If you are quite sure, as a teacher, that none of your pupils will wish to go beyond the elementary theory, there is something to be said for this view. I believe, however, that it becomes untenable if you wish to prepare any of your students for more advanced ideas. The difficulty is that, with the best will in the world, anyone who has worked through the full theory of Lebesgue measure, and then comes to the abstract theory of general measure spaces, is likely to go through it too fast, and at the end finds himself uncertain about just which ninety percent of the facts he knows are generally applicable. I believe it is safer to keep the special properties of Lebesgue measure clearly labelled as such from the beginning.”

\[D. \text{ Fremlin}\]

The “reinventing the wheel” approach seems to me like presenting Linear Algebra by first stating and proving all the theorems of the subject for \(\mathbb{R}^n\), and then restating and partially reproving them for general vector spaces.

Bogachev, Measure Theory \[Bog07\]

A very thorough two-volume treatment of measure theory. In order to cover the huge volume of material, many topics, some of them important, are relegated to “exercises” where arguments are only sketched or outlined.

The English translation was apparently done by the author himself. Although this was a commendable effort, and in most places the translation is pretty good, the results are not entirely satisfactory, and in a few places the reader must puzzle out what the author is trying to say (of course, in many books written by native English speakers, probably including my own, the reader must do the same thing!)

Doob, Measure Theory \[Doo94\]

This book, written by an eminent probabilist, is somewhat idiosyncratic, but valuable. It has nice, efficient treatments of a number of topics not readily found in standard texts. The approach is not especially

\[^{6}\ [?] \text{, V. 1, p. 9.}\]
wedded to probability theory, but the chapter on martingales (invented by the author) is especially valuable for probabilists. Many arguments represented as proofs in this book are really only outlines, and need many details to be filled in by the reader (not necessarily a bad thing!)

Dunford and Schwartz, Linear Operators I [DS88]

This first of a three-volume set is usually regarded as an encyclopedic treatise on Operator Theory. But chapters 3 and 4 give an extensive treatment of measure theory and related topics. One of the best places to read about finitely additive measures and integration (from the beginning, a “measure” is a finitely additive complex measure), and vector measures.

Folland, Real Analysis [Fol99]

This is one of my personal favorite Real Analysis texts. It contains elegant and comprehensive treatments of many important topics. However, the basic measure theory part is much too concise to be a good choice as a text for a beginning course, especially for masters students, as I discovered when I tried it. The approach of this book is the exact opposite of the “reinventing the wheel” plan, in my opinion too much so: it treats Lebesgue measure almost as an afterthought, as a special case of Stieltjes (Radon) measures.

Fremlin, Measure Theory [Fre04]–[Fre06]

A monumental five-volume treatise on almost all aspects of the subject. Not suitable as a text, although the first two volumes are designed to be accessible to students, but a detailed and comprehensive reference work.

Halmos, Measure Theory [Ha50]

The first (1950) “modern” treatment of measure theory, and still one of the best. It is an excellent treatment of the basics of measure theory (done from the standpoint of what we call “premeasures on rings”), and includes some more advanced topics such as Haar measure. In fact, one can hardly improve on the exposition of many of the topics covered.

Hewitt and Stromberg, Real and Abstract Analysis [HS75]

A classic book. It is dense in places and not always easy going, but well written and contains a wealth of material.

Lieb and Loss, Analysis [LL97]

This is a valuable book with a little different emphasis than most. The authors’ philosophy is to seamlessly merge classical and functional analysis, which they have done well. The treatments of measure theory and basic functional analysis are rather sketchy, but the coverage of $L^p$-spaces and Sobolev spaces is excellent, among the most complete in any general analysis book. The general approach is a low-tech and computational one, emphasizing inequalities and “hard analysis” arguments rather than abstract general theorems (quite the opposite of Pedersen), an approach which has much merit and should particularly appeal to applied analysts; the primary intended audience for the book is evidently applied mathematicians and mathematical physicists, and there are many physics-related applications woven throughout. The book is written in somewhat of a stream-of-consciousness style, and would have benefited from some more organizational attention and effort on the part of the authors; but on the other hand, it is full of insightful comments and observations. This is not a book I would be likely to use as a text, but it is a great supplementary text.
McDonald and Weiss, A Course in Real Analysis [MW99]

This is overall one of the better measure theory texts, and one that I have used several times. The treatments of most standard subjects range from fairly good to excellent. Some of the writing is somewhat turgid, and some of the definitions (e.g. Borel sets, absolutely continuous functions) are nonstandard but equivalent to the usual ones. A few proofs are more laborious or roundabout than necessary. A few important basic results, e.g. the Lebesgue Density Theorem, are not covered. This text follows the “reinventing the wheel” plan. This book has a lot of exercises, some of which are fairly challenging, but generally with hints and guidance.

Natanson, Theory of Functions of a Real Variable

Another standard Soviet text. Well done, and covers material not found in other books (e.g. the Banach-Tarski paradox). Volume I is pretty routine, but Volume II is more interesting. The English translation leaves something to be desired, and is confusing in places: for example, in several instances the phrase “replace $x$ by $y$” should be “replace $y$ by $x$” (this is simply a translation error; the original Russian text is correct).

Pedersen, Analysis NOW [Ped89]

The author, a specialist in Operator Algebras, takes an uncompromising approach to measure and integration using functional analysis methods (he says the title means “Analysis via norms, operators, and weak topologies”). This is a valid approach with much to recommend it (in my opinion as a fellow Operator Algebraist), but should probably be complemented by a more traditional treatment. The text is clearly written, and witty in places (different readers will have different opinions about the wit).

Royden, Real Analysis [Roy88]

This is one of the “classic” texts (and the one I learned measure theory from as a student). Despite its age, it is still a viable and worthwhile option. The third edition has some additions and improvements, but also more errors than the second edition. This book follows the “reinventing the wheel” plan, to an even greater extent than most. An expanded and updated fourth edition, prepared by M. Fitzpatrick, has recently appeared.

Rudin, Real and Complex Analysis [Rud87]

“Big Rudin.” Also a classic book, widely used. The author rather uniquely covers and integrates Real Analysis (measure and integration) with Complex Analysis. These subjects are usually treated separately, but there is much merit in the author’s approach which highlights the overlap and interplay of the two theories (although I have chosen not to follow this path in this book). Rudin does not follow the “reinventing the wheel” plan; general measure and integration is developed in Chapter 1, and Lebesgue measure is obtained by applying the Riesz Representation Theorem to the Riemann integral.

Saks, Theory of the Integral [Sak64]

This was the first comprehensive book on measure and integration (1933), and is very impressive in its breadth and depth, as well as giving a nice overall view of the logic and philosophy of the subject. While the abstract Lebesgue integral on general measure spaces is treated, it is somewhat in passing and the discussion is practically limited to subsets of Euclidean space. At least half the book is devoted to a study of the interrelated topics of indefinite integration, bounded variation, and absolute continuity. The book
is still well worth careful study. Treatments of some topics have been improved since this book, but many
have not, and indeed this book covers a number of important topics not treated adequately in most newer
references. It is overall quite well written and readable, although dense and detailed. Some of the notation
and terminology is outdated and requires care on the reader’s part to decipher. Definitely a book for adults.

Swartz, Measure, Integration and Function Spaces [Swa94]

This book has a rather unconventional approach to the subject with much to recommend it. It treats
finitely additive, signed, and complex measures from the beginning rather than as an afterthought. It has
a good treatment of limits of measures, the Nikodým Convergence and Boundedness Theorems, and the
Vitali-Hahn-Saks Theorem. The treatment of fundamentals of functional analysis includes ordered vector
spaces.

Taylor, Measure Theory and Integration [Tay06]

A good general treatment of the subject, with topics leading in the direction of geometric measure theory
not well covered in other texts in this group. However, the author is not as careful as he should be in
specifying finite or $\sigma$-finite restrictions in some results and problems (I found several places where such
restrictions are necessary and not mentioned).

Complex Analysis

There are dozens, perhaps hundreds, of complex analysis texts. We will only describe a few of the best.
Many others are at least decent.

Ahlfors, Complex Analysis [Ahl78]

Generally regarded as the classic complex analysis text, and still worth studying.

Brown and Churchill

One of the classic complex analysis texts (I used an early edition as a textbook as a student, when it was
just Churchill). Oriented primarily for nonmathematics students, e.g. physicists and engineers, but unlike
some books of this type it is mathematically solid.

Gamelin, Complex Analysis [Gam01]

One of the very best complex analysis texts.

Greene and Krantz, Function Theory of One Complex Variable [GK06]

I had a somewhat better opinion of this book before I used it (third edition) as a text for a class. It
is overall quite good, especially the later chapters (e.g. the treatment of the Riemann Mapping Theorem is
generally excellent). But the first three chapters with the basic theory are somewhat of an organizational
mess, and some of the exercises are not well thought out (in fact, I had some trouble finding enough good
and suitable exercises for my class). And the authors beat the $\frac{\partial f}{\partial z}$ and $\frac{\partial}{\partial z}$ notation to death, and insist on
writing $\frac{\partial f}{\partial z}$ instead of $f'$ for the derivative of a holomorphic function $f$ throughout most of the book, and
even $\left(\frac{\partial}{\partial z}\right)^n f$ for the $n$’th derivative, making for confusing formulas which are hard to read.
Krantz, Complex Variables, A Physical Approach with Applications and MATLAB

This book seems designed as a modern replacement for Brown and Churchill. It is a rather slow-paced treatment of basic complex analysis aimed at a nonmathematical audience. Far less detailed and sophisticated than [GK06]. Generally well-done, with good insights and explanations of standard results. The MATLAB part (also covering some other similar packages) is not at all deep, barely going beyond how to do complex arithmetic. A good book to read as an introduction to the subject.

Remmert, Theory of Complex Functions [Rem91] and Classical Topics in Complex Function Theory [Rem98]

This is more or less a two-volume sequence, but there is a vast difference in depth and sophistication between the two volumes. The first is a treatment of the basic theory, suitable for undergraduates, and the second is a dense, detailed study of more advanced topics. Both are entertainingly written with extensive historical notes, including anecdotes only loosely related to the subject at hand (such as noting that Pringsheim was at first denied Habilitation “on account of the great ignorance of the candidate”; he apparently refused to explain how to solve quadratic equations.) The books are probably not suitable as primary texts, but are excellent supplemental texts.

Stein and Sakarchi, Complex Analysis [SS03]

Part of the authors’ series of books on analysis. This book stands on its own as well as a volume in the series. Covers basic complex analysis, not as thoroughly as in [Gam01] or [GK06] but well, and then has detailed treatments of number-theoretic functions and the Prime Number Theorem, and of elliptic functions. A very worthwhile addition to the literature, but perhaps not the best choice as a text for an introductory complex analysis course.

Functional Analysis

Kreyszig, Introductory Functional Analysis with Applications [Kre89]

A classic text, and one of the few options for a class without a measure theory prerequisite. It has a vaguely applied orientation, but mainly just covers the basic theory, and covers it well. Still a good choice as a text for an introductory course at the advanced undergraduate level.

Lax, Functional Analysis [Lax02]

This is overall one of the best texts on functional analysis, by one of the leading applied analysts of the 20th century. It includes many topics not covered in other standard texts, and the treatment of most topics is excellent. It has thorough coverage of theoretical functional analysis as well as some applied areas. (There are, however, a rather large number of typographical errors, most of which are inconsequential.) The author’s applied orientation shows through not only in his choice of material, but also in the style and tone of some of the presentation and his notation and terminology; however, this does not make the book less useful or valuable for “pure” mathematicians. Some of the exposition is more reflective of the 1950s than 2002, when the book was published; for one example close to my heart, the antiquated term “B*-algebra” is used, a term I have not otherwise seen since about 1980; specialists in this field have almost universally used the name “C*-algebra” since the 1960s.
**Specialized Topics in Real Analysis**

Diestel and Uhl, Vector Measures [DU77]

Despite the name, only the first two chapters cover the theory of vector measures and integration *per se*; the rest of the book consists of applications to the structure theory of Banach spaces. The first two chapters give a nice introduction to vector measures and Bochner integration. The authors’ personalities and strong opinions make for interesting reading.

König, Measure and Integration [Kön97]

This is a highly original book and contains much material not available elsewhere about construction of measures from primitive data. But the author’s opaque writing style and frequently nonstandard notation and terminology make for tough reading.

Krantz and Parks, The Implicit Function Theorem [KP02a]

The most thorough discussion available of the various versions of the Implicit Function Theorem and Inverse Function Theorem and their histories.

Krantz and Parks, A Primer of Real Analytic Functions [KP02b]

This is a quite thorough treatment of the theory of real analytic functions, covering not only the basic one-variable and multivariable theories, but also more advanced topics such as the Cauchy–Kovalevskaya Theorem, quasi-analytic and Gevrey classes, and the embedding theorem for real analytic manifolds. Complete (or at least semi-complete) proofs are given for most elementary results, but as the topics become more advanced the book becomes more of a survey. Generally well written and readable, although, as with many of Krantz’s books, there is an annoying lack of quality control and attention to detail in places. (I do not mean to slight the contributions of Parks, who has long been an acquaintance of mine.)

Oxtoby, Measure and Category [Oxt71]

This book is rather unique and does not directly overlap with any other books I know of (except for [BJ95], which is a sequel of sorts), although most topics can be found scattered elsewhere. The theme is the duality between the theories of Lebesgue measure and Baire category. Many results in one of these theories have an analog in the other, which is often more than merely an analog. Along the way, topics like the Banach-Mazur game are discussed and woven into the story in an illuminating way.

Schechter, Handbook of Analysis and its Foundations [Sch97]

I think the review of this book on MathSciNet is far too negative. This book, if properly viewed and used, is a valuable resource. Despite the name, not too much classical analysis is covered; instead, the author concentrates on an efficient and logical development of much of the mathematics a modern analyst needs to know, beginning with sets, relations, and orderings, proceeding through algebra and logic, topology and uniformity, and concluding with an extensive treatment of topological vector spaces. This skeleton outline does not do justice to the range of mathematics covered.

Globally, this book is very much in the style of *BOURBAKI* in its organization, attempting to present the most efficient logical path through the development of the material. But the local style is very different: Schechter allows himself much room for motivation, explanation, entertaining writing, and even some humor, things famously and deliberately excluded from *BOURBAKI* (although there is some subtle humor in
Bourbaki reflecting the personalities of the members, beginning with the name itself; indeed, according to [RS09], Grothendieck was not a successful member of Bourbaki in part because he “lacked humor.”

I discovered this book when my manuscript was already well along, and I was struck by how many passages were similar to things I had written, tried to write, or planned to write, often done better than I could have myself. Needless to say, I approve of the author’s views about many things in mathematics. (This is not to say that this book is overall very similar to my own; there are very significant differences.)

That said, I do not think this book would be an appropriate place to learn about most of the material covered; it would be most suitable for someone who had already been introduced to most of the major topics. The author did not assume any prerequisites beyond calculus except for a degree of mathematical maturity, but the book would be best used to fill in holes in one’s background and to see how everything fits together logically.

This book would have been much more effective as an online book with hyperlinks, organized more in the manner I envision for my own manuscript (i.e. less linearly). The author’s logical progression of development means that many results and ideas are introduced before they are actually used, and there are numerous cross-references to other parts of the book. If one simply wants to understand a certain topic or result, a maze of tedious searches through the book via cross-references is often necessary to get the full picture, which would be much more efficient and less annoying if done via hyperlinks. For example, there is a very nice and complete proof of the Brouwer Fixed-Point Theorem contained in the book, but pieces of the proof are scattered through several parts of the book. I hope eventually an online edition of this book can be produced with hyperlinks (there is currently a CD-ROM version, but it appears to be identical to the printed version).

van Rooij and Schikhof, A Second Course on Real Functions [vRS82]

This nice book concentrates on a detailed study of differentiation and integration of functions on the real numbers. The material is restricted in scope, but the treatment is quite complete, containing facts, results, and examples otherwise scattered through the literature.

Measure and Probability

These texts are generally very good, and could serve nicely for a first-semester course in measure theory, as well as for a probability course.

Billingsley, Probability and Measure [Bil95]

A current standard for this genre. Contains a lot of material on probabilistic number theory as well as general measure theory and, of course, standard probability theory.

