Texts and Readings in Mathematics 37

Terence Tao

Analysis I

Fourth Edition

HINDUSTAN
BOOK AGENCY

Springer

Texts and Readings in Mathematics

The **Texts and Readings in Mathematics** series publishes high-quality textbooks, research-level monographs, lecture notes and contributed volumes. Undergraduate and graduate students of mathematics, research scholars and teachers would find this book series useful. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Books in this series are co-published with Hindustan Book Agency, New Delhi, India.

Terence Tao

Analysis I

Fourth Edition

HINDUSTAN
BOOK AGENCY

Springer

Terence Tao
Department of Mathematics
University of California Los Angeles
Los Angeles, CA, USA

ISSN 2366-8717 ISSN 2366-8725 (electronic)
Texts and Readings in Mathematics
ISBN ISBN 978-981-19-7261-4 (eBook)
https://doi.org/10.1007/978-981-19-7261-4

Jointly published with Hindustan Book Agency
This work is a co-publication with Hindustan Book Agency, New Delhi, licensed for sale in all countries in electronic form only. Sold and distributed in print across the world by Hindustan Book Agency, P-19 Green Park Extension, New Delhi 110016, India.
ISBN: 978-81-951961-9-7 © Hindustan Book Agency 2022

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To my parents, for everything

Preface to the First Edition

This text originated from the lecture notes I gave teaching the honours undergraduate-level real analysis sequence at the University of California, Los Angeles, in 2003. Among the undergraduates here, real analysis was viewed as being one of the most difficult courses to learn, not only because of the abstract concepts being introduced for the first time (e.g., topology, limits, measurability, etc.), but also because of the level of rigour and proof demanded of the course. Because of this perception of difficulty, one was often faced with the difficult choice of either reducing the level of rigour in the course in order to make it easier, or to maintain strict standards and face the prospect of many undergraduates, even many of the bright and enthusiastic ones, struggling with the course material.

Faced with this dilemma, I tried a somewhat unusual approach to the subject. Typically, an introductory sequence in real analysis assumes that the students are already familiar with the real numbers, with mathematical induction, with elementary calculus, and with the basics of set theory, and then quickly launches into the heart of the subject, for instance the concept of a limit. Normally, students entering this sequence do indeed have a fair bit of exposure to these prerequisite topics, though in most cases the material is not covered in a thorough manner. For instance, very few students were able to actually *define* a real number, or even an integer, properly, even though they could visualize these numbers intuitively and manipulate them algebraically. This seemed to me to be a missed opportunity. Real analysis is one of the first subjects (together with linear algebra and abstract algebra) that a student encounters, in which one truly has to grapple with the subtleties of a truly rigorous mathematical proof. As such, the course offered an excellent chance to go back to the foundations of mathematics, and in particular the opportunity to do a proper and thorough construction of the real numbers.

Thus the course was structured as follows. In the first week, I described some well-known "paradoxes" in analysis, in which standard laws of the subject (e.g., interchange of limits and sums, or sums and integrals) were applied in a non-rigorous way to give nonsensical results such as $0 = 1$. This motivated the need to go back to the very beginning of the subject, even to the very definition of the natural numbers, and check all the foundations from scratch. For instance, one of the first homework assignments was to check (using only the Peano axioms) that addition was associative for natural numbers (i.e., that $(a + b) + c = a + (b + c)$ for all natural numbers a, b, c: see Exercise 2.2.1). Thus even in the first week, the students had to write rigorous proofs using mathematical induction. After we had derived all the basic properties of the natural numbers, we then moved on to the integers (initially defined as formal differences of natural numbers); once the students had verified all the basic properties of the integers, we moved on to the rationals (initially defined as formal quotients of integers); and then from there we moved on (via formal limits of Cauchy sequences) to the reals. Around the same time, we covered the basics of set theory, for instance demonstrating the uncountability of the reals. Only then (after about ten lectures) did we begin what one normally considers the heart of undergraduate real analysis—limits, continuity, differentiability, and so forth.

The response to this format was quite interesting. In the first few weeks, the students found the material very easy on a conceptual level, as we were dealing only with the basic properties of the standard number systems. But on an intellectual level it was very challenging, as one was analyzing these number systems from a foundational viewpoint, in order to rigorously derive the more advanced facts about these number systems from the more primitive ones. One student told me how difficult it was to explain to his friends in the non-honours real analysis sequence (a) why he was still learning how to show why all rational numbers are either positive, negative, or zero (Exercise 4.2.4), while the non-honours sequence was already distinguishing absolutely convergent and convergent series, and (b) why, despite this, he thought his homework was significantly harder than that of his friends. Another student commented to me, quite wryly, that while she could obviously *see* why one could always divide a natural number n into a positive integer q to give a quotient a and a remainder r less than q (Exercise 2.3.5), she still had, to her frustration, much difficulty in writing down a proof of this fact. (I told her that later in the course she would have to prove statements for which it would not be as obvious to see that the statements were true; she did not seem to be particularly consoled by this.) Nevertheless, these students greatly enjoyed the homework, as when they did perservere and obtain a rigorous proof of an intuitive fact, it solidified the link in their minds between the abstract manipulations of formal mathematics and their informal intuition of mathematics (and of the real world), often in a very satisfying way. By the time they were assigned the task of giving the infamous "epsilon and delta" proofs in real analysis, they had already had so much experience with formalizing intuition, and in discerning the subtleties of mathematical logic (such as the distinction between the "for all" quantifier and the "there exists" quantifier), that the transition to these proofs was fairly smooth, and we were able to cover material both thoroughly and rapidly. By the tenth week, we had caught up with the non-honours class, and the students were verifying the change of variables formula for Riemann–Stieltjes integrals, and showing that piecewise continuous functions were Riemann integrable. By the conclusion of the sequence in the twentieth week, we had covered (both in lecture and in homework) the convergence theory of Taylor and Fourier series, the inverse and implicit function theorem for continuously differentiable functions of several variables, and established the dominated convergence theorem for the Lebesgue integral.

In order to cover this much material, many of the key foundational results were left to the student to prove as homework; indeed, this was an essential aspect of the course, as it ensured the students truly appreciated the concepts as they were being introduced. This format has been retained in this text; the majority of the exercises consist of proving lemmas, propositions and theorems in the main text. Indeed, I would strongly recommend that one do as many of these exercises as possible—and this includes those exercises proving "obvious" statements—if one wishes to use this text to learn real analysis; this is not a subject whose subtleties are easily appreciated just from passive reading. Most of the chapter sections have a number of exercises, which are listed at the end of the section.

To the expert mathematician, the pace of this book may seem somewhat slow, especially in early chapters, as there is a heavy emphasis on rigour (except for those discussions explicitly marked "Informal"), and justifying many steps that would ordinarily be quickly passed over as being self-evident. The first few chapters develop (in painful detail) many of the "obvious" properties of the standard number systems, for instance that the sum of two positive real numbers is again positive (Exercise 5.4.1), or that given any two distinct real numbers, one can find rational number between them (Exercise 5.4.5). In these foundational chapters, there is also an emphasis on *non-circularity*—not using later, more advanced results to prove earlier, more primitive ones. In particular, the usual laws of algebra are not used until they are derived (and they have to be derived separately for the natural numbers, integers, rationals, and reals). The reason for this is that it allows the students to learn the art of abstract reasoning, deducing true facts from a limited set of assumptions, in the friendly and intuitive setting of number systems; the payoff for this practice comes later, when one has to utilize the same type of reasoning techniques to grapple with more advanced concepts (e.g., the Lebesgue integral).

The text here evolved from my lecture notes on the subject, and thus is very much oriented towards a pedagogical perspective; much of the key material is contained inside exercises, and in many cases I have chosen to give a lengthy and tedious, but instructive, proof instead of a slick abstract proof. In more advanced textbooks, the student will see shorter and more conceptually coherent treatments of this material, and with more emphasis on intuition than on rigour; however, I feel it is important to know how to do analysis rigorously and "by hand" first, in order to truly appreciate the more modern, intuitive and abstract approach to analysis that one uses at the graduate level and beyond.

The exposition in this book heavily emphasizes rigour and formalism; however this does not necessarily mean that lectures based on this book have to proceed the same way. Indeed, in my own teaching I have used the lecture time to present the intuition behind the concepts (drawing many informal pictures and giving examples), thus providing a complementary viewpoint to the formal presentation in the text. The exercises assigned as homework provide an essential bridge between the two, requiring the student to combine both intuition and formal understanding together in order to locate correct proofs for a problem. This I found to be the most difficult task for the students, as it requires the subject to be genuinely *learnt*, rather than merely memorized or vaguely absorbed. Nevertheless, the feedback I received from the students was that the homework, while very demanding for this reason, was also very rewarding, as it allowed them to connect the rather abstract manipulations of formal mathematics with their innate intuition on such basic concepts as numbers, sets, and functions. Of course, the aid of a good teaching assistant is invaluable in achieving this connection.

With regard to examinations for a course based on this text, I would recommend either an open-book, open-notes examination with problems similar to the exercises given in the text (but perhaps shorter, with no unusual trickery involved), or else a take-home examination that involves problems comparable to the more intricate exercises in the text. The subject matter is too vast to force the students to memorize the definitions and theorems, so I would not recommend a closed-book examination, or an examination based on regurgitating extracts from the book. (Indeed, in my own examinations I gave a supplemental sheet listing the key definitions and theorems which were relevant to the examination problems.) Making the examinations similar to the homework assigned in the course will also help motivate the students to work through and understand their homework problems as thoroughly as possible (as opposed to, say, using flash cards or other such devices to memorize material), which is good preparation not only for examinations but for doing mathematics in general.

Some of the material in this textbook is somewhat peripheral to the main theme and may be omitted for reasons of time constraints. For instance, as set theory is not as fundamental to analysis as are the number systems, the chapters on set theory (Chapters 3, 8) can be covered more quickly and with substantially less rigour, or be given as reading assignments. The appendices on logic and the decimal system are intended as optional or supplemental reading and would probably not be covered in the main course lectures; the appendix on logic is particularly suitable for reading concurrently with the first few chapters. Also, Chapter 5 (on Fourier series) is not needed elsewhere in the text and can be omitted.

For reasons of length, this textbook has been split into two volumes. The first volume is slightly longer, but can be covered in about thirty lectures if the peripheral material is omitted or abridged. The second volume refers at times to the first, but can also be taught to students who have had a first course in analysis from other sources. It also takes about thirty lectures to cover.

I am deeply indebted to my students, who over the progression of the real analysis course corrected several errors in the lectures notes from which this text is derived, and gave other valuable feedback. I am also very grateful to the many anonymous referees who made several corrections and suggested many important improvements to the text. I also thank Adam, James Ameril, Quentin Batista, Biswaranjan Behara, José Antonio Lara Benítez, Dingjun Bian, Petrus Bianchi, Phillip Blagoveschensky, Tai-Danae Bradley, Brian, Eduardo Buscicchio, Carlos, cebismellim, Matheus Silva Costa, Gonzales Castillo Cristhian, Ck, William Deng, Kevin Doran, Lorenzo Dragani, EO, Florian, Gyao Gamm, Evangelos Georgiadis, Aditya Ghosh, Elie Goudout, Ti Gong, Ulrich Groh, Gökhan Güçlü, Yaver Gulusoy, Christian Gz., Kyle Hambrook, Minyoung Jeong, Bart Kleijngeld, Erik Koelink, Brett Lane, David Latorre, Matthis Lehmkühler, Bin Li, Percy Li, Ming Li, Mufei Li, Zijun Liu, Rami Luisto, Jason M., Manoranjan Majji, Mercedes Mata, Simon Mayer, Geoff Mess, Pieter Naaijkens, Vineet Nair, Jorge Peña-Vélez, Cristina Pereyra, Huaying Qiu, David Radnell, Tim Reijnders, Issa Rice, Eric Rodriquez, Pieter Roffelsen, Luke Rogers, Feras Saad, Gabriel Salmerón, Vijay Sarthak, Leopold Schlicht, Marc Schoolderman, SkysubO, Rainer aus dem Spring, Sundar, Rafał Szlendak, Karim Taya, Chaitanya Tappu, Winston Tsai, Kent Van Vels, Andrew Verras, Murtaza Wani, Daan Wanrooy, John Waters, Yandong Xiao, Sam Xu, Xueping, Hongjiang Ye, Luqing Ye, Muhammad Atif Zaheer, Zelin, and the students of Math 401/501 and Math 402/502 at the University of New Mexico for corrections to the first, second, and third editions.

Terence Tao

Preface to Subsequent Editions

Since the publication of the first edition, many students and lecturers have communicated a number of minor typos and other corrections to me. There was also some demand for a hardcover edition of the texts. Because of this, the publishers and I have decided to incorporate the corrections and issue a hardcover second edition of the textbooks. The layout, page numbering, and indexing of the texts have also been changed; in particular the two volumes are now numbered and indexed separately. However, the chapter and exercise numbering, as well as the mathematical content, remains the same as the first edition, and so the two editions can be used more or less interchangeably for homework and study purposes.

The third edition contains a number of corrections that were reported for the second edition, together with a few new exercises, but are otherwise essentially the same text. The fourth edition similarly incorporates a large number of additional corrections reported since the release of the third edition, as well as some additional exercises.

Los Angeles, USA   Terence Tao

Contents

1 Introduction 1
1.1 What Is Analysis? 1
1.2 Why Do Analysis? 2

2 Starting at the Beginning: The Natural Numbers 11
2.1 The Peano Axioms 12
2.2 Addition 20
2.3 Multiplication 24

3 Set Theory 27
3.1 Fundamentals 27
3.2 Russell's Paradox (Optional) 38
3.3 Functions 40
3.4 Images and Inverse Images 47
3.5 Cartesian Products 52
3.6 Cardinality of Sets 57

4 Integers and Rationals 63
4.1 The Integers 63
4.2 The Rationals 69
4.3 Absolute Value and Exponentiation 73
4.4 Gaps in the Rational Numbers 76

5 The Real Numbers 81
5.1 Cauchy Sequences 82
5.2 Equivalent Cauchy Sequences 86
5.3 The Construction of the Real Numbers 88
5.4 Ordering the Reals 95
5.5 The Least Upper Bound Property 100
5.6 Real Exponentiation, Part I 104

6 Limits of Sequences . . . 109
6.1 Convergence and Limit Laws . . . 109
6.2 The Extended Real Number System . . . 115
6.3 Suprema and Infima of Sequences . . . 118
6.4 Limsup, Liminf, and Limit Points . . . 120
6.5 Some Standard Limits . . . 127
6.6 Subsequences . . . 128
6.7 Real Exponentiation, Part II . . . 131

7 Series . . . 135
7.1 Finite Series . . . 135
7.2 Infinite Series . . . 144
7.3 Sums of Non-negative Numbers . . . 148
7.4 Rearrangement of Series . . . 152
7.5 The Root and Ratio Tests . . . 155

8 Infinite Sets . . . 159
8.1 Countability . . . 159
8.2 Summation on Infinite Sets . . . 165
8.3 Uncountable Sets . . . 171
8.4 The Axiom of Choice . . . 173
8.5 Ordered Sets . . . 176

9 Continuous Functions on R . . . 183
9.1 Subsets of the Real Line . . . 183
9.2 The Algebra of Real-Valued Functions . . . 189
9.3 Limiting Values of Functions . . . 191
9.4 Continuous Functions . . . 197
9.5 Left and Right Limits . . . 200
9.6 The Maximum Principle . . . 203
9.7 The Intermediate Value Theorem . . . 206
9.8 Monotonic Functions . . . 208
9.9 Uniform Continuity . . . 210
9.10 Limits at Infinity . . . 215

10 Differentiation of Functions . . . 217
10.1 Basic Definitions . . . 217
10.2 Local Maxima, Local Minima, and Derivatives . . . 222
10.3 Monotone Functions and Derivatives . . . 224
10.4 Inverse Functions and Derivatives . . . 225
10.5 L'Hôpital's Rule . . . 228

11 The Riemann Integral . . . 231
11.1 Partitions . . . 232
11.2 Piecewise Constant Functions . . . 235
11.3 Upper and Lower Riemann Integrals . . . 238
11.4 Basic Properties of the Riemann Integral . . . 242

11.5 Riemann Integrability of Continuous Functions 246
11.6 Riemann Integrability of Monotone Functions 249
11.7 A Non-riemann Integrable Function 251
11.8 The Riemann–Stieltjes Integral 252
11.9 The Two Fundamental Theorems of Calculus 255
11.10 Consequences of the Fundamental Theorems 259

Appendix A: The Basics of Mathematical Logic 265

Appendix B: The Decimal System 287

Index 295

About the Author

Terence Tao has been a professor of Mathematics at the University of California Los Angeles (UCLA), USA, since 1999, having completed his Ph.D. under Prof. Elias Stein at Princeton University, USA, in 1996. Tao's areas of research include harmonic analysis, partial differential equations, combinatorics, and number theory. He has received a number of awards, including the Salem Prize in 2000, the Bochner Prize in 2002, the Fields Medal in 2006, the MacArthur Fellowship in 2007, the Waterman Award in 2008, the Nemmers Prize in 2010, the Crafoord Prize in 2012, and the Breakthrough Prize in Mathematics in 2015. Terence Tao also currently holds the James and Carol Collins chair in Mathematics at UCLA and is a fellow of the Royal Society, the Australian Academy of Sciences (the corresponding member), the National Academy of Sciences (a foreign member), and the American Academy of Arts and Sciences. He was born in Adelaide, Australia, in 1975.

Chapter 1
Introduction

1.1 What Is Analysis?

This text is an honors-level undergraduate introduction to *real analysis*: the analysis of the real numbers, sequences and series of real numbers, and real-valued functions. This is related to, but is distinct from, *complex analysis*, which concerns the analysis of the complex numbers and complex functions, *harmonic analysis*, which concerns the analysis of harmonics (waves) such as sine waves, and how they synthesize other functions via the Fourier transform, *functional analysis*, which focuses much more heavily on functions (and how they form things like vector spaces), and so forth. *Analysis* is the rigorous study of such objects, with a focus on trying to pin down precisely and accurately the qualitative and quantitative behavior of these objects. Real analysis is the theoretical foundation which underlies *calculus*, which is the collection of computational algorithms which one uses to manipulate functions.

In this text we will be studying many objects which will be familiar to you from freshman calculus: numbers, sequences, series, limits, functions, definite integrals, derivatives, and so forth. You already have a great deal of experience of *computing* with these objects; however here we will be focused more on the underlying theory for these objects. We will be concerned with questions such as the following:

1. What is a real number? Is there a largest real number? After 0, what is the "next" real number (i.e., what is the smallest positive real number)? Can you cut a real number into pieces infinitely many times? Why does a number such as 2 have a square root, while a number such as -2 does not? If there are infinitely many reals and infinitely many rationals, how come there are "more" real numbers than rational numbers?
2. How do you take the limit of a sequence of real numbers? Which sequences have limits and which ones don't? If you can stop a sequence from escaping to infinity, does this mean that it must eventually settle down and converge? Can you add infinitely many real numbers together and still get a finite real number? Can you add infinitely many rational numbers together and end up with a non-rational

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_1

number? If you rearrange the elements of an infinite sum, is the sum still the same?

3. What is a function? What does it mean for a function to be continuous? differentiable? integrable? bounded? Can you add infinitely many functions together? What about taking limits of sequences of functions? Can you differentiate an infinite series of functions? What about integrating? If a function $f(x)$ takes the value 3 when $x = 0$ and 5 when $x = 1$ (i.e., $f(0) = 3$ and $f(1) = 5$), does it have to take every intermediate value between 3 and 5 when x goes between 0 and 1? Why?

You may already know how to answer some of these questions from your calculus classes, but most likely these sorts of issues were only of secondary importance to those courses; the emphasis was on getting you to perform computations, such as computing the integral of $x\sin(x^2)$ from $x = 0$ to $x = 1$. But now that you are comfortable with these objects and already know how to do all the computations, we will go back to the theory and try to *really* understand what is going on.

1.2 Why Do Analysis?

It is a fair question to ask, "why bother?", when it comes to analysis. There is a certain philosophical satisfaction in knowing *why* things work, but a pragmatic person may argue that one only needs to know *how* things work to do real-life problems. The calculus training you receive in introductory classes is certainly adequate for you to begin solving many problems in physics, chemistry, biology, economics, computer science, finance, engineering, or whatever else you end up doing—and you can certainly use things like the chain rule, L'Hôpital's rule, or integration by parts without knowing why these rules work, or whether there are any exceptions to these rules. However, one can get into trouble if one applies rules without knowing where they came from and what the limits of their applicability are. Let me give some examples in which several of these familiar rules, if applied blindly without knowledge of the underlying analysis, can lead to disaster.

Example 1.2.1 (Division by zero). This is a very familiar one to you: the cancellation law $ac = bc \implies a = b$ does not work when $c = 0$. For instance, the identity $1 \times 0 = 2 \times 0$ is true, but if one blindly cancels the 0 then one obtains $1 = 2$, which is false. In this case it was obvious that one was dividing by zero; but in other cases it can be more hidden.

Example 1.2.2 (Divergent series). You have probably seen geometric series such as the infinite sum

$$S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots .$$

You have probably seen the following trick to sum this series: if we call the above sum S, then if we multiply both sides by 2, we obtain

$$2S = 2 + 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 2 + S$$

and hence $S = 2$, so the series sums to 2. However, if you apply the same trick to the series

$$S = 1 + 2 + 4 + 8 + 16 + \cdots$$

one gets nonsensical results:

$$2S = 2 + 4 + 8 + 16 + \cdots = S - 1 \implies S = -1.$$

So the same reasoning that shows that $1 + \frac{1}{2} + \frac{1}{4} + \cdots = 2$ also gives that $1 + 2 + 4 + 8 + \cdots = -1$. Why is it that we trust the first equation but not the second? A similar example arises with the series

$$S = 1 - 1 + 1 - 1 + 1 - 1 + \cdots;$$

we can write

$$S = 1 - (1 - 1 + 1 - 1 + \cdots) = 1 - S$$

and hence that $S = 1/2$; or instead we can write

$$S = (1 - 1) + (1 - 1) + (1 - 1) + \cdots = 0 + 0 + \cdots$$

and hence that $S = 0$; or instead we can write

$$S = 1 + (-1 + 1) + (-1 + 1) + \cdots = 1 + 0 + 0 + \cdots$$

and hence that $S = 1$. Which one is correct? (See Exercise 7.2.1 for an answer.)

Example 1.2.3 (Divergent sequences). Here is a slight variation of the previous example. Let x be a real number, and let L be the limit

$$L = \lim_{n\to\infty} x^n.$$

Changing variables $n = m + 1$, we have

$$L = \lim_{m+1\to\infty} x^{m+1} = \lim_{m+1\to\infty} x \times x^m = x \lim_{m+1\to\infty} x^m.$$

But if $m + 1 \to \infty$, then $m \to \infty$, thus

$$\lim_{m+1\to\infty} x^m = \lim_{m\to\infty} x^m = \lim_{n\to\infty} x^n = L,$$

and thus

$$xL = L.$$

At this point we could cancel the L's and conclude that $x = 1$ for an arbitrary real number x, which is absurd. But since we are already aware of the division by zero problem, we could be a little smarter and conclude instead that either $x = 1$, or $L = 0$. In particular we seem to have shown that

$$\lim_{n\to\infty} x^n = 0 \text{ for all } x \neq 1.$$

But this conclusion is absurd if we apply it to certain values of x, for instance by specializing to the case $x = 2$ we could conclude that the sequence $1, 2, 4, 8, \ldots$ converges to zero, and by specializing to the case $x = -1$ we conclude that the sequence $1, -1, 1, -1, \ldots$ also converges to zero. These conclusions appear to be absurd; what is the problem with the above argument? (See Exercise 6.3.4 for an answer.)

Example 1.2.4 (Limiting values of functions). Start with the expression $\lim_{x\to\infty} \sin(x)$, make the change of variable $x = y + \pi$ and recall that $\sin(y + \pi) = -\sin(y)$ to obtain

$$\lim_{x\to\infty} \sin(x) = \lim_{y+\pi\to\infty} \sin(y + \pi) = \lim_{y\to\infty} (-\sin(y)) = -\lim_{y\to\infty} \sin(y).$$

Since $\lim_{x\to\infty} \sin(x) = \lim_{y\to\infty} \sin(y)$ we thus have

$$\lim_{x\to\infty} \sin(x) = -\lim_{x\to\infty} \sin(x)$$

and hence

$$\lim_{x\to\infty} \sin(x) = 0.$$

If we then make the change of variables $x = \pi/2 + z$ and recall that $\sin(\pi/2 + z) = \cos(z)$ we conclude that

$$\lim_{x\to\infty} \cos(x) = 0.$$

Squaring both of these limits and adding we see that

$$\lim_{x\to\infty} (\sin^2(x) + \cos^2(x)) = 0^2 + 0^2 = 0.$$

On the other hand, we have $\sin^2(x) + \cos^2(x) = 1$ for all x. Thus we have shown that $1 = 0$! What is the difficulty here?

Example 1.2.5 (Interchanging sums). Consider the following fact of arithmetic. Consider any matrix of numbers, e.g.,

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

and compute the sums of all the rows and the sums of all the columns, and then total all the row sums and total all the column sums. In both cases you will get the same number—the total sum of all the entries in the matrix:

$$\begin{array}{cc} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} & \begin{matrix} 6 \\ 15 \\ 24 \end{matrix} \\ \begin{matrix} 12 & 15 & 18 \end{matrix} & 45 \end{array}$$

To put it another way, if you want to add all the entries in an $m \times n$ matrix together, it doesn't matter whether you sum the rows first or sum the columns first, you end up with the same answer. (Before the invention of computers, accountants and bookkeepers would use this fact to guard against making errors when balancing their books.) In series notation, this fact would be expressed as

$$\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij},$$

if a_{ij} denoted the entry in the ith row and jth column of the matrix.

Now one might think that this rule should extend easily to infinite series:

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}.$$

Indeed, if you use infinite series a lot in your work, you will find yourself having to switch summations like this fairly often. Another way of saying this fact is that in an infinite matrix, the sum of the row totals should equal the sum of the column totals. However, despite the reasonableness of this statement, it is actually false! Here is a counterexample:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \dots \\ -1 & 1 & 0 & 0 \dots \\ 0 & -1 & 1 & 0 \dots \\ 0 & 0 & -1 & 1 \dots \\ 0 & 0 & 0 & -1 \dots \\ \vdots & \vdots & \vdots & \vdots \ddots \end{pmatrix}.$$

If you sum up all the rows, and then add up all the row totals, you get 1; but if you sum up all the columns, and add up all the column totals, you get 0! So, does this mean that summations for infinite series should not be swapped and that any argument using such a swapping should be distrusted? (See Theorem 8.2.2 for an answer.)

Example 1.2.6 (Interchanging integrals). The interchanging of integrals is a trick which occurs in mathematics just as commonly as the interchanging of sums. Suppose one wants to compute the volume under a surface $z = f(x, y)$ (let us ignore the limits of integration for the moment). One can do it by slicing parallel to the x-axis: for each fixed value of y, we can compute an area $\int f(x, y)\, \mathrm{d}x$, and then we integrate the area in the y variable to obtain the volume

$$V = \int \int f(x, y) \mathrm{d}x \mathrm{d}y.$$

Or we could slice parallel to the y-axis for each fixed x and compute an area $\int f(x, y)\, \mathrm{d}y$ and then integrate in the x-axis to obtain

$$V = \int \int f(x, y) \mathrm{d}y \mathrm{d}x.$$

This seems to suggest that one should always be able to swap integral signs:

$$\int \int f(x, y)\, \mathrm{d}x \mathrm{d}y = \int \int f(x, y)\, \mathrm{d}y \mathrm{d}x.$$

And indeed, people swap integral signs all the time, because sometimes one variable is easier to integrate in first than the other. However, just as infinite sums sometimes cannot be swapped, integrals are also sometimes dangerous to swap. An example is with the integrand $\mathrm{e}^{-xy} - xy\mathrm{e}^{-xy}$. Suppose we believe that we can swap the integrals:

$$\int_0^\infty \int_0^1 (\mathrm{e}^{-xy} - xy\mathrm{e}^{-xy})\, \mathrm{d}y\, \mathrm{d}x = \int_0^1 \int_0^\infty (\mathrm{e}^{-xy} - xy\mathrm{e}^{-xy})\, \mathrm{d}x\, \mathrm{d}y. \tag{1.1}$$

Since

$$\int_0^1 (\mathrm{e}^{-xy} - xy\mathrm{e}^{-xy})\, \mathrm{d}y = y\mathrm{e}^{-xy}|_{y=0}^{y=1} = \mathrm{e}^{-x},$$

the left-hand side of (1.1) is $\int_0^\infty \mathrm{e}^{-x}\, \mathrm{d}x = -\mathrm{e}^{-x}|_0^\infty = 1$. But since

$$\int_0^\infty (\mathrm{e}^{-xy} - xy\mathrm{e}^{-xy})\, \mathrm{d}x = x\mathrm{e}^{-xy}|_{x=0}^{x=\infty} = 0,$$

the right-hand side of (1.1) is $\int_0^1 0\, \mathrm{d}x = 0$. Clearly $1 \neq 0$, so there is an error somewhere; but you won't find one anywhere except in the step where we interchanged the integrals. So how do we know when to trust the interchange of integrals? (See Theorem 8.5.1 of *Analysis II* for a partial answer.)

Example 1.2.7 (Interchanging limits). Suppose we start with the plausible looking statement

$$\lim_{x\to 0}\lim_{y\to 0}\frac{x^2}{x^2+y^2} = \lim_{y\to 0}\lim_{x\to 0}\frac{x^2}{x^2+y^2}. \tag{1.2}$$

But we have

$$\lim_{y\to 0}\frac{x^2}{x^2+y^2} = \frac{x^2}{x^2+0^2} = 1,$$

so the left-hand side of (1.2) is 1; on the other hand, we have

$$\lim_{x\to 0}\frac{x^2}{x^2+y^2} = \frac{0^2}{0^2+y^2} = 0,$$

so the right-hand side of (1.2) is 0. Since 1 is clearly not equal to zero, this suggests that interchange of limits is untrustworthy. But are there any other circumstances in which the interchange of limits is legitimate? (See Exercise 2.2.9 of *Analysis II* for a partial answer.)

Example 1.2.8 (Interchanging limits, again). Consider the plausible looking statement

$$\lim_{x\to 1^-}\lim_{n\to\infty} x^n = \lim_{n\to\infty}\lim_{x\to 1^-} x^n$$

where the notation $x \to 1^-$ means that x is approaching 1 from the left. When x is to the left of 1, then $\lim_{n\to\infty} x^n = 0$, and hence the left-hand side is zero. But we also have $\lim_{x\to 1^-} x^n = 1$ for all n, and so the right-hand side limit is 1. Does this demonstrate that this type of limit interchange is always untrustworthy? (See Proposition 3.3.3 of *Analysis II* for an answer.)

Example 1.2.9 (Interchanging limits and integrals). For any real number y, we have

$$\int_{-\infty}^{\infty}\frac{1}{1+(x-y)^2}\,\mathrm{d}x = \arctan(x-y)|_{x=-\infty}^{\infty} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

Taking limits as $y \to \infty$, we should obtain

$$\int_{-\infty}^{\infty}\lim_{y\to\infty}\frac{1}{1+(x-y)^2}\,\mathrm{d}x = \lim_{y\to\infty}\int_{-\infty}^{\infty}\frac{1}{1+(x-y)^2}\,\mathrm{d}x = \pi.$$

But for every x, we have $\lim_{y\to\infty}\frac{1}{1+(x-y)^2} = 0$. So we seem to have concluded that $0 = \pi$. What was the problem with the above argument? Should one abandon the (very useful) technique of interchanging limits and integrals? (See Theorem 3.6.1 of *Analysis II* for a partial answer.)

Example 1.2.10 (Interchanging limits and derivatives). Observe that if $\varepsilon > 0$, then

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^3}{\varepsilon^2+x^2}\right) = \frac{3x^2(\varepsilon^2+x^2)-2x^4}{(\varepsilon^2+x^2)^2}$$

and in particular that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^3}{\varepsilon^2+x^2}\right)|_{x=0} = 0.$$

Taking limits as $\varepsilon \to 0$, one might then expect that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^3}{0+x^2}\right)|_{x=0} = 0.$$

But the right-hand side is $\frac{\mathrm{d}}{\mathrm{d}x}x = 1$. Does this mean that it is always illegitimate to interchange limits and derivatives? (See Theorem 3.7.1 of *Analysis II* for an answer.)

Example 1.2.11 (Interchanging derivatives). Let[1] $f(x, y)$ be the function $f(x, y) := \frac{xy^3}{x^2+y^2}$. A common maneuver in analysis is to interchange two partial derivatives, thus one expects

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = \frac{\partial^2 f}{\partial y \partial x}(0, 0).$$

But from the quotient rule we have

$$\frac{\partial f}{\partial y}(x, y) = \frac{3xy^2}{x^2+y^2} - \frac{2xy^4}{(x^2+y^2)^2}$$

and in particular

$$\frac{\partial f}{\partial y}(x, 0) = \frac{0}{x^2} - \frac{0}{x^4} = 0.$$

Thus

$$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = 0.$$

On the other hand, from the quotient rule again we have

$$\frac{\partial f}{\partial x}(x, y) = \frac{y^3}{x^2+y^2} - \frac{2x^2y^3}{(x^2+y^2)^2}$$

[1] One might object that this function is not defined at $(x, y) = (0, 0)$, but if we set $f(0, 0) := 0$ then this function becomes continuous and differentiable for all (x, y), and in fact both partial derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ are also continuous and differentiable for all (x, y)!

and hence

$$\frac{\partial f}{\partial x}(0, y) = \frac{y^3}{y^2} - \frac{0}{y^4} = y.$$

Thus

$$\frac{\partial^2 f}{\partial y \partial x}(0, 0) = 1.$$

Since $1 \neq 0$, we thus seem to have shown that interchange of derivatives is untrustworthy. But are there any other circumstances in which the interchange of derivatives is legitimate? (See Theorem 6.5.4 and Exercise 6.5.1 of *Analysis II* for some answers.)

Example 1.2.12 (L'Hôpital's rule). We are all familiar with the beautifully simple L'Hôpital's rule

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f'(x)}{g'(x)},$$

but one can still get led to incorrect conclusions if one applies it incorrectly. For instance, applying it to $f(x) := x$, $g(x) := 1 + x$, and $x_0 := 0$ we would obtain

$$\lim_{x \to 0} \frac{x}{1 + x} = \lim_{x \to 0} \frac{1}{1} = 1,$$

but this is the incorrect answer, since $\lim_{x \to 0} \frac{x}{1+x} = \frac{0}{1+0} = 0$. Of course, all that is going on here is that L'Hôpital's rule is only applicable when both $f(x)$ and $g(x)$ go to zero as $x \to x_0$, a condition which was violated in the above example. But even when $f(x)$ and $g(x)$ do go to zero as $x \to x_0$ there is still a possibility for an incorrect conclusion. For instance, consider the limit

$$\lim_{x \to 0} \frac{x^2 \sin(x^{-4})}{x}.$$

Both numerator and denominator go to zero as $x \to 0$, so it seems pretty safe to apply L'Hôpital's rule, to obtain

$$\begin{aligned} \lim_{x \to 0} \frac{x^2 \sin(x^{-4})}{x} &= \lim_{x \to 0} \frac{2x \sin(x^{-4}) - 4x^{-3} \cos(x^{-4})}{1} \\ &= \lim_{x \to 0} 2x \sin(x^{-4}) - \lim_{x \to 0} 4x^{-3} \cos(x^{-4}). \end{aligned}$$

The first limit converges to zero by the squeeze test (since the function $2x \sin(x^{-4})$ is bounded above by $2|x|$ and below by $-2|x|$, both of which go to zero at 0). But the second limit is divergent (because x^{-3} goes to infinity as $x \to 0$, and $\cos(x^{-4})$ does not go to zero). So the limit $\lim_{x \to 0} \frac{2x \sin(x^{-4}) - 4x^{-2} \cos(x^{-4})}{1}$ diverges. One might then conclude using L'Hôpital's rule that $\lim_{x \to 0} \frac{x^2 \sin(x^{-4})}{x}$ also diverges; however we can clearly rewrite this limit as $\lim_{x \to 0} x \sin(x^{-4})$, which goes to zero when $x \to 0$ by the squeeze test again. This does not show that L'Hôpital's rule is untrustworthy (indeed, it is quite rigorous; see Sect. 10.5), but it still requires some care when applied.

Example 1.2.13 (Limits and lengths). When you learn about integration and how it relates to the area under a curve, you were probably presented with some picture in which the area under the curve was approximated by a bunch of rectangles, whose area was given by a Riemann sum, and then one somehow "took limits" to replace that Riemann sum with an integral, which then presumably matched the actual area under the curve. Perhaps a little later, you learnt how to compute the length of a curve by a similar method—approximate the curve by a bunch of line segments, compute the length of all the line segments, and then take limits again to see what you get.

However, it should come as no surprise by now that this approach also can lead to nonsense if used incorrectly. Consider the right-angled triangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$, and suppose we wanted to compute the length of the hypotenuse of this triangle. Pythagoras' theorem tells us that this hypotenuse has length $\sqrt{2}$, but suppose for some reason that we did not know about Pythagoras' theorem, and wanted to compute the length using calculus methods. Well, one way to do so is to approximate the hypotenuse by horizontal and vertical edges. Pick a large number N, and approximate the hypotenuse by a "staircase" consisting of N horizontal edges of equal length, alternating with N vertical edges of equal length. Clearly these edges all have length $1/N$, so the total length of the staircase is $2N/N = 2$. If one takes limits as N goes to infinity, the staircase clearly approaches the hypotenuse, and so in the limit we should get the length of the hypotenuse. However, as $N \to \infty$, the limit of $2N/N$ is 2, not $\sqrt{2}$, so we have an incorrect value for the length of the hypotenuse. How did this happen?

The analysis you learn in this text will help you resolve these questions, and will let you know when these rules (and others) are justified, and when they are illegal, thus separating the useful applications of these rules from the nonsense. Thus they can prevent you from making mistakes and can help you place these rules in a wider context. Moreover, as you learn analysis you will develop an "analytical way of thinking", which will help you whenever you come into contact with any new rules of mathematics, or when dealing with situations which are not quite covered by the standard rules. For instance, what if your functions are complex-valued instead of real-valued? What if you are working on the sphere instead of the plane? What if your functions are not continuous, but are instead things like square waves and delta functions? What if your functions, or limits of integration, or limits of summation, are occasionally infinite? You will develop a sense of *why* a rule in mathematics (e.g., the chain rule) works, how to adapt it to new situations, and what its limitations (if any) are; this will allow you to apply the mathematics you have already learnt more confidently and correctly.

Chapter 2
Starting at the Beginning: The Natural Numbers

In this text, we will review the material you have learnt in high school and in elementary calculus classes, but as rigorously as possible. To do so we will have to begin at the very basics - indeed, we will go back to the concept of *numbers* and what their properties are. Of course, you have dealt with numbers for over ten years and you know how to manipulate the rules of algebra to simplify any expression involving numbers, but we will now turn to a more fundamental issue, which is: *why* do the rules of algebra work at all? For instance, why is it true that $a(b + c)$ is equal to $ab + ac$ for any three numbers a, b, c? This is not an arbitrary choice of rule; it can be proven from more primitive, and more fundamental, properties of the number system. This will teach you a new skill - how to prove complicated properties from simpler ones. You will find that even though a statement may be "obvious", it may not be easy to prove; the material here will give you plenty of practice in doing so, and in the process will lead you to think about *why* an obvious statement really is obvious. One skill in particular that you will pick up here is the use of *mathematical induction*, which is a basic tool in proving things in many areas of mathematics.

So in the first few chapters we will re-acquaint you with various number systems that are used in real analysis. In increasing order of sophistication, they are the *natural numbers* **N**; the *integers* **Z**; the *rationals* **Q**, and the *real numbers* **R**. (There are other number systems such as the *complex numbers* **C**, but we will not study them until Sect. 4.6.) The natural numbers $\{0, 1, 2, \ldots\}$ are the most primitive of the number systems, but they are used to build the integers, which in turn are used to build the rationals. Furthermore, the rationals are used to build the real numbers, which are in turn used to build the complex numbers. Thus to begin at the very beginning, we must look at the natural numbers. We will consider the following question: how does one actually *define* the natural numbers? (This is a very different question from how to *use* the natural numbers, which is something you of course know how to do very well. It's like the difference between knowing how to use, say, a computer, versus knowing how to *build* that computer.)

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_2

This question is more difficult to answer than it looks. The basic problem is that you have used the natural numbers for so long that they are embedded deeply into your mathematical thinking, and you can make various implicit assumptions about these numbers (e.g., that $a + b$ is always equal to $b + a$) without even being aware that you are doing so; it is difficult to let go and try to inspect this number system as if it is the first time you have seen it. So in what follows I will have to ask you to perform a rather difficult task: try to set aside, for the moment, everything you know about the natural numbers; forget that you know how to count, to add, to multiply, to manipulate the rules of algebra, etc. We will try to introduce these concepts one at a time and identify explicitly what our assumptions are as we go along—and not allow ourselves to use more "advanced" tricks such as the rules of algebra until we have actually proven them. This may seem like an irritating constraint, especially as we will spend a lot of time proving statements which are "obvious", but it is necessary to do this suspension of known facts to avoid *circularity* (e.g., using an advanced fact to prove a more elementary fact, and then later using the elementary fact to prove the advanced fact). Also, this exercise will be an excellent way to affirm the foundations of your mathematical knowledge. Furthermore, practicing your proofs and abstract thinking here will be invaluable when we move on to more advanced concepts, such as real numbers, functions, sequences and series, differentials and integrals, and so forth. In short, the results here may seem trivial, but the journey is much more important than the destination, for now. (Once the number systems are constructed properly, we can resume using the laws of algebra, etc., without having to rederive them each time.)

We will also forget that we know the decimal system, which of course is an extremely convenient way to manipulate numbers, but it is not something which is fundamental to what numbers are. (For instance, one could use an octal or binary system instead of the decimal system, or even the Roman numeral system, and still get exactly the same set of numbers.) Besides, if one tries to fully explain what the decimal number system is, it isn't as natural as you might think. Why is 00423 the same number as 423, but 32400 isn't the same number as 324? Why is 123.4444... a real number, while ...444.321 is not? And why do we have to carry of digits when adding or multiplying? Why is 0.999... the same number as 1? What is the smallest positive real number? Isn't it just 0.00...001? So to set aside these problems, we will not try to assume any knowledge of the decimal system, though we will of course still refer to numbers by their familiar names such as 1, 2, and 3 instead of using other notation such as I, II, III or 0++, (0++)++, ((0++)++)++ (see below) so as not to be needlessly artificial. For completeness, we review the decimal system in Appendix B.

2.1 The Peano Axioms

We now present one standard way to define the natural numbers, in terms of the *Peano axioms*, which were first laid out by Giuseppe Peano (1858–1932). This is not the only way to define the natural numbers. For instance, another approach is to talk about the cardinality of finite sets; for instance one could take a set of five elements and define 5 to be the number of elements in that set. We shall discuss this alternate approach in Sect. 3.6. However, we shall stick with the Peano axiomatic approach for now.

How are we to define what the natural numbers are? Informally, we could say

Definition 2.1.1 (*Informal*) A *natural number* is any element of the set

$$\mathbf{N} := \{0, 1, 2, 3, 4, \ldots\},$$

which is the set of all the numbers created by starting with 0 and then counting forward indefinitely. We call **N** the *set of natural numbers*.

Remark 2.1.2 In some texts the natural numbers start at 1 instead of 0, but this is a matter of notational convention more than anything else. In this text we shall refer to the set $\{1, 2, 3, \ldots\}$ as the *positive integers* $\mathbf{Z}^+$ rather than the natural numbers. Natural numbers are sometimes also known as *whole numbers*.

In a sense, this definition solves the problem of what the natural numbers are: a natural number is any element of the set[1] **N**. However, it is not really that satisfactory, because it begs the question of what **N** is. This definition of "start at 0 and count indefinitely" seems like an intuitive enough definition of **N**, but it is not entirely acceptable, because it leaves many questions unanswered. For instance: how do we know we can keep counting indefinitely, without cycling back to 0? Also, how do you perform operations such as addition, multiplication, or exponentiation?

We can answer the latter question first: we can define complicated operations in terms of simpler operations. Exponentiation is nothing more than repeated multiplication: 5^3 is nothing more than three fives multiplied together. Multiplication is nothing more than repeated addition; 5×3 is nothing more than three fives added together. (Subtraction and division will not be covered here, because they are not operations which are well-suited to the natural numbers; they will have to wait for the integers and rationals, respectively.) And addition? It is nothing more than the repeated operation of *counting forward*, or *incrementing*. If you add three to five, what you are doing is incrementing five three times. On the other hand, incrementing seems to be a fundamental operation, not reducible to any simpler operation; indeed, it is the first operation one learns on numbers, even before learning to add.

Thus, to define the natural numbers, we will use two fundamental concepts: the zero number 0 and the increment operation (also known as the *successor operation*). In deference to modern computer languages, we will use $n{+}{+}$ to denote[2] the

[1] Strictly speaking, there is another problem with this informal definition: we have not yet defined what a "set" is or what "element of" is. Thus for the rest of this chapter we shall avoid mention of sets and their elements as much as possible, except in informal discussion.

[2] The notation Sn or $S(n)$ is also often used in the literature to denote the successor $n{+}{+}$ of n. One may be tempted to use the more familiar notation $n + 1$ in place of $n{+}{+}$ to denote the successor of n, but this would introduce a circularity in our foundations, since the notion of addition will be defined in terms of the successor operation.

increment or *successor* of n, thus for instance $3{++} = 4$, $(3{++}){++} = 5$, etc. This is a slightly different usage from that in computer languages such as C, where $n{++}$ actually *redefines* the value of n to be its successor; however in mathematics we try not to define a variable more than once in any given setting, as it can often lead to confusion; many of the statements which were true for the old value of the variable can now become false, and vice versa.

So, it seems like we want to say that $\mathbf{N}$ consists of 0 and everything which can be obtained from 0 by incrementing: $\mathbf{N}$ should consist of the objects

$$0, 0{++}, (0{++}){++}, ((0{++}){++}){++}, \text{etc.}$$

If we start writing down what this means about the natural numbers, we thus see that we should have the following axioms concerning 0 and the increment operation ++:

Axiom 2.1 0 is a natural number.

Axiom 2.2 If n is a natural number, then $n{++}$ is also a natural number.

Thus for instance, from Axiom 2.1 and two applications of Axiom 2.2, we see that $(0{++}){++}$ is a natural number. Of course, this notation will begin to get unwieldy, so we adopt a convention to write these numbers in more familiar notation:

Definition 2.1.3 We define[3] 1 to be the number $0{++}$, 2 to be the number $(0{++}){++}$, 3 to be the number $((0{++}){++}){++}$, etc. (In other words, $1 := 0{++}$, $2 := 1{++}$, $3 := 2{++}$, etc. In this text I use "$x := y$" to denote the statement that x is *defined* to equal y.)

Thus for instance, we have

Proposition 2.1.4 *3 is a natural number.*

Proof By Axiom 2.1, 0 is a natural number. By Axiom 2.2, $0{++} = 1$ is a natural number. By Axiom 2.2 again, $1{++} = 2$ is a natural number. By Axiom 2.2 again, $2{++} = 3$ is a natural number. □

It may seem that this is enough to describe the natural numbers. However, we have not pinned down completely the behavior of $\mathbf{N}$:

Example 2.1.5 Consider a number system which consists of the numbers 0, 1, 2, 3, in which the increment operation wraps back from 3 to 0. More precisely $0{++}$ is equal to 1, $1{++}$ is equal to 2, $2{++}$ is equal to 3, but $3{++}$ is equal to 0 (and also equal to 4, by definition of 4). This type of thing actually happens in real life, when one uses a computer to try to store a natural number: if one starts at 0 and performs the increment operation repeatedly, eventually the computer will overflow its memory and the number will wrap around back to 0 (though this may take quite a large number

[3] This convention is actually an oversimplification. To see how to properly merge the usual decimal notation for numbers with the natural numbers given by the Peano axioms, see Appendix B.

of incrementation operations, for instance a two-byte representation of an integer will wrap around only after 65,536 increments). Note that this type of number system obeys Axiom 2.1 and Axiom 2.2, even though it clearly does not correspond to what we intuitively believe the natural numbers to be like.

To prevent this sort of "wrap-around issue" we will impose another axiom:

Axiom 2.3 0 is not the successor of any natural number; i.e., we have $n{+}{+} \neq 0$ for every natural number n.

Now we can show that certain types of wrap around do not occur: for instance we can now rule out the type of behavior in Example 2.1.5 using.

Proposition 2.1.6 4 *is not equal to* 0.

Don't laugh! Because of the way we have defined 4—it is the increment of the increment of the increment of the increment of 0—it is not necessarily true *a priori* that this number is not the same as zero, even if it is "obvious". ("*a priori*" is Latin for "beforehand"—it refers to what one already knows or assumes to be true before one begins a proof or argument. The opposite is "*a posteriori*"—what one knows to be true after the proof or argument is concluded.) Note for instance that in Example 2.1.5, 4 was indeed equal to 0, and that in a standard two-byte computer representation of a natural number, for instance, 65,536 is equal to 0 (using our definition of 65,536 as equal to 0 incremented sixty-five thousand, five hundred and thirty-six times).

Proof By definition, $4 = 3{+}{+}$. By Axioms 2.1 and 2.2, 3 is a natural number. Thus by Axiom 2.3, $3{+}{+} \neq 0$, i.e., $4 \neq 0$. □

However, even with our new axiom, it is still possible that our number system behaves in other pathological ways:

Example 2.1.7 Consider a number system consisting of five numbers 0, 1, 2, 3, 4, in which the increment operation hits a "ceiling" at 4. More precisely, suppose that $0{+}{+} = 1$, $1{+}{+} = 2$, $2{+}{+} = 3$, $3{+}{+} = 4$, but $4{+}{+} = 4$ (or in other words that $5 = 4$, and hence $6 = 4$, $7 = 4$, etc.). This does not contradict Axioms 2.1, 2.2 and 2.3. Another number system with a similar problem is one in which incrementation wraps around, but not to zero, e.g., suppose that $4{+}{+} = 1$ (so that $5 = 1$, then $6 = 2$, etc.).

There are many ways to prohibit the above types of behavior from happening, but one of the simplest is to assume the following axiom:

Axiom 2.4 Different natural numbers must have different successors; i.e., if n, m are natural numbers and $n \neq m$, then $n{+}{+} \neq m{+}{+}$. Equivalently,[4] if $n{+}{+} = m{+}{+}$ then we must have $n = m$.

[4] This is an example of reformulating an implication using its *contrapositive*; see Sect. A.2 for more details. In the converse direction, if $n = m$, then $n{+}{+} = m{+}{+}$; this is the *axiom of substitution* (see Sect. A.7) applied to the operation ++.

Thus, for instance, we have

Proposition 2.1.8 6 *is not equal to* 2.

Proof Suppose for sake of contradiction that $6 = 2$. Then $5{++} = 1{++}$, so by Axiom 2.4 we have $5 = 1$, so that $4{++} = 0{++}$. By Axiom 2.4 again we then have $4 = 0$, which contradicts our previous proposition. □

As one can see from this proposition, it now looks like we can keep all of the natural numbers distinct from each other. There is however still one more problem: while the axioms (particularly Axioms 2.3 and 2.4) allow us to confirm that $0, 1, 2, 3, \ldots$ are distinct elements of $\mathbf{N}$, there is the problem that there may be other "rogue" elements in our number system which are not of this form:

Example 2.1.9 (Informal) Suppose that our number system $\mathbf{N}$ consisted of the following collection of integers and half-integers:

$$\mathbf{N} := \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, \ldots\}.$$

(This example is marked "informal" since we are using real numbers, which we're not supposed to use yet.) One can check that Axioms 2.1–2.4 are still satisfied for this set.

What we want is some axiom which says that the only numbers in $\mathbf{N}$ are those which can be obtained from 0 and the increment operation—in order to exclude elements such as 0.5. But it is difficult to quantify what we mean by "can be obtained from" without already using the natural numbers, which we are trying to define. Fortunately, there is an ingenious solution to try to capture this fact:

Axiom 2.5 (*Principle of mathematical induction*). Let $P(n)$ be any property pertaining to a natural number n. Suppose that $P(0)$ is true, and suppose that whenever $P(n)$ is true, $P(n{++})$ is also true. Then $P(n)$ is true for every natural number n.

Remark 2.1.10 We are a little vague on what "property" means at this point, but some possible examples of $P(n)$ might be "n is even"; "n is equal to 3"; "n solves the equation $(n + 1)^2 = n^2 + 2n + 1$"; and so forth. Of course we haven't defined many of these concepts yet, but when we do, Axiom 2.5 will apply to these properties. (A logical remark: Because this axiom refers not just to *variables*, but also *properties*, it is of a different nature than the other four axioms; indeed, Axiom 2.5 should technically be called an *axiom schema* rather than an *axi*—it is a template for producing an (infinite) number of axioms, rather than being a single axiom in its own right. To discuss this distinction further is far beyond the scope of this text, though, and falls in the realm of mathematical logic.)

The informal intuition behind this axiom is the following. Suppose $P(n)$ is such that $P(0)$ is true, and such that whenever $P(n)$ is true, then $P(n{++})$ is true. Then since $P(0)$ is true, $P(0{++}) = P(1)$ is true. Since $P(1)$ is true, $P(1{++}) = P(2)$

is true. Repeating this indefinitely, we see that $P(0)$, $P(1)$, $P(2)$, $P(3)$, etc., are all true—however this line of reasoning will never let us conclude that $P(0.5)$, for instance, is true. Thus Axiom 2.5 should not hold for number systems which contain "unnecessary" elements such as 0.5. (Indeed, one can give a "proof" of this fact as follows. Apply Axiom 2.5 to the property $P(n) = n$ "is not a half-integer", i.e., an integer plus 0.5. Then $P(0)$ is true, and if $P(n)$ is true, then $P(n{+}{+})$ is true. Thus Axiom 2.5 asserts that $P(n)$ is true for all natural numbers n, i.e., no natural number can be a half-integer. In particular, 0.5 cannot be a natural number. This "proof" is not quite genuine, because we have not defined such notions as "integer", "half-integer", and "0.5" yet, but it should give you some idea as to how the principle of induction is supposed to prohibit any numbers other than the "true" natural numbers from appearing in **N**.)

The principle of induction gives us a way to prove that a property $P(n)$ is true for every natural number n. Thus in the rest of this text we will see many proofs which have a form like this:

Proposition Template 2.1.11 *A certain property $P(n)$ is true for every natural number n.*

Proof Template We use induction. We first verify the base case $n = 0$, i.e., we prove $P(0)$. (Insert proof of $P(0)$ here.) Now suppose inductively that n is a natural number, and $P(n)$ has already been proven. We now prove $P(n{+}{+})$. (Insert proof of $P(n{+}{+})$, assuming that $P(n)$ is true, here.) This closes the induction, and thus $P(n)$ is true for all numbers n. □

Of course we will not necessarily use the exact template, wording, or order in the above type of proof, but the proofs using induction will generally be something like the above form. There are also some other variants of induction which we shall encounter later, such as backwards induction (Exercise 2.2.6), strong induction (Proposition 2.2.14), and transfinite induction (Lemma 8.5.15).

Axioms 2.1–2.5 are known as the *Peano axioms* for the natural numbers. They are all very plausible, and so we shall make

Assumption 2.6 (Informal) There exists a number system **N**, whose elements we will call *natural numbers*, for which Axioms 2.1–2.5 are true.

We will make this assumption a bit more precise once we have laid down our notation for sets and functions in the next chapter.

Remark 2.1.12 We will refer to this number system **N** as *the* natural number system. One could of course consider the possibility that there is more than one natural number system, e.g., we could have the Hindu-Arabic number system $\{0, 1, 2, 3, \dots\}$ and the Roman number system $\{O, I, II, III, IV, V, VI, \dots\}$ (augmented by adding a zero symbol O), and if we really wanted to be annoying we could view these number systems as different. But these number systems are clearly equivalent (the technical term is *isomorphic*), because one can create a one-to-one correspondence $0 \leftrightarrow O$, $1 \leftrightarrow I$, $2 \leftrightarrow II$, etc., which maps the zero of the Hindu-Arabic system with the zero of the Roman system, and which is preserved by the increment operation (e.g., if 2 corresponds to II, then 2++ will correspond to II++). For a more precise statement of this type of equivalence, see Exercise 3.5.13. Since all versions of the natural number system are equivalent, there is no point in having distinct natural number systems, and we will just use a single natural number system to do mathematics.

We will not prove Assumption 2.6 (though we will eventually include it in our axioms for set theory, see Axiom 3.8), and it will be the only assumption we will ever make about our numbers. A remarkable accomplishment of modern analysis is that just by starting from these five very primitive axioms, and some additional axioms from set theory, we can build all the other number systems, create functions, and do all the algebra and calculus that we are used to.

Remark 2.1.13 (Informal) One interesting feature about the natural numbers is that while each individual natural number is finite, the *set* of natural numbers is infinite; i.e., **N** is infinite but consists of individually finite elements. (The whole is greater than any of its parts.) There are no infinite natural numbers; one can even prove this using Axiom 2.5, provided one is comfortable with the notions of finite and infinite. (Clearly 0 is finite. Also, if n is finite, then clearly n++ is also finite. Hence by Axiom 2.5, all natural numbers are finite.) So the natural numbers can *approach* infinity, but never actually reach it; infinity is not one of the natural numbers. (There are other number systems which admit "infinite" numbers, such as the cardinals, ordinals, and p-adics, but they do not obey the principle of induction, and in any event are beyond the scope of this text.)

Remark 2.1.14 Note that our definition of the natural numbers is *axiomatic* rather than *constructive*. We have not told you what the natural numbers *are* (so we do not address such questions as what the numbers are made of, are they physical objects, what do they measure, etc.)—we have only listed some things you can do with them (in fact, the only operation we have defined on them right now is the increment one) and some of the properties that they have. This is how mathematics works—it treats its objects *abstractly*, caring only about what properties the objects have, not what the objects are or what they mean. If one wants to do mathematics, it does not matter whether a natural number means a certain arrangement of beads on an abacus, or a certain organization of bits in a computer's memory, or some more abstract concept with no physical substance; as long as you can increment them, see if two of them are equal, and later on do other arithmetic operations such as add and multiply, they qualify as numbers for mathematical purposes (provided they obey the requisite axioms, of course). It is possible to construct the natural numbers from other mathematical objects—from sets, for instance—but there are multiple ways to construct a working model of the natural numbers, and it is pointless, at least from a mathematician's standpoint, as to argue about which model is the "true" one; as long as it obeys all the axioms and does all the right things, that's good enough to do maths.

Remark 2.1.15 Historically, the realization that numbers could be treated axiomatically is very recent, not much more than a hundred years old. Before then, numbers were generally understood to be inextricably connected to some external concept, such as counting the cardinality of a set, measuring the length of a line segment, or the mass of a physical object. This worked reasonably well, until one was forced to move from one number system to another; for instance, understanding numbers in terms of counting beads, for instance, is great for conceptualizing the numbers 3 and 5, but doesn't work so well for -3 or $1/3$ or $\sqrt{2}$ or $3 + 4i$; thus each great advance in the theory of numbers—negative numbers, irrational numbers, complex numbers, even the number zero—led to a lot of unnecessary philosophical anguish. The great discovery of the late nineteenth century was that numbers can be understood abstractly via axioms, without necessarily needing a concrete model; of course a mathematician can use any of these models when it is convenient, to aid his or her intuition and understanding, but they can also be just as easily discarded when they begin to get in the way.

One consequence of the axioms is that we can now define sequences *recursively*. Suppose we want to build a sequence $a_0, a_1, a_2, \ldots$ of numbers by first defining a_0 to be some base value, e.g., $a_0 := c$ for some number c, and then by letting a_1 be some function of a_0, $a_1 := f_0(a_0)$, a_2 be some function of a_1, $a_2 := f_1(a_1)$, and so forth. In general, we set $a_{n++} := f_n(a_n)$ for some function f_n from $\mathbf{N}$ to $\mathbf{N}$. By using all the axioms together we will now conclude that this procedure will give a single value to the sequence element a_n for each natural number n. More precisely[5]:

Proposition 2.1.16 (Recursive definitions). *Suppose for each natural number n, we have some function $f_n : \mathbf{N} \to \mathbf{N}$ from the natural numbers to the natural numbers. Let c be a natural number. Then we can assign a unique natural number a_n to each natural number n, such that $a_0 = c$ and $a_{n++} = f_n(a_n)$ for each natural number n.*

Proof (Informal) We use induction. We first observe that this procedure gives a single value to a_0, namely c. (None of the other definitions $a_{n++} := f_n(a_n)$ will redefine the value of a_0, because of Axiom 2.3.) Now suppose inductively that the procedure gives a single value to a_n. Then it gives a single value to a_{n++}, namely $a_{n++} := f_n(a_n)$. (None of the other definitions $a_{m++} := f_m(a_m)$ will redefine the value of a_{n++}, because of Axiom 2.4.) This completes the induction, and so a_n is defined for each natural number n, with a single value assigned to each a_n. □

Note how all of the axioms had to be used here. In a system which had some sort of wrap-around, recursive definitions would not work because some elements of the sequence would constantly be redefined. For instance, in Example 2.1.5, in which $3{++} = 0$, then there would be (at least) two conflicting definitions for a_0, either c or $f_3(a_3)$. In a system which had superfluous elements such as 0.5, the element $a_{0.5}$ would never be defined.

[5] Strictly speaking, this proposition requires one to define the notion of a *function*, which we shall do in the next chapter. However, this will not be circular, as the concept of a function does not require the Peano axioms. Proposition 2.1.16 can be formalized more rigorously in the language of set theory; see Exercise 3.5.12.

Recursive definitions are very powerful; for instance, we can use them to define addition and multiplication, to which we now turn.

2.2 Addition

The natural number system is very bare right now: we have only one operation—incrementation—and a handful of axioms. But now we can build up more complex operations, such as addition.

The way it works is the following. To add three to five should be the same as incrementing five three times—this is one increment more than adding two to five, which is one increment more than adding one to five, which is one increment more than adding zero to five, which should just give five. So we give a recursive definition for addition as follows.

Definition 2.2.1 (*Addition of natural numbers*). Let m be a natural number. To add zero to m, we define $0 + m := m$. Now suppose inductively that we have defined how to add n to m. Then we can add $n{+}{+}$ to m by defining $(n{+}{+}) + m := (n + m){+}{+}$.

Thus $0 + m$ is m, $1 + m = (0{+}{+}) + m$ is $m{+}{+}$; $2 + m = (1{+}{+}) + m = (m{+}{+}){+}{+}$; and so forth; for instance we have $2 + 3 = (3{+}{+}){+}{+} = 4{+}{+} = 5$. From our discussion of recursion in the previous section we see that we have defined $n + m$ for every natural number n. Here we are specializing the previous general discussion to the setting where $a_n = n + m$ and $f_n(a_n) = a_n{+}{+}$. Note that this definition is asymmetric: $3 + 5$ is incrementing 5 three times, while $5 + 3$ is incrementing 3 five times. Of course, they both yield the same value of 8. More generally, it is a fact (which we shall prove shortly) that $a + b = b + a$ for all natural numbers a, b, although this is not immediately clear from the definition.

Notice that we can prove easily, using Axioms 2.1, 2.2, and induction (Axiom 2.5), that the sum of two natural numbers is again a natural number (why?).

Right now we only have two facts about addition: that $0 + m = m$, and that $(n{+}{+}) + m = (n + m){+}{+}$. Remarkably, this turns out to be enough to deduce everything else we know about addition. We begin with some basic lemmas.[6]

Lemma 2.2.2 *For any natural number n, $n + 0 = n$.*

Note that we cannot deduce this immediately from $0 + m = m$ because we have not yet established the commutative property $a + b = b + a$ of addition.

[6] From a logical point of view, there is no difference between a lemma, proposition, theorem, or corollary—they are all claims waiting to be proved. However, we use these terms to suggest different levels of importance and difficulty. A lemma is an easily proved claim which is helpful for proving other propositions and theorems, but is usually not particularly interesting in its own right. A proposition is a statement which is interesting in its own right, while a theorem is a more important statement than a proposition which says something definitive on the subject, and often takes more effort to prove than a proposition or lemma. A corollary is a quick consequence of a proposition or theorem that was proven recently.

Proof We use induction. The base case $0 + 0 = 0$ follows since we know that $0 + m = m$ for every natural number m, and 0 is a natural number. Now suppose inductively that $n + 0 = n$. We wish to show that $(n{+}{+}) + 0 = n{+}{+}$. But by definition of addition, $(n{+}{+}) + 0$ is equal to $(n + 0){+}{+}$, which is equal to $n{+}{+}$ since $n + 0 = n$. This closes the induction. □

Lemma 2.2.3 *For any natural numbers n and m,* $n + (m{+}{+}) = (n + m){+}{+}$.

Again, we cannot deduce this yet from $(n{+}{+}) + m = (n + m){+}{+}$ because we do not know yet that $a + b = b + a$.

Proof We induct on n (keeping m fixed). We first consider the base case $n = 0$. In this case we have to prove $0 + (m{+}{+}) = (0 + m){+}{+}$. But by definition of addition, $0 + (m{+}{+}) = m{+}{+}$ and $0 + m = m$, so both sides are equal to $m{+}{+}$ and are thus equal to each other. Now we assume inductively that $n + (m{+}{+}) = (n + m){+}{+}$; we now have to show that $(n{+}{+}) + (m{+}{+}) = ((n{+}{+}) + m){+}{+}$. The left-hand side is $(n + (m{+}{+})){+}{+}$ by definition of addition, which is equal to $((n + m){+}{+}){+}{+}$ by the inductive hypothesis. Similarly, we have $(n{+}{+}) + m = (n + m){+}{+}$ by the definition of addition, and so the right-hand side is also equal to $((n + m){+}{+}){+}{+}$. Thus both sides are equal to each other, and we have closed the induction. □

As a particular corollary of Lemma 2.2.2 and Lemma 2.2.3 we see that $n{+}{+} = n + 1$ (why?).

As promised earlier, we can now prove that $a + b = b + a$.

Proposition 2.2.4 (Addition is commutative). *For any natural numbers n and m,* $n + m = m + n$.

Proof We shall use induction on n (keeping m fixed). First we do the base case $n = 0$, i.e., we show $0 + m = m + 0$. By the definition of addition, $0 + m = m$, while by Lemma 2.2.2, $m + 0 = m$. Thus the base case is done. Now suppose inductively that $n + m = m + n$, now we have to prove that $(n{+}{+}) + m = m + (n{+}{+})$ to close the induction. By the definition of addition, $(n{+}{+}) + m = (n + m){+}{+}$. By Lemma 2.2.3, $m + (n{+}{+}) = (m + n){+}{+}$, but this is equal to $(n + m){+}{+}$ by the inductive hypothesis $n + m = m + n$. Thus $(n{+}{+}) + m = m + (n{+}{+})$ and we have closed the induction. □

Proposition 2.2.5 (Addition is associative). *For any natural numbers* a, b, c, *we have* $(a + b) + c = a + (b + c)$.

Proof See Exercise 2.2.1. □

Because of this associativity we can write sums such as $a + b + c$ without having to worry about which order the numbers are being added together.

Now we develop a cancellation law.

Proposition 2.2.6 (Cancellation law). *Let* a, b, c *be natural numbers such that* $a + b = a + c$. *Then we have* $b = c$.

Note that we cannot use subtraction or negative numbers yet to prove this proposition, because we have not developed these concepts yet. In fact, this cancellation law is crucial in letting us define subtraction (and the integers) later on in this text, because it allows for a sort of "virtual subtraction" even before subtraction is officially defined.

Proof We prove this by induction on a. First consider the base case $a = 0$. Then we have $0 + b = 0 + c$, which by definition of addition implies that $b = c$ as desired. Now suppose inductively that we have the cancellation law for a (so that $a + b = a + c$ implies $b = c$); we now have to prove the cancellation law for $a{+}{+}$. In other words, we assume that $(a{+}{+}) + b = (a{+}{+}) + c$ and need to show that $b = c$. By the definition of addition, $(a{+}{+}) + b = (a + b){+}{+}$ and $(a{+}{+}) + c = (a + c){+}{+}$ and so we have $(a + b){+}{+} = (a + c){+}{+}$. By Axiom 2.4, we have $a + b = a + c$. Since we already have the cancellation law for a, we thus have $b = c$ as desired. This closes the induction. □

We now discuss how addition interacts with positivity.

Definition 2.2.7 (*Positive natural numbers*). A natural number n is said to be *positive* iff it is not equal to 0. ("iff" is shorthand for "if and only if"; see Sect. A.1.)

Proposition 2.2.8 *If a is a positive natural number, and b is a natural number, then $a + b$ is positive (and hence $b + a$ is also, by Proposition 2.2.4).*

Proof We use induction on b. If $b = 0$, then $a + b = a + 0 = a$, which is positive, so this proves the base case. Now suppose inductively that $a + b$ is positive. Then $a + (b{+}{+}) = (a + b){+}{+}$, which cannot be zero by Axiom 2.3, and is hence positive. This closes the induction. □

Corollary 2.2.9 *If a and b are natural numbers such that $a + b = 0$, then $a = 0$ and $b = 0$.*

Proof Suppose for sake of contradiction that $a \neq 0$ or $b \neq 0$. If $a \neq 0$ then a is positive, and hence $a + b = 0$ is positive by Proposition 2.2.8, a contradiction. Similarly if $b \neq 0$ then b is positive, and again $a + b = 0$ is positive by Proposition 2.2.8, a contradiction. Thus a and b must both be zero. □

Lemma 2.2.10 *Let a be a positive number. Then there exists exactly one natural number b such that $b{+}{+} = a$.*

Proof See Exercise 2.2.2. □

Once we have a notion of addition, we can begin defining a notion of *order*.

Definition 2.2.11 (*Ordering of the natural numbers*) Let n and m be natural numbers. We say that n is *greater than or equal to m*, and write $n \geq m$ or $m \leq n$, iff we have $n = m + a$ for some natural number a. We say that n is *strictly greater than m*, and write $n > m$ or $m < n$, iff $n \geq m$ and $n \neq m$.

Thus for instance $8 > 5$, because $8 = 5 + 3$ and $8 \neq 5$. Also note that $n{+}{+} > n$ for any n; thus there is no largest natural number n, because the next number $n{+}{+}$ is always larger still.

Proposition 2.2.12 (Basic properties of order for natural numbers). *Let a, b, c be natural numbers. Then*

(a) (*Order is reflexive*) $a \geq a$.
(b) (*Order is transitive*) *If $a \geq b$ and $b \geq c$, then $a \geq c$.*
(c) (*Order is antisymmetric*) *If $a \geq b$ and $b \geq a$, then $a = b$.*
(d) (*Addition preserves order*) *$a \geq b$ if and only if $a + c \geq b + c$.*
(e) *$a < b$ if and only if $a{+}{+} \leq b$.*
(f) *$a < b$ if and only if $b = a + d$ for some* positive *number d.*

Proof See Exercise 2.2.3. □

Proposition 2.2.13 (Trichotomy of order for natural numbers). *Let a and b be natural numbers. Then exactly one of the following statements is true: $a < b$, $a = b$, or $a > b$.*

Proof This is only a sketch of the proof; the gaps will be filled in Exercise 2.2.4.

First we show that we cannot have more than one of the statements $a < b$, $a = b$, $a > b$ holding at the same time. If $a < b$ then $a \neq b$ by definition, and if $a > b$ then $a \neq b$ by definition. If $a > b$ and $a < b$ then by Proposition 2.2.12 we have $a = b$, a contradiction. Thus no more than one of the statements is true.

Now we show that at least one of the statements is true. We keep b fixed and induct on a. When $a = 0$ we have $0 \leq b$ for all b (why?), so we have either $0 = b$ or $0 < b$, which proves the base case. Now suppose we have proven the proposition for a, and now we prove the proposition for $a{+}{+}$. From the trichotomy for a, there are three cases: $a < b$, $a = b$, and $a > b$. If $a > b$, then $a{+}{+} > b$ (why?). If $a = b$, then $a{+}{+} > b$ (why?). Now suppose that $a < b$. Then by Proposition 2.2.12, we have $a{+}{+} \leq b$. Thus either $a{+}{+} = b$ or $a{+}{+} < b$, and in either case we are done. This closes the induction. □

The properties of order allow one to obtain a stronger version of the principle of induction:

Proposition 2.2.14 (Strong principle of induction). *Let m_0 be a natural number, and let $P(m)$ be a property pertaining to an arbitrary natural number m. Suppose that for each $m \geq m_0$, we have the following implication: if $P(m')$ is true for all natural numbers $m_0 \leq m' < m$, then $P(m)$ is also true. (In particular, this means that $P(m_0)$ is true, since in this case the hypothesis is vacuous.) Then we can conclude that $P(m)$ is true for all natural numbers $m \geq m_0$.*

Remark 2.2.15 In applications we usually use this principle with $m_0 = 0$ or $m_0 = 1$.

Proof See Exercise 2.2.5. □

— Exercises —

Exercise 2.2.1 Prove Proposition 2.2.5. (*Hint:* fix two of the variables and induct on the third.)

Exercise 2.2.2 Prove Lemma 2.2.10. (*Hint:* use induction. The induction here is somewhat degenerate, in that the induction hypothesis is not actually used, but this does not prevent the argument from remaining valid; cf. the discussion on implication and causality in Appendix A.2.)

Exercise 2.2.3 Prove Proposition 2.2.12. (*Hint:* you will need many of the preceding propositions, corollaries, and lemmas.)

Exercise 2.2.4 Justify the three statements marked (why?) in the proof of Proposition 2.2.13.

Exercise 2.2.5 Prove Proposition 2.2.14. (*Hint:* define $Q(n)$ to be the property that $P(m)$ is true for all $m_0 \leq m < n$; note that $Q(n)$ is vacuously true when $n \leq m_0$.)

Exercise 2.2.6 Let n be a natural number, and let $P(m)$ be a property pertaining to the natural numbers such that whenever $P(m{+}{+})$ is true, then $P(m)$ is true. Suppose that $P(n)$ is also true. Prove that $P(m)$ is true for all natural numbers $m \leq n$; this is known as the *principle of backwards induction*. (*Hint:* apply induction to the variable n.)

Exercise 2.2.7 Let n be a natural number, and let $P(m)$ be a property pertaining to the natural numbers such that whenever $P(m)$ is true, $P(m{+}{+})$ is true. Show that if $P(n)$ is true, then $P(m)$ is true for all $m \geq n$. (This principle is sometimes referred to as the *principle of induction starting from the base case* n.)

2.3 Multiplication

In the previous section we have proven all the basic facts that we know to be true about addition and order. To save space and to avoid belaboring the obvious, we will now allow ourselves to use all the rules of algebra concerning addition and order that we are familiar with, without further comment. Thus for instance we may write things like $a + b + c = c + b + a$ without supplying any further justification. Now we introduce multiplication. Just as addition is the iterated increment operation, multiplication is iterated addition:

Definition 2.3.1 (*Multiplication of natural numbers*). Let m be a natural number. To multiply zero to m, we define $0 \times m := 0$. Now suppose inductively that we have defined how to multiply n to m. Then we can multiply $n{+}{+}$ to m by defining $(n{+}{+}) \times m := (n \times m) + m$.

Thus for instance $0 \times m = 0$, $1 \times m = 0 + m$, $2 \times m = 0 + m + m$, etc. By induction one can easily verify that the product of two natural numbers is a natural number.

Lemma 2.3.2 (Multiplication is commutative). *Let n, m be natural numbers. Then $n \times m = m \times n$.*

Proof See Exercise 2.3.1. □

We will now abbreviate $n \times m$ as nm and use the usual convention that multiplication takes precedence over addition, thus for instance $ab + c$ means $(a \times b) + c$, not $a \times (b + c)$. (We will also use the usual notational conventions of precedence for the other arithmetic operations when they are defined later, to save on using parentheses all the time.)

Lemma 2.3.3 (Positive natural numbers have no zero divisors). *Let n, m be natural numbers. Then $n \times m = 0$ if and only if at least one of n, m is equal to zero. In particular, if n and m are both positive, then nm is also positive.*

Proof See Exercise 2.3.2. □

Proposition 2.3.4 (Distributive law). *For any natural numbers a, b, c, we have $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$.*

Proof Since multiplication is commutative we only need to show the first identity $a(b + c) = ab + ac$. We keep a and b fixed, and use induction on c. Let's prove the base case $c = 0$, i.e., $a(b + 0) = ab + a0$. The left-hand side is ab, while the right-hand side is $ab + 0 = ab$, so we are done with the base case. Now let us suppose inductively that $a(b + c) = ab + ac$, and let us prove that $a(b + (c{+}{+})) = ab + a(c{+}{+})$. The left-hand side is $a((b + c){+}{+}) = a(b + c) + a$, while the right-hand side is $ab + ac + a = a(b + c) + a$ by the induction hypothesis, and so we can close the induction. □

Proposition 2.3.5 (Multiplication is associative). *For any natural numbers a, b, c, we have $(a \times b) \times c = a \times (b \times c)$.*

Proof See Exercise 2.3.3. □

Proposition 2.3.6 (Multiplication preserves order). *If a, b are natural numbers such that $a < b$, and c is positive, then $ac < bc$.*

Proof Since $a < b$, we have $b = a + d$ for some positive d. Multiplying by c and using the distributive law we obtain $bc = ac + dc$. Since d is positive, and c is positive, dc is positive, and hence $ac < bc$ as desired. □

Corollary 2.3.7 (Cancellation law). *Let a, b, c be natural numbers such that $ac = bc$ and c is non-zero. Then $a = b$.*

Remark 2.3.8 Just as Proposition 2.2.6 will allow for a "virtual subtraction" which will eventually let us define genuine subtraction, this corollary provides a "virtual division" which will be needed to define genuine division later on.