Dudley, Real Analysis and Probability [Dud02]

An excellent treatment of both measure theory and probability. The two subjects are separated more strictly than in other books of this type, and emphasis is placed on Borel measures on metric spaces, which are especially important in probability. The advanced topics discussed are primarily, but not exclusively, directed towards probability. The writing is clear and readable, and in places rather informal and even chatty, with some important definitions and secondary results buried in passages of text. The writing style somewhat obscures the very careful treatment of the mathematics; I particularly admire the efforts the author made to present the most elegant proofs available of the results. He has a good sense of how to present material in a mathematically honest but non-overwhelming way. There are also extensive and valuable historical notes.

Kingman and Taylor, Introduction to Measure and Probability [KT66]
Geometric Measure Theory

Geometric Measure Theory fairly rapidly splits into two rather distinct parts: “smooth” geometric measure theory and the study of fractals. Some of these references (e.g. Morgan) concentrate on the first, and others (e.g. Edgar) the second. But the basic tools like Hausdorff measures are common to both, and all these references have at least some relevance to both parts.

Edgar, Measure, Topology, and Fractal Geometry [Edg08]

An undergraduate-level introduction to fractals and the tools needed to study them. Contains nice but somewhat “light” (focused on the needs of the book) introductions to metric spaces, topological dimension, and measure theory (Lebesgue and Hausdorff measures). A good place to begin reading about the subjects covered.

Edgar, Integral, Probability, and Fractal Measures [Edg98]

A “sequel” to Measure, Topology, and Fractal Geometry. The title gives a good summary of the topics covered.

Falconer

Descriptive Set Theory

Bartoszyński and Judah, Set Theory: On the Structure of the Real Line [BJ95]

This book more or less picks up where Oxt71 ends. But the treatment is quite technical, and mostly of interest only to specialists in Set Theory.
Kechris, Classical Descriptive Set Theory [Kec95]

The best introductory treatment of Descriptive Set Theory. Covers the basics of the subject thoroughly from many points of view, and makes a convincing case for the importance of the subject and its place in the world of mathematics. More readable than Moschovakis for nonexperts in set theory, although it is more of a complement than a substitute.

Moschovakis, Descriptive Set Theory [Mos80]

The standard treatment of Descriptive Set Theory before Kechris. Requires more detailed familiarity with set theory and logic. Well written and readable, however, and a valuable supplement to Kechris.

Foundations, Set Theory, Logic, Large Cardinals

Kanamori, The Higher Infinite [Kan03]

This is, and will likely remain (additional volumes are planned), the definitive treatment of the theory of large cardinals. Much of the text is tough going for readers not expert in set theory and mathematical logic, but there are also extensive explanations and motivation for important topics and results, so the book is valuable even for nonspecialists. The order of topics follows the historical development, so the origins of the ideas can be traced. There are also numerous and quite candid descriptions and evaluations of the work of individual mathematicians involved in this subject, which make for interesting reading.

Landau, Foundations of Analysis [Lan51]

The first complete treatment of the construction of the real and complex numbers from the Peano axioms (although the various constructions had long since been known and carefully described). The text is rather dry and formal with little explanatory material (although the introduction is engagingly written), but not too much is really necessary and the arguments are easy to follow.

Lawvere and Rosebrugh, Sets for Mathematics

This is an appealing and readable treatment of set theory and the foundations of mathematics using categories, and makes a good case that this is the most natural and fruitful approach to foundations for most working mathematicians. However, I find somewhat unconvincing the authors’ cavalier assertions that category theory and “naive set theory” (which is what is obtained when the category framework is stripped away) as presented in this way are free from logical contradiction and paradox.

Lucas, The Conceptual Roots of Mathematics [Luc00]

This is an entertaining and enlightening book, as well as a good source of pithy quotes. The theme is an argument for a “chastened logicism” as the correct philosophy of mathematics; but there is much more of interest, such as one of the best discussions I have read of second-order logic. I know nothing of the author, but he comes across as something of a character. The quality of both the mathematics and the writing is somewhat uneven, ranging from silly (arguing that Euclidean geometry is “better” than noneuclidean, or invoking stereotypes of Scotsmen and Welshmen to explain some terms) to trenchant.

Nagel and Newman, Gödel’s Proof [NN01]

A nice introduction to Gödel’s Incompleteness Theorem, readable for nonexperts, although it contains a serious misstatement (see H. Putnam’s review in Philos. Sci. 27(1960), p. 205-207).
Potter, Set Theory and its Philosophy [Pot04]

This is an impressive and valuable treatment of set theory which bridges philosophy and mathematics. The mathematical part of the book contains good treatments of the construction of number systems, cardinals, ordinals, and the axiom of choice, as well as set theory itself. Set theory is axiomatized by the theory of levels, but there is also extensive discussion of other systems such as ZF which give theories essentially equivalent for the purposes of providing a foundation for mathematics. Includes extensive historical discussions of the development of the various topics and the philosophy underlying them. Very nicely written, and the philosophical part is varied and balanced. Nonetheless, mathematicians may not agree with all of the philosophical arguments given, and there are a few minor mathematical errors (perhaps no more than in any other book on the list!), such as the formula on p. 232 and accompanying comments.

Sierpiński, Cardinal and Ordinal Numbers [Sie65]

An encyclopedic treatment of ordered sets, ordinals, and cardinals. Contains much material about the order structure of the real line. Interesting reading, with many clever arguments and observations. Some of the language used is a little unusual mathematically, and logically ambiguous in places; for example, a number of results are phrased as “We can prove with the aid of the Axiom of Choice that . . . .” Apparently this language was used because at the time of the publication of this book, it was not definitely established that these results were independent of ZF set theory (most, if not all, have been subsequently shown to be independent).

Smullyan, Gödel’s Incompleteness Theorems [Smi92]

A very readable and mathematically complete treatment of various versions of Gödel’s Incompleteness Theorem. Perhaps the best source to learn about the details of the theorem.

Nonstandard Analysis

Davis, Applied Nonstandard Analysis [Dav77]

Goldblatt, Lectures on the Hyperreals [Gol98]

Hurd and Loeb, An Introduction to Nonstandard Real Analysis [HL85]

Robinson, Non-Standard Analysis [Rob66]

The original treatment of Nonstandard Analysis, and as such of historical and conceptual value. But it is outdated as an exposition of the subject: for example, the hyperreals are constructed using the Compactness Theorem of model theory rather than via ultrafilters (the distinction is more apparent than real, since ultrafilters are used to prove the Compactness Theorem). No longer the best reference for learning about the subject. The last chapter is a discussion of Robinson’s interpretation of the history and philosophy of calculus; while interesting and valuable as motivation for his development of Nonstandard Analysis, some of his opinions are questionable at best.

Constructive Analysis

Bishop and Bridges, Constructive Analysis [BB85]

This is a “second edition” of Bishop’s treatise, largely prepared by Bridges after Bishop’s death.
Bridges and Viţă, Techniques of Constructive Analysis [BV06]

A sequel of sorts to [BB85], although it does not presuppose any knowledge about constructive analysis. It includes much material that postdates the publication of [BB85], and is more reflective of the state of the art in this subject. Nicely written, and a good place to learn about some constructive analysis without becoming embroiled in the philosophical controversies underlying it: the authors present the subject in the right way, in my opinion, as an interesting and useful part of mathematics rather than the only true way to do all of mathematics.

The authors’ presentation does, however, have some foundational obscurities which must be sorted out. In their approach, a “set” is a collection of objects constructed according to some property, from objects which have “already” been constructed [if a construction is done once, is it permanent, or can it be “lost”, e.g. with the death of its constructor?], along with an equivalence relation of “equality” on the objects. In practice the mathematician cannot distinguish between equivalent objects. This approach implicitly presupposes two levels of existence and knowledge: an omniscient level where the objects really exist and where true equality of objects makes sense, and where the equivalence relation can be truly defined, and a lower level of perception on which human mathematicians operate where objects can only be described or distinguished by constructive arguments. Viewed this way, constructive analysis nicely models the way analysis is actually carried out by mathematicians and especially by computers, and the algorithmic arguments forced by this approach are of obvious importance.

From the point of view I take in the last paragraph, what the authors call the “equality” relation should be called “indistinguishability,” and what the authors call “inequality” (constructive proof that two objects are not “equal”) should be called “distinguishability.” But there are still some logical difficulties: for example, in practice it is not obvious that “indistinguishability” is an equivalence relation, specifically that it is transitive. For (perhaps oversimplified) example, suppose we say we can only distinguish real numbers if they differ by at least $10^{-6}$ (this is reasonable from a computer architecture point of view, perhaps with a different constant). We can then have three numbers $a, b, c$ with $a$ and $b$ indistinguishable, $b$ and $c$ indistinguishable, but $a$ and $c$ distinguishable. In fact, any type of archimedean property of the real numbers would seem to be incompatible with indistinguishability being transitive; real numbers indistinguishable from 0 would be analogous to infinitesimals in nonstandard analysis (ironically, constructivists, notably including Bishop [Bis77], are some of the most vocal critics of nonstandard analysis).

Another somewhat jarring feature of the presentation of the book is the rather brief and unmotivated adoption of the set $\mathbb{N}$ as the foundation for all of analysis. The only justification given for this is two quotes from Kronecker and Bishop in effect saying that they regarded $\mathbb{N}$ as a gift from God; the authors seem to suggest that if it was good enough for them, it is good enough for us too. This approach is somewhat of an injustice: in fact, there are reasonable arguments in favor of viewing $\mathbb{N}$ as a starting point, such as ones described in [Hey66]. But I still think the worship of $\mathbb{N}$ as the starting point of mathematics is a philosophical weak point of intuitionism/constructivism.

Heyting, Intuitionism: an Introduction [Hey66]

Number Theory

Baker, Transcendental Number Theory [Bak90]

Perhaps the definitive work on the subject, covering the author’s revolutionary contributions. The techniques cover a broad range, from algebra and number theory to geometry and hard analysis. Worth looking at, if only to get a feeling for the character of the subject.
Hardy and Wright, Introduction to the Theory of Numbers [HW08]

Khinchin, Continued Fractions [Khi97]

This little gem is probably the best introduction to the theory of continued fractions. The first 2/3 of the book is an elementary and very readable introduction to the algebraic theory. The last part is on a considerably more sophisticated level, giving a nice and still very readable treatment of the probabilistic aspects of the theory and full proofs of the author’s important results about the probabilistic distribution of terms in continued fractions.

Niven, Irrational Numbers [Niv56]

Geometry

Berger, Geometry I, II [Ber09]

Originally a single volume in French. A wide-ranging and informative treatment of geometry, including such topics as map projections for the earth.

Blumenthal, Theory and Applications of Distance Geometry [Blu70]

This is an interesting, valuable, and rather unique treatment of geometry from the metric space point of view. The subject matter is covered thoroughly, not only metric characterizations of Euclidean space, but also curvature and elliptic and hyperbolic spaces. Some of the notation and terminology is outdated, and some sections and arguments seem a bit laborious and unfocused. Nonetheless, it is the best reference on these topics, and virtually the only one on some of them.

Greenberg, Euclidean and Non-Euclidean Geometry [Gre93]

An elementary college-level text on axiomatic geometry, with emphasis on the history and nature of the Parallel Postulate. Very readable, and mostly carefully written, although EUCLID’S postulates are rather mangled (they are eventually written properly, but seemingly only as an afterthought).

Hartshorne, Geometry: Euclid and Beyond [Har00]

A careful and comprehensive axiomatic treatment of elementary geometry, both Euclidean and Noneuclidean. Written on a more sophisticated level than Greenberg, but very readable. The connections between geometry and modern algebra are extensively discussed and explained. The author has included nice touches such as hand-drawing most of the diagrams to maintain the spirit of EUCLID.

Moise, Elementary Geometry from an Advanced Standpoint [Moi90]

Topology

Topology, as covered in the literature, seems to be essentially two separate subjects: point-set topology and algebraic topology. Few books treat both parts of topology in any detail. Even within each part, standard references vary greatly in which topics are covered in detail, since the subjects are too large to be covered comprehensively in any one reference. Different standard references are better or more complete for different topics, to a much greater extent than in basic real analysis, and it is not feasible in brief reviews to give a full
summary of which references are good for which topics (and this is a somewhat subjective matter anyway). Readers are encouraged to explore broadly.

Alexandroff (Aleksandrov) and Hopf, Topologie I [AH74]

In German. The first comprehensive book on topology, both point-set and algebraic (1935). I was rather surprised how much of the basics of topology had reached its modern form (some of it for the first time) in this book, since the whole subject had barely begun by the beginning of the twentieth century, although Hausdorff’s book [Hau49] must be regarded as more revolutionary. In particular, much of the topology we cover in this book appears here in essentially modern form. Although this is denoted “Volume I”, there was never a Volume II or, unfortunately, an English translation.

Borsuk, Theory of Retracts [Bor67]

Dugundji, Topology [Dug78]

Engelking, General Topology [Eng89]

The third iteration of this book. This is the most comprehensive text on general topology, and seems to have become the standard reference, although there is essentially no mention of algebraic topology concepts, even the fundamental group (in fact, homotopy is barely mentioned at all). It is also nicely written, with considerable explanation, motivation, and historical notes. However, it is probably too detailed (and esoteric in many places) to be a good text for an introductory course. Copies are also rather difficult and expensive to come by.

Engelking, Theory of Dimensions Finite and Infinite [Eng95]

Revised version of Dimension Theory (available online). A modern version of Hurewicz and Wallman, covering the theory of dimension beyond the separable metric case, and the theory of infinite-dimensional spaces. Nicely written, and a good place to learn about dimension theory even for separable metric spaces, the case of most general interest.

Hatcher, Algebraic Topology [Hat02]

One of the best modern treatments of classical algebraic topology, and freely available online to boot. Very clearly written, with a highly readable exposition of most important basic aspects of algebraic topology. Perhaps the best source for learning the fundamentals of the subject, and a model for what I hope will someday be available online in all parts of mathematics, a goal to which I hope my own book will make a contribution. The author promises additional volumes, which I sincerely hope to see.

Hausdorff, Grundzüge der Mengenlehre [Fundamentals of Set Theory] [Hau49]

This was a path-breaking book (1914), and still merits study. The first half of the book is an exposition of set theory, ordered sets, ordinals, and cardinals; while it is undoubtedly a better treatment of these subjects than was available previously, this part does not really break new ground. But the second half of the book can be fairly regarded as the beginning of the modern subject of point-set topology, although the notion of an abstract topological space was first described by Fréchet in 1906. It contains the first modern definition of a metric space (modeled on previous work of Fréchet as well as Hausdorff himself), and essentially the modern definition of a topological space (there had also been previous work on this, notably by Hilbert). A quite remarkable percentage of the topology we cover in this book appears here. While the topology of
the real numbers and Euclidean space was already well understood (primarily via convergence of sequences and series and continuity of functions), the transition to abstract spaces as we view them today sprang forth rather suddenly in amazingly detailed and polished form in this book.

Later editions of this book (e.g. [Hau44]), and the English translations (e.g. [Hau57]), are considerably abbreviated, leaving out much of the text of the first edition including most of the topology (but with a more refined exposition of what remains), with added material on Descriptive Set Theory. So the later editions are really a different book whose title more accurately describes the contents. This book is mainly of historical interest today, especially the first edition.

Hurewicz and Wallman, Dimension Theory [HW41]
Kelley, General Topology [Kel75]

Kunen and Vaughan, ed., Handbook of Set-Theoretic Topology [KV84]

The title of this collection of articles should be taken seriously (at least the “set-theoretic topology” part; the “handbook” part is pretty misleading – this is a massive 1273-page work which would qualify as a handbook only for the likes of Paul Bunyan, although I note that several “handbooks” of similar size on other subjects are also published by Oxford University Press, so maybe the British have bigger hands). Set-theoretic topology is not quite the same thing as point-set topology; the topology treated in this book is based on the intricacies of set theory, including set-theoretic axioms and large cardinal hypotheses. Almost all the results discussed are either trivial or inapplicable for metrizable spaces, and many of them even for spaces which are separable and/or first countable. The articles seem to be carefully written and comprehensive, but will be of interest primarily to specialists in this field, who would be characterized more as set theorists than mainstream topologists. Even the articles on Borel measures, Banach spaces, and topological groups, which should be of interest to a functional analyst such as myself, have a heavy dose of set-theoretic content, although I found parts of each to be of interest. Besides these, the only article which evinced more than a passing interest from me is P. Nyikos’s discussion of nonmetrizable manifolds, and these still strike me as being mostly a curiosity. Several other articles do have preliminary sections which are well-written overviews of parts of the subject for nonexperts.

Kuratowski, Topology I, II [Kur66], [Kur68]
Massey, Algebraic Topology [Mas77], [Mas91]
Munkres, Algebraic Topology [Mun84]
Munkres, Differential Topology [Mun66]
Munkres, Topology [Mun75]
Seifert and Threlfall, A Textbook of Topology [ST80]

An early (1934) treatment of algebraic topology. Nicely written and still very readable, and of the early topology texts it is the only one which is still worthwhile studying for content instead of just historical value. Parts are rather outdated today; the English translation, published in 1980, describes some but not all of these (and even an exposition current in 1980 is now a bit dated). The small amount of point-set topology covered is particularly rudimentary and outdated. A few things in the exposition are at least misleading if not inaccurate:
In §52, an argument is given that purports to show that every finitely generated group occurs as the fundamental group of a closed 4-manifold (homogeneous 4-complex). But the argument clearly only works as written for finitely \textit{presented} groups, a much smaller class; and in fact the fundamental group of a closed manifold (and, more generally, a compact metrizable ANR, cf. ()), is always finitely presented. Actually the statement in the book is somewhat ambiguous since their complexes are not required to be finite; the question of whether every finitely generated group, or more generally any countable group, occurs as the fundamental group of a (possibly noncompact) 4-manifold appears to be up in the air, but the answer is positive for open subsets of $\mathbb{R}^5$; cf. \url{http://math.stackexchange.com/questions/1362655/fundamental-group-of-open-subsets-of-mathbb{R}^5}.

The definition of “manifold” in the book is what is now called a \textit{combinatorial homology manifold}, which is significantly different from a topological manifold: a combinatorial homology manifold is compact, with no boundary, triangulable, but not necessarily homogeneous. Thus only closed manifolds are combinatorial homology manifolds, and only those which can be triangulated, but on the other hand there are combinatorial homology manifolds which are not manifolds, such as the suspension of the Poincaré homology 3-sphere ().

Spanier, Algebraic Topology [Spa81]

My immediate impression of this book when I first picked it up as a student was that it is quite heavy for its size. When I started reading it, this impression persisted in a metaphorical sense: it is pretty dense and detailed. But it is also pretty readable, and it was a standard reference and text at that time for good reason. It may not still be state of the art, but it is still a worthwhile reference.

Willard, General Topology [Wil04]

\textbf{Numerical Analysis}

Acton, Numerical Methods That (Usually) Work [Act90]

A lively and informative book. The contents could be uncharitably described as a large bag of tricks, but the book is really much more than that: it is a detailed discussion of the art of computation (with emphasis on the \textit{art}), meaning finding, adapting, and using appropriate numerical procedures for computational problems. The author was an engineer, with a fairly typical engineering perspective on mathematics, but informed by extensive experience in computation, and there is much wisdom and common (even uncommon) sense here. His writing style is refreshing and somewhat self-effacing; although he was an acknowledged expert in computation, he never talks down to the reader. For example, he states, “I have tried to make my explanations clear, but sad experience has shown that you will not really understand what I am talking about until you have made some of the same mistakes that I have made.” There is not much attention to mathematical rigor, and even the mathematical details of standard procedures are not discussed in depth, but there are beautiful heuristic descriptions of many procedures. This book is a must-read for anyone preparing to work in computation.

Hamming, Numerical Methods for Scientists and Engineers [Ham86]

As a nonexpert in numerical analysis, I found this to be one of the most valuable books for understanding the philosophy of the subject. While many of the nuts and bolts of numerical analysis are treated in detail, and in a way I mostly found mathematically satisfactory, the book’s main strength is extensive explanations
of *why* things are done as they are, from the point of view of the user. Pure (and even applied) mathematicians may be occasionally offended by sarcastic comments of the author (his colleague BRIAN KERNIGHAN has fondly called Hamming a “classic curmudgeon”). For example, he describes the difference between Numerical Analysis and Numerical Methods thus:

> “Numerical analysis seems to be the study in depth of a few, somewhat arbitrarily selected, topics and is carried out in a formal mathematical way devoid of relevance to the real world. Numerical methods, on the other hand, try to meet the need for methods to cope with the potentially infinite variety of problems that can arise in practice. The methods given are generally chosen for their wide applicability in creating formulas and algorithms as well as for the particular result being found at that point.”

(In fact, I disagree with much of the author’s philosophy of mathematics as described in [Ham98], although I appreciate his point of view.) But from this book one really gets a feeling for how engineers and other consumers of numerical analysis view the subject and how their needs are accommodated. The author’s stated motto is “The purpose of computing is insight, not numbers,” and he delivers on this goal.

**Miscellaneous Topics**

Wagon, The Banach-Tarski Paradox [Wag93]

This is one of the best mathematics books I have read. It covers an enormous range of fascinating topics in a very engaging way, and has succeeded in giving a quite readable treatment of some deep topics without pulling punches. Not all of the topics covered are directly related to measure theory, but many are, sometimes in surprising ways. This book really gives the reader a feeling for the way many diverse parts of mathematics fit together and interact.

**History of Mathematics**

Ball, A Short Account of the History of Mathematics [Bal60]

This book, first published in 1891, is badly out of date and inconsistent with modern historical scholarship in many parts; thus its modern-day value as a serious historical work is limited. It is still interesting reading, however, with many personal tidbits which are generally more reliable than those in bell’s books.

Bell, The Development of Mathematics [Bel92]

This is a more solid and serious book than *Men of Mathematics*, and has been generally praised by historians of mathematics, although the early chapters have faced some criticism. It is written in an engaging style, with many interesting insights and strong opinions. But it should not be the only book one reads about the history of mathematics.

Boyer, A History of Mathematics [Boy89]

Eves, An Introduction to the History of Mathematics [Eve90]

---

The thesis of this book is that “Name Worshipping,” a sect of the Russian Orthodox Church, played a significant role in the mathematical careers of Dmitri Egorov and Nikolai Luzin, and in the development of Descriptive Set Theory by the Moscow school. Part of the book is a detailed account of the lives of these mathematicians and their colleagues and students, which is one of the most detailed studies yet of the important group in Moscow commonly called “Lusitania.” Other parts describe the practice of Name Worshipping and attempt to make the connection with the mathematical philosophies of Egorov and, especially, Luzin.

I found the book somewhat disappointing until I reached the last chapter, where the authors put things together nicely; this chapter is a glimpse of what the earlier part of the book could have been. Unfortunately, it seems that the authors, who are themselves certainly well versed in mathematics, shied away from depth and technicalities in most of the book in a misguided effort to make it accessible and interesting to non-mathematicians, and thus made their arguments less convincing to mathematicians. I found many of their arguments in the earlier part of the book strained and unconvincing. For example, they often seem to argue that due to the Name Worshipping philosophy, Luzin et al. conflated naming mathematical objects with establishing their existence. No one with a significant background in mathematics could accept this, and it is clear from a quote of Luzin in the Appendix that this was not what he thought (the authors also explain in the last chapter that they did not believe this argument or intend to make it). Several analogies proposed such as the comparison with Grothendieck are quite strained: Grothendieck’s talent described in the book (only one of his many talents) was his ability to come up with good names for mathematical objects which would inform and inspire himself and others; in the philosophy described in the book, where simply naming or explicitly describing mathematical objects was the crucial matter, the quality of the names chosen would seem to be irrelevant. And although the book draws extensive comparisons between the mathematics in Moscow and developments in France during the same period and immediately preceding, I was surprised that the religious influences on the work and philosophy of Cantor, who got the whole subject of set theory going, were almost totally ignored (it seems clear from other sources that Cantor was indeed significantly influenced by Christian theology and/or Jewish traditions).

This is overall an interesting book to read, but it could have been a lot better.
the principal remaining parts of traditional real analysis with an active research component (one might add probability). In addition, we now have an ever-increasing collection of things regarded by at least some mathematicians as specific concrete objects, e.g. manifolds, well-ordered sets, Hilbert space. But even the subjects superficially unchanged from 200 years ago, such as number theory and differential equations, use a range of techniques far expanded from the traditional ones which could be developed or even contemplated only as a result of the modernist revolution. For example, the modern study of differential equations heavily uses methods from functional analysis, listed by the author as one of the shining examples of modernist mathematics. (Of course, traditional techniques also continue to be used; after all, even calculus as known and understood in the eighteenth century is still extremely important and useful.) The subject of differential equations also heavily uses sophisticated modeling and simulation techniques made possible by the existence of powerful computers, a development not anticipated even at the end of the modernist revolution; the present use of computation in mathematics may mark another revolution which will turn out to be as profound as the modernist one.

Although parts of this book are dense and a challenge to slog through, it is highly recommended.

Jahnke (Ed.), A History of Analysis [Jah03]

Kline, Mathematical Thought from Ancient to Modern Times [Kli72]

This is Kline’s *magnum opus* and principal scholarly work on the history of mathematics; he has a number of other books written more at the popular level. In this, as well as his other books, there is an undercurrent of Kline’s philosophy of mathematics which I personally take some issue with, although this does not seriously undercut the value of these books as historical works. There are also small doses of sexism here and there in Kline’s books.

Moore, Zermelo’s Axiom of Choice [Moo82]

A careful and thorough study of the Axiom of Choice – its prehistory, history, controversy, variations, and intimate connections with the very nature of modern mathematics.

Pier, Mathematical Analysis During the 20th Century [Pie01]

The author has set himself a monumental and perhaps impossible task in trying to summarize and organize the development of analysis during the twentieth century, and he has partially succeeded, perhaps as well as any one individual could in a book of this length published at the immediate end of the period studied (he is well aware of the inherent problems of such a project, and indeed divides his description of each part of analysis into “evolution” covering the period 1900-1950 and “flashes” covering 1950-2000). But due to the sheer volume and scope of the material, the coverage is necessarily rather shallow, and the author’s judgment about what the most important developments were is highly subjective. For example, in the area I personally know best, operator algebras, he hits some of the high points of the subject (with occasionally muddled description) but misses others, instead discussing some developments at length which are generally regarded as peripheral, such as the axioms for a C*-algebra; I doubt that any leading operator algebra specialist would regard the author’s treatment as an adequate or satisfactory history of the subject of operator algebras during the 20th century. (To be fair, he seems to have done a more accurate and thorough job on some other parts of analysis he may be more personally familiar with.)

One of the most interesting parts of the book is the introduction, where the definition, meaning, and scope of the term “analysis” is discussed. The author then takes a fairly broad interpretation of its meaning, including subjects from topology to algebraic geometry under the umbrella of 20th century “analysis.”
Much of the material is presented via extensive quotations from the writing of leading mathematicians. These make for interesting and thought-provoking reading. The book is very helpful in giving the reader a feeling for the overall flavor of twentieth-century analysis, although it is far from a definitive study.

Roberts, King of Infinite Space [Rob06]

A biography of Donald Coxeter. This was a more engaging book than I expected. It is nicely written, and the author has really done her homework, with extensive and trenchant quotes from many leading mathematicians she apparently interviewed. The mathematics is pretty much right, both in detail and philosophical form, especially impressive from an author who does not claim any particular expertise in mathematics; it is conveyed more accurately than by some highly-trained authors who should know better. The general theme of the book, which I think is slightly (but only slightly) overblown, is that classical geometry was a dying subject kept alive primarily by Coxeter before finding a renaissance in such modern areas as computer graphics. A subplot which is greatly oversimplified and rather inaccurate, in my opinion, is that there was a war between formal and visual mathematics in the twentieth century, which is roughly characterized as Bourbaki vs. Coxeter. Bourbaki was indeed a champion of a formal, logical approach to mathematics, but it would be totally wrong to claim that the members of Bourbaki (even Dieudonné, portrayed as the principal villain) did not approve of use of pictures, diagrams, and other visual aids in understanding and creating mathematics; they just felt that the ultimate logical justification for results which may have been found by such methods must be done formally, a sentiment with which even Coxeter would have largely agreed. And it is quite another matter that Bourbaki chose to leave pictures and diagrams almost completely out of their exposition of mathematics, which was never intended to supplant all other approaches. Blaming Bourbaki for the failures of the “new math” is about as misguided as blaming Jesus for present-day fundamentalist Christianity. A new exposition of part of mathematics can never be a negative, except possibly if it is mathematically incorrect; it can only help a competent reader, in combination with other expositions, gain a better understanding of the subject. The problem comes when some people who lack the vision and expertise of the authors decide that a particular work is gospel and should be followed (often with a twisted interpretation due to a lack of understanding) to the exclusion of all others, and a fashion (which mathematics is certainly not immune from) is established. This was the principal problem with the “new math,” and even (as the author points out) with the mathematics curriculum at some leading universities, although it is questionable whether all these problems were even directly a result of the influence of Bourbaki. In any event, it is nice to see that visual mathematics is making a comeback, and while Coxeter was probably not as single-handedly responsible as this book portrays, he was certainly a major influence.

Struik, A Concise History of Mathematics [Str87]

Philosophy of Mathematics

Some of the books in other sections above are partially concerned with the philosophy of mathematics.

Chaitin, Information-Theoretic Incompleteness [Cha92] and The Unknowable [Cha99]

Entertainingly written accounts of Chaitin’s philosophy of mathematics and knowledge, focusing on complexity. The presentation is unusual, intermixing autobiographical passages with the mathematics. It is easy to dismiss the autobiographical parts as egotistical self-promotion (e.g. “Gödel discovered incompleteness, Turing discovered uncomputability, and I discovered randomness”), but, although I have no personal knowledge about the author, I see a certain naiveté about the stories that suggests such a conclusion may
be unfair. The mathematical part is heavy stuff, and I do not feel comfortable with my own understanding of it, but it is certainly a lot of food for thought.

Corfield, Towards a Philosophy of Real Mathematics [Cor03]

This is not the first criticism of the preoccupation of philosophers of mathematics with foundations at the expense of studying its actual practice, but it is the most detailed and articulate one to date. The author’s knowledge and understanding of mathematics is clearly much broader and deeper than that of most other philosophers of mathematics, and is quite impressive. He makes a compelling case against traditional philosophy of mathematics, but I have a hard time figuring out exactly what he proposes as a replacement: he says he is arguing for a Bayesian theory, but spends a good bit of time pointing out the absurdities of a strict Bayesian interpretation of mathematics and science (e.g. that existing evidence cannot help confirm the plausibility of a new theory that explains it). In any event, as evidenced by his choice of title, he clearly thinks that answering the questions he raises is a work in progress. Although this cannot be considered the definitive work on the subject, it is a significant and thought-provoking contribution to the process.

Lakatos, Proofs and Refutations [Lak76]

A thought-provoking discussion of Lakatos’ ideas on the nature of mathematical proof, and by transfer his whole philosophy of mathematics. Written primarily as a dialog based around Euler’s formula for the relation of the number of vertices, edges, and faces of a polyhedron \((V - E + F = 2)\). Interesting reading even for those, like me, who are somewhat skeptical of his conclusions. See [] and [Cor03] for detailed analyses of Lakatos’ work.