Proof By the trichotomy of order (Proposition 2.2.13), we have three cases: $a < b$, $a = b$, $a > b$. Suppose first that $a < b$, then by Proposition 2.3.6 we have $ac < bc$, a contradiction. We can obtain a similar contradiction when $a > b$. Thus the only possibility is that $a = b$, as desired. □

With these propositions it is easy to deduce all the familiar rules of algebra involving addition and multiplication, see for instance Exercise 2.3.4.

Now that we have the familiar operations of addition and multiplication, the more primitive notion of increment will begin to fall by the wayside, and we will see it rarely from now on. In any event we can always use addition to describe incrementation, since $n{+}{+} = n + 1$.

Proposition 2.3.9 (Euclid's division lemma). *Let n be a natural number, and let q be a positive number. Then there exist natural numbers m, r such that $0 \leq r < q$ and $n = mq + r$.*

Remark 2.3.10 In other words, we can divide a natural number n by a positive number q to obtain a quotient m (which is another natural number) and a remainder r (which is less than q). This algorithm marks the beginning of *number theory*, which is a beautiful and important subject but one which is beyond the scope of this text.

Proof See Exercise 2.3.5. □

Just like one uses the increment operation to recursively define addition, and addition to recursively define multiplication, one can use multiplication to recursively define *exponentiation*:

Definition 2.3.11 (*Exponentiation for natural numbers*). Let m be a natural number. To raise m to the power 0, we define $m^0 := 1$; in particular, we define $0^0 := 1$. Now suppose recursively that m^n has been defined for some natural number n, then we define $m^{n{+}{+}} := m^n \times m$.

Examples 2.3.12 Thus for instance $x^1 = x^0 \times x = 1 \times x = x$; $x^2 = x^1 \times x = x \times x$; $x^3 = x^2 \times x = x \times x \times x$; and so forth. By induction we see that this recursive definition defines x^n for all natural numbers n.

We will not develop the theory of exponentiation too deeply here, but instead wait until after we have defined the integers and rational numbers; see in particular Proposition 4.3.10.

— Exercises —

Exercise 2.3.1 Prove Lemma 2.3.2. (*Hint:* modify the proofs of Lemmas 2.2.2, 2.2.3 and Proposition 2.2.4.)

Exercise 2.3.2 Prove Lemma 2.3.3. (*Hint:* prove the second statement first.)

Exercise 2.3.3 Prove Proposition 2.3.5. (*Hint:* modify the proof of Proposition 2.2.5 and use the distributive law.)

Exercise 2.3.4 Prove the identity $(a + b)^2 = a^2 + 2ab + b^2$ for all natural numbers a, b.

Exercise 2.3.5 Prove Proposition 2.3.9. (*Hint:* fix q and induct on n.)

Chapter 3
Set Theory

Modern analysis, like most other subfields of modern mathematics, is concerned with numbers, sets, and geometry. We have already introduced one type of number system, the natural numbers. We will introduce the other number systems shortly, but for now we pause to introduce the concepts and notation of set theory, as they will be used increasingly heavily in later chapters. (We will not pursue a rigorous description of Euclidean geometry in this text, preferring instead to describe that geometry in terms of the real number system by means of the Cartesian co-ordinate system.)

While set theory is not the main focus of this text, almost every other branch of mathematics relies on set theory as part of its foundation, so it is important to get at least some grounding in set theory before doing other advanced areas of mathematics. In this chapter we present the more elementary aspects of axiomatic set theory, leaving more advanced topics such as a discussion of infinite sets and the axiom of choice to Chap. 8. A full treatment of the finer subtleties of set theory (of which there are many!) is unfortunately well beyond the scope of this text.

3.1 Fundamentals

In this section we shall set out some axioms for sets, just as we did for the natural numbers. For pedagogical reasons, we will use a somewhat overcomplete list of axioms for set theory, in the sense that some of the axioms can be used to deduce others, but there is no real harm in doing this. We begin with an informal description of what sets should be.

Definition 3.1.1 (*Informal*) We define a *set* A to be any unordered collection of objects, e.g., $\{3, 8, 5, 2\}$ is a set. If x is an object, we say that *x is an element of*

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_3

A or $x \in A$ if x lies in the collection; otherwise we say that $x \notin A$. For instance, $3 \in \{1, 2, 3, 4, 5\}$ but $7 \notin \{1, 2, 3, 4, 5\}$.

This definition is intuitive enough, but it doesn't answer a number of questions, such as which collections of objects are considered to be sets, which sets are equal to other sets, and how one defines operations on sets (e.g., unions, intersections, etc.). Also, we have no axioms yet on what sets do, or what their elements do. Obtaining these axioms and defining these operations will be the purpose of the remainder of this section.

We first clarify one point: we consider sets themselves to be a type of object.

Axiom 3.1 (*Sets are objects*). If A is a set, then A is also an object. In particular, given two sets A and B, it is meaningful to ask whether A is also an element of B.

Example 3.1.2 (Informal) The set $\{3, \{3, 4\}, 4\}$ is a set of three distinct elements, one of which happens to itself be a set of two elements. See Example 3.1.9 for a more formal version of this example.

Remark 3.1.3 There is a special case of set theory, called "pure set theory", in which *all* objects are sets; for instance the number 0 might be identified with the empty set $\emptyset = \{\}$, the number 1 might be identified with $\{0\} = \{\{\}\}$, the number 2 might be identified with $\{0, 1\} = \{\{\}, \{\{\}\}\}$, and so forth. From a logical point of view, pure set theory is a simpler theory, since one only has to deal with sets and not with objects; however, from a conceptual point of view it is often easier to deal with impure set theories in which some objects are not considered to be sets. The two types of theories are more or less equivalent for the purposes of doing mathematics, and so we shall take an agnostic position as to whether all objects are sets or not. For instance, we do not insist that a natural number such as 3 be identified with a set as indicated above. (The more accurate and mathematically useful statement is that natural numbers can be the *cardinalities* of sets, rather than necessarily being sets themselves. See Sect. 3.6.)

To summarize so far, among all the objects studied in mathematics, some of the objects happen to be sets; and if x is an object and A is a set, then either $x \in A$ is true or $x \in A$ is false. (If A is not a set, we leave the statement $x \in A$ undefined; for instance, we consider the statement $3 \in 4$ to neither be true or false, but simply meaningless, since 4 is not a set.)

Next, we try to capture the notion of equality: when are two sets considered to be equal? We do not consider the order of the elements inside a set to be important; thus we think of $\{3, 8, 5, 2\}$ and $\{2, 3, 5, 8\}$ as the same set. On the other hand, $\{3, 8, 5, 2\}$ and $\{3, 8, 5, 2, 1\}$ are different sets, because the latter set contains an element that the former one does not, namely the element 1. For similar reasons $\{3, 8, 5, 2\}$ and $\{3, 8, 5\}$ are different sets. We formalize this by a further axiom:

Axiom 3.2 (*Equality of sets*). Two sets A and B are *equal*, $A = B$, iff every element of A is an element of B and vice versa. To put it another way, $A = B$ if and only if every element x of A belongs also to B, and every element y of B belongs also to A.

Example 3.1.4 Thus, for instance, $\{1, 2, 3, 4, 5\}$ and $\{3, 4, 2, 1, 5\}$ are the same set, since they contain exactly the same elements. (The set $\{3, 3, 1, 5, 2, 4, 2\}$ is also equal to $\{1, 2, 3, 4, 5\}$; the repetition of 3 and 2 is irrelevant as it does not further change the status of 2 and 3 being elements of the set.)

The "is an element of" relation $\in$ obeys the axiom of substitution (see Section A.7). Because of this, any new operation we define on sets will also obey the axiom of substitution, as long as we can define that operation purely in terms of the relation $\in$. This is for instance the case for the remaining definitions in this section. (On the other hand, we cannot use the notion of the "first" or "last" element in a set in a well-defined manner, because this would not respect the axiom of substitution; for instance the sets $\{1, 2, 3, 4, 5\}$ and $\{3, 4, 2, 1, 5\}$ are the same set, but have different first elements.)

Next, we turn to the issue of exactly which objects are sets and which objects are not. The situation is analogous to how we defined the natural numbers in the previous chapter; we started with a single natural number, 0, and started building more numbers out of 0 using the increment operation. We will try something similar here, starting with a single set, the *empty set*, and building more sets out of the empty set by various operations. We begin by postulating the existence of the empty set.

Axiom 3.3 (*Empty set*). There exists a set $\emptyset$, known as the *empty set*, which contains no elements, i.e., for every object x we have $x \notin \varnothing$.

The empty set is also denoted $\{\}$. Note that there can only be one empty set; if there were two sets $\emptyset$ and $\emptyset'$ which were both empty, then by Axiom 3.2 they would be equal to each other (why?).

If a set is not equal to the empty set, we call it *non-empty*. The following statement is very simple, but worth stating nevertheless:

Lemma 3.1.5 (Single choice). *Let A be a non-empty set. Then there exists an object x such that $x \in A$.*

Proof We prove by contradiction. Suppose there does not exist any object x such that $x \in A$. Then for all objects x, we have $x \notin A$. Also, by Axiom 3.3 we have $x \notin \emptyset$. Thus $x \in A \iff x \in \emptyset$ (both statements are equally false), and so $A = \emptyset$ by Axiom 3.2, a contradiction. □

Remark 3.1.6 The above Lemma asserts that given any non-empty set A, we are allowed to "choose" an element x of A which demonstrates this non-emptyness. Later on (in Lemma 3.5.11) we will show that given any finite number of non-empty sets, say $A_1, \ldots, A_n$, it is possible to choose one element $x_1, \ldots, x_n$ from each set $A_1, \ldots, A_n$; this is known as "finite choice". However, in order to choose elements from an infinite number of sets, we need an additional axiom, the *axiom of choice*, which we will discuss in Sect. 8.4.

Remark 3.1.7 Note that the empty set is *not* necessarily the same thing as the natural number 0. One is a set; the other is a number. However, it is true that the *cardinality* of the empty set is 0; see Sect. 3.6.

If Axiom 3.3 was the only axiom that set theory had, then set theory could be quite boring, as there might be just a single set in existence, the empty set. We now present further axioms to enrich the class of sets available.

Axiom 3.4 (*Singleton sets and pair sets*). If a is an object, then there exists a set $\{a\}$ whose only element is a, i.e., for every object y, we have $y \in \{a\}$ if and only if $y = a$; we refer to $\{a\}$ as the *singleton set* whose element is a. Furthermore, if a and b are objects, then there exists a set $\{a, b\}$ whose only elements are a and b; i.e., for every object y, we have $y \in \{a, b\}$ if and only if $y = a$ or $y = b$; we refer to this set as the *pair set* formed by a and b.

Remarks 3.1.8 Just as there is only one empty set, there is only one singleton set for each object a, thanks to Axiom 3.2 (why?). Similarly, given any two objects a and b, there is only one pair set formed by a and b. Also, Axiom 3.2 also ensures that $\{a, b\} = \{b, a\}$ (why?) and $\{a, a\} = \{a\}$ (why?). Thus the singleton set axiom is in fact redundant, being a consequence of the pair set axiom. Conversely, the pair set axiom will follow from the singleton set axiom and the pairwise union axiom below (see Lemma 3.1.12). One may wonder why we don't go further and create triplet axioms, quadruplet axioms, etc.; however there will be no need for this once we introduce the pairwise union axiom below.

Examples 3.1.9 Since $\emptyset$ is a set (and hence an object), the singleton set $\{\emptyset\}$, i.e., the set whose only element is $\emptyset$, is also a set. Similarly, the singleton set $\{\{\emptyset\}\}$ and the pair set $\{\emptyset, \{\emptyset\}\}$ are also sets. These four sets are not equal to each other (Exercise 3.1.2).

As the above examples show, we can now create quite a few sets; however, the sets we make are still fairly small (each set that we can build consists of no more than two elements, so far). The next axiom allows us to build somewhat larger sets than before.

Axiom 3.5 (*Pairwise union*). Given any two sets A, B, there exists a set $A \cup B$, called the *union* of A and B, which consists of all the elements which belong to A or B or both. In other words, for any object x,

$$x \in A \cup B \iff (x \in A \text{ or } x \in B).$$

Recall that "or" refers by default in mathematics to *inclusive* or: "X or Y is true" means that "either X is true, or Y is true, or both are true". See Sect. A.1.

Example 3.1.10 The set $\{1, 2\} \cup \{2, 3\}$ consists of those elements which either lie on $\{1, 2\}$ or in $\{2, 3\}$ or in both, or in other words the elements of this set are simply 1, 2, and 3. Because of this, we denote this set as $\{1, 2\} \cup \{2, 3\} = \{1, 2, 3\}$.

Remark 3.1.11 If A, B, A' are sets, and A is equal to A', then $A \cup B$ is equal to $A' \cup B$ (why? One needs to use Axiom 3.5 and Axiom 3.2). Similarly if B' is a set which is equal to B, then $A \cup B$ is equal to $A \cup B'$. Thus the operation of union obeys the axiom of substitution and is thus well-defined on sets.

We now give some basic properties of unions.

Lemma 3.1.12 *If a and b are objects, then $\{a, b\} = \{a\} \cup \{b\}$. If A, B, C are sets, then the union operation is commutative* (*i.e., $A \cup B = B \cup A$*) *and associative* (*i.e., $(A \cup B) \cup C = A \cup (B \cup C)$*)*. Also, we have $A \cup A = A \cup \emptyset = \emptyset \cup A = A$.*

Proof We prove just the associativity identity $(A \cup B) \cup C = A \cup (B \cup C)$ and leave the remaining claims to Exercise 3.1.3. By Axiom 3.2, we need to show that every element x of $(A \cup B) \cup C$ is an element of $A \cup (B \cup C)$, and vice versa. So suppose first that x is an element of $(A \cup B) \cup C$. By Axiom 3.5, this means that at least one of $x \in A \cup B$ or $x \in C$ is true. We now divide into two cases. If $x \in C$, then by Axiom 3.5 again $x \in B \cup C$, and so by Axiom 3.5 again we have $x \in A \cup (B \cup C)$. Now suppose instead $x \in A \cup B$, then by Axiom 3.5 again $x \in A$ or $x \in B$. If $x \in A$ then $x \in A \cup (B \cup C)$ by Axiom 3.5, while if $x \in B$ then by consecutive applications of Axiom 3.5 we have $x \in B \cup C$ and hence $x \in A \cup (B \cup C)$. Thus in all cases we see that every element of $(A \cup B) \cup C$ lies in $A \cup (B \cup C)$. A similar argument shows that every element of $A \cup (B \cup C)$ lies in $(A \cup B) \cup C$, and so $(A \cup B) \cup C = A \cup (B \cup C)$ as desired. □

Because of the above lemma, we do not need to use parentheses to denote multiple unions, thus for instance we can write $A \cup B \cup C$ instead of $(A \cup B) \cup C$ or $A \cup (B \cup C)$. Similarly for unions of four sets, $A \cup B \cup C \cup D$, etc.

Remark 3.1.13 While the operation of union has some similarities with addition, the two operations are *not* identical. For instance, $\{2\} \cup \{3\} = \{2, 3\}$ and $2 + 3 = 5$, whereas $\{2\} + \{3\}$ is meaningless (addition pertains to numbers, not sets) and $2 \cup 3$ is also meaningless (union pertains to sets, not numbers).

This axiom allows us to define triplet sets, quadruplet sets, and so forth: if a, b, c are three objects, we define $\{a, b, c\} := \{a\} \cup \{b\} \cup \{c\}$; if a, b, c, d are four objects, then we define $\{a, b, c, d\} := \{a\} \cup \{b\} \cup \{c\} \cup \{d\}$, and so forth. On the other hand, we are not yet in a position to define sets consisting of n objects for any given natural number n; this would require iterating the above construction "n times", but the concept of n-fold iteration has not yet been rigorously defined. For similar reasons, we cannot yet define sets consisting of infinitely many objects, because that would require iterating the axiom of pairwise union infinitely often, and it is not clear at this stage that one can do this rigorously. Later on, we will introduce other axioms of set theory which allow one to construct arbitrarily large, and even infinite, sets.

Clearly, some sets seem to be larger than others. One way to formalize this concept is through the notion of a *subset*.

Definition 3.1.14 (*Subsets*). Let A, B be sets. We say that A is a *subset* of B, denoted $A \subseteq B$, iff every element of A is also an element of B, i.e.,

$$\text{For any object } x, \quad x \in A \quad \Longrightarrow \quad x \in B.$$

We say that A is a *proper subset* of B, denoted $A \subsetneq B$, if $A \subseteq B$ and $A \neq B$.

Remark 3.1.15 Because these definitions involve only the notions of equality and the "is an element of" relation, both of which already obey the axiom of substitution, the notion of subset also automatically obeys the axiom of substitution. Thus for instance if $A \subseteq B$ and $A = A'$, then $A' \subseteq B$.

Examples 3.1.16 We have $\{1, 2, 4\} \subseteq \{1, 2, 3, 4, 5\}$, because every element of $\{1, 2, 4\}$ is also an element of $\{1, 2, 3, 4, 5\}$. In fact we also have $\{1, 2, 4\} \subsetneq \{1, 2, 3, 4, 5\}$, since the two sets $\{1, 2, 4\}$ and $\{1, 2, 3, 4, 5\}$ are not equal. Given any set A, we always have $A \subseteq A$ (why?) and $\emptyset \subseteq A$ (why?).

The notion of subset in set theory is similar to the notion of "less than or equal to" for numbers, as the following proposition demonstrates (for a more precise statement, see Definition 8.5.1):

Proposition 3.1.17 (Sets are partially ordered by set inclusion). *Let A, B, C be sets. If $A \subseteq B$ and $B \subseteq C$ then $A \subseteq C$. If $A \subseteq B$ and $B \subseteq A$, then $A = B$. Finally, if $A \subsetneq B$ and $B \subsetneq C$ then $A \subsetneq C$.*

Proof We shall just prove the first claim. Suppose that $A \subseteq B$ and $B \subseteq C$. To prove that $A \subseteq C$, we have to prove that every element of A is an element of C. So, let us pick an arbitrary element x of A. Then, since $A \subseteq B$, x must then be an element of B. But then since $B \subseteq C$, x is an element of C. Thus every element of A is indeed an element of C, as claimed. □

Remark 3.1.18 The subset relation and the union operation are related to each other: see for instance Exercise 3.1.7.

Remark 3.1.19 There is one important difference between the subset relation $\subsetneq$ and the less than relation $<$. Given any two distinct natural numbers n, m, we know that one of them is smaller than the other (Proposition 2.2.13); however, given two distinct sets, it is not in general true that one of them is a subset of the other. For instance, take $A := \{2n : n \in \mathbf{N}\}$ to be the set of even natural numbers, and $B := \{2n + 1 : n \in \mathbf{N}\}$ to be the set of odd natural numbers. Then neither set is a subset of the other. This is why we say that sets are only *partially ordered*, whereas the natural numbers are *totally ordered* (see Definitions 8.5.1, 8.5.3).

Remark 3.1.20 We should also caution that the subset relation $\subseteq$ is *not* the same as the element relation $\in$. The number 2 is an element of $\{1, 2, 3\}$ but not a subset; thus $2 \in \{1, 2, 3\}$, but $2 \not\subseteq \{1, 2, 3\}$. Indeed, 2 is not even a set. Conversely, while $\{2\}$ is a subset of $\{1, 2, 3\}$, it is not an element; thus $\{2\} \subseteq \{1, 2, 3\}$ but $\{2\} \notin \{1, 2, 3\}$. The point is that the number 2 and the set $\{2\}$ are distinct objects. It is important to distinguish sets from their elements, as they can have different properties. For instance, it is possible to have an infinite set consisting of finite numbers (the set $\mathbf{N}$ of natural numbers is one such example), and it is also possible to have a finite set consisting of infinite objects (consider for instance the finite set $\{\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}\}$, which has four elements, all of which are infinite).

We now give an axiom which easily allows us to create subsets out of larger sets.

Axiom 3.6 (*Axiom of specification*). Let A be a set, and for each $x \in A$, let $P(x)$ be a property pertaining to x (i.e., for each $x \in A$, $P(x)$ is either a true statement or a false statement). Then there exists a set, called $\{x \in A : P(x) \text{ is true}\}$ (or simply $\{x \in A : P(x)\}$ for short), whose elements are precisely the elements x in A for which $P(x)$ is true. In other words, for any object y,

$$y \in \{x \in A : P(x) \text{ is true}\} \iff (y \in A \text{ and } P(y) \text{ is true}).$$

This axiom is also known as the *axiom of separation*. Note that $\{x \in A : P(x) \text{ is true}\}$ is always a subset of A (why?), though it could be as large as A or as small as the empty set. One can verify that the axiom of substitution works for specification, thus if $A = A'$ then $\{x \in A : P(x)\} = \{x \in A' : P(x)\}$ (why?).

Example 3.1.21 Let $S := \{1, 2, 3, 4, 5\}$. Then the set $\{n \in S : n < 4\}$ is the set of those elements n in S for which $n < 4$ is true, i.e., $\{n \in S : n < 4\} = \{1, 2, 3\}$. Similarly, the set $\{n \in S : n < 7\}$ is the same as S itself, while $\{n \in S : n < 1\}$ is the empty set.

We sometimes write $\{x \in A \big| P(x)\}$ instead of $\{x \in A : P(x)\}$; this is useful when we are using the colon ":" to denote something else, for instance to denote the domain and codomain of a function $f \colon X \to Y$. We can also describe $\{x \in A : P(x)\}$ in words as "the set of all x in A such that $P(x)$ is true".

We can use this axiom of specification to define some further operations on sets, namely intersections and difference sets.

Definition 3.1.22 (*Intersections*). The *intersection* $S_1 \cap S_2$ of two sets is defined to be the set

$$S_1 \cap S_2 := \{x \in S_1 : x \in S_2\}.$$

In other words, $S_1 \cap S_2$ consists of all the elements which belong to both S_1 and S_2. Thus, for all objects x,

$$x \in S_1 \cap S_2 \iff x \in S_1 \text{ and } x \in S_2.$$

Remark 3.1.23 Note that this definition is well-defined (i.e., it obeys the axiom of substitution, see Sect. A.7) because it is defined in terms of more primitive operations which were already known to obey the axiom of substitution. Similar remarks apply to future definitions in this chapter and will usually not be mentioned explicitly again.

Examples 3.1.24 We have $\{1, 2, 4\} \cap \{2, 3, 4\} = \{2, 4\}$, $\{1, 2\} \cap \{3, 4\} = \emptyset$, $\{2, 3\} \cup \emptyset = \{2, 3\}$, and $\{2, 3\} \cap \emptyset = \emptyset$.

Remark 3.1.25 By the way, one should be careful with the English word "and": rather confusingly, it can mean either union or intersection, depending on context. For instance, if one talks about a set of "boys and girls", one means the *union* of a set of boys with a set of girls, but if one talks about the set of people who are single and male, then one means the *intersection* of the set of single people with the set of male people. (Can you work out the rule of grammar that determines when "and" means union and when "and" means intersection?) Another problem is that "and" is also used in English to denote addition, thus for instance one could say that "2 and 3 is 5", while also saying that "the elements of $\{2\}$ and the elements of $\{3\}$ form the set $\{2, 3\}$" and "the elements in $\{2\}$ and $\{3\}$ form the set $\emptyset$". This can certainly get confusing! One reason we resort to mathematical symbols instead of English words such as "and" is that mathematical symbols always have a precise and unambiguous meaning, whereas one must often look very carefully at the context in order to work out what an English word means.

Two sets A, B are said to be *disjoint* if $A \cap B = \emptyset$. Note that this is not the same concept as being *distinct*, $A \neq B$. For instance, the sets $\{1, 2, 3\}$ and $\{2, 3, 4\}$ are distinct (there are elements of one set which are not elements of the other) but not disjoint (because their intersection is non-empty). Meanwhile, the sets $\emptyset$ and $\emptyset$ are disjoint but not distinct (why?).

There is an operation on sets that is somewhat analogous to subtraction:

Definition 3.1.26 (*Difference sets*). Given two sets A and B, we define the set $A - B$ or $A \backslash B$ to be the set A with any elements of B removed:

$$A \backslash B := \{x \in A : x \notin B\};$$

for instance, $\{1, 2, 3, 4\} \backslash \{2, 4, 6\} = \{1, 3\}$. In many cases B will be a subset of A, but not necessarily.

We now give some basic properties of unions, intersections, and difference sets.

Proposition 3.1.27 (Sets form a boolean algebra). *Let A, B, C be sets, and let X be a set containing A, B, C as subsets.*

(a) *(Minimal element) We have $A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$.*
(b) *(Maximal element) We have $A \cup X = X$ and $A \cap X = A$.*
(c) *(Identity) We have $A \cap A = A$ and $A \cup A = A$.*
(d) *(Commutativity) We have $A \cup B = B \cup A$ and $A \cap B = B \cap A$.*
(e) *(Associativity) We have $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.*
(f) *(Distributivity) We have $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.*
(g) *(Partition) We have $A \cup (X \backslash A) = X$ and $A \cap (X \backslash A) = \emptyset$.*
(h) *(De Morgan laws) We have $X \backslash (A \cup B) = (X \backslash A) \cap (X \backslash B)$ and $X \backslash (A \cap B) = (X \backslash A) \cup (X \backslash B)$.*

Remark 3.1.28 The de Morgan laws are named after the logician Augustus De Morgan (1806–1871), who identified them as one of the basic laws of set theory.

Proof See Exercise 3.1.6. □

Remark 3.1.29 The reader may observe a certain symmetry in the above laws between $\cup$ and $\cap$, and between X and $\emptyset$. This is an example of *duality*—two distinct properties or objects being dual to each other. In this case, the duality is manifested by the complementation relation $A \mapsto X\backslash A$; the de Morgan laws assert that this relation converts unions into intersections and vice versa. (It also interchanges X and the empty set.) The above laws are collectively known as the *laws of Boolean algebra*, after the mathematician George Boole (1815–1864), and are also applicable to a number of other objects other than sets; they play a particularly important rôle in mathematical logic.

We have now accumulated a number of axioms and results about sets, but there are still many things we are not able to do yet. One of the basic things we wish to do with a set is take each of the objects of that set, and somehow transform each such object into a new object; for instance we may wish to start with a set of numbers, say $\{3, 5, 9\}$, and increment each one, creating a new set $\{4, 6, 10\}$. This is not something we can do directly using only the axioms we already have, so we need a new axiom:

Axiom 3.7 (*Replacement*). Let A be a set. For any object $x \in A$, and any object y, suppose we have a statement $P(x, y)$ pertaining to x and y, such that for each $x \in A$ there is at most one y for which $P(x, y)$ is true. Then there exists a set $\{y : P(x, y) \text{ is true for some } x \in A\}$, such that for any object z,

$$\begin{aligned} z \in &\{y : P(x, y) \text{ is true for some } x \in A\} \\ &\iff P(x, z) \text{ is true for some } x \in A. \end{aligned}$$

Example 3.1.30 Let $A := \{3, 5, 9\}$, and let $P(x, y)$ be the statement $y = x{+}{+}$, i.e., y is the successor of x. Observe that for every $x \in A$, there is exactly one y for which $P(x, y)$ is true—specifically, the successor of x. Thus the above axiom asserts that the set $\{y : y = x{+}{+} \text{ for some } x \in \{3, 5, 9\}\}$ exists; in this case, it is clearly the same set as $\{4, 6, 10\}$ (why?).

Example 3.1.31 Let $A = \{3, 5, 9\}$, and let $P(x, y)$ be the statement $y = 1$. Then again for every $x \in A$, there is exactly one y for which $P(x, y)$ is true—specifically, the number 1. In this case $\{y : y = 1 \text{ for some } x \in \{3, 5, 9\}\}$ is just the singleton set $\{1\}$; we have replaced each element 3, 5, 9 of the original set A by the same object, namely 1. Thus this rather silly example shows that the set obtained by the above axiom can be "smaller" than the original set.

We often abbreviate a set of the form

$$\{y : y = f(x) \text{ for some } x \in A\}$$

as $\{f(x) : x \in A\}$ or $\{f(x)\big| x \in A\}$. Thus for instance, if $A = \{3, 5, 9\}$, then $\{x{+}{+} : x \in A\}$ is the set $\{4, 6, 10\}$. We can of course combine the axiom of replacement with the axiom of specification, thus for instance we can create sets such as $\{f(x) :$

$x \in A$; $P(x)$ is true} by starting with the set A, using the axiom of specification to create the set $\{x \in A : P(x) \text{ is true}\}$, and then applying the axiom of replacement to create $\{f(x) : x \in A; P(x) \text{ is true}\}$. Thus for instance $\{n{+}{+} : n \in \{3, 5, 9\}; n < 6\} = \{4, 6\}$.

In many of our examples we have implicitly assumed that natural numbers are in fact objects. Let us formalize this as follows.

Axiom 3.8 (*Infinity*). There exists a set $\mathbf{N}$, whose elements are called natural numbers, as well as an object 0 in $\mathbf{N}$, and an object $n{+}{+}$ assigned to every natural number $n \in \mathbf{N}$, such that the Peano axioms (Axioms 2.1–2.5) hold.

This is the more formal version of Assumption 2.6. It is called the axiom of infinity because it introduces the most basic example of an infinite set, namely the set of natural numbers $\mathbf{N}$. (We will formalize what finite and infinite mean in Sect. 3.6.) From the axiom of infinity we see that numbers such as 3, 5, and 7 are indeed objects in set theory, and so (from the pair set axiom and pairwise union axiom) we can indeed legitimately construct sets such as $\{3, 5, 9\}$ as we have been doing in our examples.

One has to keep the concept of a set distinct from the elements of that set; for instance, the set $\{n + 3 : n \in \mathbf{N}, 0 \le n \le 5\}$ is not the same thing as the expression or function $n + 3$. We emphasize this with an example:

Example 3.1.32 (Informal) This example requires the notion of subtraction, which has not yet been formally introduced. The following two sets are equal,

$$\{n + 3 : n \in \mathbf{N}, 0 \le n \le 5\} = \{8 - n : n \in \mathbf{N}, 0 \le n \le 5\}, \tag{3.1}$$

(see below), even though the expressions $n + 3$ and $8 - n$ are never equal to each other for any natural number n. Thus, it is a good idea to remember to use those curly braces {} when you talk about sets, lest you accidentally confuse a set with its elements. One reason for this counterintuitive situation is that the letter n is being used in two different ways on the two sides of (3.1). To clarify the situation, let us rewrite the set $\{8 - n : n \in \mathbf{N}, 0 \le n \le 5\}$ by replacing the letter n by the letter m, thus giving $\{8 - m : m \in \mathbf{N}, 0 \le m \le 5\}$. This is exactly the same set as before (why?), so we can rewrite (3.1) as

$$\{n + 3 : n \in \mathbf{N}, 0 \le n \le 5\} = \{8 - m : m \in \mathbf{N}, 0 \le m \le 5\}.$$

Now it is easy to see (using Axiom 3.2) why this identity is true: every number of the form $n + 3$, where n is a natural number between 0 and 5, is also of the form $8 - m$ where $m := 5 - n$ (note that m is therefore also a natural number between 0 and 5); conversely, every number of the form $8 - m$, where m is a natural number between 0 and 5, is also of the form $n + 3$, where $n := 5 - m$ (note that n is therefore a natural number between 0 and 5). Observe how much more confusing the above explanation of (3.1) would have been if we had not changed one of the n's to an m first!

Formally, we can refer to **N** as "the set of natural numbers", but we shall often abbreviate this to simply "the natural numbers". Similarly for some other sets that we will introduce later in this text; for instance **Z** will be "the set of integers" but also the "integers", **R** will be the "set of real numbers" but also "the real numbers" or even just "the reals", and so forth.

— Exercises —

Exercise 3.1.1 Let a, b, c, d be objects such that $\{a, b\} = \{c, d\}$. Show that at least one of the two statements "$a = c$ and $b = d$" and "$a = d$ and $b = c$" hold.

Exercise 3.1.2 Using only Axiom 3.2, Axiom 3.1, Axiom 3.3, and Axiom 3.4, prove that the sets $\emptyset$, $\{\emptyset\}$, $\{\{\emptyset\}\}$, and $\{\emptyset, \{\emptyset\}\}$ are all distinct (i.e., no two of them are equal to each other).

Exercise 3.1.3 Prove the remaining claims in Lemma 3.1.12.

Exercise 3.1.4 Prove the remaining claims in Proposition 3.1.17.

Exercise 3.1.5 Let A, B be sets. Show that the three statements $A \subseteq B$, $A \cup B = B$, $A \cap B = A$ are logically equivalent (any one of them implies the other two).

Exercise 3.1.6 Prove Proposition 3.1.27. (*Hint:* one can use some of these claims to prove others. Some of the claims have also appeared previously in Lemma 3.1.12.)

Exercise 3.1.7 Let A, B, C be sets. Show that $A \cap B \subseteq A$ and $A \cap B \subseteq B$. Furthermore, show that $C \subseteq A$ and $C \subseteq B$ if and only if $C \subseteq A \cap B$. In a similar spirit, show that $A \subseteq A \cup B$ and $B \subseteq A \cup B$, and furthermore that $A \subseteq C$ and $B \subseteq C$ if and only if $A \cup B \subseteq C$.

Exercise 3.1.8 Let A, B be sets. Prove the *absorption laws* $A \cap (A \cup B) = A$ and $A \cup (A \cap B) = A$.

Exercise 3.1.9 Let A, B, X be sets such that $A \cup B = X$ and $A \cap B = \emptyset$. Show that $A = X \backslash B$ and $B = X \backslash A$.

Exercise 3.1.10 Let A and B be sets. Show that the three sets $A \backslash B$, $A \cap B$, and $B \backslash A$ are disjoint, and that their union is $A \cup B$.

Exercise 3.1.11 Show that the axiom of replacement implies the axiom of specification.

Exercise 3.1.12 Suppose that A, B, A', B' are sets such that $A' \subseteq A$ and $B' \subseteq B$.

(i) Show that $A' \cup B' \subseteq A \cup B$ and $A' \cap B' \subseteq A \cap B$.
(ii) Give a counterexample to show that the statement $A' \backslash B' \subseteq A \backslash B$ is false. Can you find a modification of this statement involving the set difference operation $\backslash$ which is true given the stated hypotheses? Justify your answer.

Exercise 3.1.13 Euclid famously defined a point to be "that which has no part". This exercise should be reminiscent of that definition. Define a *proper subset* of a set A to be a subset B of A with $B \neq A$. Let A be a non-empty set. Show that A does not have any non-empty proper subsets if and only if A is of the form $A = \{x\}$ for some object x.

3.2 Russell's Paradox (Optional)

Many of the axioms introduced in the previous section have a similar flavor: they allow us to form a set consisting of all the elements which have a certain property. These axioms are plausible, but one might think that they could be unified, for instance by introducing the following axiom:

Axiom 3.9 (*Universal specification*). (Dangerous!) Suppose for every object x we have a property $P(x)$ pertaining to x (so that for every x, $P(x)$ is either a true statement or a false statement). Then there exists a set $\{x : P(x) \text{ is true}\}$ such that for every object y,

$$y \in \{x : P(x) \text{ is true}\} \iff P(y) \text{ is true.}$$

This axiom is also known as the *axiom of comprehension*. It asserts that every property corresponds to a set; if we assumed that axiom, we could talk about the set of all blue objects, the set of all natural numbers, the set of all sets, and so forth. This axiom also implies most of the axioms in the previous section (Exercise 3.2.1). Unfortunately, this axiom cannot be introduced into set theory, because it creates a logical contradiction known as *Russell's paradox*, discovered by the philosopher and logician Bertrand Russell (1872–1970) in 1901. The paradox runs as follows. Let $P(x)$ be the statement

$$P(x) \iff \text{“}x \text{ is a set, and } x \notin x\text{”;}$$

i.e., $P(x)$ is true only when x is a set which does not contain itself. For instance, $P(\{2, 3, 4\})$ is true, since the set $\{2, 3, 4\}$ is not one of the three elements 2, 3, 4 of $\{2, 3, 4\}$. On the other hand, if we let S be the set of all sets (which we would know to exist from the axiom of universal specification), then since S is itself a set, it is an element of S, and so $P(S)$ is false. Now use the axiom of universal specification to create the set

$$\Omega := \{x : P(x) \text{ is true}\} = \{x : x \text{ is a set and } x \notin x\},$$

i.e., the set of all sets which do not contain themselves. Now ask the question: does Ω contain itself, i.e. is $\Omega \in \Omega$? If Ω did contain itself, then by the definition of Ω this means that $P(\Omega)$ is true, i.e., Ω is a set and $\Omega \notin \Omega$. On the other hand, if Ω did not contain itself, then by the definition of P $P(\Omega)$ would be true, and hence by the definition of Ω we have $\Omega \in \Omega$. Thus in either case we have both $\Omega \in \Omega$ and $\Omega \notin \Omega$, which is absurd.

The problem with the above axiom is that it creates sets which are far too "large"—for instance, we can use that axiom to talk about the set of *all* objects (a so-called universal set). Since sets are themselves objects (Axiom 3.1), this means that sets are allowed to contain themselves, which is a somewhat silly state of affairs. One way to informally resolve this issue is to think of objects as being arranged in a hierarchy. At the bottom of the hierarchy are the *primitive objects*—the objects that are not sets,[1] such as the natural number 37. Then on the next rung of the hierarchy there are sets whose elements consist only of primitive objects, such as $\{3, 4, 7\}$ or the empty set $\emptyset$; let's call these "primitive sets" for now. Then there are sets whose elements consist only of primitive objects and primitive sets, such as $\{3, 4, 7, \{3, 4, 7\}\}$. Then we can form sets out of these objects, and so forth. The point is that at each stage of the hierarchy we only see sets whose elements consist of objects at lower stages of the hierarchy, and so at no stage do we ever construct a set which contains itself.

To actually formalize the above intuition of a hierarchy of objects is actually rather complicated, and we will not do so here. Instead, we shall simply postulate an axiom which ensures that absurdities such as Russell's paradox do not occur.

Axiom 3.10 (*Regularity*). If A is a non-empty set, then there is at least one element x of A which is either not a set, or is disjoint from A.

The point of this axiom (which is also known as the *axiom of foundation*) is that it is asserting that at least one of the elements of A is so low on the hierarchy of objects that it does not contain any of the other elements of A. For instance, if $A = \{\{3, 4\}, \{3, 4, \{3, 4\}\}\}$, then the element $\{3, 4\} \in A$ does not contain any of the elements of A (neither 3 nor 4 lies in A), although the element $\{3, 4, \{3, 4\}\}$, being somewhat higher in the hierarchy, does contain an element of A, namely $\{3, 4\}$. One particular consequence of this axiom is that sets are no longer allowed to contain themselves (Exercise 3.2.2).

One can legitimately ask whether we really need this axiom in our set theory, as it is certainly less intuitive than our other axioms. For the purposes of doing analysis, it turns out in fact that this axiom is never needed; all the sets we consider in analysis are typically very low on the hierarchy of objects, for instance being sets of primitive objects, or sets of sets of primitive objects, or at worst sets of sets of sets of primitive objects. However it is necessary to include this axiom in order to perform more advanced set theory, and so we have included this axiom in the text (but in an optional section) for sake of completeness.

[1] In pure set theory, there will be no primitive objects, but there will be one primitive set $\emptyset$ on the next rung of the hierarchy.

— Exercises —

Exercise 3.2.1 Show that the universal specification axiom, Axiom 3.9, if assumed to be true, would imply Axioms 3.3, 3.4, 3.5, 3.6, and 3.7. (If we assume that all natural numbers are objects, we also obtain Axiom 3.8.) Thus, this axiom, if permitted, would simplify the foundations of set theory tremendously (and can be viewed as one basis for an intuitive model of set theory known as "naive set theory"). Unfortunately, as we have seen, Axiom 3.9 is "too good to be true"!

Exercise 3.2.2 Use the axiom of regularity (and the singleton set axiom) to show that if A is a set, then $A \notin A$. Furthermore, show that if A and B are two sets, then either $A \notin B$ or $B \notin A$ (or both). (One corollary of this exercise is worth noting: given any set A, there exists a mathematical object that is not an element in A, namely A itself. Thus one can always "add one more element" to a set A to create a larger set, namely $A \cup \{A\}$.)

Exercise 3.2.3 Show (assuming the other axioms of set theory) that the universal specification axiom, Axiom 3.9, is equivalent to an axiom postulating the existence of a "universal set" Ω consisting of all objects (i.e., for all objects x, we have $x \in \Omega$). In other words, if Axiom 3.9 is true, then a universal set exists, and conversely, if a universal set exists, then Axiom 3.9 is true. (This helps explain why Axiom 3.9 is called the axiom of *universal* specification.) Note that if a universal set Ω existed, then we would have $\Omega \in \Omega$ by Axiom 3.1, contradicting Exercise 3.2.2. Thus the axiom of foundation specifically rules out the axiom of universal specification.

3.3 Functions

In order to do analysis, it is not particularly useful to just have the notion of a set; we also need the notion of a *function* from one set to another. Informally, a function $f: X \to Y$ from one set X to another set Y is an operation which assigns to each element (or "input") x in X, a single element (or "output") $f(x)$ in Y; we have already used this informal concept in the previous chapter when we discussed the natural numbers. The formal definition is as follows.

Definition 3.3.1 (*Functions*) Let X, Y be sets, and let $P(x, y)$ be a property pertaining to an object $x \in X$ and an object $y \in Y$, such that for every $x \in X$, there is exactly one $y \in Y$ for which $P(x, y)$ is true (this is sometimes known as the *vertical line test*). Then we define the *function $f: X \to Y$ defined by P on the domain X and codomain*[2] to be the object which, given any input $x \in X$, assigns an output $f(x) \in Y$, defined to be the unique object $f(x) \in Y$ for which $P(x, f(x))$ is true. Thus, for any $x \in X$ and $y \in Y$,

$$y = f(x) \iff P(x, y) \text{ is true.}$$

Functions are also referred to as *maps* or *transformations*, depending on the context. They are also sometimes called *morphisms*, although to be more precise, a morphism refers to a more general class of object, which may or may not correspond to actual functions, depending on the context.

[2] In some texts the codomain is referred to as the *range*; however we will use the term range to refer instead to the image $f(X)$ of the domain, defined after Definition 3.4.1.

Remark 3.3.2 Implicit in the above definition is an assumption that whenever one is given two sets X, Y and a property P obeying the vertical line test, one can form a function object f. Strictly speaking, the assumption of the existence of such a function object f should be stated as an explicit axiom. However, we will not do so here, as it turns out to be redundant. (More precisely, in view of Exercise 3.5.10, it is always possible to encode a function f as an ordered triple $(X, Y, \{(x, f(x)) : x \in X\})$ consisting of the domain, codomain, and graph of the function, which gives a way to build functions as objects using the operations provided by the preceding axioms of set theory.) Also implicit in the above definition is the understanding that every function f is automatically associated with a domain X, a codomain Y, and a defining property P.

Example 3.3.3 Let $X = \mathbf{N}$, $Y = \mathbf{N}$, and let $P(x, y)$ be the property that $y = x{+}{+}$. Then for each $x \in \mathbf{N}$ there is exactly one $y \in \mathbf{N}$ for which $P(x, y)$ is true, namely $y = x{+}{+}$. Thus we can define a function $f : \mathbf{N} \to \mathbf{N}$ associated to this property, so that $f(x) = x{+}{+}$ for all x; this is the *increment* function on $\mathbf{N}$, which takes a natural number as input and returns its increment as output. Thus for instance $f(4) = 5$, $f(2n + 3) = 2n + 4$ and so forth. One might also hope to define a *decrement* function $g : \mathbf{N} \to \mathbf{N}$ associated to the property $P(x, y)$ defined by $y{+}{+} = x$, i.e., $g(x)$ would be the number whose increment is x. Unfortunately this does not define a function, because when $x = 0$ there is no natural number y whose increment is equal to x (Axiom 2.3). On the other hand, we can legitimately define a decrement function $h : \mathbf{N}\backslash\{0\} \to \mathbf{N}$ associated to the property $P(x, y)$ defined by $y{+}{+} = x$, because when $x \in \mathbf{N}\backslash\{0\}$ there is indeed exactly one natural number y such that $y{+}{+} = x$, thanks to Lemma 2.2.10. Thus for instance $h(4) = 3$ and $h(2n + 3) = 2n + 2$, but $h(0)$ is undefined since 0 is not in the domain $\mathbf{N}\backslash\{0\}$.

Example 3.3.4 (Informal) This example requires the real numbers $\mathbf{R}$, which we will define in Chap. 5. One could try to define a square root function $\sqrt{\ }: \mathbf{R} \to \mathbf{R}$ by associating it to the property $P(x, y)$ defined by $y^2 = x$, i.e., we would want $\sqrt{x}$ to be the number y such that $y^2 = x$. Unfortunately there are two problems which prohibit this definition from actually creating a function. The first is that there exist real numbers x for which $P(x, y)$ is never true, for instance if $x = -1$ then there is no real number y such that $y^2 = x$. This problem however can be solved by restricting the domain from $\mathbf{R}$ to the right half-line $[0, +\infty)$. The second problem is that even when $x \in [0, +\infty)$, it is possible for there to be more than one y in the codomain $\mathbf{R}$ for which $y^2 = x$, for instance if $x = 4$ then both $y = 2$ and $y = -2$ obey the property $P(x, y)$, i.e., both $+2$ and -2 are square roots of 4. This problem can however be solved by restricting the codomain of $\mathbf{R}$ to $[0, +\infty)$. Once one does this, then one can correctly define a square root function $\sqrt{\ } : [0, +\infty) \to [0, +\infty)$ using the relation $y^2 = x$; thus $\sqrt{x}$ is the unique number $y \in [0, +\infty)$ such that $y^2 = x$.

One common way to define a function is simply to specify its domain, its codomain, and how one generates the output $f(x)$ from each input; this is known as an *explicit* definition of a function. For instance, the function f in Example 3.3.3 could be defined explicitly by saying that f has domain and codomain equal to $\mathbf{N}$, and $f(x) := x{++}$ for all $x \in \mathbf{N}$. In other cases we only define a function f by specifying what property $P(x, y)$ links the input x with the output $f(x)$; this is an *implicit* definition of a function. For instance, the square root function $\sqrt{x}$ in Example 3.3.4 was defined implicitly by the relation $(\sqrt{x})^2 = x$. Note that an implicit definition is only valid if we know that for every input there is exactly one output which obeys the implicit relation. In many cases we omit specifying the domain and codomain of a function for brevity, and thus for instance we could refer to the function f in Example 3.3.3 as "the function $f(x) := x{++}$", "the function $x \mapsto x{++}$", "the function $x{++}$", or even the extremely abbreviated "++". However, too much of this abbreviation can be dangerous; sometimes it is important to know what the domain and codomain of the function is.

We observe that functions obey the axiom of substitution: if $x = x'$, then $f(x) = f(x')$ (why?). In other words, equal inputs imply equal outputs. On the other hand, unequal inputs do not necessarily ensure unequal outputs, as the following example shows:

Example 3.3.5 Let $X = \mathbf{N}$, $Y = \mathbf{N}$, and let $P(x, y)$ be the property that $y = 7$. Then certainly for every $x \in \mathbf{N}$ there is exactly one y for which $P(x, y)$ is true, namely the number 7. Thus we can create a function $f : \mathbf{N} \to \mathbf{N}$ associated to this property; it is simply the *constant function* which assigns the output of $f(x) = 7$ to each input $x \in \mathbf{N}$. Thus it is certainly possible for different inputs to generate the same output.

Remark 3.3.6 We are now using parentheses () to denote several different things in mathematics; on one hand, we are using them to clarify the order of operations (compare for instance $2 + (3 \times 4) = 14$ with $(2 + 3) \times 4 = 20$), but on the other hand we also use parentheses to enclose the argument x of a function $f(x)$ or of a property such as $P(x)$. However, the two usages of parentheses usually are unambiguous from context. For instance, if a is a number, then $a(b + c)$ denotes the expression $a \times (b + c)$, whereas if f is a function, then $f(b + c)$ denotes the output of f when the input is $b + c$. Sometimes the argument of a function is denoted by subscripting instead of parentheses; for instance, a sequence of natural numbers $a_0, a_1, a_2, a_3, \ldots$ is, strictly speaking, a function from $\mathbf{N}$ to $\mathbf{N}$, but is denoted by $n \mapsto a_n$ rather than $n \mapsto a(n)$.

Remark 3.3.7 We do not necessarily require functions to be sets, nor do we require sets to be functions. Thus, it does not necessarily make sense to ask whether an object x is an element of a function f, and it does not necessarily make sense to apply a set A to an input x to create an output $A(x)$. On the other hand, it is permissible to start with a function $f : X \to Y$ and construct its *graph* $\{(x, f(x)) : x \in X\}$, which describes the function completely once the domain X and codomain Y are specified: see Sect. 3.5.

We now define some basic concepts and notions for functions. The first notion is that of equality.

Definition 3.3.8 (*Equality of functions*). Two functions $f : X \to Y$, $g : X' \to Y'$ are said to be equal if their domains and codomains agree (i.e., $X = X'$ and $Y = Y'$), and furthermore that $f(x) = g(x)$ for *all* $x \in X$. If $f(x)$ and $g(x)$ agree for some values of x in the domain, but not others, then we do not consider f and g to be equal.[3] If two functions f, g have different domains, or different ranges, we also do not consider them to be equal.

Remark 3.3.9 According to this definition, two functions that have different domains or different codomains are, strictly speaking, distinct functions. However, when it is safe to do so without causing confusion, it is sometimes useful to "abuse notation" by identifying together functions of different domains or codomains if their values agree ontheir common domain of definition; this is analogous to the practice of "overloading" an operator in software engineering. See the discussion after Definition 9.4.1 for one instance of this.

Example 3.3.10 (Informal) The functions $x \mapsto x^2 + 2x + 1$ and $x \mapsto (x + 1)^2$ are equal on the domain $\mathbf{R}$. The functions $x \mapsto x$ and $x \mapsto |x|$ are equal on the positive real axis, but are not equal on $\mathbf{R}$; thus the concept of equality of functions can depend on the choice of domain.

Example 3.3.11 A rather boring example of a function is the *empty function* $f : \emptyset \to X$ from the empty set to a given set X. Since the empty set has no elements, we do not need to specify what f does to any input. Nevertheless, just as the empty set is a set, the empty function is a function, albeit not a particularly interesting one. Note that for each set X, there is only one function from $\emptyset$ to X, since Definition 3.3.8 asserts that all functions from $\emptyset$ to X are equal (why?).

Remark 3.3.12 It is not immediately apparent that Definition 3.3.8 is compatible with the axioms of equality in Appendix A.7, although Exercise 3.3.1 provides evidence toward this compatibility. There are at least three ways to address this issue. One is to regard Definition 3.3.8 as an *axiom* about equality of functions rather than a definition. Another is to provide a more explicit definition of a function in which Definition 3.3.8 becomes a theorem; for instance, one can define a function $f : X \to Y$ to be an ordered triple (X, Y, G) consisting of a domain set X, a codomain set Y, and a graph $G = \{(x, f(x)) : x \in X\}$ that obeys the vertical line test and use this latter graph to define the value of $f(x) \in Y$ for each element x of the domain; see Exercise 3.5.10. A third way is to start with a mathematical universe $\mathcal{U}$ without any functions in it and use Definition 3.3.8 to create a larger extension of this universe that contains function objects that behave as specified as in Definition 3.3.8. This final procedure however requires a bit more of the formalism of logic and model theory than is provided by this text, and so will not be detailed here.

A fundamental operation available for functions is *composition*.

Definition 3.3.13 (*Composition*). Let $f : X \to Y$ and $g : Y \to Z$ be two functions, such that the codomain of f is the same set as the domain of g. We then define the

[3] In Chap. 8 of *Analysis II*, we shall introduce a weaker notion of equality, that of two functions being *equal almost everywhere*.

composition $g \circ f \colon X \to Z$ of the two functions g and f to be the function defined explicitly by the formula

$$(g \circ f)(x) := g(f(x)).$$

If the codomain of f does not match the domain of g, we leave the composition $g \circ f$ undefined.

It is easy to check that composition obeys the axiom of substitution (Exercise 3.3.1).

Example 3.3.14 Let $f \colon \mathbf{N} \to \mathbf{N}$ be the function $f(n) := 2n$, and let $g \colon \mathbf{N} \to \mathbf{N}$ be the function $g(n) := n + 3$. Then $g \circ f$ is the function

$$g \circ f(n) = g(f(n)) = g(2n) = 2n + 3,$$

thus for instance $g \circ f(1) = 5$, $g \circ f(2) = 7$, and so forth. Meanwhile, $f \circ g$ is the function

$$f \circ g(n) = f(g(n)) = f(n + 3) = 2(n + 3) = 2n + 6,$$

thus for instance $f \circ g(1) = 8$, $f \circ g(2) = 10$, and so forth.

The above example shows that composition is not commutative: $f \circ g$ and $g \circ f$ are not necessarily the same function. However, composition is still associative:

Lemma 3.3.15 (Composition is associative). *Let $f \colon Z \to W$, $g \colon Y \to Z$, and $h \colon X \to Y$ be functions. Then $f \circ (g \circ h) = (f \circ g) \circ h$.*

Proof Since $g \circ h$ is a function from X to Z, $f \circ (g \circ h)$ is a function from X to W. Similarly $f \circ g$ is a function from Y to W, and hence $(f \circ g) \circ h$ is a function from X to W. Thus $f \circ (g \circ h)$ and $(f \circ g) \circ h$ have the same domain and codomain. In order to check that they are equal, we see from Definition 3.3.8 that we have to verify that $(f \circ (g \circ h))(x) = ((f \circ g) \circ h)(x)$ for all $x \in X$. But by Definition 3.3.13

$$\begin{aligned}(f \circ (g \circ h))(x) &= f((g \circ h)(x)) \\ &= f(g(h(x)) \\ &= (f \circ g)(h(x)) \\ &= ((f \circ g) \circ h)(x)\end{aligned}$$

as desired. □

Remark 3.3.16 Note that while g appears to the left of f in the expression $g \circ f$, the function $g \circ f$ applies the right-most function f first, before applying g. This is often confusing at first; it arises because we traditionally place a function f to the left of its input x rather than to the right. (There are some alternate mathematical notations in which the function is placed to the right of the input; thus we would write xf instead of $f(x)$, but this notation has often proven to be more confusing than clarifying and has not as yet become particularly popular.)

We now describe certain special types of functions: *one-to-one* functions, *onto* functions, and *invertible* functions.

Definition 3.3.17 (*One-to-one functions*). A function f is *one-to-one* (or *injective*) if different elements map to different elements:

$$x \neq x' \quad \Longrightarrow \quad f(x) \neq f(x').$$

Equivalently, a function is one-to-one if

$$f(x) = f(x') \quad \Longrightarrow \quad x = x'.$$

Example 3.3.18 (Informal) The function $f: \mathbf{Z} \to \mathbf{Z}$ defined by $f(n) := n^2$ is not one-to-one because the distinct elements -1, 1 map to the same element 1. On the other hand, if we restrict this function to the natural numbers, defining the function $g: \mathbf{N} \to \mathbf{Z}$ by $g(n) := n^2$, then g is now a one-to-one function. Thus the notion of a one-to-one function depends not just on what the function does, but also what its domain is.

Remark 3.3.19 If a function $f: X \to Y$ is not one-to-one, then one can find distinct x and x' in the domain X such that $f(x) = f(x')$, thus one can find two inputs which map to one output. Because of this, we say that f is *two-to-one* instead of *one-to-one*.

Definition 3.3.20 (*Onto functions*). A function f is *onto* (or *surjective*) if every element in Y comes from applying f to some element in X:

$$\text{For every } y \in Y, \text{ there exists } x \in X \text{ such that } f(x) = y.$$

Example 3.3.21 (Informal) The function $f: \mathbf{Z} \to \mathbf{Z}$ defined by $f(n) := n^2$ is not onto because the negative numbers are not in the image of f. However, if we restrict the codomain $\mathbf{Z}$ to the set $A := \{n^2 : n \in \mathbf{Z}\}$ of square numbers, then the function $g: \mathbf{Z} \to A$ defined by $g(n) := n^2$ is now onto. Thus the notion of an onto function depends not just on what the function does, but also what its range is.

Remark 3.3.22 The concepts of injectivity and surjectivity are in many ways dual to each other; see Exercises 3.3.2, 3.3.4, 3.3.5 for some evidence of this.

Definition 3.3.23 (*Bijective functions*). Functions $f: X \to Y$ which are both one-to-one and onto are also called *bijective* or *invertible*.

Example 3.3.24 Let $f: \{0, 1, 2\} \to \{3, 4\}$ be the function $f(0) := 3$, $f(1) := 3$, $f(2) := 4$. This function is not bijective because if we set $y = 3$, then there is more than one x in $\{0, 1, 2\}$ such that $f(x) = y$ (this is a failure of injectivity). Now let $g: \{0, 1\} \to \{2, 3, 4\}$ be the function $g(0) := 2$, $g(1) := 3$; then g is not bijective because if we set $y = 4$, then there is no x for which $g(x) = y$ (this is a failure of surjectivity). Now let $h: \{0, 1, 2\} \to \{3, 4, 5\}$ be the function $h(0) := 3$, $h(1) := 4$, $h(2) := 5$. Then h is bijective, because each of the elements 3, 4, 5 comes from exactly one element from 0, 1, 2.

Example 3.3.25 The function $f\colon \mathbf{N} \to \mathbf{N}\backslash\{0\}$ defined by $f(n) := n{+}{+}$ is a bijection (in fact, this fact is simply restating Lemma 2.2.10). On the other hand, the function $g\colon \mathbf{N} \to \mathbf{N}$ defined by the same definition $g(n) := n{+}{+}$ is not a bijection. Thus the notion of a bijective function depends not just on what the function does, but also what its domain and codomain are.

Remark 3.3.26 If a function $x \mapsto f(x)$ is bijective, then we sometimes call f a *perfect matching* or a *one-to-one correspondence* (not to be confused with the notion of a one-to-one function) and denote the action of f using the notation $x \leftrightarrow f(x)$ instead of $x \mapsto f(x)$. Thus for instance the function h in the above example is the one-to-one correspondence $0 \leftrightarrow 3$, $1 \leftrightarrow 4$, $2 \leftrightarrow 5$.

Remark 3.3.27 A common error is to say that a function $f\colon X \to Y$ is bijective iff "for every x in X, there is exactly one y in Y such that $y = f(x)$". This is not what it means for f to be bijective; rather, this is merely stating what it means for f to be a *function*. A function cannot map one element to two different elements, for instance one cannot have a function f for which $f(0) = 1$ and also $f(0) = 2$. The functions f, g given in Example 3.3.25 are not bijective, but they are still functions, since each input still gives exactly one output.

If f is bijective, then for every $y \in Y$, there is exactly one x such that $f(x) = y$ (there is at least one because of surjectivity, and at most one because of injectivity). This value of x is denoted $f^{-1}(y)$; thus f^{-1} is a function from Y to X. We call f^{-1} the *inverse* of f.

— Exercises —

Exercise 3.3.1 Show that the definition of equality in Definition 3.3.8 is reflexive, symmetric, and transitive. Also verify the substitution property: if $f, \tilde{f}\colon X \to Y$ and $g, \tilde{g}\colon Y \to Z$ are functions such that $f = \tilde{f}$ and $g = \tilde{g}$, then $g \circ f = \tilde{g} \circ \tilde{f}$. (Of course, these statements are immediate from the axioms of equality in Appendix A.7 applied directly to the functions in question, but the point of the exercise is to show that they can also be established by instead applying the axioms of equality to elements of the domain and codomain of these functions, rather than to the functions itself.)

Exercise 3.3.2 Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions. Show that if f and g are both injective, then so is $g \circ f$; similarly, show that if f and g are both surjective, then so is $g \circ f$.

Exercise 3.3.3 When is the empty function into a given set X injective? surjective? bijective?

Exercise 3.3.4 In this section we give some cancellation laws for composition. Let $f\colon X \to Y$, $\tilde{f}\colon X \to Y$, $g\colon Y \to Z$, and $\tilde{g}\colon Y \to Z$ be functions. Show that if $g \circ f = g \circ \tilde{f}$ and g is injective, then $f = \tilde{f}$. Is the same statement true if g is not injective? Show that if $g \circ f = \tilde{g} \circ f$ and f is surjective, then $g = \tilde{g}$. Is the same statement true if f is not surjective?

Exercise 3.3.5 Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions. Show that if $g \circ f$ is injective, then f must be injective. Is it true that g must also be injective? Show that if $g \circ f$ is surjective, then g must be surjective. Is it true that f must also be surjective?

Exercise 3.3.6 Let $f\colon X \to Y$ be a bijective function, and let $f^{-1}: Y \to X$ be its inverse. Verify the cancellation laws $f^{-1}(f(x)) = x$ for all $x \in X$ and $f(f^{-1}(y)) = y$ for all $y \in Y$. Conclude that f^{-1} is also invertible and has f as its inverse (thus $(f^{-1})^{-1} = f$).

Exercise 3.3.7 Let $f\colon X \to Y$ and $g\colon Y \to Z$ be functions. Show that if f and g are bijective, then so is $g \circ f$, and we have $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Exercise 3.3.8 If X is a subset of Y, let $\iota_{X \to Y} : X \to Y$ be the *inclusion map from X to Y*, defined by mapping $x \mapsto x$ for all $x \in X$, i.e., $\iota_{X \to Y}(x) := x$ for all $x \in X$. The map $\iota_{X \to X}$ is in particular called the *identity map* on X.

(a) Show that if $X \subseteq Y \subseteq Z$ then $\iota_{Y \to Z} \circ \iota_{X \to Y} = \iota_{X \to Z}$.
(b) Show that if $f\colon A \to B$ is any function, then $f = f \circ \iota_{A \to A} = \iota_{B \to B} \circ f$.
(c) Show that, if $f\colon A \to B$ is a bijective function, then $f \circ f^{-1} = \iota_{B \to B}$ and $f^{-1} \circ f = \iota_{A \to A}$.
(d) Show that if X and Y are disjoint sets, and $f\colon X \to Z$ and $g\colon Y \to Z$ are functions, then there is a unique function $h\colon X \cup Y \to Z$ such that $h \circ \iota_{X \to X \cup Y} = f$ and $h \circ \iota_{Y \to X \cup Y} = g$.
(e) Show that the hypothesis that X and Y are disjoint can be dropped in (d) if one adds the additional hypothesis that $f(x) = g(x)$ for all $x \in X \cap Y$.

3.4 Images and Inverse Images

We know that a function $f\colon X \to Y$ from a set X to a set Y can take individual elements $x \in X$ to elements $f(x) \in Y$. Functions can also take subsets in X to subsets in Y:

Definition 3.4.1 (*Images of sets*). If $f\colon X \to Y$ is a function from X to Y, and S is a subset of X, we define[4] $f(S)$ to be the set

$$f(S) := \{f(x) : x \in S\};$$

this set is a subset of Y and is sometimes called the *image* of S under the map f. We sometimes call $f(S)$ the *forward image* of S to distinguish it from the concept of the *inverse image* $f^{-1}(S)$ of S, which is defined below.

Note that the set $f(S)$ is well-defined thanks to the axiom of replacement (Axiom 3.7). One can also define $f(S)$ using the axiom of specification (Axiom 3.6) instead of replacement, but we leave this as an exercise to the reader. The image $f(X)$ of the domain is also known as the *range* of the function $f\colon X \to Y$; it is a subset of the codomain Y.

Example 3.4.2 If $f\colon \mathbf{N} \to \mathbf{N}$ is the map $f(x) = 2x$, then the forward image of $\{1, 2, 3\}$ is $\{2, 4, 6\}$:

$$f(\{1, 2, 3\}) = \{2, 4, 6\}.$$

More informally, to compute $f(S)$, we take every element x of S and apply f to each element individually, and then put all the resulting objects together to form a new set.

[4] In principle this notation could collide with the existing notation $f(x)$ for the evaluation of f at x, if S turns out to both be a subset of X and an element of X. However, we will ignore this potential collision as it rarely occurs in practice.

In the above example, the image had the same size as the original set. But sometimes the image can be smaller, because f is not one-to-one (see Definition 3.3.17):

Example 3.4.3 (Informal) Let $\mathbf{Z}$ be the set of integers (which we will define rigorously in the next section) and let $f: \mathbf{Z} \to \mathbf{Z}$ be the map $f(x) = x^2$, then

$$f(\{-1, 0, 1, 2\}) = \{0, 1, 4\}.$$

Note that f is not one-to-one because $f(-1) = f(1)$.

Note that

$$x \in S \implies f(x) \in f(S)$$

but in general

$$f(x) \in f(S) \not\Rightarrow x \in S;$$

for instance in the above informal example, $f(-2)$ lies in the set $f(\{-1, 0, 1, 2\})$, but -2 is not in $\{-1, 0, 1, 2\}$. The correct statement is

$$y \in f(S) \iff y = f(x) \text{ for some } x \in S$$

(why?).

Example 3.4.4 From Definition 3.3.20 we see that a function $f: X \to Y$ is onto if and only if $f(X) = Y$.

Definition 3.4.5 (*Inverse images*) If U is a subset of Y, we define the set $f^{-1}(U)$ to be the set

$$f^{-1}(U) := \{x \in X : f(x) \in U\}.$$

In other words, $f^{-1}(U)$ consists of all the elements of X which map into U:

$$f(x) \in U \iff x \in f^{-1}(U).$$

We call $f^{-1}(U)$ the *inverse image* of U.

Example 3.4.6 If $f: \mathbf{N} \to \mathbf{N}$ is the map $f(x) = 2x$, then $f(\{1, 2, 3\}) = \{2, 4, 6\}$, but $f^{-1}(\{1, 2, 3\}) = \{1\}$. Thus the forward image of $\{1, 2, 3\}$ and the backwards image of $\{1, 2, 3\}$ are quite different sets. Also note that

$$f(f^{-1}(\{1, 2, 3\})) \neq \{1, 2, 3\}$$

(why?).

Example 3.4.7 (Informal) If $f: \mathbf{Z} \to \mathbf{Z}$ is the map $f(x) = x^2$, then

$$f^{-1}(\{0, 1, 4\}) = \{-2, -1, 0, 1, 2\}.$$

Note that f does not have to be invertible in order for $f^{-1}(U)$ to make sense. Also note that images and inverse images do not quite invert each other, for instance we have

$$f^{-1}(f(\{-1, 0, 1, 2\})) \neq \{-1, 0, 1, 2\}$$

(why?).

Remark 3.4.8 If f is a bijective function, then we have defined f^{-1} in two slightly different ways, but this is not an issue because the two definitions agree in this case (Exercise 3.4.1).

As remarked earlier, functions are not necessarily sets. However, we do consider functions to be a type of object, and in particular we should be able to consider sets of functions. In particular, we should be able to consider the set of *all* functions from a set X to a set Y. To do this we need to introduce another axiom to set theory:

Axiom 3.11 (*Power set axiom*). Let X and Y be sets. Then there exists a set, denoted Y^X, which consists of all the functions from X to Y, thus

$$f \in Y^X \iff (f \text{ is a function with domain } X \text{ and codomain } Y).$$

Example 3.4.9 Let $X = \{4, 7\}$ and $Y = \{0, 1\}$. Then the set Y^X consists of four functions: the function that maps $4 \mapsto 0$ and $7 \mapsto 0$; the function that maps $4 \mapsto 0$ and $7 \mapsto 1$; the function that maps $4 \mapsto 1$ and $7 \mapsto 0$; and the function that maps $4 \mapsto 1$ and $7 \mapsto 1$. The reason we use the notation Y^X to denote this set is that if Y has n elements and X has m elements, then one can show that Y^X has n^m elements; see Proposition 3.6.14(f).

One consequence of this axiom is

Lemma 3.4.10 *Let X be a set. Then the set*

$$\{Y : Y \textit{ is a subset of } X\}$$

is a set. That is to say, there exists a set Z such that

$$Y \in Z \iff Y \subseteq X$$

for all objects Y.

Proof See Exercise 3.4.6. □

Remark 3.4.11 The set $\{Y : Y$ is a subset of $X\}$ is known as the *power set* of X and is denoted 2^X. For instance, if a, b, c are distinct objects, we have

$$2^{\{a,b,c\}} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Note that while $\{a, b, c\}$ has 3 elements, $2^{\{a,b,c\}}$ has $2^3 = 8$ elements. This gives a hint as to why we refer to the power set of X as 2^X; we return to this issue in Chap. 8.

For sake of completeness, let us now add one further axiom to our set theory, in which we enhance the axiom of pairwise union to allow unions of much larger collections of sets.

Axiom 3.12 (*Union*). Let A be a set, all of whose elements are themselves sets. Then there exists a set $\bigcup A$ whose elements are precisely those objects which are elements of the elements of A, thus for all objects x

$$x \in \bigcup A \iff (x \in S \text{ for some } S \in A).$$

Example 3.4.12 If $A = \{\{2, 3\}, \{3, 4\}, \{4, 5\}\}$, then $\bigcup A = \{2, 3, 4, 5\}$ (why?).

The axiom of union, combined with the axiom of pair set, implies the axiom of pairwise union (Exercise 3.4.8). Another important consequence of this axiom is that if one has some set I, and for every element $\alpha \in I$ we have some set A_α, then we can form the union set $\bigcup_{\alpha \in I} A_\alpha$ by defining

$$\bigcup_{\alpha \in I} A_\alpha := \bigcup\{A_\alpha : \alpha \in I\},$$

which is a set thanks to the axiom of replacement and the axiom of union. Thus for instance, if $I = \{1, 2, 3\}$, $A_1 := \{2, 3\}$, $A_2 := \{3, 4\}$, and $A_3 := \{4, 5\}$, then $\bigcup_{\alpha \in \{1,2,3\}} A_\alpha = \{2, 3, 4, 5\}$. More generally, we see that for any object y,

$$y \in \bigcup_{\alpha \in I} A_\alpha \iff (y \in A_\alpha \text{ for some } \alpha \in I). \tag{3.2}$$

In situations like this, we often refer to I as an *index set*, and the elements α of this index set as *labels*; the sets A_α are then called a *family of sets* and are *indexed* by the labels $\alpha \in I$. Note that if I was empty, then $\bigcup_{\alpha \in I} A_\alpha$ would automatically also be empty (why?).

We can similarly form intersections of families of sets, as long as the index set is non-empty. More specifically, given any non-empty set I, and given an assignment of a set A_α to each $\alpha \in I$, we can define the intersection $\bigcap_{\alpha \in I} A_\alpha$ by first choosing some element β of I (which we can do since I is non-empty), and setting

$$\bigcap_{\alpha \in I} A_\alpha := \{x \in A_\beta : x \in A_\alpha \text{ for all } \alpha \in I\}, \tag{3.3}$$

which is a set by the axiom of specification. This definition may look like it depends on the choice of β, but it does not (Exercise 3.4.9). Observe that for any object y,

$$y \in \bigcap_{\alpha \in I} A_\alpha \iff (y \in A_\alpha \text{ for all } \alpha \in I) \tag{3.4}$$

(compare with (3.2)).

Remark 3.4.13 The axioms of set theory that we have introduced (Axioms 3.1 and 3.12, excluding the dangerous Axiom 3.9) are known as the *Zermelo–Fraenkel axioms of set theory*,[5] after Ernest Zermelo (1871–1953) and Abraham Fraenkel (1891–1965). There is one further axiom we will eventually need, the famous *axiom of choice* (see Sect. 8.4), giving rise to the *Zermelo–Fraenkel–Choice* (*ZFC*) *axioms of set theory*, but we will not need this axiom for some time.

— Exercises —

Exercise 3.4.1 Let $f: X \to Y$ be a bijective function, and let $f^{-1}: Y \to X$ be its inverse. Let V be any subset of Y. Prove that the forward image of V under f^{-1} is the same set as the inverse image of V under f; thus the fact that both sets are denoted by $f^{-1}(V)$ will not lead to any inconsistency.

Exercise 3.4.2 Let $f: X \to Y$ be a function from one set X to another set Y, let S be a subset of X, and let U be a subset of Y.

(i) What, in general, can one say about $f^{-1}(f(S))$ and S?
(ii) What about $f(f^{-1}(U))$ and U?
(iii) What about $f^{-1}(f(f^{-1}(U)))$ and $f^{-1}(U)$?

Exercise 3.4.3 Let A, B be two subsets of a set X, and let $f: X \to Y$ be a function. Show that $f(A \cap B) \subseteq f(A) \cap f(B)$, that $f(A) \backslash f(B) \subseteq f(A \backslash B)$, $f(A \cup B) = f(A) \cup f(B)$. For the first two statements, is it true that the $\subseteq$ relation can be improved to $=$?

Exercise 3.4.4 Let $f: X \to Y$ be a function from one set X to another set Y, and let U, V be subsets of Y. Show that $f^{-1}(U \cup V) = f^{-1}(U) \cup f^{-1}(V)$, that $f^{-1}(U \cap V) = f^{-1}(U) \cap f^{-1}(V)$, and that $f^{-1}(U \backslash V) = f^{-1}(U) \backslash f^{-1}(V)$.

Exercise 3.4.5 Let $f: X \to Y$ be a function from one set X to another set Y. Show that $f(f^{-1}(S)) = S$ for every $S \subseteq Y$ if and only if f is surjective. Show that $f^{-1}(f(S)) = S$ for every $S \subseteq X$ if and only if f is injective.

Exercise 3.4.6

(i) Prove Lemma 3.4.10. (*Hint:* start with the set $\{0, 1\}^X$ and apply the replacement axiom, replacing each function f with the object $f^{-1}(\{1\})$.) See also Exercise 3.5.11.
(ii) Conversely, show that Axiom 3.11 can be deduced the preceding axioms of set theory if one accepts Lemma 3.4.10 as an axiom. (This may help explain why we refer to Axiom 3.11 as the "power set axiom".)

Exercise 3.4.7 Let X, Y be sets. Define a *partial function* from X to Y to be any function $f: X' \to Y'$ whose domain X' is a subset of X, and whose codomain Y' is a subset of Y. Show that the collection of all partial functions from X to Y is itself a set. (*Hint:* use Exercise 3.4.6, the power set axiom, the replacement axiom, and the union axiom.)

Exercise 3.4.8 Show that Axiom 3.5 can be deduced from Axiom 3.1, Axiom 3.4, and Axiom 3.12.

Exercise 3.4.9 Show that if β and β' are two elements of a set I, and to each $\alpha \in I$ we assign a set A_α, then

$$\{x \in A_\beta : x \in A_\alpha \text{ for all } \alpha \in I\} = \{x \in A_{\beta'} : x \in A_\alpha \text{ for all } \alpha \in I\},$$

and so the definition of $\bigcap_{\alpha \in I} A_\alpha$ defined in (3.3) does not depend on β. Also explain why (3.4) is true.

[5] These axioms are formulated slightly differently in other texts, but all the formulations can be shown to be equivalent to each other.

Exercise 3.4.10 Suppose that I and J are two sets, and for all $\alpha \in I \cup J$ let A_α be a set. Show that $(\bigcup_{\alpha\in I} A_\alpha) \cup (\bigcup_{\alpha\in J} A_\alpha) = \bigcup_{\alpha\in I\cup J} A_\alpha$. If I and J are non-empty, show that $(\bigcap_{\alpha\in I} A_\alpha) \cap (\bigcap_{\alpha\in J} A_\alpha) = \bigcap_{\alpha\in I\cup J} A_\alpha$.

Exercise 3.4.11 Let X be a set, let I be a non-empty set, and for all $\alpha \in I$ let A_α be a subset of X. Show that

$$X \backslash \bigcup_{\alpha\in I} A_\alpha = \bigcap_{\alpha\in I} (X \backslash A_\alpha)$$

and

$$X \backslash \bigcap_{\alpha\in I} A_\alpha = \bigcup_{\alpha\in I} (X \backslash A_\alpha).$$

This should be compared with De Morgan's laws in Proposition 3.1.27 (although one cannot derive the above identities directly from De Morgan's laws, as I could be infinite).

3.5 Cartesian Products

In addition to the basic operations of union, intersection, and differencing, another fundamental operation on sets is that of the *Cartesian product*. To define this notion, we first need the concept of an *ordered pair*.

Definition 3.5.1 (*Ordered pair*). If x and y are any objects (possibly equal), we define the *ordered pair* (x, y) to be a new object, consisting of x as its first component and y as its second component. Two ordered pairs (x, y) and (x', y') are considered equal if and only if both their components match, i.e.,

$$(x, y) = (x', y') \iff (x = x' \text{ and } y = y'). \tag{3.5}$$

This notion of equality is consistent with the usual axioms of equality (Exercise 3.5.3). Thus for instance, the pair $(3, 5)$ is equal to the pair $(2 + 1, 3 + 2)$, but is distinct from the pairs $(5, 3)$, $(3, 3)$, and $(2, 5)$. (This is in contrast to sets, where $\{3, 5\}$ and $\{5, 3\}$ are equal.)

Remark 3.5.2 Strictly speaking, this definition is partly an axiom, because we have simply postulated that given any two objects x and y, that an object of the form (x, y) exists. However, it is possible to define an ordered pair using the axioms of set theory in such a way that we do not need any further postulates (see Exercise 3.5.1).

Remark 3.5.3 We have now "overloaded" the parenthesis symbols () once again; they now are not only used to denote grouping of operators and arguments of functions, but also to enclose ordered pairs. This is usually not a problem in practice as one can still determine what usage the symbols () were intended for from context.

Definition 3.5.4 (*Cartesian product*). If X and Y are sets, then we define the *Cartesian product* $X \times Y$ to be the collection of ordered pairs, whose first component lies in X and second component lies in Y, thus

$$X \times Y = \{(x, y) : x \in X, y \in Y\}$$

or equivalently

$$a \in (X \times Y) \iff (a = (x, y) \text{ for some } x \in X \text{ and } y \in Y).$$

One can show that the Cartesian product $X \times Y$ is in fact a set; see Exercise 3.5.1.

Example 3.5.5 If $X := \{1, 2\}$ and $Y := \{3, 4, 5\}$, then

$$X \times Y = \{(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5)\}$$

and

$$Y \times X = \{(3, 1), (4, 1), (5, 1), (3, 2), (4, 2), (5, 2)\}.$$

Thus, strictly speaking, $X \times Y$ and $Y \times X$ are different sets, although they are very similar. For instance, they always have the same number of elements (Exercise 3.6.5).

Let $f : X \times Y \to Z$ be a function whose domain $X \times Y$ is a Cartesian product of two other sets X and Y. Then f can either be thought of as a function of one variable, mapping the single input of an ordered pair (x, y) in $X \times Y$ to an output[6] $f(x, y)$ in Z, or as a function of two variables, mapping an input $x \in X$ and another input $y \in Y$ to a single output $f(x, y)$ in Z. While the two notions are technically different, we will not bother to distinguish the two, and think of f simultaneously as a function of one variable with domain $X \times Y$ and as a function of two variables with domains X and Y. Thus for instance the addition operation $+$ on the natural numbers can now be re-interpreted as a function $+ : \mathbf{N} \times \mathbf{N} \to \mathbf{N}$, defined by $(x, y) \mapsto x + y$.

Once one has the notion of an ordered pair, one can also define an ordered triple (x, y, z) of three objects (x, y, z) by the formula $(x, y, z) := ((x, y), z)$. One could continue in this fashion and define ordered quadruples, etc., but we shall instead use a different construction to build ordered n-tuples:

Definition 3.5.6 (*Ordered n-tuple and n-fold Cartesian product*). Let n be a natural number. An *ordered n-tuple* $(x_i)_{1 \leq i \leq n}$ (also denoted $(x_1, \ldots, x_n)$) is a collection of objects x_i, one for every natural number i between 1 and n; we refer to x_i as the i^{th} *component* of the ordered n-tuple. Two ordered n-tuples $(x_i)_{1 \leq i \leq n}$ and $(y_i)_{1 \leq i \leq n}$ are said to be equal iff $x_i = y_i$ for all $1 \leq i \leq n$. If $(X_i)_{1 \leq i \leq n}$ is an ordered n-tuple of sets, we define their *Cartesian product* $\prod_{1 \leq i \leq n} X_i$ (also denoted $\prod_{i=1}^{n} X_i$ or $X_1 \times \ldots \times X_n$) by

$$\prod_{1 \leq i \leq n} X_i := \{(x_i)_{1 \leq i \leq n} : x_i \in X_i \text{ for all } 1 \leq i \leq n\}.$$

[6] Here (and in the rest of this text) we adopt the very common practice of abbreviating $f((x, y))$ as $f(x, y)$.

Again, this definition simply postulates that an ordered n-tuple and a Cartesian product always exist when needed, but using the axioms of set theory one can explicitly construct these objects; see Exercise 3.5.2.

Remark 3.5.7 One can generalize this construction to infinite Cartesian products; see Definition 8.4.1.

Example 3.5.8 Let $a_1, b_1, a_2, b_2, a_3, b_3$ be objects, and let $X_1 := \{a_1, b_1\}$, $X_2 := \{a_2, b_2\}$, and $X_3 := \{a_3, b_3\}$. Then we have

$$\begin{aligned}
X_1 \times X_2 \times X_3 = \; & \{(a_1, a_2, a_3), (a_1, a_2, b_3), (a_1, b_2, a_3), (a_1, b_2, b_3), \\
& (b_1, a_2, a_3), (b_1, a_2, b_3), (b_1, b_2, a_3), (b_1, b_2, b_3)\} \\
(X_1 \times X_2) \times X_3 = & \{((a_1, a_2), a_3), ((a_1, a_2), b_3), ((a_1, b_2), a_3), ((a_1, b_2), b_3), \\
& ((b_1, a_2), a_3), ((b_1, a_2), b_3), ((b_1, b_2), a_3), ((b_1, b_2), b_3)\} \\
X_1 \times (X_2 \times X_3) = & \{(a_1, (a_2, a_3)), (a_1, (a_2, b_3)), (a_1, (b_2, a_3)), (a_1, (b_2, b_3)), \\
& (b_1, (a_2, a_3)), (b_1, (a_2, b_3)), (b_1, (b_2, a_3)), (b_1, (b_2, b_3))\}.
\end{aligned}$$

Thus, strictly speaking, the sets $X_1 \times X_2 \times X_3$, $(X_1 \times X_2) \times X_3$, and $X_1 \times (X_2 \times X_3)$ are distinct. However, they are clearly very related to each other (for instance, there are obvious bijections between any two of the three sets), and it is common in practice to neglect the minor distinctions between these sets and pretend that they are in fact equal. Thus a function $f : X_1 \times X_2 \times X_3 \to Y$ can be thought of as a function of one variable $(x_1, x_2, x_3) \in X_1 \times X_2 \times X_3$, or as a function of three variables $x_1 \in X_1$, $x_2 \in X_2$, $x_3 \in X_3$, or as a function of two variables $x_1 \in X_1$, $(x_2, x_3) \in X_2 \times X_3$, and so forth; we will not bother to distinguish between these different perspectives.

Remark 3.5.9 An ordered n-tuple $(x_1, \ldots, x_n)$ of objects is also called an *ordered sequence* of n elements, or a *finite sequence* for short. In Chap. 5 we shall also introduce the very useful concept of an *infinite sequence*.

Example 3.5.10 If x is an object, then (x) is a 1-tuple, which we shall identify with x itself (even though the two are, strictly speaking, not the same object). Then if X_1 is any set, then the Cartesian product $\prod_{1 \le i \le 1} X_i$ is just X_1 (why?). Also, the *empty Cartesian product* $\prod_{1 \le i \le 0} X_i$ gives, not the empty set $\{\}$, but rather the singleton set $\{()\}$ whose only element is the 0-*tuple* $()$, also known as the *empty tuple*.

If n is a natural number, we often write X^n as shorthand for the n-fold Cartesian product $X^n := \prod_{1 \le i \le n} X$. Thus X^1 is essentially the same set as X (if we ignore the distinction between an object x and the 1-tuple (x)), while X^2 is essentially the Cartesian product $X \times X$. The set X^0 is a singleton set $\{()\}$ (why?).

We can now generalize the single choice lemma (Lemma 3.1.5) to allow for multiple (but finite) number of choices.

Lemma 3.5.11 (Finite choice). *Let $n \ge 1$ be a natural number, and for each natural number $1 \le i \le n$, let X_i be a non-empty set. Then there exists an n-tuple $(x_i)_{1 \le i \le n}$*

such that $x_i \in X_i$ for all $1 \leq i \leq n$. In other words, if each X_i is non-empty, then the set $\prod_{1 \leq i \leq n} X_i$ is also non-empty.

Proof We induct on n (starting with the base case $n = 1$; the claim is also vacuously true with $n = 0$ but is not particularly interesting in that case). When $n = 1$ the claim follows from Lemma 3.1.5 (why?). Now suppose inductively that the claim has already been proven for some n; we will now prove it for $n{+}{+}$. Let $X_1, \ldots, X_{n++}$ be a collection of non-empty sets. By induction hypothesis, we can find an n-tuple $(x_i)_{1 \leq i \leq n}$ such that $x_i \in X_i$ for all $1 \leq i \leq n$. Also, since X_{n++} is non-empty, by Lemma 3.1.5 we may find an object a such that $a \in X_{n++}$. If we thus define the $n{+}{+}$-tuple $(y_i)_{1 \leq i \leq n++}$ by setting $y_i := x_i$ when $1 \leq i \leq n$ and $y_i := a$ when $i = n{+}{+}$ it is clear that $y_i \in X_i$ for all $1 \leq i \leq n{+}{+}$, thus closing the induction. □

Remark 3.5.12 It is intuitively plausible that this lemma should be extended to allow for an infinite number of choices, but this cannot be done automatically; it requires an additional axiom, the *axiom of choice*. See Section 8.4.

— Exercises —

Exercise 3.5.1 (i) Suppose we *define* the ordered pair (x, y) for any objects x and y by the formula $(x, y) := \{\{x\}, \{x, y\}\}$ (thus using several applications of Axiom 3.4). Thus for instance $(1, 2)$ is the set $\{\{1\}, \{1, 2\}\}$, $(2, 1)$ is the set $\{\{2\}, \{2, 1\}\}$, and $(1, 1)$ is the set $\{\{1\}\}$. Show that such a definition (known as the *Kuratowski definition* of an ordered pair) indeed obeys the property (3.5).

(ii) Suppose we instead define an ordered pair using the alternate definition $(x, y) := \{x, \{x, y\}\}$. Show that this definition (known as the *short definition* of an ordered pair) also verifies (3.5) and is thus also an acceptable definition of ordered pair. (Warning: this is tricky; one needs the axiom of regularity, and in particular Exercise 3.2.2.)

(iii) Show that regardless of the definition of ordered pair, the Cartesian product $X \times Y$ of any two sets X, Y is again a set. (*Hint:* first use the axiom of replacement to show that for any $x \in X$, that $\{(x, y) : y \in Y\}$ is a set, and then apply the axiom of union.)

Exercise 3.5.2 Suppose we *define*[7] an ordered n-tuple to be a surjective function $x : \{i \in \mathbf{N} : 1 \leq i \leq n\} \to X$ whose codomain is some arbitrary set X (so different ordered n-tuples are allowed to have different ranges); we then write x_i for $x(i)$ and also write x as $(x_i)_{1 \leq i \leq n}$. Using this definition, verify that we have $(x_i)_{1 \leq i \leq n} = (y_i)_{1 \leq i \leq n}$ if and only if $x_i = y_i$ for all $1 \leq i \leq n$. Also, show that if $(X_i)_{1 \leq i \leq n}$ are an ordered n-tuple of sets, then the Cartesian product, as defined in Definition 3.5.6, is indeed a set. (*Hint*: use Exercise 3.4.7 and the axiom of specification.)

Exercise 3.5.3 Show that the definitions of equality for ordered pair and ordered n-tuple are consistent with the reflexivity, symmetry, and transitivity axioms, in the sense that if these axioms are assumed to hold for the individual components x, y of an ordered pair (x, y), then they hold for the ordered pair itself.

Exercise 3.5.4 Let A, B, C be sets. Show that $A \times (B \cup C) = (A \times B) \cup (A \times C)$, that $A \times (B \cap C) = (A \times B) \cap (A \times C)$, and that $A \times (B \backslash C) = (A \times B) \backslash (A \times C)$. (One can of course prove similar identities in which the rôles of the left and right factors of the Cartesian product are reversed.)

[7] Technically, this construction of ordered n-tuple is not compatible with the constructions of ordered pairs in Exercise 3.5.1, but this does not cause a difficulty in practice; for instance, one can use the definition of an ordered 2-tuple here to replace the construction in Exercise 3.5.1, or one can make a rather pedantic distinction between an ordered 2-tuple and an ordered pair in one's mathematical arguments.

Exercise 3.5.5 Let A, B, C, D be sets. Show that $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$. Is it true that $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$? Is it true that $(A \times B) \backslash (C \times D) = (A \backslash C) \times (B \backslash D)$?

Exercise 3.5.6 Let A, B, C, D be non-empty sets. Show that $A \times B \subseteq C \times D$ if and only if $A \subseteq C$ and $B \subseteq D$, and that $A \times B = C \times D$ if and only if $A = C$ and $B = D$. What happens if some or all of the hypotheses that the A, B, C, D are non-empty are removed?

Exercise 3.5.7 Let X, Y be sets, and let $\pi_{X \times Y \to X} \colon X \times Y \to X$ and $\pi_{X \times Y \to Y} \colon X \times Y \to Y$ be the maps $\pi_{X \times Y \to X}(x, y) := x$ and $\pi_{X \times Y \to Y}(x, y) := y$; these maps are known as the *co-ordinate functions* on $X \times Y$. Show that for any functions $f \colon Z \to X$ and $g \colon Z \to Y$, there exists a unique function $h \colon Z \to X \times Y$ such that $\pi_{X \times Y \to X} \circ h = f$ and $\pi_{X \times Y \to Y} \circ h = g$. (Compare this to the last part of Exercise 3.3.8, and to Exercise 3.1.7.) This function h is known as the *pairing* of f and g and is denoted $h = (f, g)$.

Exercise 3.5.8 Let $X_1, \ldots, X_n$ be sets. Show that the Cartesian product $\prod_{i=1}^{n} X_i$ is empty if and only if at least one of the X_i is empty.

Exercise 3.5.9 Suppose that I and J are two sets, and for all $\alpha \in I$ let A_α be a set, and for all $\beta \in J$ let B_β be a set. Show that $(\bigcup_{\alpha \in I} A_\alpha) \cap (\bigcup_{\beta \in J} B_\beta) = \bigcup_{(\alpha,\beta) \in I \times J} (A_\alpha \cap B_\beta)$. What happens if one interchanges all the union and intersection symbols here?

Exercise 3.5.10 If $f \colon X \to Y$ is a function, define the *graph* of f to be the subset of $X \times Y$ defined by $\{(x, f(x)) : x \in X\}$.

(i) Show that two functions $f \colon X \to Y$, $\tilde{f} \colon X \to Y$ are equal if and only if they have the same graph.
(ii) Conversely, if G is any subset of $X \times Y$ with the property that for each $x \in X$, the set $\{y \in Y : (x, y) \in G\}$ has exactly one element (or in other words, G obeys the *vertical line test*), show that there is exactly one function $f \colon X \to Y$ whose graph is equal to G.
(iii) Suppose we *define*[8] a function f to be an ordered triple $f = (X, Y, G)$, where X, Y are sets, and G is a subset of $X \times Y$ that obeys the vertical line test. We then define the domain of such a triple to be X, the codomain to be Y and for every $x \in X$, we *define* $f(x)$ to be the unique $y \in Y$ such that $(x, y) \in G$. Show that this definition is compatible with Definition 3.3.1 in the sense that every choice of domain X, codomain Y, and property $P(x, y)$ obeying the vertical line test produces a function as defined here that obeys all the properties required of it in that definition, and is also similarly compatible with Definition 3.3.8.

Exercise 3.5.11 Show that Axiom 3.11 can in fact be deduced from Lemma 3.4.10 and the other axioms of set theory, and thus Lemma 3.4.10 can be used as an alternate formulation of the power set axiom. (*Hint:* for any two sets X and Y, use Lemma 3.4.10 and the axiom of specification to construct the set of all subsets of $X \times Y$ which obey the vertical line test. Then use Exercise 3.5.10 and the axiom of replacement.)

Exercise 3.5.12 This exercise will establish a rigorous version of Proposition 2.1.16 that avoids circularity (in particular, avoiding the use of any object that required Proposition 2.1.16 to construct).

(i) Let X be a set, let $f \colon \mathbf{N} \times X \to X$ be a function, and let c be an element of X. Show that there exists a function $a \colon X \to X$ such that

$$a(0) = c$$

and

$$a(n{+}{+}) = f(n, a(n)) \text{ for all } n \in \mathbf{N},$$

[8] Note that this definition is not circular, because the notion of a function was not used to define ordered triples or a Cartesian product of two sets.

and furthermore that this function is unique. (*Hint:* first show inductively, by a modification of the proof of Lemma 3.5.11, that for every natural number $N \in \mathbf{N}$, there exists a unique function $a_N : \{n \in \mathbf{N} : n \leq N\} \to X$ such that $a_N(0) = c$ and $a_N(n{+}{+}) = f(n, a_N(n))$ for all $n \in \mathbf{N}$ such that $n < N$.)

(ii) (Warning: this is challenging.) Prove (i) without using any properties of the natural numbers other than the Peano axioms directly (in particular, without using the ordering of the natural numbers, and without appealing to Proposition 2.1.16). (*Hint:* first show inductively, using only the Peano axioms and basic set theory, that for every natural number $N \in \mathbf{N}$, there exists a unique pair A_N, B_N of subsets of $\mathbf{N}$ which obeys the following properties: (a) $A_N \cap B_N = \emptyset$, (b) $A_N \cup B_N = \mathbf{N}$, (c) $0 \in A_N$, (d) $N{+}{+} \in B_N$, (e) Whenever $n \in B_N$, we have $n{+}{+} \in B_N$. (f) Whenever $n \in A_N$ and $n \neq N$, we have $n{+}{+} \in A_N$. Once one obtains these sets, use A_N as a substitute for $\{n \in \mathbf{N} : n \leq N\}$ in the previous argument.)

Exercise 3.5.13 The purpose of this exercise is to show that there is essentially only one version of the natural number system in set theory (cf. the discussion in Remark 2.1.12). Suppose we have a set $\mathbf{N}'$ of "alternative natural numbers", an "alternative zero" $0'$, and an "alternative increment operation" which takes any alternative natural number $n' \in \mathbf{N}'$ and returns another alternative natural number $n'{+}{+}' \in \mathbf{N}'$, such that the Peano axioms (Axioms 2.1-2.5) all hold with the natural numbers, zero, and increment replaced by their alternative counterparts. Show that there exists a bijection $f : \mathbf{N} \to \mathbf{N}'$ from the natural numbers to the alternative natural numbers such that $f(0) = 0'$, and such that for any $n \in \mathbf{N}$ and $n' \in \mathbf{N}'$, we have $f(n) = n'$ if and only if $f(n{+}{+}) = n'{+}{+}'$. (*Hint:* use Exercise 3.5.12.)

3.6 Cardinality of Sets

In the previous chapter we defined the natural numbers axiomatically, assuming that they were equipped with a 0 and an increment operation, and assuming five axioms on these numbers. Philosophically, this is quite different from one of our main conceptualizations of natural numbers—that of *cardinality*, or measuring *how many* elements there are in a set. Indeed, the Peano axiom approach treats natural numbers more like *ordinals* than *cardinals*. (The cardinals are One, Two, Three, ..., and are used to count how many things there are in a set. The *ordinals* are First, Second, Third, ..., and are used to order a sequence of objects. There is a subtle difference between the two, especially when comparing infinite cardinals with infinite ordinals, but this is beyond the scope of this text.) We paid a lot of attention to what number came *next* after a given number *n*—which is an operation which is quite natural for ordinals, but less so for cardinals—but did not address the issue of whether these numbers could be used to *count* sets. The purpose of this section is to address this issue by noting that the natural numbers *can* be used to count the cardinality of sets, as long as the set is finite.

The first thing is to work out when two sets have the same size. For instance, it seems clear that the sets $\{1, 2, 3\}$ and $\{4, 5, 6\}$ have the same size, but that both have a different size from $\{8, 9\}$. As an initial attempt to define a notion of size, we could try to say that two sets have the same size if they have the same number of elements, but we have not yet defined what the "number of elements" in a set is. Besides, this runs into problems when a set is infinite.

The right way to define the concept of "two sets having the same size" is not immediately obvious, but can be worked out with some thought. One intuitive reason why the sets $\{1, 2, 3\}$ and $\{4, 5, 6\}$ have the same size is that one can match the elements of the first set with the elements in the second set in a one-to-one correspondence: $1 \leftrightarrow 4$, $2 \leftrightarrow 5$, $3 \leftrightarrow 6$. (Indeed, this is how we first learn to count a set: we correspond the set we are trying to count with another set, such as a set of fingers on your hand.) We will use this intuitive understanding as our rigorous basis for "having the same size".

Definition 3.6.1 (*Equal cardinality*) We say that two sets X and Y have *equal cardinality* iff there exists a bijection $f: X \to Y$ from X to Y.

Example 3.6.2 The sets $\{0, 1, 2\}$ and $\{3, 4, 5\}$ have equal cardinality, since we can find a bijection between the two sets. Note that we do not yet know whether $\{0, 1, 2\}$ and $\{3, 4\}$ have equal cardinality; we know that one of the functions f from $\{0, 1, 2\}$ to $\{3, 4\}$ is not a bijection, but we have not proven yet that there might still be some other bijection from one set to the other. (It turns out that they do not have equal cardinality, but we will prove this a little later.) Note that this definition makes sense regardless of whether X is finite or infinite (in fact, we haven't even defined what finite means yet).

Remark 3.6.3 The fact that two sets have equal cardinality does not preclude one of the sets from containing the other. For instance, if X is the set of natural numbers and Y is the set of even[9] natural numbers, then the map $f: X \to Y$ defined by $f(n) := 2n$ is a bijection from X to Y (why?), and so X and Y have equal cardinality, despite Y being a subset of X and seeming intuitively as if it should only have "half" of the elements of X.

The notion of having equal cardinality is an equivalence relation:

Proposition 3.6.4 *Let X, Y, Z be sets. Then X has equal cardinality with X. If X has equal cardinality with Y, then Y has equal cardinality with X. If X has equal cardinality with Y and Y has equal cardinality with Z, then X has equal cardinality with Z.*

Proof See Exercise 3.6.1. □

Let n be a natural number. Now we want to say when a set X has n elements. Certainly we want the set $\{i \in \mathbf{N} : 1 \leq i \leq n\} = \{1, 2, \ldots, n\}$ to have n elements. (This is true even when $n = 0$; the set $\{i \in N : 1 \leq i \leq 0\}$ is just the empty set.) Using our notion of equal cardinality, we thus define:

Definition 3.6.5 Let n be a natural number. A set X is said to have *cardinality n*, iff it has equal cardinality with $\{i \in \mathbf{N} : 1 \leq i \leq n\}$. We also say that X *has n elements* iff it has cardinality n.

[9] A natural number is *even* if it is of the form $2n$ for some natural number n.

Remark 3.6.6 One can use the set $\{i \in \mathbf{N} : i < n\}$ instead of $\{i \in \mathbf{N} : 1 \leq i \leq n\}$, since these two sets clearly have equal cardinality. (Why? What is the bijection?)

Example 3.6.7 Let a, b, c, d be distinct objects. Then $\{a, b, c, d\}$ has the same cardinality as $\{i \in \mathbf{N} : i < 4\} = \{0, 1, 2, 3\}$ or $\{i \in \mathbf{N} : 1 \leq i \leq 4\} = \{1, 2, 3, 4\}$ and thus has cardinality 4. Similarly, the set $\{a\}$ has cardinality 1.

There might be one problem with this definition: a set might have two different cardinalities. But this is not possible:

Proposition 3.6.8 (Uniqueness of cardinality) *Let X be a set with some cardinality n. Then X cannot have any other cardinality, i.e., X cannot have cardinality m for any $m \neq n$.*

Before we prove this proposition, we need a lemma.

Lemma 3.6.9 *Suppose that $n \geq 1$, and X has cardinality n. Then X is non-empty, and if x is any element of X, then the set $X - \{x\}$ (i.e., X with the element x removed) has cardinality*[10] *$n - 1$.*

Proof If X is empty then it clearly cannot have the same cardinality as the non-empty set $\{i \in \mathbf{N} : 1 \leq i \leq n\}$, as there is no bijection from the empty set to a non-empty set (why?). Now let x be an element of X. Since X has the same cardinality as $\{i \in \mathbf{N} : 1 \leq i \leq n\}$, we thus have a bijection f from X to $\{i \in \mathbf{N} : 1 \leq i \leq n\}$. In particular, $f(x)$ is a natural number between 1 and n. Now define the function $g : X - \{x\} \to \{i \in \mathbf{N} : 1 \leq i \leq n - 1\}$ by the following rule: for any $y \in X - \{x\}$, we define $g(y) := f(y)$ if $f(y) < f(x)$, and define $g(y) := f(y) - 1$ if $f(y) > f(x)$. (Note that $f(y)$ cannot equal $f(x)$ since $y \neq x$ and f is a bijection.) It is easy to check that this map is also a bijection (why?), and so $X - \{x\}$ has equal cardinality with $\{i \in \mathbf{N} : 1 \leq i \leq n - 1\}$. In particular $X - \{x\}$ has cardinality $n - 1$, as desired. □

Now we prove the proposition.

Proof of Proposition 3.6.8 We induct on n. First suppose that $n = 0$. Then X must be empty, and so X cannot have any non-zero cardinality. Now suppose that the proposition is already proven for some n; we now prove it for $n{+}{+}$. Let X have cardinality $n{+}{+}$; and suppose that X also has some other cardinality $m \neq n{+}{+}$. By Lemma 3.6.9, X is non-empty, and if x is any element of X, then $X - \{x\}$ has cardinality n and also has cardinality $m - 1$, by Lemma 3.6.9. By induction hypothesis, this means that $n = m - 1$, which implies that $m = n{+}{+}$, a contradiction. This closes the induction. □

[10] Strictly speaking, $n - 1$ has not yet been defined in this text. For the purposes of this lemma, we define $n - 1$ to be the unique natural number m such that $m{+}{+} = n$; this m is given by Lemma 2.2.10.

Thus, for instance, we now know, thanks to Propositions 3.6.4 and 3.6.8, that the sets $\{0, 1, 2\}$ and $\{3, 4\}$ do not have equal cardinality, since the first set has cardinality 3 and the second set has cardinality 2.

Definition 3.6.10 (*Finite sets*). A set is *finite* iff it has cardinality n for some natural number n; otherwise, the set is called *infinite*. If X is a finite set, we use $\#(X)$ to denote the cardinality of X.

Example 3.6.11 The sets $\{0, 1, 2\}$ and $\{3, 4\}$ are finite, as is the empty set (0 is a natural number), and $\#(\{0, 1, 2\}) = 3$, $\#(\{3, 4\}) = 2$, and $\#(\emptyset) = 0$.

Now we give an example of an infinite set.

Theorem 3.6.12 *The set of natural numbers* $\mathbf{N}$ *is infinite.*

Proof Suppose for sake of contradiction that the set of natural numbers $\mathbf{N}$ was finite, so it had some cardinality $\#(\mathbf{N}) = n$. By Lemma 3.6.9, $\mathbf{N}\backslash\{0\}$ would then have cardinality $n - 1$. But $\mathbf{N}$ has equal cardinality with $\mathbf{N}\backslash\{0\}$ (using $x \mapsto x + 1$ as the bijection from the latter to the former), hence $n = n - 1$, which gives the desired contradiction. □

Remark 3.6.13 One can also use similar arguments to show that any unbounded set[11] is infinite; for instance the rationals $\mathbf{Q}$ and the reals $\mathbf{R}$ (which we will construct in later chapters) are infinite. However, it is possible for some sets to be "more" infinite than others; see Sect. 8.3.

Now we relate cardinality with the arithmetic of natural numbers.

Proposition 3.6.14 (Cardinal arithmetic).

(*a*) *Let X be a finite set, and let x be an object which is not an element of X. Then $X \cup \{x\}$ is finite and $\#(X \cup \{x\}) = \#(X) + 1$.*

(*b*) *Let X and Y be finite sets. Then $X \cup Y$ is finite and $\#(X \cup Y) \leq \#(X) + \#(Y)$. If in addition X and Y are disjoint (i.e., $X \cap Y = \emptyset$), then $\#(X \cup Y) = \#(X) + \#(Y)$.*

(*c*) *Let X be a finite set, and let Y be a subset of X. Then Y is finite, and $\#(Y) \leq \#(X)$. If in addition $Y \neq X$ (i.e., Y is a proper subset of X), then we have $\#(Y) < \#(X)$.*

(*d*) *If X is a finite set, and $f : X \to Y$ is a function, then $f(X)$ is a finite set with $\#(f(X)) \leq \#(X)$. One has equality $\#(f(X)) = \#(X)$ if and only if f is one-to-one.*

(*e*) *Let X and Y be finite sets. Then Cartesian product $X \times Y$ is finite and $\#(X \times Y) = \#(X) \times \#(Y)$.*

(*f*) *Let X and Y be finite sets. Then the set Y^X (defined in Axiom 3.11) is finite and $\#(Y^X) = \#(Y)^{\#(X)}$.*

Proof See Exercise 3.6.4. □

[11] The notion of a bounded or unbounded set is defined in Definition 9.1.22.

Remark 3.6.15 Proposition 3.6.14 suggests that there is another way to define the arithmetic operations of natural numbers; not defined recursively as in Definitions 2.2.1, 2.3.1, 2.3.11, but instead using the notions of union, Cartesian product, and power set. This is the basis of *cardinal arithmetic*, which is an alternative foundation to arithmetic than the Peano arithmetic we have developed here; we will not develop this arithmetic in this text, but we give some examples of how one would work with this arithmetic in Exercises 3.6.5, 3.6.6.

This concludes our discussion of finite sets. We shall discuss infinite sets in Chap. 8, once we have constructed a few more examples of infinite sets (such as the integers, rationals, and reals).

— Exercises —

Exercise 3.6.1 Prove Proposition 3.6.4.

Exercise 3.6.2 Show that a set X has cardinality 0 if and only if X is the empty set.

Exercise 3.6.3 Let n be a natural number, and let $f\colon \{i \in \mathbf{N} : 1 \leq i \leq n\} \to \mathbf{N}$ be a function. Show that there exists a natural number M such that $f(i) \leq M$ for all $1 \leq i \leq n$. (*Hint:* induct on n. You may also want to peek at Lemma 5.1.14.) Thus finite subsets of the natural numbers are bounded. Use this to give an alternate proof of Theorem 3.6.12 that does not use Lemma 3.6.9.

Exercise 3.6.4 Prove Proposition 3.6.14.

Exercise 3.6.5 Let A and B be sets. Show that $A \times B$ and $B \times A$ have equal cardinality by constructing an explicit bijection between the two sets. Then use Proposition 3.6.14 to conclude an alternate proof of Lemma 2.3.2.

Exercise 3.6.6 Let A, B, C be sets. Show that the sets $(A^B)^C$ and $A^{B\times C}$ have equal cardinality by constructing an explicit bijection between the two sets. Conclude that $(a^b)^c = a^{bc}$ for any natural numbers a, b, c. Use a similar argument to also conclude $a^b \times a^c = a^{b+c}$.

Exercise 3.6.7 Let A and B be sets. Let us say that A has *lesser or equal* cardinality to B if there exists an injection $f\colon A \to B$ from A to B. Show that if A and B are finite sets, then A has lesser or equal cardinality to B if and only if $\#(A) \leq \#(B)$.

Exercise 3.6.8 Let A and B be sets such that there exists an injection $f\colon A \to B$ from A to B (i.e., A has lesser or equal cardinality to B). Assume also that A is non-empty. Show that there exists a surjection $g\colon B \to A$ from B to A. (The converse to this statement requires the axiom of choice; see Exercise 8.4.3.)

Exercise 3.6.9 Let A and B be finite sets. Show that $A \cup B$ and $A \cap B$ are also finite sets, and that $\#(A) + \#(B) = \#(A \cup B) + \#(A \cap B)$.

Exercise 3.6.10 Let $A_1, \ldots, A_n$ be finite sets such that $\#(\bigcup_{i\in\{1,\ldots,n\}} A_i) > n$. Show that there exists $i \in \{1, \ldots, n\}$ such that $\#(A_i) \geq 2$. (This is known as the *pigeonhole principle*.)

Exercise 3.6.11 Let $f\colon X \to Y$ be a function between two sets X, Y. Show that the following are equivalent:

(a) f is injective.
(b) Whenever $E \subseteq X$ has cardinality $\#(E)$ equal to 2, then the image $f(E)$ also has cardinality $\#(f(E)) = 2$.

(Note that if X has cardinality less than 2 then the claim in (b) is vacuously true; nevertheless, the equivalence still holds in this case!) Because of this equivalence, one could refer to an injective function as a *two-to-two function*. (This observation is due to John Conway (1937–2020).)

Exercise 3.6.12 For any natural number n, let S_n be the set of all bijections $\phi\colon \{i \in \mathbf{N} : 1 \leq i \leq n\} \to \{i \in \mathbf{N} : 1 \leq i \leq n\}$ from the set $\{i \in \mathbf{N} : 1 \leq i \leq n\}$ to itself (such bijections are also known as *permutations* of $\{i \in \mathbf{N} : 1 \leq i \leq n\}$.

(i) For any natural number n, show that S_n is finite, and $\#(S_{n{+}{+}}) = (n{+}{+}) \times \#(S_n)$. (*Hint:* partition $S_{n{+}{+}}$ into $n{+}{+}$ subsets, depending on the value $\phi(n{+}{+})$ a permutation $\phi : \{i \in \mathbf{N} : 1 \leq i \leq n{+}{+}\} \to \{i \in \mathbf{N} : 1 \leq i \leq n\}$ from the set $\{i \in \mathbf{N} : 1 \leq i \leq n{+}{+}\}$ assigns to $n{+}{+}$.

(ii) Define the *factorial* $n!$ of a natural number n recursively by $0! := 1$ and $(n{+}{+})! := (n{+}{+}) \times n!$ for all natural numbers n. Show that $\#(S_n) = n!$ for all natural numbers n.

Chapter 4
Integers and Rationals

4.1 The Integers

In Chap. 2 we built up most of the basic properties of the natural number system, but we have reached the limits of what one can do with just addition and multiplication. We would now like to introduce a new operation, that of subtraction, but to do that properly we will have to pass from the natural number system to a larger number system, that of the *integers*.

Informally, the integers are what you can get by subtracting two natural numbers; for instance, $3 - 5$ should be an integer, as should $6 - 2$. This is not a complete definition of the integers, because (a) it doesn't say when two differences are equal (for instance we should know why $3 - 5$ is equal to $2 - 4$, but is not equal to $1 - 6$), and (b) it doesn't say how to do arithmetic on these differences (how does one add $3 - 5$ to $6 - 2$?). Furthermore, (c) this definition is circular because it requires a notion of subtraction, which we can only adequately define once the integers are constructed. Fortunately, because of our prior experience with integers we know what the answers to these questions should be. To answer (a), we know from our advanced knowledge in algebra that $a - b = c - d$ happens exactly when $a + d = c + b$, so we can characterize equality of differences using only the concept of addition. Similarly, to answer (b) we know from algebra that $(a - b) + (c - d) = (a + c) - (b + d)$ and that $(a - b)(c - d) = (ac + bd) - (ad + bc)$. So we will take advantage of our foreknowledge by building all this into the *defn* of the integers, as we shall do shortly.

We still have to resolve (c). To get around this problem we will use the following work around: we will temporarily write integers not as a difference $a - b$, but instead use a new notation $a\text{——}b$ to define integers, where the —— is a meaningless placeholder, similar to the comma in the Cartesian co-ordinate notation (x, y) for points in the plane. Later when we define subtraction we will see that $a\text{——}b$ is in fact equal to $a - b$, and so we can discard the notation ——; it is only needed right now to avoid circularity. (These devices are similar to the scaffolding used to construct a building; they are temporarily essential to make sure the building is built correctly, but once

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_4

the building is completed they are thrown away and never used again.) This may seem unnecessarily complicated in order to define something that we already are very familiar with, but we will use this device again to construct the rationals, and knowing these kinds of constructions will be very helpful in later chapters.

Definition 4.1.1 (*Integers*). An *integer* is an expression[1] of the form $a—b$, where a and b are natural numbers. Two integers are considered to be equal, $a—b = c—d$, if and only if $a + d = c + b$. We let $\mathbf{Z}$ denote the set of all integers.

Thus for instance $3—5$ is an integer, and is equal to $2—4$, because $3 + 4 = 2 + 5$. On the other hand, $3—5$ is not equal to $2—3$ because $3 + 3 \neq 2 + 5$. This notation is strange looking and has a few deficiencies; for instance, 3 is not yet an integer, because it is not of the form $a—b$! We will rectify these problems later.

We have to check that this is a legitimate notion of equality. We need to verify the reflexivity, symmetry, transitivity, and substitution axioms (see Sect. A.7). We leave reflexivity and symmetry to Exercise 4.1.1 and instead verify the transitivity axiom. Suppose we know that $a—b = c—d$ and $c—d = e—f$. Then we have $a + d = c + b$ and $c + f = d + e$. Adding the two equations together we obtain $a + d + c + f = c + b + d + e$. By Proposition 2.2.6 we can cancel the c and d, obtaining $a + f = b + e$, i.e., $a—b = e—f$. Thus the cancellation law was needed to make sure that our notion of equality is sound. As for the substitution axiom, we cannot verify it at this stage because we have not yet defined any operations on the integers. However, when we do define our basic operations on the integers, such as addition, multiplication, and order, we will have to verify the substitution axiom at that time in order to ensure that the definition is valid. (We will only need to do this for the basic operations; more advanced operations on the integers, such as exponentiation, will be defined in terms of the basic ones, and so we do not need to reverify the substitution axiom for the advanced operations.)

Now we define two basic arithmetic operations on integers: addition and multiplication.

Definition 4.1.2 The sum of two integers, $(a—b) + (c—d)$, is defined by the formula

$$(a—b) + (c—d) := (a + c)—(b + d).$$

The product of two integers, $(a—b) \times (c—d)$, is defined by

$$(a—b) \times (c—d) := (ac + bd)—(ad + bc).$$

[1] In the language of set theory, what we are doing here is starting with the space $\mathbf{N} \times \mathbf{N}$ of ordered pairs (a, b) of natural numbers. Then we place an *equivalence relation* $\sim$ on these pairs by declaring $(a, b) \sim (c, d)$ iff $a + d = c + b$. The set-theoretic interpretation of the symbol $a—b$ is that it is the space of all pairs equivalent to (a, b): $a—b := \{(c, d) \in \mathbf{N} \times \mathbf{N} : (a, b) \sim (c, d)\}$; the existence of the set $\mathbf{Z} = \{a—b : (a, b) \in \mathbf{N} \times \mathbf{N}\}$ of integers then follows from two applications of the axiom of replacement. However, this interpretation plays no role in how we manipulate the integers and we will not refer to it again. A similar set-theoretic interpretation can be given to the construction of the rational numbers later in this chapter, or the real numbers in the next chapter.

Thus for instance, $(3—5) + (1—4)$ is equal to $(4—9)$. There is however one thing we have to check before we can accept these definitions—we have to check that if we replace one of the integers by an equal integer, that the sum or product does not change. For instance, $(3—5)$ is equal to $(2—4)$, so $(3—5) + (1—4)$ ought to have the same value as $(2—4) + (1—4)$, otherwise this would not give a consistent definition of addition. Fortunately, this is the case:

Lemma 4.1.3 (Addition and multiplication are well-defined). *Let a, b, a', b', c, d be natural numbers. If $(a—b) = (a'—b')$, then $(a—b) + (c—d) = (a'—b') + (c—d)$ and $(a—b) \times (c—d) = (a'—b') \times (c—d)$, and also $(c—d) + (a—b) = (c—d) + (a'—b')$ and $(c—d) \times (a—b) = (c—d) \times (a'—b')$. Thus addition and multiplication are well-defined operations* (*equal inputs give equal outputs*).

Proof To prove that $(a—b) + (c—d) = (a'—b') + (c—d)$, we evaluate both sides as $(a + c)—(b + d)$ and $(a' + c)—(b' + d)$. Thus we need to show that $a + c + b' + d = a' + c + b + d$. But since $(a—b) = (a'—b')$, we have $a + b' = a' + b$, and so by adding $c + d$ to both sides we obtain the claim. Now we show that $(a—b) \times (c—d) = (a'—b') \times (c—d)$. Both sides evaluate to $(ac + bd)—(ad + bc)$ and $(a'c + b'd)—(a'd + b'c)$, so we have to show that $ac + bd + a'd + b'c = a'c + b'd + ad + bc$. But the left-hand side factors as $c(a + b') + d(a' + b)$, while the right factors as $c(a' + b) + d(a + b')$. Since $a + b' = a' + b$, the two sides are equal. The other two identities are proven similarly. □

The integers $n—0$ behave in the same way as the natural numbers n; indeed one can check that $(n—0) + (m—0) = (n + m)—0$ and $(n—0) \times (m—0) = nm—0$. Furthermore, $(n—0)$ is equal to $(m—0)$ if and only if $n = m$. (The mathematical term for this is that there is an *isomorphism* between the natural numbers n and those integers of the form $n—0$.) Thus we may *identify* the natural numbers with integers by setting $n \equiv n—0$; this does not affect our definitions of addition or multiplication or equality since they are consistent with each other. For instance the natural number 3 is now considered to be the same as the integer $3—0$, thus $3 = 3—0$. In particular 0 is equal to $0—0$ and 1 is equal to $1—0$. Of course, if we set n equal to $n—0$, then it will also be equal to any other integer which is equal to $n—0$, for instance 3 is equal not only to $3—0$, but also to $4—1$, $5—2$, etc.

We can now define incrementation on the integers by defining $x{++} := x + 1$ for any integer x; this is of course consistent with our definition of the increment operation for natural numbers. However, this is no longer an important operation for us, as it has been now superceded by the more general notion of addition.

Now we consider some other basic operations on the integers.

Definition 4.1.4 (*Negation of integers*). If $(a—b)$ is an integer, we define the negation $-(a—b)$ to be the integer $(b—a)$. In particular if $n = n—0$ is a positive natural number, we can define its negation $-n = 0—n$.

For instance $-(3—5) = (5—3)$. One can check this definition is well-defined (Exercise 4.1.2).

We can now show that the integers correspond exactly to what we expect.

Lemma 4.1.5 (Trichotomy of integers). *Let x be an integer. Then exactly one of the following three statements is true: (a) x is zero; (b) x is equal to a positive natural number n; or (c) x is the negation $-n$ of a positive natural number n.*

Proof We first show that at least one of (a), (b), (c) is true. By definition, $x = a -\!\!- b$ for some natural numbers a, b. We have three cases: $a > b$, $a = b$, or $a < b$. If $a > b$ then $a = b + c$ for some positive natural number c, which means that $a -\!\!- b = c -\!\!- 0 = c$, which is (b). If $a = b$, then $a -\!\!- b = a -\!\!- a = 0 -\!\!- 0 = 0$, which is (a). If $a < b$, then $b > a$, so that $b -\!\!- a = n$ for some natural number n by the previous reasoning, and thus $a -\!\!- b = -n$, which is (c).

Now we show that no more than one of (a), (b), (c) can hold at a time. By definition, a positive natural number is non-zero, so (a) and (b) cannot simultaneously be true. If (a) and (c) were simultaneously true, then $0 = -n$ for some positive natural n; thus $(0 -\!\!- 0) = (0 -\!\!- n)$, so that $0 + n = 0 + 0$, so that $n = 0$, a contradiction. If (b) and (c) were simultaneously true, then $n = -m$ for some positive n, m, so that $(n -\!\!- 0) = (0 -\!\!- m)$, so that $n + m = 0 + 0$, which contradicts Proposition 2.2.8. Thus exactly one of (a), (b), (c) is true for any integer x. □

If n is a positive natural number, we call n a *positive integer*, and $-n$ a *negative integer*. Thus every integer is positive, zero, or negative, but not more than one of these at a time.

One could well ask why we don't use Lemma 4.1.5 to *define* the integers; i.e., why didn't we just say an integer is anything which is either a positive natural number, zero, or the negative of a natural number. The reason is that if we did so, the rules for adding and multiplying integers would split into many different cases (e.g., negative times positive equals positive; negative plus positive is either negative, positive, or zero, depending on which term is larger, etc.) and to verify all the properties would end up being much messier.

We now summarize the algebraic properties of the integers.

Proposition 4.1.6 (Laws of algebra for integers). *Let x, y, z be integers. Then we have*

$$\begin{aligned}
x + y &= y + x \\
(x + y) + z &= x + (y + z) \\
x + 0 = 0 + x &= x \\
x + (-x) = (-x) + x &= 0 \\
xy &= yx \\
(xy)z &= x(yz) \\
x1 = 1x &= x \\
x(y + z) &= xy + xz \\
(y + z)x &= yx + zx.
\end{aligned}$$

Remark 4.1.7 The above set of nine identities have a name; they are asserting that the integers form a *commutative ring*. (If one deleted the identity $xy = yx$, then they would only assert that the integers form a *ring*). Note that some of these identities were already proven for the natural numbers, but this does not automatically mean that they also hold for the integers because the integers are a larger set than the natural numbers. On the other hand, this proposition supercedes many of the propositions derived earlier for natural numbers.

Proof There are two ways to prove these identities. One is to use Lemma 4.1.5 and split into a lot of cases depending on whether x, y, z are zero, positive, or negative. This becomes very messy. A shorter way is to write $x = (a—b)$, $y = (c—d)$, and $z = (e—f)$ for some natural numbers a, b, c, d, e, f, and expand these identities in terms of a, b, c, d, e, f and use the algebra of the natural numbers. This allows each identity to be proven in a few lines. We shall just prove the longest one, namely $(xy)z = x(yz)$:

$$\begin{aligned}
(xy)z &= ((a—b)(c—d))\,(e—f)\\
&= ((ac+bd)—(ad+bc))\,(e—f)\\
&= ((ace+bde+adf+bcf)—(acf+bdf+ade+bce))\,;\\
x(yz) &= (a—b)\,((c—d)(e—f))\\
&= (a—b)\,((ce+df)—(cf+de))\\
&= ((ace+adf+bcf+bde)—(acf+ade+bce+bdf))
\end{aligned}$$

and so one can see that $(xy)z$ and $x(yz)$ are equal. The other identities are proven in a similar fashion; see Exercise 4.1.4. □

We now define the operation of *subtraction* $x - y$ of two integers by the formula

$$x - y := x + (-y).$$

We do not need to verify the substitution axiom for this operation, since we have defined subtraction in terms of two other operations on integers, namely addition and negation, and we have already verified that those operations are well-defined.

One can easily check now that if a and b are natural numbers, then

$$a - b = a + -b = (a—0) + (0—b) = a—b,$$

and so $a—b$ is just the same thing as $a - b$. Because of this we can now discard the — notation, and use the familiar.

We can now generalize Lemma 2.3.3 and Corollary 2.3.7 from the natural numbers to the integers:

Proposition 4.1.8 (Integers have no zero divisors) *Let a and b be integers such that $ab = 0$. Then either $a = 0$ or $b = 0$ (or both).*

Proof See Exercise 4.1.5. □

Corollary 4.1.9 (Cancellation law for integers) *If a, b, c are integers such that $ac = bc$ and c is non-zero, then $a = b$.*

Proof See Exercise 4.1.6. □

We now extend the notion of order, which was defined on the natural numbers, to the integers by repeating the definition verbatim:

Definition 4.1.10 (*Ordering of the integers*) Let n and m be integers. We say that n is *greater than or equal to* m, and write $n \geq m$ or $m \leq n$, iff we have $n = m + a$ for some natural number a. We say that n is *strictly greater than* m, and write $n > m$ or $m < n$, iff $n \geq m$ and $n \neq m$.

Thus for instance $5 > -3$, because $5 = -3 + 8$ and $5 \neq -3$. Clearly this definition is consistent with the notion of order on the natural numbers, since we are using the same definition.

Using the laws of algebra in Proposition 4.1.6 it is not hard to show the following properties of order:

Lemma 4.1.11 (Properties of order). *Let a, b, c be integers.*

(*a*) *$a > b$ if and only if $a - b$ is a positive natural number.*
(*b*) (*Addition preserves order*) *If $a > b$, then $a + c > b + c$.*
(*c*) (*Positive multiplication preserves order*) *If $a > b$ and c is positive, then $ac > bc$.*
(*d*) (*Negation reverses order*) *If $a > b$, then $-a < -b$.*
(*e*) (*Order is transitive*) *If $a > b$ and $b > c$, then $a > c$.*
(*f*) (*Order trichotomy*) *Exactly one of the statements $a > b$, $a < b$, or $a = b$ is true.*

Proof See Exercise 4.1.7. □

— Exercises —

Exercise 4.1.1 Verify that the definition of equality on the integers is both reflexive and symmetric.

Exercise 4.1.2 Show that the definition of negation on the integers is well-defined in the sense that if $(a—b) = (a'—b')$, then $-(a—b) = -(a'—b')$ (so equal integers have equal negations).

Exercise 4.1.3 Show that $(-1) \times a = -a$ for every integer a.

Exercise 4.1.4 Prove the remaining identities in Proposition 4.1.6. (*Hint:* one can save some work by using some identities to prove others. For instance, once you know that $xy = yx$, you get for free that $x1 = 1x$, and once you also prove $x(y + z) = xy + xz$, you automatically get $(y + z)x = yx + zx$ for free.)

Exercise 4.1.5 Prove Proposition 4.1.8. (*Hint:* while this proposition is not quite the same as Lemma 2.3.3, it is certainly legitimate to use Lemma 2.3.3 in the course of proving Proposition 4.1.8.)

Exercise 4.1.6 Prove Corollary 4.1.9. (*Hint:* there are two ways to do this. One is to use Proposition 4.1.8 to conclude that $a - b$ must be zero. Another way is to combine Corollary 2.3.7 with Lemma 4.1.5.)

Exercise 4.1.7 Prove Lemma 4.1.11. (*Hint:* use the first part of this lemma to prove all the others.)

Exercise 4.1.8 Show that the principle of induction (Axiom 2.5) does not apply directly to the integers. More precisely, give an example of a property $P(n)$ pertaining to an integer n such that $P(0)$ is true, and that $P(n)$ implies $P(n{+}{+})$ for all integers n, but that $P(n)$ is not true for all integers n. Thus induction is not as useful a tool for dealing with the integers as it is with the natural numbers. (The situation becomes even worse with the rational and real numbers, which we shall define shortly.)

Exercise 4.1.9 Show that the square of an integer is always a natural number. That is to say, prove that $n^2 \geq 0$ for every integer n.

4.2 The Rationals

We have now constructed the integers, with the operations of addition, subtraction, multiplication, and order and verified all the expected algebraic and order-theoretic properties. Now we will use a similar construction to build the rationals, adding division to our mix of operations.

Just like the integers were constructed by subtracting two natural numbers, the rationals can be constructed by dividing two integers, though of course we have to make the usual caveat that the denominator should be non-zero.[2] Of course, just as two differences $a - b$ and $c - d$ can be equal if $a + d = c + b$, we know (from more advanced knowledge) that two quotients a/b and c/d can be equal if $ad = bc$. Thus, in analogy with the integers, we create a new meaningless symbol // (which will eventually be superceded by division), and define

Definition 4.2.1 A *rational number* is an expression of the form $a//b$, where a and b are integers and b is non-zero; $a//0$ is not considered to be a rational number. Two rational numbers are considered to be equal, $a//b = c//d$, if and only if $ad = cb$. The set of all rational numbers is denoted $\mathbf{Q}$.

Thus for instance $3//4 = 6//8 = -3// - 4$, but $3//4 \neq 4//3$. This is a valid definition of equality (Exercise 4.2.1). Now we need a notion of addition, multiplication, and negation. Again, we will take advantage of our pre-existing knowledge, which tells us that $a/b + c/d$ should equal $(ad + bc)/(bd)$ and that $a/b * c/d$ should equal ac/bd, while $-(a/b)$ equals $(-a)/b$. Motivated by this foreknowledge, we define

Definition 4.2.2 If $a//b$ and $c//d$ are rational numbers, we define their sum

$$(a//b) + (c//d) := (ad + bc)//(bd)$$

[2] There is no reasonable way we can divide by zero, since one cannot have both the identities $(a/b) * b = a$ and $c * 0 = 0$ hold simultaneously if b is allowed to be zero and a is non-zero. Similarly, the identities $a/a = 1$ and $2 * (a/a) = (2 * a)/a$ cannot simultaneously hold if $0/0$ is defined. However, we can eventually get a reasonable notion of dividing by a quantity which *approaches* zero-think of L'Hôpital's rule (see Sect. 10.5), which suffices for doing things like defining differentiation.

their product

$$(a//b) * (c//d) := (ac)//(bd)$$

and the negation

$$-(a//b) := (-a)//b.$$

Note that if b and d are non-zero, then bd is also non-zero, by Proposition 4.1.8, so the sum or product of two rational numbers remains a rational number.

Lemma 4.2.3 *The sum, product, and negation operations on rational numbers are well-defined, in the sense that if one replaces $a//b$ with another rational number $a'//b'$ which is equal to $a//b$, then the output of the above operations remains unchanged, and similarly for $c//d$.*

Proof We just verify this for addition; we leave the remaining claims to Exercise 4.2.2. Suppose $a//b = a'//b'$, so that b and b' are non-zero and $ab' = a'b$. We now show that $a//b + c//d = a'//b' + c//d$. By definition, the left-hand side is $(ad + bc)//bd$ and the right-hand side is $(a'd + b'c)//b'd$, so we have to show that

$$(ad + bc)b'd = (a'd + b'c)bd,$$

which expands to

$$ab'd^2 + bb'cd = a'bd^2 + bb'cd.$$

But since $ab' = a'b$, the claim follows. Similarly if one replaces $c//d$ by $c'//d'$. □

We note that the rational numbers $a//1$ behave in a manner identical to the integers a:

$$(a//1) + (b//1) = (a + b)//1;$$
$$(a//1) \times (b//1) = (ab//1);$$
$$-(a//1) = (-a)//1.$$

Also, $a//1$ and $b//1$ are only equal when a and b are equal. Because of this, we will identify a with $a//1$ for each integer a: $a \equiv a//1$; the above identities then guarantee that the arithmetic of the integers is consistent with the arithmetic of the rationals. Thus just as we embedded the natural numbers inside the integers, we embed the integers inside the rational numbers. In particular, all natural numbers are rational numbers, for instance 0 is equal to $0//1$ and 1 is equal to $1//1$.

Observe that a rational number $a//b$ is equal to $0 = 0//1$ if and only if $a \times 1 = b \times 0$, i.e., if the numerator a is equal to 0. Thus if a and b are non-zero then so is $a//b$.

We now define a new operation on the rationals: reciprocal. If $x = a//b$ is a non-zero rational (so that $a, b \neq 0$) then we define the *reciprocal* x^{-1} of x to be the rational number $x^{-1} := b//a$. It is easy to check that this operation is consistent with our notion of equality: if two rational numbers $a//b$, $a'//b'$ are equal, then their reciprocals are also equal. (In contrast, an operation such as "numerator" is not well-defined: the rationals $3//4$ and $6//8$ are equal, but have unequal numerators, so we have to be careful when referring to such terms as "the numerator of x".) We however leave the reciprocal of 0 undefined.

We now summarize the algebraic properties of the rationals.

Proposition 4.2.4 (Laws of algebra for rationals) *Let x, y, z be rationals. Then the following laws of algebra hold:*

$$\begin{aligned} x+y &= y+x \\ (x+y)+z &= x+(y+z) \\ x+0 = 0+x &= x \\ x+(-x) = (-x)+x &= 0 \\ xy &= yx \\ (xy)z &= x(yz) \\ x1 = 1x &= x \\ x(y+z) &= xy+xz \\ (y+z)x &= yx+zx. \end{aligned}$$

If x is non-zero, we also have

$$xx^{-1} = x^{-1}x = 1.$$

Remark 4.2.5 The above set of ten identities have a name; they are asserting that the rationals **Q** form a *field*. This is better than being a commutative ring because of the tenth identity $xx^{-1} = x^{-1}x = 1$. Note that this proposition supercedes Proposition 4.1.6.

Proof To prove this identity, one writes $x = a//b$, $y = c//d$, $z = e//f$ for some integers a, c, e and non-zero integers b, d, f, and verifies each identity in turn using the algebra of the integers. We shall just prove the longest one, namely $(x+y)+z = x+(y+z)$:

$$\begin{aligned} (x+y)+z &= ((a//b)+(c//d))+(e//f) \\ &= ((ad+bc)//bd)+(e//f) \\ &= (adf+bcf+bde)//bdf; \\ x+(y+z) &= (a//b)+((c//d)+(e//f)) \\ &= (a//b)+((cf+de)//df) \\ &= (adf+bcf+bde)//bdf \end{aligned}$$

and so one can see that $(x+y)+z$ and $x+(y+z)$ are equal. The other identities are proven in a similar fashion and are left to Exercise 4.2.3. □

We can now define the *quotient* x/y of two rational numbers x and y, *provided that* y is non-zero, by the formula

$$x/y := x \times y^{-1}.$$

Thus, for instance

$$(3//4)/(5//6) = (3//4) \times (6//5) = (18//20) = (9//10).$$

Using this formula, it is easy to see that $a/b = a//b$ for every integer a and every non-zero integer b. Thus we can now discard the // notation, and use the more customary a/b instead of $a//b$.

In a similar spirit, we define subtraction on the rationals by the formula

$$x - y := x + (-y),$$

just as we did with the integers.

Proposition 4.2.4 allows us to use all the normal rules of algebra; we will now proceed to do so without further comment.

In the previous section we organized the integers into positive, zero, and negative numbers. We now do the same for the rationals.

Definition 4.2.6 A rational number x is said to be *positive* iff we have $x = a/b$ for some positive integers a and b. It is said to be *negative* iff we have $x = -y$ for some positive rational y (i.e., $x = (-a)/b$ for some positive integers a and b).

Thus for instance, every positive integer is a positive rational number, and every negative integer is a negative rational number, so our new definition is consistent with our old one.

Lemma 4.2.7 (Trichotomy of rationals) *Let x be a rational number. Then exactly one of the following three statements is true:* (a) *x is equal to 0.* (b) *x is a positive rational number.* (c) *x is a negative rational number.*

Proof See Exercise 4.2.4. □

Definition 4.2.8 (*Ordering of the rationals*) Let x and y be rational numbers. We say that $x > y$ iff $x - y$ is a positive rational number, and $x < y$ iff $x - y$ is a negative rational number. We write $x \geq y$ iff either $x > y$ or $x = y$, and similarly define $x \leq y$.

Proposition 4.2.9 (Basic properties of order on the rationals) *Let x, y, z be rational numbers. Then the following properties hold.*

(a) *(Order trichotomy) Exactly one of the three statements $x = y$, $x < y$, or $x > y$ is true.*

(b) *(Order is antisymmetric) One has $x < y$ if and only if $y > x$.*

(*c*) *(Order is transitive) If $x < y$ and $y < z$, then $x < z$.*
(*d*) *(Addition preserves order) If $x < y$, then $x + z < y + z$.*
(*e*) *(Positive multiplication preserves order) If $x < y$ and z is positive, then $xz < yz$.*

Proof See Exercise 4.2.5. □

Remark 4.2.10 The above five properties in Proposition 4.2.9, combined with the field axioms in Proposition 4.2.4, have a name: they assert that the rationals $\mathbf{Q}$ form an *ordered field*. It is important to keep in mind that Proposition 4.2.9(e) only works when z is positive, see Exercise 4.2.6.

— Exercises —

Exercise 4.2.1 Show that the definition of equality for the rational numbers is reflexive, symmetric, and transitive. (*Hint:* for transitivity, use Corollary 4.1.9.)

Exercise 4.2.2 Prove the remaining components of Lemma 4.2.3.

Exercise 4.2.3 Prove the remaining components of Proposition 4.2.4. (*Hint:* as with Proposition 4.1.6, you can save some work by using some identities to prove others.)

Exercise 4.2.4 Prove Lemma 4.2.7. (Note that, as in Proposition 2.2.13, you have to prove two different things: firstly, that *at least* one of (a), (b), (c) is true; and secondly, that *at most* one of (a), (b), (c) is true.)

Exercise 4.2.5 Prove Proposition 4.2.9.

Exercise 4.2.6 Show that if x, y, z are rational numbers such that $x < y$ and z is *negative*, then $xz > yz$.

4.3 Absolute Value and Exponentiation

We have already introduced the four basic arithmetic operations of addition, subtraction, multiplication, and division on the rationals. (Recall that subtraction and division came from the more primitive notions of negation and reciprocal by the formulae $x - y := x + (-y)$ and $x/y := x \times y^{-1}$.) We also have a notion of order $<$, and have organized the rationals into the positive rationals, the negative rationals, and zero. In short, we have shown that the rationals $\mathbf{Q}$ form an *ordered field*.

One can now use these basic operations to construct more operations. There are many such operations we can construct, but we shall just introduce two particularly useful ones: absolute value and exponentiation.

Definition 4.3.1 (*Absolute value*) If x is a rational number, the *absolute value* $|x|$ of x is defined as follows. If x is positive, then $|x| := x$. If x is negative, then $|x| := -x$. If x is zero, then $|x| := 0$.

Definition 4.3.2 (*Distance*) Let x and y be rational numbers. The quantity $|x - y|$ is called the *distance between x and y* and is sometimes denoted $d(x, y)$, thus $d(x, y) := |x - y|$. For instance, $d(3, 5) = 2$.

Proposition 4.3.3 (Basic properties of absolute value and distance*) Let x, y, z be rational numbers.*

(*a*) (*Non-degeneracy of absolute value*) *We have* $|x| \geq 0$. *Also,* $|x| = 0$ *if and only if x is 0.*
(*b*) (*Triangle inequality for absolute value*) *We have* $|x + y| \leq |x| + |y|$.
(*c*) *We have the inequalities* $-y \leq x \leq y$ *if and only if* $y \geq |x|$. *In particular, we have* $-|x| \leq x \leq |x|$.
(*d*) (*Multiplicativity of absolute value*) *We have* $|xy| = |x|\,|y|$. *In particular,* $|-x| = |x|$.
(*e*) (*Non-degeneracy of distance*) *We have* $d(x, y) \geq 0$. *Also,* $d(x, y) = 0$ *if and only if* $x = y$.
(*f*) (*Symmetry of distance*) $d(x, y) = d(y, x)$.
(*g*) (*Triangle inequality for distance*) $d(x, z) \leq d(x, y) + d(y, z)$.

Proof See Exercise 4.3.1. □

Absolute value is useful for measuring how "close" two numbers are. Let us make a somewhat artificial definition:

Definition 4.3.4 (*ε-closeness*) Let $\varepsilon > 0$ be a rational number, and let x, y be rational numbers. We say that y is *ε-close* to x iff we have $d(y, x) \leq \varepsilon$.

Remark 4.3.5 This definition is not standard in mathematics textbooks; we will use it as "scaffolding" to construct the more important notions of limits (and of Cauchy sequences) later on, and once we have those more advanced notions we will discard the notion of ε-close.

Examples 4.3.6 The numbers 0.99 and 1.01 are 0.1-close, but they are not 0.01 close, because $d(0.99, 1.01) = |0.99 - 1.01| = 0.02$ is larger than 0.01. The numbers 2 and 2 are ε-close for every positive ε.

We do not bother defining a notion of ε-close when ε is zero or negative, because if ε is zero then x and y are only ε-close when they are equal, and when ε is negative then x and y are never ε-close. (In any event it is a long-standing tradition in analysis that the Greek letters ε, δ should only denote small positive numbers.)

Some basic properties of ε-closeness are the following.

Proposition 4.3.7 *Let x, y, z, w be rational numbers.*

(*a*) *If $x = y$, then x is ε-close to y for every* $\varepsilon > 0$. *Conversely, if x is ε-close to y for every* $\varepsilon > 0$, *then we have* $x = y$.
(*b*) *Let* $\varepsilon > 0$. *If x is ε-close to y, then y is ε-close to x.*
(*c*) *Let* $\varepsilon, \delta > 0$. *If x is ε-close to y, and y is δ-close to z, then x and z are $(\varepsilon + \delta)$-close.*
(*d*) *Let* $\varepsilon, \delta > 0$. *If x and y are ε-close, and z and w are δ-close, then $x + z$ and $y + w$ are $(\varepsilon + \delta)$-close, and $x - z$ and $y - w$ are also $(\varepsilon + \delta)$-close.*
(*e*) *Let* $\varepsilon > 0$. *If x and y are ε-close, they are also ε'-close for every* $\varepsilon' > \varepsilon$.

(*f*) *Let $\varepsilon > 0$. If y and z are both ε-close to x, and w is between y and z (i.e., $y \leq w \leq z$ or $z \leq w \leq y$), then w is also ε-close to x.*

(*g*) *Let $\varepsilon > 0$. If x and y are ε-close, and z is non-zero, then xz and yz are $\varepsilon|z|$-close.*

(*h*) *Let $\varepsilon, \delta > 0$. If x and y are ε-close, and z and w are δ-close, then xz and yw are $(\varepsilon|z| + \delta|x| + \varepsilon\delta)$-close.*

Proof We only prove the most difficult one, (h); we leave (a)–(g) to Exercise 4.3.2. Let $\varepsilon, \delta > 0$, and suppose that x and y are ε-close. If we write $a := y - x$, then we have $y = x + a$ and that $|a| \leq \varepsilon$. Similarly, if z and w are δ-close, and we define $b := w - z$, then $w = z + b$ and $|b| \leq \delta$.

Since $y = x + a$ and $w = z + b$, we have

$$yw = (x + a)(z + b) = xz + az + xb + ab.$$

Thus

$$|yw - xz| = |az + bx + ab| \leq |az| + |bx| + |ab| = |a||z| + |b||x| + |a||b|.$$

Since $|a| \leq \varepsilon$ and $|b| \leq \delta$, we thus have

$$|yw - xz| \leq \varepsilon|z| + \delta|x| + \varepsilon\delta$$

and thus that yw and xz are $(\varepsilon|z| + \delta|x| + \varepsilon\delta)$-close. □

Remark 4.3.8 One should compare statements (a)–(c) of this proposition with the reflexive, symmetric, and transitive axioms of equality. It is often useful to think of the notion of "ε-close" as an approximate substitute for that of equality in analysis.

Now we recursively define exponentiation for natural number exponents, extending the previous definition in Definition 2.3.11.

Definition 4.3.9 (*Exponentiation to a natural number*) Let x be a rational number. To raise x to the power 0, we define $x^0 := 1$; in particular we define $0^0 := 1$. Now suppose inductively that x^n has been defined for some natural number n, then we define $x^{n+1} := x^n \times x$.

Proposition 4.3.10 (Properties of exponentiation, I) *Let x, y be rational numbers, and let n, m be natural numbers.*

(*a*) *We have $x^n x^m = x^{n+m}$, $(x^n)^m = x^{nm}$, and $(xy)^n = x^n y^n$.*

(*b*) *Suppose $n > 0$. Then we have $x^n = 0$ if and only if $x = 0$.*

(*c*) *If $x \geq y \geq 0$, then $x^n \geq y^n \geq 0$. If $x > y \geq 0$ and $n > 0$, then $x^n > y^n \geq 0$.*

(*d*) *We have $|x^n| = |x|^n$.*

Proof See Exercise 4.3.3. □

Now we define exponentiation for negative integer exponents.

Definition 4.3.11 (*Exponentiation to a negative number*) Let x be a non-zero rational number. Then for any negative integer $-n$, we define $x^{-n} := 1/x^n$.

Thus for instance $x^{-3} = 1/x^3 = 1/(x \times x \times x)$. Note that when $n = 1$, the definition of x^{-1} provided by Definition 4.3.11 coincides with the reciprocal of x defined in Sect. 4.2, so there is no incompatibility of notation caused by this new definition.

We now have x^n defined for any integer n, whether n is positive, negative, or zero. Exponentiation with integer exponents has the following properties (which supercede Proposition 4.3.10):

Proposition 4.3.12 (Properties of exponentiation, II) *Let x, y be non-zero rational numbers, and let n, m be integers.*

(*a*) *We have $x^n x^m = x^{n+m}$, $(x^n)^m = x^{nm}$, and $(xy)^n = x^n y^n$.*
(*b*) *If $x \geq y > 0$, then $x^n \geq y^n > 0$ if n is positive, and $0 < x^n \leq y^n$ if n is negative.*
(*c*) *If $x, y > 0$, $n \neq 0$, and $x^n = y^n$, then $x = y$.*
(*d*) *We have $|x^n| = |x|^n$.*

Proof See Exercise 4.3.4. □

— Exercises —

Exercise 4.3.1 Prove Proposition 4.3.3. (*Hint:* while all of these claims can be proven by dividing into cases, such as when x is positive, negative, or zero, several parts of the proposition can be proven without such a tedious division into cases. For instance one can use earlier parts of the proposition to prove later ones.)

Exercise 4.3.2 Prove the remaining claims in Proposition 4.3.7.

Exercise 4.3.3 Prove Proposition 4.3.10. (*Hint:* use induction.)

Exercise 4.3.4 Prove Proposition 4.3.12. (*Hint:* induction is not suitable here. Instead, use Proposition 4.3.10.)

Exercise 4.3.5 Prove that $2^N \geq N$ for all positive integers N. (*Hint:* use induction.)

4.4 Gaps in the Rational Numbers

Imagine that we arrange the rationals on a line, arranging x to the right of y if $x > y$. (This is a non-rigorous arrangement, since we have not yet defined the concept of a line, but this discussion is only intended to motivate the more rigorous propositions below.) Inside the rationals we have the integers, which are thus also arranged on the line. Now we work out how the rationals are arranged with respect to the integers.

Proposition 4.4.1 (Interspersing of integers by rationals). *Let x be a rational number. Then there exists an integer n such that $n \leq x < n + 1$. In fact, this integer is unique (i.e., for each x there is only one n for which $n \leq x < n + 1$). In particular, there exists a natural number N such that $N > x$ (i.e., there is no such thing as a rational number which is larger than all the natural numbers).*

Remark 4.4.2 The integer n for which $n \leq x < n + 1$ is sometimes referred to as the *integer part* of x and is sometimes denoted $n = \lfloor x \rfloor$.

Proof See Exercise 4.4.1. □

Also, between every two rational numbers there is at least one additional rational:

Proposition 4.4.3 (Interspersing of rationals by rationals). *If x and y are two rationals such that $x < y$, then there exists a third rational z such that $x < z < y$.*

Proof We set $z := (x + y)/2$. Since $x < y$, and $1/2 = 1//2$ is positive, we have from Proposition 4.2.9 that $x/2 < y/2$. If we add $y/2$ to both sides using Proposition 4.2.9 we obtain $x/2 + y/2 < y/2 + y/2$, i.e., $z < y$. If we instead add $x/2$ to both sides we obtain $x/2 + x/2 < y/2 + x/2$, i.e., $x < z$. Thus $x < z < y$ as desired. □

Despite the rationals having this denseness property, they are still incomplete; there are still an infinite number of “gaps” or “holes” between the rationals, although this denseness property does ensure that these holes are in some sense infinitely small. For instance, we will now show that the rational numbers do not contain any square root of two.

Proposition 4.4.4 *There does not exist any rational number x for which $x^2 = 2$.*

Proof We only give a sketch of a proof; the gaps will be filled in Exercise 4.4.3. Suppose for sake of contradiction that we had a rational number x for which $x^2 = 2$. Clearly x is not zero. We may assume that x is positive, for if x were negative then we could just replace x by $-x$ (since $x^2 = (-x)^2$). Thus $x = p/q$ for some positive integers p, q, so $(p/q)^2 = 2$, which we can rearrange as $p^2 = 2q^2$. Define a natural number p to be *even* if $p = 2k$ for some natural number k, and *odd* if $p = 2k + 1$ for some natural number k. Every natural number is either even or odd, but not both (why?). If p is odd, then p^2 is also odd (why?), which contradicts $p^2 = 2q^2$. Thus p is even, i.e., $p = 2k$ for some natural number k. Since p is positive, k must also be positive. Inserting $p = 2k$ into $p^2 = 2q^2$ we obtain $4k^2 = 2q^2$, so that $q^2 = 2k^2$.

To summarize, we started with a pair (p, q) of positive integers such that $p^2 = 2q^2$, and ended up with a pair (q, k) of positive integers such that $q^2 = 2k^2$. Since $p^2 = 2q^2$, we have $q < p$ (why?). If we rewrite $p' := q$ and $q' := k$, we thus can pass from one solution (p, q) to the equation $p^2 = 2q^2$ to a new solution (p', q') to the same equation which has a smaller value of p. But then we can repeat this procedure again and again, obtaining a sequence (p'', q''), (p''', q'''), etc., of solutions to $p^2 = 2q^2$, each one with a smaller value of p than the previous, and each one consisting of positive integers. But this contradicts the principle of infinite descent (see Exercise 4.4.2). This contradiction shows that we could not have had a rational x for which $x^2 = 2$. □

On the other hand, we can get rational numbers which are arbitrarily close to a square root of 2:

Proposition 4.4.5 *For every rational number $\varepsilon > 0$, there exists a non-negative rational number x such that $x^2 < 2 < (x+\varepsilon)^2$.*

Proof Let $\varepsilon > 0$ be rational. Suppose for sake of contradiction that there is no non-negative rational number x for which $x^2 < 2 < (x+\varepsilon)^2$. This means that whenever x is non-negative and $x^2 < 2$, we must also have $(x+\varepsilon)^2 < 2$ (note that $(x+\varepsilon)^2$ cannot equal 2, by Proposition 4.4.4). Since $0^2 < 2$, we thus have $\varepsilon^2 < 2$, which then implies $(2\varepsilon)^2 < 2$, and indeed a simple induction shows that $(n\varepsilon)^2 < 2$ for every natural number n. (Note that $n\varepsilon$ is non-negative for every natural number n - why?) But, by Proposition 4.4.1 we can find an integer n such that $n > 2/\varepsilon$, which implies that $n\varepsilon > 2$, which implies that $(n\varepsilon)^2 > 4 > 2$, contradicting the claim that $(n\varepsilon)^2 < 2$ for all natural numbers n. This contradiction gives the proof. □

Example 4.4.6 If[3] $\varepsilon = 0.001$, we can take $x = 1.414$, since $x^2 = 1.999396$ and $(x+\varepsilon)^2 = 2.002225$.

Proposition 4.4.5 indicates that, while the set **Q** of rationals does not actually have $\sqrt{2}$ as a member, we can get as close as we wish to $\sqrt{2}$. For instance, the sequence of rationals

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

seem to get closer and closer to $\sqrt{2}$, as their squares indicate:

$$1.96, 1.9881, 1.99396, 1.99996164, 1.9999899241, \ldots$$

Thus it seems that we can create a square root of 2 by taking a "limit" of a sequence of rationals. This is how we shall construct the real numbers in the next chapter. (There is another way to do so, using something called "Dedekind cuts", which we will not pursue here. One can also proceed using infinite decimal expansions, but there are some sticky issues when doing so, e.g., one has to make $0.999\ldots$ equal to $1.000\ldots$, and this approach, despite being the most familiar, is actually *more* complicated than other approaches; see Appendix B.)

— Exercises —

Exercise 4.4.1 Prove Proposition 4.4.1. (*Hint:* use Proposition 2.3.9.)

Exercise 4.4.2 A definition: a sequence $a_0, a_1, a_2, \ldots$ of numbers (natural numbers, integers, rationals, or reals) is said to be in *infinite descent* if we have $a_n > a_{n+1}$ for all natural numbers n (i.e., $a_0 > a_1 > a_2 > \ldots$).

(a) Prove the *principle of infinite descent*: that it is not possible to have a sequence of *natural numbers* which is in infinite descent. (*Hint:* assume for sake of contradiction that you can find a sequence of natural numbers which is in infinite descent. Since all the a_n are natural numbers, you know that $a_n \geq 0$ for all n. Now use induction to show in fact that $a_n \geq k$ for all $k \in \mathbf{N}$ and all $n \in \mathbf{N}$, and obtain a contradiction.)

[3] We will use the decimal system for defining terminating decimals, for instance 1.414 is defined to equal the rational number 1414/1000. For a formal discussion on the decimal system, see Appendix B.

(b) Does the principle of infinite descent work if the sequence $a_1, a_2, a_3, \ldots$ is allowed to take integer values instead of natural number values? What about if it is allowed to take positive rational values instead of natural numbers? Explain.

Exercise 4.4.3 Fill in the gaps marked (why?) in the proof of Proposition 4.4.4. Is the axiom of choice required to establish this proposition?

Chapter 5
The Real Numbers

To review our progress to date, we have rigorously constructed three fundamental number systems: the natural number system **N**, the integers **Z**, and the rationals **Q**.[1] We defined the natural numbers using the five Peano axioms and postulated that such a number system existed; this is plausible, since the natural numbers correspond to the very intuitive and fundamental notion of *sequential counting*. Using that number system one could then recursively define addition and multiplication, and verify that they obeyed the usual laws of algebra. We then constructed the integers by taking formal[2] differences of the natural numbers, $a \text{—} b$. We then constructed the rationals by taking formal quotients of the integers, $a//b$, although we need to exclude division by zero in order to keep the laws of algebra reasonable. (You are of course free to design your own number system, possibly including one where division by zero is permitted; but you will have to give up one or more of the field axioms from Proposition 4.2.4, among other things, and you will probably get a less useful number system in which to do any real-world problems.)