Nontechnical and Popular Books

Klein, Elementary Mathematics from an Advanced Standpoint: Arithmetic, Algebra, Analysis [Kle04]

This is not intended to be a popular book, and is not exactly nontechnical. Its goal was twofold: to describe Klein’s ideas on how mathematics should be presented and taught in schools, and to give schoolteachers knowledge and insight on the true nature of the mathematics they teach (he says on p. 162: “. . . the teacher’s knowledge should be far greater than that which he presents to his pupils. He must be familiar with the cliffs and the whirlpools in order to guide his pupils safely past them.”) Although it was written more than a hundred years ago, most of it is as relevant today as when it first appeared, and it is full of valuable insights even today from a first-rate mathematical mind which paid unusually great attention to how mathematics should be taught. I myself learned some things about mathematics and its development, such as a beautiful proof of Taylor’s theorem with Lagrange remainder via difference calculus, regarded by Klein as the best and most natural proof of the theorem (and the way Taylor himself found the theorem, although he did not come close to giving what would be today recognized as a proof), which I have not found in modern texts, probably because difference calculus is rarely discussed except in numerical analysis texts. See () for the argument. And Hilbert in his Grundlagen der Geometrie described something close to the nonstandard real numbers (they may have even been found earlier by someone else). This book is well worth reading and studying by any potential mathematician or mathematics teacher. The reader is cautioned that the English translation in this edition contains an unusually large number of misprints, most of which are harmless.

Livio, Is God a Mathematician? [Liv09]

This is generally a well-written, informative, and entertaining study of the “unreasonable effectiveness of mathematics” as a basis for science and the underlying principles of the universe. There is little actual religious content, mainly just a discussion of the religious beliefs of major figures such as Galileo, Descartes,
Rucker, Infinity and the Mind \[\text{Ruc82}\]

In my opinion, the most interesting of Rucker’s books. Contains accurate and readable, although non-rigorous, treatments of the theory of ordinals and of Gödel’s Incompleteness Theorem.

Smullyan, The Lady or the Tiger? \[?\]

The first half of this book is a collection of Smullyan’s famous logic puzzles. The second half is a delightful treatment of Gödel’s Incompleteness Theorem in the form of a “mathematical novel.”

Smullyan, Forever Undecided \[?\]

This book falls somewhere between The Lady or the Tiger? and Gödel’s Incompleteness Theorems. It is not a technical book, although it covers some symbolic logic, but it is also not nearly as lighthearted as Smullyan’s other popular books. The topics discussed and explained are somewhere between philosophy and mathematics, based on various versions of Gödel’s Incompleteness Theorem.

The two books The Lady or the Tiger? and Forever Undecided are available together in a single paperback volume.

Books to be Wary of

Alexander, Infinitesimal: How a Dangerous Mathematical Theory Shaped the Modern World \[\text{Ale15}\]

The theme of this book is that the mathematical “theory of indivisibles” precipitated an intense battle in the sixteenth and seventeenth centuries, not just among mathematicians but even among political and social theorists. Throughout the book the author tries to characterize the opposition to the theory of indivisibles as a manifestation of a rigid, authoritarian social philosophy, and the belief in the theory of indivisibles as representing ideals of free inquiry and, ultimately, democracy. While this may be somewhat of a stretch, oversimplification, or extrapolation, it certainly has an element of truth, and the historical events are engagingly described in considerable detail in support of the thesis.

I saw and browsed this book in a bookstore and it looked interesting, so I got a copy and began reading it systematically. I could hardly get past the introduction, since it was so full of mathematical misconceptions. The most egregious, affecting the entire thesis of the book, is confusion of the mathematical terms indivisible and infinitesimal (which mean entirely different things – whatever an infinitesimal is, it is not indivisible! Granted, before the late seventeenth century the difference was not well understood even by mathematicians.

and Newton. The author is an astrophysicist and not a mathematician, and his primary interest (at least in this book) is the relation between mathematics and science; but he has deep knowledge and understanding of mathematics and gets it almost all right. There are a few misstatements, such as that “a fully consistent theory of analysis was only formulated in the 1960s” (an apparent reference to nonstandard analysis), and the crucial words finitely describable (technically, recursively axiomatizable) are omitted from the statements of Gödel’s theorems (he is hardly the first author to make this mistake, and in his case it may well be an inadvertent error rather than a misunderstanding). His discussion of the philosophy of mathematics is extensive and intelligently argued, although he repeatedly makes rather unconvincing arguments against Platonism, and he rather too glowingly refers to [\text{LN00}], which in my opinion is hardly more than a work of fiction. He buys a little too much into the “crisis” characterization of the modernist revolution precipitated by the discovery of noneuclidean geometry, but his discussion of this subject is more accurate and intelligent than many others. He does not pretend to have the answers to the main questions he discusses (no one else does either), but makes worthwhile contributions to the discussion.
See [She18] for a detailed study.) In fact, the book probably should have been titled *Indivisible*, since this term is the one causing most of the controversy discussed in the book. The only mention of infinitesimals as mathematicians since Archimedes have almost always thought of them is a few oblique references to the development of calculus, which really has nothing to do with the theory of indivisibles; the only relevance to the book’s thesis is that (early) calculus was based on controversial (although not actually new) ideas which were not rigorously established.

But there are other serious mathematical inadequacies, such as the attempt to describe what the theory of indivisibles means (it essentially means that a line is a set of points, or, what is not exactly the same thing, that a line is “made up” of a collection of points, and similarly a plane figure is “made up” of a collection of parallel line segments, etc., not that a line segment is actually indivisible). The author is a historian, not a mathematician, and it is perhaps unrealistic to expect him to get the mathematics right, but many authors of other books about mathematics have done better with no more formal mathematical training. Perhaps he was simply presenting the idea as it was (mis)understood by people like Renaissance Jesuits as well as many mathematicians of the time, but if so he should have said so and pointed out the misconception. And he never mentioned the resolution of the paradoxes discussed, not found until the twentieth century using measure theory and the difference between countably and uncountably infinite sets (see I.3.1.2. for a discussion). The author seems to subscribe to the common picture of a line as a string of beads, each point having a “next” point, a picture every modern mathematician has had to try to correct in the classroom. This picture was common at one time even among mathematicians: Galileo believed that a line segment was made up of infinitely many infinitely small points separated by infinitely many infinitely small gaps.

Pushing on, I found the rest of the book more interesting, with much detailed history I didn’t know which I assume is accurately portrayed since this should be the author’s forte. The book traces the formation of the Jesuit order of the Catholic Church, largely as a response to the Reformation and resulting loss of influence and control of the Church, and the incorporation of mathematics into the curriculum of the Jesuit colleges and, more generally, as a subject as worthy of study as philosophy and theology. Mathematics as taught by the Jesuits was in a form compatible with overall Jesuit philosophy, as a body of eternal truths and not as an intellectual pursuit in which new ideas were sought or even tolerated. As the author puts it on p. 78:

“But why study mathematics? Only because it provided a model of perfect rational order and certainty, and an example of how universal truths governed the world. If mathematics was to become a field of far-reaching innovation, in which new truths were proposed and then subjected to challenge and debate, then it would be worse than useless. It would be dangerous, as it would compromise the very foundations of the truth it was meant to buttress.”

The Jesuits had enormous influence on the intellectual life of continental Europe in the sixteenth and seventeenth centuries, partially through their colleges, which were in some respects the best system of higher education at the time.

A good deal of the book is devoted to Thomas Hobbes and John Wallis. The mathematical part of their longstanding intellectual battle (there was a considerable nonmathematical component too) is described as being mainly a disagreement about mathematical philosophy. While there is probably some truth to this, I think there is a more important reason. The author does point out that Hobbes’s (later) mathematical work was not taken seriously by the English mathematical establishment, but attributes this to political prejudice. In fact, Hobbes was a hopeless mathematical charlatan, and his mathematical work was no more worthy of being taken seriously than that of modern-day cranks (cf. Figure I.8). Wallis was an excellent mathematician, one of the world’s leading mathematicians of the seventeenth century (although overshadowed in Britain by Newton and Gregory), and very likely had the attitude widespread among mathematicians of not suffering fools gladly. The author may perhaps again be excused for not having the

1962
ability to make a judgment about the quality of Hobbes’s mathematics. The statement on p. 213 that Hobbes was “one of the most respected English mathematicians of the day” [1640] is highly doubtful; I am not aware of a single (correct) original contribution he made to mathematics.

The author’s assertion that Wallis created a new kind of experimental mathematics, and especially that he did so to conform to the philosophy of the Royal Society, is nonsense. All mathematics is experimental when first being developed, and mathematical results have been found by induction since ancient Greece and even earlier. Wallis used techniques going back to Archimedes and probably even Euclid’s predecessors (ancient Greeks apparently did not record how they obtained their results, and in any event knowledge about their methods has been lost). Wallis may not have paid sufficient attention to carefully proving his results, but he undoubtedly could have (and probably did mentally) in many cases using the elaborate double *reductio ad absurdum* technique of Archimedes, but evidently was more interested in using his time and effort to explore new things. He was surely aware, as contemporaries like Newton were, that some of the techniques did not (at that time) have a sound logical footing, but these mathematicians had strong intuition and their results invariably turned out to be correct and logically justifiable (unlike Hobbes). It seems farfetched that Wallis declined to try to prove his results simply in order to adhere to the Royal Society philosophy. The author seems to be obsessively insistent on attributing the primary motivation for every mathematical or scientific study to adherence to one side or the other of the philosophical dichotomy which is the theme of the book.

Incidentally, Wallis did not initiate the idea that mathematical objects exist and only need to be explored, and that mathematical truths are discovered; this philosophy of mathematics is called Platonism since it goes back (at least) to Plato. (Hobbes was also a Platonist and, contrary to what is suggested in the book, Euclid probably was too.)

There is at least one historical inaccuracy I noticed. On p. 290, it is stated:

“Following the war over the infinitely small, advanced mathematics in Italy came to a standstill, whereas English mathematics quickly became one of the dominant national traditions in Europe, rivaled only by the French.”

This may be an accurate assessment about Italy. But although England in many respects, including technology, industrialization, and certain intellectual and social areas, became a world leader, it notably remained a mathematical backwater for nearly 200 years after Newton, particularly in analysis. In fact, there was a nineteenth century analysis renaissance in Italy before there was one in England. It was French, Swiss, and German mathematicians, not English, who were at the forefront of analysis in the eighteenth and nineteenth centuries (I count Lagrange as French although he was actually born in Italy).

There have been criticisms of the reasons the author attributes for the Jesuit opposition to the theory of indivisibles, that it is inconsistent with Aristotle and Euclid and the notion that mathematics represents eternal truth. In [She18] (which I find an articulate criticism of much of Alexander’s book, although I cannot evaluate the historical aspects), it is argued that the principal reason for the opposition was that it contradicted the Catholic doctrine of transubstantiation of Christ’s body and blood in the Eucharist (this idea was first presented in [Red87]). Note that I only discovered and read these other references after I had written this review.

It is remarkable that so many people, both mathematicians and nonmathematicians, spent so much time and effort and generated such acrimony over a dispute of such little ultimate (mathematical) consequence, with both sides advancing arguments which today would be regarded as foolishly misguided. I wish the author had been able to convey the obsolescence better. The book is worth reading for its historical content, as long as the author’s attributions and conclusions are taken with a degree of skepticism.
Bell, Men of Mathematics [?]

This book is entertaining reading, but unreliable as history; some parts are downright fanciful. The chapters on Galois and Cantor are the worst from the standpoint of historical accuracy. Although Bell was a respectable mathematician, some critics have asserted that as a biographer he seems to have been the type of writer who never lets the facts get in the way of a good story. Overall, however, the book is not nearly as bad as these critics have made it out to be: the vast majority of the book is factually accurate, and the errors appear to be entirely the result of careless or inadequate research and unwarranted extrapolation rather than deliberate falsification; the accounts are much closer to the actual events than in many modern “historical” movies. In fact, the author has gone to some pains to qualify historical assertions he felt were not adequately established. Despite the title, the book does not have a sexist tone, and in fact was fairly progressive for its time (1937), giving sympathetic treatment to several women mathematicians and the obstacles they faced; for example, Bell gave the first popular (if dated and flawed) account of Sofia Kovalevskaya. There is some gratuitous antisemitism in the chapter on Cantor (there is controversy over to what degree, if any, Cantor himself was Jewish, although he clearly had at least some Jewish ancestry). This book may be inspiring to a young reader who is aware of its historical inaccuracies, and it also serves a useful purpose in countering the stereotype of mathematicians as nerds of borderline sanity and little personality or variety. The description of Bell in [New88] seems fairly apt: “lively, stimulating, inoffensively crotchety and opinionated, with a good sense of historical circumstance, a fine impatience with humbug, a sound grasp of the entire mathematical scene, and a gift for clear and orderly explanation.”

Griffiths and Hilton, A Comprehensive Textbook of Classical Mathematics [GH78]

This book is hard to categorize. The aim is to give a unified exposition of “classical mathematics” (or what may be loosely described as “undergraduate mathematics”) from a modern perspective: logic and set theory, abstract algebra, elementary geometry and topology, and undergraduate analysis. Some things (generally simpler ones) are proved in detail, but most of the deeper or more difficult results are just stated with some informal or intuitive justification given (which is clearly represented as such; the book is pretty honest mathematically). I would not describe the treatment as inspired, and it is generally pretty conventional, but there is much wisdom and insight in the descriptions. And, importantly, the authors explicitly describe, and follow, what they call the spiral principle, which underlies both the historical development of mathematics and the way it is, or should be, taught: fundamental topics and principles are first introduced and developed in a natural and intuitive way, and after they have become somewhat familiar are redeveloped in a more careful way. This is nicely summarized in this statement from p. 249, in the chapter “The Logic of Geometry”, after Euclid’s geometry (with modern improvements) is described:

“A person might object to all this procedure on the grounds that the definitions given are artificial without a prior knowledge of Euclidean geometry; to which we can only reply that once they are given, a logically watertight system of geometry can be developed, and it matters not a whit how the definitions happened to be suggested. This is typical of the way in which a great deal of mathematics has grown. First a terminology and body of results is obtained by logically unsatisfactory (but often aesthetically pleasing) methods. Then the theory is seen to have flaws which prevent further progress. Finally the theory is reset in a possibly different language, and derived in a correct logical manner, usually gaining greater power.”

On the other hand, the book contains surprisingly many (especially for an apparently revised second edition) missteps, most of which are fairly minor and many of little real consequence; I think at least nearly all of them could be described as “careless.” For example, both the statement of the Implicit Function
Theorem (37.4.2) and Theorem 15.5.2 are false as stated; additional hypotheses are needed (continuity of the relevant derivatives suffices). In contrast, the statement of the Mean Value Theorem, or more accurately of “Rolle’s conditions” on p. 483 (ROLLE never stated, and probably never considered, such conditions), contains a superfluous continuity assumption on the derivative, which is not used in the proof or mentioned when the theorem is quoted in several other places. It is stated that the definition of “finite set” in 7.2.4 (what is usually called Dedekind-finite) is equivalent to the one in 7.1.4 (apparently meaning 2.11.1), but some form of the Axiom of Choice, which is only mentioned later, is needed. (Theorem 7.1.4 also has an obvious superfluous assumption.) And the authors seem to like to invent their own terminology, which is sometimes at odds with convention. Some of this invented terminology is picturesque, such as their name “mostest theorem” for what is normally called the Extreme Value Theorem, but some is not an improvement, such as “flat” for “affine subspace” or “laws of indices” for “rules of exponents” (maybe some of their terms are in more general use in Britain than in the U.S.). And what they call the “Axiom of Choice” (that an element can be chosen out of one nonempty set) is either trivial or a nonmathematical assertion depending on one’s point of view (cf. II.6.1.6.); what is usually called the Axiom of Choice is called the “Strong Axiom of Choice”.

This book could be quite helpful to an advanced undergraduate or knowledgeable nonmathematician in helping to “put the pieces together”, somewhat in the manner of [Sch97] (although the books are very different), as long as it is recognized that the small inaccuracies mean that the details of the exposition must be regarded with a degree of skepticism. But, although this book is described and clearly intended as a textbook, I find it hard to think it could be suitable as a textbook for a course it would be worthwhile for a mathematics student to take; it might be suitable for a “capstone” course for students preparing to be math teachers at the high-school or community college level.

Guillen, Bridges to Infinity [Gui83]

This book is not entirely about mathematics, and I will not pass judgment on the nonmathematical part. But it is so riddled with mathematical errors and misconceptions that the credibility of the entire work is questionable. There are also sloppy historical errors, such as calling von Neumann a “German mathematician.”