The rational system is already sufficient to do a lot of mathematics—much of high school algebra, for instance, works just fine if one only knows about the rationals. However, there is a fundamental area of mathematics where the rational number system does not suffice—that of *geometry* (the study of lengths, areas, etc.). For instance, a right-angled triangle with both sides equal to 1 gives a hypotenuse of $\sqrt{2}$, which is an *irrational* number, i.e., not a rational number; see Proposition 4.4.4.

[1] The symbols **N**, **Q**, and **R** stand for "natural", "quotient", and "real" respectively. **Z** stands for "Zahlen", the German word for "numbers". There is also the *complex numbers* **C**, which obviously stands for "complex", which you will see in Sect. 4.6 of *Analysis II*.

[2] *Formal* means "having the form of"; at the beginning of our construction the expression $a \text{—} b$ did not actually *mean* the difference $a - b$, since the symbol — was meaningless. It only had the *form* of a difference. Later on we defined subtraction and verified that the formal difference was equal to the actual difference, so this eventually became a non-issue, and our symbol for formal differencing was discarded. Somewhat confusingly, this use of the term "formal" is unrelated to the notions of a formal argument and an informal argument.

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_5

Things get even worse when one starts to deal with the subfield of geometry known as *trigonometry*, when one sees numbers such as π or $\cos(1)$, which turn out to be in some sense "even more" irrational than $\sqrt{2}$. (These numbers are known as *transcendental numbers*, but to discuss this further would be far beyond the scope of this text.) Thus, in order to have a number system which can adequately describe geometry—or even something as simple as measuring lengths on a line—one needs to replace the rational number system with the real number system. Since differential and integral calculus is also intimately tied up with geometry—think of slopes of tangents, or areas under a curve—calculus also requires the real number system in order to function properly.

However, a rigorous way to construct the reals from the rationals turns out to be somewhat difficult—requiring a bit more machinery than what was needed to pass from the naturals to the integers, or the integers to the rationals. In those two constructions, the task was to introduce one more *algebraic* operation to the number system—e.g., one can get integers from naturals by introducing subtraction, and get the rationals from the integers by introducing division. But to get the reals from the rationals is to pass from a "discrete" system to a "continuous" one and requires the introduction of a somewhat different notion—that of a *limit*. The limit is a concept which on one level is quite intuitive, but to pin down rigorously turns out to be quite challenging. (Even such great mathematicians as Euler and Newton had difficulty with this concept. It was only in the nineteenth century that mathematicians such as Cauchy and Dedekind figured out how to deal with limits rigorously.)

In Sect. 4.4 we explored the "gaps" in the rational numbers; now we shall fill in these gaps using limits to create the real numbers. The real number system will end up being a lot like the rational numbers but will have some new operations—notably that of *supremum*, which can then be used to define limits and thence to everything else that calculus needs.

The procedure we give here of obtaining the real numbers as the limit of sequences of rational numbers may seem rather complicated. However, it is in fact an instance of a very general and useful procedure, that of *completing* one metric space to form another; see Exercise 1.4.8 of *Analysis II*.

5.1 Cauchy Sequences

Our construction of the real numbers shall rely on the concept of a *Cauchy sequence*. Before we define this notion formally, let us first define the concept of a sequence.

Definition 5.1.1 *(Sequences).* Let m be an integer. A *sequence* $(a_n)_{n=m}^{\infty}$ *of rational numbers* is any function from the set $\{n \in \mathbf{Z} : n \geq m\}$ to $\mathbf{Q}$, i.e., a mapping which assigns to each integer n greater than or equal to m, a rational number a_n. More informally, a sequence $(a_n)_{n=m}^{\infty}$ of rational numbers is a collection of rationals a_m, $a_{m+1}, a_{m+2}, \ldots$.

Example 5.1.2 The sequence $(n^2)_{n=0}^{\infty}$ is the collection $0, 1, 4, 9, \ldots$ of natural numbers; the sequence $(3)_{n=0}^{\infty}$ is the collection 3, 3, 3, ... of natural numbers. These sequences are indexed starting from 0, but we can of course make sequences starting from 1 or any other number; for instance, the sequence $(a_n)_{n=3}^{\infty}$ denotes the sequence $a_3, a_4, a_5, \ldots$, so $(n^2)_{n=3}^{\infty}$ is the collection 9, 16, 25, ... of natural numbers.

We want to define the real numbers as the limits of sequences of rational numbers. To do so, we have to distinguish which sequences of rationals are convergent and which ones are not. For instance, the sequence

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

looks like it is trying to converge to something, as does

$$0.1, 0.01, 0.001, 0.0001, \ldots$$

while other sequences such as

$$1, 2, 4, 8, 16, \ldots$$

or

$$1, 0, 1, 0, 1, \ldots$$

do not. To do this we use the definition of ε-closeness defined earlier. Recall from Definition 4.3.4 that two rational numbers x, y are ε-close if $d(x, y) = |x - y| \leq \varepsilon$.

Definition 5.1.3 *(ε-steadiness).* Let $\varepsilon > 0$. A sequence $(a_n)_{n=0}^{\infty}$ is said to be *ε-steady* iff each pair a_j, a_k of sequence elements is ε-close for every natural number j, k. In other words, the sequence $a_0, a_1, a_2, \ldots$ is ε-steady iff $|a_j - a_k| \leq \varepsilon$ for all j, k.

Remark 5.1.4 This definition is not standard in the literature; we will not need it outside of this section; similarly for the concept of "eventual ε-steadiness" below. We have defined ε-steadiness for sequences whose index starts at 0, but clearly we can make a similar notion for sequences whose indices start from any other number: a sequence $a_N, a_{N+1}, \ldots$ is ε-steady if one has $|a_j - a_k| \leq \varepsilon$ for all $j, k \geq N$.

Example 5.1.5 The sequence $1, 0, 1, 0, 1, \ldots$ is 1-steady but is not 1/2-steady. The sequence $0.1, 0.01, 0.001, 0.0001, \ldots$ is 0.1-steady, but is not 0.01-steady (why?). The sequence 1, 2, 4, 8, 16, ... is not ε-steady for any ε (why?). The sequence $2, 2, 2, 2, \ldots$ is ε-steady for every $\varepsilon > 0$.

The notion of ε-steadiness of a sequence is simple, but does not really capture the *limiting* behavior of a sequence, because it is too sensitive to the initial members of the sequence. For instance, the sequence

$$10, 0, 0, 0, 0, 0, \ldots$$

is 10-steady, but is not ε-steady for any smaller value of ε, despite the sequence converging almost immediately to zero. So we need a more robust notion of steadiness that does not care about the initial members of a sequence.

Definition 5.1.6 *(Eventual ε-steadiness).* Let $\varepsilon > 0$. A sequence $(a_n)_{n=0}^{\infty}$ is said to be *eventually ε-steady* iff the sequence $a_N, a_{N+1}, a_{N+2}, \dots$ is ε-steady for some natural number $N \geq 0$. In other words, the sequence $a_0, a_1, a_2, \dots$ is eventually ε-steady iff there exists an $N \geq 0$ such that $|a_j - a_k| \leq \varepsilon$ for all $j, k \geq N$.

Example 5.1.7 The sequence $a_1, a_2, \dots$ defined by $a_n := 1/n$, (i.e., the sequence $1, 1/2, 1/3, 1/4, \dots$) is not 0.1-steady, but is eventually 0.1-steady, because the sequence $a_{10}, a_{11}, a_{12}, \dots$ (i.e., $1/10, 1/11, 1/12, \dots$) is 0.1-steady. The sequence $10, 0, 0, 0, 0, \dots$ is not ε-steady for any ε less than 10, but it is eventually ε-steady for every $\varepsilon > 0$ (why?).

Now we can finally define the correct notion of what it means for a sequence of rationals to "want" to converge.

Definition 5.1.8 *(Cauchy sequences).* A sequence $(a_n)_{n=0}^{\infty}$ of rational numbers is said to be a *Cauchy sequence* iff for every rational $\varepsilon > 0$, the sequence $(a_n)_{n=0}^{\infty}$ is eventually ε-steady. In other words, the sequence $a_0, a_1, a_2, \dots$ is a Cauchy sequence iff for every $\varepsilon > 0$, there exists an $N \geq 0$ such that $d(a_j, a_k) \leq \varepsilon$ for all $j, k \geq N$.

Remark 5.1.9 At present, the parameter ε is restricted to be a positive rational; we cannot take ε to be an arbitrary positive real number, because the real numbers have not yet been constructed. However, once we do construct the real numbers, we shall see that the above definition will not change if we require ε to be real instead of rational. In other words, we will eventually prove that a sequence is eventually ε-steady for every rational $\varepsilon > 0$ if and only if it is eventually ε-steady for every real $\varepsilon > 0$; see Proposition 6.1.4. This rather subtle distinction between a rational ε and a real ε turns out not to be very important in the long run, and the reader is advised not to pay too much attention as to what type of number ε should be.

Example 5.1.10 *(Informal)* Consider the sequence

$$1.4, 1.41, 1.414, 1.4142, \dots$$

mentioned earlier. This sequence is already 0.1-steady. If one discards the first element 1.4, then the remaining sequence

$$1.41, 1.414, 1.4142, \dots$$

is now 0.01-steady, which means that the original sequence was eventually 0.01-steady. Discarding the next element gives the 0.001-steady sequence 1.414, 1.4142, ...; thus the original sequence was eventually 0.001-steady. Continuing in this way it seems plausible that this sequence is in fact ε-steady for every $\varepsilon > 0$, which seems to suggest that this is a Cauchy sequence. However, this discussion is not rigorous for several reasons, for instance we have not precisely defined what this sequence 1.4, 1.41, 1.414, ... really is. An example of a rigorous treatment follows next.

Proposition 5.1.11 *The sequence $a_1, a_2, a_3, \ldots$ defined by $a_n := 1/n$ (i.e., the sequence $1, 1/2, 1/3, \ldots$) is a Cauchy sequence.*

Proof We have to show that for every $\varepsilon > 0$, the sequence $a_1, a_2, \ldots$ is eventually ε-steady. So let $\varepsilon > 0$ be arbitrary. We now have to find a number $N \geq 1$ such that the sequence $a_N, a_{N+1}, \ldots$ is ε-steady. Let us see what this means. This means that $d(a_j, a_k) \leq \varepsilon$ for every $j, k \geq N$, i.e.

$$|1/j - 1/k| \leq \varepsilon \text{ for every } j, k \geq N.$$

Now since $j, k \geq N$, we know that $0 < 1/j, 1/k \leq 1/N$, so that $|1/j - 1/k| \leq 1/N$. So in order to force $|1/j - 1/k|$ to be less than or equal to ε, it would be sufficient for $1/N$ to be less than ε. So all we need to do is choose an N such that $1/N$ is less than ε, or in other words that N is greater than $1/\varepsilon$. But this can be done thanks to Proposition 4.4.1. □

As you can see, verifying from first principles (i.e., without using any of the machinery of limits, etc.) that a sequence is a Cauchy sequence requires some effort, even for a sequence as simple as $1/n$. The part about selecting an N can be particularly difficult for beginners—one has to think in reverse, working out what conditions on N would suffice to force the sequence $a_N, a_{N+1}, a_{N+2}, \ldots$ to be ε-steady, and then finding an N which obeys those conditions. Later we will develop some limit laws which allow us to determine when a sequence is Cauchy more easily.

We now relate the notion of a Cauchy sequence to another basic notion, that of a bounded sequence.

Definition 5.1.12 *(Bounded sequences).* Let $M \geq 0$ be rational. A finite sequence $a_1, a_2, \ldots, a_n$ is *bounded by M* iff $|a_i| \leq M$ for all $1 \leq i \leq n$. An infinite sequence $(a_n)_{n=1}^{\infty}$ is *bounded by M* iff $|a_i| \leq M$ for all $i \geq 1$. A sequence is said to be *bounded* iff it is bounded by M for some rational $M \geq 0$.

Example 5.1.13 The finite sequence $1, -2, 3, -4$ is bounded (in this case, it is bounded by 4, or indeed by any M greater than or equal to 4). But the infinite sequence $1, -2, 3, -4, 5, -6, \ldots$ is unbounded. (Can you prove this? Use Proposition 4.4.1.) The sequence $1, -1, 1, -1, \ldots$ is bounded (e.g., by 1), but is not a Cauchy sequence.

Lemma 5.1.14 (Finite sequences are bounded). *Every finite sequence $a_1, a_2, \ldots, a_n$ is bounded.*

Proof We prove this by induction on n. When $n = 1$ the sequence a_1 is clearly bounded, for if we choose $M := |a_1|$ then clearly we have $|a_i| \leq M$ for all $1 \leq i \leq n$. Now suppose that we have already proved the lemma for some $n \geq 1$; we now prove it for $n + 1$, i.e., we prove every sequence $a_1, a_2, \ldots, a_{n+1}$ is bounded. By the induction hypothesis we know that $a_1, a_2, \ldots, a_n$ is bounded by some $M \geq 0$;

in particular, it must be bounded by $M + |a_{n+1}|$. On the other hand, a_{n+1} is also bounded by $M + |a_{n+1}|$. Thus $a_1, a_2, \ldots, a_n, a_{n++}$ is bounded by $M + |a_{n+1}|$, and is hence bounded. This closes the induction. □

Note that while this argument shows that every finite sequence is bounded, no matter how long the finite sequence is, it does not say anything about whether an infinite sequence is bounded or not; infinity is not a natural number. However, we have

Lemma 5.1.15 (Cauchy sequences are bounded). *Every Cauchy sequence $(a_n)_{n=1}^\infty$ is bounded.*

Proof See Exercise 5.1.1. □

— Exercises —

Exercise 5.1.1 Prove Lemma 5.1.15. (*Hint:* use the fact that a_n is eventually 1-steady, and thus can be split into a finite sequence and a 1-steady sequence. Then use Lemma 5.1.14 for the finite part. Note there is nothing special about the number 1 used here; any other positive number would have sufficed.)

Exercise 5.1.2 If $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are bounded sequences, show that $(a_n + b_n)_{n=1}^\infty$, $(a_n - b_n)_{n=1}^\infty$, and $(a_n b_n)_{n=1}^\infty$ are also bounded.

5.2 Equivalent Cauchy Sequences

Consider the two Cauchy sequences of rational numbers:

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

and

$$1.5, 1.42, 1.415, 1.4143, 1.41422, \ldots$$

Informally, both of these sequences seem to be converging to the same number, the square root $\sqrt{2} = 1.41421\ldots$ (though this statement is not yet rigorous because we have not defined real numbers yet). If we are to define the real numbers from the rationals as limits of Cauchy sequences, we have to know when two Cauchy sequences of rationals give the same limit, without first defining a real number (since that would be circular). To do this we use a similar set of definitions to those used to define a Cauchy sequence in the first place.

Definition 5.2.1 *(ε-close sequences).* Let $(a_n)_{n=0}^\infty$ and $(b_n)_{n=0}^\infty$ be two sequences, and let $\varepsilon > 0$. We say that the sequence $(a_n)_{n=0}^\infty$ is *ε-close* to $(b_n)_{n=0}^\infty$ iff a_n is ε-close to b_n for each $n \in \mathbf{N}$. In other words, the sequence $a_0, a_1, a_2, \ldots$ is ε-close to the sequence $b_0, b_1, b_2, \ldots$ iff $|a_n - b_n| \le \varepsilon$ for all $n = 0, 1, 2, \ldots$.

Example 5.2.2 The two sequences

$$1, -1, 1, -1, 1, \dots$$

and

$$1.1, -1.1, 1.1, -1.1, 1.1, \dots$$

are 0.1-close to each other. (Note however that neither of them are 0.1-steady).

Definition 5.2.3 *(Eventually ε-close sequences).* Let $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ be two sequences, and let $\varepsilon > 0$. We say that the sequence $(a_n)_{n=0}^{\infty}$ is *eventually ε-close* to $(b_n)_{n=0}^{\infty}$ iff there exists an $N \geq 0$ such that the sequences $(a_n)_{n=N}^{\infty}$ and $(b_n)_{n=N}^{\infty}$ are ε-close. In other words, $a_0, a_1, a_2, \dots$ is eventually ε-close to $b_0, b_1, b_2, \dots$ iff there exists an $N \geq 0$ such that $|a_n - b_n| \leq \varepsilon$ for all $n \geq N$.

Remark 5.2.4 Again, the notions of Oε-close sequences and eventually ε-close sequences are not standard in the literature, and we will not use them outside of this section.

Example 5.2.5 The two sequences

$$1.1, 1.01, 1.001, 1.0001, \dots$$

and

$$0.9, 0.99, 0.999, 0.9999, \dots$$

are not 0.1-close (because the first elements of both sequences are not 0.1-close to each other). However, the sequences are still eventually 0.1-close, because if we start from the second elements onwards in the sequence, these sequences are 0.1-close. A similar argument shows that the two sequences are eventually 0.01-close (by starting from the third element onwards), and so forth.

Definition 5.2.6 *(Equivalent sequences).* Two sequences $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ are *equivalent* iff for each rational $\varepsilon > 0$, the sequences $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ are eventually ε-close. In other words, $a_0, a_1, a_2, \dots$ and $b_0, b_1, b_2, \dots$ are equivalent iff for every rational $\varepsilon > 0$, there exists an $N \geq 0$ such that $|a_n - b_n| \leq \varepsilon$ for all $n \geq N$.

Remark 5.2.7 As with Definition 5.1.8, the quantity $\varepsilon > 0$ is currently restricted to be a positive rational, rather than a positive real. However, we shall eventually see that it makes no difference whether ε ranges over the positive rationals or positive reals; see Exercise 6.1.10.

From Definition 5.2.6 it seems that the two sequences given in Example 5.2.5 appear to be equivalent. We now prove this rigorously.

Proposition 5.2.8 *Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be the sequences $a_n = 1 + 10^{-n}$ and $b_n = 1 - 10^{-n}$. Then the sequences a_n, b_n are equivalent.*

Remark 5.2.9 This proposition, in decimal notation, asserts that $1.0000\ldots = 0.9999\ldots$; see Proposition B.2.3.

Proof We need to prove that for every $\varepsilon > 0$, the two sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are eventually ε-close to each other. So we fix an $\varepsilon > 0$. We need to find an $N > 0$ such that $(a_n)_{n=N}^\infty$ and $(b_n)_{n=N}^\infty$ are ε-close; in other words, we need to find an $N > 0$ such that

$$|a_n - b_n| \leq \varepsilon \text{ for all } n \geq N.$$

However, we have

$$|a_n - b_n| = |(1 + 10^{-n}) - (1 - 10^{-n})| = 2 \times 10^{-n}.$$

Since 10^{-n} is a decreasing function of n (i.e., $10^{-m} < 10^{-n}$ whenever $m > n$; this is easily proven by induction), and $n \geq N$, we have $2 \times 10^{-n} \leq 2 \times 10^{-N}$. Thus we have

$$|a_n - b_n| \leq 2 \times 10^{-N} \text{ for all } n \geq N.$$

Thus in order to obtain $|a_n - b_n| \leq \varepsilon$ for all $n \geq N$, it will be sufficient to choose N so that $2 \times 10^{-N} \leq \varepsilon$. This is easy to do using logarithms, but we have not yet developed logarithms yet, so we will use a cruder method. First, we observe 10^N is always greater than N for any $N \geq 1$ (see Exercise 4.3.5). Thus $10^{-N} \leq 1/N$, and so $2 \times 10^{-N} \leq 2/N$. Thus to get $2 \times 10^{-N} \leq \varepsilon$, it will suffice to choose N so that $2/N \leq \varepsilon$, or equivalently that $N \geq 2/\varepsilon$. But by Proposition 4.4.1 we can always choose such an N, and the claim follows. □

— Exercises —

Exercise 5.2.1 Show that if $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are equivalent sequences of rationals, then $(a_n)_{n=1}^\infty$ is a Cauchy sequence if and only if $(b_n)_{n=1}^\infty$ is a Cauchy sequence.

Exercise 5.2.2 Let $\varepsilon > 0$. Show that if $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are eventually ε-close, then $(a_n)_{n=1}^\infty$ is bounded if and only if $(b_n)_{n=1}^\infty$ is bounded.

5.3 The Construction of the Real Numbers

We are now ready to construct the real numbers. We shall introduce a new formal symbol LIM, similar to the formal notations — and // defined earlier; as the notation suggests, this will eventually match the familiar operation of lim, at which point the formal limit symbol can be discarded.

Definition 5.3.1 *(Real numbers).* A *real number* is defined to be an object of the form $\mathrm{LIM}_{n\to\infty} a_n$, where $(a_n)_{n=1}^\infty$ is a Cauchy sequence of rational numbers. Two real numbers $\mathrm{LIM}_{n\to\infty} a_n$ and $\mathrm{LIM}_{n\to\infty} b_n$ are said to be equal iff $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ are equivalent Cauchy sequences. The set of all real numbers is denoted $\mathbf{R}$.

Example 5.3.2 *(Informal)* Let $a_1, a_2, a_3, \ldots$ denote the sequence

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

and let $b_1, b_2, b_3, \ldots$ denote the sequence

$$1.5, 1.42, 1.415, 1.4143, 1.41422, \ldots$$

then $\mathrm{LIM}_{n\to\infty} a_n$ is a real number, and is the same real number as $\mathrm{LIM}_{n\to\infty} b_n$, because $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are equivalent Cauchy sequences: $\mathrm{LIM}_{n\to\infty} a_n = \mathrm{LIM}_{n\to\infty} b_n$.

We will refer to $\mathrm{LIM}_{n\to\infty} a_n$ as the *formal limit* of the sequence $(a_n)_{n=1}^{\infty}$. Later on we will define a genuine notion of limit, and show that the formal limit of a Cauchy sequence is the same as the limit of that sequence; after that, we will not need formal limits ever again. (The situation is much like what we did with formal subtraction — and formal division //.)

In order to ensure that this definition is valid, we need to check that the notion of equality in the definition obeys the first three axioms of equality:

Proposition 5.3.3 (Formal limits are well-defined). *Let $x = \mathrm{LIM}_{n\to\infty} a_n$, $y = \mathrm{LIM}_{n\to\infty} b_n$, and $z = \mathrm{LIM}_{n\to\infty} c_n$ be real numbers. Then, with the above definition of equality for real numbers, we have $x = x$. Also, if $x = y$, then $y = x$. Finally, if $x = y$ and $y = z$, then $x = z$.*

Proof See Exercise 5.3.1. □

Because of this proposition, we know that our definition of equality between two real numbers is legitimate. Of course, when we define other operations on the reals, we have to check that they obey the axiom of substitution: two real number inputs which are equal should give equal outputs when applied to any operation on the real numbers.

Now we want to define on the real numbers all the usual arithmetic operations, such as addition and multiplication. We begin with addition.

Definition 5.3.4 *(Addition of reals).* Let $x = \mathrm{LIM}_{n\to\infty} a_n$ and $y = \mathrm{LIM}_{n\to\infty} b_n$ be real numbers. Then we define the sum $x + y$ to be $x + y := \mathrm{LIM}_{n\to\infty}(a_n + b_n)$.

Example 5.3.5 The sum of $\mathrm{LIM}_{n\to\infty} 1 + 1/n$ and $\mathrm{LIM}_{n\to\infty} 2 + 3/n$ is $\mathrm{LIM}_{n\to\infty} 3 + 4/n$.

We now check that this definition is valid. The first thing we need to do is to confirm that the sum of two real numbers is in fact a real number:

Lemma 5.3.6 (Sum of Cauchy sequences is Cauchy). *Let $x = \mathrm{LIM}_{n\to\infty} a_n$ and $y = \mathrm{LIM}_{n\to\infty} b_n$ be real numbers. Then $x + y$ is also a real number (i.e., $(a_n + b_n)_{n=1}^{\infty}$ is a Cauchy sequence of rationals).*

Proof We need to show that for every $\varepsilon > 0$, the sequence $(a_n + b_n)_{n=1}^{\infty}$ is eventually ε-steady. Now from hypothesis we know that $(a_n)_{n=1}^{\infty}$ is eventually ε-steady, and $(b_n)_{n=1}^{\infty}$ is eventually ε-steady, but it turns out that this is not quite enough (this can be used to imply that $(a_n + b_n)_{n=1}^{\infty}$ is eventually 2ε-steady, but that's not what we want). So we need to do a little trick, which is to play with the value of ε.

We know that $(a_n)_{n=1}^{\infty}$ is eventually δ-steady for every value of δ. This implies not only that $(a_n)_{n=1}^{\infty}$ is eventually ε-steady, but it is also eventually $\varepsilon/2$-steady. Similarly, the sequence $(b_n)_{n=1}^{\infty}$ is also eventually $\varepsilon/2$-steady. This will turn out to be enough to conclude that $(a_n + b_n)_{n=1}^{\infty}$ is eventually ε-steady.

Since $(a_n)_{n=1}^{\infty}$ is eventually $\varepsilon/2$-steady, we know that there exists an $N \geq 1$ such that $(a_n)_{n=N}^{\infty}$ is $\varepsilon/2$-steady, i.e., a_n and a_m are $\varepsilon/2$-close for every $n, m \geq N$. Similarly there exists an $M \geq 1$ such that $(b_n)_{n=M}^{\infty}$ is $\varepsilon/2$-steady, i.e., b_n and b_m are $\varepsilon/2$-close for every $n, m \geq M$.

Let $\max(N, M)$ be the larger of N and M (we know from Proposition 2.2.13 that one has to be greater than or equal to the other). If $n, m \geq \max(N, M)$, then we know that a_n and a_m are $\varepsilon/2$-close, and b_n and b_m are $\varepsilon/2$-close, and so by Proposition 4.3.7 we see that $a_n + b_n$ and $a_m + b_m$ are ε-close for every $n, m \geq \max(N, M)$. This implies that the sequence $(a_n + b_n)_{n=1}^{\infty}$ is eventually ε-steady, as desired. □

The other thing we need to check is the axiom of substitution (see Sect. A.7): if we replace a real number x by another number equal to x, this should not change the sum $x + y$ (and similarly if we substitute y by another number equal to y).

Lemma 5.3.7 (Sums of equivalent Cauchy sequences are equivalent). *Let $x = \mathrm{LIM}_{n\to\infty} a_n$, $y = \mathrm{LIM}_{n\to\infty} b_n$, and $x' = \mathrm{LIM}_{n\to\infty} a'_n$ be real numbers. Suppose that $x = x'$. Then we have $x + y = x' + y$.*

Proof Since x and x' are equal, we know that the Cauchy sequences $(a_n)_{n=1}^{\infty}$ and $(a'_n)_{n=1}^{\infty}$ are equivalent, so in other words they are eventually ε-close for each $\varepsilon > 0$. We need to show that the sequences $(a_n + b_n)_{n=1}^{\infty}$ and $(a'_n + b_n)_{n=1}^{\infty}$ are eventually ε-close for each $\varepsilon > 0$. But we already know that there is an $N \geq 1$ such that $(a_n)_{n=N}^{\infty}$ and $(a'_n)_{n=N}^{\infty}$ are ε-close, i.e., that a_n and a'_n are ε-close for each $n \geq N$. Since b_n is of course 0-close to b_n (where we extend the notion of ε-closeness to the $\varepsilon = 0$ case in the obvious fashion), we thus see from Proposition 4.3.7 (extended to cover the 0-close case) that $a_n + b_n$ and $a'_n + b_n$ are ε-close for each $n \geq N$. This implies that $(a_n + b_n)_{n=1}^{\infty}$ and $(a'_n + b_n)_{n=1}^{\infty}$ are eventually ε-close for each $\varepsilon > 0$, and we are done. □

Remark 5.3.8 The above lemma verifies the axiom of substitution for the "x" variable in $x + y$, but one can similarly prove the axiom of substitution for the "y" variable. (A quick way is to observe from the definition of $x + y$ that we certainly have $x + y = y + x$, since $a_n + b_n = b_n + a_n$.)

We can define multiplication of real numbers in a manner similar to that of addition:

Definition 5.3.9 *(Multiplication of reals).* Let $x = \mathrm{LIM}_{n\to\infty} a_n$ and $y = \mathrm{LIM}_{n\to\infty} b_n$ be real numbers. Then we define the product xy to be $xy := \mathrm{LIM}_{n\to\infty} a_n b_n$.

The following proposition ensures that this definition is valid, and that the product of two real numbers is in fact a real number:

Proposition 5.3.10 (Multiplication is well-defined). *Let $x = \mathrm{LIM}_{n\to\infty} a_n$, $y = \mathrm{LIM}_{n\to\infty} b_n$, and $x' = \mathrm{LIM}_{n\to\infty} a'_n$ be real numbers. Then xy is also a real number. Furthermore, if $x = x'$, then $xy = x'y$.*

Proof See Exercise 5.3.2. □

Of course we can prove a similar substitution rule when y is replaced by a real number y' which is equal to y.

At this point we embed the rationals back into the reals, by equating every rational number q with the real number $\mathrm{LIM}_{n\to\infty} q$. For instance, if $a_1, a_2, a_3, \ldots$ is the sequence

$$0.5, 0.5, 0.5, 0.5, 0.5, \ldots$$

then we set $\mathrm{LIM}_{n\to\infty} a_n$ equal to 0.5. This embedding is consistent with our definitions of addition and multiplication, since for any rational numbers a, b we have

$$(\mathrm{LIM}_{n\to\infty} a) + (\mathrm{LIM}_{n\to\infty} b) = \mathrm{LIM}_{n\to\infty}(a + b) \text{ and}$$
$$(\mathrm{LIM}_{n\to\infty} a) \times (\mathrm{LIM}_{n\to\infty} b) = \mathrm{LIM}_{n\to\infty}(ab);$$

this means that when one wants to add or multiply two rational numbers a, b it does not matter whether one thinks of these numbers as rationals or as the real numbers $\mathrm{LIM}_{n\to\infty} a$, $\mathrm{LIM}_{n\to\infty} b$. Also, this identification of rational numbers and real numbers is consistent with our definitions of equality (Exercise 5.3.3).

We can now easily define negation $-x$ for real numbers x by the formula

$$-x := (-1) \times x,$$

since -1 is a rational number and is hence real. Note that this is clearly consistent with our negation for rational numbers since we have $-q = (-1) \times q$ for all rational numbers q. Also, from our definitions it is clear that

$$-\mathrm{LIM}_{n\to\infty} a_n = \mathrm{LIM}_{n\to\infty}(-a_n)$$

(why?). Once we have addition and negation, we can define subtraction as usual by

$$x - y := x + (-y),$$

note that this implies

$$\mathrm{LIM}_{n\to\infty} a_n - \mathrm{LIM}_{n\to\infty} b_n = \mathrm{LIM}_{n\to\infty}(a_n - b_n).$$

We can now easily show that the real numbers obey all the usual rules of algebra (except perhaps for the laws involving division, which we shall address shortly):

Proposition 5.3.11 *All the laws of algebra from Proposition 4.1.6 hold not only for the integers, but for the reals as well.*

Proof We illustrate this with one such rule: $x(y+z) = xy + xz$. Let $x = \mathrm{LIM}_{n\to\infty} a_n$, $y = \mathrm{LIM}_{n\to\infty} b_n$, and $z = \mathrm{LIM}_{n\to\infty} c_n$ be real numbers. Then by definition, $xy = \mathrm{LIM}_{n\to\infty} a_n b_n$ and $xz = \mathrm{LIM}_{n\to\infty} a_n c_n$, and so $xy + xz = \mathrm{LIM}_{n\to\infty}(a_n b_n + a_n c_n)$. A similar line of reasoning shows that $x(y+z) = \mathrm{LIM}_{n\to\infty} a_n(b_n + c_n)$. But we already know that $a_n(b_n + c_n)$ is equal to $a_n b_n + a_n c_n$ for the rational numbers a_n, b_n, c_n, and the claim follows. The other laws of algebra are proven similarly. □

The last basic arithmetic operation we need to define is reciprocation: $x \to x^{-1}$. This one is a little more subtle. One obvious first guess for how to proceed would be define

$$(\mathrm{LIM}_{n\to\infty} a_n)^{-1} := \mathrm{LIM}_{n\to\infty} a_n^{-1},$$

but there are a few problems with this. For instance, let $a_1, a_2, a_3, \ldots$ be the Cauchy sequence

$$0.1, 0.01, 0.001, 0.0001, \ldots,$$

and let $x := \mathrm{LIM}_{n\to\infty} a_n$. Then by this definition, x^{-1} would be $\mathrm{LIM}_{n\to\infty} b_n$, where $b_1, b_2, b_3, \ldots$ is the sequence

$$10, 100, 1000, 10000, \ldots$$

but this is not a Cauchy sequence (it isn't even bounded). Of course, the problem here is that our original Cauchy sequence $(a_n)_{n=1}^\infty$ was equivalent to the zero sequence $(0)_{n=1}^\infty$ (why?), and hence that our real number x was in fact equal to 0. So we should only allow the operation of reciprocal when x is non-zero.

However, even when we restrict ourselves to non-zero real numbers, we have a slight problem, because a non-zero real number might be the formal limit of a Cauchy sequence which contains zero elements. For instance, the number 1, which is rational and hence real, is the formal limit $1 = \mathrm{LIM}_{n\to\infty} a_n$ of the Cauchy sequence

$$0, 0.9, 0.99, 0.999, 0.9999, \ldots$$

but using our naive definition of reciprocal, we cannot invert the real number 1, because we can't invert the first element 0 of this Cauchy sequence!

To get around these problems we need to keep our Cauchy sequence away from zero. To do this we first need a definition.

Definition 5.3.12 *(Sequences bounded away from zero).* A sequence $(a_n)_{n=1}^\infty$ of rational numbers is said to be *bounded away from zero* iff there exists a rational number $c > 0$ such that $|a_n| \geq c$ for all $n \geq 1$.

Examples 5.3.13 The sequence $1, -1, 1, -1, 1, -1, 1, \ldots$ is bounded away from zero (all the coefficients have absolute value at least 1). But the sequence

0.1, 0.01, 0.001, ... is not bounded away from zero, and neither is 0, 0.9, 0.99, 0.999, 0.9999, ... The sequence 10, 100, 1000, ... is bounded away from zero, but is not bounded.

We now show that every non-zero real number is the formal limit of a Cauchy sequence bounded away from zero:

Lemma 5.3.14 *Let x be a non-zero real number. Then $x = \mathrm{LIM}_{n\to\infty} a_n$ for some Cauchy sequence $(a_n)_{n=1}^{\infty}$ which is bounded away from zero.*

Proof Since x is real, we know that $x = \mathrm{LIM}_{n\to\infty} b_n$ for some Cauchy sequence $(b_n)_{n=1}^{\infty}$. But we are not yet done, because we do not know that b_n is bounded away from zero. On the other hand, we are given that $x \neq 0 = \mathrm{LIM}_{n\to\infty} 0$, which means that the sequence $(b_n)_{n=1}^{\infty}$ is **not** equivalent to $(0)_{n=1}^{\infty}$. Thus the sequence $(b_n)_{n=1}^{\infty}$ cannot be eventually ε-close to $(0)_{n=1}^{\infty}$ for *every* $\varepsilon > 0$. Therefore we can find an $\varepsilon > 0$ such that $(b_n)_{n=1}^{\infty}$ is **not** eventually ε-close to $(0)_{n=1}^{\infty}$.

Let us fix this ε. We know that $(b_n)_{n=1}^{\infty}$ is a Cauchy sequence, so it is eventually ε-steady. Moreover, it is eventually $\varepsilon/2$-steady, since $\varepsilon/2 > 0$. Thus there is an $N \geq 1$ such that $|b_n - b_m| \leq \varepsilon/2$ for all $n, m \geq N$.

On the other hand, we cannot have $|b_n| \leq \varepsilon$ for all $n \geq N$, since this would imply that $(b_n)_{n=1}^{\infty}$ is eventually ε-close to $(0)_{n=1}^{\infty}$. Thus there must be some $n_0 \geq N$ for which $|b_{n_0}| > \varepsilon$. Since we already know that $|b_{n_0} - b_n| \leq \varepsilon/2$ for all $n \geq N$, we thus conclude from the triangle inequality (how?) that $|b_n| \geq \varepsilon/2$ for all $n \geq N$.

This almost proves that $(b_n)_{n=1}^{\infty}$ is bounded away from zero. Actually, what it does is show that $(b_n)_{n=1}^{\infty}$ is *eventually* bounded away from zero. But this is easily fixed, by defining a new sequence a_n, by setting $a_n := \varepsilon/2$ if $n < N$ and $a_n := b_n$ if $n \geq N$. Since b_n is a Cauchy sequence, it is not hard to verify that a_n is also a Cauchy sequence which is equivalent to b_n (because the two sequences are eventually the same), and so $x = \mathrm{LIM}_{n\to\infty} a_n$. And since $|b_n| \geq \varepsilon/2$ for all $n \geq N$, we know that $|a_n| \geq \varepsilon/2$ for all $n \geq 1$ (splitting into the two cases $n \geq N$ and $n < N$ separately). Thus we have a Cauchy sequence which is bounded away from zero (by $\varepsilon/2$ instead of ε, but that's still OK since $\varepsilon/2 > 0$), and which has x as a formal limit, and so we are done. □

Once a sequence is bounded away from zero, we can take its reciprocal without any difficulty:

Lemma 5.3.15 *Suppose that $(a_n)_{n=1}^{\infty}$ is a Cauchy sequence which is bounded away from zero. Then the sequence $(a_n^{-1})_{n=1}^{\infty}$ is also a Cauchy sequence.*

Proof Since $(a_n)_{n=1}^{\infty}$ is bounded away from zero, we know that there is a $c > 0$ such that $|a_n| \geq c$ for all $n \geq 1$. Now we need to show that $(a_n^{-1})_{n=1}^{\infty}$ is eventually ε-steady for each $\varepsilon > 0$. Thus let us fix an $\varepsilon > 0$; our task is now to find an $N \geq 1$ such that $|a_n^{-1} - a_m^{-1}| \leq \varepsilon$ for all $n, m \geq N$. But

$$|a_n^{-1} - a_m^{-1}| = \left|\frac{a_m - a_n}{a_m a_n}\right| \leq \frac{|a_m - a_n|}{c^2}$$

(since $|a_m|, |a_n| \geq c$), and so to make $|a_n^{-1} - a_m^{-1}|$ less than or equal to ε, it will suffice to make $|a_m - a_n|$ less than or equal to $c^2\varepsilon$. But since $(a_n)_{n=1}^{\infty}$ is a Cauchy sequence, and $c^2\varepsilon > 0$, we can certainly find an N such that the sequence $(a_n)_{n=N}^{\infty}$ is $c^2\varepsilon$-steady, i.e., $|a_m - a_n| \leq c^2\varepsilon$ for all $n \geq N$. By what we have said above, this shows that $|a_n^{-1} - a_m^{-1}| \leq \varepsilon$ for all $m, n \geq N$, and hence the sequence $(a_n^{-1})_{n=1}^{\infty}$ is eventually ε-steady. Since we have proven this for every ε, we have that $(a_n^{-1})_{n=1}^{\infty}$ is a Cauchy sequence, as desired. □

We are now ready to define reciprocation:

Definition 5.3.16 *(Reciprocals of real numbers).* Let x be a non-zero real number. Let $(a_n)_{n=1}^{\infty}$ be a Cauchy sequence bounded away from zero such that $x = \mathrm{LIM}_{n\to\infty} a_n$ (such a sequence exists by Lemma 5.3.14). Then we define the reciprocal x^{-1} by the formula $x^{-1} := \mathrm{LIM}_{n\to\infty} a_n^{-1}$. (From Lemma 5.3.15 we know that x^{-1} is a real number.)

We need to check one thing before we are sure this definition makes sense: what if there are two different Cauchy sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ which have x as their formal limit, $x = \mathrm{LIM}_{n\to\infty} a_n = \mathrm{LIM}_{n\to\infty} b_n$. The above definition might conceivably give *two* different reciprocals x^{-1}, namely $\mathrm{LIM}_{n\to\infty} a_n^{-1}$ and $\mathrm{LIM}_{n\to\infty} b_n^{-1}$. Fortunately, this never happens:

Lemma 5.3.17 (Reciprocation is well-defined). *Let* $(a_n)_{n=1}^{\infty}$ *and* $(b_n)_{n=1}^{\infty}$ *be two Cauchy sequences bounded away from zero such that* $\mathrm{LIM}_{n\to\infty} a_n = \mathrm{LIM}_{n\to\infty} b_n$ (*i.e., the two sequences are equivalent*). *Then* $\mathrm{LIM}_{n\to\infty} a_n^{-1} = \mathrm{LIM}_{n\to\infty} b_n^{-1}$.

Proof Consider the following product P of three real numbers:

$$P := (\mathrm{LIM}_{n\to\infty} a_n^{-1}) \times (\mathrm{LIM}_{n\to\infty} a_n) \times (\mathrm{LIM}_{n\to\infty} b_n^{-1}).$$

If we multiply this out, we obtain

$$P = \mathrm{LIM}_{n\to\infty} a_n^{-1} a_n b_n^{-1} = \mathrm{LIM}_{n\to\infty} b_n^{-1}.$$

On the other hand, since $\mathrm{LIM}_{n\to\infty} a_n = \mathrm{LIM}_{n\to\infty} b_n$, we can write P in another way as

$$P = (\mathrm{LIM}_{n\to\infty} a_n^{-1}) \times (\mathrm{LIM}_{n\to\infty} b_n) \times (\mathrm{LIM}_{n\to\infty} b_n^{-1})$$

(cf. Proposition 5.3.10). Multiplying things out again, we get

$$P = \mathrm{LIM}_{n\to\infty} a_n^{-1} b_n b_n^{-1} = \mathrm{LIM}_{n\to\infty} a_n^{-1}.$$

Comparing our different formulae for P we see that $\mathrm{LIM}_{n\to\infty} a_n^{-1} = \mathrm{LIM}_{n\to\infty} b_n^{-1}$, as desired. □

Thus reciprocal is well-defined (for each non-zero real number x, we have exactly one definition of the reciprocal x^{-1}). Note it is clear from the definition that

$xx^{-1} = x^{-1}x = 1$ (why?); thus all the field axioms (Proposition 4.2.4) apply to the reals as well as to the rationals. We of course cannot give 0 a reciprocal, since 0 multiplied by anything gives 0, not 1. Also note that if q is a non-zero rational, and hence equal to the real number $\mathrm{LIM}_{n\to\infty} q$, then the reciprocal of $\mathrm{LIM}_{n\to\infty} q$ is $\mathrm{LIM}_{n\to\infty} q^{-1} = q^{-1}$; thus the operation of reciprocal on real numbers is consistent with the operation of reciprocal on rational numbers.

Once one has reciprocal, one can define division x/y of two real numbers x, y, provided y is non-zero, by the formula

$$x/y := x \times y^{-1},$$

just as we did with the rationals. In particular, we have the *cancelation law*: if x, y, z are real numbers such that $xz = yz$, and z is non-zero, then by dividing by z we conclude that $x = y$. Note that this cancelation law does not work when z is zero.

We now have all four of the basic arithmetic operations on the reals: addition, subtraction, multiplication, and division, with all the usual rules of algebra. Next we turn to the notion of order on the reals.

— Exercises —

Exercise 5.3.1 Prove Proposition 5.3.3. (*Hint:* you may find Proposition 4.3.7 to be useful.)

Exercise 5.3.2 Prove Proposition 5.3.10. (*Hint:* again, Proposition 4.3.7 may be useful.)

Exercise 5.3.3 Let a, b be rational numbers. Show that $a = b$ if and only if $\mathrm{LIM}_{n\to\infty} a = \mathrm{LIM}_{n\to\infty} b$ (i.e., the Cauchy sequences $a, a, a, a, \ldots$ and $b, b, b, b \ldots$ equivalent if and only if $a = b$). This allows us to embed the rational numbers inside the real numbers in a well-defined manner.

Exercise 5.3.4 Let $(a_n)_{n=0}^{\infty}$ be a sequence of rational numbers which is bounded. Let $(b_n)_{n=0}^{\infty}$ be another sequence of rational numbers which is equivalent to $(a_n)_{n=0}^{\infty}$. Show that $(b_n)_{n=0}^{\infty}$ is also bounded. (*Hint:* use Exercise 5.2.2.)

Exercise 5.3.5 Show that $\mathrm{LIM}_{n\to\infty} 1/n = 0$.

5.4 Ordering the Reals

We know that every rational number is positive, negative, or zero. We now want to say the same thing for the reals: each real number should be positive, negative, or zero. Since a real number x is just a formal limit of rationals a_n, it is tempting to make the following definition: a real number $x = \mathrm{LIM}_{n\to\infty} a_n$ is positive if all of the a_n are positive, and negative if all of the a_n are negative (and zero if all of the a_n are zero). However, one soon realizes some problems with this definition. For instance, the sequence $(a_n)_{n=1}^{\infty}$ defined by $a_n := 10^{-n}$, thus

$$0.1, 0.01, 0.001, 0.0001, \ldots$$

consists entirely of positive numbers, but this sequence is equivalent to the zero sequence $0, 0, 0, 0, \ldots$ and thus $\mathrm{LIM}_{n\to\infty} a_n = 0$. Thus even though all the rationals were positive, the real formal limit of these rationals was zero rather than positive. Another example is

$$0.1, -0.01, 0.001, -0.0001, \ldots;$$

this sequence is a hybrid of positive and negative numbers, but again the formal limit is zero.

The trick, as with the reciprocals in the previous section, is to limit one's attention to sequences which are bounded away from zero.

Definition 5.4.1 Let $(a_n)_{n=1}^{\infty}$ be a sequence of rationals. We say that this sequence is *positively bounded away from zero* iff we have a positive rational $c > 0$ such that $a_n \geq c$ for all $n \geq 1$ (in particular, the sequence is entirely positive). The sequence is *negatively bounded away from zero* iff we have a negative rational $-c < 0$ such that $a_n \leq -c$ for all $n \geq 1$ (in particular, the sequence is entirely negative).

Examples 5.4.2 The sequence $1.1, 1.01, 1.001, 1.0001, \ldots$ is positively bounded away from zero (all terms are greater than or equal to 1). The sequence $-1.1, -1.01, -1.001, -1.0001, \ldots$ is negatively bounded away from zero. The sequence $1, -1, 1, -1, 1, -1, \ldots$ is bounded away from zero but is neither positively bounded away from zero nor negatively bounded away from zero.

It is clear that any sequence which is positively or negatively bounded away from zero is bounded away from zero. Also, a sequence cannot be both positively bounded away from zero and negatively bounded away from zero at the same time.

Definition 5.4.3 A real number x is said to be *positive* iff it can be written as $x = \mathrm{LIM}_{n\to\infty} a_n$ for some Cauchy sequence $(a_n)_{n=1}^{\infty}$ which is positively bounded away from zero. x is said to be *negative* iff it can be written as $x = \mathrm{LIM}_{n\to\infty} a_n$ for some sequence $(a_n)_{n=1}^{\infty}$ which is negatively bounded away from zero.

Proposition 5.4.4 (Basic properties of positive reals). *For every real number x, exactly one of the following three statements is true: (a) x is zero; (b) x is positive; (c) x is negative. A real number x is negative if and only if $-x$ is positive. If x and y are positive, then so are $x + y$ and xy.*

Proof See Exercise 5.4.1. □

Note that if q is a positive rational number, then the Cauchy sequence $q, q, q, \ldots$ is positively bounded away from zero, and hence $\mathrm{LIM}_{n\to\infty} q = q$ is a positive real number. Thus the notion of positivity for rationals is consistent with that for reals. Similarly, the notion of negativity for rationals is consistent with that for reals.

Once we have defined positive and negative numbers, we can define absolute value and order.

Definition 5.4.5 *(Absolute value).* Let x be a real number. We define the *absolute value* $|x|$ of x to equal x if x is positive, $-x$ when x is negative, and 0 when x is zero.

Definition 5.4.6 *(Ordering of the real numbers).* Let x and y be real numbers. We say that x is *greater than* y, and write $x > y$, iff $x - y$ is a positive real number, and $x < y$ iff $x - y$ is a negative real number. We define $x \geq y$ iff $x > y$ or $x = y$, and similarly define $x \leq y$.

Comparing this with the definition of order on the rationals from Definition 4.2.8 we see that order on the reals is consistent with order on the rationals, i.e., if two rational numbers q, q' are such that q is less than q' in the rational number system, then q is still less than q' in the real number system, and similarly for "greater than". In the same way we see that the definition of absolute value given here is consistent with that in Definition 4.3.1.

Proposition 5.4.7 *All the claims in Proposition 4.2.9 which held for rationals continue to hold for real numbers.*

Proof We just prove one of the claims and leave the rest to Exercise 5.4.2. Suppose we have $x < y$ and z a positive real, and want to conclude that $xz < yz$. Since $x < y$, $y - x$ is positive, hence by Proposition 5.4.4 we have $(y - x)z = yz - xz$ is positive, hence $xz < yz$. □

As an application of these propositions, we prove

Proposition 5.4.8 *Let x be a positive real number. Then x^{-1} is also positive. Also, if y is another positive number and $x > y$, then $x^{-1} < y^{-1}$.*

Proof Let x be positive. Since $xx^{-1} = 1$, the real number x^{-1} cannot be zero (since $x0 = 0 \neq 1$). Also, from Proposition 5.4.4 it is easy to see that a positive number times a negative number is negative; this shows that x^{-1} cannot be negative, since this would imply that $xx^{-1} = 1$ is negative, a contradiction. Thus, by Proposition 5.4.4, the only possibility left is that x^{-1} is positive.

Now let y be positive as well, so x^{-1} and y^{-1} are also positive. Suppose that $x > y$. If $x^{-1} \geq y^{-1}$, then by Proposition 5.4.7 we have $xx^{-1} > yx^{-1} \geq yy^{-1}$, thus $1 > 1$, which is a contradiction. Thus we must have $x^{-1} < y^{-1}$. □

Another application is that the laws of exponentiation (Proposition 4.3.12) that were previously proven for rationals, are also true for reals; see Sect. 5.6.

We have already seen that the formal limit of positive rationals need not be positive; it could be zero, as the example $0.1, 0.01, 0.001, \ldots$ showed. However, the formal limit of *non-negative* rationals (i.e., rationals that are either positive or zero) is non-negative.

Proposition 5.4.9 *Let $a_1, a_2, a_3, \ldots$ be a Cauchy sequence of non-negative rational numbers. Then $\mathrm{LIM}_{n\to\infty} a_n$ is a non-negative real number.*

Eventually, we will see a better explanation of this fact: the set of non-negative reals is *closed*, whereas the set of positive reals is *open*. See Sect. 1.2.

Proof We argue by contradiction, and suppose that the real number $x := \mathrm{LIM}_{n\to\infty} a_n$ is a negative number. Then by definition of negative real number, we have $x = \mathrm{LIM}_{n\to\infty} b_n$ for some sequence b_n which is negatively bounded away from zero, i.e., there is a negative rational $-c < 0$ such that $b_n \leq -c$ for all $n \geq 1$. On the other hand, we have $a_n \geq 0$ for all $n \geq 1$, by hypothesis. Thus the numbers a_n and b_n are never $c/2$-close, since $c/2 < c$. Thus the sequences $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are not eventually $c/2$-close. Since $c/2 > 0$, this implies that $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are not equivalent. But this contradicts the fact that both these sequences have x as their formal limit. □

Corollary 5.4.10 *Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be Cauchy sequences of rationals such that $a_n \geq b_n$ for all $n \geq 1$. Then $\mathrm{LIM}_{n\to\infty} a_n \geq \mathrm{LIM}_{n\to\infty} b_n$.*

Proof Apply Proposition 5.4.9 to the sequence $a_n - b_n$. □

Remark 5.4.11 Note that the above corollary does not work if the $\geq$ signs are replaced by $>$: for instance if $a_n := 1 + 1/n$ and $b_n := 1 - 1/n$, then a_n is always strictly greater than b_n, but the formal limit of a_n is not greater than the formal limit of b_n, instead they are equal.

We now define distance $d(x, y) := |x - y|$ just as we did for the rationals. In fact, Propositions 4.3.3 and 4.3.7 hold not only for the rationals, but for the reals; the proof is identical, since the real numbers obey all the laws of algebra and order that the rationals do.

We now observe that while positive real numbers can be arbitrarily large or small, they cannot be larger than all of the positive integers, or smaller in magnitude than all of the positive rationals:

Proposition 5.4.12 (Bounding of reals by rationals). *Let x be a positive real number. Then there exists a positive rational number q such that $q \leq x$, and there exists a positive integer N such that $x \leq N$.*

Proof Since x is a positive real, it is the formal limit of some Cauchy sequence $(a_n)_{n=1}^{\infty}$ which is positively bounded away from zero. Also, by Lemma 5.1.15, this sequence is bounded. Thus we have rationals $q > 0$ and r such that $q \leq a_n \leq r$ for all $n \geq 1$. But by Proposition 4.4.1 we know that there is some integer N such that $r \leq N$; since q is positive and $q \leq r \leq N$, we see that N is positive. Thus $q \leq a_n \leq N$ for all $n \geq 1$. Applying Corollary 5.4.10 we obtain that $q \leq x \leq N$, as desired. □

Corollary 5.4.13 (Archimedean property). *Let x be a real number, and let ε be a positive real number. Then there exists a positive integer M such that $M\varepsilon > x$.*

Proof If x is zero or negative, one can just take $M = 1$, so suppose that x is positive. Then the number x/ε is positive, and hence by Proposition 5.4.12 there exists a positive integer N such that $x/\varepsilon \leq N$. If we set $M := N + 1$, then $x/\varepsilon < M$. Now multiply by ε. □

This property is quite important; it says that no matter how large x is and how small ε is, if one keeps adding ε to itself, one will eventually overtake x.

Proposition 5.4.14 *Given any two real numbers $x < y$, we can find a rational number q such that $x < q < y$.*

Proof See Exercise 5.4.5. □

We have now completed our construction of the real numbers. This number system contains the rationals and has almost everything that the rational number system has: the arithmetic operations, the laws of algebra, the laws of order. However, we have not yet demonstrated any *advantages* that the real numbers have over the rationals; so far, even after much effort, all we have done is shown that they are *at least as good as* the rational number system. But in the next few sections we show that the real numbers can do more things than rationals: for example, we can take square roots in a real number system.

Remark 5.4.15 Up until now, we have not addressed the fact that real numbers can be expressed using the decimal system. For instance, the formal limit of

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

is more conventionally represented as the decimal $1.41421\ldots$. We will address this in an Appendix (B), but for now let us just remark that there are some subtleties in the decimal system, for instance $0.9999\ldots$ and $1.000\ldots$ are in fact the same real number.

— Exercises —

Exercise 5.4.1 Prove Proposition 5.4.4. (*Hint:* if x is not zero, and x is the formal limit of some sequence $(a_n)_{n=1}^\infty$, then this sequence cannot be eventually ε-close to the zero sequence $(0)_{n=1}^\infty$ for every single $\varepsilon > 0$. Use this to show that the sequence $(a_n)_{n=1}^\infty$ is eventually either positively bounded away from zero or negatively bounded away from zero.)

Exercise 5.4.2 Prove the remaining claims in Proposition 5.4.7.

Exercise 5.4.3 Show that for every real number x there is exactly one integer N such that $N \leq x < N + 1$. (This integer N is called the *integer part* of x and is sometimes denoted $N = \lfloor x \rfloor$.)

Exercise 5.4.4 Show that for any positive real number $x > 0$ there exists a positive integer N such that $x > 1/N > 0$.

Exercise 5.4.5 Prove Proposition 5.4.14. (*Hint:* use Exercise 5.4.4. You may also need to argue by contradiction.)

Exercise 5.4.6 Let x, y be real numbers and let $\varepsilon > 0$ be a positive real. Show that $|x - y| < \varepsilon$ if and only if $y - \varepsilon < x < y + \varepsilon$, and that $|x - y| \leq \varepsilon$ if and only if $y - \varepsilon \leq x \leq y + \varepsilon$.

Exercise 5.4.7 Let x and y be real numbers. Show that $x \leq y + \varepsilon$ for all real numbers $\varepsilon > 0$ if and only if $x \leq y$. Show that $|x - y| \leq \varepsilon$ for all real numbers $\varepsilon > 0$ if and only if $x = y$.

Exercise 5.4.8 Let $(a_n)_{n=1}^{\infty}$ be a Cauchy sequence of rationals, and let x be a real number. Show that if $a_n \leq x$ for all $n \geq 1$, then $\mathrm{LIM}_{n\to\infty} a_n \leq x$. Similarly, show that if $a_n \geq x$ for all $n \geq 1$, then $\mathrm{LIM}_{n\to\infty} a_n \geq x$. (*Hint:* prove by contradiction. Use Proposition 5.4.14 to find a rational between $\mathrm{LIM}_{n\to\infty} a_n$ and x, and then use Proposition 5.4.9 or Corollary 5.4.10.)

Exercise 5.4.9 If x, y are real numbers, define the *maximum* $\max(x, y)$ of x and y to equal x if $x \geq y$, and y if $x < y$. Similarly, define the *minimum* $\min(x, y)$ of x and y to equal x if $x \leq y$, and y if $x > y$.

(i) If x, y are real numbers, show that $\max(x, y) = -\min(-x, -y)$ and $\min(x, y) = -\max(-x, -y)$.
(ii) x, y, z are real numbers, show that $\max(x, y) = \max(y, x)$, $\max(x, x) = x$, and $\max(x + z, y + z) = \max(x, y) + z$. If z is non-negative, show that $\max(xz, yz) = z\max(x, y)$. What happens to the last claim if z is negative?
(iii) Show that all the claims in (ii) also hold if max is replaced with min.
(iv) If x, y are positive real numbers, show that $\max(x, y)^{-1} = \min(x^{-1}, y^{-1})$ and $\min(x, y)^{-1} = \max(x^{-1}, y^{-1})$.

5.5 The Least Upper Bound Property

We now give one of the most basic advantages of the real numbers over the rationals; one can take the *least upper bound* $\sup(E)$ of any (non-empty, upper-bounded) subset E of the real numbers $\mathbf{R}$.

Definition 5.5.1 *(Upper bound).* Let E be a subset of $\mathbf{R}$, and let M be a real number. We say that M is an *upper bound* for E, iff we have $x \leq M$ for every element x in E.

Example 5.5.2 Let E be the interval $E := \{x \in \mathbf{R} : 0 \leq x \leq 1\}$. Then 1 is an upper bound for E, since every element of E is less than or equal to 1. It is also true that 2 is an upper bound for E, and indeed every number greater or equal to 1 is an upper bound for E. On the other hand, any other number, such as 0.5, is not an upper bound, because 0.5 is not larger than *every* element in E. (Merely being larger than *some* elements of E is not necessarily enough to make 0.5 an upper bound.)

Example 5.5.3 Let $\mathbf{R}^+$ be the set of positive reals: $\mathbf{R}^+ := \{x \in \mathbf{R} : x > 0\}$. Then $\mathbf{R}^+$ does not have any upper bounds[3] at all (why?).

Example 5.5.4 Let $\emptyset$ be the empty set. Then every number M is an upper bound for $\emptyset$, because M is greater than every element of the empty set (this is a vacuously true statement, but still true).

It is clear that if M is an upper bound of E, then any larger number $M' \geq M$ is also an upper bound of E. On the other hand, it is not so clear whether it is also possible for any number smaller than M to also be an upper bound of E. This motivates the following definition:

[3] More precisely, $\mathbf{R}^+$ has no upper bounds which are real numbers. In Sect. 6.2 we shall introduce the *extended real number system* $\mathbf{R}^*$, which allows one to give the upper bound of $+\infty$ for sets such as $\mathbf{R}^+$.

Definition 5.5.5 *(Least upper bound).* Let E be a subset of $\mathbf{R}$, and M be a real number. We say that M is a *least upper bound* for E iff (a) M is an upper bound for E, and also (b) any other upper bound M' for E must be larger than or equal to M.

Example 5.5.6 Let E be the interval $E := \{x \in \mathbf{R} : 0 \leq x \leq 1\}$. Then, as noted before, E has many upper bounds, indeed every number greater than or equal to 1 is an upper bound. But only 1 is the *least* upper bound; all other upper bounds are larger than 1.

Example 5.5.7 The empty set does not have a least upper bound (why?).

Proposition 5.5.8 (Uniqueness of least upper bound). *Let E be a subset of* $\mathbf{R}$. *Then E can have at most one least upper bound.*

Proof Let M_1 and M_2 be two least upper bounds, say M_1 and M_2. Since M_1 is a least upper bound and M_2 is an upper bound, then by definition of least upper bound we have $M_2 \geq M_1$. Since M_2 is a least upper bound and M_1 is an upper bound, we similarly have $M_1 \geq M_2$. Thus $M_1 = M_2$. Thus there is at most one least upper bound. □

Now we come to an important property of the real numbers:

Theorem 5.5.9 (Existence of least upper bound). *Let E be a non-empty subset of* $\mathbf{R}$. *If E has an upper bound, (i.e., E has some upper bound M), then it must have exactly one* least *upper bound.*

Proof This theorem will take quite a bit of effort to prove, and many of the steps will be left as exercises.

Let E be a non-empty subset of $\mathbf{R}$ with an upper bound M. By Proposition 5.5.8, we know that E has at most one least upper bound; we have to show that E has at least one least upper bound. Since E is non-empty, we can choose some element x_0 in E.

Let $n \geq 1$ be a positive integer. We know that E has an upper bound M. By the Archimedean property (Corollary 5.4.13), we can find an integer K such that $K/n \geq M$, and hence K/n is also an upper bound for E. By the Archimedean property again, there exists another integer L such that $L/n < x_0$. Since x_0 lies in E, we see that L/n is *not* an upper bound for E. Since K/n is an upper bound but L/n is not, we see that $K \geq L$.

Since K/n is an upper bound for E and L/n is not, we can find an integer $L < m_n \leq K$ with the property that m_n/n is an upper bound for E, but $(m_n - 1)/n$ is not (see Exercise 5.5.2). In fact, this integer m_n is unique (Exercise 5.5.3). We subscript m_n by n to emphasize the fact that this integer m depends on the choice of n. This gives a well-defined (and unique) sequence $m_1, m_2, m_3, \ldots$ of integers, with each of the m_n/n being upper bounds and each of the $(m_n - 1)/n$ not being upper bounds.

Now let $N \geq 1$ be a positive integer, and let $n, n' \geq N$ be integers larger than or equal to N. Since m_n/n is an upper bound for E and $(m_{n'} - 1)/n'$ is not, we must have $m_n/n > (m_{n'} - 1)/n'$ (why?). After a little algebra, this implies that

$$\frac{m_n}{n} - \frac{m_{n'}}{n'} > -\frac{1}{n'} \geq -\frac{1}{N}.$$

Similarly, since $m_{n'}/n'$ is an upper bound for E and $(m_n - 1)/n$ is not, we have $m_{n'}/n' > (m_n - 1)/n$, and hence

$$\frac{m_n}{n} - \frac{m_{n'}}{n'} \leq \frac{1}{n} \leq \frac{1}{N}.$$

Putting these two bounds together, we see that

$$\left|\frac{m_n}{n} - \frac{m_{n'}}{n'}\right| \leq \frac{1}{N} \text{ for all } n, n' \geq N \geq 1.$$

This implies that $\frac{m_n}{n}$ is a Cauchy sequence (Exercise 5.5.4). Since the $\frac{m_n}{n}$ are rational numbers, we can now define the real number S as

$$S := \mathrm{LIM}_{n\to\infty} \frac{m_n}{n}.$$

From Exercise 5.3.5 we conclude that

$$S = \mathrm{LIM}_{n\to\infty} \frac{m_n - 1}{n}.$$

To finish the proof of the theorem, we need to show that S is the least upper bound for E. First we show that it is an upper bound. Let x be any element of E. Then, since m_n/n is an upper bound for E, we have $x \leq m_n/n$ for all $n \geq 1$. Applying Exercise 5.4.8, we conclude that $x \leq \mathrm{LIM}_{n\to\infty} m_n/n = S$. Thus S is indeed an upper bound for E.

Now we show it is a least upper bound. Suppose y is an upper bound for E. Since $(m_n - 1)/n$ is not an upper bound, we conclude that $y \geq (m_n - 1)/n$ for all $n \geq 1$. Applying Exercise 5.4.8, we conclude that $y \geq \mathrm{LIM}_{n\to\infty}(m_n - 1)/n = S$. Thus the upper bound S is less than or equal to every upper bound of E, and S is thus a least upper bound of E. □

Definition 5.5.10 *(Supremum).* Let E be a subset of the real numbers. If E is non-empty and has some upper bound, we define $\sup(E)$ to be the least upper bound of E (this is well-defined by Theorem 5.5.9). We introduce two additional symbols, $+\infty$ and $-\infty$. If E is non-empty and has no upper bound, we set $\sup(E) := +\infty$; if E is empty, we set $\sup(E) := -\infty$. We refer to $\sup(E)$ as the *supremum* of E, and also denote it by $\sup E$.

Remark 5.5.11 At present, $+\infty$ and $-\infty$ are meaningless symbols; we have no operations on them at present, and none of our results involving real numbers apply to $+\infty$ and $-\infty$, because these are not real numbers. In Sect. 6.2 we add $+\infty$ and $-\infty$ to the reals to form the *extended real number system*, but this system is not as convenient to work with as the real number system, because many of the laws of algebra break down. For instance, it is not a good idea to try to define $+\infty + -\infty$; setting this equal to 0 causes some problems.

Now we give an example of how the least upper bound property is useful.

Proposition 5.5.12 *There exists a positive real number x such that $x^2 = 2$.*

Remark 5.5.13 Comparing this result with Proposition 4.4.4, we see that certain numbers are real but not rational. The proof of this proposition also shows that the rationals $\mathbf{Q}$ do not obey the least upper bound property, otherwise one could use that property to construct a square root of 2, which by Proposition 4.4.4 is not possible.

Proof Let E be the set $\{y \in \mathbf{R} : y \geq 0 \text{ and } y^2 < 2\}$; thus E is the set of all non-negative real numbers whose square is less than 2. Observe that E has an upper bound of 2 (because if $y > 2$, then $y^2 > 4 > 2$ and hence $y \notin E$). Also, E is non-empty (for instance, 1 is an element of E). Thus by the least upper bound property, we have a real number $x := \sup(E)$ which is the least upper bound of E. Then x is greater than or equal to 1 (since $1 \in E$) and less than or equal to 2 (since 2 is an upper bound for E). So x is positive. Now we show that $x^2 = 2$.

We argue this by contradiction. We show that both $x^2 < 2$ and $x^2 > 2$ lead to contradictions. First suppose that $x^2 < 2$. Let $0 < \varepsilon < 1$ be a small number; then we have

$$(x + \varepsilon)^2 = x^2 + 2\varepsilon x + \varepsilon^2 \leq x^2 + 4\varepsilon + \varepsilon = x^2 + 5\varepsilon$$

since $x \leq 2$ and $\varepsilon^2 \leq \varepsilon$. Since $x^2 < 2$, we see that we can choose an $0 < \varepsilon < 1$ such that $x^2 + 5\varepsilon < 2$, thus $(x + \varepsilon)^2 < 2$. By construction of E, this means that $x + \varepsilon \in E$; but this contradicts the fact that x is an upper bound of E.

Now suppose that $x^2 > 2$. Let $0 < \varepsilon < 1$ be a small number; then we have

$$(x - \varepsilon)^2 = x^2 - 2\varepsilon x + \varepsilon^2 \geq x^2 - 2\varepsilon x \geq x^2 - 4\varepsilon$$

since $x \leq 2$ and $\varepsilon^2 \geq 0$. Since $x^2 > 2$, we can choose $0 < \varepsilon < 1$ such that $x^2 - 4\varepsilon > 2$, and thus $(x - \varepsilon)^2 > 2$. But then this implies that $x - \varepsilon \geq y$ for all $y \in E$. (Why? If $x - \varepsilon < y$ then $(x - \varepsilon)^2 < y^2 \leq 2$, a contradiction.) Thus $x - \varepsilon$ is an upper bound for E, which contradicts the fact that x is the *least* upper bound of E. From these two contradictions we see that $x^2 = 2$, as desired. □

Remark 5.5.14 In Chap. 6 we will use the least upper bound property to develop the theory of limits, which allows us to do many more things than just take square roots.

Remark 5.5.15 We can of course talk about lower bounds and greatest lower bounds, of sets E; the greatest lower bound of a set E is also known as the *infimum*[4] of E and

[4] Supremum means "highest" and infimum means "lowest", and the plurals are suprema and infima. Supremum is to superior, and infimum to inferior, as maximum is to major, and minimum to minor. The root words are "super", which means "above", and "infer", which means "below" (this usage only survives in a few rare English words such as "infernal", with the Latin prefix "sub" having mostly replaced "infer" in English).

is denoted $\inf(E)$ or $\inf E$. Everything we say about suprema has a counterpart for infima; we will usually leave such statements to the reader. A precise relationship between the two notions is given by Exercise 5.5.1. See also Sect. 6.2.

— Exercises —

Exercise 5.5.1 Let E be a subset of the real numbers $\mathbf{R}$, and suppose that E has a least upper bound M which is a real number, i.e., $M = \sup(E)$. Let $-E$ be the set

$$-E := \{-x : x \in E\}.$$

Show that $-M$ is the greatest lower bound of $-E$, i.e., $-M = \inf(-E)$.

Exercise 5.5.2 Let E be a non-empty subset of $\mathbf{R}$, let $n \geq 1$ be an integer, and let $L < K$ be integers. Suppose that K/n is an upper bound for E, but that L/n is not an upper bound for E. Without using Theorem 5.5.9, show that there exists an integer $L < m \leq K$ such that m/n is an upper bound for E, but that $(m-1)/n$ is not an upper bound for E. (*Hint:* prove by contradiction, and use induction. It may also help to draw a picture of the situation.)

Exercise 5.5.3 Let E be a non-empty subset of $\mathbf{R}$, let $n \geq 1$ be an integer, and let m, m' be integers with the properties that m/n and m'/n are upper bounds for E, but $(m-1)/n$ and $(m'-1)/n$ are not upper bounds for E. Show that $m = m'$. This shows that the integer m constructed in Exercise 5.5.2 is unique. (*Hint:* again, drawing a picture will be helpful.)

Exercise 5.5.4 Let $q_1, q_2, q_3, \ldots$ be a sequence of rational numbers with the property that $|q_n - q_{n'}| \leq \frac{1}{M}$ whenever $M \geq 1$ is an integer and $n, n' \geq M$. Show that $q_1, q_2, q_3, \ldots$ is a Cauchy sequence. Furthermore, if $S := \mathrm{LIM}_{n\to\infty} q_n$, show that $|q_M - S| \leq \frac{1}{M}$ for every $M \geq 1$. (*Hint:* use Corollary 5.4.10 or Exercise 5.4.8.)

Exercise 5.5.5 Establish an analogue of Proposition 5.4.14, in which “rational” is replaced by “irrational”.

5.6 Real Exponentiation, Part I

In Sect. 4.3 we defined exponentiation x^n when x is rational and n is a natural number, or when x is a non-zero rational and n is an integer. Now that we have all the arithmetic operations on the reals (and Proposition 5.4.7 assures us that the arithmetic properties of the rationals that we are used to, continue to hold for the reals) we can similarly define exponentiation of the reals.

Definition 5.6.1 *(Exponentiating a real by a natural number).* Let x be a real number. To raise x to the power 0, we define $x^0 := 1$. Now suppose recursively that x^n has been defined for some natural number n, then we define $x^{n+1} := x^n \times x$.

Definition 5.6.2 *(Exponentiating a real by an integer).* Let x be a non-zero real number. Then for any negative integer $-n$, we define $x^{-n} := 1/x^n$.

Clearly these definitions are consistent with the definition of rational exponentiation given earlier. We can then assert

Proposition 5.6.3 *All the properties in Propositions 4.3.10 and 4.3.12 remain valid if x and y are assumed to be real numbers instead of rational numbers.*

Instead of giving an actual proof of this proposition, we shall give a meta-proof (an argument appealing to the nature of proofs, rather than the nature of real and rational numbers).

Meta-proof. If one inspects the proof of Propositions 4.3.10 and 4.3.12 we see that they rely on the laws of algebra and the laws of order for the rationals (Propositions 4.2.4 and 4.2.9). But by Propositions 5.3.11 and 5.4.7, and the identity $xx^{-1} = x^{-1}x = 1$ we know that all these laws of algebra and order continue to hold for real numbers as well as rationals. Thus we can modify the proof of Proposition 4.3.10 and 4.3.12 to hold in the case when x and y are real. □

Now we consider exponentiation to exponents which are not integers. We begin with the notion of an nth root, which we can define using our notion of supremum.

Definition 5.6.4 Let $x \geq 0$ be a non-negative real, and let $n \geq 1$ be a positive integer. We define $x^{1/n}$, also known as the nth *root of* x, by the formula

$$x^{1/n} := \sup\{y \in \mathbf{R} : y \geq 0 \text{ and } y^n \leq x\}.$$

We often write $\sqrt{x}$ for $x^{1/2}$.

Note we do not define the nth root of a negative number. In fact, we will leave the nth roots of negative numbers undefined for the rest of the text (one can define these nth roots once one defines the complex numbers, but we shall refrain from doing so).

Lemma 5.6.5 (Existence of nth roots). *Let $x \geq 0$ be a non-negative real, and let $n \geq 1$ be a positive integer. Then the set $E := \{y \in \mathbf{R} : y \geq 0$ and $y^n \leq x\}$ is non-empty and is also bounded above. In particular, $x^{1/n}$ is a real number.*

Proof The set E contains 0 (why?), so it is certainly not empty. Now we show it has an upper bound. We divide into two cases: $x \leq 1$ and $x > 1$. First suppose that we are in the case where $x \leq 1$. Then we claim that the set E is bounded above by 1. To see this, suppose for sake of contradiction that there was an element $y \in E$ for which $y > 1$. But then $y^n > 1$ (why?), and hence $y^n > x$, a contradiction. Thus E has an upper bound. Now suppose that we are in the case where $x > 1$. Then we claim that the set E is bounded above by x. To see this, suppose for contradiction that there was an element $y \in E$ for which $y > x$. Since $x > 1$, we thus have $y > 1$. Since $y > x$ and $y > 1$, we have $y^n > x$ (why?), a contradiction. Thus in both cases E has an upper bound, and so $x^{1/n}$ is finite. □

We list some basic properties of nth roots below.

Lemma 5.6.6 *Let $x, y \geq 0$ be non-negative reals, and let $n, m \geq 1$ be positive integers.*

(*a*) *If* $y = x^{1/n}$, *then* $y^n = x$.
(*b*) *Conversely, if* $y^n = x$, *then* $y = x^{1/n}$.
(*c*) $x^{1/n}$ *is a non-negative real number, and is positive iff* x *is positive.*
(*d*) *We have* $x > y$ *if and only if* $x^{1/n} > y^{1/n}$.
(*e*) *If* $x > 1$, *then* $x^{1/k}$ *is a decreasing function of* k, *where* k *ranges over the positive integers; that is to say,* $x^{1/k} < x^{1/l}$ *whenever* $k > l$. *If* $0 < x < 1$, *then* $x^{1/k}$ *is an increasing function of* k *(i.e.,* $x^{1/k} > x^{1/l}$ *whenever* $k > l$*). If* $x = 1$, *then* $x^{1/k} = 1$ *for all* k.
(*f*) *We have* $(xy)^{1/n} = x^{1/n}y^{1/n}$.
(*g*) *We have* $(x^{1/n})^{1/m} = x^{1/nm}$.

Proof See Exercise 5.6.1. □

The observant reader may note that this definition of $x^{1/n}$ might possibly be inconsistent with our previous notion of x^n when $n = 1$, but it is easy to check that $x^{1/1} = x = x^1$ (why?), so there is no inconsistency.

One consequence of Lemma 5.6.6(b) is another proof of the cancelation law from Proposition 4.3.12(c) and Proposition 5.6.3: if y and z are positive and $y^n = z^n$, then $y = z$. (Why does this follow from Lemma 5.6.6(b)?) Note that this only works when y and z are positive; for instance, $(-3)^2 = 3^2$, but we cannot conclude from this that $-3 = 3$.

Now we define how to raise a positive number x to a *rational* exponent q.

Definition 5.6.7 Let $x > 0$ be a positive real number, and let q be a rational number. To define x^q, we write $q = a/b$ for some integer a and positive integer b, and define

$$x^q := (x^{1/b})^a.$$

Note that every rational q, whether positive, negative, or zero, can be written in the form a/b where a is an integer and b is positive (why?). However, the rational number q can be expressed in the form a/b in more than one way, for instance $1/2$ can also be expressed as $2/4$ or $3/6$. So to ensure that this definition is well-defined, we need to check that different expressions a/b give the same formula for x^q:

Lemma 5.6.8 *Let* a, a' *be integers and* b, b' *be positive integers such that* $a/b = a'/b'$, *and let* x *be a positive real number. Then we have* $(x^{1/b'})^{a'} = (x^{1/b})^a$.

Proof There are three cases: $a = 0$, $a > 0$, $a < 0$. If $a = 0$, then we must have $a' = 0$ (why?) and so both $(x^{1/b'})^{a'}$ and $(x^{1/b})^a$ are equal to 1, so we are done.

Now suppose that $a > 0$. Then $a' > 0$ (why?), and $ab' = ba'$. Write $y := x^{1/(ab')} = x^{1/(ba')}$. By Lemma 5.6.6(g) we have $y = (x^{1/b'})^{1/a}$ and $y = (x^{1/b})^{1/a'}$; by Lemma 5.6.6(a) we thus have $y^a = x^{1/b'}$ and $y^{a'} = x^{1/b}$. Thus we have

$$(x^{1/b'})^{a'} = (y^a)^{a'} = y^{aa'} = (y^{a'})^a = (x^{1/b})^a$$

as desired.

Finally, suppose that $a < 0$. Then we have $(-a)/b = (-a')/b'$. But $-a$ is positive, so the previous case applies and we have $(x^{1/b'})^{-a'} = (x^{1/b})^{-a}$. Taking the reciprocal of both sides we obtain the result. □

Thus x^q is well-defined for every rational q. Note that this new definition is consistent with our old definition for $x^{1/n}$ (why?) and is also consistent with our old definition for x^n (why?).

Some basic facts about rational exponentiation:

Lemma 5.6.9 *Let $x, y > 0$ be positive reals, and let q, r be rationals.*

(*a*) *x^q is a positive real.*
(*b*) *$x^{q+r} = x^q x^r$ and $(x^q)^r = x^{qr}$.*
(*c*) *$x^{-q} = 1/x^q$.*
(*d*) *If $q > 0$, then $x > y$ if and only if $x^q > y^q$.*
(*e*) *If $x > 1$, then $x^q > x^r$ if and only if $q > r$. If $x < 1$, then $x^q > x^r$ if and only if $q < r$.*
(*f*) *$(xy)^q = x^q y^q$.*

Proof See Exercise 5.6.2. □

We still have to do real exponentiation; in other words, we still have to define x^y where $x > 0$ and y is a real number—but we will defer that until Sect. 6.7, once we have formalized the concept of limit.

In the rest of the text we shall now just assume the real numbers to obey all the usual laws of algebra, order, and exponentiation.

— Exercises —

Exercise 5.6.1 Prove Lemma 5.6.6. (*Hints:* review the proof of Proposition 5.5.12. Also, you will find proof by contradiction a useful tool, especially when combined with the trichotomy of order in Proposition 5.4.7 and Proposition 5.4.12. The earlier parts of the lemma can be used to prove later parts of the lemma. With part (e), first show that if $x > 1$ then $x^{1/n} > 1$, and if $x < 1$ then $x^{1/n} < 1$.)

Exercise 5.6.2 Prove Lemma 5.6.9. (*Hint:* you should rely mainly on Lemma 5.6.6 and on algebra.)

Exercise 5.6.3 If x is a real number and n is an even natural number (thus $n = 2m$ for some natural number m), show that $x^n \geq 0$.

Exercise 5.6.4 If x is a real number, show that $|x| = (x^2)^{1/2}$.

Exercise 5.6.5 If x, y are positive reals, and q is a positive rational with $q \geq 1$, show that $\max(x^q, y^q) = \max(x, y)^q$ and $\min(x^q, y^q) = \min(x, y)^q$, where the operations min, max were defined in Exercise 5.4.9. What happens if we have $q < 1$ instead of $q \geq 1$?

Chapter 6
Limits of Sequences

6.1 Convergence and Limit Laws

In the previous chapter, we defined the real numbers as formal limits of rational (Cauchy) sequences, and we then defined various operations on the real numbers. However, unlike our work in constructing the integers (where we eventually replaced formal differences with actual differences) and rationals (where we eventually replaced formal quotients with actual quotients), we did not completely finish the job of constructing the real numbers, because we never got around to replacing formal limits $\mathrm{LIM}_{n\to\infty} a_n$ with actual limits $\lim_{n\to\infty} a_n$. In fact, we haven't defined limits at all yet. This will now be rectified.

We begin by repeating much of the machinery of ε-close sequences, etc., again—but this time, we do it for sequences of *real* numbers, not rational numbers. Thus this discussion will supercede what we did in the previous chapter. First, we define distance for real numbers:

Definition 6.1.1 *(Distance between two real numbers).* Given two real numbers x and y, we define their distance $d(x, y)$ to be $d(x, y) := |x - y|$.

Clearly this definition is consistent with Definition 4.3.2. Further, Proposition 4.3.3 works just as well for real numbers as it does for rationals, because the real numbers obey all the rules of algebra that the rationals do.

Definition 6.1.2 *(ε-close real numbers).* Let $\varepsilon > 0$ be a real number. We say that two real numbers x, y are *ε-close* iff we have $d(y, x) \le \varepsilon$.

Again, it is clear that this definition of ε-close is consistent with Definition 4.3.4.

Now let $(a_n)_{n=m}^{\infty}$ be a sequence of *real* numbers; i.e., we assign a real number a_n for every integer $n \ge m$. The starting index m is some integer; usually this will be 1, but in some cases we will start from some index other than 1. (The choice of label used to index this sequence is unimportant; we could use for instance $(a_k)_{k=m}^{\infty}$ and this would represent exactly the same sequence as $(a_n)_{n=m}^{\infty}$.) We can define the notion of a Cauchy sequence in the same manner as before.

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_6

Definition 6.1.3 *(Cauchy sequences of reals).* Let $\varepsilon > 0$ be a real number. A sequence $(a_n)_{n=N}^{\infty}$ of real numbers starting at some integer index N is said to be *ε-steady* iff a_j and a_k are ε-close for every $j, k \geq N$. A sequence $(a_n)_{n=m}^{\infty}$ starting at some integer index m is said to be *eventually ε-steady* iff there exists an $N \geq m$ such that $(a_n)_{n=N}^{\infty}$ is ε-steady. We say that $(a_n)_{n=m}^{\infty}$ is a *Cauchy sequence* iff it is eventually ε-steady for every $\varepsilon > 0$.

To put it another way, a sequence $(a_n)_{n=m}^{\infty}$ of real numbers is a Cauchy sequence if, for every real $\varepsilon > 0$, there exists an $N \geq m$ such that $|a_n - a_{n'}| \leq \varepsilon$ for all $n, n' \geq N$. These definitions are consistent with the corresponding definitions for rational numbers (Definitions 5.1.3, 5.1.6, 5.1.8), although verifying consistency for Cauchy sequences takes a little bit of care.

Proposition 6.1.4 *Let $(a_n)_{n=m}^{\infty}$ be a sequence of rational numbers starting at some integer index m. Then $(a_n)_{n=m}^{\infty}$ is a Cauchy sequence in the sense of Definition 5.1.8 if and only if it is a Cauchy sequence in the sense of Definition 6.1.3.*

Proof Suppose first that $(a_n)_{n=m}^{\infty}$ is a Cauchy sequence in the sense of Definition 6.1.3; then it is eventually ε-steady for every *real* $\varepsilon > 0$. In particular, it is eventually ε-steady for every *rational* $\varepsilon > 0$, which makes it a Cauchy sequence in the sense of Definition 5.1.8.

Now suppose that $(a_n)_{n=m}^{\infty}$ is a Cauchy sequence in the sense of Definition 5.1.8; then it is eventually ε-steady for every *rational* $\varepsilon > 0$. If $\varepsilon > 0$ is a real number, then there exists a *rational* $\varepsilon' > 0$ which is smaller than ε, by Proposition 5.4.12. Since ε' is rational, we know that $(a_n)_{n=m}^{\infty}$ is eventually ε'-steady; since $\varepsilon' < \varepsilon$, this implies that $(a_n)_{n=m}^{\infty}$ is eventually ε-steady. Since ε is an arbitrary positive real number, we thus see that $(a_n)_{n=m}^{\infty}$ is a Cauchy sequence in the sense of Definition 6.1.3. □

Because of this proposition, we will no longer care about the distinction between Definition 5.1.8 and Definition 6.1.3 and view the concept of a Cauchy sequence as a single unified concept.

Now we talk about what it means for a sequence of real numbers to converge to some limit L.

Definition 6.1.5 *(Convergence of sequences).* Let $\varepsilon > 0$ be a real number, and let L be a real number. A sequence $(a_n)_{n=N}^{\infty}$ of real numbers is said to be *ε-close to L* iff a_n is ε-close to L for every $n \geq N$, i.e., we have $|a_n - L| \leq \varepsilon$ for every $n \geq N$. We say that a sequence $(a_n)_{n=m}^{\infty}$ is *eventually ε-close to L* iff there exists an $N \geq m$ such that $(a_n)_{n=N}^{\infty}$ is ε-close to L. We say that a sequence $(a_n)_{n=m}^{\infty}$ *converges to L* iff it is eventually ε-close to L for every real $\varepsilon > 0$.

One can unwrap all the definitions here and write the concept of convergence more directly; see Exercise 6.1.2.

Example 6.1.6 The sequence

$$0.9, 0.99, 0.999, 0.9999, \ldots$$

is 0.1-close to 1 but is not 0.01-close to 1, because of the first element of the sequence. However, it is eventually 0.01-close to 1. In fact, for every real $\varepsilon > 0$, this sequence is eventually ε-close to 1, hence is convergent to 1.

Proposition 6.1.7 (Uniqueness of limits). *Let $(a_n)_{n=m}^\infty$ be a real sequence starting at some integer index m, and let $L \neq L'$ be two distinct real numbers. Then it is not possible for $(a_n)_{n=m}^\infty$ to converge to L while also converging to L'.*

Proof Suppose for sake of contradiction that $(a_n)_{n=m}^\infty$ was converging to both L and L'. Let $\varepsilon = |L - L'|/3$; note that ε is positive since $L \neq L'$. Since $(a_n)_{n=m}^\infty$ converges to L, we know that $(a_n)_{n=m}^\infty$ is eventually ε-close to L; thus there is an $N \geq m$ such that $d(a_n, L) \leq \varepsilon$ for all $n \geq N$. Similarly, there is an $M \geq m$ such that $d(a_n, L') \leq \varepsilon$ for all $n \geq M$. In particular, if we set $n := \max(N, M)$, then we have $d(a_n, L) \leq \varepsilon$ and $d(a_n, L') \leq \varepsilon$, hence by the triangle inequality $d(L, L') \leq 2\varepsilon = 2|L - L'|/3$. But then we have $|L - L'| \leq 2|L - L'|/3$, which contradicts the fact that $|L - L'| > 0$. Thus it is not possible to converge to both L and L'. □

Now that we know limits are unique, we can set up notation to specify them:

Definition 6.1.8 *(Limits of sequences).* If a sequence $(a_n)_{n=m}^\infty$ converges to some real number L, we say that $(a_n)_{n=m}^\infty$ is *convergent* and that its *limit* is L; we write

$$L = \lim_{n\to\infty} a_n$$

to denote this fact. If a sequence $(a_n)_{n=m}^\infty$ is not converging to any real number L, we say that the sequence $(a_n)_{n=m}^\infty$ is *divergent* and we leave $\lim_{n\to\infty} a_n$ undefined.

Note that Proposition 6.1.7 ensures that a sequence can have at most one limit. Thus, if the limit exists, it is a single real number, otherwise it is undefined.

Remark 6.1.9 The notation $\lim_{n\to\infty} a_n$ does not give any indication about the starting index m of the sequence, but the starting index is irrelevant (Exercise 6.1.3). Thus in the rest of this discussion we shall not be too careful as to where these sequences start, as we shall be mostly focused on their limits.

We sometimes use the phrase "$a_n \to x$ as $n \to \infty$" as an alternate way of writing the statement "$(a_n)_{n=m}^\infty$ converges to x". Bear in mind, though, that the individual statements $a_n \to x$ and $n \to \infty$ do not have any rigorous meaning; this phrase is just a convention, though of course a very suggestive one.

Remark 6.1.10 The exact choice of letter used to denote the index (in this case n) is irrelevant: the phrase $\lim_{n\to\infty} a_n$ has exactly the same meaning as $\lim_{k\to\infty} a_k$, for instance. Sometimes it will be convenient to change the label of the index to avoid conflicts of notation; for instance, we might want to change n to k because n is simultaneously being used for some other purpose, and we want to reduce confusion. See Exercise 6.1.4.

As an example of a limit, we present

Proposition 6.1.11 *We have* $\lim_{n\to\infty} 1/n = 0$.

Proof We have to show that the sequence $(a_n)_{n=1}^\infty$ converges to 0, where $a_n := 1/n$. In other words, for every $\varepsilon > 0$, we need to show that the sequence $(a_n)_{n=1}^\infty$ is eventually ε-close to 0. So, let $\varepsilon > 0$ be an arbitrary real number. We have to find an N such that $|a_n - 0| \leq \varepsilon$ for every $n \geq N$. But if $n \geq N$, then

$$|a_n - 0| = |1/n - 0| = 1/n \leq 1/N.$$

Thus, if we pick $N > 1/\varepsilon$ (which we can do by the Archimedean principle), then $1/N < \varepsilon$, and so $(a_n)_{n=N}^\infty$ is ε-close to 0. Thus $(a_n)_{n=1}^\infty$ is eventually ε-close to 0. Since ε was arbitrary, $(a_n)_{n=1}^\infty$ converges to 0. □

Proposition 6.1.12 (Convergent sequences are Cauchy). *Suppose that* $(a_n)_{n=m}^\infty$ *is a convergent sequence of real numbers. Then* $(a_n)_{n=m}^\infty$ *is also a Cauchy sequence.*

Proof See Exercise 6.1.5. □

Example 6.1.13 The sequence $1, -1, 1, -1, 1, -1, \ldots$ is not a Cauchy sequence (because it is not eventually 1-steady), and is hence not a convergent sequence, by Proposition 6.1.12.

Remark 6.1.14 For a converse to Proposition 6.1.12, see Theorem 6.4.18.

Now we show that formal limits can be superceded by actual limits, just as formal subtraction was superceded by actual subtraction when constructing the integers, and formal division superceded by actual division when constructing the rational numbers.

Proposition 6.1.15 (Formal limits are genuine limits). *Suppose that* $(a_n)_{n=1}^\infty$ *is a Cauchy sequence of rational numbers. Then* $(a_n)_{n=1}^\infty$ *converges to* $\mathrm{LIM}_{n\to\infty} a_n$, *i.e.*

$$\underset{n\to\infty}{\mathrm{LIM}}\, a_n = \lim_{n\to\infty} a_n.$$

Proof See Exercise 6.1.6. □

Definition 6.1.16 *(Bounded sequences).* A sequence $(a_n)_{n=m}^\infty$ of real numbers is *bounded by* a real number M iff we have $|a_n| \leq M$ for all $n \geq m$. We say that $(a_n)_{n=m}^\infty$ is *bounded* iff it is bounded by M for some real number $M > 0$.

This definition is consistent with Definition 5.1.12; see Exercise 6.1.7.

Recall from Lemma 5.1.15 that every Cauchy sequence of rational numbers is bounded. An inspection of the proof of that Lemma shows that the same argument works for real numbers; every Cauchy sequence of real numbers is bounded. In particular, from Proposition 6.1.12 we have

Corollary 6.1.17 *Every convergent sequence of real numbers is bounded.*

Example 6.1.18 The sequence $1, 2, 3, 4, 5, \ldots$ is not bounded, and hence is not convergent.

We can now prove the usual limit laws.

Theorem 6.1.19 (Limit Laws). *Let $(a_n)_{n=m}^{\infty}$ and $(b_n)_{n=m}^{\infty}$ be convergent sequences of real numbers, and let x, y be the real numbers $x := \lim_{n\to\infty} a_n$ and $y := \lim_{n\to\infty} b_n$.*

(a) *The sequence $(a_n + b_n)_{n=m}^{\infty}$ converges to $x + y$; in other words,*

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n.$$

(b) *The sequence $(a_n b_n)_{n=m}^{\infty}$ converges to xy; in other words,*

$$\lim_{n\to\infty}(a_n b_n) = \left(\lim_{n\to\infty} a_n\right)\left(\lim_{n\to\infty} b_n\right).$$

(c) *For any real number c, the sequence $(ca_n)_{n=m}^{\infty}$ converges to cx; in other words,*

$$\lim_{n\to\infty}(ca_n) = c\lim_{n\to\infty} a_n.$$

(d) *The sequence $(a_n - b_n)_{n=m}^{\infty}$ converges to $x - y$; in other words,*

$$\lim_{n\to\infty}(a_n - b_n) = \lim_{n\to\infty} a_n - \lim_{n\to\infty} b_n.$$

(e) *Suppose that $y \neq 0$, and that $b_n \neq 0$ for all $n \geq m$. Then the sequence $(b_n^{-1})_{n=m}^{\infty}$ converges to y^{-1}; in other words,*

$$\lim_{n\to\infty} b_n^{-1} = \left(\lim_{n\to\infty} b_n\right)^{-1}.$$

(f) *Suppose that $y \neq 0$, and that $b_n \neq 0$ for all $n \geq m$. Then the sequence $(a_n/b_n)_{n=m}^{\infty}$ converges to x/y; in other words,*

$$\lim_{n\to\infty}\frac{a_n}{b_n} = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n}.$$

(g) *The sequence*[1] $(\max(a_n, b_n))_{n=m}^{\infty}$ *converges to* $\max(x, y)$*; in other words,*

$$\lim_{n\to\infty}\max(a_n, b_n) = \max\left(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n\right).$$

[1] The operations min, max are defined in Exercise 5.4.9.

(*h*) *The sequence* $(\min(a_n, b_n))_{n=m}^\infty$ *converges to* $\min(x, y)$*; in other words,*

$$\lim_{n\to\infty} \min(a_n, b_n) = \min\left(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n\right).$$

Proof See Exercise 6.1.8. □

— Exercises —

Exercise 6.1.1 Let $(a_n)_{n=0}^\infty$ be a sequence of real numbers, such that $a_{n+1} > a_n$ for each natural number n. Prove that whenever n and m are natural numbers such that $m > n$, then we have $a_m > a_n$. (We refer to these sequences as *strictly increasing* sequences.)

Exercise 6.1.2 Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers, and let L be a real number. Show that $(a_n)_{n=m}^\infty$ converges to L if and only if, given any real $\varepsilon > 0$, one can find an $N \geq m$ such that $|a_n - L| \leq \varepsilon$ for all $n \geq N$.

Exercise 6.1.3 Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers, let c be a real number, and let $m' \geq m$ be an integer. Show that $(a_n)_{n=m}^\infty$ converges to c if and only if $(a_n)_{n=m'}^\infty$ converges to c.

Exercise 6.1.4 Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers, let c be a real number, and let $k \geq 0$ be a non-negative integer. Show that $(a_n)_{n=m}^\infty$ converges to c if and only if $(a_{n+k})_{n=m}^\infty$ converges to c.

Exercise 6.1.5 Prove Proposition 6.1.12. (*Hint:* use the triangle inequality, or Proposition 4.3.7.)

Exercise 6.1.6 Prove Proposition 6.1.15, using the following outline. Let $(a_n)_{n=1}^\infty$ be a Cauchy sequence of rationals, and write $L := \mathrm{LIM}_{n\to\infty} a_n$. We have to show that $(a_n)_{n=1}^\infty$ converges to L. Let $\varepsilon > 0$. Assume for sake of contradiction that sequence a_n is *not* eventually ε-close to L. Use this, and the fact that $(a_n)_{n=1}^\infty$ is Cauchy, to show that there is an $N \geq m$ such that either $a_n > L + \varepsilon/2$ for all $n \geq N$, or $a_n < L - \varepsilon/2$ for all $n \geq N$. Then use Exercise 5.4.8.

Exercise 6.1.7 Show that Definition 6.1.16 is consistent with Definition 5.1.12 (i.e., prove an analogue of Proposition 6.1.4 for bounded sequences instead of Cauchy sequences).

Exercise 6.1.8 Prove Theorem 6.1.19. (*Hint:* you can use some parts of the theorem to prove others, e.g., (b) can be used to prove (c); (a),(c) can be used to prove (d); and (b), (e) can be used to prove (f). The proofs are similar to those of Lemma 5.3.6, Proposition 5.3.10, and Lemma 5.3.15. For (e), you may need to first prove the auxiliary result that any sequence whose elements are non-zero, and which converges to a non-zero limit, is bounded away from zero.)

Exercise 6.1.9 Explain why Theorem 6.1.19(f) fails when the limit of the denominator is 0. (To repair that problem requires *L'Hôpital's rule*, see Section 10.5.)

Exercise 6.1.10 Show that the concept of equivalent Cauchy sequence, as defined in Definition 5.2.6, does not change if ε is required to be positive real instead of positive rational. More precisely, if $(a_n)_{n=0}^\infty$ and $(b_n)_{n=0}^\infty$ are sequences of reals, show that $(a_n)_{n=0}^\infty$ and $(b_n)_{n=0}^\infty$ are eventually ε-close for every rational $\varepsilon > 0$ if and only if they are eventually ε-close for every real $\varepsilon > 0$. (*Hint:* modify the proof of Proposition 6.1.4.)

6.2 The Extended Real Number System

There are some sequences which do not converge to any real number but instead seem to be wanting to converge to $+\infty$ or $-\infty$. For instance, it seems intuitive that the sequence

$$1, 2, 3, 4, 5, \ldots$$

should be converging to $+\infty$, while

$$-1, -2, -3, -4, -5, \ldots$$

should be converging to $-\infty$. Meanwhile, the sequence

$$1, -1, 1, -1, 1, -1, \ldots$$

does not seem to be converging to anything (although we shall see later that it does have $+1$ and -1 as "limit points"—see below). Similarly the sequence

$$1, -2, 3, -4, 5, -6, \ldots$$

does not converge to any real number, and also does not appear to be converging to $+\infty$ or converging to $-\infty$. To make this precise we need to talk about something called the *extended real number system*.

Definition 6.2.1 *(Extended real number system)*. The *extended real number system* $\mathbf{R}^*$ is the real line $\mathbf{R}$ with two additional elements attached, called $+\infty$ and $-\infty$. These elements are distinct from each other and also distinct from every real number. An extended real number x is called *finite* iff it is a real number, and *infinite* iff it is equal to $+\infty$ or $-\infty$. (This definition is not directly related to the notion of finite and infinite sets in Section 3.6, though it is of course similar in spirit.)

These new symbols, $+\infty$ and $-\infty$, at present do not have much meaning, since we have no operations to manipulate them (other than equality $=$ and inequality $\neq$). As with many of the other mathematical concepts considered here, the precise construction of $+\infty$ and $-\infty$ is not important, but (by Exercise 3.2.2) one could for instance set $+\infty := \{\mathbf{R}\}$ and $-\infty := \{\mathbf{R} \cup \{\infty\}\}$ if desired. Now we place a few operations on the extended real number system.

Definition 6.2.2 *(Negation of extended reals)*. The operation of negation $x \mapsto -x$ on $\mathbf{R}$, we now extend to $\mathbf{R}^*$ by defining $-(+\infty) := -\infty$ and $-(-\infty) := +\infty$.

Thus every extended real number x has a negation, and $-(-x)$ is always equal to x.

Definition 6.2.3 *(Ordering of extended reals)*. Let x and y be extended real numbers. We say that $x \leq y$, i.e., x is less than or equal to y, iff one of the following three statements is true:

(a) x and y are real numbers, and $x \leq y$ as real numbers.
(b) $y = +\infty$.
(c) $x = -\infty$.

We say that $x < y$ if we have $x \leq y$ and $x \neq y$. We sometimes write $x < y$ as $y > x$, and $x \leq y$ as $y \geq x$.

Example 6.2.4 $3 \leq 5$, $3 < +\infty$, and $-\infty < +\infty$, but $3 \not\leq -\infty$.

Some basic properties of order and negation on the extended real number system:

Proposition 6.2.5 *Let x, y, z be extended real numbers. Then the following statements are true:*

(a) (*Reflexivity*) *We have $x \leq x$.*
(b) (*Trichotomy*) *Exactly one of the statements $x < y$, $x = y$, or $x > y$ is true.*
(c) (*Transitivity*) *If $x \leq y$ and $y \leq z$, then $x \leq z$.*
(d) (*Negation reverses order*) *If $x \leq y$, then $-y \leq -x$.*

Proof See Exercise 6.2.1. □

One could also introduce other operations on the extended real number system, such as addition and multiplication. However, this is somewhat dangerous as these operations will almost certainly fail to obey the familiar rules of algebra. For instance, to define addition it seems reasonable (given one's intuitive notion of infinity) to set $+\infty + 5 = +\infty$ and $+\infty + 3 = +\infty$, but then this implies that $+\infty + 5 = +\infty + 3$, while $5 \neq 3$. So things like the cancelation law begin to break down once we try to operate involving infinity. To avoid these issues we shall simply not define any arithmetic operations on the extended real number system other than negation and order.

Remember that we defined the notion of *supremum* or *least upper bound* of a set E of reals; this gave an extended real number $\sup(E)$, which was either finite or infinite. We now extend this notion slightly.

Definition 6.2.6 *(Supremum of sets of extended reals).* Let E be a subset of $\mathbf{R}^*$. Then we define the *supremum* $\sup(E)$ or *least upper bound* of E by the following rule.

(a) If E is contained in $\mathbf{R}$ (i.e., $+\infty$ and $-\infty$ are not elements of E), then we let $\sup(E)$ be as defined in Definition 5.5.10.
(b) If E contains $+\infty$, then we set $\sup(E) := +\infty$.
(c) If E does not contain $+\infty$ but does contain $-\infty$, then we set $\sup(E) := \sup(E\backslash\{-\infty\})$ (which is a subset of $\mathbf{R}$ and thus falls under case (a)).

We also define the *infimum* $\inf(E)$ of E (also known as the *greatest lower bound* of E) by the formula

$$\inf(E) := -\sup(-E)$$

where $-E$ is the set $-E := \{-x : x \in E\}$.

Example 6.2.7 Let E be the negative integers, together with $-\infty$:

$$E = \{-1, -2, -3, -4, \ldots\} \cup \{-\infty\}.$$

Then $\sup(E) = \sup(E\backslash\{-\infty\}) = -1$, while

$$\inf(E) = -\sup(-E) = -(+\infty) = -\infty.$$

Example 6.2.8 The set $\{0.9, 0.99, 0.999, 0.9999, \ldots\}$ has infimum 0.9 and supremum 1. Note that in this case the supremum does not actually belong to the set, but it is in some sense "touching it" from the right.

Example 6.2.9 The set $\{1, 2, 3, 4, 5 \ldots\}$ has infimum 1 and supremum $+\infty$.

Example 6.2.10 Let E be the empty set. Then $\sup(E) = -\infty$ and $\inf(E) = +\infty$ (why?). This is the only case in which the supremum can be less than the infimum (why?).

One can intuitively think of the supremum of E as follows. Imagine the real line with $+\infty$ somehow on the far right, and $-\infty$ on the far left. Imagine a piston at $+\infty$ moving leftward until it is stopped by the presence of a set E; the location where it stops is the supremum of E. Similarly if one imagines a piston at $-\infty$ moving rightward until it is stopped by the presence of E, the location where it stops is the infimum of E. In the case when E is the empty set, the pistons pass through each other, the supremum landing at $-\infty$ and the infimum landing at $+\infty$.

The following theorem justifies the terminology "least upper bound" and "greatest lower bound":

Theorem 6.2.11 *Let E be a subset of $\mathbf{R}^*$. Then the following statements are true.*

(*a*) *For every $x \in E$ we have $x \leq \sup(E)$ and $x \geq \inf(E)$.*
(*b*) *Suppose that $M \in \mathbf{R}^*$ is an upper bound for E, i.e., $x \leq M$ for all $x \in E$. Then we have $\sup(E) \leq M$.*
(*c*) *Suppose that $M \in \mathbf{R}^*$ is a lower bound for E, i.e., $x \geq M$ for all $x \in E$. Then we have $\inf(E) \geq M$.*

Proof See Exercise 6.2.2. □

— Exercises —

Exercise 6.2.1 Prove Proposition 6.2.5. (*Hint:* you may need Proposition 5.4.7.)

Exercise 6.2.2 Prove Theorem 6.2.11. (*Hint:* you may need to break into cases depending on whether $+\infty$ or $-\infty$ belongs to E. You can of course use Definition 5.5.10, *provided that E consists only of real numbers*.)

6.3 Suprema and Infima of Sequences

Having defined the notion of a supremum and infimum of sets of reals, we can now also talk about the supremum and infimum of a sequence.

Definition 6.3.1 *(Suprema and infima of sequences).* Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers. Then we define $\sup(a_n)_{n=m}^\infty$ to be the supremum of the set $\{a_n : n \geq m\}$, and $\inf(a_n)_{n=m}^\infty$ to the infimum of the same set $\{a_n : n \geq m\}$.

Remark 6.3.2 The quantities $\sup(a_n)_{n=m}^\infty$ and $\inf(a_n)_{n=m}^\infty$ are sometimes written as $\sup_{n\geq m} a_n$ and $\inf_{n\geq m} a_n$ respectively.

Example 6.3.3 Let $a_n := (-1)^n$; thus $(a_n)_{n=1}^\infty$ is the sequence $-1, 1, -1, 1, \ldots$. Then the set $\{a_n : n \geq 1\}$ is just the two-element set $\{-1, 1\}$, and hence $\sup(a_n)_{n=1}^\infty$ is equal to 1. Similarly $\inf(a_n)_{n=1}^\infty$ is equal to -1.

Example 6.3.4 Let $a_n := 1/n$; thus $(a_n)_{n=1}^\infty$ is the sequence $1, 1/2, 1/3, \ldots$. Then the set $\{a_n : n \geq 1\}$ is the countable set $\{1, 1/2, 1/3, 1/4, \ldots\}$. Thus $\sup(a_n)_{n=1}^\infty = 1$ and $\inf(a_n)_{n=1}^\infty = 0$ (Exercise 6.3.1). Notice here that the infimum of the sequence is not actually a member of the sequence, though it becomes very close to the sequence eventually. (So it is a little inaccurate to think of the supremum and infimum as the "largest element of the sequence" and "smallest element of the sequence", respectively.)

Example 6.3.5 Let $a_n := n$; thus $(a_n)_{n=1}^\infty$ is the sequence $1, 2, 3, 4, \ldots$. Then the set $\{a_n : n \geq 1\}$ is just the positive integers $\{1, 2, 3, 4, \ldots\}$. Then $\sup(a_n)_{n=1}^\infty = +\infty$ and $\inf(a_n)_{n=1}^\infty = 1$.

As the last example shows, it is possible for the supremum or infimum of a sequence to be $+\infty$ or $-\infty$. However, if a sequence $(a_n)_{n=m}^\infty$ is bounded, say bounded by M, then all the elements a_n of the sequence lie between $-M$ and M, so that the set $\{a_n : n \geq m\}$ has M as an upper bound and $-M$ as a lower bound. Since this set is clearly non-empty, we can thus conclude that the supremum and infimum of a bounded sequence are real numbers (i.e., not $+\infty$ and $-\infty$).

Proposition 6.3.6 (Least upper bound property). *Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers, and let x be the extended real number $x := \sup(a_n)_{n=m}^\infty$. Then we have $a_n \leq x$ for all $n \geq m$. Also, whenever $M \in \mathbf{R}^*$ is an upper bound for a_n (i.e., $a_n \leq M$ for all $n \geq m$), we have $x \leq M$. Finally, for every extended real number y for which $y < x$, there exists at least one $n \geq m$ for which $y < a_n \leq x$.*

Proof See Exercise 6.3.2. □

Remark 6.3.7 There is a corresponding Proposition for infima, but with all the references to order reversed, e.g., all upper bounds should now be lower bounds, etc. The proof is exactly the same.

Now we give an application of these concepts of supremum and infimum. In the previous section we saw that all convergent sequences are bounded. It is natural to ask whether the converse is true: are all bounded sequences convergent? The answer is no; for instance, the sequence $1, -1, 1, -1, \dots$ is bounded, but not Cauchy and hence not convergent. However, if we make the sequence both bounded and *monotone* (i.e., increasing or decreasing), then it is true that it must converge:

Proposition 6.3.8 (Monotone bounded sequences converge). *Let $(a_n)_{n=m}^{\infty}$ be a sequence of real numbers which has some finite upper bound $M \in \mathbf{R}$, and which is also increasing (i.e., $a_{n+1} \geq a_n$ for all $n \geq m$). Then $(a_n)_{n=m}^{\infty}$ is convergent, and in fact*

$$\lim_{n\to\infty} a_n = \sup(a_n)_{n=m}^{\infty} \leq M.$$

Proof See Exercise 6.3.3. □

One can similarly prove that if a sequence $(a_n)_{n=m}^{\infty}$ is bounded below and decreasing (i.e., $a_{n+1} \leq a_n$), then it is convergent, and that the limit is equal to the infimum.

A sequence is said to be *monotone* if it is either increasing or decreasing. From Proposition 6.3.8 and Corollary 6.1.17 we see that a monotone sequence converges if and only if it is bounded.

Example 6.3.9 The sequence $3, 3.1, 3.14, 3.141, 3.1415, \dots$ is increasing, and is bounded above by 4. Hence by Proposition 6.3.8 it must have a limit, which is a real number less than or equal to 4.

Proposition 6.3.8 asserts that the limit of a monotone sequence exists, but does not directly say what that limit is. Nevertheless, with a little extra work one can often find the limit once one is given that the limit does exist. For instance:

Proposition 6.3.10 *Let* $0 < x < 1$. *Then we have* $\lim_{n\to\infty} x^n = 0$.

Proof Since $0 < x < 1$, one can show that the sequence $(x^n)_{n=1}^{\infty}$ is decreasing (why?). On the other hand, the sequence $(x^n)_{n=1}^{\infty}$ has a lower bound of 0. Thus by Proposition 6.3.8 (for infima instead of suprema) the sequence $(x^n)_{n=1}^{\infty}$ converges to some limit L. Since $x^{n+1} = x \times x^n$, we thus see from the limit laws (Theorem 6.1.19) that $(x^{n+1})_{n=1}^{\infty}$ converges to xL. But the sequence $(x^{n+1})_{n=1}^{\infty}$ is just the sequence $(x^n)_{n=2}^{\infty}$ shifted by one, and so they must have the same limits (why?). So $xL = L$. Since $x \neq 1$, we can solve for L to obtain $L = 0$. Thus $(x^n)_{n=1}^{\infty}$ converges to 0. □

Note that this proof does not work when $x > 1$ (Exercise 6.3.4).

— Exercises —

Exercise 6.3.1 Verify the claim in Example 6.3.4.

Exercise 6.3.2 Prove Proposition 6.3.6. (*Hint:* use Theorem 6.2.11.)

Exercise 6.3.3 Prove Proposition 6.3.8. (*Hint:* use Proposition 6.3.6, together with the assumption that a_n is increasing, to show that a_n converges to $\sup(a_n)_{n=m}^{\infty}$.)

Exercise 6.3.4 Explain why Proposition 6.3.10 fails when $x > 1$. In fact, show that the sequence $(x^n)_{n=1}^{\infty}$ diverges when $x > 1$. (*Hint:* prove by contradiction and use the identity $(1/x)^n x^n = 1$ and the limit laws in Theorem 6.1.19.) Compare this with the argument in Example 1.2.3; can you now explain the flaws in the reasoning in that example?

6.4 Limsup, Liminf, and Limit Points

Consider the sequence

$$1.1, -1.01, 1.001, -1.0001, 1.00001, \ldots .$$

If one plots this sequence, then one sees (informally, of course) that this sequence does not converge; half the time the sequence is getting close to 1, and half the time the sequence is getting close to -1, but it is not converging to either of them; for instance, it never gets eventually 1/2-close to 1, and never gets eventually 1/2-close to -1. However, even though -1 and $+1$ are not quite limits of this sequence, it does seem that in some vague way they "want" to be limits. To make this notion precise we introduce the notion of a *limit point*.

Definition 6.4.1 *(Limit points).* Let $(a_n)_{n=m}^{\infty}$ be a sequence of real numbers, let x be a real number, and let $\varepsilon > 0$ be a real number. We say that x is ε-*adherent* to $(a_n)_{n=m}^{\infty}$ iff there exists an $n \geq m$ such that a_n is ε-close to x. We say that x is *continually* ε-*adherent* to $(a_n)_{n=m}^{\infty}$ iff it is ε-adherent to $(a_n)_{n=N}^{\infty}$ for every $N \geq m$. We say that x is a *limit point* or *adherent point* of $(a_n)_{n=m}^{\infty}$ iff it is continually ε-adherent to $(a_n)_{n=m}^{\infty}$ for every $\varepsilon > 0$.

Remark 6.4.2 The verb "to adhere" means much the same as "to stick to"; hence the term "adhesive".

Unwrapping all the definitions, we see that x is a limit point of $(a_n)_{n=m}^{\infty}$ if, for every $\varepsilon > 0$ and every $N \geq m$, there exists an $n \geq N$ such that $|a_n - x| \leq \varepsilon$. (Why is this the same definition?) Note the difference between a sequence being ε-close to L (which means that *all* the elements of the sequence stay within a distance ε of L) and L being ε-adherent to the sequence (which only needs a *single* element of the sequence to stay within a distance ε of L). Also, for L to be continually ε-adherent to $(a_n)_{n=m}^{\infty}$, it has to be ε-adherent to $(a_n)_{n=N}^{\infty}$ for *all* $N \geq m$, whereas for $(a_n)_{n=m}^{\infty}$ to be eventually ε-close to L, we only need $(a_n)_{n=N}^{\infty}$ to be ε-close to L for *some* $N \geq m$. Thus there are some subtle differences in quantifiers between limits and limit points.

Note that limit points are only defined for finite real numbers. It is also possible to rigorously define the concept of $+\infty$ or $-\infty$ being a limit point; see Exercise 6.4.8.

Example 6.4.3 Let $(a_n)_{n=1}^{\infty}$ denote the sequence

$$0.9, 0.99, 0.999, 0.9999, 0.99999, \ldots .$$

The number 0.8 is 0.1-adherent to this sequence, since 0.8 is 0.1-close to 0.9, which is a member of this sequence. However, it is not *continually* 0.1-adherent to this sequence, since once one discards the first element of this sequence there is no member of the sequence to be 0.1-close to. In particular, 0.8 is not a limit point of this sequence. On the other hand, the number 1 is 0.1-adherent to this sequence, and in fact is continually 0.1-adherent to this sequence, since no matter how many initial members of the sequence one discards, there is still something for 1 to be 0.1-close to. In fact, it is continually ε-adherent for every ε, and is hence a limit point of this sequence.

Example 6.4.4 Now consider the sequence

$$1.1, -1.01, 1.001, -1.0001, 1.00001, \ldots.$$

The number 1 is 0.1-adherent to this sequence; in fact it is continually 0.1-adherent to this sequence, because no matter how many elements of the sequence one discards, there are some elements of the sequence that 1 is 0.1-close to. (As discussed earlier, one does not need *all* the elements to be 0.1-close to 1, just some; thus 0.1-adherent is weaker than 0.1-close, and continually 0.1-adherent is a different notion from eventually 0.1-close.) In fact, for every $\varepsilon > 0$, the number 1 is continually ε-adherent to this sequence and is thus a limit point of this sequence. Similarly, -1 is a limit point of this sequence; however 0 (say) is not a limit point of this sequence, since it is not continually 0.1-adherent to it.

Limits are of course a special case of limit points:

Proposition 6.4.5 (Limits are limit points). *Let $(a_n)_{n=m}^\infty$ be a sequence which converges to a real number c. Then c is a limit point of $(a_n)_{n=m}^\infty$, and in fact it is the only limit point of $(a_n)_{n=m}^\infty$.*

Proof See Exercise 6.4.1. □

Now we will look at two special types of limit points: the limit superior (lim sup) and limit inferior (lim inf).

Definition 6.4.6 *(Limit superior and limit inferior).* Suppose that $(a_n)_{n=m}^\infty$ is a sequence. We define a new sequence $(a_N^+)_{N=m}^\infty$ by the formula

$$a_N^+ := \sup(a_n)_{n=N}^\infty.$$

More informally, a_N^+ is the supremum of all the elements in the sequence from a_N onwards. We then define the *limit superior* of the sequence $(a_n)_{n=m}^\infty$, denoted $\limsup_{n\to\infty} a_n$, by the formula

$$\limsup_{n\to\infty} a_n := \inf(a_N^+)_{N=m}^\infty.$$

Similarly, we can define

$$a_N^- := \inf(a_n)_{n=N}^\infty$$

and define the *limit inferior* of the sequence $(a_n)_{n=m}^\infty$, denoted $\liminf_{n\to\infty} a_n$, by the formula

$$\liminf_{n\to\infty} a_n := \sup(a_N^-)_{N=m}^\infty.$$

Example 6.4.7 Let $a_1, a_2, a_3, \ldots$ denote the sequence

$$1.1, -1.01, 1.001, -1.0001, 1.00001, \ldots.$$

Then $a_1^+, a_2^+, a_3^+, \ldots$ is the sequence

$$1.1, 1.001, 1.001, 1.00001, 1.00001, \ldots$$

(why?), and its infimum is 1. Hence the limit superior of this sequence is 1. Similarly, $a_1^-, a_2^-, a_3^-, \ldots$ is the sequence

$$-1.01, -1.01, -1.0001, -1.0001, -1.000001, \ldots$$

(why?), and the supremum of this sequence is -1. Hence the limit inferior of this sequence is -1. One should compare this with the supremum and infimum of the sequence, which are 1.1 and -1.01 respectively.

Example 6.4.8 Let $a_1, a_2, a_3, \ldots$ denote the sequence

$$1, -2, 3, -4, 5, -6, 7, -8, \ldots$$

Then $a_1^+, a_2^+, \ldots$ is the sequence

$$+\infty, +\infty, +\infty, +\infty, \ldots$$

(why?) and so the limit superior is $+\infty$. Similarly, $a_1^-, a_2^-, \ldots$ is the sequence

$$-\infty, -\infty, -\infty, -\infty, \ldots$$

and so the limit inferior is $-\infty$.

Example 6.4.9 Let $a_1, a_2, a_3, \ldots$ denote the sequence

$$1, -1/2, 1/3, -1/4, 1/5, -1/6, \ldots$$

Then $a_1^+, a_2^+, \ldots$ is the sequence

$$1, 1/3, 1/3, 1/5, 1/5, 1/7, \ldots$$

which has an infimum of 0 (why?), so the limit superior is 0. Similarly, $a_1^-, a_2^-, \ldots$ is the sequence

$$-1/2, -1/2, -1/4, -1/4, -1/6, -1/6, \ldots$$

which has a supremum of 0. So the limit inferior is also 0.

Example 6.4.10 Let $a_1, a_2, a_3, \ldots$ denote the sequence

$$1, 2, 3, 4, 5, 6, \ldots$$

Then $a_1^+, a_2^+, \ldots$ is the sequence

$$+\infty, +\infty, +\infty, \ldots$$

so the limit superior is $+\infty$. Similarly, $a_1^-, a_2^-, \ldots$ is the sequence

$$1, 2, 3, 4, 5, \ldots$$

which has a supremum of $+\infty$. So the limit inferior is also $+\infty$.

Remark 6.4.11 Some authors use the notation $\overline{\lim}_{n\to\infty} a_n$ instead of $\limsup_{n\to\infty} a_n$, and $\underline{\lim}_{n\to\infty} a_n$ instead of $\liminf_{n\to\infty} a_n$. Note that the starting index m of the sequence is irrelevant (see Exercise 6.4.2).

Returning to the piston analogy, imagine a piston at $+\infty$ moving leftward until it is stopped by the presence of the sequence $a_1, a_2, \ldots$. The place it will stop is the supremum of $a_1, a_2, a_3, \ldots$, which in our new notation is a_1^+. Now let us remove the first element a_1 from the sequence; this may cause our piston to slip leftward, to a new point a_2^+ (though in many cases the piston will not move and a_2^+ will just be the same as a_1^+). Then we remove the second element a_2, causing the piston to slip a little more. If we keep doing this the piston will keep slipping, but there will be some point where it cannot go any further, and this is the limit superior of the sequence. A similar analogy can describe the limit inferior of the sequence.

We now describe some basic properties of limit superior and limit inferior.

Proposition 6.4.12 *Let $(a_n)_{n=m}^\infty$ be a sequence of real numbers, let L^+ be the limit superior of this sequence, and let L^- be the limit inferior of this sequence (thus both L^+ and L^- are extended real numbers).*

(*a*) *For every $x > L^+$, there exists an $N \geq m$ such that $a_n < x$ for all $n \geq N$. (In other words, for every $x > L^+$, the elements of the sequence $(a_n)_{n=m}^\infty$ are eventually less than x.) Similarly, for every $y < L^-$ there exists an $N \geq m$ such that $a_n > y$ for all $n \geq N$.*

(*b*) *For every $x < L^+$, and every $N \geq m$, there exists an $n \geq N$ such that $a_n > x$. (In other words, for every $x < L^+$, the elements of the sequence $(a_n)_{n=m}^\infty$ exceed x infinitely often.) Similarly, for every $y > L^-$ and every $N \geq m$, there exists an $n \geq N$ such that $a_n < y$.*

(*c*) *We have* $\inf(a_n)_{n=m}^\infty \le L^- \le L^+ \le \sup(a_n)_{n=m}^\infty$.
(*d*) *If* c *is any limit point of* $(a_n)_{n=m}^\infty$, *then we have* $L^- \le c \le L^+$.
(*e*) *If* L^+ *is finite, then it is a limit point of* $(a_n)_{n=m}^\infty$. *Similarly, if* L^- *is finite, then it is a limit point of* $(a_n)_{n=m}^\infty$.
(*f*) *Let* c *be a real number. If* $(a_n)_{n=m}^\infty$ *converges to* c, *then we must have* $L^+ = L^- = c$. *Conversely, if* $L^+ = L^- = c$, *then* $(a_n)_{n=m}^\infty$ *converges to* c.

Proof We shall prove (a) and (b) and leave the remaining parts to the exercises. Suppose first that $x > L^+$. Then by definition of L^+, we have $x > \inf(a_N^+)_{N=m}^\infty$. By Proposition 6.3.6, there must then exist an integer $N \ge m$ such that $x > a_N^+$. By definition of a_N^+, this means that $x > \sup(a_n)_{n=N}^\infty$. Thus by Proposition 6.3.6 again, we have $x > a_n$ for all $n \ge N$, as desired. This proves the first part of (a); the second part of (a) is proven similarly.

Now we prove (b). Suppose that $x < L^+$. Then we have $x < \inf(a_N^+)_{N=m}^\infty$. If we fix any $N \ge m$, then by Proposition 6.3.6, we thus have $x < a_N^+$. By definition of a_N^+, this means that $x < \sup(a_n)_{n=N}^\infty$. By Proposition 6.3.6 again, there must thus exist $n \ge N$ such that $a_n > x$, as desired. This proves the first part of (b), the second part of (b) is proven similarly.

The proofs of (c), (d), (e), (f) are left to Exercise 6.4.3. □

Parts (d) and (e) of Proposition 6.4.12 say, in particular, that L^+ is the largest limit point of $(a_n)_{n=m}^\infty$, and L^- is the smallest limit point (provided that L^+ and L^- are finite). Proposition 6.4.12 (f) then says that if L^+ and L^- coincide (so there is only one limit point) and are finite, then the sequence in fact converges. This gives a way to test if a sequence converges: compute its limit superior and limit inferior, and see if they are equal.

We now give a basic comparison property of limit superior and limit inferior.

Lemma 6.4.13 (Comparison principle). *Suppose that* $(a_n)_{n=m}^\infty$ *and* $(b_n)_{n=m}^\infty$ *are two sequences of real numbers such that* $a_n \le b_n$ *for all* $n \ge m$. *Then we have the inequalities*

$$\begin{aligned}\sup(a_n)_{n=m}^\infty &\le \sup(b_n)_{n=m}^\infty\\ \inf(a_n)_{n=m}^\infty &\le \inf(b_n)_{n=m}^\infty\\ \limsup_{n\to\infty} a_n &\le \limsup_{n\to\infty} b_n\\ \liminf_{n\to\infty} a_n &\le \liminf_{n\to\infty} b_n\end{aligned}$$

Proof See Exercise 6.4.4. □

Corollary 6.4.14 (Squeeze test). *Let* $(a_n)_{n=m}^\infty$, $(b_n)_{n=m}^\infty$, *and* $(c_n)_{n=m}^\infty$ *be sequences of real numbers such that*

$$a_n \le b_n \le c_n$$

for all $n \ge m$. *Suppose also that* $(a_n)_{n=m}^\infty$ *and* $(c_n)_{n=m}^\infty$ *both converge to the same limit* L. *Then* $(b_n)_{n=m}^\infty$ *is also convergent to* L.

Proof See Exercise 6.4.5. □

Example 6.4.15 We already know (see Proposition 6.1.11) that $\lim_{n\to\infty} 1/n = 0$. By the limit laws (Theorem 6.1.19), this also implies that $\lim_{n\to\infty} 2/n = 0$ and $\lim_{n\to\infty} -2/n = 0$. The squeeze test then shows that any sequence $(b_n)_{n=1}^{\infty}$ for which

$$-2/n \le b_n \le 2/n \text{ for all } n \ge 1$$

is convergent to 0. For instance, we can use this to show that the sequence $(-1)^n/n + 1/n^2$ converges to zero, or that 2^{-n} converges to zero. Note one can use induction to show that $0 \le 2^{-n} \le 1/n$ for all $n \ge 1$.

Remark 6.4.16 The squeeze test, combined with the limit laws and the principle that monotone bounded sequences always have limits, allows one to compute a large number of limits. We give some examples in the next chapter.

One commonly used consequence of the squeeze test is

Corollary 6.4.17 (Zero test for sequences). *Let $(a_n)_{n=M}^{\infty}$ be a sequence of real numbers. Then the limit $\lim_{n\to\infty} a_n$ exists and is equal to zero if and only if the limit $\lim_{n\to\infty} |a_n|$ exists and is equal to zero.*

Proof See Exercise 6.4.7. □

We close this section with the following improvement to Proposition 6.1.12.

Theorem 6.4.18 *(Completeness of the reals). A sequence $(a_n)_{n=1}^{\infty}$ of real numbers is a Cauchy sequence if and only if it is convergent.*

Remark 6.4.19 Note that while this is very similar in spirit to Proposition 6.1.15, it is a bit more general, since Proposition 6.1.15 refers to Cauchy sequences of rationals instead of real numbers.

Proof Proposition 6.1.12 already tells us that every convergent sequence is Cauchy, so it suffices to show that every Cauchy sequence is convergent.

Let $(a_n)_{n=1}^{\infty}$ be a Cauchy sequence. We know from Lemma 5.1.15 (or more precisely, from the extension of this lemma to the real numbers, which is proven in exactly the same fashion) that the sequence $(a_n)_{n=1}^{\infty}$ is bounded; by Lemma 6.4.13 (or Proposition 6.4.12(c)) this implies that $L^- := \liminf_{n\to\infty} a_n$ and $L^+ := \limsup_{n\to\infty} a_n$ of the sequence are both finite. To show that the sequence converges, it will suffice by Proposition 6.4.12(f) to show that $L^- = L^+$.

Now let $\varepsilon > 0$ be any real number. Since $(a_n)_{n=1}^{\infty}$ is a Cauchy sequence, it must be eventually ε-steady, so in particular there exists an $N \ge 1$ such that the sequence $(a_n)_{n=N}^{\infty}$ is ε-steady. In particular, we have $a_N - \varepsilon \le a_n \le a_N + \varepsilon$ for all $n \ge N$. By Proposition 6.3.6 (or Lemma 6.4.13) this implies that

$$a_N - \varepsilon \le \inf(a_n)_{n=N}^{\infty} \le \sup(a_n)_{n=N}^{\infty} \le a_N + \varepsilon$$

and hence by the definition of L^- and L^+ (and Proposition 6.3.6 again)

$$a_N - \varepsilon \leq L^- \leq L^+ \leq a_N + \varepsilon.$$

Thus we have

$$0 \leq L^+ - L^- \leq 2\varepsilon.$$

But this is true for all $\varepsilon > 0$, and L^+ and L^- do not depend on ε; so we must therefore have $L^+ = L^-$. (If $L^+ > L^-$ then we could set $\varepsilon := (L^+ - L^-)/3$ and obtain a contradiction.) By Proposition 6.4.12(f) we thus see that the sequence converges. □

Remark 6.4.20 In the language of metric spaces (see Chap. 1 of *Analysis II*), Theorem 6.4.18 asserts that the real numbers are a *complete* metric space—that they do not contain "holes" the same way the rationals do. (Certainly the rationals have lots of Cauchy sequences which do not converge to other rationals; take for instance the sequence $1, 1.4, 1.41, 1.414, 1.4142, \ldots$ which converges to the irrational $\sqrt{2}$.) This property is closely related to the least upper bound property (Theorem 5.5.9), and is one of the principal characteristics which make the real numbers superior to the rational numbers for the purposes of doing analysis (taking limits, taking derivatives and integrals, finding zeroes of functions, that kind of thing), as we shall see in later chapters.

— Exercises —

Exercise 6.4.1 Prove Proposition 6.4.5.

Exercise 6.4.2 State and prove analogues of Exercises 6.1.3 and 6.1.4 for limit points, limit superior, and limit inferior.

Exercise 6.4.3 Prove parts (c), (d), (e), (f) of Proposition 6.4.12. (*Hint:* you can use earlier parts of the proposition to prove later ones.)

Exercise 6.4.4 Prove Lemma 6.4.13.

Exercise 6.4.5 Use Lemma 6.4.13 to prove Corollary 6.4.14.

Exercise 6.4.6 Give an example of two bounded sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ such that $a_n < b_n$ for all $n \geq 1$, but that $\sup(a_n)_{n=1}^\infty \not< \sup(b_n)_{n=1}^\infty$. Explain why this does not contradict Lemma 6.4.13.

Exercise 6.4.7 Prove Corollary 6.4.17. Is the corollary still true if we replace zero in the statement of this corollary by some other number?

Exercise 6.4.8 Let us say that a sequence $(a_n)_{n=M}^\infty$ of real numbers has $+\infty$ as a limit point iff it has no finite upper bound, and that it has $-\infty$ as a limit point iff it has no finite lower bound. With this definition, show that $\limsup_{n\to\infty} a_n$ is a limit point of $(a_n)_{n=M}^\infty$, and furthermore that it is larger than all the other limit points of $(a_n)_{n=M}^\infty$; in other words, the limit superior is the largest limit point of a sequence. Similarly, show that the limit inferior is the smallest limit point of a sequence. (One can use Proposition 6.4.12 in the course of the proof.)

Exercise 6.4.9 Using the definition in Exercise 6.4.8, construct a sequence $(a_n)_{n=1}^\infty$ which has exactly three limit points, at $-\infty$, 0, and $+\infty$.

Exercise 6.4.10 Let $(a_n)_{n=N}^{\infty}$ be a sequence of real numbers, and let $(b_m)_{m=M}^{\infty}$ be another sequence of real numbers such that each b_m is a limit point of $(a_n)_{n=N}^{\infty}$. Let c be a limit point of $(b_m)_{m=M}^{\infty}$. Prove that c is also a limit point of $(a_n)_{n=N}^{\infty}$. (In other words, limit points of limit points are themselves limit points of the original sequence.)

6.5 Some Standard Limits

Armed now with the limit laws and the squeeze test, we can now compute a large number of limits.

A particularly simple limit is that of the *constant sequence* $c, c, c, c, \ldots$; we clearly have

$$\lim_{n\to\infty} c = c$$

for any constant c (why?).

Also, in Proposition 6.1.11, we proved that $\lim_{n\to\infty} 1/n = 0$. This now implies

Corollary 6.5.1 *We have* $\lim_{n\to\infty} 1/n^{1/k} = 0$ *for every integer* $k \geq 1$.

Proof From Lemma 5.6.6 we know that $1/n^{1/k}$ is a decreasing function of n, while being bounded below by 0. By Proposition 6.3.8 (for decreasing sequences instead of increasing sequences) we thus know that this sequence converges to some limit $L \geq 0$:

$$L = \lim_{n\to\infty} 1/n^{1/k}.$$

Raising this to the kth power and using the limit laws (or more precisely, Theorem 6.1.19(b) and induction), we obtain

$$L^k = \lim_{n\to\infty} 1/n.$$

By Proposition 6.1.11 we thus have $L^k = 0$; but this means that L cannot be positive (else L^k would be positive), so $L = 0$, and we are done. □

Some other basic limits:

Lemma 6.5.2 *Let* x *be a real number. Then the limit* $\lim_{n\to\infty} x^n$ *exists and is equal to zero when* $|x| < 1$, *exists and is equal to 1 when* $x = 1$, *and diverges when* $x = -1$ *or when* $|x| > 1$.

Proof See Exercise 6.5.2. □

Lemma 6.5.3 *For any* $x > 0$, *we have* $\lim_{n\to\infty} x^{1/n} = 1$.

Proof See Exercise 6.5.3. □

We will derive a few more standard limits later on, once we develop the root and ratio tests for series and for sequences.

— Exercises —

Exercise 6.5.1 Show that $\lim_{n\to\infty} 1/n^q = 0$ for any rational $q > 0$. (*Hint:* use Corollary 6.5.1 and the limit laws, Theorem 6.1.19.) Conclude that the limit $\lim_{n\to\infty} n^q$ does not exist. (*Hint:* argue by contradiction using Theorem 6.1.19(e).)

Exercise 6.5.2 Prove Lemma 6.5.2. (*Hint:* use Proposition 6.3.10, Exercise 6.3.4, and the squeeze test.)

Exercise 6.5.3 Prove Lemma 6.5.3. (*Hint:* you may need to treat the cases $x \geq 1$ and $x < 1$ separately. You might wish to first use Lemma 6.5.2 to prove the preliminary result that for every $\varepsilon > 0$ and every real number $M > 0$, there exists an n such that $M^{1/n} \leq 1 + \varepsilon$.)

6.6 Subsequences

This chapter has been devoted to the study of sequences $(a_n)_{n=m}^\infty$ of real numbers, and their limits. Some sequences were convergent to a single limit, while others had multiple limit points. For instance, the sequence

$$1.1, 0.1, 1.01, 0.01, 1.001, 0.001, 1.0001, \ldots$$

has two limit points at 0 and 1 (which are incidentally also the lim inf and lim sup respectively), but is not actually convergent (since the lim sup and lim inf are not equal). However, while this sequence is not convergent, it does appear to contain convergent components; it seems to be a mixture of two convergent subsequences, namely

$$1.1, 1.01, 1.001, \ldots$$

and

$$0.1, 0.01, 0.001, \ldots.$$

To make this notion more precise, we need a notion of subsequence.

Definition 6.6.1 *(Subsequences).* Let $(a_n)_{n=0}^\infty$ and $(b_n)_{n=0}^\infty$ be sequences of real numbers. We say that $(b_n)_{n=0}^\infty$ is a *subsequence* of $(a_n)_{n=0}^\infty$ iff there exists a function $f\colon \mathbf{N} \to \mathbf{N}$ which is strictly increasing (i.e., $f(n+1) > f(n)$ for all $n \in \mathbf{N}$) such that

$$b_n = a_{f(n)} \text{ for all } n \in \mathbf{N}.$$

More generally, we say that $(b_n)_{n=m'}^\infty$ is a subsequence of $(a_n)_{n=m}^\infty$ if there exists a strictly increasing function $f\colon \{n \in \mathbf{N} : n \geq m'\} \to \{n \in \mathbf{N} : n \geq m\}$ such that $b_n = a_{f(n)}$ for all $n \in \mathbf{N}$ with $n \geq m'$.

Example 6.6.2 If $(a_n)_{n=0}^{\infty}$ is a sequence, then $(a_{2n})_{n=0}^{\infty}$ is a subsequence of $(a_n)_{n=0}^{\infty}$, since the function $f : \mathbf{N} \to \mathbf{N}$ defined by $f(n) := 2n$ is a strictly increasing function from $\mathbf{N}$ to $\mathbf{N}$. Note that we do not assume f to be obijective, although it is necessarily injective (why?). More informally, the sequence

$$a_0, a_2, a_4, a_6, \ldots$$

is a subsequence of

$$a_0, a_1, a_2, a_3, a_4, \ldots.$$

Example 6.6.3 The two sequences

$$1.1, 1.01, 1.001, \ldots$$

and

$$0.1, 0.01, 0.001, \ldots$$

mentioned earlier are both subsequences of

$$1.1, 0.1, 1.01, 0.01, 1.001, 0.001, 1.0001, \ldots$$

The property of being a subsequence is reflexive and transitive, though not symmetric:

Lemma 6.6.4 *Let $(a_n)_{n=0}^{\infty}$, $(b_n)_{n=0}^{\infty}$, and $(c_n)_{n=0}^{\infty}$ be sequences of real numbers. Then $(a_n)_{n=0}^{\infty}$ is a subsequence of $(a_n)_{n=0}^{\infty}$. Furthermore, if $(b_n)_{n=0}^{\infty}$ is a subsequence of $(a_n)_{n=0}^{\infty}$, and $(c_n)_{n=0}^{\infty}$ is a subsequence of $(b_n)_{n=0}^{\infty}$, then $(c_n)_{n=0}^{\infty}$ is a subsequence of $(a_n)_{n=0}^{\infty}$.*

Proof See Exercise 6.6.1. □

We now relate the concept of subsequences to the concept of limits and limit points.

Proposition 6.6.5 (Subsequences related to limits). *Let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers, and let L be a real number. Then the following two statements are logically equivalent* (*each one implies the other*)*:*

(*a*) *The sequence $(a_n)_{n=0}^{\infty}$ converges to L.*
(*b*) *Every subsequence of $(a_n)_{n=0}^{\infty}$ converges to L.*

Proof See Exercise 6.6.4. □

Proposition 6.6.6 (Subsequences related to limit points). *Let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers, and let L be a real number. Then the following two statements are logically equivalent.*

(*a*) *L is a limit point of $(a_n)_{n=0}^{\infty}$.*

(*b*) *There exists a subsequence of* $(a_n)_{n=0}^{\infty}$ *which converges to* L.

Proof See Exercise 6.6.5. □

Remark 6.6.7 The above two propositions give a sharp contrast between the notion of a limit and that of a limit point. When a sequence has a limit L, then *all* subsequences also converge to L. But when a sequence has L as a limit point, then only *some* subsequences converge to L.

We can now prove an important theorem in real analysis, due to Bernard Bolzano (1781–1848) and Karl Weierstrass (1815–1897): every bounded sequence has a convergent subsequence.

Theorem 6.6.8 *(Bolzano–Weierstrass theorem) Let* $(a_n)_{n=0}^{\infty}$ *be a bounded sequence* (*i.e., there exists a real number* $M > 0$ *such that* $|a_n| \leq M$ *for all* $n \in \mathbf{N}$). *Then there is at least one subsequence of* $(a_n)_{n=0}^{\infty}$ *which converges.*

Proof Let L be the limit superior of the sequence $(a_n)_{n=0}^{\infty}$. Since we have $-M \leq a_n \leq M$ for all natural numbers n, it follows from the comparison principle (Lemma 6.4.13) that $-M \leq L \leq M$. In particular, L is a real number (not $+\infty$ or $-\infty$). By Proposition 6.4.12(e), L is thus a limit point of $(a_n)_{n=0}^{\infty}$. Thus by Proposition 6.6.6, there exists a subsequence of $(a_n)_{n=0}^{\infty}$ which converges (in fact, it converges to L). □

Note that we could as well have used the limit inferior instead of the limit superior in the above argument.

Remark 6.6.9 The Bolzano–Weierstrass theorem says that if a sequence is bounded, then eventually it has no choice but to converge in some places; it has "no room" to spread out and stop itself from acquiring limit points. It is not true for unbounded sequences; for instance, the sequence $1, 2, 3, \ldots$ has no convergent subsequences whatsoever (why?). In the language of topology, this means that the interval $\{x \in \mathbf{R} : -M \leq x \leq M\}$ is *compact*, whereas an unbounded set such as the real line $\mathbf{R}$ is not compact. The distinction between compact sets and non-compact sets will be very important in later chapters - of similar importance to the distinction between finite sets and infinite sets.

— Exercises —

Exercise 6.6.1 Prove Lemma 6.6.4.

Exercise 6.6.2 Can you find two sequences $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ which are not the same sequence, but such that each is a subsequence of the other?

Exercise 6.6.3 (For this exercise you may assume the well-ordering principle, Proposition 8.1.4.) Let $(a_n)_{n=0}^{\infty}$ be a sequence which is not bounded. Show that there exists a subsequence $(b_n)_{n=0}^{\infty}$ of $(a_n)_{n=0}^{\infty}$ such that $\lim_{n\to\infty} 1/b_n$ exists and is equal to zero. (*Hint:* for each natural number j, recursively introduce the quantity $n_j := \min\{n \in \mathbf{N} : |a_n| \geq j; n > n_{j-1}\}$ (omitting the condition $n > n_{j-1}$ when $j = 0$), first explaining why the set $\{n \in \mathbf{N} : |a_n| \geq j; n > n_{j-1}\}$ is non-empty. Then set $b_j := a_{n_j}$. To ensure the existence and uniqueness of the minimum, one either needs to invoke the well-ordering principle (which we have placed in Proposition 8.1.4, but whose proof does not rely on any material not already presented), or the least upper bound principle (Theorem 5.5.9).)

Exercise 6.6.4 Prove Proposition 6.6.5. (Note that one of the two implications has a very short proof.)

Exercise 6.6.5 Prove Proposition 6.6.6. (*Hint:* to show that (a) implies (b), define the numbers n_j for each natural numbers j by the formula $n_j := \min\{n > n_{j-1} : |a_n - L| \leq 1/j\}$, with the convention $n_0 := 0$, explaining why the set $\{n > n_{j-1} : |a_n - L| \leq 1/j\}$ is non-empty. Then consider the sequence a_{n_j}.)

6.7 Real Exponentiation, Part II

We finally return to the topic of exponentiation of real numbers that we started in Sec. 5.6. In that section we defined x^q for all rational q and positive real numbers x, but we have not yet defined x^α when α is real. We now rectify this situation using limits (in a similar way as to how we defined all the other standard operations on the real numbers). First, we need a lemma:

Lemma 6.7.1 *[Continuity of exponentiation] Let $x > 0$, and let α be a real number. Let $(q_n)_{n=1}^\infty$ be any sequence of rational numbers converging to α. Then $(x^{q_n})_{n=1}^\infty$ is also a convergent sequence. Furthermore, if $(q'_n)_{n=1}^\infty$ is any other sequence of rational numbers converging to α, then $(x^{q'_n})_{n=1}^\infty$ has the same limit as $(x^{q_n})_{n=1}^\infty$:*

$$\lim_{n\to\infty} x^{q_n} = \lim_{n\to\infty} x^{q'_n}.$$

Proof There are three cases: $x < 1$, $x = 1$, and $x > 1$. The case $x = 1$ is rather easy (because then $x^q = 1$ for all rational q). We shall just do the case $x > 1$, and leave the case $x < 1$ (which is very similar) to the reader.

Let us first prove that $(x^{q_n})_{n=1}^\infty$ converges. By Proposition 6.4.18 it is enough to show that $(x^{q_n})_{n=1}^\infty$ is a Cauchy sequence.

To do this, we need to estimate the distance between x^{q_n} and x^{q_m}; let us say for the time being that $q_n \geq q_m$, so that $x^{q_n} \geq x^{q_m}$ (since $x > 1$). We have

$$d(x^{q_n}, x^{q_m}) = x^{q_n} - x^{q_m} = x^{q_m}(x^{q_n - q_m} - 1).$$

Since $(q_n)_{n=1}^\infty$ is a convergent sequence, it has some upper bound M; since $x > 1$, we have $x^{q_m} \leq x^M$. Thus

$$d(x^{q_n}, x^{q_m}) = |x^{q_n} - x^{q_m}| \leq x^M(x^{q_n - q_m} - 1).$$

Now let $\varepsilon > 0$. We know by Lemma 6.5.3 that the sequence $(x^{1/k})_{k=1}^\infty$ is eventually εx^{-M}-close to 1. Thus there exists some $K \geq 1$ such that

$$|x^{1/K} - 1| \leq \varepsilon x^{-M}.$$

Now since $(q_n)_{n=1}^{\infty}$ is convergent, it is a Cauchy sequence, and so there is an $N \geq 1$ such that q_n and q_m are $1/K$-close for all $n, m \geq N$. Thus we have

$$d(x^{q_n}, x^{q_m}) \leq x^M(x^{q_n - q_m} - 1) \leq x^M(x^{1/K} - 1) \leq x^M \varepsilon x^{-M} = \varepsilon$$

for every $n, m \geq N$ such that $q_n \geq q_m$. By symmetry we also have this bound when $n, m \geq N$ and $q_n \leq q_m$. Thus the sequence $(x^{q_n})_{n=N}^{\infty}$ is ε-steady. Thus the sequence $(x^{q_n})_{n=1}^{\infty}$ is eventually ε-steady for every $\varepsilon > 0$, and is thus a Cauchy sequence as desired. This proves the convergence of $(x^{q_n})_{n=1}^{\infty}$.

Now we prove the second claim. It will suffice to show that

$$\lim_{n\to\infty} x^{q_n - q_n'} = 1,$$

since the claim would then follow from limit laws (since $x^{q_n} = x^{q_n - q_n'} x^{q_n'}$).

Write $r_n := q_n - q_n'$; by limit laws we know that $(r_n)_{n=1}^{\infty}$ converges to 0. We have to show that for every $\varepsilon > 0$, the sequence $(x^{r_n})_{n=1}^{\infty}$ is eventually ε-close to 1. But from Lemma 6.5.3 we know that the sequence $(x^{1/k})_{k=1}^{\infty}$ is eventually ε-close to 1. Since $\lim_{k\to\infty} x^{-1/k}$ is also equal to 1 by Lemma 6.5.3, we know that $(x^{-1/k})_{k=1}^{\infty}$ is also eventually ε-close to 1. Thus we can find a K such that $x^{1/K}$ and $x^{-1/K}$ are both ε-close to 1. But since $(r_n)_{n=1}^{\infty}$ is convergent to 0, it is eventually $1/K$-close to 0, so that eventually $-1/K \leq r_n \leq 1/K$, and thus $x^{-1/K} \leq x^{r_n} \leq x^{1/K}$. In particular x^{r_n} is also eventually ε-close to 1 (see Proposition 4.3.7(f)), as desired. □

We may now make the following definition.

Definition 6.7.2 *(Exponentiation to a real exponent).* Let $x > 0$ be real, and let α be a real number. We define the quantity x^α by the formula $x^\alpha = \lim_{n\to\infty} x^{q_n}$, where $(q_n)_{n=1}^{\infty}$ is any sequence of rational numbers converging to α.

Let us check that this definition is well-defined. First of all, given any real number α we always have at least one sequence $(q_n)_{n=1}^{\infty}$ of rational numbers converging to α, by the definition of real numbers (and Proposition 6.1.15). Secondly, given any such sequence $(q_n)_{n=1}^{\infty}$, the limit $\lim_{n\to\infty} x^{q_n}$ exists by Lemma 6.7.1. Finally, even though there can be multiple choices for the sequence $(q_n)_{n=1}^{\infty}$, they all give the same limit by Lemma 6.7.1. Thus this definition is well-defined.

If α is not just real but rational, i.e., $\alpha = q$ for some rational q, then this definition could in principle be inconsistent with our earlier definition of exponentiation in Section 5.6. But in this case α is clearly the limit of the sequence $(q)_{n=1}^{\infty}$, so by definition $x^\alpha = \lim_{n\to\infty} x^q = x^q$. Thus the new definition of exponentiation is consistent with the old one.

Proposition 6.7.3 *All the results of Lemma 5.6.9, which held for rational numbers q and r, continue to hold for real numbers q and r.*

Proof We demonstrate this for the identity $x^{q+r} = x^q x^r$ (i.e., the first part of Lemma 5.6.9(b)); the other parts are similar and are left to Exercise 6.7.1. The idea is to start with Lemma 5.6.9 for rationals and then take limits to obtain the corresponding results for reals.

Let q and r be real numbers. Then we can write $q = \lim_{n\to\infty} q_n$ and $r = \lim_{n\to\infty} r_n$ for some sequences $(q_n)_{n=1}^{\infty}$ and $(r_n)_{n=1}^{\infty}$ of rationals, by the definition of real numbers (and Proposition 6.1.15). Then by the limit laws, $q + r$ is the limit of $(q_n + r_n)_{n=1}^{\infty}$. By definition of real exponentiation, we have

$$x^{q+r} = \lim_{n\to\infty} x^{q_n+r_n}; \quad x^q = \lim_{n\to\infty} x^{q_n}; \quad x^r = \lim_{n\to\infty} x^{r_n}.$$

But by Lemma 5.6.9(b) (applied to *rational* exponents) we have $x^{q_n+r_n} = x^{q_n}x^{r_n}$. Thus by limit laws we have $x^{q+r} = x^q x^r$, as desired. □

— Exercises —

Exercise 6.7.1 Prove the remaining components of Proposition 6.7.3.

Chapter 7
Series

Now that we have developed a reasonable theory of limits of sequences, we will use that theory to develop a theory of infinite series

$$\sum_{n=m}^{\infty} a_n = a_m + a_{m+1} + a_{m+2} + \ldots .$$

But before we develop infinite series, we must first develop the theory of finite series.

7.1 Finite Series

Definition 7.1.1 *(Finite series)* Let m, n be integers, and let $(a_i)_{i=m}^n$ be a finite sequence of real numbers, assigning a real number a_i to each integer i between m and n inclusive (i.e., $m \leq i \leq n$). Then we define the finite sum (or finite series) $\sum_{i=m}^n a_i$ by the recursive formula

$$\sum_{i=m}^{n} a_i := 0 \text{ whenever } n < m;$$

$$\sum_{i=m}^{n+1} a_i := \left(\sum_{i=m}^{n} a_i\right) + a_{n+1} \text{ whenever } n \geq m - 1.$$

Thus for instance we have the identities

$$\sum_{i=m}^{m-2} a_i = 0; \quad \sum_{i=m}^{m-1} a_i = 0; \quad \sum_{i=m}^{m} a_i = a_m;$$

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_7

$$\sum_{i=m}^{m+1} a_i = a_m + a_{m+1}; \quad \sum_{i=m}^{m+2} a_i = a_m + a_{m+1} + a_{m+2}$$

(why?). Because of this, we sometimes express $\sum_{i=m}^{n} a_i$ less formally as

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + \ldots + a_n.$$

Remark 7.1.2 The difference between "sum" and "series" is a subtle linguistic one. Strictly speaking, a series is an *expression* of the form $\sum_{i=m}^{n} a_i$; this series is mathematically (but not semantically) equal to a real number, which is then the *sum* of that series. For instance, $1 + 2 + 3 + 4 + 5$ is a series, whose sum is 15; if one were to be very picky about semantics, one would not consider 15 a series and one would not consider $1 + 2 + 3 + 4 + 5$ a sum, despite the two expressions having the same value. However, we will not be very careful about this distinction as it is purely linguistic and has no bearing on the mathematics; the expressions $1 + 2 + 3 + 4 + 5$ and 15 are the same number, and thus *mathematically* interchangeable, in the sense of the axiom of substitution (see Sect. A.7), even if they are not semantically interchangeable.

Remark 7.1.3 Note that the variable i (sometimes called the *index of summation*) is a *bound variable* (sometimes called a *dummy variable*); the expression $\sum_{i=m}^{n} a_i$ does not actually depend on any quantity named i. In particular, one can replace the index of summation i with any other symbol, and obtain the same sum:

$$\sum_{i=m}^{n} a_i = \sum_{j=m}^{n} a_j.$$

We list some basic properties of summation below.

Lemma 7.1.4 *(a) Let $m \leq n < p$ be integers, and let a_i be a real number assigned to each integer $m \leq i \leq p$. Then we have*

$$\sum_{i=m}^{n} a_i + \sum_{i=n+1}^{p} a_i = \sum_{i=m}^{p} a_i.$$

(b) Let $m \leq n$ be integers, k be another integer, and let a_i be a real number assigned to each integer $m \leq i \leq n$. Then we have

$$\sum_{i=m}^{n} a_i = \sum_{j=m+k}^{n+k} a_{j-k}.$$

(c) Let $m \leq n$ be integers, and let a_i, b_i be real numbers assigned to each integer $m \leq i \leq n$. Then we have

$$\sum_{i=m}^{n}(a_i + b_i) = \left(\sum_{i=m}^{n} a_i\right) + \left(\sum_{i=m}^{n} b_i\right).$$

(d) Let $m \leq n$ be integers, and let a_i be a real number assigned to each integer $m \leq i \leq n$, and let c be another real number. Then we have

$$\sum_{i=m}^{n}(ca_i) = c\left(\sum_{i=m}^{n} a_i\right).$$

(e) (Triangle inequality for finite series) Let $m \leq n$ be integers, and let a_i be a real number assigned to each integer $m \leq i \leq n$. Then we have

$$\left|\sum_{i=m}^{n} a_i\right| \leq \sum_{i=m}^{n} |a_i|.$$

(f) (Comparison test for finite series) Let $m \leq n$ be integers, and let a_i, b_i be real numbers assigned to each integer $m \leq i \leq n$. Suppose that $a_i \leq b_i$ for all $m \leq i \leq n$. Then we have

$$\sum_{i=m}^{n} a_i \leq \sum_{i=m}^{n} b_i.$$

Proof See Exercise 7.1.1. □

Remark 7.1.5 In the future we may omit some of the parentheses in series expressions, for instance we may write $\sum_{i=m}^{n}(a_i + b_i)$ simply as $\sum_{i=m}^{n} a_i + b_i$. This is reasonably safe from being mis-interpreted, because the alternative interpretation $(\sum_{i=m}^{n} a_i) + b_i$ does not make any sense (the index i in b_i is meaningless outside of the summation, since i is only a dummy variable).