Lakoff and Nuñez, Where Mathematics Comes From [LN00]

This book purports to be a study of how the human mind perceives and understands mathematics. But the authors, a linguist and a cognitive psychologist, repeatedly demonstrate an embarrassing lack of understanding of basic undergraduate mathematics (e.g. limits); they would be well advised to study a book like Closer and Closer to see what they are missing. They regard the equation $e^{i\pi} = -1$ as “one of the deepest equations in all of mathematics,” and suggest that even most modern mathematicians do not understand it. There is also much discussion of the philosophy of mathematics, which is shallow and logically flawed, sometimes with a vicious undertone. While I cannot evaluate the authors’ cognitive psychology arguments, this is not a book that mathematicians should take very seriously. See () for a more extensive review.

Polya, Induction and Analogy in Mathematics [Pól90a] and Patterns of Plausible Inference [Pól90b]

This two-volume work is in many ways an admirable book. It is a thorough discussion of mathematical reasoning and problem solving by one of the masters, and it is filled with solid and valuable advice on how to approach and solve mathematical problems. The techniques are illustrated by numerous examples, containing clever and insightful arguments (originally dating back as far as ancient Greece) from which much can be learned.
But the book is unfortunately riddled with what I call “mathematical dishonesty.” Many of the arguments represented as proofs are logically incomplete or nonrigorous (see 1.5.1.5. for a discussion of some of these; earlier, he seems to represent the principle of induction as so self-evident that it merits no comment.) This is a much more serious problem in a book of this sort than in books like [Doo94] or [Mor09] (where there are gaps in arguments but no attempt is made to hide them, and the gaps are generally routine to fill in), since the main subject of this book is mathematical reasoning. Polya himself certainly knew better. Some rather minor changes in the language used, and a few added comments, could have maintained mathematical honesty while still making the main points the author wanted to convey. But he seemed to deliberately go out of his way to obscure or gloss over some of the mathematical difficulties in his arguments, and as a result many readers could be misled about the nature of full demonstrations in mathematics.
Bibliography


1969

[Bla] Bruce Blackadar. A family of mutually nonhomeomorphic separable contractible 2-manifolds of cardinality $2^\omega$. \url{http://arxiv.org/abs/1505.01441}.


1975


1979


1980


1985


1989


1991


1994

1995
# Index

\(A \otimes \mathbb{K}, 1889\)  
\(A^{-1}, 1825\)  
\(A_+^\times, 1873\)  
\(A_{o}^{-1}, 1825\)  
\(A_{\infty}, 1863\)  
\(C_{0}(X), 1862\)  
\(M_{\infty}(A), 1889\)  
\(M_{n}(A), 1889\)  
\(W(T), 1829\)  
\(GL_{1}(A), 1825\)  
\(GL_{1}(A)_{\circ}, 1825\)  
\(\text{Irr}_{\mathbb{H}}(A), 1888\)  
\(\mathbb{K}, 1861\)  
\(\Lambda_{I}, 1876\)  
\(\text{Prim}(A), 1886\)  
\(\text{Prim}^{-1}(A), 1887\)  
\(\text{Prim}_{J}(A), 1887\)  
\(\mathbb{M}_{n}, 1861\)  
\(\mathcal{H}^{\infty}, 1803\)  
\(\mathcal{H}^{n}, 1803\)  
\(\mathcal{H}_{\phi}, 1883\)  
\(\mathcal{P}(A), 1882\)  
\(\mathcal{S}(A), 1880\)  
\(\bar{x}, 1887\)  
\(\ell^{2}, 1786\)  
\(\ell^{2}(S), 1786\)  
\(\Lambda, 1888\)  
\(\pi_{\phi}, 1883\)  
\(\sigma(T), 1827\)  
\(\sigma_{A}(x), 1816\)  
\(\xi_{\phi}, 1884\)  
\(e^{x}, 1825\)  
\(\text{exp}(A), 1825\)  
\(u(T), 1829\)  
\(x_{+}, 1873\)

\(x_{-}, 1873\)

Abel’s Test, 471  
Abel’s Theorem, 678  
\textsc{Abel, Niels Henrik}, 274, 1932  
abelian group, 274  
abelian semigroup, 274  
absolute \(G_{\delta}, 1102\)  
absolutely continuous function, 1712  
absolutely convergent infinite series, 459  
abstract simplicial complex, 1347  
abstraction, 11  
AC, 161  
accumulation point, 1014  
\text{of a subset of } \mathbb{R}, 723  
adaptive procedure, 1548  
Addison, John, 1391  
additivity  
\text{countable}, 1393  
\text{finite}, 1394  
adele, 1183  
a.e., 1400  
affine map  
\text{of convex sets}, 1728  
affinely independent, 394  
Al-Khwarizmi, 1927  
Aleksandrov, Pavel, 1390  
alpheh, 186  
Alexander horned sphere, 1244  
algebra  
\text{of functions}, 1377  
algebraic number, 158, 331, 343  
algebraically independent numbers, 343  
amost disjoint, 1400  
amost everywhere, 1400, 1519

1996
alternating series, 463
Alternating Series Test, 464
amplification
  of a Hilbert space, 1803
  of a representation, 1878
analysis, 2
  complex, 300
analytic continuation, 969
analytic dimension, 1305
analytic function, 683
  at a point, 699
  complete, 964
  complex, 683
  real, 683
  total, 963
analytic function element, 963
analytic set, 1386–1392
analytic subring, 1304
annihilator, 1915
annulus, 940
antiderivative, 601
Antiderivative Theorem, 914
  for convex sets, 913
Antoine’s necklace, 1248
ANTOINE, LOUIS, 1245
A-operation, 1432
applied mathematics, 60
approximate midpoint property, 776
approximate unit, 1875
  almost idempotent, 1875
  continuous, 1875
  idempotent, 1875
  sequential, 1875
approximation property, 1837
arc, 770, 795, 1161, 1244
arc length, 766, 784
Archimedean property, 238
ARCHIMEDES, 35, 49, 1483
Archimedes’ cattle problem, 49
arcsine, 572
arcwise connected space, 1159, 1161
area
  of a circle, 576
ARISTOTLE, 44
ARMSTRONG, NEIL, 71
Artin-Schreier Theorem, 278
ascent, 1835
ASKEY, RICHARD, 56
associative law, 215, 254
ATTYAH, MICHAEL, 60, 82
Atkinson’s Theorem, 1842
atom, 201, 365, 1404
atomic statement, 102
atomless measure, 201
attaching, 1081
authority, 9
autonomous differential equation, 878
average value
  of a function, 1527
axiom, 11, 12, 214
Axiom of Choice, 161, 172, 176–177, 188
  countable, 165
  for finite sets, 167
Axiom of Constructibility, 183
Axiom of Countable Choice, 165
Axiom of Dependent Choice, 165
Axiom of Determinacy, 164
Axiom of Projective Determinacy, 164
Axiom of Regularity, 172
C*-algebra, 1861
abstract, 1861
stable, 1889
C*-axiom, 1861
calculus
multivariable, 796
vector, 796
cancellation, 255, 258, 271
canonical embedding, 1741
Cantor set, 1177, 1227, 1248, 1382, 1651
of positive measure, 1382
Cantor-Bendixson Theorem, 1020
Carathéodory extension process, 1414–1434
Carathéodory Extension Theorem, 1428
Carathéodory’s Theorem, 1420
Carathéodory, Constantin, 1421
Cardano, Hieronimo, 75
cardinal, 184–197
aleph, 186
continuum, 153
inaccessible, 193
infinite, 189
infinite in the sense of Dedekind, 189
κ-additive, 198
large, 193
limit, 186
Mahlo, 202
measurable, 204
Ramsey, 207
real-valued measurable, 199
regular, 192
singular, 192
strongly inaccessible, 193
successor, 186
weakly inaccessible, 192
cardinal invariant
of a topological space, 1007
Cartan, Henri, 1034, 1116, 1932
Casorati-Weierstrass Theorem, 955
catalog, 1652
Cauchy Criterion
for infinite series, 434
Cauchy integral, 1483–1485
Cauchy Integral Formula, 916
for derivatives, 919
Cauchy product of infinite series, 494
Cauchy sequence, 288
Cauchy’s Condensation Test, 456
Cauchy’s Inequalities, 920
Cauchy’s Test, 447
Cauchy’s Theorem, 925
for convex sets, 915
for simply connected sets, 926
homology version, 927
CAUCHY, AUGUSTIN-Louis, 6, 288, 508, 590, 591, 700, 1481, 1781
Cauchy-Lipschitz Theorem, 882
Cauchy-Riemann equations, 901
Cauchy-Schwarz Inequality, 1781
CauLIERI, BonAventuRDa, 1537
Cayley number, 308
CBS Inequality, 1781, 1849
CBS inequality
for states, 1881
ČECH, EDUARD, 1127
cell, 1338
cell complex, 1345
centered collection of subsets, 1034
Chain Rule
one variable, 535–538
chained arrow notation, 183
Chaitin, Gregory, 39, 48
Champernowne’s constant, 337
Chapman, C. H., 78
character
of a locally compact abelian group, 1914
characteristic
of a field, 277
characteristic function, 1489
Chebyshev nodes, 645
Chebyshev polynomial, 647
Chebyshev quadrature, 1541
Chebyshev, Pafnuty, 644
Chernavskii, A. V., 854
Chernoff, Paul, 1116
Chevalley, Claude, 1912
Choice, 161
Choquet’s Theorem, 1760
CHRÉTIEN, JEAN, 40
Church, Alonzo, 92
1999
circular function, 578

circumference, 17

Clairaut’s Theorem, 807
classical adjoint, 408
clopen set, 1150
closed face, 1752
Closed Graph Theorem, 1747
  for locally compact groups, 1902
closed interval, 236
closed manifold, 1265
closed set, 1013
  in $\mathbb{R}$, 721
closed statement, 108–109
closer and closer, 417
closure
  of a subset of $\mathbb{R}$, 724
closure of a set, 1014
closure operation, 1018
cluster point
  of a net, 1027
  of a sequence, 1023
cobounded filter, 205
cofinal condition, 145
Cohen, Paul, 157
coisometry, 1863
Cole, Frank N., 109
Cole, K. C., 11
collapsing, 1073
column vector, 371
comeager set, 1142
commutative law, 215, 254
compact group, 1900
compact manifold, 1265
compact operator, 1835
compact representation, 1127
compact space, 1008, 1107–1118
compact support, 1125
compact-open topology, 1193
compactification, 1126–1130
  one-point, 1121, 1139
  Stone-Čech, 1127–1130
  Wallman, 1130
compactly generated, 1135
Compactness Theorem of model theory, 166
comparable elements, 135

Comparison Test, 440, 463
  Extended, 441
  General Limit, 444
  Limit, 443
complement, 1355
complete analytic function, 964
complete induction, 139–140, 224–225
complete measure space, 1399
complete topological group, 1899
completely continuous operator, 1835
completely Hausdorff space, 1059
completely normal space, 1093
completely regular space, 1053–1054
completely separated, 1741
Completeness Axiom, 136, 234, 294
completion
  of a measure space, 1400
  of a pre-Hilbert space, 1787
  of a topological group, 1899
  of an inner product space, 1786
complex
  cell, 1345
    simplicial, 1345
complex analysis, 300
complex differentiable, 900
complex number, 300–308
complex plane, 302
component, 1154
composite quadrature procedure, 1536
compression
  of a measure, 1406
computable function, 28
concatenation, 764, 782
concave function, 596
conditionally convergent infinite series, 459
cone, 1341
  over a topological space, 1074
conjecture, 47
conjugate exponent, 1847
connected
  simply, 959
connected component, 1154
connected space, 1150
connective, 100
constant, 100
Constant Multiple Rule, 531
Daniell, P. J., 1492
Darboux continuous function, 595
Darboux property, 595
Darboux sum, 1499
interior, 1501
Darboux, Jean-Gaston, 1484
Da vis, Martin, 250
Da vis, Philip J., 660
DC, 165
de Leeuw, Karel, 1763
de Moivre, Abraham, 661
decadent rational number, 317
decimal expansion, 315–320
decomposable measure, 1402
decomposition dimension, 1299
Dedekind cut, 279
Dedekind, Richard, 157, 189, 245, 247, 279
Dedekind-finite set, 149
definition, 114
deleted neighborhood
of a point of $\mathbb{R}$, 723
Denjoy integral, 1493
Denjoy, Arnaud, 561
Denjoy-Young-Saks Theorem, 561
dense, 1015
dense-in-itself, 1020
density
of a topological space, 1017
Denver, John, 52
Dependent Choice, 165
derangement, 473
derivation, 986
derivative
Dini, 558–562
directional, 809
higher, 541
one-sided, 557–558
partial, 802
second, 541
derived set, 1020
higher, 1021
Descartes, René, 42, 57, 94
Descending Chain Condition, 172
descent, 1835
descriptive set theory, 1652
determinant, 396–410
expansion by minors, 401–402
geometric interpretation, 404–408
multiplicative property, 402–403
of a linear transformation, 404
diagonal property, 1024
diagonalization argument, 152, 1654
Dieudonné, Jean, 82, 1130
difference of sets, 1355
difference quotient, 526
differences, 552
differential equation, 876
autonomous, 878
ordinary, 876
partial, 876
differential field, 989
differential function field, 991
differential homomorphism, 998
differential ideal, 999
diffomorphism, 998
dilation
of a measure, 1406
dim, 1295
dimension
analytic, 1305
covering, 1295
decomposition, 1299
essential, 1286
large inductive, 1292
linear (vector space), 385
of a cell, 1338
of a complex, 1345
orthogonal, 1796
partition, 1298
small inductive, 1290
dimension-counting theorem, 393
dimensional kernel, 1313
Dini derivative, 558–562
Dini’s Theorem, 1141
Dini, Ulisse, 558, 828
direct limit
of sets, 1166–1170
of topological spaces, 1174–1175
direct sum
of Hilbert spaces, 1803
direct system, 1166
of sets, 1166
of topological spaces, 1174
directed equivalence, 145
directed function, 145
directed set, 137, 144–147
directional derivative, 809
Dirichlet’s Test, 471
discontinuity
  jump, 1665
discrete metric, 1004
discrete topology, 1004
dishonesty, 94
disk of convergence, 933
distributive law, 216, 254
Divergence Criterion, 434
divergent sequence, 414
divergent series, 432
divided difference, 633
  higher-order, 634
division ring, 307
domain of discourse, 103, 117
Domain-Straightening Theorem, 855
Dominated Convergence Theorem
  nonnegative version, 1622
  dominating function, 1622
dot product, 394
double sum of an array, 483
Dowker’s Theorem, 1297
downward continuity, 1397
drastic revision
  of an infinite series, 490
  of an unordered sum, 491
dual group, 1914
dyadic rational number, 278
\(e\), 566–568
disk of convergence, 933
distributive law, 216, 254
Divergence Criterion, 434
divergent sequence, 414
divergent series, 432
divided difference, 633
  higher-order, 634
division ring, 307
domain of discourse, 103, 117
Domain-Straightening Theorem, 855
Dominated Convergence Theorem
  nonnegative version, 1622
  dominating function, 1622
dot product, 394
double sum of an array, 483
Dowker’s Theorem, 1297
downward continuity, 1397
drastic revision
  of an infinite series, 490
  of an unordered sum, 491
dual group, 1914
dyadic rational number, 278
\(e\), 566–568
disk of convergence, 933
distributive law, 216, 254
Divergence Criterion, 434
divergent sequence, 414
divergent series, 432

epsilon-
star, 1342

equiconsistent, 207

equipotent, 148
equivalence
  of representations, 1878
equivalence class, 132
equivalence relation, 131
Ermakoff, see Yermakov
error, 41
essential dimension, 1286
essential map, 1274
essential singularity, 954
essential subspace, 1879
Euclid, 29, 230, 264
Euclid’s Parallel Postulate, 29
Euclidean Four-Point Axiom
  Weak, 750
Euler, Leonhard, 84, 87, 335, 339, 344, 609, 654, 1485
Eves, Howard, 35
Existence and Uniqueness Theorem
  for first-order ODEs, 882
    global version, 885
    local version, 884
    standard version, 885
  for higher-order ODEs
    global version, 889
    local version, 888
    standard version, 888, 895
  for systems of first-order ODEs
    global version, 887
    local version, 887
    standard version, 886, 895
existential quantifier, 106
exponential
  in a differential field, 992
exponential function, 563–571
exponents
  rules of, 563
exposed point, 1765
Extended Banach Fixed-Point Theorem, 744

2003
Extended Comparison Test, 441
Extended First-Order Transfer Principle, 364
extended metric, 733
extended simple language, 121
Extended Simple Transfer Principle, 364
extended topologist’s sine curve, 1224
extension
  of a measure, 1406, 1407
Extension Axiom, 747
Extension Theorem, 1426
extrapolation, 632
Extrapolation Property, 749
extreme point, 1751
Extreme Value Theorem, 313, 521

Faà di Bruno’s formula, 542, 696
Fabius function, 708
face, 1339, 1751
  closed, 1752
Face Partition Theorem, 1275
factorial polynomial, 642
Fatou’s Lemma, 1620
Fermat’s Last Theorem, 344
Fermat, Pierre de, 335
Field, 218, 276, 370
  differential, 989
  differential function, 991
  formally real, 278
  nonarchimedean ordered, 295, 353
  of fractions, 278
  ordered, 220, 276, 352
  quotient, 278, 296
  real-closed, 245, 278, 297
  skew, 307
Field axioms, 218
Filter, 205, 1034, 1365
  cobounded, 205
  finer, 1036
  fixed, 1366
  free, 1366
  κ-complete, 205
  neighborhood, 1034
  normal, 206
  prime, 1367
Filter Extension Theorem, 1366
Filterbase, 1034
  maximal, 1037
  subordinate, 1036
finite additivity, 1394
finite ascent, 1835
finite complement topology, 1006
finite descent, 1835
Finite Intersection Property, 1108, 1366
finite measure, 1401
finite set, 149
finite subadditivity, 1395, 1415
Finite-Dimensional Spectral Theorem, 1845
finitely additive measure, 1394
Finslerian length structure, 778
first category set, 1142
first countable space, 1006
First Decomposition Theorem, 1294
First Derivative Test, 593
first differences, 552
first-order logic, 103
first-order statement, 103
first-order theory, 103
First-Order Transfer Principle, 363
fixed filter, 1366
fixed-point theorem
  Banach, 742
  extended, 744
  Brouwer, 1272–1282
foot, 751
formal Laurent series, 296
formal power series, 295
formal proof, 40
formal statement, 100–117
formalism, 76–82
Forsyth, A. R., 62
forward differences, 642
Four-Color Theorem, 86, 1224
Fourier coefficient, 1792
Fourier series, 508
Fourier, Joseph, 339
Fraenkel, Abraham, 171
Franklin, Benjamin, 62
Fréchet, Maurice, 1428
Fréchet ordering, 180, 295
Fréchet-Urysohn space, 1038
Fréchet-von Neumann-Jordan Theorem, 1783
Fredholm Alternative, 1840

2004
Fredholm index, 1837
Fredholm operator, 1837
Fredholm, Irav, 1834
free cyclic group, 265
free filter, 1366
free variable, 106
Frege, Gottlob, 8, 75, 76
\( F_\sigma \), 1371, 1382
Fubini’s Differentiation Theorem, 1321
Fubini’s Series Theorem, 1668
Fubini’s Theorem, 1321
continuous form, 1574
for double summations, 487
product form, 1634
Riemann integral form, 1573
upper and lower form, 1572
Fubini, Guido, 488, 1321, 1668
full partition, 1586
full real-valued measure, 198
fully normal, 1139
fully separated, 1741
function
analytic, 683
at a point, 699
arcsine, 572
Borel, 1361
Borel measurable, 1372–1380, 1652
characteristic, 1489
circular, 578
convex, 1673
cosecant, 574
cosine, 573
cotangent, 574
Darboux continuous, 595
directed, 145
exponential, 563–571
Fabius, 708
hyperbolic, 578–582
increasing, 1665
indicator, 1489
inverse hyperbolic, 580
inverse trigonometric, 575–576
log-convex, 1692
logarithm, 565
measurable, 1361
midpoint-convex, 1682
monotone, 1665
multiplicity, 1380
natural logarithm, 567
nondecreasing, 1665
nonincreasing, 1665
rational, 278, 294, 409
regulated, 606, 1594
secant, 574
simple, 1489, 1598
simplicial, 1346
sine, 573
step, 601
strictly increasing, 1665
strictly monotone, 1665
tangent, 574
trigonometric, 571–582
functional calculus
continuous, 1870–1871
holomorphic, 1824–1825
functionally dependent, 821
functionally Hausdorff spaces, 1059
functionally independent, 821
Fundamental Principle of Infinite Series
First, 431
Fundamental Principle of Nonnegative Series, 440
fundamental system of neighborhoods, 1894
Fundamental Theorem of Algebra, 306, 906, 981, 1281
Fundamental Theorem of Calculus, 230, 264
Fundamental Theorem of Calculus, 432
general HK version, 1640
HK version, 1639
Version 1, 1523
Version 2, 1526
Fundamental Theorem of Interval Length, 238
fuzzy logic, 74
galaxy, 733
Galileo, 10, 45, 149
Gall, Franz Joseph, 1933
Galois, Évariste, 1932
Galois’s Lemma, 348
Gauss’s Lemma, 348
Gauss’s Test, 456
2005
GAUSS, CARL FRIEDRICH, 29, 47, 223, 465, 659
Gauss-Kronrod integration, 1548
Gaussian quadrature, 1540
G¹, 1101, 1371, 1382
Gelfand transform, 1869
Gelfand’s Theorem, 1865
Gelfand-Mazur Theorem, 1822
Gelfand-Naimark Theorem, 1885
GELFOND, ALEKSANDR, 344
Gelfand-Schneider Theorem, 344
General Complex Analytic Implicit/Inverse Function Theorem, 861
General Implicit/Inverse Function Theorem, 834
General Limit Comparison Test, 444
general position, 393
General Real Analytic Implicit/Inverse Function Theorem, 860
generalization, 854
Generalized Continuum Hypothesis, 192
generic set, 1142
gerometric realization, 1347
gemetric series, 435
germ, 975, 976
GLEASON, ANDREW, 1912
GLOECKNER, HELGE, 1905
gling, 1075
gling scheme, 1089
GNS construction, 1883–1884
GNS representation, 1883
GÖDEL, KURT, 78, 152, 157, 171, 1935
Gödel constructible universe, 183
GOETHE, 77
GOLDBBLATT, ROBERT, 82, 351
golden ratio, 335
good approximation to an irrational number, 328
GORDAN, PAUL, 74
GOSPER, R. WILLIAM, 664
Goursat’s Theorem, 911
gradient, 818
Gram-Schmidt procedure, 1797
greatest lower bound, 136, 233, 281
GREENBERG, MARVIN J., 61
GREGORY, JAMES, 622, 635
Gregory-Newton interpolation formula, 642
grid, 1553
Grothendieck group, 274
GROTHENDIECK, ALEXANDRE, 211, 1935, 1947, 1957
group, 265
abelian, 274
compact, 1900
cyclic, 265
dual, 1914
free cyclic, 265
Grothendieck, 274
locally compact, 1900
locally Euclidean, 1912
group action, 1085, 1895
faithful, 1085
grouping of an infinite series, 438
Haar measure, 1907
HADAMARD, JACQUES, 61, 81
Hahn Series, 297
HAHN, HANS, 1428, 1476
Hahn-Banach Theorem
   Version 1, 1737
   Version 1A, 1737
   Version 2, 1740
half-open interval, 236
HALMOS, PAUL, 1360
Hamel basis, 383
HAMILTON, WILLIAM ROWAN, 307
handles, 1082
Hanf-Tarski Theorem, 204
HARDY, G. H., 44, 89
harmonic analysis, 1914
harmonic series, 441
HART, W. D., 68
Hartog’s Theorem, 977
Hausdorff Maximal Principle, 163
Hausdorff metric, 1104–1105
Hausdorff space, 1046, 1049
HAUSDORFF, FELIX, 1371, 1750, 1931
Hausdorffization, 1069
Hawaiian earring, 1226
HEATH, THOMAS L., 31
Heaviside function
   smooth, 701
   of a topological space, 1021
Helly space, 1137
Helmholtz, H., 62
Henstock-Kurzweil integral, 1493, 1638
Herbrand, Jacques, 1932
hereditarily indecomposable continuum, 1239
hereditarily normal space, 1093
hereditary quotient map, 1071
hereditary set, 279
Hermite polynomial, 641
Hermite, Charles, 339, 344
Hersh, Reuben, 81
higher derivatives, 541
higher differences, 552
higher-order logic, 103
higher-order theory, 103
Hilbert cube, 1651
Hilbert space, 1785
real, 1803
Hilbert, David, 29, 30, 74, 77, 148, 344, 1080, 1786
Hippasus, 44
HK integral, 1638
Hobbes, Thomas, 1931
Hölder, Otto, 659
Hölder's Inequality, 1849
generalized, 1858
Holmes, Sherlock, 37, 43
holomorphic function, 900
holomorphically simply connected, 960
homogeneous space, 1087, 1894
homology
of contours, 924
homomorphism
differential, 998
homotopic, 1315
homotopy, 1315
of contours, 921
Homotopy Theorem for Contour Integrals, 921
honesty, 94
hull-kernel topology, 1058
Hurwitz's Theorem, 957
Hurwitz, Adolf, 341
hyperbolic function, 578–582
inverse, 580
hypergeometric series, 465–468
hypernatural numbers, 358
hyperreal numbers, 354
hyperspace, 1104
ideal, 170, 273, 1862
differential, 999
maximal, 170
prime, 170
primitive, 1886
principal, 170
identification space, 1073
if and only if, 105
image, 127, 129
imaginary part
of an element of a $C^*$-algebra, 1863
implication, 104–105
Implicit Function Theorem
complex analytic version, 860
continuous version, 830
real analytic version, 859
smooth version, 830
Implicit/Inverse Function Theorem
General, 834
General Complex Analytic, 861
General Real Analytic, 860
impossibility theorem, 44
improper Riemann integral, 1530
inaccessible cardinal, 193
incomparable elements, 135
increasing function, 1665
Ind, 1292
ind, 1290
indecomposable continuum, 1239
index
Fredholm, 1837
indicator function, 1489
indiscrete topology, 1004
induction, 222–227, 251–252
complete, 139–140, 224–225
transfinite, 138–141, 175–176
inductive limit, 1167
inductive limit topology, 1174
infimum, 136, 233, 281
infinite broom, 1178
infinite comb, 1255
infinite ladder, 1259
infinite product
of numbers, 496–501
infinite regress, 225–226
infinite series
of numbers, 431–507
absolutely convergent, 459
alternating, 463
complex, 501–502
conditionally convergent, 459
convergent, 432
divergent, 432
hypergeometric, 465–468
nonnegative, 439
infinite set, 149
infinitesimal, 295, 296, 350, 352
initial value problem, 878
inner measure, 1431
integer, 227–230, 265–274
integrable function, 1623
nonnegative, 1604
integral
Cauchy, 1483–1485
Daniell, 1492
Denjoy, 1493
Henstock-Kurzweil, 1493, 1638
improper Riemann, 1530
iterated, 1572
Lebesgue, 1487–1492, 1615
Perron, 1493
Riemann, 6–7, 1481–1482, 1486, 1496–1596
in \( \mathbb{R}^n \), 1552–1585
integral domain, 274
Integral Test, 447
integration, 6–7, 1481–1493
interior, 1017
of a cell, 1339
interior Darboux sum, 1501
interleaving, 1029
Intermediate Value Property, 595
Intermediate Value Theorem, 519, 1152
interpolating polynomial, 632
Lagrange form, 633
Newton form, 635
interpolation, 632
Interpolation Property, 749
interpolatory quadrature procedure, 1536
interpretable, 119
intersection, 1355
interval, 236
closed, 236
half-open, 236
open, 236
interval of convergence, 675
intrinsic metric, 771
intuitionism, 73–76
Invariance of Domain Theorem, 1204, 1289
Inverse Function Theorem
complex analytic, 904
complex analytic multivariable, 861
multivariable, 834
real analytic multivariable, 860
inverse hyperbolic function, 580
inverse image, 127, 129
inverse limit, 1175
of sets, 1170–1174
of topological spaces, 1175–1180
inverse limit topology, 1175
inverse system
of sets, 1170
of topological spaces, 1175
inverse trigonometric function, 575–576
inversion, 1893
involution
Banach algebra, 1861
irrational number, 44, 245, 339–349, 570
isolated point, 1014
of a subset of \( \mathbb{R} \), 723
isolated singularity, 949
isometry
in a C*-algebra, 1863
isomorphism
Borel, 1361
of convex sets, 1728
of vector spaces, 379
isotopy, 1319
iterated integral, 1572
iterated sum of an array, 485
IVANOV, SERGEI, 1331
IVP, 878
IVT, 519
JABUKA, STANISLAV, 1009
Jacobson topology, 1058
Jensen’s inequality, 1691

2008
Jones’s Lemma, 1056
Jones, Vaughan, 25
Jones, William, 1930
Jordan content, 1562
Jordan curve, 1201
Jordan Curve Theorem, 1202
in $\mathbb{R}^n$, 1203
Jordan region, 1562
Jordan, Camille, 1202
Jordan-Schönflies Theorem, 1203
jump discontinuity, 1665
Junior Monotone Convergence Theorem, 1602
Junior Tonelli Theorem, 1632
Kaczynski, Ted, 1935
$\kappa$-additive cardinal, 198
$\kappa$-complete filter, 205
Kasner, Edward, 149
Kelley, John, 1000
Kepler, Johannes, 1537
Khintchine, Aleksandr, 330
Klein bottle, 1078
Klein, Felix, 49, 80
Knaster bucket-handle, 1180, 1235
Knéebone, G. T., 11, 92, 172
Knot, 1204, 1241
Knuth, Donald, 181
Kolmogorov quotient, 1066
Kolmogorov, Andrei, 49, 1390, 1428
Körner, T. W., 67
Korovkin’s Theorem, 1767
Kovalevskaya, Sofia, 1964
Kramp, Christian, 1929
Kronecker, Leopold, 74, 1951
Kummer’s Test, 451
Kuratowski closure operation, 1018
Kuratowski’s Lemma, 163
Kuratowski, Kazimierz, 200

l’Hôpital’s Rule, 609–620
l’Hôpital, Marquis Guillaume de, 609
Lagrange, Joseph, 335, 591
Lagrange, Joseph-Louis, 1927, 1929
Lakatos, Imre, 41, 508
Lakes of Wada, 1240
$\lambda$-system, 1359