One can use finite series to also define summations over finite sets:

Definition 7.1.6 *(Summations over finite sets)* Let X be a finite set with n elements (where $n \in \mathbf{N}$), and let $f: X \to \mathbf{R}$ be a function from X to the real numbers (i.e., f assigns a real number $f(x)$ to each element x of X). Then we can define the finite sum $\sum_{x \in X} f(x)$ as follows. We first select any bijection g from $\{i \in \mathbf{N} : 1 \leq i \leq n\}$ to X; such a bijection exists since X is assumed to have n elements. We then define

$$\sum_{x \in X} f(x) := \sum_{i=1}^{n} f(g(i)).$$

The same definition also permits us to define $\sum_{x\in X} f(x)$ when f is defined on a larger set Y than X.

Example 7.1.7 Let X be the three-element set $X := \{a, b, c\}$, where a, b, c are distinct objects, and let $f: X \to \mathbf{R}$ be the function $f(a) := 2$, $f(b) := 5$, $f(c) := -1$. In order to compute the sum $\sum_{x\in X} f(x)$, we select a bijection $g: \{1, 2, 3\} \to X$, e.g., $g(1) := a$, $g(2) := b$, $g(3) := c$. We then have

$$\sum_{x\in X} f(x) = \sum_{i=1}^{3} f(g(i)) = f(a) + f(b) + f(c) = 6.$$

One could pick another bijection from $\{1, 2, 3\}$ to X, e.g., $h(1) := c$, $h(2) := b$, $h(3) = a$, but the end result is still the same:

$$\sum_{x\in X} f(x) = \sum_{i=1}^{3} f(h(i)) = f(c) + f(b) + f(a) = 6.$$

To verify that this definition actually does give a single, well-defined value to $\sum_{x\in X} f(x)$, one has to check that different bijections g from $\{i \in \mathbf{N} : 1 \le i \le n\}$ to X give the same sum. In other words, we must prove

Proposition 7.1.8 (Finite summations are well-defined) *Let X be a finite set with n elements* (*where $n \in \mathbf{N}$*)*, let $f: X \to \mathbf{R}$ be a function, and let $g: \{i \in \mathbf{N} : 1 \le i \le n\} \to X$ and $h: \{i \in \mathbf{N} : 1 \le i \le n\} \to X$ be bijections. Then we have*

$$\sum_{i=1}^{n} f(g(i)) = \sum_{i=1}^{n} f(h(i)).$$

Remark 7.1.9 The issue is somewhat more complicated when summing over infinite sets; see Section 8.2.

Proof We use induction on n; more precisely, we let $P(n)$ be the assertion that "For any set X of n elements, any function $f: X \to \mathbf{R}$, and any two bijections g, h from $\{i \in \mathbf{N} : 1 \le i \le n\}$ to X, we have $\sum_{i=1}^{n} f(g(i)) = \sum_{i=1}^{n} f(h(i))$". (More informally, $P(n)$ is the assertion that Proposition 7.1.8 is true for that value of n.) We want to prove that $P(n)$ is true for all natural numbers n.

We first check the base case $P(0)$. In this case $\sum_{i=1}^{0} f(g(i))$ and $\sum_{i=1}^{0} f(h(i))$ both equal to 0, by definition of finite series, so we are done.

Now suppose inductively that $P(n)$ is true; we now prove that $P(n+1)$ is true. Thus, let X be a set with $n+1$ elements, let $f: X \to \mathbf{R}$ be a function, and let g and h be bijections from $\{i \in \mathbf{N} : 1 \le i \le n+1\}$ to X. We have to prove that

$$\sum_{i=1}^{n+1} f(g(i)) = \sum_{i=1}^{n+1} f(h(i)). \tag{7.1}$$

Let $x := g(n+1)$; thus x is an element of X. By definition of finite series, we can expand the left-hand side of (7.1) as

$$\sum_{i=1}^{n+1} f(g(i)) = \left(\sum_{i=1}^{n} f(g(i))\right) + f(x).$$

Now let us look at the right-hand side of (7.1). Ideally we would like to have $h(n+1)$ also equal to x—this would allow us to use the inductive hypothesis $P(n)$ much more easily—but we cannot assume this. However, since h is a bijection, we do know that there is *some* index j, with $1 \le j \le n+1$, for which $h(j) = x$. We now use Lemma 7.1.4 and the definition of finite series to write

$$\begin{aligned}\sum_{i=1}^{n+1} f(h(i)) &= \left(\sum_{i=1}^{j} f(h(i))\right) + \left(\sum_{i=j+1}^{n+1} f(h(i))\right)\\ &= \left(\sum_{i=1}^{j-1} f(h(i))\right) + f(h(j)) + \left(\sum_{i=j+1}^{n+1} f(h(i))\right)\\ &= \left(\sum_{i=1}^{j-1} f(h(i))\right) + f(x) + \left(\sum_{i=j}^{n} f(h(i+1))\right).\end{aligned}$$

We now define the function $\tilde{h}\colon \{i \in \mathbf{N} : 1 \le i \le n\} \to X - \{x\}$ by setting $\tilde{h}(i) := h(i)$ when $i < j$ and $\tilde{h}(i) := h(i+1)$ when $i \ge j$. We can thus write the right-hand side of (7.1) as

$$= \left(\sum_{i=1}^{j-1} f(\tilde{h}(i))\right) + f(x) + \left(\sum_{i=j}^{n} f(\tilde{h}(i))\right) \quad = \left(\sum_{i=1}^{n} f(\tilde{h}(i))\right) + f(x)$$

where we have used Lemma 7.1.4 once again. Thus to finish the proof of (7.1) we have to show that

$$\sum_{i=1}^{n} f(g(i)) = \sum_{i=1}^{n} f(\tilde{h}(i)). \tag{7.2}$$

But the function g (when restricted to $\{i \in \mathbf{N} : 1 \le i \le n\}$) is a bijection from $\{i \in \mathbf{N} : 1 \le i \le n\} \to X - \{x\}$ (why?). The function $\tilde{h}$ is also a bijection from $\{i \in \mathbf{N} : 1 \le i \le n\} \to X - \{x\}$ (why? cf. Lemma 3.6.9). Since $X - \{x\}$ has n elements (by Lemma 3.6.9), the claim 7.2 then follows directly from the induction hypothesis $P(n)$. □

Remark 7.1.10 Suppose that X is a set, that $P(x)$ is a property pertaining to an element x of X, and $f\colon \{y \in X : P(y) \text{ is true}\} \to \mathbf{R}$ is a function. Then we will often abbreviate

$$\sum_{x \in \{y \in X : P(y) \text{ is true}\}} f(x)$$

as $\sum_{x \in X : P(x) \text{ is true}} f(x)$ or even as $\sum_{P(x) \text{ is true}} f(x)$ when there is no chance of confusion. For instance, $\sum_{n \in \mathbf{N} : 2 \leq n \leq 4} f(n)$ or $\sum_{2 \leq n \leq 4} f(n)$ are both short-hand for $\sum_{n \in \{2,3,4\}} f(n) = f(2) + f(3) + f(4)$. (This convention is currently limited to cases in which $\{y \in X : P(y) \text{ is true}\}$ is finite, but in later sections we will also define sums over infinite sets, in which case this convention will also extend to such settings.)

The following properties of summation on finite sets are fairly obvious but do require a rigorous proof:

Proposition 7.1.11 (Basic properties of summation over finite sets)

(a) If X is empty, and $f : X \to \mathbf{R}$ is a function (i.e., f is the empty function), we have

$$\sum_{x \in X} f(x) = 0.$$

(b) If X consists of a single element, $X = \{x_0\}$, and $f : X \to \mathbf{R}$ is a function, we have

$$\sum_{x \in X} f(x) = f(x_0).$$

(c) (Substitution, part I) If X is a finite set, $f : X \to \mathbf{R}$ is a function, and $g : Y \to X$ is a bijection, then

$$\sum_{x \in X} f(x) = \sum_{y \in Y} f(g(y)).$$

(d) (Substitution, part II) Let $n \leq m$ be integers, and let X be the set $X := \{i \in \mathbf{Z} : n \leq i \leq m\}$. If a_i is a real number assigned to each integer $i \in X$, then we have

$$\sum_{i=n}^{m} a_i = \sum_{i \in X} a_i.$$

(e) Let X, Y be disjoint finite sets (so $X \cap Y = \emptyset$), and $f : X \cup Y \to \mathbf{R}$ is a function. Then we have

$$\sum_{z \in X \cup Y} f(z) = \left(\sum_{x \in X} f(x) \right) + \left(\sum_{y \in Y} f(y) \right).$$

(f) (Linearity, part I) Let X be a finite set, and let $f : X \to \mathbf{R}$ and $g : X \to \mathbf{R}$ be functions. Then

$$\sum_{x \in X} (f(x) + g(x)) = \sum_{x \in X} f(x) + \sum_{x \in X} g(x).$$

(g) (*Linearity, part II*) *Let X be a finite set, let $f\colon X \to \mathbf{R}$ be a function, and let c be a real number. Then*

$$\sum_{x\in X} cf(x) = c\sum_{x\in X} f(x).$$

(h) (*Monotonicity*) *Let X be a finite set, and let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be functions such that $f(x) \le g(x)$ for all $x \in X$. Then we have*

$$\sum_{x\in X} f(x) \le \sum_{x\in X} g(x).$$

(i) (*Triangle inequality*) *Let X be a finite set, and let $f\colon X \to \mathbf{R}$ be a function, then*

$$|\sum_{x\in X} f(x)| \le \sum_{x\in X} |f(x)|.$$

Proof See Exercise 7.1.2. □

Remark 7.1.12 The substitution rule in Proposition 7.1.11(c) can be thought of as making the substitution $x := g(y)$ (hence the name). Note that the assumption that g is a bijection is essential; can you see why the rule may fail when g is not one-to-one or not onto? From Proposition 7.1.11(c) and (d) we see that

$$\sum_{i=n}^{m} a_i = \sum_{i=n}^{m} a_{f(i)}$$

for any bijection f from the set $\{i \in \mathbf{Z} : n \le i \le m\}$ to itself. Informally, this means that we can rearrange the elements of a finite sequence at will and still obtain the same value.

Now we look at double finite series—finite series of finite series—and how they connect with Cartesian products.

Lemma 7.1.13 *Let X, Y be finite sets, and let $f\colon X \times Y \to \mathbf{R}$ be a function. Then*

$$\sum_{x\in X}\left(\sum_{y\in Y} f(x, y)\right) = \sum_{(x,y)\in X\times Y} f(x, y).$$

Proof Let n be the number of elements in X. We will use induction on n (cf. Proposition 7.1.8); i.e., we let $P(n)$ be the assertion that Lemma 7.1.13 is true for any set X with n elements, and any finite set Y and any function $f\colon X \times Y \to \mathbf{R}$. We wish to prove $P(n)$ for all natural numbers n.

The base case $P(0)$ is easy, following from Proposition 7.1.11(a) (why?). Now suppose that $P(n)$ is true; we now show that $P(n+1)$ is true. Let X be a set with $n+1$ elements. In particular, by Lemma 3.6.9, we can write $X = X' \cup \{x_0\}$, where

x_0 is an element of X and $X' := X - \{x_0\}$ has n elements. Then by Proposition 7.1.11(e) we have

$$\sum_{x\in X}\left(\sum_{y\in Y} f(x, y)\right) = \left(\sum_{x\in X'}\left(\sum_{y\in Y} f(x, y)\right)\right) + \left(\sum_{y\in Y} f(x_0, y)\right);$$

by the induction hypothesis this is equal to

$$\sum_{(x,y)\in X'\times Y} f(x, y) + \left(\sum_{y\in Y} f(x_0, y)\right).$$

By Proposition 7.1.11(c) this is equal to

$$\sum_{(x,y)\in X'\times Y} f(x, y) + \left(\sum_{(x,y)\in\{x_0\}\times Y} f(x, y)\right).$$

By Proposition 7.1.11(e) this is equal to

$$\sum_{(x,y)\in X\times Y} f(x, y)$$

(why?) as desired. □

Corollary 7.1.14 (Fubini's theorem for finite series) *Let X, Y be finite sets, and let $f: X \times Y \to \mathbf{R}$ be a function. Then*

$$\begin{aligned}\sum_{x\in X}\left(\sum_{y\in Y} f(x, y)\right) &= \sum_{(x,y)\in X\times Y} f(x, y)\\ &= \sum_{(y,x)\in Y\times X} f(x, y)\\ &= \sum_{y\in Y}\left(\sum_{x\in X} f(x, y)\right).\end{aligned}$$

Proof In light of Lemma 7.1.13, it suffices to show that

$$\sum_{(x,y)\in X\times Y} f(x, y) = \sum_{(y,x)\in Y\times X} f(x, y).$$

But this follows from Proposition 7.1.11(c) by applying the bijection $h\colon Y \times X \to X \times Y$ defined by $h(y, x) := (x, y)$. (Why is this a bijection, and why does Proposition 7.1.11(c) give us what we want?) □

Remark 7.1.15 This should be contrasted with Example 1.2.5; thus we anticipate something interesting to happen when we move from finite sums to infinite sums. However, see Theorem 8.2.2.

— Exercises —

Exercise 7.1.1 Prove Lemma 7.1.4. (*Hint:* you will need to use induction, but the base case might not necessarily be at 0.)

Exercise 7.1.2 Prove Proposition 7.1.11. (*Hint:* this is not as lengthy as it may first appear. It is largely a matter of choosing the right bijections to turn these sums over sets into finite series, and then applying Lemma 7.1.4.)

Exercise 7.1.3 Form a definition for the finite products $\prod_{i=1}^{n} a_i$ and $\prod_{x \in X} f(x)$. Which of the above results for finite series have analogues for finite products? (Note that it is dangerous to apply logarithms because some of the a_i or $f(x)$ could be zero or negative. Besides, we haven't defined logarithms yet.)

Exercise 7.1.4 Define the *factorial function* $n!$ for natural numbers n by the recursive definition $0! := 1$ and $(n+1)! := n! \times (n+1)$. If x and y are real numbers, prove the *binomial formula*

$$(x+y)^n = \sum_{j=0}^{n} \frac{n!}{j!(n-j)!} x^j y^{n-j}$$

for all natural numbers n. (*Hint:* induct on n.)

Exercise 7.1.5 Let X be a finite set, let m be an integer, and for each $x \in X$ let $(a_n(x))_{n=m}^{\infty}$ be a convergent sequence of real numbers. Show that the sequence $(\sum_{x \in X} a_n(x))_{n=m}^{\infty}$ is convergent, and

$$\lim_{n \to \infty} \sum_{x \in X} a_n(x) = \sum_{x \in X} \lim_{n \to \infty} a_n(x).$$

(*Hint:* induct on the cardinality of X, and use Theorem 6.1.19(a).) Thus we may always interchange finite sums with convergent limits. Things however get trickier with infinite sums; see Corollary 8.2.11 of *Analysis II.*

Exercise 7.1.6 Let I be a finite set, and for each $i \in I$, let E_i be a finite set. Suppose that the E_i are pairwise disjoint, which means that $E_i \cap E_j = \emptyset$ whenever $i, j \in I$ are distinct. For each $x \in \bigcup_{i \in I} E_i$, let $f(x)$ be a real number. Show that $\sum_{x \in \bigcup_{i \in I} E_i} f(x) = \sum_{i \in I} \sum_{x \in E_i} f(x)$.

Exercise 7.1.7 Let n, m be natural numbers, and for each $1 \leq i \leq n$ let a_i be a natural number with $a_i \leq m$. Establish the identity

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} \#(\{1 \leq i \leq n : a_i \geq j\}).$$

(*Hint:* apply Corollary 7.1.14 to compute a sum $\sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j}$ in two different ways, for a well chosen choice of summands $c_{i,j}$.) Use of identities such as this is known as the *double counting method*, and is often useful in combinatorics.

7.2 Infinite Series

We are now ready to sum infinite series.

Definition 7.2.1 *(Formal infinite series)* A (formal) infinite series is any expression of the form

$$\sum_{n=m}^{\infty} a_n,$$

where m is an integer, and a_n is a real number for any integer $n \geq m$. We sometimes write this series as

$$a_m + a_{m+1} + a_{m+2} + \ldots.$$

At present, this series is only defined *formally*; we have not set this sum equal to any real number; the notation $a_m + a_{m+1} + a_{m+2} + \ldots$ is of course designed to look very suggestively like a sum, but is not actually a finite sum because of the "..." symbol. To rigorously define what the series actually sums to, we need another definition.

Definition 7.2.2 *(Convergence of series)* Let $\sum_{n=m}^{\infty} a_n$ be a formal infinite series. For any integer $N \geq m$, we define the N^{th} *partial sum* S_N of this series to be $S_N := \sum_{n=m}^{N} a_n$; of course, S_N is a real number. If the sequence $(S_N)_{N=m}^{\infty}$ converges to some limit L as $N \to \infty$, then we say that the infinite series $\sum_{n=m}^{\infty} a_n$ is *convergent*, and *converges to* L; we also write $L = \sum_{n=m}^{\infty} a_n$, and say that L is the *sum* of the infinite series $\sum_{n=m}^{\infty} a_n$. If the partial sums S_N diverge, then we say that the infinite series $\sum_{n=m}^{\infty} a_n$ is *divergent*, and we do not assign any real number value to that series.

Remark 7.2.3 Note that Proposition 6.1.7 shows that if a series converges, then it has a unique sum, so it is safe to talk about *the* sum $L = \sum_{n=m}^{\infty} a_n$ of a convergent series.

Example 7.2.4 Consider the formal infinite series

$$\sum_{n=1}^{\infty} 2^{-n} = 2^{-1} + 2^{-2} + 2^{-3} + \cdots.$$

The partial sums can be verified to equal

$$S_N = \sum_{n=1}^{N} 2^{-n} = 1 - 2^{-N}$$

by an easy induction argument (or by Lemma 7.3.3); the sequence $1 - 2^{-N}$ converges to 1 as $N \to \infty$, and hence we have

$$\sum_{n=1}^{\infty} 2^{-n} = 1.$$

In particular, this series is convergent. On the other hand, if we consider the series

$$\sum_{n=1}^{\infty} 2^{n} = 2^1 + 2^2 + 2^3 + \cdots$$

then the partial sums are

$$S_N = \sum_{n=1}^{N} 2^{n} = 2^{N+1} - 2$$

and this is easily shown to be an unbounded sequence, and hence divergent. Thus the series $\sum_{n=1}^{\infty} 2^n$ is divergent.

Now we address the question of when a series converges. The following proposition shows that a series converges iff the "tail" of the sequence is eventually less than ε for any $\varepsilon > 0$:

Proposition 7.2.5 *Let $\sum_{n=m}^{\infty} a_n$ be a formal series of real numbers. Then $\sum_{n=m}^{\infty} a_n$ converges if and only if, for every real number $\varepsilon > 0$, there exists an integer $N \geq m$ such that*

$$\left| \sum_{n=p}^{q} a_n \right| \leq \varepsilon \textit{ for all } p, q \geq N.$$

Proof See Exercise 7.2.2. □

This proposition, by itself, is not very handy, because it is not so easy to compute the partial sums $\sum_{n=p}^{q} a_n$ in practice. However, it has a number of useful corollaries. For instance.

Corollary 7.2.6 (Zero test) *Let $\sum_{n=m}^{\infty} a_n$ be a convergent series of real numbers. Then we must have $\lim_{n\to\infty} a_n = 0$. To put this another way, if $\lim_{n\to\infty} a_n$ is non-zero or divergent, then the series $\sum_{n=m}^{\infty} a_n$ is divergent.*

Proof See Exercise 7.2.3. □

Example 7.2.7 The sequence $a_n := 1$ does not converge to 0 as $n \to \infty$, so we know that $\sum_{n=1}^{\infty} 1$ is a divergent series. (Note however that $1, 1, 1, 1, \ldots$ is a convergent *sequence*; convergence of series is a different notion from convergence of sequences.) Similarly, the sequence $a_n := (-1)^n$ diverges, and in particular does not converge to zero; thus the series $\sum_{n=1}^{\infty} (-1)^n$ is also divergent.

If a sequence $(a_n)_{n=m}^{\infty}$ *does* converge to zero, then the series $\sum_{n=m}^{\infty} a_n$ may or may not be convergent; it depends on the series. For instance, we will soon see that the series $\sum_{n=1}^{\infty} 1/n$ is divergent despite the fact that $1/n$ converges to 0 as $n \to \infty$.

Definition 7.2.8 *(Absolute convergence)* Let $\sum_{n=m}^{\infty} a_n$ be a formal series of real numbers. We say that this series is *absolutely convergent* iff the series $\sum_{n=m}^{\infty} |a_n|$ is convergent.

Proposition 7.2.9 (Absolute convergence test) *Let $\sum_{n=m}^{\infty} a_n$ be a formal series of real numbers. If this series is absolutely convergent, then it is also convergent. Furthermore, in this case we have the triangle inequality*

$$\left|\sum_{n=m}^{\infty} a_n\right| \leq \sum_{n=m}^{\infty} |a_n|.$$

Proof See Exercise 7.2.4. □

Remark 7.2.10 The converse to this proposition is not true; there exist series which are convergent but not absolutely convergent. See Example 7.2.12. Series that are convergent but not absolutely convergent are also known as *conditionally convergent* series.

Proposition 7.2.11 (Alternating series test) *Let $(a_n)_{n=m}^{\infty}$ be a sequence of real numbers which are non-negative and decreasing, thus $a_n \geq 0$ and $a_n \geq a_{n+1}$ for every $n \geq m$. Then the series $\sum_{n=m}^{\infty}(-1)^n a_n$ is convergent if and only if the sequence a_n converges to 0 as $n \to \infty$.*

Proof From the zero test, we know that if $\sum_{n=m}^{\infty}(-1)^n a_n$ is a convergent series, then the sequence $((-1)^n a_n)_{n=m}^{\infty}$ converges to 0, which implies that $(a_n)_{n=m}^{\infty}$ also converges to 0, since $(-1)^n a_n$ and a_n have the same distance from 0.

Now suppose conversely that $(a_n)_{n=m}^{\infty}$ converges to 0. For each $N \geq m$, let S_N be the partial sum $S_N := \sum_{n=m}^{N}(-1)^n a_n$; our job is to show that $(S_N)_{N=m}^{\infty}$ converges. Observe that

$$\begin{aligned} S_{N+2} &= S_N + (-1)^{N+1} a_{N+1} + (-1)^{N+2} a_{N+2} \\ &= S_N + (-1)^{N+1}(a_{N+1} - a_{N+2}). \end{aligned}$$

But by hypothesis, $(a_{N+1} - a_{N+2})$ is non-negative. Thus we have $S_{N+2} \geq S_N$ when N is odd and $S_{N+2} \leq S_N$ if N is even.

Now suppose that N is even. From the above discussion and induction we see that $S_{N+2k} \leq S_N$ for all natural numbers k (why?). Also we have $S_{N+2k+1} \geq S_{N+1} = S_N - a_{N+1}$ (why?). Finally, we have $S_{N+2k+1} = S_{N+2k} - a_{N+2k+1} \leq S_{N+2k}$ (why?). Thus we have

$$S_N - a_{N+1} \leq S_{N+2k+1} \leq S_{N+2k} \leq S_N$$

for all k. In particular, we have

$$S_N - a_{N+1} \leq S_n \leq S_N \text{ for all } n \geq N$$

(why?). In particular, the sequence S_n is eventually a_{N+1}-steady. But the sequence $(a_N)_{N=m}^{\infty}$ converges to 0 as $N \to \infty$, thus this implies that S_n is eventually ε-steady for every $\varepsilon > 0$ (why?). Thus $(S_n)_{n=m}^{\infty}$ converges, and so the series $\sum_{n=m}^{\infty}(-1)^n a_n$ is convergent. □

Example 7.2.12 The sequence $(1/n)_{n=1}^{\infty}$ is non-negative, decreasing, and converges to zero. Thus $\sum_{n=1}^{\infty}(-1)^n/n$ is convergent (but it is not absolutely convergent, because $\sum_{n=1}^{\infty} 1/n$ diverges, see Corollary 7.3.7). Thus lack of absolute convergence does not imply lack of convergence, even though absolute convergence implies convergence.

Some basic identities concerning convergent series are collected below.

Proposition 7.2.13 (Series laws)

(a) If $\sum_{n=m}^{\infty} a_n$ is a series of real numbers converging to x, and $\sum_{n=m}^{\infty} b_n$ is a series of real numbers converging to y, then $\sum_{n=m}^{\infty}(a_n + b_n)$ is also a convergent series, and converges to $x + y$. In particular, we have

$$\sum_{n=m}^{\infty}(a_n + b_n) = \sum_{n=m}^{\infty} a_n + \sum_{n=m}^{\infty} b_n.$$

(b) If $\sum_{n=m}^{\infty} a_n$ is a series of real numbers converging to x, and c is a real number, then $\sum_{n=m}^{\infty}(ca_n)$ is also a convergent series, and converges to cx. In particular, we have

$$\sum_{n=m}^{\infty}(ca_n) = c\sum_{n=m}^{\infty} a_n.$$

(c) Let $\sum_{n=m}^{\infty} a_n$ be a series of real numbers, and let $k \geq 0$ be an integer. If one of the two series $\sum_{n=m}^{\infty} a_n$ and $\sum_{n=m+k}^{\infty} a_n$ are convergent, then the other one is also, and we have the identity

$$\sum_{n=m}^{\infty} a_n = \sum_{n=m}^{m+k-1} a_n + \sum_{n=m+k}^{\infty} a_n.$$

(d) Let $\sum_{n=m}^{\infty} a_n$ be a series of real numbers converging to x, and let k be an integer. Then $\sum_{n=m+k}^{\infty} a_{n-k}$ also converges to x.

Proof See Exercise 7.2.5. □

From Proposition 7.2.13(c) we see that the convergence of a series does not depend on the first few elements of the series (though of course those elements do influence which value the series converges to). Because of this, we will usually not pay much attention as to what the initial index m of the series is.

There is one type of series, called *telescoping series*, which are easy to sum:

Lemma 7.2.14 (Telescoping series) *Let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers which converge to 0, i.e., $\lim_{n\to\infty} a_n = 0$. Then the series $\sum_{n=0}^{\infty}(a_n - a_{n+1})$ converges to a_0.*

Proof See Exercise 7.2.6. □

— Exercises —

Exercise 7.2.1 Is the series $\sum_{n=1}^{\infty}(-1)^n$ convergent or divergent? Justify your answer. Can you now resolve the difficulty in Example 1.2.2?

Exercise 7.2.2 Prove Proposition 7.2.5. (*Hint:* use Proposition 6.1.12 and Theorem 6.4.18.)

Exercise 7.2.3 Use Proposition 7.2.5 to prove Corollary 7.2.6.

Exercise 7.2.4 Prove Proposition 7.2.9. (*Hint:* use Proposition 7.2.5 and Proposition 7.1.4(e).)

Exercise 7.2.5 Prove Proposition 7.2.13. (*Hint:* use Theorem 6.1.19.)

Exercise 7.2.6 Prove Lemma 7.2.14. (*Hint:* First work out what the partial sums $\sum_{n=0}^{N}(a_n - a_{n+1})$ should be, and prove your assertion using induction.) How does the proposition change if we assume that a_n does not converge to zero, but instead converges to some other real number L?

7.3 Sums of Non-negative Numbers

Now we specialize the preceding discussion in order to consider sums $\sum_{n=m}^{\infty} a_n$ where all the terms a_n are non-negative. This situation comes up, for instance, from the absolute convergence test, since the absolute value $|a_n|$ of a real number a_n is always non-negative. Note that when all the terms in a series are non-negative, there is no distinction between convergence and absolute convergence.

Suppose $\sum_{n=m}^{\infty} a_n$ is a series of non-negative numbers. Then the partial sums $S_N := \sum_{n=m}^{N} a_n$ are increasing, i.e., $S_{N+1} \geq S_N$ for all $N \geq m$ (why?). From Proposition 6.3.8 and Corollary 6.1.17, we thus see that the sequence $(S_N)_{N=m}^{\infty}$ is convergent if and only if it has an upper bound M. In other words, we have just shown

Proposition 7.3.1 *Let $\sum_{n=m}^{\infty} a_n$ be a formal series of non-negative real numbers. Then this series is convergent if and only if there is a real number M such that*

$$\sum_{n=m}^{N} a_n \leq M \text{ for all integers } N \geq m.$$

A simple corollary of this is

Corollary 7.3.2 (Comparison test) *Let $\sum_{n=m}^{\infty} a_n$ and $\sum_{n=m}^{\infty} b_n$ be two formal series of real numbers, and suppose that $|a_n| \leq b_n$ for all $n \geq m$. Then if $\sum_{n=m}^{\infty} b_n$ is convergent, then $\sum_{n=m}^{\infty} a_n$ is absolutely convergent, and in fact*

$$\left|\sum_{n=m}^{\infty} a_n\right| \leq \sum_{n=m}^{\infty} |a_n| \leq \sum_{n=m}^{\infty} b_n.$$

Proof See Exercise 7.3.1. □

We can also run the comparison test in the contrapositive: if we have $|a_n| \leq b_n$ for all $n \geq m$, and $\sum_{n=m}^{\infty} a_n$ is not absolutely convergent, then $\sum_{n=m}^{\infty} b_n$ is not convergent. (Why does this follow immediately from Corollary 7.3.2?)

A useful series to use in the comparison test is the *geometric series*

$$\sum_{n=0}^{\infty} x^n,$$

where x is some real number:

Lemma 7.3.3 (Geometric series) *Let x be a real number. If $|x| \geq 1$, then the series $\sum_{n=0}^{\infty} x^n$ is divergent. If however $|x| < 1$, then the series is absolutely convergent and*

$$\sum_{n=0}^{\infty} x^n = 1/(1-x).$$

Proof See Exercise 7.3.2. □

We now give a useful criterion, known as the *Cauchy criterion*, to test whether a series of non-negative but decreasing terms is convergent.

Proposition 7.3.4 (Cauchy criterion) *Let $(a_n)_{n=1}^{\infty}$ be a decreasing sequence of non-negative real numbers (so $a_n \geq 0$ and $a_{n+1} \leq a_n$ for all $n \geq 1$). Then the series $\sum_{n=1}^{\infty} a_n$ is convergent if and only if the series*

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + 8a_8 + \ldots$$

is convergent.

Remark 7.3.5 An interesting feature of this criterion is that it only uses a small number of elements of the sequence a_n (namely, those elements whose index n is a power of 2, $n = 2^k$) in order to determine whether the whole series is convergent or not.

Proof Let $S_N := \sum_{n=1}^{N} a_n$ be the partial sums of $\sum_{n=1}^{\infty} a_n$, and let $T_K := \sum_{k=0}^{K} 2^k a_{2^k}$ be the partial sums of $\sum_{k=0}^{\infty} 2^k a_{2^k}$. In light of Proposition 7.3.1, our task is to show that the sequence $(S_N)_{N=1}^{\infty}$ is bounded if and only if the sequence $(T_K)_{K=0}^{\infty}$ is bounded. To do this we need the following claim: □

Lemma 7.3.6 *For any natural number K, we have $S_{2^{K+1}-1} \leq T_K \leq 2S_{2^K}$.*

Proof We use induction on K. First we prove the claim when $K = 0$, i.e.

$$S_1 \le T_0 \le 2S_1.$$

This becomes

$$a_1 \le a_1 \le 2a_1$$

which is clearly true, since a_1 is non-negative.

Now suppose the claim has been proven for K, and now we try to prove it for $K + 1$:

$$S_{2^{K+2}-1} \le T_{K+1} \le 2S_{2^{K+1}}.$$

Clearly we have

$$T_{K+1} = T_K + 2^{K+1} a_{2^{K+1}}.$$

Also, we have (using Lemma 7.1.4(a) and (f), and the hypothesis that the a_n are decreasing)

$$S_{2^{K+1}} = S_{2^K} + \sum_{n=2^K+1}^{2^{K+1}} a_n \ge S_{2^K} + \sum_{n=2^K+1}^{2^{K+1}} a_{2^{K+1}} = S_{2^K} + 2^K a_{2^{K+1}}$$

and hence

$$2S_{2^{K+1}} \ge 2S_{2^K} + 2^{K+1} a_{2^{K+1}}.$$

Similarly we have

$$\begin{aligned} S_{2^{K+2}-1} &= S_{2^{K+1}-1} + \sum_{n=2^{K+1}}^{2^{K+2}-1} a_n \\ &\le S_{2^{K+1}-1} + \sum_{n=2^{K+1}}^{2^{K+2}-1} a_{2^{K+1}} \\ &= S_{2^{K+1}-1} + 2^{K+1} a_{2^{K+1}}. \end{aligned}$$

Combining these inequalities with the induction hypothesis

$$S_{2^{K+1}-1} \le T_K \le 2S_{2^K}$$

we obtain

$$S_{2^{K+2}-1} \le T_{K+1} \le 2S_{2^{K+1}}$$

as desired. This proves the claim.

From this claim we see that if $(S_N)_{N=1}^{\infty}$ is bounded, then $(S_{2^K})_{K=0}^{\infty}$ is bounded, and hence $(T_K)_{K=0}^{\infty}$ is bounded. Conversely, if $(T_K)_{K=0}^{\infty}$ is bounded, then the claim implies that $S_{2^{K+1}-1}$ is bounded, i.e., there is an M such that $S_{2^{K+1}-1} \leq M$ for all natural numbers K. But one can easily show (using induction) that $2^{K+1} - 1 \geq K + 1$, and hence that $S_{K+1} \leq M$ for all natural numbers K, hence $(S_N)_{N=1}^{\infty}$ is bounded. □

Corollary 7.3.7 *Let $q > 0$ be a real number. Then the series $\sum_{n=1}^{\infty} 1/n^q$ is convergent when $q > 1$ and divergent when $q \leq 1$.*

Proof The sequence $(1/n^q)_{n=1}^{\infty}$ is non-negative and decreasing (by Lemma 5.6.9(d) and Lemma 6.7.3), and so the Cauchy criterion applies. Thus this series is convergent if and only if

$$\sum_{k=0}^{\infty} 2^k \frac{1}{(2^k)^q}$$

is convergent. But by the laws of exponentiation (Lemma 5.6.9 and Lemma 6.7.3) we can rewrite this as the geometric series

$$\sum_{k=0}^{\infty} (2^{1-q})^k.$$

As mentioned earlier, the geometric series $\sum_{k=0}^{\infty} x^k$ converges if and only if $|x| < 1$. Thus the series $\sum_{n=1}^{\infty} 1/n^q$ will converge if and only if $|2^{1-q}| < 1$, which happens if and only if $q > 1$ (why? Try proving it just using Lemma 5.6.9 and Lemma 6.7.3, and without using logarithms). □

In particular, the series $\sum_{n=1}^{\infty} 1/n$ (also known as the *harmonic series*) is divergent, as claimed earlier. However, the series $\sum_{n=1}^{\infty} 1/n^2$ is convergent.

Remark 7.3.8 The quantity $\sum_{n=1}^{\infty} 1/n^q$, when it converges, is called $\zeta(q)$, the *Riemann-zeta function of q*. This function is very important in number theory, and in particular in the distribution of the primes; there is a very famous unsolved problem regarding this function, called the *Riemann hypothesis*, but to discuss it further is far beyond the scope of this text. I will mention however that there is a US$ 1 million prize—and instant fame among all mathematicians—attached to the solution to this problem.

— Exercises —

Exercise 7.3.1 Use Proposition 7.3.1 to prove Corollary 7.3.2.

Exercise 7.3.2 Prove Lemma 7.3.3. (*Hint:* for the first part, use the zero test. For the second part, first use induction to establish the *geometric series formula*

$$\sum_{n=0}^{N} x^n = (1 - x^{N+1})/(1 - x)$$

and then apply Lemma 6.5.2.)

Exercise 7.3.3 Let $\sum_{n=0}^{\infty} a_n$ be an absolutely convergent series of real numbers such that $\sum_{n=0}^{\infty} |a_n| = 0$. Show that $a_n = 0$ for every natural number n.

7.4 Rearrangement of Series

One feature of finite sums is that no matter how one rearranges the terms in a sequence, the total sum is the same. For instance,

$$a_1 + a_2 + a_3 + a_4 + a_5 = a_4 + a_3 + a_5 + a_1 + a_2.$$

A more rigorous statement of this, involving bijections, has already appeared earlier, see Remark 7.1.12.

One can ask whether the same thing is true for infinite series. If all the terms are non-negative, the answer is yes:

Proposition 7.4.1 *Let $\sum_{n=0}^{\infty} a_n$ be a convergent series of non-negative real numbers, and let $f : \mathbf{N} \to \mathbf{N}$ be a bijection. Then $\sum_{m=0}^{\infty} a_{f(m)}$ is also convergent, and has the same sum:*

$$\sum_{n=0}^{\infty} a_n = \sum_{m=0}^{\infty} a_{f(m)}.$$

Proof We introduce the partial sums $S_N := \sum_{n=0}^{N} a_n$ and $T_M := \sum_{m=0}^{M} a_{f(m)}$. We know that the sequences $(S_N)_{N=0}^{\infty}$ and $(T_M)_{M=0}^{\infty}$ are increasing. Write $L := \sup(S_N)_{n=0}^{\infty}$ and $L' := \sup(T_M)_{M=0}^{\infty}$. By Proposition 6.3.8 we know that L is finite, and in fact $L = \sum_{n=0}^{\infty} a_n$; by Proposition 6.3.8 again we see that we will thus be done as soon as we can show that $L' = L$.

Fix M, and let Y be the set $Y := \{m \in \mathbf{N} : m \leq M\}$. Note that f is a bijection between Y and $f(Y)$. By Proposition 7.1.11, we have

$$T_M = \sum_{m=0}^{M} a_{f(m)} = \sum_{m \in Y} a_{f(m)} = \sum_{n \in f(Y)} a_n.$$

The sequence $(f(m))_{m=0}^{M}$ is finite, hence bounded, i.e., there exists an N such that $f(m) \leq N$ for all $m \leq M$. In particular $f(Y)$ is a subset of $\{n \in \mathbf{N} : n \leq N\}$, and so by Proposition 7.1.11 again (and the assumption that all the a_n are non-negative)

$$T_M = \sum_{n \in f(Y)} a_n \leq \sum_{n \in \{n \in \mathbf{N}: n \leq N\}} a_n = \sum_{n=0}^{N} a_n = S_N.$$

But since $(S_N)_{N=0}^{\infty}$ has a supremum of L, we thus see that $S_N \leq L$, and hence that $T_M \leq L$ for all M. Since L' is the least upper bound of $(T_M)_{M=0}^{\infty}$, this implies that $L' \leq L$.

A very similar argument (using the inverse f^{-1} instead of f) shows that every S_N is bounded above by L', and hence $L \leq L'$. Combining these two inequalities we obtain $L = L'$, as desired. □

Example 7.4.2 From Corollary 7.3.7 we know that the series

$$\sum_{n=1}^{\infty} 1/n^2 = 1 + 1/4 + 1/9 + 1/16 + 1/25 + 1/36 + \cdots$$

is convergent. Thus, if we interchange every pair of terms, to obtain

$$1/4 + 1 + 1/16 + 1/9 + 1/36 + 1/25 + \cdots$$

we know that this series is also convergent, and has the same sum. (It turns out that the value of this sum is $\zeta(2) = \pi^2/6$, a fact which we shall prove in Exercise 5.5.2.)

Now we ask what happens when the series is not non-negative. Then as long as the series is *absolutely* convergent, we can still do rearrangements:

Proposition 7.4.3 (Rearrangement of series) *Let $\sum_{n=0}^{\infty} a_n$ be an absolutely convergent series of real numbers, and let $f : \mathbf{N} \to \mathbf{N}$ be a bijection. Then $\sum_{m=0}^{\infty} a_{f(m)}$ is also absolutely convergent, and has the same sum:*

$$\sum_{n=0}^{\infty} a_n = \sum_{m=0}^{\infty} a_{f(m)}.$$

Proof (Optional) We apply Proposition 7.4.1 to the infinite series $\sum_{n=0}^{\infty} |a_n|$, which by hypothesis is a convergent series of non-negative numbers. If we write $L := \sum_{n=0}^{\infty} |a_n|$, then by Proposition 7.4.1 we know that $\sum_{m=0}^{\infty} |a_{f(m)}|$ also converges to L.

Now write $L' := \sum_{n=0}^{\infty} a_n$. We have to show that $\sum_{m=0}^{\infty} a_{f(m)}$ also converges to L'. In other words, given any $\varepsilon > 0$, we have to find an M such that $\sum_{m=0}^{M'} a_{f(m)}$ is ε-close to L' for every $M' \geq M$.

Since $\sum_{n=0}^{\infty} |a_n|$ is convergent, we can use Proposition 7.2.5 and find an N_1 such that $\sum_{n=p}^{q} |a_n| \le \varepsilon/2$ for all $p, q \ge N_1$. Since $\sum_{n=0}^{\infty} a_n$ converges to L', the partial sums $\sum_{n=0}^{N} a_n$ also converge to L', and so there exists $N \ge N_1$ such that $\sum_{n=0}^{N} a_n$ is $\varepsilon/2$-close to L'.

Now the sequence $(f^{-1}(n))_{n=0}^{N}$ is finite, hence bounded, so there exists an M such that $f^{-1}(n) \le M$ for all $0 \le n \le N$. In particular, for any $M' \ge M$, the set $\{f(m) : m \in \mathbf{N}; m \le M'\}$ contains $\{n \in \mathbf{N} : n \le N\}$ (why?). So by Proposition 7.1.11, for any $M' \ge M$

$$\sum_{m=0}^{M'} a_{f(m)} = \sum_{n \in \{f(m): m \in \mathbf{N}; m \le M'\}} a_n = \sum_{n=0}^{N} a_n + \sum_{n \in X} a_n$$

where X is the set

$$X = \{f(m) : m \in \mathbf{N}; m \le M'\} \backslash \{n \in \mathbf{N} : n \le N\}.$$

The set X is finite, and is therefore bounded by some natural number q; we must therefore have

$$X \subseteq \{n \in \mathbf{N} : N + 1 \le n \le q\}$$

(why?). Thus

$$\left| \sum_{n \in X} a_n \right| \le \sum_{n \in X} |a_n| \le \sum_{n=N+1}^{q} |a_n| \le \varepsilon/2$$

by our choice of N. Thus $\sum_{m=0}^{M'} a_{f(m)}$ is $\varepsilon/2$-close to $\sum_{n=0}^{N} a_n$, which as mentioned before is $\varepsilon/2$-close to L'. Thus $\sum_{m=0}^{M'} a_{f(m)}$ is ε-close to L' for all $M' \ge M$, as desired. □

Surprisingly, when the series is not absolutely convergent, then the rearrangements are very badly behaved.

Example 7.4.4 Consider the series

$$1/3 - 1/4 + 1/5 - 1/6 + 1/7 - 1/8 + \cdots .$$

This series is not absolutely convergent (why?), but is convergent by the alternating series test, and in fact the sum can be seen to converge to a positive number (in fact, it converges to $\ln(2) - 1/2 = 0.193147\ldots$, see Example 4.5.7). Basically, the reason why the sum is positive is because the quantities $(1/3 - 1/4)$, $(1/5 - 1/6)$, $(1/7 - 1/8)$ are all positive, which can then be used to show that every partial sum is positive. (Why? you have to break into two cases, depending on whether there are an even or odd number of terms in the partial sum.)

If, however, we rearrange the series to have two negative terms to each positive term, thus

$$1/3 - 1/4 - 1/6 + 1/5 - 1/8 - 1/10 + 1/7 - 1/12 - 1/14 + \cdots$$

then the partial sums quickly become negative (this is because $(1/3 - 1/4 - 1/6)$, $(1/5 - 1/8 - 1/9)$, and more generally $(1/(2n + 1) - 1/4n - 1/(4n + 2))$ are all negative), and so this series converges to a negative quantity; in fact, it converges to

$$(\ln(2) - 1)/2 = -.153426 \ldots.$$

There is in fact a surprising result of Riemann, which shows that a series which is conditionally convergent (that is, convergent but not absolutely convergent) can in fact be rearranged to converge to *any* value (or rearranged to diverge, in fact—see Exercise 8.2.6); see Theorem 8.2.8.

To summarize, rearranging series is safe when the series is absolutely convergent, but is somewhat dangerous otherwise. (This is not to say that rearranging a series that is not absolutely convergent necessarily gives you the wrong answer—for instance, in theoretical physics one often performs similar maneuvres, and one still (usually) obtains a correct answer at the end—but doing so is risky, unless it is backed by a rigorous result such as Proposition 7.4.3.)

— Exercises —

Exercise 7.4.1 Let $\sum_{n=0}^{\infty} a_n$ be an absolutely convergent series of real numbers. Let $f : \mathbf{N} \to \mathbf{N}$ be an increasing function (i.e., $f(n + 1) > f(n)$ for all $n \in \mathbf{N}$). Show that $\sum_{n=0}^{\infty} a_{f(n)}$ is also an absolutely convergent series. (*Hint:* try to compare each partial sum of $\sum_{n=0}^{\infty} a_{f(n)}$ with a (slightly different) partial sum of $\sum_{n=0}^{\infty} a_n$.) What happens if we assume f is merely one-to-one, rather than increasing?

Exercise 7.4.2 Obtain an alternate proof of Proposition 7.4.3 using Proposition 7.4.1, Proposition 7.2.13, and expressing a_n as the difference of $a_n + |a_n|$ and $|a_n|$. (This argument is due to Will Ballard.)

7.5 The Root and Ratio Tests

Now we can state and prove the famous root and ratio tests for convergence.

Theorem 7.5.1 (Root test) *Let $\sum_{n=m}^{\infty} a_n$ be a series of real numbers, and let $\alpha :=$* $\limsup_{n\to\infty} |a_n|^{1/n}$.

(a) If $\alpha < 1$, then the series $\sum_{n=m}^{\infty} a_n$ is absolutely convergent (and hence convergent).
(b) If $\alpha > 1$, then the series $\sum_{n=m}^{\infty} a_n$ is not convergent (and hence cannot be absolutely convergent either).
(c) If $\alpha = 1$, we cannot assert any conclusion.

Proof By Proposition 7.2.13(c), we may assume without loss of generality that $m \geq 1$; in particular $|a_n|^{1/n}$ is well-defined for any $n \geq m$.

First suppose that $\alpha < 1$. Note that we must have $\alpha \geq 0$, since $|a_n|^{1/n} \geq 0$ for every n. Then we can find an $\varepsilon > 0$ such that $0 < \alpha + \varepsilon < 1$ (for instance, we can set $\varepsilon := (1-\alpha)/2$). By Proposition 6.4.12(a), there exists an $N \geq m$ such that $|a_n|^{1/n} \leq \alpha + \varepsilon$ for all $n \geq N$. In other words, we have $|a_n| \leq (\alpha+\varepsilon)^n$ for all $n \geq N$. But from the geometric series we have that $\sum_{n=N}^{\infty} (\alpha+\varepsilon)^n$ is absolutely convergent, since $0 < \alpha + \varepsilon < 1$ (note that the fact that we start from N is irrelevant by Proposition 7.2.13(c)). Thus by the comparison test, we see that $\sum_{n=N}^{\infty} a_n$ is absolutely convergent, and thus $\sum_{n=m}^{\infty} a_n$ is absolutely convergent, by Proposition 7.2.13(c) again.

Now suppose that $\alpha > 1$. Then by Proposition 6.4.12(b), we see that for every $N \geq m$ there exists an $n \geq N$ such that $|a_n|^{1/n} > 1$, and hence that $|a_n| > 1$. In particular, $(a_n)_{n=N}^{\infty}$ is not 1-close to 0 for any N, and hence $(a_n)_{n=m}^{\infty}$ is not eventually 1-close to 0. In particular, $(a_n)_{n=m}^{\infty}$ does not converge to zero. Thus by the zero test, $\sum_{n=m}^{\infty} a_n$ is not convergent.

For $\alpha = 1$, see Exercise 7.5.3. □

The root test is phrased using the limit superior, but of course if $\lim_{n\to\infty} |a_n|^{1/n}$ converges then the limit is the same as the limit superior. Thus one can phrase the root test using the limit instead of the limit superior, but *only when the limit exists*.

The root test is sometimes difficult to use; however we can replace roots by ratios using the following lemma.

Lemma 7.5.2 *Let $(c_n)_{n=m}^{\infty}$ be a sequence of positive numbers. Then we have*

$$\liminf_{n\to\infty} \frac{c_{n+1}}{c_n} \leq \liminf_{n\to\infty} c_n^{1/n} \leq \limsup_{n\to\infty} c_n^{1/n} \leq \limsup_{n\to\infty} \frac{c_{n+1}}{c_n}.$$

Proof There are three inequalities to prove here. The middle inequality follows from Proposition 6.4.12(c). We shall prove the last inequality, and leave the first one to Exercise 7.5.1.

Write $L := \limsup_{n\to\infty} \frac{c_{n+1}}{c_n}$. If $L = +\infty$ then there is nothing to prove (since $x \leq +\infty$ for every extended real number x), so we may assume that L is a finite real number. (Note that L cannot equal $-\infty$; why?). Since $\frac{c_{n+1}}{c_n}$ is always positive, we know that $L \geq 0$.

Let $\varepsilon > 0$. By Proposition 6.4.12(a), we know that there exists an $N \geq m$ such that $\frac{c_{n+1}}{c_n} \leq L + \varepsilon$ for all $n \geq N$. without loss of generality we may assume that $N \geq 1$. This implies that $c_{n+1} \leq c_n(L+\varepsilon)$ for all $n \geq N$. By induction this implies that

$$c_n \leq c_N (L+\varepsilon)^{n-N} \text{ for all } n \geq N$$

(why?). If we write $A := c_N(L+\varepsilon)^{-N}$, then we have

$$c_n \leq A(L+\varepsilon)^n$$

and thus

$$c_n^{1/n} \leq A^{1/n}(L + \varepsilon)$$

for all $n \geq N$. But we have

$$\lim_{n\to\infty} A^{1/n}(L + \varepsilon) = L + \varepsilon$$

by the limit laws (Theorem 6.1.19) and Lemma 6.5.3. Thus by the comparison principle (Lemma 6.4.13) we have

$$\limsup_{n\to\infty} c_n^{1/n} \leq L + \varepsilon.$$

But this is true for all $\varepsilon > 0$, so this must imply that

$$\limsup_{n\to\infty} c_n^{1/n} \leq L$$

(why? prove by contradiction), as desired. □

From Theorem 7.5.1 and Lemma 7.5.2 (and Exercise 7.5.3) we have

Corollary 7.5.3 (Ratio test) *Let* $\sum_{n=m}^{\infty} a_n$ *be a series of* non-zero *numbers.* (*The non-zero hypothesis is required so that the ratios* $|a_{n+1}|/|a_n|$ *appearing below are well-defined.*)

- *If* $\limsup_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} < 1$, *then the series* $\sum_{n=m}^{\infty} a_n$ *is absolutely convergent* (*hence convergent*).
- *If* $\liminf_{n\to\infty} \frac{|a_{n+1}|}{|a_n|} > 1$, *then the series* $\sum_{n=m}^{\infty} a_n$ *is not convergent* (*and thus cannot be absolutely convergent*).
- *In the remaining cases, we cannot assert any conclusion.*

Another consequence of Lemma 7.5.2 is the following limit:

Proposition 7.5.4 *We have* $\lim_{n\to\infty} n^{1/n} = 1$.

Proof By Lemma 7.5.2 we have

$$\limsup_{n\to\infty} n^{1/n} \leq \limsup_{n\to\infty} (n+1)/n = \limsup_{n\to\infty} 1 + 1/n = 1$$

by Proposition 6.1.11 and limit laws (Theorem 6.1.19). Similarly we have

$$\liminf_{n\to\infty} n^{1/n} \geq \liminf_{n\to\infty} (n+1)/n = \liminf_{n\to\infty} 1 + 1/n = 1.$$

The claim then follows from Proposition 6.4.12(c) and (f). □

Remark 7.5.5 In addition to the ratio and root tests, another very useful convergence test is the *integral test*, which we will cover in Proposition 11.6.4.

— Exercises —

Exercise 7.5.1 Prove the first inequality in Lemma 7.5.2.

Exercise 7.5.2 Let x be a real number with $|x| < 1$, and q be a real number. Show that the series $\sum_{n=1}^{\infty} n^q x^n$ is absolutely convergent, and that $\lim_{n\to\infty} n^q x^n = 0$.

Exercise 7.5.3 Give an example of a divergent series $\sum_{n=1}^{\infty} a_n$ of positive numbers a_n such that $\lim_{n\to\infty} a_{n+1}/a_n = \lim_{n\to\infty} a_n^{1/n} = 1$, and give an example of a convergent series $\sum_{n=1}^{\infty} b_n$ of positive numbers b_n such that $\lim_{n\to\infty} b_{n+1}/b_n = \lim_{n\to\infty} b_n^{1/n} = 1$. (*Hint:* use Corollary 7.3.7.) This shows that the ratio and root tests can be inconclusive even when the summands are positive and all the limits converge.

Chapter 8
Infinite Sets

We now return to the study of set theory, and specifically to the study of cardinality of sets which are infinite (i.e., sets which do not have cardinality n for any natural number n), a topic which was initiated in Sect. 3.6.

8.1 Countability

From Proposition 3.6.14c we know that if X is a finite set, and Y is a proper subset of X, then Y does not have equal cardinality with X. However, this is not the case for infinite sets. For instance, from Theorem 3.6.12 we know that the set $\mathbf{N}$ of natural numbers is infinite. The set $\mathbf{N} - \{0\}$ is also infinite, thanks to Proposition 3.6.14a (why?), and is a proper subset of $\mathbf{N}$. However, the set $\mathbf{N} - \{0\}$, despite being "smaller" than $\mathbf{N}$, still has the same cardinality as $\mathbf{N}$, because the function $f : \mathbf{N} \to \mathbf{N} - \{0\}$ defined by $f(n) := n + 1$, is a bijection from $\mathbf{N}$ to $\mathbf{N} - \{0\}$. (Why?) This is one characteristic of infinite sets; see Exercise 8.1.1.

We now distinguish two types of infinite sets: the countable sets and the uncountable sets.

Definition 8.1.1 *(Countable sets)* A set X is said to be *countably infinite* (or just *countable*) iff it has equal cardinality with the natural numbers $\mathbf{N}$. A set X is said to be *at most countable* iff it is either countable or finite. We say that a set is *uncountable* if it is infinite but not countable.

Remark 8.1.2 Countably infinite sets are also called *denumerable* sets.

Examples 8.1.3 From the preceding discussion we see that $\mathbf{N}$ is countable, and so is $\mathbf{N} - \{0\}$. Another example of a countable set is the even natural numbers $\{2n : n \in \mathbf{N}\}$, since the function $f(n) := 2n$ provides a bijection between $\mathbf{N}$ and the even natural numbers (why?).

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_8

Let X be a countable set. Then, by definition, we know that there exists a bijection $f : \mathbf{N} \to X$. Thus, every element of X can be written in the form $f(n)$ for exactly one natural number n. Informally, we thus have

$$X = \{f(0), f(1), f(2), f(3), \ldots\}.$$

Thus, a countable set can be arranged in a sequence, so that we have a zeroth element $f(0)$, followed by a first element $f(1)$, then a second element $f(2)$, and so forth, in such a way that all these elements $f(0), f(1), f(2), \ldots$ are all distinct, and together they fill out all of X. (This is why these sets are called *countable*; because we can literally count them one by one, starting from $f(0)$, then $f(1)$, and so forth.)

Viewed in this way, it is clear why the natural numbers

$$\mathbf{N} = \{0, 1, 2, 3, \ldots\},$$

the positive integers

$$\mathbf{N} - \{0\} = \{1, 2, 3, \ldots\},$$

and the even natural numbers

$$\{0, 2, 4, 6, 8, \ldots\}$$

are countable. However, it is not as obvious whether the integers

$$\mathbf{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$$

or the rationals

$$\mathbf{Q} = \{0, 1/4, -2/3, \ldots\}$$

or the reals

$$\mathbf{R} = \{0, \sqrt{2}, -\pi, 2.5, \ldots\}$$

are countable or not; for instance, it is not yet clear whether we can arrange the real numbers in a sequence $f(0), f(1), f(2), \ldots$. We will answer these questions shortly.

From Proposition 3.6.4 and Theorem 3.6.12, we know that countable sets are infinite; however it is not so clear whether all infinite sets are countable. Again, we will answer those questions shortly. We first need the following important principle.

Proposition 8.1.4 (Well-ordering principle) *Let X be a non-empty subset of the natural numbers* $\mathbf{N}$. *Then there exists exactly one element $n \in X$ such that $n \leq m$ for all $m \in X$. In other words, every non-empty set of natural numbers has a minimum element.*

Proof See Exercise 8.1.2. □

We will refer to the element n given by the well-ordering principle as the *minimum* of X, and write it as $\min(X)$. Thus for instance the minimum of the set $\{2, 4, 6, 8, \ldots\}$ is 2. This minimum is clearly the same as the infimum of X, as defined in Definition 5.5.10 (why?).

Proposition 8.1.5 *Let X be an infinite subset of the natural numbers* $\mathbf{N}$. *Then there exists a unique bijection $f : \mathbf{N} \to X$ which is increasing, in the sense that $f(n+1) > f(n)$ for all $n \in \mathbf{N}$. In particular, X has equal cardinality with* $\mathbf{N}$ *and is hence countable.*

Proof We will give an incomplete sketch of the proof, with some gaps marked by a question mark (?); these gaps will be filled in Exercise 8.1.3.

We now define a sequence $a_0, a_1, a_2, \ldots$ of natural numbers recursively by the formula

$$a_n := \min\{x \in X : x \neq a_m \text{ for all } m < n\}.$$

Intuitively speaking, a_0 is the smallest element of X; a_1 is the second smallest element of X, i.e., the smallest element of X once a_0 is removed; a_2 is the third smallest element of X; and so forth. Observe that in order to define a_n, one only needs to know the values of a_m for all $m < n$, so this definition is recursive. Also, since X is infinite, the set $\{x \in X : x \neq a_m \text{ for all } m < n\}$ is infinite(?), hence non-empty. Thus by the well-ordering principle, the minimum, $\min\{x \in X : x \neq a_m \text{ for all } m < n\}$ is always well-defined.

One can show(?) that a_n is an increasing sequence, i.e.,

$$a_0 < a_1 < a_2 < \ldots$$

and in particular that(?) $a_n \neq a_m$ for all $n \neq m$. Also, we have(?) $a_n \in X$ for each natural number n.

Now define the function $f : \mathbf{N} \to X$ by $f(n) := a_n$. From the previous paragraph we know that f is one-to-one. Now we show that f is onto. In other words, we claim that for every $x \in X$, there exists an n such that $a_n = x$.

Let $x \in X$. Suppose for sake of contradiction that $a_n \neq x$ for every natural number n. Then this implies(?) that x is an element of the set $\{x \in X : x \neq a_m \text{ for all } m < n\}$ for all n. By definition of a_n, this implies that $x \geq a_n$ for every natural number n. However, since a_n is an increasing sequence, we have $a_n \geq n$ (?), and hence $x \geq n$ for every natural number n. In particular we have $x \geq x + 1$, which is a contradiction. Thus we must have $a_n = x$ for some natural number n, and hence f is onto.

Since $f : \mathbf{N} \to X$ is both one-to-one and onto, it is a bijection. We have thus found at least one increasing bijection f from $\mathbf{N}$ to X. Now suppose for sake of contradiction that there was at least one other increasing bijection g from $\mathbf{N}$ to X which was not equal to f. Then the set $\{n \in \mathbf{N} : g(n) \neq f(n)\}$ is non-empty, and define $m := \min\{n \in \mathbf{N} : g(n) \neq f(n)\}$, thus in particular $g(m) \neq f(m) = a_m$, and $g(n) = f(n) = a_n$ for all $n < m$. But we then must have(?)

$$g(m) = \min\{x \in X : x \neq a_t \text{ for all } t < m\} = a_m,$$

a contradiction. Thus there is no other increasing bijection from $\mathbf{N}$ to X other than f. □

Since finite sets are at most countable by definition, we thus have

Corollary 8.1.6 *All subsets of the natural numbers are at most countable.*

Corollary 8.1.7 *If X is an at most countable set, and Y is a subset of X, then Y is at most countable.*

Proof If X is finite then this follows from Proposition 3.6.14c, so assume X is countable. Then there is a bijection $f: X \to \mathbf{N}$ between X and $\mathbf{N}$. Since Y is a subset of X, and f is a bijection from X and $\mathbf{N}$, then when we restrict f to Y, we obtain a bijection between Y and $f(Y)$. (Why is this a bijection?) Thus $f(Y)$ has equal cardinality with Y. But $f(Y)$ is a subset of $\mathbf{N}$, and hence at most countable by Corollary 8.1.6. Hence Y is also at most countable. □

Proposition 8.1.8 *Let Y be a set, and let $f: \mathbf{N} \to Y$ be a function. Then $f(\mathbf{N})$ is at most countable.*

Proof See Exercise 8.1.4. □

Corollary 8.1.9 *Let X be a countable set, and let $f: X \to Y$ be a function. Then $f(X)$ is at most countable.*

Proof See Exercise 8.1.5. □

Proposition 8.1.10 *Let X be a countable set, and let Y be a countable set. Then $X \cup Y$ is a countable set.*

Proof See Exercise 8.1.7. □

To summarize, any subset or image of a countable set is at most countable, and any finite union of countable sets is still countable. We can now establish countability of the integers.

Corollary 8.1.11 *The integers $\mathbf{Z}$ are countable.*

Proof We already know that the set $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$ of natural numbers are countable. The set $-\mathbf{N}$ defined by

$$-\mathbf{N} := \{-n : n \in \mathbf{N}\} = \{0, -1, -2, -3, \ldots\}$$

is also countable, since the map $f(n) := -n$ is a bijection between $\mathbf{N}$ and this set. Since the integers are the union of $\mathbf{N}$ and $-\mathbf{N}$, the claim follows from Proposition 8.1.10 □

To establish countability of the rationals, we need to relate countability with Cartesian products. In particular, we need to show that the set $\mathbf{N} \times \mathbf{N}$ is countable. We first need a preliminary lemma:

Lemma 8.1.12 *The set*

$$A := \{(n, m) \in \mathbf{N} \times \mathbf{N} : 0 \leq m \leq n\}$$

is countable.

Proof Define the sequence $a_0, a_1, a_2, \ldots$ recursively by setting $a_0 := 0$, and $a_{n+1} := a_n + n + 1$ for all natural numbers n. Thus

$$a_0 = 0; a_1 = 0 + 1; a_2 = 0 + 1 + 2; a_3 = 0 + 1 + 2 + 3; \ldots.$$

By induction one can show that a_n is increasing, i.e., that $a_n > a_m$ whenever $n > m$ (why?).

Now define the function $f : A \to \mathbf{N}$ by

$$f(n, m) := a_n + m.$$

We claim that f is one-to-one. In other words, if (n, m) and (n', m') are any two distinct elements of A, then we claim that $f(n, m) \neq f(n', m')$.

To prove this claim, let (n, m) and (n', m') be two distinct elements of A. There are three cases: $n' = n$, $n' > n$, and $n' < n$. First suppose that $n' = n$. Then we must have $m \neq m'$, otherwise (n, m) and (n', m') would not be distinct. Thus $a_n + m \neq a_n + m'$, and hence $f(n, m) \neq f(n', m')$, as desired.

Now suppose that $n' > n$. Then $n' \geq n + 1$, and hence

$$f(n', m') = a_{n'} + m' \geq a_{n'} \geq a_{n+1} = a_n + n + 1.$$

But since $(n, m) \in A$, we have $m \leq n < n + 1$, and hence

$$f(n', m') \geq a_n + n + 1 > a_n + m = f(n, m),$$

and thus $f(n', m') \neq f(n, m)$.

The case $n' < n$ is proven similarly, by switching the rôles of n and n' in the previous argument. Thus we have shown that f is one-to-one. Thus f is a bijection from A to $f(A)$, and so A has equal cardinality with $f(A)$. But $f(A)$ is a subset of $\mathbf{N}$, and hence by Corollary 8.1.6 $f(A)$ is at most countable. Therefore A is at most countable. But, A is clearly not finite. (Why? Hint: if A was finite, then every subset of A would be finite, and in particular $\{(n, 0) : n \in \mathbf{N}\}$ would be finite, but this is clearly countably infinite, a contradiction.) Thus, A must be countable. □

Corollary 8.1.13 *The set* $\mathbf{N} \times \mathbf{N}$ *is countable.*

Proof We already know that the set

$$A := \{(n, m) \in \mathbf{N} \times \mathbf{N} : 0 \leq m \leq n\}$$

is countable. This implies that the set

$$B := \{(n, m) \in \mathbf{N} \times \mathbf{N} : 0 \leq n \leq m\}$$

is also countable, since the map $f\colon A \to B$ given by $f(n, m) := (m, n)$ is a bijection from A to B (why?). But since $\mathbf{N} \times \mathbf{N}$ is the union of A and B (why?), the claim then follows from Proposition 8.1.10. □

Corollary 8.1.14 *If X and Y are countable, then $X \times Y$ is countable.*

Proof See Exercise 8.1.8. □

Corollary 8.1.15 *The rationals $\mathbf{Q}$ are countable.*

Proof We already know that the integers $\mathbf{Z}$ are countable, which implies that the non-zero integers $\mathbf{Z} - \{0\}$ are countable (why?). By Corollary 8.1.14, the set

$$\mathbf{Z} \times (\mathbf{Z} - \{0\}) = \{(a, b) : a, b \in \mathbf{Z}, b \neq 0\}$$

is thus countable. If one lets $f\colon \mathbf{Z} \times (\mathbf{Z} - \{0\}) \to \mathbf{Q}$ be the function $f(a, b) := a/b$ (note that f is well-defined since we prohibit b from being equal to 0), we see from Corollary 8.1.9 that $f(\mathbf{Z} \times (\mathbf{Z} - \{0\}))$ is at most countable. But we have $f(\mathbf{Z} \times (\mathbf{Z} - \{0\})) = \mathbf{Q}$ (why? This is basically the definition of the rationals $\mathbf{Q}$). Thus $\mathbf{Q}$ is at most countable. However, $\mathbf{Q}$ cannot be finite, since it contains the infinite set $\mathbf{N}$. Thus $\mathbf{Q}$ is countable. □

Remark 8.1.16 Because the rationals are countable, we know *in principle* that it is possible to arrange the rational numbers as a sequence:

$$\mathbf{Q} = \{a_0, a_1, a_2, a_3, \ldots\}$$

such that every element of the sequence is different from every other element, and that the elements of the sequence exhaust $\mathbf{Q}$ (i.e., every rational number turns up as one of the elements a_n of the sequence). However, it is quite difficult (though not impossible) to actually try and come up with an explicit sequence $a_0, a_1, \ldots$ which does this; see Exercise 8.1.10.

— Exercises —

Exercise 8.1.1 Let X be a set. Show that X is infinite if and only if there exists a proper subset $Y \subsetneq X$ of X which has the same cardinality as X. (This exercise requires the axiom of choice, Axiom 8.1)

Exercise 8.1.2 Prove Proposition 8.1.4. (*Hint:* you can either use induction, or use the principle of infinite descent, Exercise 4.4.2, or use the least upper bound (or greatest lower bound) principle, Theorem 5.5.9.) Does the well-ordering principle work if we replace the natural numbers by the integers? What if we replace the natural numbers by the positive rationals? Explain.

Exercise 8.1.3 Fill in the gaps marked (?) in Proposition 8.1.5.

Exercise 8.1.4 Prove Proposition 8.1.8. (*Hint:* the basic problem here is that f is not assumed to be one-to-one. Define A to be the set

$$A := \{n \in \mathbf{N} : f(m) \neq f(n) \text{ for all } 0 \leq m < n\};$$

informally speaking, A is the set of natural numbers n for which $f(n)$ does not appear in the sequence $f(0), f(1), \ldots f(n-1)$. Prove that when f is restricted to A, it becomes a bijection from A to $f(\mathbf{N})$. Then use Corollary 8.1.6.)

Exercise 8.1.5 Use Proposition 8.1.8 to prove Corollary 8.1.9.

Exercise 8.1.6 Let A be a set. Show that A is at most countable if and only if there exists an injective map $f : A \to \mathbf{N}$ from A to $\mathbf{N}$.

Exercise 8.1.7 Prove Proposition 8.1.10. (*Hint:* by hypothesis, we have a bijection $f : \mathbf{N} \to X$, and a bijection $g : \mathbf{N} \to Y$. Now define $h : \mathbf{N} \to X \cup Y$ by setting $h(2n) := f(n)$ and $h(2n+1) := g(n)$ for every natural number n, and show that $h(\mathbf{N}) = X \cup Y$. Then use Corollary 8.1.9, and show that $X \cup Y$ cannot possibly be finite.)

Exercise 8.1.8 Use Corollary 8.1.13 to prove Corollary 8.1.14.

Exercise 8.1.9 Suppose that I is an at most countable set, and for each $\alpha \in I$, let A_α be an at most countable set. Show that the set $\bigcup_{\alpha \in I} A_\alpha$ is also at most countable. In particular, countable unions of countable sets are countable. (This exercise requires the axiom of choice, see Sect. 8.4.)

Exercise 8.1.10 Find a bijection $f : \mathbf{N} \to \mathbf{Q}$ from the natural numbers to the rationals. (Warning: this is actually rather tricky to do explicitly; it is difficult to get f to be simultaneously injective and surjective.)

8.2 Summation on Infinite Sets

We now introduce the concept of summation on *countable sets*, which will be well-defined provided that the sum is absolutely convergent.

Definition 8.2.1 *(Series on countable sets)* Let X be a countable set, and let $f : X \to \mathbf{R}$ be a function. We say that the series $\sum_{x \in X} f(x)$ is *absolutely convergent* iff for some bijection $g : \mathbf{N} \to X$, the sum $\sum_{n=0}^{\infty} f(g(n))$ is absolutely convergent. We then define the sum of $\sum_{x \in X} f(x)$ by the formula

$$\sum_{x \in X} f(x) = \sum_{n=0}^{\infty} f(g(n)).$$

From Proposition 7.4.3, one can show that these definitions do not depend on the choice of g, and so are well-defined.

We can now give an important theorem about double summations.

Theorem 8.2.2 (Fubini's theorem for infinite sums) *Let $f : \mathbf{N} \times \mathbf{N} \to \mathbf{R}$ be a function such that $\sum_{(n,m) \in \mathbf{N} \times \mathbf{N}} f(n, m)$ is absolutely convergent. Then we have*

$$\sum_{n=0}^{\infty}\left(\sum_{m=0}^{\infty} f(n,m)\right) = \sum_{(n,m)\in\mathbf{N}\times\mathbf{N}} f(n,m)$$
$$= \sum_{(m,n)\in\mathbf{N}\times\mathbf{N}} f(n,m)$$
$$= \sum_{m=0}^{\infty}\left(\sum_{n=0}^{\infty} f(n,m)\right).$$

In other words, we can switch the order of infinite sums *provided that the entire sum is absolutely convergent*. You should go back and compare this with Example 1.2.5.

Proof (A sketch only; this proof is considerably more complex than the other proofs, and is optional reading.) The second equality follows easily from Proposition 7.4.3 (and Proposition 3.6.4). We shall just prove the first equality, as the third is very similar (basically one switches the rôle of n and m).

Let us first consider the case when $f(n,m)$ is always non-negative (we will deal with the general case later). Write

$$L := \sum_{(n,m)\in\mathbf{N}\times\mathbf{N}} f(n,m);$$

our task is to show that the series $\sum_{n=0}^{\infty}(\sum_{m=0}^{\infty} f(n,m))$ converges to L.

One can easily show that $\sum_{(n,m)\in X} f(n,m) \le L$ for all finite sets $X \subseteq \mathbf{N}\times\mathbf{N}$. (Why? Use a bijection g between $\mathbf{N}\times\mathbf{N}$ and $\mathbf{N}$, and then use the fact that $g(X)$ is finite, hence bounded.) In particular, for every $n \in \mathbf{N}$ and $M \in \mathbf{N}$ we have $\sum_{m=0}^{M} f(n,m) \le L$, which implies by Proposition 6.3.8 that $\sum_{m=0}^{\infty} f(n,m)$ is convergent for each m. Similarly, for any $N \in \mathbf{N}$ and $M \in \mathbf{N}$ we have (by Corollary 7.1.14)

$$\sum_{n=0}^{N}\sum_{m=0}^{M} f(n,m) \le \sum_{(n,m)\in X} f(n,m) \le L$$

where X is the set $\{(n,m)\in\mathbf{N}\times\mathbf{N} : n \le N, m \le M\}$ which is finite by Proposition 3.6.14. Taking suprema of this as $M \to \infty$ we have (by limit laws, and an induction on N)

$$\sum_{n=0}^{N}\sum_{m=0}^{\infty} f(n,m) \le L.$$

By Proposition 6.3.8, this implies that $\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(n,m)$ converges, and

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(n,m) \le L.$$

To finish the proof, it will suffice to show that

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(n,m) \geq L - \varepsilon$$

for every $\varepsilon > 0$. (Why will this be enough? Prove by contradiction.) So, let $\varepsilon > 0$. By definition of L, we can then find a finite set $X \subseteq \mathbf{N} \times \mathbf{N}$ such that $\sum_{(n,m)\in X} f(n,m) \geq L - \varepsilon$. (Why?) This set, being finite, must be contained in some set of the form $Y := \{(n,m) \in \mathbf{N} \times \mathbf{N} : n \leq N; m \leq M\}$. (Why? Use induction.) Thus by Corollary 7.1.14

$$\sum_{n=0}^{N}\sum_{m=0}^{M} f(n,m) = \sum_{(n,m)\in Y} f(n,m) \geq \sum_{(n,m)\in X} f(n,m) \geq L - \varepsilon$$

and hence

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f(n,m) \geq \sum_{n=0}^{N}\sum_{m=0}^{\infty} f(n,m) \geq \sum_{n=0}^{N}\sum_{m=0}^{M} f(n,m) \geq L - \varepsilon$$

as desired.

This proves the claim when the $f(n,m)$ are all non-negative. A similar argument works when the $f(n,m)$ are all non-positive (in fact, one can simply apply the result just obtained to the function $-f(n,m)$, and then use limit laws to remove the $-$. For the general case, note that any function $f(n,m)$ can be written (why?) as $f_+(n,m) + f_-(n,m)$, where $f_+(n,m)$ is the positive part of $f(n,m)$ (i.e., it equals $f(n,m)$ when $f(n,m)$ is positive, and 0 otherwise), and f_- is the negative part of $f(n,m)$ (it equals $f(n,m)$ when $f(n,m)$ is negative, and 0 otherwise). It is easy to show that if $\sum_{(n,m)\in\mathbf{N}\times\mathbf{N}} f(n,m)$ is absolutely convergent, then so are $\sum_{(n,m)\in\mathbf{N}\times\mathbf{N}} f_+(n,m)$ and $\sum_{(n,m)\in\mathbf{N}\times\mathbf{N}} f_-(n,m)$. So now one applies the results just obtained to f_+ and to f_- and adds them together using limit laws to obtain the result for a general f. □

There is another characterization of absolutely convergent series.

Lemma 8.2.3 *Let X be a countable set, and let $f : X \to \mathbf{R}$ be a function. Then the series $\sum_{x\in X} f(x)$ is absolutely convergent if and only if*

$$\sup\left\{\sum_{x\in A} |f(x)| : A \subseteq X, A \textit{ finite}\right\} < \infty.$$

Proof See Exercise 8.2.1. □

Inspired by this lemma, we may now define the concept of an absolutely convergent series even when the set X could be uncountable. (We give some examples of uncountable sets in the next section.)

Definition 8.2.4 Let X be a set (which could be uncountable), and let $f\colon X \to \mathbf{R}$ be a function. We say that the series $\sum_{x\in X} f(x)$ is absolutely convergent iff

$$\sup\left\{\sum_{x\in A} |f(x)| : A \subseteq X, A \text{ finite}\right\} < \infty.$$

Note that we have not yet said what the series $\sum_{x\in X} f(x)$ is equal to. This shall be accomplished by the following lemma.

Lemma 8.2.5 *Let X be a set (which could be uncountable), and let $f\colon X \to \mathbf{R}$ be a function such that the series $\sum_{x\in X} f(x)$ is absolutely convergent. Then the set $\{x \in X : f(x) \neq 0\}$ is at most countable. (This result requires the axiom of choice, see Sect. 8.4.)*

Proof See Exercise 8.2.2. □

Because of this, we can define the value of $\sum_{x\in X} f(x)$ for any absolutely convergent series on an uncountable set X by the formula

$$\sum_{x\in X} f(x) := \sum_{x\in X: f(x)\neq 0} f(x),$$

since we have replaced a sum on an uncountable set X by a sum on the at most countable set $\{x \in X : f(x) \neq 0\}$. (Note that if the former sum is absolutely convergent, then the latter one is also.) Note also that this definition is consistent with the definitions we already have for series on countable sets.

We give some laws for absolutely convergent series on arbitrary sets.

Proposition 8.2.6 (Absolutely convergent series laws) *Let X be an arbitrary set (possibly uncountable), and let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be functions such that the series $\sum_{x\in X} f(x)$ and $\sum_{x\in X} g(x)$ are both absolutely convergent.*

(a) The series $\sum_{x\in X}(f(x) + g(x))$ is absolutely convergent, and

$$\sum_{x\in X}(f(x) + g(x)) = \sum_{x\in X} f(x) + \sum_{x\in X} g(x).$$

(b) If c is a real number, then $\sum_{x\in X} cf(x)$ is absolutely convergent, and

$$\sum_{x\in X} cf(x) = c\sum_{x\in X} f(x).$$

(c) If $X = X_1 \cup X_2$ for some disjoint sets X_1 and X_2, then $\sum_{x\in X_1} f(x)$ and $\sum_{x\in X_2} f(x)$ are absolutely convergent, and

$$\sum_{x\in X_1\cup X_2} f(x) = \sum_{x\in X_1} f(x) + \sum_{x\in X_2} f(x).$$

Conversely, if $h\colon X \to \mathbf{R}$ is such that $\sum_{x\in X_1} h(x)$ and $\sum_{x\in X_2} h(x)$ are absolutely convergent, then $\sum_{x\in X_1\cup X_2} h(x)$ is also absolutely convergent, and

$$\sum_{x\in X_1\cup X_2} h(x) = \sum_{x\in X_1} h(x) + \sum_{x\in X_2} h(x).$$

(d) *If Y is another set, and $\phi : Y \to X$ is a bijection, then $\sum_{y\in Y} f(\phi(y))$ is absolutely convergent, and*

$$\sum_{y\in Y} f(\phi(y)) = \sum_{x\in X} f(x).$$

(This result requires the axiom of choice when X is uncountable, see Sect. 8.4.)

Proof See Exercise 8.2.3. □

Recall in Example 7.4.4 that if a series was conditionally convergent, then its behavior with respect to rearrangements was bad. We now analyze this phenomenon further.

Lemma 8.2.7 *Let $\sum_{n=0}^{\infty} a_n$ be a series of real numbers which is conditionally convergent (convergent but not absolutely convergent). Define the sets $A_+ := \{n \in \mathbf{N} : a_n \geq 0\}$ and $A_- := \{n \in \mathbf{N} : a_n < 0\}$, thus $A_+ \cup A_- = \mathbf{N}$ and $A_+ \cap A_- = \emptyset$. Then both of the series $\sum_{n\in A_+} a_n$ and $\sum_{n\in A_-} a_n$ are not absolutely convergent.*

Proof See Exercise 8.2.4. □

We are now ready to present a remarkable theorem of Georg Riemann (1826–1866), which asserts that a series which converges conditionally but not absolutely can be rearranged to converge to any value one pleases!

Theorem 8.2.8 *Let $\sum_{n=0}^{\infty} a_n$ be a series which is conditionally convergent (i.e., convergent, but not absolutely convergent), and let L be any real number. Then there exists a bijection $f : \mathbf{N} \to \mathbf{N}$ such that $\sum_{m=0}^{\infty} a_{f(m)}$ converges conditionally to L.*

Proof (Optional) We give a sketch of the proof, leaving the details to be filled in in Exercise 8.2.5. Let A_+ and A_- be the sets in Lemma 8.2.7; from that lemma we know that $\sum_{n\in A_+} a_n$ and $\sum_{n\in A_-} a_n$ both fail to be absolutely convergent. In particular A_+ and A_- are infinite (why?). By Proposition 8.1.5 we can then find increasing bijections $f_+ : \mathbf{N} \to A_+$ and $f_- : \mathbf{N} \to A_-$. Thus the sums $\sum_{m=0}^{\infty} a_{f_+(m)}$ and $\sum_{m=0}^{\infty} a_{f_-(m)}$ both fail to be absolutely convergent (why?). The plan shall be to select terms from the divergent series $\sum_{m=0}^{\infty} a_{f_+(m)}$ and $\sum_{m=0}^{\infty} a_{f_-(m)}$ in a well-chosen order in order to keep their difference converging toward L.

We define the sequence $n_0, n_1, n_2, \ldots$ of natural numbers recursively as follows. Suppose that j is a natural number, and that n_i has already been defined for all $i < j$ (this is vacuously true if $j = 0$). We then define n_j by the following rule:

(I) If $\sum_{0 \leq i < j} a_{n_i} < L$, then we set

$$n_j := \min\{n \in A_+ : n \neq n_i \text{ for all } i < j\}.$$

(II) If instead $\sum_{0 \leq i < j} a_{n_i} \geq L$, then we set

$$n_j := \min\{n \in A_- : n \neq n_i \text{ for all } i < j\}.$$

Note that this recursive definition is well-defined because A_+ and A_- are infinite, and so the sets $\{n \in A_+ : n \neq n_i$ for all $i < j\}$ and $n_j := \min\{n \in A_- : n \neq n_i$ for all $i < j\}$ are never empty. (Intuitively, we add a non-negative number to the series whenever the partial sum is too low, and add a negative number when the sum is too high.) One can then verify the following claims:

- The map $j \mapsto n_j$ is injective. (Why?)
- Case I occurs an infinite number of times, and Case II also occurs an infinite number of times. (Why? prove by contradiction.)
- The map $j \mapsto n_j$ is surjective. (Why?)
- We have $\lim_{j \to \infty} a_{n_j} = 0$. (Why? Note from Corollary 7.2.6 that $\lim_{n \to \infty} a_n = 0$.)
- We have $\lim_{j \to \infty} \sum_{0 \leq i \leq j} a_{n_i} = L$. (Why?)

The claim then follows by setting $f(i) := n_i$ for all i. □

— Exercises —

Exercise 8.2.1 Prove Lemma 8.2.3. (*Hint:* you may find Exercise 3.6.3 to be useful.)

Exercise 8.2.2 Prove Lemma 8.2.5. (*Hint:* first show that if M is the quantity

$$M := \sup \left\{ \sum_{x \in A} |f(x)| : A \subseteq X, A \text{ finite} \right\}$$

then the sets $\{x \in X : |f(x)| > 1/n\}$ are finite with cardinality at most Mn for every positive integer n. Then use Exercise 8.1.9 (which uses the axiom of choice, see Sect. 8.4).)

Exercise 8.2.3 Prove Proposition 8.2.6. (*Hint:* you may of course use all the results from Chap. 7 to do this.)

Exercise 8.2.4 Prove Lemma 8.2.7. (*Hint:* prove by contradiction, and use limit laws.)

Exercise 8.2.5 Explain the gaps marked (why?) in the proof of Theorem 8.2.8.

Exercise 8.2.6 Let $\sum_{n=0}^{\infty} a_n$ be a series which is conditionally convergent (i.e., convergent but not absolutely convergent). Show that there exists a bijection $f : \mathbf{N} \to \mathbf{N}$ such that $\sum_{m=0}^{\infty} a_{f(m)}$ diverges to $+\infty$, or more precisely that

$$\liminf_{N \to \infty} \sum_{m=0}^{N} a_{f(m)} = \limsup_{N \to \infty} \sum_{m=0}^{N} a_{f(m)} = +\infty.$$

(Of course, a similar statement holds with $+\infty$ replaced by $-\infty$.)

8.3 Uncountable Sets

We have just shown that a lot of infinite sets are countable - even such sets as the rationals, for which it is not obvious how to arrange as a sequence. After such examples, one may begin to hope that other infinite sets, such as the real numbers, are also countable - after all, the real numbers are nothing more than (formal) limits of the rationals, and we've already shown the rationals are countable, so it seems plausible that the reals are also countable.

It was thus a great shock when Georg Cantor (1845–1918) showed in 1873 that certain sets—including the real numbers **R** are in fact uncountable—no matter how hard you try, you cannot arrange the real numbers **R** as a sequence $a_0, a_1, a_2, \ldots$. (Of course, the real numbers **R** can *contain* many infinite sequences, e.g., the sequence $0, 1, 2, 3, 4, \ldots$. However, what Cantor proved is that no such sequence can ever *exhaust* the real numbers; no matter what sequence of real numbers you choose, there will always be some real numbers that are not covered by that sequence.)

Recall from Remark 3.4.11 that if X is a set, then the *power set* of X, denoted $2^X := \{A : A \subseteq X\}$, is the set of all subsets of X. Thus for instance $2^{\{1,2\}} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$. The reason for the notation 2^X is given in Exercise 8.3.1.

Theorem 8.3.1 (Cantor's theorem) *Let X be an arbitrary set (finite or infinite). Then the sets X and 2^X cannot have equal cardinality.*

Proof Suppose for sake of contradiction that the sets X and 2^X had equal cardinality. Then there exists a bijection $f : X \to 2^X$ between X and the power set of X. Now consider the set

$$A := \{x \in X : x \notin f(x)\}.$$

Note that this set is well-defined since $f(x)$ is an element of 2^X and is hence a subset of X. Clearly A is a subset of X, hence is an element of 2^X. Since f is a bijection, there must therefore exist $x \in X$ such that $f(x) = A$. There are now two cases, depending on whether $x \in A$ or $x \notin A$. If $x \in A$, then by definition of A we have $x \notin f(x)$, hence $x \notin A$, a contradiction. But if $x \notin A$, then $x \notin f(x)$, hence by definition of A we have $x \in A$, a contradiction. Thus in either case we have a contradiction. □

Remark 8.3.2 The reader should compare the proof of Cantor's theorem with the statement of Russell's paradox (Sect. 3.2). The point is that a bijection between X and 2^X would come dangerously close to the concept of a set X "containing itself".

Corollary 8.3.3 *$2^{\mathbf{N}}$ is uncountable.*

Proof By Theorem 8.3.1, $2^{\mathbf{N}}$ cannot have equal cardinality with **N**, hence is either uncountable or finite. However, $2^{\mathbf{N}}$ contains as a subset the set of singletons $\{\{n\} : n \in \mathbf{N}\}$, which is clearly bijective to **N** and hence countably infinite. Thus $2^{\mathbf{N}}$ cannot be finite (by Proposition 3.6.14) and is hence uncountable. □

Cantor's theorem has the following important (and unintuitive) consequence.

Corollary 8.3.4 **R** *is uncountable.*

Proof Let us define the map $f: 2^{\mathbf{N}} \to \mathbf{R}$ by the formula

$$f(A) := \sum_{n \in A} 10^{-n}.$$

Observe that since $\sum_{n=0}^{\infty} 10^{-n}$ is an absolutely convergent series (by Lemma 7.3.3), the series $\sum_{n \in A} 10^{-n}$ is also absolutely convergent (by Proposition 8.2.6c). Thus the map f is well-defined. We now claim that f is injective. Suppose for sake of contradiction that there were two distinct sets $A, B \in 2^{\mathbf{N}}$ such that $f(A) = f(B)$. Since $A \neq B$, the set $(A \backslash B) \cup (B \backslash A)$ is a non-empty subset of $\mathbf{N}$. By the well-ordering principle (Proposition 8.1.4), we can then define the minimum of this set, say $n_0 := \min(A \backslash B) \cup (B \backslash A)$. Thus n_0 either lies in $A \backslash B$ or $B \backslash A$. By symmetry we may assume it lies in $A \backslash B$. Then $n_0 \in A$, $n_0 \notin B$, and for all $n < n_0$ we either have $n \in A, B$ or $n \notin A, B$. Thus

$$\begin{aligned}
0 &= f(A) - f(B) \\
&= \sum_{n \in A} 10^{-n} - \sum_{n \in B} 10^{-n} \\
&= \left(\sum_{n < n_0 : n \in A} 10^{-n} + 10^{-n_0} + \sum_{n > n_0 : n \in A} 10^{-n} \right) \\
&\quad - \left(\sum_{n < n_0 : n \in B} 10^{-n} + \sum_{n > n_0 : n \in B} 10^{-n} \right) \\
&= 10^{-n_0} + \sum_{n > n_0 : n \in A} 10^{-n} - \sum_{n > n_0 : n \in B} 10^{-n} \\
&\geq 10^{-n_0} + 0 - \sum_{n > n_0} 10^{-n} \\
&\geq 10^{-n_0} - \frac{1}{9} 10^{-n_0} \\
&> 0,
\end{aligned}$$

a contradiction, where we have used the geometric series lemma (Lemma 7.3.3) to sum

$$\sum_{n > n_0} 10^{-n} = \sum_{m=0}^{\infty} 10^{-(n_0+1+m)} = 10^{-n_0-1} \sum_{m=0}^{\infty} 10^{-m} = \frac{1}{9} 10^{-n_0}.$$

Thus f is injective, which means that $f(2^{\mathbf{N}})$ has the same cardinality as $2^{\mathbf{N}}$ and is thus uncountable. Since $f(2^{\mathbf{N}})$ is a subset of $\mathbf{R}$, this forces $\mathbf{R}$ to be uncountable also (otherwise this would contradict Corollary 8.1.7), and we are done. □

Remark 8.3.5 We will give another proof of this result using measure theory in Exercise 7.2.6 of *Analysis II*.

Remark 8.3.6 Corollary 8.3.4 shows that the reals have strictly larger cardinality than the natural numbers (in the sense of Exercise 3.6.7). One could ask whether there exist any sets which have strictly larger cardinality than the natural numbers, but strictly smaller cardinality than the reals. The *Continuum Hypothesis* asserts that no such sets exist. Interestingly, it was shown in separate works of Kurt Gödel (1906–1978) and Paul Cohen (1934–2007) that this hypothesis is independent of the other axioms of set theory; it can neither be proved nor disproved in that set of axioms (unless those axioms are inconsistent, which is highly unlikely).

— Exercises —

Exercise 8.3.1 Let X be a finite set of cardinality n. Show that 2^X is a finite set of cardinality 2^n. (*Hint:* use induction on n.)

Exercise 8.3.2 Let A, B, C be sets such that $A \subseteq B \subseteq C$, and suppose that there is a injection $f: C \to A$. Define the sets $D_0, D_1, D_2, \ldots$ recursively by setting $D_0 := B \backslash A$, and then $D_{n+1} := f(D_n)$ for all natural numbers n. Prove that the sets $D_0, D_1, \ldots$ are all disjoint from each other (i.e., $D_n \cap D_m = \emptyset$ whenever $n \neq m$). Also show that if $g: A \to B$ is the function defined by setting $g(x) := f^{-1}(x)$ when $x \in \bigcup_{n=1}^{\infty} D_n$, and $g(x) := x$ when $x \notin \bigcup_{n=1}^{\infty} D_n$, then g does indeed map A to B and is a bijection between the two. In particular, A and B have the same cardinality.

Exercise 8.3.3 Recall from Exercise 3.6.7 that a set A is said to have lesser or equal cardinality than a set B iff there is an injective map $f: A \to B$ from A to B. Using Exercise 8.3.2, show that if A, B are sets such that A has lesser or equal cardinality to B and B has lesser or equal cardinality to A, then A and B have equal cardinality. (This is known as the *Schröder–Bernstein theorem*, after Ernst Schröder (1841–1902) and Felix Bernstein (1878–1956).)

Exercise 8.3.4 Let us say that a set A has *strictly lesser cardinality* than a set B if A has lesser than or equal cardinality to B (in the sense of Exercise 3.6.7) but A does not have equal cardinality to B. Show that for any set X, that X has strictly lesser cardinality than 2^X. Also, show that if A has strictly lesser cardinality than B, and B has strictly lesser cardinality than C, then A has strictly lesser cardinality than C.

Exercise 8.3.5 Show that no power set (i.e., a set of the form 2^X for some set X) can be countably infinite.

8.4 The Axiom of Choice

We now discuss the final axiom of the standard Zermelo–Fraenkel–Choice system of set theory, namely the *axiom of choice*. We have delayed introducing this axiom for a while now, to demonstrate that a large portion of the foundations of analysis can be constructed without appealing to this axiom. However, in many further developments of the theory, it is very convenient (and in some cases even essential) to employ this powerful axiom. On the other hand, the axiom of choice can lead to a number of unintuitive consequences (for instance the *Banach–Tarski paradox*, a simplified version of which we will encounter in Sect. 7.3) and can lead to proofs that are philosophically somewhat unsatisfying. Nevertheless, the axiom is almost universally accepted by mathematicians. One reason for this confidence is a theorem due to the great logician Kurt Gödel, who showed that a result proven using the axiom of choice will never contradict a result proven without the axiom of choice (unless all the other axioms of set theory are themselves inconsistent, which is highly unlikely). More precisely, Gödel demonstrated that the axiom of choice is *undecidable*; it can neither be proved nor disproved from the other axioms of set theory, so long as those axioms are themselves consistent. (From a set of inconsistent axioms one can prove that every statement is both true and false.) In practice, this means that any "real-life" application of analysis (more precisely, any application involving only "decidable" questions) which can be rigorously supported using the axiom of choice, can also be rigorously supported without the axiom of choice, though in many cases it would take a much more complicated and lengthier argument to do so if one were not allowed to use the axiom of choice. Thus one can view the axiom of choice as a convenient and safe labor-saving device in analysis. In other disciplines of mathematics, notably in set theory in which many of the questions are not decidable, the issue of whether to accept the axiom of choice is more open to debate and involves some philosophical concerns as well as mathematical and logical ones. However, we will not discuss these issues in this text.

We begin by generalizing the notion of finite Cartesian products from Definition 3.5.6 to infinite Cartesian products.

Definition 8.4.1 *(Infinite Cartesian products)* Let I be a set (possibly infinite), and for each $\alpha \in I$ let X_α be a set. We then define the Cartesian product $\prod_{\alpha \in I} X_\alpha$ to be the set

$$\prod_{\alpha \in I} X_\alpha = \left\{ (x_\alpha)_{\alpha \in I} \in (\bigcup_{\beta \in I} X_\beta)^I : x_\alpha \in X_\alpha \text{ for all } \alpha \in I \right\},$$

where we recall (from Axiom 3.11) that $(\bigcup_{\alpha \in I} X_\alpha)^I$ is the set of all functions $(x_\alpha)_{\alpha \in I}$ which assign an element $x_\alpha \in \bigcup_{\beta \in I} X_\beta$ to each $\alpha \in I$. Thus $\prod_{\alpha \in I} X_\alpha$ is a subset of that set of functions, consisting instead of those functions $(x_\alpha)_{\alpha \in I}$ which assign an element $x_\alpha \in X_\alpha$ to each $\alpha \in I$.

Example 8.4.2 For any sets I and X, we have $\prod_{\alpha \in I} X = X^I$ (why?). If I is a set of the form $I := \{i \in \mathbf{N} : 1 \le i \le n\}$, then $\prod_{\alpha \in I} X_\alpha$ is essentially the same set as the set $\prod_{1 \le i \le N} X_i$ defined in Definition 3.5.6, in the sense that there is a canonical bijection between the two sets (why?).

Recall from Lemma 3.5.11 that if $X_1, \ldots, X_n$ were any finite collection of non-empty sets, then the finite Cartesian product $\prod_{1 \le i \le n} X_i$ was also non-empty. The axiom of choice asserts that this statement is also true for infinite Cartesian products:

Axiom 8.1 (Choice) Let I be a set, and for each $\alpha \in I$, let X_α be a non-empty set. Then $\prod_{\alpha \in I} X_\alpha$ is also non-empty. In other words, there exists a function $(x_\alpha)_{\alpha \in I}$ which assigns to each $\alpha \in I$ an element $x_\alpha \in X_\alpha$.

Remark 8.4.3 The intuition behind this axiom is that given a (possibly infinite) collection of non-empty sets X_α, one should be able to choose a single element x_α from each one, and then form the possibly infinite tuple $(x_\alpha)_{\alpha \in I}$ from all the choices one has made. On the one hand, this is a very intuitively appealing axiom; in some sense one is just applying Lemma 3.1.5 over and over again. On the other hand, the fact that one is making an infinite number of arbitrary choices, with no explicit rule as to *how* to make these choices, is a little disconcerting. Indeed, there are many theorems proven using the axiom of choice which assert the abstract existence of some object x with certain properties, without saying at all *what* that object is, or how to construct it. Thus the axiom of choice can lead to proofs which are *non-constructive*—demonstrating existence of an object without actually constructing the object explicitly. This problem is not unique to the axiom of choice—it already appears for instance in Lemma 3.1.5—but the objects shown to exist using the axiom of choice tend to be rather extreme in their level of non-constructiveness. However, as long as one is aware of the distinction between a non-constructive existence statement, and a constructive existence statement (with the latter being preferable, but not strictly necessary in many cases), there is no difficulty here, except perhaps on a philosophical level.

Remark 8.4.4 There are many equivalent formulations of the axiom of choice; we give some of these in the exercises below.

In analysis one often does not need the full power of the axiom of choice. Instead, one often only needs the *axiom of countable choice*, which is the same as the axiom of choice but with the index set I restricted to be at most countable. We give a typical example of this below.

Lemma 8.4.5 *Let E be a non-empty subset of the real line with* $\sup(E) < \infty$ (*i.e., E is bounded from above*). *Then there exists a sequence* $(a_n)_{n=1}^\infty$ *whose elements a_n all lie in E, such that* $\lim_{n\to\infty} a_n = \sup(E)$.

Proof For each positive natural number n, let X_n denote the set

$$X_n := \{x \in E : \sup(E) - 1/n \le x \le \sup(E)\}.$$

Since $\sup(E)$ is the least upper bound for E, then $\sup(E) - 1/n$ cannot be an upper bound for E, and so X_n is non-empty for each n. Using the axiom of choice (or the axiom of countable choice), we can then find a sequence $(a_n)_{n=1}^\infty$ such that $a_n \in X_n$ for all $n \ge 1$. In particular $a_n \in E$ for all n, and $\sup(E) - 1/n \le a_n \le \sup(E)$ for all n. But then we have $\lim_{n\to\infty} a_n = \sup(E)$ by the squeeze test (Corollary 6.4.14). □

Remark 8.4.6 In many special cases, one can obtain the conclusion of this lemma without using the axiom of choice. For instance, if E is a closed set (Definition 1.2.12), then one can define a_n without choice by the formula $a_n := \inf(X_n)$; the extra hypothesis that E is closed will ensure that a_n lies in E.

Another formulation of the axiom of choice is as follows.

Proposition 8.4.7 *Let X and Y be sets, and let $P(x, y)$ be a property pertaining to an object $x \in X$ and an object $y \in Y$ such that for every $x \in X$ there is at least one $y \in Y$ such that $P(x, y)$ is true. Then there exists a function $f : X \to Y$ such that $P(x, f(x))$ is true for all $x \in X$.*

Proof See Exercise 8.4.1. □

— Exercises —

Exercise 8.4.1 Show that the axiom of choice implies Proposition 8.4.7. (*Hint:* consider the sets $Y_x := \{y \in Y : P(x, y) \text{ is true}\}$ for each $x \in X$.) Conversely, show that if Proposition 8.4.7 is true, then the axiom of choice is also true.

Exercise 8.4.2 Let I be a set, and for each $\alpha \in I$ let X_α be a non-empty set. Suppose that all the sets X_α are disjoint from each other, i.e., $X_\alpha \cap X_\beta = \emptyset$ for all distinct $\alpha, \beta \in I$. Using the axiom of choice, show that there exists a set Y such that $\#(Y \cap X_\alpha) = 1$ for all $\alpha \in I$ (i.e., Y intersects each X_α in exactly one element). Conversely, show that if the above statement was true for an arbitrary choice of sets I and non-empty disjoint sets X_α, then the axiom of choice is true. (*Hint:* the problem is that in Axiom 8.1 the sets X_α are not assumed to be disjoint. But this can be fixed by the trick by looking at the sets $\{\alpha\} \times X_\alpha = \{(\alpha, x) : x \in X_\alpha\}$ instead.)

Exercise 8.4.3 Let A and B be sets such that there exists a surjection $g : B \to A$. Using the axiom of choice, show that there then exists an injection $f : A \to B$ with $g \circ f : A \to A$ the identity map; in particular, A has lesser or equal cardinality to B in the sense of Exercise 3.6.7. (*Hint:* consider the inverse images $g^{-1}(\{a\})$ for each $a \in A$.) Compare this with Exercise 3.6.8. Conversely, show that if the above statement is true for arbitrary sets A, B and surjections $g : B \to A$, then the axiom of choice is true. (*Hint:* use Exercise 8.4.2.)

8.5 Ordered Sets

The axiom of choice is intimately connected to the theory of *ordered sets*. There are actually many types of ordered sets; we will concern ourselves with three such types, the *partially ordered sets*, the *totally ordered sets*, and the *well-ordered sets*.

Definition 8.5.1 *(Partially ordered sets)* A *partially ordered set* (or *poset*) is a set X, together[1] with a relation $\leq_X$ on X (thus for any two objects $x, y \in X$, the statement $x \leq_X y$ is either a true statement or a false statement). Furthermore, this relation is assumed to obey the following three properties:

[1] Strictly speaking, a partially ordered set is not a set X, but rather a pair $(X, \leq_X)$. But in many cases the ordering $\leq_X$ will be clear from context, and so we shall refer to X itself as the partially ordered set even though this is technically incorrect.

- (Reflexivity) For any $x \in X$, we have $x \leq_X x$.
- (Antisymmetry) If $x, y \in X$ are such that $x \leq_X y$ and $y \leq_X x$, then $x = y$.
- (Transitivity) If $x, y, z \in X$ are such that $x \leq_X y$ and $y \leq_X z$, then $x \leq_X z$.

We refer to $\leq_X$ as the *ordering relation*. In most situations it is understood what the set X is from context, and in those cases we shall simply write $\leq$ instead of $\leq_X$. We write $x <_X y$ (or $x < y$ for short) if $x \leq_X y$ and $x \neq y$.

Examples 8.5.2 The natural numbers **N** together with the usual less-than-or-equal-to relation $\leq$ (as defined in Definition 2.2.11) forms a partially ordered set, by Proposition 2.2.12. Similar arguments (using the appropriate definitions and propositions) show that the integers **Z**, the rationals **Q**, the reals **R**, and the extended reals $\mathbf{R}^*$ are also partially ordered sets. Meanwhile, if X is any collection of sets, and one uses the relation of is-a-subset-of $\subseteq$ (as defined in Definition 3.1.14) for the ordering relation $\leq_X$, then X is also partially ordered (Proposition 3.1.17). Note that it is certainly possible to give these sets a different partial ordering than the standard one; see for instance Exercise 8.5.3.

Definition 8.5.3 *(Totally ordered set)* Let X be a partially ordered set with some order relation $\leq_X$. A subset Y of X is said to be *totally ordered* if, given any two $y, y' \in Y$, we either have $y \leq_X y'$ or $y' \leq_X y$ (or both). If X itself is totally ordered, we say that X is a *totally ordered set* (or *chain*) with order relation $\leq_X$.

Examples 8.5.4 The natural numbers **N**, the integers **Z**, the rationals **Q**, reals **R**, and the extended reals $\mathbf{R}^*$, all with the usual ordering relation $\leq$, are totally ordered (by Proposition 2.2.13, Lemma 4.1.11, Proposition 4.2.9, Proposition 5.4.7, and Proposition 6.2.5, respectively). Also, any subset of a totally ordered set is again totally ordered (why?). On the other hand, a collection of sets with the $\subseteq$ relation is usually not totally ordered. For instance, if X is the set $\{\{1, 2\}, \{2\}, \{2, 3\}, \{2, 3, 4\}, \{5\}\}$, ordered by the set inclusion relation $\subseteq$, then the elements $\{1, 2\}$ and $\{2, 3\}$ of X are not comparable to each other (i.e., $\{1, 2\} \not\subseteq \{2, 3\}$ and $\{2, 3\} \not\subseteq \{1, 2\}$).

Definition 8.5.5 *(Maximal and minimal elements)* Let X be a partially ordered set, and let Y be a subset of X. We say that y is a *minimal element of* Y if $y \in Y$ and there is no element $y' \in Y$ such that $y' < y$. We say that y is a *maximal element of* Y if $y \in Y$ and there is no element $y' \in Y$ such that $y < y'$.

Example 8.5.6 Using the set X from the previous example, $\{2\}$ is a minimal element, $\{1, 2\}$ and $\{2, 3, 4\}$ are maximal elements, $\{5\}$ is both a minimal and a maximal element, and $\{2, 3\}$ is neither a minimal nor a maximal element. This example shows that a partially ordered set can have multiple maxima and minima; however, a totally ordered set cannot (Exercise 8.5.7).

Example 8.5.7 The natural numbers **N** (ordered by $\leq$) have a minimal element, namely 0, but no maximal element. The set of integers **Z** has no maximal and no minimal element.

Definition 8.5.8 *(Well-ordered sets)* Let X be a partially ordered set, and let Y be a totally ordered subset of X. We say that Y is *well-ordered* if every non-empty subset Z of Y has a minimal element $\min(Z)$.

Examples 8.5.9 The natural numbers $\mathbf{N}$ are well-ordered by Proposition 8.1.4. However, the integers $\mathbf{Z}$, the rationals $\mathbf{Q}$, and the real numbers $\mathbf{R}$ are not (see Exercise 8.1.2). Every finite totally ordered set is well-ordered (Exercise 8.5.8). Every subset of a well-ordered set is again well-ordered (why?).

One advantage of well-ordered sets is that they automatically obey a principle of strong induction (cf. Proposition 2.2.14):

Proposition 8.5.10 (Principle of strong induction) *Let X be a well-ordered set with an ordering relation $\leq$, and let $P(n)$ be a property pertaining to an element $n \in X$ (i.e., for each $n \in X$, $P(n)$ is either a true statement or a false statement). Suppose that for every $n \in X$, we have the following implication: if $P(m)$ is true for all $m \in X$ with $m <_X n$, then $P(n)$ is also true. Then $P(n)$ is true for all $n \in X$.*

Remark 8.5.11 It may seem strange that there is no "base" case in strong induction, corresponding to the hypothesis $P(0)$ in Axiom 2.5. However, such a base case is automatically included in the strong induction hypothesis. Indeed, if 0 is the minimal element of X, then by specializing the hypothesis "if $P(m)$ is true for all $m \in X$ with $m <_X n$, then $P(n)$ is also true" to the $n = 0$ case, we automatically obtain that $P(0)$ is true. (Why?)

Proof See Exercise 8.5.10. □

So far we have not seen the axiom of choice play any rôle. This will come in once we introduce the notion of an upper bound and a strict upper bound.

Definition 8.5.12 *(Upper bounds and strict upper bounds)* Let X be a partially ordered set with ordering relation $\leq$, and let Y be a subset of X. If $x \in X$, we say that x is an *upper bound* for Y iff $y \leq x$ for all $y \in Y$. If in addition $x \notin Y$, we say that x is a *strict upper bound* for Y. Equivalently, x is a strict upper bound for Y iff $y < x$ for all $y \in Y$. (Why is this equivalent?)

Example 8.5.13 Let us work in the real number system $\mathbf{R}$ with the usual ordering $\leq$. Then 2 is an upper bound for the set $\{x \in \mathbf{R} : 1 \leq x \leq 2\}$ but is not a strict upper bound. The number 3, on the other hand, is a strict upper bound for this set.

Lemma 8.5.14 *Let X be a partially ordered set with ordering relation $\leq$, and let x_0 be an element of X. Then there is a well-ordered subset Y of X which has x_0 as its minimal element and which has no strict upper bound.*

Proof The intuition behind this lemma is that one is trying to perform the following algorithm: we initalize $Y := \{x_0\}$. If Y has no strict upper bound, then we are done; otherwise, we choose a strict upper bound and add it to Y. Then we look again to see if Y has a strict upper bound or not. If not, we are done; otherwise we choose another strict upper bound and add it to Y. We continue this algorithm “infinitely often” until we exhaust all the strict upper bounds; the axiom of choice comes in because infinitely many choices are involved. This is however not a rigorous proof because it is quite difficult to precisely pin down what it means to perform an algorithm “infinitely often”. Instead, what we will do is that we will isolate a collection of “partially completed” sets Y, which we shall call *good sets*, and then take the union of all these good sets to obtain a “completed” object Y_∞ which will indeed have no strict upper bound.

We now begin the rigorous proof. Suppose for sake of contradiction that every well-ordered subset Y of X which has x_0 as its minimal element has at least one strict upper bound. Using the axiom of choice (in the form of Proposition 8.4.7), we can thus assign a strict upper bound $s(Y) \in X$ to each well-ordered subset Y of X which has x_0 as its minimal element.

Henceforth we fix a single such strict upper bound function s. Let us define a special class of subsets Y of X. We say that a subset Y of X is *good* iff it is well-ordered, contains x_0 as its minimal element, and obeys the property that

$$x = s(\{y \in Y : y < x\}) \text{ for all } x \in Y\backslash\{x_0\}.$$

Note that if $x \in Y\backslash\{x_0\}$ then the set $\{y \in Y : y < x\}$ is a subset of X which is well-ordered and contains x_0 as its minimal element. Let $\Omega := \{Y \subseteq X : Y \text{ is good}\}$ be the collection of all good subsets of X. This collection is not empty, since the subset $\{x_0\}$ of X is clearly good (why?).

We make the following important observation: if Y and Y' are two good subsets of X, then every element of $Y'\backslash Y$ is a strict upper bound for Y, and every element of $Y\backslash Y'$ is a strict upper bound for Y' (Exercise 8.5.13). In particular, given any two good sets Y and Y', at least one of $Y'\backslash Y$ and $Y\backslash Y'$ must be empty (since they are both strict upper bounds of each other). In other words, Ω is totally ordered by set inclusion: given any two good sets Y and Y', either $Y \subseteq Y'$ or $Y' \subseteq Y$.

Let $Y_\infty := \bigcup \Omega$, i.e., Y_∞ is the set of all elements of X which belong to at least one good subset of X. Clearly $x_0 \in Y_\infty$. Also, since each good subset of X has x_0 as its minimal element, the set Y_∞ also has x_0 as its minimal element (why?).

Next, we show that Y_∞ is totally ordered. Let x, x' be two elements of Y_∞. By definition of Y_∞, we know that x lies in some good set Y and x' lies in some good set Y'. But since Ω is totally ordered, one of these good sets contains the other. Thus x, x' are contained in a single good set (either Y or Y'); since good sets are totally ordered, we thus see that either $x \leq x'$ or $x' \leq x$ as desired.

Next, we show that Y_∞ is well-ordered. Let A be any non-empty subset of Y_∞. Then we can pick an element $a \in A$, which then lies in Y_∞. Therefore there is a good set Y such that $a \in Y$. Then $A \cap Y$ is a non-empty subset of Y; since Y is well-ordered, the set $A \cap Y$ thus has a minimal element, call it b. Now recall that for any other good set Y', every element of $Y'\backslash Y$ is a strict upper bound for Y, and in particular is larger than b. Since b is a minimal element of $A \cap Y$, this implies that b is also a minimal element of $A \cap Y'$ for any good set Y' with $A \cap Y' \neq \emptyset$ (why?).

Since every element of A belongs to Y_∞ and hence belongs to at least one good set Y', we thus see that b is a minimal element of A. Thus Y_∞ is well-ordered as claimed.

Since Y_∞ is well-ordered with x_0 as its minimal element, it has a strict upper bound $s(Y_\infty)$. But then $Y_\infty \cup \{s(Y_\infty)\}$ is well-ordered (why? see Exercise 8.5.11) and has x_0 as its minimal element (why?).

We now claim that $Y_\infty \cup \{s(Y_\infty)\}$ is good. By the preceding discussion, it suffices to show that $x = s(\{y \in Y_\infty \cup \{s(Y_\infty)\} : y < x\}$ when $x \in (Y_\infty \cup \{s(Y_\infty)\})\backslash\{x_0\}$. If $x = s(Y_\infty)$, this is clear since $\{y \in Y_\infty \cup \{s(Y_\infty)\} : y < x\} = Y_\infty$ in this case. If instead $x \in Y_\infty$, then $x \in Y$ for some good Y. Then the set $\{y \in Y_\infty \cup \{s(Y_\infty)\} : y < x\}$ is equal to $\{y \in Y : y < x\}$ (why? Use the previous observation that every element of $Y'\backslash Y$ is an upper bound for x for every good Y'), and the claim then follows since Y is good.

By definition of Y_∞, we conclude that the good set $Y_\infty \cup \{s(Y_\infty)\}$ is contained in Y_∞. But this is a contradiction since $s(Y_\infty)$ is a *strict* upper bound for Y_∞. Thus we have constructed a set with no strict upper bound, as desired. □

The above lemma has the following important consequence:

Lemma 8.5.15 (Zorn's lemma) *Let X be a non-empty partially ordered set, with the property that every non-empty totally ordered subset Y of X has an upper bound. Then X contains at least one maximal element.*

Proof See Exercise 8.5.14. □

We give some applications of Zorn's lemma in the exercises below.

— Exercises —

Exercise 8.5.1 Consider the empty set $\emptyset$ with the empty order relation $\leq_\emptyset$ (this relation is vacuous because the empty set has no elements). Is this set partially ordered? totally ordered? well-ordered? Explain.

Exercise 8.5.2 Give examples of a set X and a relation $\leq$ such that

(a) The relation $\leq$ is reflexive and antisymmetric, but not transitive;
(b) The relation $\leq$ is reflexive and transitive, but not antisymmetric;
(c) The relation $\leq$ is antisymmetric and transitive, but not reflexive.

Exercise 8.5.3 Given two positive integers $n, m \in \mathbf{N}\backslash\{0\}$, we say that n *divides* m, and write $n|m$, if there exists a positive integer a such that $m = na$. Show that the set $\mathbf{N}\backslash\{0\}$ with the ordering relation $|$ is a partially ordered set but not a totally ordered one. Note that this is a different ordering relation from the usual $\leq$ ordering of $\mathbf{N}\backslash\{0\}$.

Exercise 8.5.4 Show that the set of positive reals $\mathbf{R}^+ := \{x \in \mathbf{R} : x > 0\}$ have no minimal element.

Exercise 8.5.5 Let $f : X \to Y$ be a function from one set X to another set Y. Suppose that Y is partially ordered with some ordering relation $\leq_Y$. Define a relation $\leq_X$ on X by defining $x \leq_X x'$ if and only if $f(x) <_Y f(x')$ or $x = x'$. Show that this relation $\leq_X$ turns X into a partially ordered set. If we know in addition that the relation $\leq_Y$ makes Y totally ordered, does this mean that the relation $\leq_X$ makes X totally ordered also? If not, what additional assumption needs to be made on f in order to ensure that $\leq_X$ makes X totally ordered?

Exercise 8.5.6 Let X be a partially ordered set. For any x in X, define the *order ideal* $(x) \subseteq X$ to be the set $(x) := \{y \in X : y \leq x\}$. Let $(X) := \{(x) : x \in X\}$ be the set of all order ideals, and let $f: X \to (X)$ be the map $f(x) := (x)$ that sends every element of X to its order ideal. Show that f is a bijection, and that given any $x, y \in X$, that $x \leq_X y$ if and only if $f(x) \subseteq f(y)$. This exercise shows that any partially ordered set can be *represented* by a collection of sets whose ordering relation is given by set inclusion.

Exercise 8.5.7 Let X be a partially ordered set, and let Y be a totally ordered subset of X. Show that Y can have at most one maximum and at most one minimum.

Exercise 8.5.8 Show that every finite non-empty subset of a totally ordered set has a minimum and a maximum. (*Hint:* use induction.) Conclude in particular that every finite totally ordered set is well-ordered.

Exercise 8.5.9 Let X be a totally ordered set such that every non-empty subset of X has both a minimum and a maximum. Show that X is finite. (*Hint:* assume for sake of contradiction that X is infinite. Start with the minimal element x_0 of X and then construct an increasing sequence $x_0 < x_1 < \ldots$ in X.)

Exercise 8.5.10 Prove Proposition 8.5.10, without using the axiom of choice. (*Hint:* consider the set

$$Y := \{n \in X : P(m) \text{ is false for some } m \in X \text{ with } m \leq_X n\},$$

and show that Y being non-empty would lead to a contradiction.)

Exercise 8.5.11 Let X be a partially ordered set, and let Y and Y' be well-ordered subsets of X. Show that $Y \cup Y'$ is well-ordered if and only if it is totally ordered.

Exercise 8.5.12 Let X and Y be partially ordered sets with ordering relations $\leq_X$ and $\leq_Y$, respectively. Define a relation $\leq_{X \times Y}$ on the Cartesian product $X \times Y$ by defining $(x, y) \leq_{X \times Y} (x', y')$ if $x <_X x'$, or if $x = x'$ and $y \leq_Y y'$. (This is called the *lexicographical ordering* on $X \times Y$, and is similar to the alphabetical ordering of words; a word w appears earlier in a dictionary than another word w' if the first letter of w is earlier in the alphabet than the first letter of w', or if the first letters match and the second letter of w is earlier than the second letter of w', and so forth.) Show that $\leq_{X \times Y}$ defines a partial ordering on $X \times Y$. Furthermore, show that if X and Y are totally ordered, then so is $X \times Y$, and if X and Y are well-ordered, then so is $X \times Y$.

Exercise 8.5.13 Prove the claim in the proof of Lemma 8.5.14, namely that every element of $Y' \backslash Y$ is an upper bound for Y and vice versa. (*Hint:* Show using Proposition 8.5.10 that

$$\{y \in Y : y \leq a\} = \{y \in Y' : y \leq a\} = \{y \in Y \cap Y' : y \leq a\}$$

for all $a \in Y \cap Y'$. Conclude that $Y \cap Y'$ is good, and hence $s(Y \cap Y')$ exists. Show that $s(Y \cap Y') = \min(Y' \backslash Y)$ if $Y' \backslash Y$ is non-empty, and similarly with Y and Y' interchanged. Since $Y' \backslash Y$ and $Y \backslash Y'$ are disjoint, one can then conclude that one of these sets is empty, at which point the claim becomes easy to establish.)

Exercise 8.5.14 Use Lemma 8.5.14 to prove Lemma 8.5.15. (*Hint:* first show that if X had no maximal elements, then any subset of X which has an upper bound, also has a strict upper bound.)

Exercise 8.5.15 Let A and B be two non-empty sets such that A does not have lesser or equal cardinality to B. Using Zorn's lemma, prove that B has lesser or equal cardinality to A. (*Hint:* for every subset $X \subseteq B$, let $P(X)$ denote the property that there exists an injective map from X

to A.) This exercise (combined with Exercise 8.3.3) shows that the cardinality of any two sets is comparable, as long as one assumes the axiom of choice.

Exercise 8.5.16 Let X be a set, and let P be the set of all partial orderings of X. (For instance, if $X := \mathbf{N}\backslash\{0\}$, then both the usual partial ordering $\leq$, and the partial ordering in Exercise 8.5.3, are elements of P.) We say that one partial ordering $\leq \in P$ is *coarser* than another partial ordering $\leq' \in P$ if for any $x, y \in X$, we have the implication $(x \leq y) \implies (x \leq' y)$. Thus for instance the partial ordering in Exercise 8.5.3 is coarser than the usual ordering $\leq$. Let us write $\leq \preceq \leq'$ if $\leq$ is coarser than $\leq'$. Show that $\preceq$ turns P into a partially ordered set; thus the set of partial orderings on X is itself partially ordered. There is exactly one minimal element of P; what is it? Show that the maximal elements of P are precisely the total orderings of X. Using Zorn's lemma, show that given any partial ordering $\leq$ of X there exists a total ordering $\leq'$ such that $\leq$ is coarser than $\leq'$.

Exercise 8.5.17 Use Zorn's lemma to give another proof of the claim in Exercise 8.4.2. (*Hint:* let Ω be the set of all $Y \subseteq \bigcup_{\alpha \in I} X_\alpha$ such that $\#(Y \cap X_\alpha) \leq 1$ for all $\alpha \in I$, i.e., all sets which intersect each X_α in at most one element. Use Zorn's lemma to locate a maximal element of Ω.) Deduce that Zorn's lemma and the axiom of choice are in fact logically equivalent (i.e., they can be deduced from each other).

Exercise 8.5.18 Using Zorn's lemma, prove *Hausdorff's maximality principle*: if X is a partially ordered set, then there exists a totally ordered subset Y of X which is maximal with respect to set inclusion (i.e., there is no other totally ordered subset Y' of X which contains Y). Conversely, show that if Hausdorff's maximality principle is true, then Zorn's lemma is true. Thus by Exercise 8.5.17, these two statements are logically equivalent to the axiom of choice.

Exercise 8.5.19 Let X be a set, and let Ω be the space of all pairs $(Y, \leq)$, where Y is a subset of X and $\leq$ is a well-ordering of Y. If $(Y, \leq)$ and $(Y', \leq')$ are elements of Ω, we say that $(Y, \leq)$ is an *initial segment* of $(Y', \leq')$ if there exists an $x \in Y'$ such that $Y := \{y \in Y' : y <' x\}$ (so in particular $Y \subsetneq Y'$), and for any $y, y' \in Y$, $y \leq y'$ if and only if $y \leq' y'$. Define a relation $\preceq$ on Ω by defining $(Y, \leq) \preceq (Y', \leq')$ if either $(Y, \leq) = (Y', \leq')$, or if $(Y, \leq)$ is an initial segment of $(Y', \leq')$. Show that $\preceq$ is a partial ordering of Ω. There is exactly one minimal element of Ω; what is it? Show that the maximal elements of Ω are precisely the well-orderings $(X, \leq)$ of X. Using Zorn's lemma, conclude the *well-ordering principle*: every set X has at least one well-ordering. Conversely, use the well-ordering principle to prove the axiom of choice, Axiom 8.1. (*Hint:* place a well-ordering $\leq$ on $\bigcup_{\alpha \in I} X_\alpha$, and then consider the minimal elements of each X_α.) We thus see that the axiom of choice, Zorn's lemma, and the well-ordering principle are all logically equivalent to each other.

Exercise 8.5.20 Let X be a set, and let $\Omega \subseteq 2^X$ be a collection of subsets of X. Assume that Ω does not contain the empty set $\emptyset$. Using Zorn's lemma, show that there is a subcollection $\Omega' \subseteq \Omega$ such that all the elements of Ω' are disjoint from each other (i.e., $A \cap B = \emptyset$ whenever A, B are distinct elements of Ω'), but that all the elements of Ω intersect at least one element of Ω' (i.e., for all $C \in \Omega$ there exists $A \in \Omega'$ such that $C \cap A \neq \emptyset$). (*Hint:* consider all the subsets of Ω whose elements are all disjoint from each other, and locate a maximal element of this collection.) Conversely, if the above claim is true, show that it implies the claim in Exercise 8.4.2, and thus this is yet another claim which is logically equivalent to the axiom of choice. (*Hint:* let Ω be the set of all pair sets of the form $\{(0, \alpha), (1, x_\alpha)\}$, where $\alpha \in I$ and $x_\alpha \in X_\alpha$.)

Chapter 9
Continuous Functions on R

In previous chapters we have been focusing primarily on *sequences*. A sequence $(a_n)_{n=0}^{\infty}$ can be viewed as a function from $\mathbf{N}$ to $\mathbf{R}$, i.e., an object which assigns a real number a_n to each natural number n. We then did various things with these functions from $\mathbf{N}$ to $\mathbf{R}$, such as take their limit at infinity (if the function was convergent), or form suprema, infima, etc., or computed the sum of all the elements in the sequence (again, assuming the series was convergent).

Now we will look at functions not on the natural numbers $\mathbf{N}$, which are "discrete", but instead look at functions on a *continuum*[1] such as the real line $\mathbf{R}$, or perhaps on an interval such as $\{x \in \mathbf{R} : a \leq x \leq b\}$. Eventually we will perform a number of operations on these functions, including taking limits, computing derivatives, and evaluating integrals. In this chapter we will focus primarily on limits of functions, and on the closely related concept of a *continuous function*.

Before we discuss functions, though, we must first set out some notation for subsets of the real line.

9.1 Subsets of the Real Line

Very often in analysis we do not work on the whole real line $\mathbf{R}$, but on certain subsets of the real line, such as the positive real axis $\{x \in \mathbf{R} : x > 0\}$. Also, we occasionally work with the extended real line $\mathbf{R}^*$ defined in Sect. 6.2, or in subsets of that extended real line.

[1] We will not rigorously define the notion of a discrete set or a continuum in this text, but roughly speaking a set is discrete if each element is separated from the rest of the set by some non-zero distance, whereas a set is a *continuum* if it is connected and contains no "holes".

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_9

There are of course infinitely many subsets of the real line; indeed, Cantor's theorem (Theorem 8.3.1; see also Exercise 8.3.4) shows that there are even more such sets than there are real numbers. However, there are certain special subsets of the real line (and the extended real line) which arise quite often. One such family of sets are the *intervals*.

Definition 9.1.1 *(Intervals)* Let $a, b \in \mathbf{R}^*$ be extended real numbers. We define the *closed interval* $[a, b]$ by

$$[a, b] := \{x \in \mathbf{R}^* : a \leq x \leq b\},$$

the *half-open intervals* $[a, b)$ and $(a, b]$ by

$$[a, b) := \{x \in \mathbf{R}^* : a \leq x < b\}; \quad (a, b] := \{x \in \mathbf{R}^* : a < x \leq b\},$$

and the *open interval* (a, b) by

$$(a, b) := \{x \in \mathbf{R}^* : a < x < b\}.$$

We call a the *left endpoint* of these intervals, and b the *right endpoint*.

Remark 9.1.2 Once again, we are overloading the parenthesis notation; for instance, we are now using $(2, 3)$ to denote both an open interval from 2 to 3 and an ordered pair in the Cartesian plane $\mathbf{R}^2 := \mathbf{R} \times \mathbf{R}$. This can cause some genuine ambiguity, but the reader should still be able to resolve which meaning of the parentheses is intended from context. In some texts, this issue is resolved by using reversed brackets instead of parenthesis, and thus for instance $[a, b)$ would now be $[a, b[$, $(a, b]$ would be $]a, b]$, and (a, b) would be $]a, b[$.

Examples 9.1.3 If a and b are real numbers (i.e., not equal to $+\infty$ or $-\infty$), then all of the above intervals are subsets of the real line, for instance $[2, 3) = \{x \in \mathbf{R} : 2 \leq x < 3\}$. The positive real axis $\{x \in \mathbf{R} : x > 0\}$ is the open interval $(0, +\infty)$, while the non-negative real axis $\{x \in \mathbf{R} : x \geq 0\}$ is the half-open interval $[0, +\infty)$. Similarly, the negative real axis $\{x \in \mathbf{R} : x < 0\}$ is $(-\infty, 0)$, and the non-positive real axis $\{x \in \mathbf{R} : x \leq 0\}$ is $(-\infty, 0]$. Finally, the real line $\mathbf{R}$ itself is the open interval $(-\infty, +\infty)$, while the extended real line $\mathbf{R}^*$ is the closed interval $[-\infty, +\infty]$. We sometimes refer to an interval in which one endpoint is infinite (either $+\infty$ or $-\infty$) as *half-infinite* intervals, and intervals in which both endpoints are infinite as *doubly infinite* intervals; all other intervals are *bounded intervals*. Thus $[2, 3)$ is a bounded interval, the positive and negative real axes are half-infinite intervals, and $\mathbf{R}$ and $\mathbf{R}^*$ are infinite intervals.

Example 9.1.4 If $a > b$, then all four of the intervals $[a, b]$, $[a, b)$, $(a, b]$, and (a, b) are the empty set (why?). If $a = b$, then the three intervals $[a, b)$, $(a, b]$, and (a, b) are the empty set, while $[a, b]$ is just the singleton set $\{a\}$ (why?). Because of this, we call these intervals *degenerate*; most (but not all) of our analysis will be restricted to non-degenerate intervals.

Of course intervals are not the only interesting subsets of the real line. Other important examples include the *natural numbers* $\mathbf{N}$, the *integers* $\mathbf{Z}$, and the *rationals* $\mathbf{Q}$. One can form additional sets using such operations as union and intersection (see Sect. 3.1); for instance one could have a disconnected union of two intervals such as $(1, 2) \cup [3, 4]$, or one could consider the set $[-1, 1] \cap \mathbf{Q}$ of rational numbers between -1 and 1 inclusive. Clearly there are infinitely many possibilities of sets one could create by such operations.

Just as sequences of real numbers have limit points, sets of real numbers have *adherent points*, which we now define.

Definition 9.1.5 *(ε-adherent points)* Let X be a subset of $\mathbf{R}$, let $\varepsilon > 0$, and let $x \in \mathbf{R}$. We say that x is *ε-adherent to X* iff there exists a $y \in X$ which is ε-close to x (i.e., $|x - y| \leq \varepsilon$).

Remark 9.1.6 The terminology "ε-adherent" is not standard in the literature. However, we shall shortly use it to define the notion of an adherent point, which is standard.

Example 9.1.7 The point 1.1 is 0.5-adherent to the open interval $(0, 1)$, but is not 0.1-adherent to this interval (why?). The point 1.1 is 0.5-adherent to the finite set $\{1, 2, 3\}$. The point 1 is 0.5-adherent to $\{1, 2, 3\}$ (why?).

Definition 9.1.8 *(Adherent points)* Let X be a subset of $\mathbf{R}$, and let $x \in \mathbf{R}$. We say that x is an *adherent point* of X iff it is ε-adherent to X for every $\varepsilon > 0$.

Example 9.1.9 The number 1 is ε-adherent to the open interval $(0, 1)$ for every $\varepsilon > 0$ (why?) and is thus an adherent point of $(0, 1)$. The point 0.5 is similarly an adherent point of $(0, 1)$. However, the number 2 is not 0.5-adherent (for instance) to $(0, 1)$ and is thus not an adherent point to $(0, 1)$.

Definition 9.1.10 *(Closure)* Let X be a subset of $\mathbf{R}$. The *closure* of X, sometimes denoted $\overline{X}$ is defined to be the set of all the adherent points of X.

Lemma 9.1.11 (Elementary properties of closures) *Let X and Y be arbitrary subsets of* $\mathbf{R}$. *Then $X \subseteq \overline{X}$, $\overline{X \cup Y} = \overline{X} \cup \overline{Y}$, and $\overline{X \cap Y} \subseteq \overline{X} \cap \overline{Y}$. If $X \subseteq Y$, then $\overline{X} \subseteq \overline{Y}$.*

Proof See Exercise 9.1.1. □

We now compute some closures.

Lemma 9.1.12 (Closures of intervals) *Let $a < b$ be real numbers, and let I be any one of the four intervals (a, b), $(a, b]$, $[a, b)$, or $[a, b]$. Then the closure of I is $[a, b]$. Similarly, the closure of (a, ∞) or $[a, \infty)$ is $[a, \infty)$, while the closure of $(-\infty, a)$ or $(-\infty, a]$ is $(-\infty, a]$. Finally, the closure of $(-\infty, \infty)$ is $(-\infty, \infty)$.*

Proof We will just show one of these facts, namely that the closure of (a, b) is $[a, b]$; the other results are proven similarly (or one can use Exercise 9.1.6).

First let us show that every element of $[a, b]$ is adherent to (a, b). Let $x \in [a, b]$. If $x \in (a, b)$, then it is definitely adherent to (a, b). If $x = b$, then x is also adherent to (a, b) (why?). Similarly when $x = a$. Thus every point in $[a, b]$ is adherent to (a, b).

Now we show that every point x that is adherent to (a, b) lies in $[a, b]$. Suppose for sake of contradiction that x does not lie in $[a, b]$, then either $x > b$ or $x < a$. If $x > b$ then x is not $(x - b)$-adherent to (a, b) (why?) and is hence not an adherent point to (a, b). Similarly, if $x < a$, then x is not $(a - x)$-adherent to $(a - b)$ and is hence not an adherent point to (a, b). This contradiction shows that x is in fact in $[a, b]$ as claimed. □

Lemma 9.1.13 *The closure of* **N** *is* **N**. *The closure of* **Z** *is* **Z**. *The closure of* **Q** *is* **R**, *and the closure of* **R** *is* **R**. *The closure of the empty set* ∅ *is* ∅.

Proof See Exercise 9.1.2. □

The following lemma shows that adherent points of a set X can be obtained as the limit of elements in X:

Lemma 9.1.14 *Let X be a subset of* **R**, *and let $x \in$* **R**. *Then x is an adherent point of X if and only if there exists a sequence $(a_n)_{n=0}^{\infty}$, consisting entirely of elements in X, which converges to x.*

Proof See Exercise 9.1.4. □

Definition 9.1.15 A subset $E \subseteq \mathbf{R}$ is said to be *closed* if $\overline{E} = E$, or in other words that E contains all of its adherent points.

Examples 9.1.16 From Lemma 9.1.12 we see that if $a < b$ are real numbers, then $[a, b]$, $[a, +\infty)$, $(-\infty, a]$, and $(-\infty, +\infty)$ are closed, while (a, b), $(a, b]$, $[a, b)$, $(a, +\infty)$, and $(-\infty, a)$ are not. From Lemma 9.1.13 we see that **N**, **Z**, **R**, ∅ are closed, while **Q** is not.

From Lemma 9.1.14 we can define closure in terms of sequences:

Corollary 9.1.17 *Let X be a subset of* **R**. *If X is closed, and $(a_n)_{n=0}^{\infty}$ is a convergent sequence consisting of elements in X, then* $\lim_{n\to\infty} a_n$ *also lies in X. Conversely, if it is true that every convergent sequence $(a_n)_{n=0}^{\infty}$ of elements in X has its limit in X as well, then X is necessarily closed.*

When we study differentiation in the next chapter, we shall need to replace the concept of an adherent point by the closely related notion of a *limit point*.

Definition 9.1.18 *(Limit points)* Let X be a subset of the real line. We say that x is a *limit point* (or a *cluster point*) of X iff it is an adherent point of $X \backslash \{x\}$. We say that x is an *isolated point* of X if $x \in X$ and there exists some $\varepsilon > 0$ such that $|x - y| > \varepsilon$ for all $y \in X \backslash \{x\}$.

Example 9.1.19 Let X be the set $X = (1, 2) \cup \{3\}$. Then 3 is an adherent point of X, but it is not a limit point of X, since 3 is not adherent to $X\backslash\{3\} = (1, 2)$; instead, 3 is an isolated point of X. On the other hand, 2 is still a limit point of X, since 2 is adherent to $X\backslash\{2\} = X$; but it is not isolated (why?).

Remark 9.1.20 From Lemma 9.1.14 we see that x is a limit point of X iff there exists a sequence $(a_n)_{n=0}^{\infty}$, consisting entirely of elements in X that are distinct from x, and such that $(a_n)_{n=0}^{\infty}$ converges to x. It turns out that the set of adherent points splits into the set of limit points and the set of isolated points (Exercise 9.1.9).

Lemma 9.1.21 *Let I be an interval (possibly infinite), i.e., I is a set of the form (a, b), $(a, b]$, $[a, b)$, $[a, b]$, $(a, +\infty)$, $[a, +\infty)$, $(-\infty, a)$, or $(-\infty, a]$, with $a < b$ in the first four cases. Then every element of I is a limit point of I.*

Proof We show this for the case $I = [a, b]$; the other cases are similar and are left to the reader. Let $x \in I$; we have to show that x is a limit point of I. There are three cases: $x = a$, $a < x < b$, and $x = b$. If $x = a$, then consider the sequence $(x + \frac{1}{n})_{n=N}^{\infty}$. This sequence converges to x and will lie inside $I\backslash\{a\} = (a, b]$ if N is chosen large enough (why?). Thus by Remark 9.1.20 we see that $x = a$ is a limit point of $[a, b]$. A similar argument works when $a < x < b$. When $x = b$ one has to use the sequence $(x - \frac{1}{n})_{n=N}^{\infty}$ instead (why?) but the argument is otherwise the same. □

Next, we define the concept of a bounded set.

Definition 9.1.22 *(Bounded sets)* A subset X of the real line is said to be *bounded* if we have $X \subseteq [-M, M]$ for some real number $M > 0$. A subset X of the real line is *unbounded* if it is not bounded.

Example 9.1.23 For any real numbers a, b, the interval $[a, b]$ is bounded, because it is contained inside $[-M, M]$, where $M := \max(|a|, |b|)$. However, the half-infinite interval $[0, +\infty)$ is unbounded (why?). In fact, no half-infinite interval or doubly infinite interval can be bounded. The sets **N**, **Z**, **Q**, and **R** are all unbounded (why?).

A basic property of closed and bounded sets is the following.

Theorem 9.1.24 (Heine–Borel theorem for the line) *Let X be a subset of* **R**. *Then the following two statements are equivalent:*

(a) *X is closed and bounded.*
(b) *Given any sequence $(a_n)_{n=0}^{\infty}$ of real numbers which takes values in X (i.e., $a_n \in X$ for all n), there exists a subsequence $(a_{n_j})_{j=0}^{\infty}$ of the original sequence, which converges to some number L in X.*

Proof See Exercise 9.1.13. □

Remark 9.1.25 This theorem shall play a key rôle in subsequent sections of this chapter. In the language of metric space topology, it asserts that every subset of the real line which is closed and bounded and is also compact; see Sect. 1.5. A more general version of this theorem, due to Eduard Heine (1821–1881) and Emile Borel (1871–1956), can be found in Theorem 1.5.7.

— Exercises —

Exercise 9.1.1 Prove Lemma 9.1.11.

Exercise 9.1.2 Prove Lemma 9.1.13. (*Hint:* for computing the closure of $\mathbf{Q}$, you will need Proposition 5.4.14.)

Exercise 9.1.3 Give an example of two subsets X, Y of the real line such that $\overline{X \cap Y} \neq \overline{X} \cap \overline{Y}$.

Exercise 9.1.4 Prove Lemma 9.1.14. (*Hint:* in order to prove one of the two implications here you will need axiom of choice, as in Lemma 8.4.5.)

Exercise 9.1.5 Let X be a subset of $\mathbf{R}$. Show that $\overline{X}$ is closed (i.e., $\overline{\overline{X}} = \overline{X}$). Furthermore, show that if Y is any closed set that contains X, then Y also contains $\overline{X}$. Thus the closure $\overline{X}$ of X is the smallest closed set which contains X.

Exercise 9.1.6 Let X be any subset of the real line, and let Y be a set such that $X \subseteq Y \subseteq \overline{X}$. Show that $\overline{Y} = \overline{X}$.

Exercise 9.1.7 Let $n \geq 1$ be a positive integer, and let $X_1, \ldots, X_n$ be closed subsets of $\mathbf{R}$. Show that $X_1 \cup X_2 \cup \ldots \cup X_n$ is also closed.

Exercise 9.1.8 Let I be a non-empty set (possibly infinite), and for each $\alpha \in I$ let X_α be a closed subset of $\mathbf{R}$. Show that the intersection $\bigcap_{\alpha \in I} X_\alpha$ (defined in (3.3)) is also closed.

Exercise 9.1.9 Let X be a subset of the real line. Show that every adherent point of X is either a limit point or an isolated point of X, but cannot be both. Conversely, show that every limit point and every isolated point of X is an adherent point of X.

Exercise 9.1.10 If X is a non-empty subset of $\mathbf{R}$, show that X is bounded if and only if $\inf(X)$ and $\sup(X)$ are finite.

Exercise 9.1.11 Show that if X is a bounded subset of $\mathbf{R}$, then the closure $\overline{X}$ is also bounded.

Exercise 9.1.12 Show that the union of any finite collection of bounded subsets of $\mathbf{R}$ is still a bounded set. Is this conclusion still true if one takes an infinite collection of bounded subsets of $\mathbf{R}$?

Exercise 9.1.13 Prove Theorem 9.1.24. (*Hint:* to show (a) implies (b), use the Bolzano–Weierstrass theorem (Theorem 6.6.8) and Corollary 9.1.17. To show (b) implies (a), argue by contradiction, using Corollary 9.1.17 to establish that X is closed. You will need the axiom of choice to show that X is bounded, as in Lemma 8.4.5.)

Exercise 9.1.14 Show that any finite subset of $\mathbf{R}$ is closed and bounded.

Exercise 9.1.15 Let E be a bounded non-empty subset of $\mathbf{R}$, and let $S := \sup(E)$ be the least upper bound of E. (Note from the least upper bound principle, Theorem 5.5.9, that S is a real number.) Show that S is an adherent point of E and is also an adherent point of $\mathbf{R} \backslash E$.

9.2 The Algebra of Real-Valued Functions

You are familiar with many functions $f: \mathbf{R} \to \mathbf{R}$ from the real line to the real line. Some examples are: $f(x) := x^2 + 3x + 5$; $f(x) := 2^x/(x^2+1)$; $f(x) := \sin(x)\exp(x)$ (we will define sin and exp formally in Chap.4). These are functions from $\mathbf{R}$ to $\mathbf{R}$ since to every real number x they assign a single real number $f(x)$. We can also consider more exotic functions, e.g.,

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q}. \end{cases}$$

This function is not *algebraic* (i.e., it cannot be expressed in terms of x purely by using the standard algebraic operations of $+$, $-$, $\times$, $/$, $\sqrt{}$, etc.; we will not need this notion in this text), but it is still a function from $\mathbf{R}$ to $\mathbf{R}$, because it still assigns a real number $f(x)$ to each $x \in \mathbf{R}$.

We can take any one of the previous functions $f: \mathbf{R} \to \mathbf{R}$ defined on all of $\mathbf{R}$, and *restrict* the domain to a smaller set $X \subseteq \mathbf{R}$, creating a new function, sometimes called $f|_X$, from X to $\mathbf{R}$. This is the same function as the original function f, but is only defined on a smaller domain. (Thus $f|_X(x) := f(x)$ when $x \in X$, and $f|_X(x)$ is undefined when $x \notin X$.) For instance, we can restrict the function $f(x) := x^2$, which is initially defined from $\mathbf{R}$ to $\mathbf{R}$, to the interval $[1, 2]$, thus creating a new function $f|_{[1,2]} : [1, 2] \to \mathbf{R}$, which is defined as $f|_{[1,2]}(x) = x^2$ when $x \in [1, 2]$ but is undefined elsewhere.

One could also restrict the codomain from $\mathbf{R}$ to some smaller subset Y of $\mathbf{R}$, provided of course that all the values of $f(x)$ lie inside Y. For instance, the function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2$ could also be thought of as a function from $\mathbf{R}$ to $[0, \infty)$, instead of a function from $\mathbf{R}$ to $\mathbf{R}$. Formally, these two functions are different functions, but the distinction between them is so minor that we shall often be careless about the range of a function in our discussion.

Strictly speaking, there is a distinction between a *function* f, and its *value* $f(x)$ at a point x. f is a function; but $f(x)$ is a number (which depends on some free variable x). This distinction is rather subtle and we will not stress it too much, but there are times when one has to distinguish between the two. For instance, if $f: \mathbf{R} \to \mathbf{R}$ is the function $f(x) := x^2$, and $g := f|_{[1,2]}$ is the restriction of f to the interval $[1, 2]$, then f and g both perform the operation of squaring, i.e., $f(x) = x^2$ and $g(x) = x^2$, but the two functions f and g are not considered the same function, $f \neq g$, because they have different domains. Despite this distinction, we shall often be careless, and say things like "consider the function $x^2 + 2x + 3$" when really we should be saying "consider the function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2 + 2x + 3$". (This distinction makes more of a difference when we start doing things like differentiation. For instance, if $f: \mathbf{R} \to \mathbf{R}$ is the function $f(x) = x^2$, then of course $f(3) = 9$, but the derivative of f at 3 is 6, whereas the derivative of 9 is of course 0, so we cannot simply "differentiate both sides" of $f(3) = 9$ and conclude that $6 = 0$.)

If X is a subset of $\mathbf{R}$, and $f\colon X \to \mathbf{R}$ is a function, we can form the *graph* $\{(x, f(x)) : x \in X\}$ of the function f; this is a subset of $X \times \mathbf{R}$, and hence a subset of the Euclidean plane $\mathbf{R}^2 = \mathbf{R} \times \mathbf{R}$. One can certainly study a function through its graph, by using the geometry of the plane $\mathbf{R}^2$ (e.g., employing such concepts as tangent lines, area, and so forth). We however will pursue a more "analytic" approach, in which we rely instead on the properties of the real numbers to analyze these functions. The two approaches are complementary; the geometric approach offers more visual intuition, while the analytic approach offers rigor and precision. Both the geometric intuition and the analytic formalism become useful when extending analysis of functions of one variable to functions of many variables (or possibly even infinitely many variables).

Just as numbers can be manipulated arithmetically, so can functions: the sum of two functions is a function, the product of two functions is a function, and so forth.

Definition 9.2.1 *(Arithmetic operations on functions)* Given two functions $f\colon X \to \mathbf{R}$ and $g : X \to \mathbf{R}$, we can define their sum $f + g : X \to \mathbf{R}$ by the formula

$$(f + g)(x) := f(x) + g(x),$$

their difference $f - g\colon X \to \mathbf{R}$ by the formula

$$(f - g)(x) := f(x) - g(x),$$

their maximum $\max(f, g) : X \to \mathbf{R}$ by

$$\max(f, g)(x) := \max(f(x), g(x)),$$

their minimum $\min(f, g) : X \to \mathbf{R}$ by

$$\min(f, g)(x) := \min(f(x), g(x)),$$

their product $fg\colon X \to \mathbf{R}$ (or $f \cdot g\colon X \to \mathbf{R}$) by the formula

$$(fg)(x) := f(x)g(x),$$

and (provided that $g(x) \neq 0$ for all $x \in X$) the quotient $f/g\colon X \to \mathbf{R}$ by the formula

$$(f/g)(x) := f(x)/g(x).$$

Finally, if c is a real number, we can define the function $cf\colon X \to \mathbf{R}$ (or $c \cdot f\colon X \to \mathbf{R}$) by the formula

$$(cf)(x) := c \times f(x).$$

Example 9.2.2 If $f\colon \mathbf{R} \to \mathbf{R}$ is the function $f(x) := x^2$, and $g\colon \mathbf{R} \to \mathbf{R}$ is the function $g(x) := 2x$, then $f + g\colon \mathbf{R} \to \mathbf{R}$ is the function $(f + g)(x) := x^2 + 2x$, while $fg\colon \mathbf{R} \to \mathbf{R}$ is the function $fg(x) = 2x^3$. Similarly $f - g\colon \mathbf{R} \to \mathbf{R}$ is the function $(f - g)(x) := x^2 - 2x$, while $6f\colon \mathbf{R} \to \mathbf{R}$ is the function $(6f)(x) = 6x^2$. Observe that fg is not the same function as $f \circ g$, which maps $x \mapsto 4x^2$, nor is it the same as $g \circ f$, which maps $x \mapsto 2x^2$ (why?). Thus multiplication of functions and composition of functions are two different operations.

— Exercises —

Exercise 9.2.1 Let $f\colon \mathbf{R} \to \mathbf{R}$, $g\colon \mathbf{R} \to \mathbf{R}$, $h\colon \mathbf{R} \to \mathbf{R}$. Which of the following identities are true, and which ones are false? In the former case, give a proof; in the latter case, give a counterexample.

$$(f + g) \circ h = (f \circ h) + (g \circ h)$$
$$f \circ (g + h) = (f \circ g) + (f \circ h)$$
$$(f + g) \cdot h = (f \cdot h) + (g \cdot h)$$
$$f \cdot (g + h) = (f \cdot g) + (f \cdot h)$$

9.3 Limiting Values of Functions

In Chap. 6 we defined what it means for a sequence $(a_n)_{n=0}^{\infty}$ to converge to a limit L. We now define a similar notion for what it means for a function f defined on the real line, or on some subset of the real line, to converge to some value at a point. Just as we used the notions of ε-closeness and eventual ε-closeness to deal with limits of sequences, we shall need a notion of ε-closeness and local ε-closeness to deal with limits of functions.

Definition 9.3.1 *(ε-closeness)* Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ be a function, let L be a real number, and let $\varepsilon > 0$ be a real number. We say that the function f is *ε-close to L* iff $f(x)$ is ε-close to L for every $x \in X$.

Example 9.3.2 When the function $f(x) := x^2$ is restricted to the interval $[1, 3]$, then it is 5-close to 4, since when $x \in [1, 3]$ then $1 \leq f(x) \leq 9$, and hence $|f(x) - 4| \leq 5$. When instead it is restricted to the smaller interval $[1.9, 2.1]$, then it is 0.41-close to 4, since if $x \in [1.9, 2.1]$, then $3.61 \leq f(x) \leq 4.41$, and hence $|f(x) - 4| \leq 0.41$.

Definition 9.3.3 *(Local ε-closeness)* Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ be a function, let L be a real number, x_0 be an adherent point of X, and $\varepsilon > 0$ be a real number. We say that f is *ε-close to L near x_0* iff there exists a $\delta > 0$ such that f becomes ε-close to L when restricted to the set $\{x \in X : |x - x_0| < \delta\}$.

Example 9.3.4 Let $f\colon [1, 3] \to \mathbf{R}$ be the function $f(x) := x^2$, restricted to the interval $[1, 3]$. This function is not 0.1-close to 4, since for instance $f(1)$ is not 0.1-close to 4. However, f is 0.1-close to 4 near 2, since when restricted to the set $\{x \in [1, 3] : |x - 2| < 0.01\}$, the function f is indeed 0.1-close to 4. This is because when $|x - 2| < 0.01$, we have $1.99 < x < 2.01$, and hence $3.9601 < f(x) < 4.0401$, and in particular $f(x)$ is 0.1-close to 4.

Example 9.3.5 Continuing with the same function f used in the previous example, we observe that f is not 0.1-close to 9, since for instance $f(1)$ is not 0.1-close to 9. However, f is 0.1-close to 9 near 3, since when restricted to the set $\{x \in [1, 3] : |x - 3| < 0.01\}$—which is the same as the half-open interval $(2.99, 3]$ (why?), the function f becomes 0.1-close to 9 (since if $2.99 < x \leq 3$, then $8.9401 < f(x) \leq 9$, and hence $f(x)$ is 0.1-close to 9).

Definition 9.3.6 *(Convergence of functions at a point)* Let X be a subset of $\mathbf{R}$, let $f: X \to \mathbf{R}$ be a function, let E be a subset of X, x_0 be an adherent point of E, and let L be a real number. We say that f *converges to* L *at* x_0 *in* E and write $\lim_{x \to x_0; x \in E} f(x) = L$, iff f, after restricting to E, is ε-close to L near x_0 for every $\varepsilon > 0$. If f does not converge to any number L at x_0, we say that f *diverges* at x_0, and leave $\lim_{x \to x_0; x \in E} f(x)$ undefined.

In other words, we have $\lim_{x \to x_0; x \in E} f(x) = L$ iff for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - L| \leq \varepsilon$ for all $x \in E$ such that $|x - x_0| < \delta$. (Why is this definition equivalent to the one given above?)

Remark 9.3.7 In many cases we will omit the set E from the above notation (i.e., we will just say that f converges to L at x_0, or that $\lim_{x \to x_0} f(x) = L$), although this is slightly dangerous. For instance, it sometimes makes a difference whether E actually contains x_0 or not. To give an example, if $f: \mathbf{R} \to \mathbf{R}$ is the function defined by setting $f(x) = 1$ when $x = 0$ and $f(x) = 0$ when $x \neq 0$, then one has $\lim_{x \to 0; x \in \mathbf{R} \setminus \{0\}} f(x) = 0$, but $\lim_{x \to 0; x \in \mathbf{R}} f(x)$ is undefined. Some authors only define the limit $\lim_{x \to x_0; x \in E} f(x)$ when E does not contain x_0 (so that x_0 is now a limit point of E rather than an adherent point), or would use $\lim_{x \to x_0; x \in E} f(x)$ to denote what we would call $\lim_{x \to x_0; x \in E \setminus \{x_0\}} f(x)$, but we have chosen a slightly more general notation, which allows the possibility that E contains x_0.

Example 9.3.8 Let $f: [1, 3] \to \mathbf{R}$ be the function $f(x) := x^2$. We have seen before that f is 0.1-close to 4 near 2. A similar argument shows that f is 0.01-close to 4 near 2 (one just has to pick a smaller value of δ).

Definition 9.3.6 is rather unwieldy. However, we can rewrite this definition in terms of a more familiar one, involving limits of sequences.

Proposition 9.3.9 *Let X be a subset of* $\mathbf{R}$, *let* $f: X \to \mathbf{R}$ *be a function, let E be a subset of X, let x_0 be an adherent point of E, and let L be a real number. Then the following two statements are logically equivalent:*

(a) *f converges to L at x_0 in E.*
(b) *For every sequence $(a_n)_{n=0}^{\infty}$ which consists entirely of elements of E and converges to x_0, the sequence $(f(a_n))_{n=0}^{\infty}$ converges to L.*

Proof See Exercise 9.3.1. □

In view of the above proposition, we will sometimes write "$f(x) \to L$ as $x \to x_0$ in E" or "f has a limit L at x_0 in E" instead of "f converges to L at x_0", or "$\lim_{x \to x_0} f(x) = L$".