Lambert, Johann Heinrich, 29, 339
language, 118
extended simple, 121
simple, 118
large cardinal, 193
large inductive dimension, 1292
largest element, 137
lattice, 137, 1770
Laugwitz, Detlef, 350, 508
Laurent series
formal, 296
Lawvere, F. William, 211
Lax-Milgram Theorem, 1810
least upper bound, 136, 233, 281
Lebesgue Covering Theorem, 1276
Lebesgue Density Theorem, 1721
Lebesgue integrable function
nonnegative, 1604
Lebesgue integral, 1487–1492, 1615
Lebesgue measurable set, 1421
Lebesgue measure, 5, 1421–1423
Lebesgue number, 1112
Lebesgue outer measure, 1417–1419
Lebesgue Property, 1112
Lebesgue, Henri, 198, 1203, 1390, 1391, 1421,
1487, 1652, 1667
Lefschetz, Solomon, 1927
left inverse, 374
Legendre differential equation, 1799
Legendre polynomials, 1540, 1798
Legendre, Adrien, 339, 654
Legendre-Gauss quadrature, 1540
Leibniz, Gottfried, 66, 350, 1483, 1485, 1931
lema, 123
length metric, 771
length of an interval, 237
length structure, 777
Finslerian, 778
lens space, 1086
Levi, Beppo, 1608, 1788
Levi-Civita number, 297
Lie group, 1912–1913
Lie, Sophus, 1912
Lilly, William, 37
lim inf, 424–429
lim sup, 424–429
limit
direct
  of sets, 1166–1170
  of topological spaces, 1174–1175
inductive, 1167
inverse
  of sets, 1170–1174
  of topological spaces, 1175–1180
of a net, 1025
of a sequence, 413–417, 1022
projective, 1171
limit cardinal, 186
Limit Comparison Test, 443
limit inferior, 424–429
limit ordinal, 175
limit point, 1014
  of a subset of $\mathbb{R}$, 723
limit superior, 424–429
limited element, 352
Lindelöf space, 1008
LINDEMANN, FERDINAND, 62, 339, 342
Line Axiom, 745
Line Segment Axiom, 745
linear algebra, 370–411
linear combination, 380
linear function, 390
linear functional, 1732, 1735
  positive, 1880
linear homotopy, 1316
linear operator, 390
linear transformation, 390–391
  determinant, 404
linearly independent set, 381
link, 1242
Liouville number, 332
Liouville’s Theorem, 920
LIOUVILLE, JOSEPH, 331
LISTING, JOHANN, 1928
Little Picard Theorem, 920
LOBACHEVSKII, NIKOLAI, 29
local base
  for a topology, 1005
local extremum, 538
local maximum, 538, 560
local minimum, 538, 560
local uniformizing variable, 973
locally compact group, 1900
locally compact space, 1120–1125
locally conical, 1342
locally connected space, 1158
locally Euclidean group, 1912
locally finite cover, 1130
locally Hausdorff space, 1049
locally path-connected space, 1160
locally second countable, 1008
locally $T_0$ space, 1047
locally $T_1$ space, 1048
LOCHS, GUSTAV, 329
Loeb measure, 367
log-convex function, 1692
logarithm, 565
  in a differential field, 992
  natural, 567
logic
  first-order, 103
  higher-order, 103
  intuitionist, 114
  ordinary, 113
  second-order, 103
logicism, 72–73
long line, 1267
loop, 1317
Łoś’s Theorem
  simple version, 364
low-dimensional topology, 1218
lower bound, 136, 233
lower envelope, 1521, 1592
lower semicontinuous decomposition, 1063
lower set, 279
$L^p$-space, 1847–1860
LUCAS, J. R., 41, 68
Lusin, see Luzin
Luzin Separation Theorem, 1388
LUZIN, NIKOLAI, 1390, 1708, 1957

Mackey Borel structure, 1888
MACLANE, SAUNDERS, 211, 417
MADDY, PENELOE, 148
Mahlo cardinal, 202
manifold, 1260–1271
  closed, 1265
  compact, 1265

2010
open, 1265
topological, 1260–1266
  with boundary, 1263–1265
unbounded, 1265
manifold with boundary
  topological, 1263–1265
map
  essential, 1274
MARCIKIEWICZ, JÓZEF, 1145
Martin’s Axiom, 208
MARTIN, GEORGE E., 35
mathematical dishonesty, 94
mathematical honesty, 94
mathematics
  applied, 60
  pure, 60
matrix, 370
  addition, 370
  classical adjoint, 408
determinant, 396–410
  expansion by minors, 401–402
  geometric interpretation, 404–408
  multiplicative property, 402–403
identity, 373
inverse, 374, 408
multiplication, 372
  by blocks, 374
transpose, 374
matrix algebra, 373
matrix ring, 373
matrix units, 1889
Max-Min Theorem, 521
maximal element, 137
maximal Hausdorff quotient, 1069
maximal ideal, 170
maximal pseudometric, 734
maximal $T_0$ quotient, 1066
maximum
  endpoint, 560
  local, 560
MAY, KENNETH O., 54, 84
McCARTY, D. C., 76
meager set, 1142
Mean Value Inequality, 594, 910
Mean Value Theorem, 589
  for integrals, 1527
measurable box, 1458
measurable cardinal, 204
measurable function, 1361
  Borel, 1372–1380
measurable rectangle, 1454, 1457, 1458
measurable set, 1419
measurable space, 1360
measurable subspace, 1362
measure, 2–6, 1393–1407
  atomless, 201
  Borel, 1443, 1852
  compression, 1406
  counting, 1394
  decomposable, 1402
  dilation, 1406
  downward continuous, 1397
  extension, 1406, 1407
  finite, 1401
  finitely additive, 1394
  full real-valued, 198
  Haar, 1907
  inner, 1431
  Lebesgue, 5, 1421–1423
  Loeb, 367
  outer, 1415
  packing, 1451
  probability, 1401
  product, 1456
  push-forward, 1407
  restriction, 1406, 1407
  semifinite, 1403
  σ-finite, 1401
  spherical, 1448
  upward continuous, 1396
  zero, 1394
measure space, 1393
  complete, 1399
  completion, 1400
measure zero
  in $\mathbb{R}^n$, 1326
Menger sponge, 1231
Menger universal curve, 1231
MERSENNE, MARIN, 109
mesa function
  smooth, 701
Method I, 1416

2011
Method II, 1443
method of undetermined coefficients, 1539
metric, 725
discrete, 1004
extended, 733
intrinsic, 771
length, 771
quotient, 1067
strict length, 771
strictly intrinsic, 746, 771
metric outer measure, 1442
metric space, 725
of a measure space, 1352–1353
metrizable space, 1004, 1093–1097
midpoint, 750
Midpoint Axiom, 761
midpoint rule, 1533
midpoint-convex function, 1682
MILL, JOHN STUART, 42
MILNOR, JOHN, 906, 1283
minimax approximation, 645
minimum
endpoint, 560
local, 560
Minkowski’s Inequality, 1850
MINKOWSKI, HERMANN, 30
Möbius strip, 1076
MÖBIUS, AUGUST FERDINAND, 1933
MÖBIUS, PAUL JULIAN, 1933
modus ponens, 114
MÖLLERUP, J., 656
MONGRÉ, PAUL, 1931
Monodromy Theorem, 971
monotone class, 1359
Monotone Class Theorem, 1360
Monotone Convergence Theorem, 1607
Baby, 1396
Downward, 1610
for Riemann integration, 1589
for series, 1609
for step functions, 1589
Junior, 1602
monotone function, 1665
monotone regular outer measure, 1429
monotone sequence
of functions, 1141
Montel’s Theorem, 946
MONTGOMERY, DEANE, 1912
Moore manifold, 1268, 1269
Moore method, 1030
Moore space, 1268
MOORE, CALVIN C., 1030
MOORE, ELIAKIM HASTINGS, 1030, 1927
MOORE, GREGORY H., 167
MOORE, ROBERT L., 1030, 1268
Moore-Smith convergence, 1030
Morera’s Theorem, 919
MORSE, ANTHONY P., 1325
MORSE, MARSTON, 1325
Morse-Sard Theorem, 1325
Mostowski collapsing map, 366
multi-index, 807
multi-index notation, 869, 873
multiple
of a representation, 1878
Multiplicative Axiom, 162
multiplicity function, 1380
multivariable calculus, 796
multivariable power series, 869
NAPIER, JOHN, 37, 1928
NASH, JOHN, 1935
natural logarithm, 567
natural number, 220–222, 247–265
negation, 110–113
negative part of a function, 1624
neighborhood, 1003
of a point of \( \mathbb{R} \), 723
neighborhood filter, 1034
NEMES, GÉRGO, 664
net, 1024–1030
interleaving, 1029
universal, 1037
NEUMAN, ALFRED E., 30
NEWMAN, JAMES, 149
NEWTON, ISAAC, 350, 635, 1483, 1537, 1929, 1931, 1935
Newton-Cotes formula, 1537
open, 1539
Nikodym Convergence Theorem, 1477
NÍKODÝM, OTTO, 1476
NÍKOLSKII, SERGEI, 1931
2012
Niven, Ivan, 339
No-Contraction Theorem, 1272
No-Retraction Theorem, 1272
Nöbeling universal space, 1306
nodal set, 206
Noether, Emmy, 1834, 1927
Noether, Fritz, 1834
Noether, Max, 1834
nonanalytic point, 699
type (C), 699
type (P), 699
nonarchimedean ordered field, 295, 353
nondecreasing function, 1665
nondecreasing sequence of functions, 1141
nondegenerate representation, 1879
nonincreasing function, 1665
nonincreasing sequence of functions, 1141
nonmeasurable set, 1436–1439
nonnegative infinite series, 439
nonoverlapping, 1553
nonstandard analysis, 350
nonstandard extension of a function, 359
of a relation, 360
of a set, 357
nonstationary, 765, 783
normal element of a C*-algebra, 1863
normal convergence, 509, 928
normal family, 945
normal filter, 206
normal space, 1046, 1050–1053
nowhere dense set, 1142
NSS group, 1912
null set, 1398, 1517
null space of a linear transformation, 392
nullhomotopic, 1316
number algebraic, 158, 331, 343
algebraically independent, 343
Cayley, 308
complex, 300–308
decadent rational, 317
dyadic rational, 278
hypernatural, 358
hyperreal, 354
integer, 227–230, 265–274
irrational, 44, 245, 339–349, 570
Levi-Civita, 297
Liouville, 332
natural, 220–222, 247–265
p-adic, 299
prime, 230, 263
rational, 231, 275–278
real, 214–245, 279–300
transcendental, 158, 331, 339–349
von Neumann, 338
number line, 232
numerical analysis, 1531
numerical integration procedure, 1532
numerical integration process, 1532
numerical range, 1829
obvious, 14
ODE, 876
oganesson, 71
one-point compactification, 1121, 1139
one-sided derivative, 557–558
open cover, 1008, 1107
decomposable, 1297
open interval, 236
open manifold, 1265
Open Mapping Theorem, 1744
for σ-compact groups, 1904
for locally compact groups, 1902
open set, 1003
basic, 1005
in  R, 721
operation, 100
operator, 391
bounded, 1720
bounded below, 1827
compact, 1835
completely continuous, 1835
Fredholm, 1837
positive, 1830
strictly singular, 1837
weakly continuous, 1836
operator algebra, 1733
Oppenheimer, Frank, 61
orbit, 1085
orbit space, 1085
order
  of a zero, 948
order ideal, 1365
  prime, 1367
order of a cover, 1295
order of precision, 1535
ordered field, 220, 276, 352
  nonarchimedean, 353
ordering
  complete, 136
  equivalent, 135
  induced, 135
  ordinary, 135
  partial, 135–143
  strict, 135
  total, 135
Ordering Principle, 166
ordinal, 173–182
  limit, 175
  successor, 175
ordinary differential equation, 876
ordinary ordering, 135
ordinary differential equation, 876
Oresme, Nicole, 442, 1930
orthogonal dimension, 1796
orthogonal projection operator, 1800
orthogonality
  in Hilbert space, 1789
orthonormal basis, 1794
orthonormal set, 1791
oscillation, 1484, 1503–1504
  weighted local, 1504
osculating polynomial, 641
osculatory quadrature procedure, 1536
Osgood’s Lemma, 977
Osgood, William, 1143, 1656
Oughtred, William, 1930
outer measure, 1415
  Borel regular, 1432
  Lebesgue, 1417–1419
  metric, 1442
  monotone regular, 1429
  regular, 1429
packing measure, 1451
p-adic number, 299
parabolic rule, 1539
paracompact space, 1130–1135
paradox
  Burali-Forti, 174
  Russell’s, 160
Parallel Postulate, 29
Parallelogram Axiom, 751
Parallelogram Law, 1783
parametrization by arc length, 770, 790
parametrized curve
  in $\mathbb{R}^n$, 782
  in a metric space, 763
parentheses, 102
Parseval’s relation, 1793
partial derivative, 802
partial differential equation, 876
partial isometry
  in a C*-algebra, 1863
partial ordering, 135–143
  strict, 135
partial sum, 431
partially ordered set, 135–143
partition, 1497–1498
  between closed sets, 1050
  full, 1586
  of a rectangle, 1553
partition dimension, 1298
partition of unity, 1132–1134
Pasc, Moritz, 31
path, 1159
  in $\mathbb{R}^n$, 784
  in a metric space, 765
path component, 1160
path-connected space, 1159
Paulos, John, 10
PDE, 876
Peano Axioms, 247
Peano’s Existence Theorem for First-Order ODEs, 891
Peano, Giuseppe, 247, 1651
Peary, Robert, 71
Pedersen, Gert K., 13
Pell’s equation, 335
Penrose, Roger, 30, 70
PERELMAN, Grigori, 1935
perfect subset, 1020
perfectly normal, 1106
Perron integral, 1493
Picard Existence Theorem, 882
Picard iteration, 884
Picard’s Theorem
  Big, 955
  Little, 920
Picard-Lindelöf Theorem, 882
piecewise-linear map, 1350
PIERPONT, James, 149
\(\text{-}\)system, 1359
PL map, 1350
PLATO, 41
Platonism, 55, 68–72
Playfair’s Axiom, 29
Poincaré homology sphere, 1087
Poincaré, Henri, 29, 30, 54, 61, 79
point, 379
  singular, 699
point finite cover, 1130
point mass, 1394
point of inflection, 598
point spectrum, 1827
pointwise bounded sequence of functions, 1378
Polarization Identity, 1783
pole, 951
  simple, 951
Polish group, 1903
Polish space, 1101, 1385
POLYA, George, 52, 81
polydisk, 870
polyhedron, 1341
  noncompact, 1344
Pontryagin dual group, 1914
PONTRYAGIN, Lev, 1912
positive
  element of a C*-algebra, 1873
  linear functional, 1880
  operator, 1830
positive part of a function, 1624
POSTNIKOV, Mikhail, 80
Poulsen simplex, 1764
Power Rule, 530, 535
  complex, 903
power series, 674–682
  formal, 295
  multivariable, 869
  unordered, 690
pre-Hilbert space, 1787
preimage, 127, 129
premeasure, 1423
prime filter, 1367
prime ideal, 170
prime number, 230, 263
primitive, 601
primitive ideal, 1886
principal ideal, 170
Principle of Induction, 222
Principle of Recursion, 226
Principle of the Excluded Middle, 44
PRINGSHEIM, Alfred, 99, 712, 1945
probability, 6
probability measure, 1401
product \(\sigma\)-algebra, 1454, 1459
product measure, 1456
product of infinite series, 494
Product Rule, 532
  complex, 903
projectable region, 1576, 1578
projection, 1800
  in a C*-algebra, 1863
projective limit, 1171
projective plane, 1080
projective space
  complex, 1086
  real, 1086
proof, 40
  by contradiction, 43–44
  formal, 40
Property (N), 1708
Property (S), 1709
proposition, 102, 123
PRÜFER, Heinz, 1268
Prüfer manifold, 1268
\(p\)-series, 443
pseudo-arc, 1180, 1238
pseudo-circle, 1240
pseudocompact space, 1129, 1303
Puiseux Series, 296
pure mathematics, 60
push-forward
  of a measure, 1407
Putnam, Hilary, 77
Pythagoras, 43
Pythagorean Axiom, 751
Pythagorean Theorem, 1790
quadrature procedure, 1532
  composite, 1536
  interpolatory, 1536
  osculatory, 1536
  regular, 1532
  simple, 1536
quantifier, 106–108
  existential, 106
  negation, 111–112
  restricted, 106
  separated, 116–117
  universal, 106
quasi-compact mapping, 1062
quasi-component, 1154
quasi-inverse, 1838
quaternion, 307–308
quotient field, 278, 296
quotient metric, 1067
quotient ring, 273
Quotient Rule, 533
  complex, 903
quotient topology, 1060
Raabe’s Test, 452
  Limit Version, 452
  Second Order, 454
radius of convergence, 674, 932
Radó, Tibor, 1268
Radon-Nikodym Theorem
  for positive linear functionals, 1884
Raikov-complete, 1899
Ramanujan constant, 344
Ramsey cardinal, 207
Ramsey, Frank, 1932
range
  of a linear transformation, 392
Range-Straightening Theorem, 857
Rank Theorem, 856
Ratio Test, 449, 462
  Basic, 449
rational function, 278, 294, 409
rational number, 231, 275–278
rational ruler, 1255
real number, 214–245, 279–300
real part
  of an element of a C*-algebra, 1863
real-closed field, 245, 278, 297
real-valued measurable cardinal, 199
realism, 55, 68–72
rearrangement, 473
Recorde, Robert, 10, 1929
rectangle, 1552
rectangular sum of an array, 484
rectifiable curve, 766, 785
rectifiably path-connected, 772
recursion, 226–227, 251–252
  transfinite, 140–141
reductio ad absurdum, 43–44
Rees, Mina, 13
refinement, 1130
reflexive Banach space, 1741
region, 963
regular cardinal, 192
regular outer measure, 1429
regular point, 854
regular quadrature procedure, 1532
regular Riemann Surface, 971
regular space, 1046, 1049–1050
regulated function, 606
relation
  equivalence, 131
relational system, 117
relative closure, 1010
relative extremum, 538
relative maximum, 538
relative minimum, 538
relative topology, 1009
relatively open, 1010
religion, 38
removable singularity, 950
reparametrization, 764, 783
representation, 1878
  faithful, 1878
  GNS, 1883
  irreducible, 1878
nondegenerate, 1879
representing measure, 1757
residual set, 1142
residual spectrum, 1827
resolvent function, 1817
resolvent set, 1816
restricted direct product, 1183
restricted quantifier, 106
restriction
of a measure, 1362, 1406, 1407
revision of an infinite series, 477
drastic, 490
Ribenboim, Paulo, 61
Riemann Hypothesis, 344
Riemann integral, 6–7, 1481–1482, 1486, 1496–1596
improper, 1530
in \( \mathbb{R}^n \), 1552–1585
Riemann Mapping Theorem, 960
Riemann sum, 1484, 1498
Riemann Surface
of a function, 964
regular, 971
Riemann, Bernhard, 6, 29, 477, 1481
Riesz Representation Theorem, 1492
Riesz, Frigyes, 1018, 1786, 1827
Riesz-Frèchet theorem, 1802
right inverse, 374
Rigidity Theorem, 948
ring, 271, 370
commutative, 271
division, 307
integral domain, 274
matrix, 373
of fractions, 277
of subsets, 1356
ordered, 273
quotient, 273
unital, 271
ring of adeles, 1183
ring of fractions, 277
Robinson, Abraham, 350, 508
Rolle’s Theorem, 590
Rolle, Michel, 590, 1930
Romberg integration, 1548, 1551
root

of a number, 243
Root Test, 448, 462
Basic, 448
Roth, Klaus, 333
Rotman, Brian, 11, 172
row vector, 371
Rowbottom, Frederick, 206
Rucker, Rudy, 181
Rudin, Mary Ellen, 1131
Runge phenomenon, 643
Russell’s Paradox, 160, 171
Russell, Bertrand, 14, 57, 72, 73, 79, 152, 162
Saint Raymond, Jean, 854
Saks, Stanislaw, 561, 1476
Sard, Arthur, 1325
saturated, 1061
Scaling Test, 470
scattered, 1020
Schmidt, Erhard, 1786
Schmieden, Curt, 350
Schneider, Theodor, 344
Schönhflies, Arthur, 1203, 1786
Schröder, Ernst, 157
Schröder-Bernstein Theorem, 157, 173
for measurable spaces, 1364
Schumacher, Carol, 12
Schwarz Inequality, 1782
Schwarz, Hermann, 1782
Scott, Dana, 208
secant, 574
second category set, 1142
second countable
locally, 1008
second countable space, 1006
Second Decomposition Theorem, 1295
second derivative, 541
second derivative test, 598
second differences, 552
second-order logic, 103
second-order statement, 103
second-order theory, 103, 248
self-adjoint
element of a C*-algebra, 1863
semialgebra, 1355
semicanonical form, 1599

2017
semiclosed map, 1071
semifinite measure, 1403
semigroup
  abelian, 274
semiring, 1355
sentence, 102
separable topological space, 1016
separated
  completely, 1741
  fully, 1741
separated sets, 1093
separated union, 1011
sequence, 412
  bounded, 413
  Cauchy, 288
  convergent, 414
  divergent, 414
  limit, 413–417
  of functions, 508–510
  monotone, 1141
  nondecreasing, 1141
  nonincreasing, 1141
  pointwise bounded, 1378
sequence of first differences, 552
Sequential Criterion for continuity, 518
sequential space, 1038
sequentially closed, 1038
sequentially compact, 1110
SÉRÉE, JÉAN-PIERRE, 68
set-theoretic difference, 1355
sex, 61
SHAW, GEORGE BERNARD, 7
SHELAH, SAHARON, 208
shrinking of a cover, 1051
SIEGEL, CARL, 344, 346
Sierpiński carpet, 1227
Sierpiński gasket, 1227
Sierpiński sieve, 1227
Sierpiński space, 1054
Sierpiński triangle, 1227
σ-finite measure, 1401
σ-algebra
  product, 1454, 1459
σ-compact space, 1119
SILVER, JACK, 208
simple function, 1489, 1598
simple language, 118
simple pole, 951
simple quadrature procedure, 1536
Simple Transfer Principle, 363
simplex, 1340
simplicial complex, 1345
  abstract, 1347
simplicial isomorphism, 1346
simplicial map, 1346
simply connected, 959
Simpson’s rule, 1534
Simpson’s three-eighths rule, 1549
SIMPSON, THOMAS, 1537
sine, 573
singular cardinal, 192
singular point, 699, 968
singularity, 949
  essential, 954
  pole, 951
  removable, 950
skeleton, 1346
skew field, 307
SKOLEM, THORALF, 171
small inductive dimension, 1290
smooth bump function, 700
smooth Heaviside function, 701
smooth mesa function, 701
smooth partition of unity, 701
Smooth Urysohn Lemma, 702
solenoid, 1179, 1233
SOLOVAY, ROBERT, 201, 208
sophomore’s dream, 673
Sorgenfrey line, 1056
Suslin, see Suslin
space-filling curve, 1651–1652
  ultimate, 1651
span, 380
Spectral Mapping Theorem
  for inverses, 1819
  for polynomials, 1818
spectral resolution, 1846
Spectral Theorem
  Finite-Dimensional, 1845
  spectrum
    of an element of a Banach algebra, 1816
    of an operator, 1827
spherical measure, 1448
Spivak, Michael, 12
square root
   in a C*-algebra, 1873
square sum of an array, 483
stable algebra, 1889
stable C*-algebra, 1889
stably isomorphic C*-algebras, 1889
Stalin, Josef, 1390
standard Borel space, 1385
standard revision of an unordered power series, 691
star, 1139, 1344
star-refinement, 1139
state, 1880
   pure, 1882
   vector, 1881
state space, 1880
statement
   atomic, 102
   closed, 108–109
   first-order, 103
   formal, 100–117
   second-order, 103
stationary set, 1304
step function, 601, 1586
Stewart, Ian, 39, 58, 877
Stirling’s formula, 661
Stone-Čech compactification, 1127–1130
Stone, Arthur H., 1131
Stone, Marshall, 1127, 1768
Stone-Weierstrass Theorem, 1304, 1768
Story, William, 1933
Straightening Theorem, 857
strict contraction, 742
strict length metric, 771
strict ordering, 135
strict partial order, 135
strict power contraction, 743
strictly convex, 1807
strictly decreasing, 1665
strictly decreasing function, 1665
strictly increasing function, 1665
strictly intrinsic metric, 746, 771
strictly monotone function, 1665
strictly singular operator, 1837
Strong Law of Large Numbers, 330
strongly continuous mapping, 1061
strongly inaccessible cardinal, 193
strongly infinite-dimensional space, 1311
Struik, Dirk, 1932
subadditivity
   countable, 1395, 1415
   finite, 1395, 1415
subarrangement, 473
subbase, 1007
subcomplex, 1345
subcover, 1008, 1107
subdivision
   of a complex, 1346
submultiplicative norm, 1811
subnet, 1025–1027
   weak, 1027
subordinate filterbase, 1036
subrepresentation, 1878
subseries, 473
subspace, 379
   Borel, 1362
      complementary, 387
      measurable, 1362
subtree, 1100
successor cardinal, 186
successor ordinal, 175
sum
   of representations, 1878
sum of an infinite series, 432
Sum Rule, 531
   complex, 903
Sum Theorem
   for ind, 1293
superset property, 1034, 1365
superstructure, 365
support, 1132
   compact, 1125
   finite, 1855
   of σ-finite measure, 1855
   of almost finite measure, 1856
   of finite measure, 1855
support hyperplane, 1752
supremum, 136, 233, 281
surgery, 1082
   on a knot, 1084
Suslin operation, 1389, 1421, 1432
Suslin scheme, 1432
monotone, 1432
Suslin set, 1389, 1432
in \( \mathbb{R}^n \), 1434
Suslin, Mikhail, 1390, 1932
swelling, 1051
Swiss cheese, 1229
symmetric difference, 1355
tail
  of an infinite series, 433
tail set, 1028
tangent, 574
tautology, 113
Taylor series, 682, 938
Taylor’s Theorem
  Cauchy remainder formula, 627, 630
  integral remainder formula, 628
  Lagrange remainder formula, 626
  single variable, 624
Taylor, Brook, 622, 632
Teichmüller-Tukey Lemma, 164
Teichmüller, Oswald, 164
tensor product
  of Hilbert spaces, 1803
term
  compound, 102
  formal, 102
  simple, 102
tetration, 181
theorem, 123
Thue, Axel, 333
Tikhonov plank, 1137
Tikhonov space, 1053
Tikhonov’s Theorem, 164, 168, 1115–1118
Titchmarsh, E. C., 7, 339
Tonelli’s Theorem, 1633
  Baby, 1328
  for double summations, 485
  Junior, 1632
Tonelli, Leonida, 488
topological group
  complete, 1899
  locally compact, 1900
  locally Euclidean, 1912

NSS, 1912
Polish, 1903
topological space, 1003–1012
arcwise connected, 1159
compact, 1008, 1107–1118
compactly generated, 1135
completely collectionwise normal, 1191
completely Hausdorff, 1059
completely normal, 1093
completely regular, 1053–1054
connected, 1150
continuum, 1239
contractible, 1316
discrete, 1004
first countable, 1006
fully normal, 1139
functionally Hausdorff, 1059
Hausdorff, 1046, 1049
hereditarily normal, 1093
indiscrete, 1004
Lindelöf, 1008
locally compact, 1120–1125
locally connected, 1158
locally Hausdorff, 1049
locally path-connected, 1160
locally \( T_0 \), 1047
locally \( T_1 \), 1048
metrizable, 1004, 1093–1097
normal, 1046, 1050–1053
paracompact, 1130–1135
path-connected, 1159
Polish, 1101, 1385
pseudocompact, 1129, 1303
regular, 1046, 1049–1050
scattered, 1020
second countable, 1006
separable, 1016
sequential, 1038
\( T_0 \), 1046
\( T_1 \), 1046
\( T_2 \), 1046
\( T_{2 \frac{1}{2}} \), 1059
\( T_3 \), 1046
\( T_{3 \frac{1}{2}} \), 1053
\( T_4 \), 1046
\( T_5 \), 1093
\(T_6, 1106\)
totally disconnected, 1155
totally normal, 1056
totally path disconnected, 1160
totally separated, 1156
Urysohn, 1059
zero-dimensional, 1156
topologist’s sine curve, 1178, 1223
topology, 1003
countable complement, 1012
defined by a metric, 1004
discrete, 1004
finite complement, 1006
indiscrete, 1004
inductive limit, 1174
relative, 1009
torus, 1078
total analytic function, 963
total derivative, 812
total differential, 812
total ordering, 135
totally disconnected space, 1155
totally normal space, 1056
totally ordered set, 135
totally path disconnected space, 1160
totally separated space, 1156
transcendental number, 158, 331, 339–349
Transfer Principle, 361
Extended First-Order, 364
Extended Simple, 364
First-Order, 363
Simple, 363
transfinite dimension, 1314
transfinite induction, 138–141, 175–176
transitive set, 173, 183
translation, 1892
transpose, 374
trapezoid rule, 1533
tree, 1100
tree of mathematics, 57
triangle inequality, 1850
triangular sum of an array, 484
triangulation
of a polyhedron, 1345
of a topological space, 1348
weak, 1348
trigonometric function, 571–582
inverse, 575–576
trigonometric polynomial, 1774
TRUSS, J. K., 9, 214
truth table, 106
TUCKER, A. W., 1278
Tukey’s Lemma, 164
Turing computable, 28
Tychonoff, see Tikhonov
u.c. convergence, 509, 1192
UFD, 348
Ulam matrix, 200
ULAM, STANISLAW, 200
ultimate space-filling curve, 1651
ultrafilter, 205, 1365
fixed, 205
free, 205, 1366
normal, 206
principal, 1366
Ultrafilter Property, 166
unbounded manifold, 1265
uncountable set, 152
Uniform Boundedness Theorem, 1749
uniform structure
on a topological group, 1897
uniformly absolutely continuous family, 1475
uniformly convex, 1807
uniformly countably additive family, 1475
unilateral shift, 1829
union, 1355
unique factorization domain, 348
unitary
in a C*-algebra, 1863
unitization, 1812
universal net, 1037
universal property, 1167, 1171
universal quantifier, 106
universe
Gödel constructible, 183
von Neumann, 183
unlimited element, 352
unordered power series, 690
unordered sum, 478
upper bound, 136, 233
upper envelope, 1521, 1592

2021
upper semicontinuous decomposition, 1063
upper set, 279
upward continuity, 1396
urelement, 365
Urysohn Metrization Theorem, 1094
Urysohn space, 1059
Urysohn’s Lemma, 1052
Smooth
for $\mathbb{R}$, 702
URYSOHN, PAVEL, 1932
vacuous, 225
Vandermonde matrix, 632
variable, 100
bound, 106
free, 106
VEBLEN, OSWALD, 1202
vector, 379
  column, 371
  linear independence, 381
  row, 371
vector calculus, 796
vector space, 377–390
  basis, 383
  dimension, 385
  finite-dimensional, 381
  isomorphism, 379
  quotient, 388–390
  subspace, 379
  complementary, 387
vertex, 1338, 1345
vertical arrow notation, 181
Viéte, François, 496
Vietoris topology, 1105
VIETORIS, LEOPO LD, 1931
VILLANI, CÉDRIC, 38, 39, 897
Vitali cover
  of a metric space, 1444
VITALI, GIUSEPPE, 198, 1436, 1476
Vitali-Hahn-Saks Theorem, 1477
volume, 1552
von Neumann algebra, 1733
von Neumann numbers, 338
von Neumann universe, 183
VON NEUMANN, JOHN, 30, 65, 171, 249, 1418, 1460, 1912, 1928, 1935
WAGON, STAN, 172
WALLIS, JOHN, 75, 496, 1929, 1931
Wallman compactification, 1130
WARING, EDWARD, 633
Warsaw circle, 1224
Weak Euclidean Four-Point Axiom, 750
weak subnet, 1027
weak triangulation, 1348
weakly inaccessible cardinal, 192
weakly infinite-dimensional space, 1311
weakly $n$-dimensional, 1313
wedge sum, 1075
Weierstrass Approximation Theorem, 1766, 1931
  general, 1771
WEIERSTRASS, KARL, 9, 342, 508, 1766, 1929, 1931
weight
  of a topological space, 1007
weighted local oscillation, 1504
Weil completion, 1899
WEIL, ANDRÉ, 52, 80, 90, 207
well defined, 115
well ordering, 137–138
well-ordered set, 137–138
Well-Ordering Principle, 162
WEYL, HERMANN, 79
WHewELL, WILLIAM, 77
Whitehead continuum, 1252
Whitehead manifold, 1253
WHITEHEAD, ALFRED NORTH, 40, 99
WHITEHEAD, JOHN H. C., 1252
WHITNEY, HASSLER, 1335
WIDMANN, JOHANNES, 1930
WIGNER, EUGENE, 60
Wikipedia, 94
wild arc, 1244
WILES, ANDREW, 54
WOLF, ABRAHAM, 92
WRIGHT brothers, 71
YAMABE, HIDEHIKO, 1912
Yegorov, see Egorov
Yermakov’s Test, 457
Young’s Inequality
  for products, 1849
  General, 1857

2022
Young, Grace Chisholm, 561
Young, William H., 1849

Zahorski’s Theorem, 713
Zariski topology, 1058
Zermelo, Ernst, 152, 171
zero content, 1517, 1564
zero measure, 1394

zero set, 701
zero-dimensional space, 1156
ZF set theory, 171
ZFC set theory, 171
Zippin, Leo, 1912
Zorich, Vladimir, 45
Zorn’s Lemma, 163, 176–177