Remark 9.3.10 With the notation of Proposition 9.3.9, we have the following corollary: if $\lim_{x \to x_0; x \in E} f(x) = L$, and $\lim_{n \to \infty} a_n = x_0$, then $\lim_{n \to \infty} f(a_n) = L$.

Remark 9.3.11 We only consider limits of a function f at x_0 in the case when x_0 is an adherent point of E. When x_0 is not an adherent point then it is not worth it to define the concept of a limit. (Can you see why there will be problems?)

Remark 9.3.12 The variable x used to denote a limit is a dummy variable; we could replace it by any other variable and obtain exactly the same limit. For instance, if $\lim_{x \to x_0; x \in E} f(x) = L$, then $\lim_{y \to x_0; y \in E} f(y) = L$, and conversely (why?).

Proposition 9.3.9 has some immediate corollaries. For instance, we now know that a function can have at most one limit at each point:

Corollary 9.3.13 *Let X be a subset of* $\mathbf{R}$*, let E be a subset of X, let x_0 be an adherent point of E, and let $f: X \to \mathbf{R}$ be a function. Then f can have at most one limit at x_0 in E.*

Proof Suppose for sake of contradiction that there are two distinct numbers L and L' such that f has a limit L at x_0 in E, and such that f also has a limit L' at x_0 in E. Since x_0 is an adherent point of E, we know by Lemma 9.1.14 that there is a sequence $(a_n)_{n=0}^\infty$ consisting of elements in E which converges to x_0. Since f has a limit L at x_0 in E, we thus see by Proposition 9.3.9, that $(f(a_n))_{n=0}^\infty$ converges to L. But since f also has a limit L' at x_0 in E, we see that $(f(a_n))_{n=0}^\infty$ also converges to L'. But this contradicts the uniqueness of limits of sequences (Proposition 6.1.7). □

Using the limit laws for sequences, one can now deduce the limit laws for functions:

Proposition 9.3.14 (Limit laws for functions) *Let X be a subset of* $\mathbf{R}$*, let E be a subset of X, let x_0 be an adherent point of E, and let $f: X \to \mathbf{R}$ and $g: X \to \mathbf{R}$ be functions. Suppose that f has a limit L at x_0 in E, and g has a limit M at x_0 in E. Then $f + g$ has a limit $L + M$ at x_0 in E, $f - g$ has a limit $L - M$ at x_0 in E,* $\max(f, g)$ *has a limit* $\max(L, M)$ *at x_0 in E,* $\min(f, g)$ *has a limit* $\min(L, M)$ *at x_0 in E and fg has a limit LM at x_0 in E. If c is a real number, then cf has a limit cL at x_0 in E. Finally, if g is non-zero on E (i.e., $g(x) \neq 0$ for all $x \in E$) and M is non-zero, then f/g has a limit L/M at x_0 in E.*

Proof We just prove the first claim (that $f + g$ has a limit $L + M$); the others are very similar and are left to Exercise 9.3.2. Let $(a_n)_{n=0}^\infty$ be an arbitrary sequence of elements in E that converges to x_0. Since f has a limit L at x_0 in E, we thus see from Proposition 9.3.9 that $(f(a_n))_{n=0}^\infty$ converges to L. Similarly $(g(a_n))_{n=0}^\infty$ converges to M. By the limit laws for sequences (Theorem 6.1.19) we conclude that $((f + g)(a_n))_{n=0}^\infty$ converges to $L + M$. By Proposition 9.3.9 again, this implies that $f + g$ has a limit $L + M$ at x_0 in E as desired (since $(a_n)_{n=0}^\infty$ was an arbitrary sequence in E converging to x_0). □

Remark 9.3.15 One can phrase Proposition 9.3.14 more informally as saying that

$$\lim_{x\to x_0}(f\pm g)(x) = \lim_{x\to x_0} f(x) \pm \lim_{x\to x_0} g(x)$$

$$\lim_{x\to x_0}\max(f,g)(x) = \max\left(\lim_{x\to x_0} f(x), \lim_{x\to x_0} g(x)\right)$$

$$\lim_{x\to x_0}\min(f,g)(x) = \min\left(\lim_{x\to x_0} f(x), \lim_{x\to x_0} g(x)\right)$$

$$\lim_{x\to x_0}(fg)(x) = \lim_{x\to x_0} f(x) \lim_{x\to x_0} g(x)$$

$$\lim_{x\to x_0}(f/g)(x) = \frac{\lim_{x\to x_0} f(x)}{\lim_{x\to x_0} g(x)}$$

(where we have dropped the restriction $x \in E$ for brevity) but bear in mind that these identities are only true when the right-hand side makes sense, and furthermore for the final identity we need g to be non-zero, and also $\lim_{x\to x_0} g(x)$ to be non-zero. (See Example 1.2.4 for some examples of what goes wrong when limits are manipulated carelessly.)

Using the limit laws in Proposition 9.3.14 we can already deduce several limits. First of all, it is easy to check the basic limits

$$\lim_{x\to x_0;x\in\mathbf{R}} c = c$$

and

$$\lim_{x\to x_0;x\in\mathbf{R}} x = x_0$$

for any real numbers x_0 and c. (Why? Use Proposition 9.3.9.) By the limit laws we can thus conclude that

$$\lim_{x\to x_0;x\in\mathbf{R}} x^2 = x_0^2$$

$$\lim_{x\to x_0;x\in\mathbf{R}} cx = cx_0$$

$$\lim_{x\to x_0;x\in\mathbf{R}} x^2 + cx + d = x_0^2 + cx_0 + d$$

etc., where c, d are arbitrary real numbers.

If f converges to L at x_0 in X, and Y is any subset of X such that x_0 is still an adherent point of Y, then f will also converge to L at x_0 in Y (why?). Thus convergence on a large set implies convergence on a smaller set. The converse, however, is not true:

Example 9.3.16 Consider the *signum function* sgn: $\mathbf{R} \to \mathbf{R}$, defined by

$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Then $\lim_{x\to 0; x\in(0,\infty)} \operatorname{sgn}(x) = 1$ (why?), whereas $\lim_{x\to 0; x\in(-\infty,0)} = -1$ (why?) and $\lim_{x\to 0; x\in\mathbf{R}} \operatorname{sgn}(x)$ is undefined (why?). Thus it is sometimes dangerous to drop the set E from the notation of limit. However, in many cases it is safe to do so; for instance, since we know that $\lim_{x\to x_0; x\in\mathbf{R}} x^2 = x_0^2$, we know in fact that $\lim_{x\to x_0; x\in X} x^2 = x_0^2$ for any set X with x_0 as an adherent point (why?). Thus it is safe to write $\lim_{x\to x_0} x^2 = x_0^2$.

Example 9.3.17 Let $f(x)$ be the function

$$f(x) := \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0. \end{cases}$$

Then $\lim_{x\to 0; x\in\mathbf{R}\setminus\{0\}} f(x) = 0$ (why?), but $\lim_{x\to 0; x\in\mathbf{R}} f(x)$ is undefined (why?). (When this happens, we say that f has a "removable singularity" or "removable discontinuity" at 0. Because of such singularities, it is sometimes the convention when writing $\lim_{x\to x_0} f(x)$ to automatically exclude x_0 from the set; for instance, in some textbooks, $\lim_{x\to x_0} f(x)$ is used as shorthand for $\lim_{x\to x_0; x\in X\setminus\{x_0\}} f(x)$.)

On the other hand, the limit at x_0 should only depend on the values of the function near x_0; the values away from x_0 are not relevant. The following proposition reflects this intuition:

Proposition 9.3.18 (Limits are local) *Let X be a subset of* $\mathbf{R}$*, let E be a subset of X, let x_0 be an adherent point of E, let $f: X \to \mathbf{R}$ be a function, and let L be a real number. Let $\delta > 0$. Then we have*

$$\lim_{x\to x_0; x\in E} f(x) = L$$

if and only if

$$\lim_{x\to x_0; x\in E\cap(x_0-\delta, x_0+\delta)} f(x) = L.$$

Proof See Exercise 9.3.3. □

Informally, the above proposition asserts that

$$\lim_{x\to x_0; x\in E} f(x) = \lim_{x\to x_0; x\in E\cap(x_0-\delta, x_0+\delta)} f(x).$$

Thus the limit of a function at x_0, if it exists, only depends on the values of f near x_0; the values far away do not actually influence the limit.

We now give a few more examples of limits.

Example 9.3.19 Consider the functions $f\colon \mathbf{R} \to \mathbf{R}$ and $g\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x + 2$ and $g(x) := x + 1$. Then $\lim_{x\to 2; x\in\mathbf{R}} f(x) = 4$ and $\lim_{x\to 2; x\in\mathbf{R}} g(x) = 3$. We would like to use the limit laws to conclude that $\lim_{x\to 2; x\in\mathbf{R}} f(x)/g(x) = 4/3$, or in other words that $\lim_{x\to 2; x\in\mathbf{R}} \frac{x+2}{x+1} = \frac{4}{3}$. Strictly speaking, we cannot use Proposition 9.3.14 to ensure this, because $x + 1$ is zero at $x = -1$, and so $f(x)/g(x)$ is not defined. However, this is easily solved, by restricting the domain of f and g from $\mathbf{R}$ to a smaller domain, such as $\mathbf{R}\backslash\{-1\}$. Then Proposition 9.3.14 does apply, and we have $\lim_{x\to 2; x\in\mathbf{R}\backslash\{-1\}} \frac{x+2}{x+1} = \frac{4}{3}$.

Example 9.3.20 Consider the function $f\colon \mathbf{R}\backslash\{1\} \to \mathbf{R}$ defined by $f(x) := (x^2 - 1)/(x - 1)$. This function is well-defined for every real number except 1, so $f(1)$ is undefined. However, 1 is still an adherent point of $\mathbf{R}\backslash\{1\}$ (why?), and the limit $\lim_{x\to 1; x\in\mathbf{R}-\{1\}} f(x)$ is still defined. This is because on the domain $\mathbf{R}\backslash\{1\}$ we have the identity $(x^2 - 1)/(x - 1) = (x + 1)(x - 1)/(x - 1) = x + 1$, and $\lim_{x\to 1; x\in\mathbf{R}-\{1\}} x + 1 = 2$.

Example 9.3.21 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q}. \end{cases}$$

We will show that $f(x)$ has no limit at 0 in $\mathbf{R}$. Suppose for sake of contradiction that $f(x)$ had some limit L at 0 in $\mathbf{R}$. Then we would have $\lim_{n\to\infty} f(a_n) = L$ whenever $(a_n)_{n=1}^{\infty}$ is a sequence of non-zero numbers converging to 0. Since $(1/n)_{n=1}^{\infty}$ is such a sequence, we would have

$$L = \lim_{n\to\infty} f(1/n) = \lim_{n\to\infty} 1 = 1.$$

On the other hand, since $(\sqrt{2}/n)_{n=1}^{\infty}$ is another sequence of non-zero numbers converging to 0—but now these numbers are irrational instead of rational—we have

$$L = \lim_{n\to\infty} f(\sqrt{2}/n) = \lim_{n\to\infty} 0 = 0.$$

Since $1 \neq 0$, we have a contradiction. Thus this function does not have a limit at 0.

— Exercises —

Exercise 9.3.1 Prove Proposition 9.3.9.

Exercise 9.3.2 Prove the remaining claims in Proposition 9.3.14.

Exercise 9.3.3 Prove Proposition 9.3.18.

Exercise 9.3.4 Propose a definition for limit superior $\limsup_{x\to x_0; x\in E} f(x)$ and limit inferior $\liminf_{x\to x_0; x\in E} f(x)$, and then propose an analogue of Proposition 9.3.9 for your definition. (For an additional challenge: prove that analogue.)

Exercise 9.3.5 (Continuous version of squeeze test) Let X be a subset of $\mathbf{R}$, let E be a subset of X, let x_0 be an adherent point of E, and let $f\colon X \to \mathbf{R}$, $g\colon X \to \mathbf{R}$, $h\colon X \to \mathbf{R}$ be functions such that $f(x) \leq g(x) \leq h(x)$ for all $x \in E$. If we have $\lim_{x \to x_0; x \in E} f(x) = \lim_{x \to x_0; x \in E} h(x) = L$ for some real number L, show that $\lim_{x \to x_0; x \in E} g(x) = L$.

9.4 Continuous Functions

We now introduce one of the most fundamental notions in the theory of functions - that of *continuity*.

Definition 9.4.1 *(Continuity)* Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a function. Let x_0 be an element of X. We say that f is *continuous at* x_0 iff we have

$$\lim_{x \to x_0; x \in X} f(x) = f(x_0);$$

in other words, the limit of $f(x)$ as x converges to x_0 in X exists and is equal to $f(x_0)$. We say that f is *continuous on* X (or simply *continuous*) iff f is continuous at x_0 for every $x_0 \in X$. We say that f is *discontinuous at* x_0 iff it is not continuous at x_0.

We also extend these notions to functions $f\colon X \to Y$ that take values in a subset Y of $\mathbf{R}$, by identifying such functions (by abuse of notation) with the function $\tilde{f}\colon X \to \mathbf{R}$ that agrees everywhere with f (so $\tilde{f}(x) = f(x)$ for all $x \in X$) but where the codomain has been enlarged from Y to $\mathbf{R}$.

Example 9.4.2 Let c be a real number, and let $f\colon \mathbf{R} \to \mathbf{R}$ be the constant function $f(x) := c$. Then for every real number $x_0 \in \mathbf{R}$, we have

$$\lim_{x \to x_0; x \in \mathbf{R}} f(x) = \lim_{x \to x_0; x \in \mathbf{R}} c = c = f(x_0),$$

thus f is continuous at every point $x_0 \in \mathbf{R}$, or in other words f is continuous on $\mathbf{R}$.

Example 9.4.3 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the identity function $f(x) := x$. Then for every real number $x_0 \in \mathbf{R}$, we have

$$\lim_{x \to x_0; x \in \mathbf{R}} f(x) = \lim_{x_0 \in x; x \in \mathbf{R}} x = x_0 = f(x_0),$$

thus f is continuous at every point $x_0 \in \mathbf{R}$, or in other words f is continuous on $\mathbf{R}$.

Example 9.4.4 Let $\operatorname{sgn}\colon \mathbf{R} \to \mathbf{R}$ be the signum function defined in Example 9.3.16. Then $\operatorname{sgn}(x)$ is continuous at every non-zero value of x; for instance, at 1, we have (using Proposition 9.3.18)

$$\begin{aligned}\lim_{x\to 1; x\in\mathbf{R}} \operatorname{sgn}(x) &= \lim_{x\to 1; x\in(0.9,1.1)} \operatorname{sgn}(x)\\ &= \lim_{x\to 1; x\in(0.9,1.1)} 1\\ &= 1\\ &= \operatorname{sgn}(1).\end{aligned}$$

On the other hand, sgn is not continuous at 0, since the limit $\lim_{x\to 0; x\in\mathbf{R}} \operatorname{sgn}(x)$ does not exist.

Example 9.4.5 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q}. \end{cases}$$

Then by the discussion in the previous section, f is not continuous at 0. In fact, it turns out that f is not continuous at any real number x_0 (can you see why?).

Example 9.4.6 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function

$$f(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Then f is continuous at every non-zero real number (why?), but is not continuous at 0. However, if we restrict f to the right-hand line $[0, \infty)$, then the resulting function $f|_{[0,\infty)}$ now becomes continuous everywhere in its domain, including 0. Thus restricting the domain of a function can make a discontinuous function continuous again.

There are several ways to phrase the statement that "f is continuous at x_0":

Proposition 9.4.7 (Equivalent formulations of continuity) *Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ be a function, and let x_0 be an element of X. Then the following four statements are logically equivalent:*

(a) *f is continuous at x_0.*
(b) *For every sequence $(a_n)_{n=0}^{\infty}$ consisting of elements of X with $\lim_{n\to\infty} a_n = x_0$, we have $\lim_{n\to\infty} f(a_n) = f(x_0)$.*
(c) *For every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(x_0)| < \varepsilon$ for all $x \in X$ with $|x - x_0| < \delta$.*
(d) *For every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f(x) - f(x_0)| \leq \varepsilon$ for all $x \in X$ with $|x - x_0| \leq \delta$.*

Proof See Exercise 9.4.1. □

Remark 9.4.8 A particularly useful consequence of Proposition 9.4.7 is the following: if f is continuous at x_0, and $a_n \to x_0$ as $n \to \infty$, then $f(a_n) \to f(x_0)$ as $n \to \infty$ (provided that all the elements of the sequence $(a_n)_{n=0}^{\infty}$ lie in the domain of f, of course). Thus continuous functions are very useful in computing limits.

The limit laws in Proposition 9.3.14, combined with the definition of continuity in Definition 9.4.1, immediately imply

Proposition 9.4.9 (Arithmetic preserves continuity) *Let X be a subset of* $\mathbf{R}$, *and let $f: X \to \mathbf{R}$ and $g: X \to \mathbf{R}$ be functions. Let $x_0 \in X$. Then if f and g are both continuous at x_0, then the functions $f+g$, $f-g$,* $\max(f,g)$, $\min(f,g)$ *and fg are also continuous at x_0. If g is non-zero on X, then f/g is also continuous at x_0.*

In particular, the sum, difference, maximum, minimum, and product of continuous functions are continuous; and the quotient of two continuous functions is continuous as long as the denominator does not become zero.

One can use Proposition 9.4.9 to show that a lot of functions are continuous. For instance, just by starting from the fact that constant functions are continuous, and the identity function $f(x) = x$ is continuous (Exercise 9.4.2), one can show that the function $g(x) := \max(x^3 + 4x^2 + x + 5, x^4 - x^3)/(x^2 - 4)$, for instance, is continuous at every point of $\mathbf{R}$ except the two points $x = +2$, $x = -2$ where the denominator vanishes.

Some other examples of continuous functions are given below.

Proposition 9.4.10 (Exponentiation is continuous, I) *Let $a > 0$ be a positive real number. Then the function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := a^x$ is continuous.*

Proof See Exercise 9.4.3. □

Proposition 9.4.11 (Exponentiation is continuous, II) *Let p be a real number. Then the function $f: (0, \infty) \to \mathbf{R}$ defined by $f(x) := x^p$ is continuous.*

Proof See Exercise 9.4.4. □

There is a stronger statement than Propositions 9.4.10 and 9.4.11, namely that exponentiation is *jointly continuous* in both the exponent and the base, but this is harder to show; see Exercise 4.5.10.

Proposition 9.4.12 (Absolute value is continuous) *The function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := |x|$ is continuous.*

Proof This follows since $|x| = \max(x, -x)$ and the functions $x, -x$ are already continuous. □

The class of continuous functions is not only closed under addition, subtraction, multiplication, and division, but is also closed under composition:

Proposition 9.4.13 (Composition preserves continuity) *Let X and Y be subsets of* $\mathbf{R}$, *and let $f: X \to Y$ and $g: Y \to \mathbf{R}$ be functions. Let x_0 be a point in X. If f is continuous at x_0, and g is continuous at $f(x_0)$, then the composition $g \circ f: X \to \mathbf{R}$ is continuous at x_0.*

Proof See Exercise 9.4.5. □

Example 9.4.14 Since the function $f(x) := 3x + 1$ is continuous on all of **R**, and the function $g(x) := 5^x$ is continuous on all of **R**, the function $g \circ f(x) = 5^{3x+1}$ is continuous on all of **R**. By several applications of the above propositions, one can show that far more complicated functions, e.g., $h(x) := |x^2 - 8x + 7|^{\sqrt{2}}/(x^2 + 1)$, are also continuous. (Why is this function continuous?) There are still a few functions though that are not yet easy to test for continuity, such as $k(x) := x^x$; this function can be dealt with more easily once we have the machinery of logarithms, which we will see in Sect. 4.5 of *Analysis II*.

— Exercises —

Exercise 9.4.1 Prove Proposition 9.4.7. (*Hint:* this can largely be done by applying the previous propositions and lemmas. Note that to prove (a),(b), and (c) are equivalent, you do not have to prove all six implications, but you do have to prove at least three; for instance, showing that (a) implies (b), (b) implies (c), and (c) implies (a) will suffice, although this is not necessarily the shortest or simplest way to do this question.)

Exercise 9.4.2 Let X be a subset of **R**, and let $c \in \mathbf{R}$. Show that the constant function $f: X \to \mathbf{R}$ defined by $f(x) := c$ is continuous, and show that the identity function $g: X \to \mathbf{R}$ defined by $g(x) := x$ is also continuous.

Exercise 9.4.3 Prove Proposition 9.4.10. (*Hint:* you can use Lemma 6.5.3, combined with the squeeze test (Corollary 6.4.14) and Proposition 6.7.3.)

Exercise 9.4.4 Prove Proposition 9.4.11. (*Hint:* from limit laws (Proposition 9.3.14) one can show that $\lim_{x\to 1} x^n = 1$ for all integers n. From this and the squeeze test (Corollary 6.4.14) deduce that $\lim_{x\to 1} x^p = 1$ for all real numbers p. Finally, apply Proposition 6.7.3.)

Exercise 9.4.5 Prove Proposition 9.4.13.

Exercise 9.4.6 Let X be a subset of **R**, and let $f: X \to \mathbf{R}$ be a continuous function. If Y is a subset of X, show that the restriction $f|_Y : Y \to \mathbf{R}$ of f to Y is also a continuous function. (*Hint:* this is a simple result, but it requires you to follow the definitions carefully.)

Exercise 9.4.7 Let $n \geq 0$ be an integer, and for each $0 \leq i \leq n$ let c_i be a real number. Let P : $\mathbf{R} \to \mathbf{R}$ be the function

$$P(x) := \sum_{i=0}^{n} c_i x^i;$$

such a function is known as a *polynomial of one variable*; a typical example is $P(x) = 6x^4 - 3x^2 + 4$. Show that P is continuous.

9.5 Left and Right Limits

We now introduce the notion of left and right limits, which can be thought of as two seperate "halves" of the complete limit $\lim_{x\to x_0; x\in X} f(x)$.

Definition 9.5.1 *(Left and right limits)* Let X be a subset of **R**, $f: X \to \mathbf{R}$ be a function, and let x_0 be a real number. If x_0 is an adherent point of $X \cap (x_0, \infty)$, then we define the *right limit* $f(x_0+)$ of f at x_0 by the formula

$$f(x_0+) := \lim_{x \to x_0; x \in X \cap (x_0, \infty)} f(x),$$

provided of course that this limit exists. Similarly, if x_0 is an adherent point of $X \cap (-\infty, x_0)$, then we define the *left limit* $f(x_0-)$ of f at x_0 by the formula

$$f(x_0-) := \lim_{x \to x_0; x \in X \cap (-\infty, x_0)} f(x),$$

again provided that the limit exists. (Thus in many cases $f(x_0+)$ and $f(x_0-)$ will not be defined.)

Sometimes we use the shorthand notations

$$\lim_{x \to x_0+} f(x) := \lim_{x \to x_0; x \in X \cap (x_0, \infty)} f(x);$$
$$\lim_{x \to x_0-} f(x) := \lim_{x \to x_0; x \in X \cap (-\infty, x_0)} f(x)$$

when the domain X of f is clear from context.

Example 9.5.2 Consider the signum function $\mathrm{sgn}\colon \mathbf{R} \to \mathbf{R}$ defined in Example 9.3.16. We have

$$\mathrm{sgn}(0+) = \lim_{x \to 0; x \in \mathbf{R} \cap (0, \infty)} \mathrm{sgn}(x) = \lim_{x \to 0; x \in \mathbf{R} \cap (0, \infty)} 1 = 1$$

and

$$\mathrm{sgn}(0-) = \lim_{x \to 0; x \in \mathbf{R} \cap (-\infty, 0)} \mathrm{sgn}(x) = \lim_{x \to 0; x \in \mathbf{R} \cap (-\infty, 0)} -1 = -1,$$

while $\mathrm{sgn}(0) = 0$ by definition.

Note that f does not necessarily have to be defined at x_0 in order for $f(x_0+)$ or $f(x_0-)$ to be defined. For instance, if $f\colon \mathbf{R} \backslash \{0\} \to \mathbf{R}$ is the function $f(x) := x/|x|$, then $f(0+) = 1$ and $f(0-) = -1$ (why?), even though $f(0)$ is undefined.

From Proposition 9.3.9 we see that if the right limit $f(x_0+)$ exists, and $(a_n)_{n=0}^{\infty}$ is a sequence in X converging to x_0 from the right (i.e., $a_n > x_0$ for all $n \in \mathbf{N}$), then $\lim_{n\to\infty} f(a_n) = f(x_0+)$. Similarly, if $(b_n)_{n=0}^{\infty}$ is a sequence converging to x_0 from the left (i.e., $b_n < x_0$ for all $n \in \mathbf{N}$), then $\lim_{n\to\infty} f(b_n) = f(x_0-)$.

Let x_0 be an adherent point of both $X \cap (x_0, \infty)$ and $X \cap (-\infty, x_0)$. If f is continuous at x_0, it is clear from Proposition 9.4.7 that $f(x_0+)$ and $f(x_0-)$ both exist and are equal to $f(x_0)$. (Can you see why?) A converse is also true (compare this with Proposition 6.4.12f):

Proposition 9.5.3 *Let X be a subset of $\mathbf{R}$ containing a real number x_0, and suppose that x_0 is an adherent point of both $X \cap (x_0, \infty)$ and $X \cap (-\infty, x_0)$. Let $f\colon X \to \mathbf{R}$ be a function. If $f(x_0+)$ and $f(x_0-)$ both exist and are both equal to $f(x_0)$, then f is continuous at x_0.*

Proof Let us write $L := f(x_0)$. Then by hypothesis we have

$$\lim_{x \to x_0; x \in X \cap (x_0, \infty)} f(x) = L \tag{9.1}$$

and

$$\lim_{x \to x_0; x \in X \cap (-\infty, x_0)} f(x) = L. \tag{9.2}$$

Let $\varepsilon > 0$ be given. From (9.1), Definition 9.3.6, and Definition 9.3.3 (applied to the restriction of f to $X \cap (x_0, +\infty)$), we know that there exists a $\delta_+ > 0$ such that $|f(x) - L| < \varepsilon$ for all $x \in X \cap (x_0, \infty)$ for which $|x - x_0| < \delta_+$. From (9.2) we similarly know that there exists a $\delta_- > 0$ such that $|f(x) - L| < \varepsilon$ for all $x \in X \cap (-\infty, x_0)$ for which $|x - x_0| < \delta_-$. Now let $\delta := \min(\delta_-, \delta_+)$; then $\delta > 0$ (why?), and suppose that $x \in X$ is such that $|x - x_0| < \delta$. Then there are three cases: $x > x_0$, $x = x_0$, and $x < x_0$, but in all three cases we know that $|f(x) - L| < \varepsilon$. (Why? The reason is different in each of the three cases.) By Proposition 9.4.7 we thus have that f is continuous at x_0, as desired. □

As we saw with the signum function in Example 9.3.16, it is possible for the left and right limits $f(x_0-)$, $f(x_0+)$ of a function f at a point x_0 to both exist, but not be equal to each other; when this happens, we say that f has a *jump discontinuity* at x_0. Thus, for instance, the signum function has a jump discontinuity at zero. Also, it is possible for the left and right limits $f(x_0-)$, $f(x_0+)$ to exist and be equal each other, but not be equal to $f(x_0)$; when this happens we say that f has a *removable discontinuity* (or *removable singularity*) at x_0. For instance, if we take $f: \mathbf{R} \to \mathbf{R}$ to be the function

$$f(x) := \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0, \end{cases}$$

then $f(0+)$ and $f(0-)$ both exist and equal 0 (why?), but $f(0)$ equals 1; thus f has a removable discontinuity at 0.

Remark 9.5.4 Jump discontinuities and removable discontinuities are not the only way a function can be discontinuous. Another way is for a function to go to infinity at the discontinuity: for instance, the function $f: \mathbf{R}\backslash\{0\} \to \mathbf{R}$ defined by $f(x) := 1/x$ has a discontinuity at 0 which is neither a jump discontinuity or a removable singularity; informally, $f(x)$ converges to $+\infty$ when x approaches 0 from the right and converges to $-\infty$ when x approaches 0 from the left. These types of singularities are sometimes known as *asymptotic discontinuities*. There are also *oscillatory discontinuities*, where the function remains bounded but still does not have a limit near x_0. For instance, the function $f: \mathbf{R} \to \mathbf{R}$ defined by

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q} \end{cases}$$

has an oscillatory discontinuity at 0 (and in fact at any other real number also). This is because the function does not have left or right limits at 0, despite the fact that the function is bounded.

The study of discontinuities (also called *singularities*) continues further, but is beyond the scope of this text. For instance, singularities play a key rôle in complex analysis.

— Exercises —

Exercise 9.5.1 Let E be a subset of $\mathbf{R}$, let $f: E \to \mathbf{R}$ be a function, and let x_0 be an adherent point of E. Write down a definition of what it would mean for the limit $\lim_{x\to x_0; x\in E} f(x)$ to exist and equal $+\infty$ or $-\infty$. If $f: \mathbf{R}\backslash\{0\} \to \mathbf{R}$ is the function $f(x) := 1/x$, use your definition to conclude $f(0+) = +\infty$ and $f(0-) = -\infty$. Also, state and prove some analogue of Proposition 9.3.9 when $L = +\infty$ or $L = -\infty$.

9.6 The Maximum Principle

In the previous two sections we saw that a large number of functions were continuous, though certainly not all functions were continuous. We now show that continuous functions enjoy a number of other useful properties, especially if their domain is a closed interval. It is here that we shall begin exploiting the full power of the Heine–Borel theorem (Theorem 9.1.24).

Definition 9.6.1 Let X be a subset of $\mathbf{R}$, and let $f: X \to \mathbf{R}$ be a function. We say that f is *bounded from above* iff there exists a real number M such that $f(x) \leq M$ for all $x \in X$. We say that f is *bounded from below* iff there exists a real number M such that $f(x) \geq -M$ for all $x \in X$. We say that f is *bounded* iff there exists a real number M such that $|f(x)| \leq M$ for all $x \in X$.

Remark 9.6.2 A function is bounded if and only if it is bounded both from above and below. (Why? Note that one part of the "if and only if" is slightly trickier than the other.) Also, a function $f: X \to \mathbf{R}$ is bounded if and only if its image $f(X)$ is a bounded set in the sense of Definition 9.1.22 (why?).

Not all continuous functions are bounded. For instance, the function $f(x) := x$ on the domain $\mathbf{R}$ is continuous but unbounded (why?), although it is bounded on some smaller domains, such as $[1, 2]$. The function $f(x) := 1/x$ is continuous but unbounded on $(0, 1)$ (why?), though it is continuous and bounded on $[1, 2]$ (why?). However, if the domain of the continuous function is a closed and bounded interval, then we do have boundedness:

Lemma 9.6.3 *Let $a < b$ be real numbers, and let $f: [a, b] \to \mathbf{R}$ be a function continuous on $[a, b]$. Then f is a bounded function.*

Proof Suppose for sake of contradiction that f is not bounded. Thus for every real number M there exists an element $x \in [a, b]$ such that $|f(x)| \geq M$.

In particular, for every natural number n, the set $\{x \in [a, b] : |f(x)| \geq n\}$ is non-empty. We can thus choose[2] a sequence $(x_n)_{n=0}^{\infty}$ in $[a, b]$ such that $|f(x_n)| \geq n$ for all n. This sequence lies in $[a, b]$, and so by Theorem 9.1.24 there exists a subsequence $(x_{n_j})_{j=0}^{\infty}$ which converges to some limit $L \in [a, b]$, where $n_0 < n_1 < n_2 < \ldots$ is an increasing sequence of natural numbers. In particular, we see that $n_j \geq j$ for all $j \in \mathbf{N}$ (why? Use induction).

Since f is continuous on $[a, b]$, it is continuous at L, and in particular we see that

$$\lim_{j \to \infty} f(x_{n_j}) = f(L).$$

Thus the sequence $(f(x_{n_j}))_{j=0}^{\infty}$ is convergent, and hence it is bounded. On the other hand, we know from the construction that $|f(x_{n_j})| \geq n_j \geq j$ for all j, and hence the sequence $(f(x_{n_j}))_{j=0}^{\infty}$ is not bounded, a contradiction. □

Remark 9.6.4 There are two things about this proof that are worth noting. Firstly, it shows how useful the Heine–Borel theorem (Theorem 9.1.24) is. Secondly, it is an indirect proof; it doesn't say *how* to find the bound for f, but it shows that having f unbounded leads to a contradiction.

We now improve Lemma 9.6.3 to say something more.

Definition 9.6.5 *(Maxima and minima)* Let X be a set, let $f : X \to \mathbf{R}$ be a function, and let $x_0 \in X$. We say that f *attains its maximum at* x_0 if we have $f(x_0) \geq f(x)$ for all $x \in X$ (i.e., the value of f at the point x_0 is larger than or equal to the value of f at any other point in X). We say that f *attains its minimum at* x_0 if we have $f(x_0) \leq f(x)$.

Remark 9.6.6 If a function attains its maximum somewhere, then it must be bounded from above (why?). Similarly if it attains its minimum somewhere, then it must be bounded from below. These notions of maxima and minima are *global*; local versions will be defined in Definition 10.2.1.

Proposition 9.6.7 (Maximum principle) *Let $a < b$ be real numbers, and let $f : [a, b] \to \mathbf{R}$ be a function continuous on $[a, b]$. Then f attains its maximum at some point $x_{max} \in [a, b]$ and also attains its minimum at some point $x_{min} \in [a, b]$.*

Remark 9.6.8 Strictly speaking, "maximum principle" is a misnomer, since the principle also concerns the minimum. Perhaps a more precise name would have been "extremum principle"; the word "extremum" is used to denote either a maximum or a minimum.

Proof We shall just show that f attains its maximum somewhere; the proof that it attains its minimum also is similar but is left to the reader.

[2] Strictly speaking, this requires the axiom of choice, as in Lemma 8.4.5. However, one can also proceed without the axiom of choice, by defining $x_n := \sup\{x \in [a, b] : |f(x)| \geq n\}$, and using the continuity of f to show that $|f(x_n)| \geq n$. We leave the details to the reader.

From Lemma 9.6.3 we know that f is bounded, thus there exists an M such that $-M \leq f(x) \leq M$ for each $x \in [a, b]$. Now let E denote the set

$$E := \{f(x) : x \in [a, b]\}.$$

(In other words, $E := f([a, b])$.) By what we just said, this set is a subset of $[-M, M]$. It is also non-empty, since it contains for instance the point $f(a)$. Hence by the least upper bound principle, it has a supremum $\sup(E)$ which is a real number.

Write $m := \sup(E)$. By definition of supremum, we know that $y \leq m$ for all $y \in E$; by definition of E, this means that $f(x) \leq m$ for all $x \in [a, b]$. Thus to show that f attains its maximum somewhere, it will suffice to find an $x_{max} \in [a, b]$ such that $f(x_{max}) = m$. (Why will this suffice?)

Let $n \geq 1$ be any integer. Then $m - \frac{1}{n} < m = \sup(E)$. As $\sup(E)$ is the least upper bound for E, $m - \frac{1}{n}$ cannot be an upper bound for E, thus there exists a $y \in E$ such that $m - \frac{1}{n} < y$. By definition of E, this implies that there exists an $x \in [a, b]$ such that $m - \frac{1}{n} < f(x)$.

We now choose a sequence $(x_n)_{n=1}^{\infty}$ by choosing, for each n, x_n to be an element of $[a, b]$ such that $m - \frac{1}{n} < f(x_n)$. (Again, this requires the axiom of choice; however it is possible to prove this principle without the axiom of choice. For instance, you will see a better proof of this proposition using the notion of *compactness* in Proposition 2.3.2.) This is a sequence in $[a, b]$; by the Heine–Borel theorem (Theorem 9.1.24), we can thus find a subsequence $(x_{n_j})_{j=1}^{\infty}$, where $n_1 < n_2 < \ldots$, which converges to some limit $x_{max} \in [a, b]$. Since $(x_{n_j})_{j=1}^{\infty}$ converges to x_{max}, and f is continuous at x_{max}, we have as before that

$$\lim_{j \to \infty} f(x_{n_j}) = f(x_{max}).$$

On the other hand, by construction we know that

$$f(x_{n_j}) > m - \frac{1}{n_j} \geq m - \frac{1}{j},$$

and so by taking limits of both sides we see that

$$f(x_{max}) = \lim_{j \to \infty} f(x_{n_j}) \geq \lim_{j \to \infty} m - \frac{1}{j} = m.$$

On the other hand, we know that $f(x) \leq m$ for all $x \in [a, b]$, so in particular $f(x_{max}) \leq m$. Combining these two inequalities we see that $f(x_{max}) = m$ as desired. □

Note that the maximum principle does not prevent a function from attaining its maximum or minimum at more than one point. For instance, the function $f(x) := x^2$ on the interval $[-2, 2]$ attains its maximum at two different points, at -2 and at 2.

Let us write $\sup_{x\in[a,b]} f(x)$ as shorthand for $\sup\{f(x) : x \in [a,b]\}$, and similarly define $\inf_{x\in[a,b]} f(x)$. The maximum principle thus asserts that $m := \sup_{x\in[a,b]} f(x)$ is a real number and is the *maximum value* of f on $[a,b]$; i.e., there is at least one point x_{max} in $[a,b]$ for which $f(x_{max}) = m$, and for every other $x \in [a,b]$, $f(x)$ is less than or equal to m. Similarly $\inf_{x\in[a,b]} f(x)$ is the minimum value of f on $[a,b]$.

We now know that on a closed interval, every continuous function is bounded and attains its maximum at least once and minimum at least once. The same is not true for open or infinite intervals; see Exercise 9.6.1.

Remark 9.6.9 You may encounter a rather different "maximum principle" in complex analysis or partial differential equations, involving analytic functions and harmonic functions, respectively, instead of continuous functions. Those maximum principles are not directly related to this one (though they are also concerned with whether maxima exist, and where the maxima are located).

— Exercises —

Exercise 9.6.1 Give examples of

(a) a function $f: (1,2) \to \mathbf{R}$ which is continuous and bounded, attains its minimum somewhere, but does not attain its maximum anywhere;
(b) a function $f: [0,\infty) \to \mathbf{R}$ which is continuous, bounded, attains its maximum somewhere, but does not attain its minimum anywhere;
(c) a function $f: [-1,1] \to \mathbf{R}$ which is bounded but does not attain its minimum anywhere or its maximum anywhere.
(d) a function $f: [-1,1] \to \mathbf{R}$ which has no upper bound and no lower bound.

Explain why none of the examples you construct violate the maximum principle. (*Note:* read the assumptions of that principle *carefully*!)

Exercise 9.6.2 If $f, g: X \to \mathbf{R}$ are bounded functions, show that $f+g$, $f-g$, and $f\cdot g$ are also bounded functions. If we furthermore assume that $g(x) \neq 0$ for all $x \in X$, is it true that f/g is bounded? Prove this or give a counterexample.

9.7 The Intermediate Value Theorem

We have just shown that a continuous function attains both its maximum value and its minimum value. We now show that f also attains every value in between. To do this, we first prove a very intuitive theorem:

Theorem 9.7.1 (Intermediate value theorem) *Let $a < b$, and let $f: [a,b] \to \mathbf{R}$ be a continuous function on $[a,b]$. Let y be a real number between $f(a)$ and $f(b)$, i.e., either $f(a) \le y \le f(b)$ or $f(a) \ge y \ge f(b)$. Then there exists $c \in [a,b]$ such that $f(c) = y$.*

Proof We have two cases: $f(a) \le y \le f(b)$ or $f(a) \ge y \ge f(b)$. We will assume the former, that $f(a) \le y \le f(b)$; the latter is proven similarly and is left to the reader.

If $y = f(a)$ or $y = f(b)$, then the claim is easy, as one can simply set $c = a$ or $c = b$, so we will assume that $f(a) < y < f(b)$. Let E denote the set

$$E := \{x \in [a, b] : f(x) < y\}.$$

Clearly E is a subset of $[a, b]$ and is hence bounded. Also, since $f(a) < y$, we see that a is an element of E, so E is non-empty. By the least upper bound principle, the supremum

$$c := \sup(E)$$

is thus finite. Since E is bounded by b, we know that $c \leq b$; since E contains a, we know that $c \geq a$. Thus we have $c \in [a, b]$. To complete the proof we now show that $f(c) = y$. The idea is to work from the left of c to show that $f(c) \leq y$ and to work from the right of c to show that $f(c) \geq y$.

Let $n \geq 1$ be an integer. The number $c - \frac{1}{n}$ is less than $c = \sup(E)$ and hence cannot be an upper bound for E. Thus there exists a point, call it x_n, which lies in E and which is greater than $c - \frac{1}{n}$. Also $x_n \leq c$ since c is an upper bound for E. Thus

$$c - \frac{1}{n} \leq x_n \leq c.$$

By the squeeze test (Corollary 6.4.14) we thus have $\lim_{n\to\infty} x_n = c$. Since f is continuous at c, this implies that $\lim_{n\to\infty} f(x_n) = f(c)$. But since x_n lies in E for every n, we have $f(x_n) < y$ for every n. By the comparison principle (Lemma 6.4.13) we thus have $f(c) \leq y$. Since $f(b) > f(c)$, we conclude $c \neq b$.

Since $c \neq b$ and $c \in [a, b]$, we must have $c < b$. In particular there is an $N > 0$ such that $c + \frac{1}{n} < b$ for all $n > N$ (since $c + \frac{1}{n}$ converges to c as $n \to \infty$). Since c is the supremum of E and $c + \frac{1}{n} > c$, we thus have $c + \frac{1}{n} \notin E$ for all $n > N$. Since $c + \frac{1}{n} \in [a, b]$, we thus have $f(c + \frac{1}{n}) \geq y$ for all $n \geq N$. But $c + \frac{1}{n}$ converges to c, and f is continuous at c, thus $f(c) \geq y$. But we already knew that $f(c) \leq y$, thus $f(c) = y$, as desired. □

The intermediate value theorem says that if f takes the values $f(a)$ and $f(b)$, then it must also take all the values in between. Note that if f is not assumed to be continuous, then the intermediate value theorem no longer applies. For instance, if $f : [-1, 1] \to \mathbf{R}$ is the function

$$f(x) := \begin{cases} -1 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

then $f(-1) = -1$, and $f(1) = 1$, but there is no $c \in [-1, 1]$ for which $f(c) = 0$. Thus if a function is discontinuous, it can "jump" past intermediate values; however continuous functions cannot do so.

Remark 9.7.2 A continuous function may take an intermediate value multiple times. For instance, if $f : [-2, 2] \to \mathbf{R}$ is the function $f(x) := x^3 - x$, then $f(-2) = -6$

and $f(2) = 6$, so we know that there exists a $c \in [-2, 2]$ for which $f(c) = 0$. In fact, in this case there exists three such values of c: we have $f(-1) = f(0) = f(1) = 0$.

Remark 9.7.3 The intermediate value theorem gives another way to show that one can take n^{th} roots of a number. For instance, to construct the square root of 2, consider the function $f\colon [0, 2] \to \mathbf{R}$ defined by $f(x) = x^2$. This function is continuous, with $f(0) = 0$ and $f(2) = 4$. Thus there exists a $c \in [0, 2]$ such that $f(c) = 2$, i.e., $c^2 = 2$. (This argument does not show that there is just one square root of 2, but it does prove that there is *at least* one square root of 2.)

Corollary 9.7.4 (Images of continuous functions) *Let $a < b$, and let $f\colon [a, b] \to \mathbf{R}$ be a continuous function on $[a, b]$. Let $M := \sup_{x\in[a,b]} f(x)$ be the maximum value of f, and let $m := \inf_{x\in[a,b]} f(x)$ be the minimum value. Let y be a real number between m and M (i.e., $m \le y \le M$). Then there exists a $c \in [a, b]$ such that $f(c) = y$. Furthermore, we have $f([a, b]) = [m, M]$.*

Proof See Exercise 9.7.1. □

— Exercises —

Exercise 9.7.1 Prove Corollary 9.7.4. (*Hint:* you may need Exercise 9.4.6 in addition to the intermediate value theorem.)

Exercise 9.7.2 Let $f\colon [0, 1] \to [0, 1]$ be a continuous function. Show that there exists a real number x in $[0, 1]$ such that $f(x) = x$. (*Hint:* apply the intermediate value theorem to the function $f(x) - x$.) This point x is known as a *fixed point* of f, and this result is a basic example of a *fixed point theorem*, which play an important rôle in certain types of analysis.

9.8 Monotonic Functions

We now discuss a class of functions which is distinct from the class of continuous functions, but has somewhat similar properties: the class of monotone (or monotonic) functions.

Definition 9.8.1 *(Monotonic functions)* Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a function. We say that f is *monotone increasing* iff $f(y) \ge f(x)$ whenever $x, y \in X$ and $y > x$. We say that f is *strictly monotone increasing* iff $f(y) > f(x)$ whenever $x, y \in X$ and $y > x$. Similarly, we say f is *monotone decreasing* iff $f(y) \le f(x)$ whenever $x, y \in X$ and $y > x$, and *strictly monotone decreasing* iff $f(y) < f(x)$ whenever $x, y \in X$ and $y > x$. We say that f is *monotone* if it is monotone increasing or monotone decreasing, and *strictly monotone* if it is strictly monotone increasing or strictly monotone decreasing.

Examples 9.8.2 The function $f(x) := x^2$, when restricted to the domain $[0, \infty)$, is strictly monotone increasing (why?), but when restricted instead to the domain

$(-\infty, 0]$, is strictly monotone decreasing (why?). Thus the function is strictly monotone on both $(-\infty, 0]$ and $[0, \infty)$, but is not strictly monotone (or monotone) on the full real line $(-\infty, \infty)$. Note that if a function is strictly monotone on a domain X, it is automatically monotone as well on the same domain X. The constant function $f(x) := 6$, when restricted to an arbitrary domain $X \subseteq \mathbf{R}$, is both monotone increasing and monotone decreasing, but is not strictly monotone (unless X consists of at most one point - why?).

Continuous functions are not necessarily monotone (consider for instance the function $f(x) = x^2$ on $\mathbf{R}$), and monotone functions are not necessarily continuous; for instance, consider the function $f\colon [-1, 1] \to \mathbf{R}$ defined earlier by

$$f(x) := \begin{cases} -1 & \text{if } x \le 0 \\ 1 & \text{if } x > 0. \end{cases}$$

Monotone functions obey the maximum principle (Exercise 9.8.1), but not the intermediate value principle (Exercise 9.8.2). On the other hand, it is possible for a monotone function to have many, many discontinuities (Exercise 9.8.5).

If a function is both strictly monotone and continuous, then it has many nice properties. In particular, it is invertible:

Proposition 9.8.3 *Let $a < b$ be real numbers, and let $f\colon [a, b] \to \mathbf{R}$ be a function which is both continuous and strictly monotone increasing. Then f is a bijection from $[a, b]$ to $[f(a), f(b)]$, and the inverse $f^{-1} : [f(a), f(b)] \to [a, b]$ is also continuous and strictly monotone increasing.*

Proof See Exercise 9.8.4. □

There is a similar Proposition for functions which are strictly monotone decreasing; see Exercise 9.8.4.

Example 9.8.4 Let n be a positive integer and $R > 0$. Since the function $f(x) := x^n$ is strictly increasing on the interval $[0, R]$, we see from Proposition 9.8.3 that this function is a bijection from $[0, R]$ to $[0, R^n]$, and hence there is an inverse from $[0, R^n]$ to $[0, R]$. This can be used to give an alternate means to construct the n^{th} root $x^{1/n}$ of a number $x \in [0, R]$ than what was done in Lemma 5.6.5.

— Exercises —

Exercise 9.8.1 Explain why the maximum principle remains true if the hypothesis that f is continuous is replaced with f being monotone, or with f being strictly monotone. (You can use the same explanation for both cases.)

Exercise 9.8.2 Give an example to show that the intermediate value theorem becomes false if the hypothesis that f is continuous is replaced with f being monotone, or with f being strictly monotone. (You can use the same counterexample for both cases.)

Exercise 9.8.3 Let $a < b$ be real numbers, and let $f: [a, b] \to \mathbf{R}$ be a function which is both continuous and one-to-one. Show that f is strictly monotone. (*Hint:* divide into the three cases $f(a) < f(b)$, $f(a) = f(b)$, $f(a) > f(b)$. The second case leads directly to a contradiction. In the first case, use contradiction and the intermediate value theorem to show that f is strictly monotone increasing; in the third case, argue similarly to show f is strictly monotone decreasing.)

Exercise 9.8.4 Prove Proposition 9.8.3. (*Hint:* to show that f^{-1} is continuous, it is easiest to use the "epsilon-delta" definition of continuity, Proposition 9.4.7c.) Is the proposition still true if the continuity assumption is dropped, or if strict monotonicity is replaced just by monotonicity? How should one modify the proposition to deal with strictly monotone decreasing functions instead of strictly monotone increasing functions?

Exercise 9.8.5 In this exercise we give an example of a function which has a discontinuity at every rational point, but is continuous at every irrational. Since the rationals are countable, we can write them as $\mathbf{Q} = \{q(0), q(1), q(2), \ldots\}$, where $q : \mathbf{N} \to \mathbf{Q}$ is a bijection from $\mathbf{N}$ to $\mathbf{Q}$. Now define a function $g: \mathbf{Q} \to \mathbf{R}$ by setting $g(q(n)) := 2^{-n}$ for each natural number n; thus g maps $q(0)$ to 1, $q(1)$ to 2^{-1}, etc. Since $\sum_{n=0}^{\infty} 2^{-n}$ is absolutely convergent, we see that $\sum_{r \in \mathbf{Q}} g(r)$ is also absolutely convergent. Now define the function $f: \mathbf{R} \to \mathbf{R}$ by

$$f(x) := \sum_{r \in \mathbf{Q}: r < x} g(r).$$

Since $\sum_{r \in \mathbf{Q}} g(r)$ is absolutely convergent, we know that $f(x)$ is well-defined for every real number x.

(a) Show that f is strictly monotone increasing. (*Hint:* you will need Proposition 5.4.14.)
(b) Show that for every rational number r, f is discontinuous at r. (*Hint:* since r is rational, $r = q(n)$ for some natural number n. Show that $f(x) \geq f(r) + 2^{-n}$ for all $x > r$.)
(c) Show that for every irrational number x, f is continuous at x. (*Hint:* first demonstrate that the functions

$$f_n(x) := \sum_{r \in \mathbf{Q}: r < x, g(r) \geq 2^{-n}} g(r)$$

are continuous at x, and that $|f(x) - f_n(x)| \leq 2^{-n}$.)

9.9 Uniform Continuity

We know that a continuous function on a closed interval $[a, b]$ remains bounded (and in fact attains its maximum and minimum, by the maximum principle). However, if we replace the closed interval by an open interval, then continuous functions need not be bounded any more. An example is the function $f: (0, 2) \to \mathbf{R}$ defined by $f(x) := 1/x$. This function is continuous at every point in $(0, 2)$ and is hence continuous at $(0, 2)$, but is not bounded. Informally speaking, the problem here is that while the function is indeed continuous at every point in the open interval $(0, 2)$, it becomes "less and less" continuous as one approaches the endpoint 0.

Let us analyze this phenomenon further, using the "epsilon-delta" definition of continuity—Proposition 9.4.7c. We know that if $f: X \to \mathbf{R}$ is continuous at a point x_0, then for every $\varepsilon > 0$ there exists a δ such that $f(x)$ will be ε-close to $f(x_0)$

whenever $x \in X$ is δ-close to x_0. In other words, we can force $f(x)$ to be ε-close to $f(x_0)$ if we ensure that x is sufficiently close to x_0. One way of thinking about this is that around every point x_0 there is an "island of stability" $(x_0 - \delta, x_0 + \delta)$, where the function $f(x)$ doesn't stray by more than ε from $f(x_0)$.

Example 9.9.1 Take the function $f(x) := 1/x$ mentioned above at the point $x_0 = 1$. In order to ensure that $f(x)$ is 0.1-close to $f(x_0)$, it suffices to take x to be 1/11-close to x_0, since if x is 1/11-close to x_0 then $10/11 < x < 12/11$, and so $11/12 < f(x) < 11/10$, and so $f(x)$ is 0.1-close to $f(x_0)$. Thus the "δ" one needs to make $f(x)$ 0.1-close to $f(x_0)$ is about 1/11 or so, at the point $x_0 = 1$.

Now let us look instead at the point $x_0 = 0.1$. The function $f(x) = 1/x$ is still continuous here, but we shall see the continuity is much worse. In order to ensure that $f(x)$ is 0.1-close to $f(x_0)$, we need x to be 1/1010-close to x_0. Indeed, if x is 1/1010 close to x_0, then $10/101 < x < 102/1010$, and so $9.901 < f(x) < 10.1$, so $f(x)$ is 0.1-close to $f(x_0)$. Thus one needs a much smaller "δ" for the same value of ε, i.e., $f(x)$ is much more "unstable" near 0.1 than it is near 1, in the sense that there is a much smaller "island of stability" around 0.1 as there is around 1 (if one is interested in keeping $f(x)$ 0.1-stable).

On the other hand, there are other continuous functions which do not exhibit this behavior. Consider the function $g: (0, 2) \to \mathbf{R}$ defined by $g(x) := 2x$. Let us fix $\varepsilon = 0.1$ as before and investigate the island of stability around $x_0 = 1$. It is clear that if x is 0.05-close to x_0, then $g(x)$ is 0.1-close to $g(x_0)$; in this case we can take δ to be 0.05 at $x_0 = 1$. And if we move x_0 around, say if we set x_0 to 0.1 instead, the δ does not change—even when x_0 is set to 0.1 instead of 1, we see that $g(x)$ will stay 0.1-close to $g(x_0)$ whenever x is 0.05-close to x_0. Indeed, the same δ works for every x_0. When this happens, we say that the function g is *uniformly continuous*. More precisely:

Definition 9.9.2 [Uniform continuity] Let X be a subset of $\mathbf{R}$, and let $f: X \to \mathbf{R}$ be a function. We say that f is *uniformly continuous* if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $f(x)$ and $f(x_0)$ are ε-close whenever $x, x_0 \in X$ are two points in X which are δ-close.

Remark 9.9.3 This definition should be compared with the notion of continuity. From Proposition 9.4.7c, we know that a function f is *continuous* if for every $\varepsilon > 0$, and every $x_0 \in X$, there is a $\delta > 0$ such that $f(x)$ and $f(x_0)$ are ε-close whenever $x \in X$ is δ-close to x_0. The difference between uniform continuity and continuity is that in uniform continuity one can take a single δ which works for all $x_0 \in X$; for ordinary continuity, each $x_0 \in X$ might use a different δ. Thus every uniformly continuous function is continuous, but not conversely.

Example 9.9.4 (Informal) The function $f: (0, 2) \to \mathbf{R}$ defined by $f(x) := 1/x$ is continuous on $(0, 2)$, but not uniformly continuous, because the continuity (or more precisely, the dependence of δ on ε) becomes worse and worse as $x \to 0$. (We will make this more precise in Example 9.9.10.)

Recall that the notions of adherent point and of continuous function had several equivalent formulations, both had "epsilon-delta" type formulations (involving the notion of ε-closeness), and both had "sequential" formulations (involving the convergence of sequences); see Lemma 9.1.14 and Proposition 9.3.9. The concept of uniform continuity can similarly be phrased in a sequential formulation, this time using the concept of *equivalent sequences* (cf. Definition 5.2.6, but we now generalize to sequences of real numbers instead of rationals, and no longer require the sequences to be Cauchy):

Definition 9.9.5 *(Equivalent sequences)* Let m be an integer, let $(a_n)_{n=m}^{\infty}$ and $(b_n)_{n=m}^{\infty}$ be two sequences of real numbers, and let $\varepsilon > 0$ be given. We say that $(a_n)_{n=m}^{\infty}$ is *ε-close* to $(b_n)_{n=m}^{\infty}$ iff a_n is ε-close to b_n for each $n \geq m$. We say that $(a_n)_{n=m}^{\infty}$ is *eventually ε-close* to $(b_n)_{n=m}^{\infty}$ iff there exists an $N \geq m$ such that the sequences $(a_n)_{n=N}^{\infty}$ and $(b_n)_{n=N}^{\infty}$ are ε-close. Two sequences $(a_n)_{n=m}^{\infty}$ and $(b_n)_{n=m}^{\infty}$ are *equivalent* iff for each $\varepsilon > 0$, the sequences $(a_n)_{n=m}^{\infty}$ and $(b_n)_{n=m}^{\infty}$ are eventually ε-close.

Remark 9.9.6 One could debate whether ε should be assumed to be rational or real, but a minor modification of Proposition 6.1.4 shows that this does not make any difference to the above definitions.

The notion of equivalence can be phrased more succinctly using our language of limits:

Lemma 9.9.7 *Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be sequences of real numbers (not necessarily bounded or convergent). Then $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are equivalent if and only if* $\lim_{n\to\infty}(a_n - b_n) = 0$.

Proof See Exercise 9.9.1. □

Meanwhile, the notion of uniform continuity can be phrased using equivalent sequences:

Proposition 9.9.8 *Let X be a subset of* $\mathbf{R}$, *and let $f : X \to$* $\mathbf{R}$ *be a function. Then the following two statements are logically equivalent:*

(a) f is uniformly continuous on X.
(b) Whenever $(x_n)_{n=0}^{\infty}$ and $(y_n)_{n=0}^{\infty}$ are two equivalent sequences consisting of elements of X, the sequences $(f(x_n))_{n=0}^{\infty}$ and $(f(y_n))_{n=0}^{\infty}$ are also equivalent.

Proof See Exercise 9.9.2. □

Remark 9.9.9 The reader should compare this with Proposition 9.3.9. Proposition 9.3.9 asserted that if f was continuous, then f maps convergent sequences to convergent sequences. In contrast, Proposition 9.9.8 asserts that if f is *uniformly* continuous, then f maps *equivalent* pairs of sequences to equivalent pairs of sequences. To see how the two Propositions are connected, observe from Lemma 9.9.7 that $(x_n)_{n=0}^{\infty}$ will converge to x_* if and only if the sequences $(x_n)_{n=0}^{\infty}$ and $(x_*)_{n=0}^{\infty}$ are equivalent.

Example 9.9.10 Consider the function $f\colon (0,2) \to \mathbf{R}$ defined by $f(x) := 1/x$ considered earlier. From Lemma 9.9.7 we see that the sequence $(1/n)_{n=1}^{\infty}$ and $(1/2n)_{n=1}^{\infty}$ are equivalent sequences in $(0,2)$. However, the sequences $(f(1/n))_{n=1}^{\infty}$ and $(f(1/2n))_{n=1}^{\infty}$ are not equivalent (why? Use Lemma 9.9.7 again). So by Proposition 9.9.8, f is not uniformly continuous. (These sequences start at 1 instead of 0, but the reader can easily see that this makes no difference to the above discussion.)

Example 9.9.11 Consider the function $f\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2$. This is a continuous function on $\mathbf{R}$, but it turns out not to be uniformly continuous; in some sense the continuity gets "worse and worse" as one approaches infinity. One way to quantify this is via Proposition 9.9.8. Consider the sequences $(n)_{n=1}^{\infty}$ and $(n+\frac{1}{n})_{n=1}^{\infty}$. By Lemma 9.9.7, these sequences are equivalent. But the sequences $(f(n))_{n=1}^{\infty}$ and $(f(n+\frac{1}{n}))_{n=1}^{\infty}$ are not equivalent, since $f(n+\frac{1}{n}) = n^2+2+\frac{1}{n^2} = f(n)+2+\frac{1}{n^2}$ does not become eventually 2-close to $f(n)$. By Proposition 9.9.8 we can thus conclude that f is not uniformly continuous.

Another property of uniformly continuous functions is that they map Cauchy sequences to Cauchy sequences.

Proposition 9.9.12 *Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a uniformly continuous function. Let $(x_n)_{n=0}^{\infty}$ be a Cauchy sequence consisting entirely of elements in X. Then $(f(x_n))_{n=0}^{\infty}$ is also a Cauchy sequence.*

Proof See Exercise 9.9.3. □

Example 9.9.13 Once again, we demonstrate that the function $f\colon (0,2) \to \mathbf{R}$ defined by $f(x) := 1/x$ is not uniformly continuous. The sequence $(1/n)_{n=1}^{\infty}$ is a Cauchy sequence in $(0,2)$, but the sequence $(f(1/n))_{n=1}^{\infty}$ is not a Cauchy sequence (why?). Thus by Proposition 9.9.12, f is not uniformly continuous.

Corollary 9.9.14 *Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ be a uniformly continuous function, and let x_0 be an adherent point of X. Then the limit $\lim_{x\to x_0; x\in X} f(x)$ exists (in particular, it is a real number).*

Proof See Exercise 9.9.4. □

We now show that a uniformly continuous function will map bounded sets to bounded sets.

Proposition 9.9.15 *Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a uniformly continuous function. Suppose that E is a bounded subset of X. Then $f(E)$ is also bounded.*

Proof See Exercise 9.9.5. □

As we have just seen repeatedly, not all continuous functions are uniformly continuous. However, if the domain of the function is a closed interval, then continuous functions are in fact uniformly continuous:

Theorem 9.9.16 *Let $a < b$ be real numbers, and let $f : [a, b] \to \mathbf{R}$ be a function which is continuous on $[a, b]$. Then f is also uniformly continuous.*

Proof Suppose for sake of contradiction that f is not uniformly continuous. By Proposition 9.9.8, there must therefore exist two equivalent sequences $(x_n)_{n=0}^{\infty}$ and $(y_n)_{n=0}^{\infty}$ in $[a, b]$ such that the sequences $(f(x_n))_{n=0}^{\infty}$ and $(f(y_n))_{n=0}^{\infty}$ are not equivalent. In particular, we can find an $\varepsilon > 0$ such that $(f(x_n))_{n=0}^{\infty}$ and $(f(y_n))_{n=0}^{\infty}$ are not eventually ε-close.

Fix this value of ε, and let E be the set

$$E := \{n \in \mathbf{N} : f(x_n) \text{ and } f(y_n) \text{ are not } \varepsilon\text{-close}\}.$$

We must have E infinite, since if E were finite then $(f(x_n))_{n=0}^{\infty}$ and $(f(y_n))_{n=0}^{\infty}$ would be eventually ε-close (why?). By Proposition 8.1.5, E is countable; in fact from the proof of that proposition we see that we can find an infinite sequence

$$n_0 < n_1 < n_2 < \ldots$$

consisting entirely of elements in E. In particular, we have

$$|f(x_{n_j}) - f(y_{n_j})| > \varepsilon \text{ for all } j \in \mathbf{N}. \tag{9.3}$$

On the other hand, the sequence $(x_{n_j})_{j=0}^{\infty}$ is a sequence in $[a, b]$, and so by the Heine–Borel theorem (Theorem 9.1.24) there must be a subsequence $(x_{n_{j_k}})_{k=0}^{\infty}$ which converges to some limit L in $[a, b]$. In particular, f is continuous at L, and so by Proposition 9.4.7,

$$\lim_{k\to\infty} f(x_{n_{j_k}}) = f(L). \tag{9.4}$$

Note that $(x_{n_{j_k}})_{k=0}^{\infty}$ is a subsequence of $(x_n)_{n=0}^{\infty}$, and $(y_{n_{j_k}})_{k=0}^{\infty}$ is a subsequence of $(y_n)_{n=0}^{\infty}$, by Lemma 6.6.4. On the other hand, from Lemma 9.9.7 we have

$$\lim_{n\to\infty} (x_n - y_n) = 0.$$

By Proposition 6.6.5, we thus have

$$\lim_{k\to\infty} (x_{n_{j_k}} - y_{n_{j_k}}) = 0.$$

Since $x_{n_{j_k}}$ converges to L as $k \to \infty$, we thus have by limit laws

$$\lim_{k\to\infty} y_{n_{j_k}} = L$$

and hence by continuity of f at L

$$\lim_{k\to\infty} f(y_{n_{j_k}}) = f(L).$$

Subtracting this from (9.4) using limit laws, we obtain

$$\lim_{k\to\infty}(f(x_{n_{j_k}}) - f(y_{n_{j_k}})) = 0.$$

But this contradicts (9.3) (why?). From this contradiction we conclude that f is in fact uniformly continuous. □

Remark 9.9.17 One should compare Lemma 9.6.3, Proposition 9.9.15, and Theorem 9.9.16 with each other. Note in particular that the lemma follows from combining the proposition with the theorem.

— Exercises —

Exercise 9.9.1 Prove Lemma 9.9.7.

Exercise 9.9.2 Prove Proposition 9.9.8. (*Hint:* you should avoid Lemma 9.9.7, and instead go back to the definition of equivalent sequences in Definition 9.9.5.)

Exercise 9.9.3 Prove Proposition 9.9.12. (*Hint:* use Definition 9.9.2 directly.)

Exercise 9.9.4 Use Proposition 9.9.12 to prove Corollary 9.9.14. Use this corollary to give an alternate demonstration of the results in Example 9.9.10.

Exercise 9.9.5 Prove Proposition 9.9.15. (*Hint:* mimic the proof of Lemma 9.6.3. At some point you will need either Proposition 9.9.12 or Corollary 9.9.14.)

Exercise 9.9.6 Let X, Y, Z be subsets of $\mathbf{R}$. Let $f: X \to Y$ be a function which is uniformly continuous on X, and let $g: Y \to Z$ be a function which is uniformly continuous on Y. Show that the function $g \circ f: X \to Z$ is uniformly continuous on X.

9.10 Limits at Infinity

Until now, we have discussed what it means for a function $f: X \to \mathbf{R}$ to have a limit as $x \to x_0$ as long as x_0 is a *real* number. We now briefly discuss what it would mean to take limits when x_0 is equal to $+\infty$ or $-\infty$. (This is part of a more general theory of continuous functions on a topological space; see Sect. 11.12.)

First, we need a notion of what it means for $+\infty$ or $-\infty$ to be adherent to a set.

Definition 9.10.1 *(Infinite adherent points)* Let X be a subset of $\mathbf{R}$. We say that $+\infty$ is *adherent* to X iff for every $M \in \mathbf{R}$ there exists an $x \in X$ such that $x > M$; we say that $-\infty$ is *adherent* to X iff for every $M \in \mathbf{R}$ there exists an $x \in X$ such that $x < M$.

In other words, $+\infty$ is adherent to X iff X has no upper bound, or equivalently iff $\sup(X) = +\infty$. Similarly $-\infty$ is adherent to X iff X has no lower bound, or iff $\inf(X) = -\infty$. Thus a set is bounded if and only if $+\infty$ and $-\infty$ are not adherent points.

Remark 9.10.2 This definition may seem rather different from Definition 9.1.8, but can be unified using the topological structure of the extended real line $\mathbf{R}^*$, which we will not discuss here.

Definition 9.10.3 *(Limits at infinity)* Let X be a subset of $\mathbf{R}$ with $+\infty$ as an adherent point, and let $f\colon X \to \mathbf{R}$ be a function. We say that $f(x)$ *converges to* L as $x \to +\infty$ in X, and write $\lim_{x\to+\infty;x\in X} f(x) = L$, iff for every $\varepsilon > 0$ there exists an M such that f is ε-close to L on $X \cap (M, +\infty)$ (i.e., $|f(x) - L| \le \varepsilon$ for all $x \in X$ such that $x > M$). Similarly we say that $f(x)$ *converges to* L as $x \to -\infty$ iff for every $\varepsilon > 0$ there exists an M such that f is ε-close to L on $X \cap (-\infty, M)$.

Example 9.10.4 Let $f\colon (0,\infty) \to \mathbf{R}$ be the function $f(x) := 1/x$. Then we have $\lim_{x\to+\infty;x\in(0,\infty)} 1/x = 0$. (Can you see why, from the definition?)

One can do many of the same things with these limits at infinity as we have been doing with limits at other points x_0; for instance, it turns out that all of the limit laws continue to hold. However, as we will not be using these limits much in this text, we will not devote much attention to these matters. We will note though that this definition is consistent with the notion of a limit $\lim_{n\to\infty} a_n$ of a sequence (Exercise 9.10.1).

— Exercises —

Exercise 9.10.1 Let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers, then a_n can also be thought of as a function from $\mathbf{N}$ to $\mathbf{R}$, which takes each natural number n to a real number a_n. Show that

$$\lim_{n\to+\infty;n\in\mathbf{N}} a_n = \lim_{n\to\infty} a_n$$

where the left-hand limit is defined by Definition 9.10.3 and the right-hand limit is defined by Definition 6.1.8. More precisely, show that if one of the above two limits exists then so does the other, and then they both have the same value. Thus the two notions of limit here are compatible.

Chapter 10
Differentiation of Functions

10.1 Basic Definitions

We can now begin the rigorous treatment of calculus in earnest, starting with the notion of a derivative. We can now define derivatives analytically, using limits, in contrast to the geometric definition of derivatives, which uses tangents. The advantage of working analytically is that (a) we do not need to know the axioms of geometry, and (b) these definitions can be modified to handle functions of several variables, or functions whose values are vectors instead of scalar. Furthermore, one's geometric intuition becomes difficult to rely on once one has more than three dimensions in play. (Conversely, one can use one's experience in analytic rigor to extend one's geometric intuition to such abstract settings; as mentioned earlier, the two viewpoints complement rather than oppose each other.)

Definition 10.1.1 *(Differentiability at a point)* Let X be a subset of $\mathbf{R}$, and let $x_0 \in X$ be an element of X which is also a limit point of X. Let $f : X \to \mathbf{R}$ be a function. If the limit

$$\lim_{x \to x_0; x \in X \backslash \{x_0\}} \frac{f(x) - f(x_0)}{x - x_0}$$

converges to some real number L, then we say that f is *differentiable at* x_0 on X *with derivative* L and write $f'(x_0) := L$. If the limit does not exist, or if x_0 is not an element of X or not a limit point of X, we leave $f'(x_0)$ undefined and say that f is *not differentiable at* x_0 on X.

Remark 10.1.2 Note that we need x_0 to be a limit point in order for x_0 to be adherent to $X \backslash \{x_0\}$, otherwise the limit

$$\lim_{x \to x_0; x \in X \backslash \{x_0\}} \frac{f(x) - f(x_0)}{x - x_0}$$

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_10

would automatically be undefined. In particular, we do not define the derivative of a function at an isolated point; for instance, if one restricts the function $f\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2$ to the domain $X := [1, 2] \cup \{3\}$, then the restriction of the function ceases to be differentiable at 3. (See however Exercise 10.1.1 below.) In practice, the domain X will almost always be an interval, and so by Lemma 9.1.21 all elements x_0 of X will automatically be limit points and we will not have to care much about these issues.

Example 10.1.3 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function $f(x) := x^2$, and let x_0 be any real number. To see whether f is differentiable at x_0 on $\mathbf{R}$, we compute the limit

$$\lim_{x \to x_0; x \in \mathbf{R}\backslash\{x_0\}} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0; x \in \mathbf{R}\backslash\{x_0\}} \frac{x^2 - x_0^2}{x - x_0}.$$

We can factor the numerator as $(x^2 - x_0^2) = (x - x_0)(x + x_0)$. Since $x \in \mathbf{R}\backslash\{x_0\}$, we may legitimately cancel the factors of $x - x_0$ and write the above limit as

$$\lim_{x \to x_0; x \in \mathbf{R}\backslash\{x_0\}} x + x_0$$

which by limit laws is equal to $2x_0$. Thus the function $f(x)$ is differentiable at x_0 and its derivative there is $2x_0$.

Remark 10.1.4 This point is trivial, but it is worth mentioning: if $f\colon X \to \mathbf{R}$ is differentiable at x_0, and $g\colon X \to \mathbf{R}$ is equal to f (i.e., $g(x) = f(x)$ for all $x \in X$), then g is also differentiable at x_0 and $g'(x_0) = f'(x_0)$ (why?). However, if two functions f and g merely have the same *value* at x_0, i.e., $g(x_0) = f(x_0)$, this does not imply that $g'(x_0) = f'(x_0)$. (Can you see a counterexample?) Thus there is a big difference between two functions being equal on their whole domain and merely being equal at one point.

Remark 10.1.5 One sometimes writes $\frac{\mathrm{d}f}{\mathrm{d}x}$ instead of f'. This notation is of course very familiar and convenient, but one has to be a little careful, because it is only safe to use as long as x is the only variable used to represent the input for f; otherwise one can get into all sorts of trouble. For instance, the function $f\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x^2$ has derivative $\frac{\mathrm{d}f}{\mathrm{d}x} = 2x$, but the function $g\colon \mathbf{R} \to \mathbf{R}$ defined by $g(y) := y^2$ would seem to have derivative $\frac{\mathrm{d}g}{\mathrm{d}x} = 0$ if y and x are independent variables, despite the fact that g and f are exactly the same function. Because of this possible source of confusion, we will refrain from using the notation $\frac{\mathrm{d}f}{\mathrm{d}x}$ whenever it could possibly lead to confusion. (This confusion becomes even worse in the calculus of several variables, and the standard notation of $\frac{\partial f}{\partial x}$ can lead to some serious ambiguities. There are ways to resolve these ambiguities, most notably by introducing the notion of differentiation along vector fields, but this is beyond the scope of this text.)

Example 10.1.6 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function $f(x) := |x|$, and let $x_0 = 0$. To see whether f is differentiable at 0 on $\mathbf{R}$, we compute the limit

$$\lim_{x\to 0; x\in\mathbf{R}\backslash\{0\}} \frac{f(x)-f(0)}{x-0} = \lim_{x\to 0; x\in\mathbf{R}\backslash\{0\}} \frac{|x|}{x}.$$

Now we take left limits and right limits. The right limit is

$$\lim_{x\to 0; x\in(0,\infty)} \frac{|x|}{x} = \lim_{x\to 0; x\in(0,\infty)} \frac{x}{x} = \lim_{x\to 0; x\in(0,\infty)} 1 = 1,$$

while the left limit is

$$\lim_{x\to 0; x\in(-\infty,0)} \frac{|x|}{x} = \lim_{x\to 0; x\in(-\infty,0)} \frac{-x}{x} = \lim_{x\to 0; x\in(-\infty,0)} -1 = -1,$$

and these limits do not match. Thus $\lim_{x\to 0; x\in\mathbf{R}\backslash\{0\}} \frac{|x|}{x}$ does not exist, and f is not differentiable at 0 on $\mathbf{R}$. However, if one restricts f to $[0,\infty)$, then the restricted function $f|_{[0,\infty)}$ *is* differentiable at 0 on $[0,\infty)$, with derivative 1:

$$\lim_{x\to 0; x\in[0,\infty)\backslash\{0\}} \frac{f(x)-f(0)}{x-0} = \lim_{x\to 0; x\in(0,\infty)} \frac{|x|}{x} = 1.$$

Similarly, when one restricts f to $(-\infty,0]$, the restricted function $f|_{(-\infty,0]}$ is differentiable at 0 on $(-\infty,0]$, with derivative -1. Thus even when a function is not differentiable, it is sometimes possible to restore the differentiability by restricting the domain of the function.

If a function is differentiable at x_0, then it is approximately linear near x_0:

Proposition 10.1.7 (Newton's approximation) *Let X be a subset of $\mathbf{R}$, let $x_0 \in X$ be a limit point of X, let $f: X \to \mathbf{R}$ be a function, and let L be a real number. Then the following statements are logically equivalent:*

(a) f is differentiable at x_0 on X with derivative L.

(b) For every $\varepsilon > 0$, there exists a $\delta > 0$ such that $f(x)$ is $\varepsilon|x-x_0|$-close to $f(x_0) + L(x-x_0)$ whenever $x \in X$ is δ-close to x_0, i.e., we have

$$|f(x) - (f(x_0) + L(x-x_0))| \leq \varepsilon|x-x_0|$$

whenever $x \in X$ and $|x-x_0| \leq \delta$.

Remark 10.1.8 Newton's approximation is of course named after the great scientist and mathematician Isaac Newton (1642–1727), one of the founders of differential and integral calculus.

Proof See Exercise 10.1.2.

Remark 10.1.9 We can phrase Proposition 10.1.7 in a more informal way: if f is differentiable at x_0, then one has the approximation $f(x) \approx f(x_0) + f'(x_0)(x-x_0)$, and conversely.

As the example of the function $f\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := |x|$ shows, a function can be continuous at a point without being differentiable at that point. However, the converse is true:

Proposition 10.1.10 (Differentiability implies continuity) *Let X be a subset of $\mathbf{R}$, let $x_0 \in X$ be a limit point of X, and let $f\colon X \to \mathbf{R}$ be a function. If f is differentiable at x_0, then f is also continuous at x_0.*

Proof See Exercise 10.1.3.

Definition 10.1.11 *(Differentiability on a domain)* Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a function. We say that f is *differentiable on X* if, for every limit point $x_0 \in X$, the function f is differentiable at x_0 on X.

From Proposition 10.1.10 and the above definition, as well as the fact that a function is automatically continuous at every isolated point of its domain, we have an immediate corollary:

Corollary 10.1.12 *Let X be a subset of $\mathbf{R}$, and let $f\colon X \to \mathbf{R}$ be a function which is differentiable on X. Then f is also continuous on X.*

Now we state the basic properties of derivatives which you are all familiar with.

Theorem 10.1.13 *[Differential calculus] Let X be a subset of $\mathbf{R}$, let $x_0 \in X$ be a limit point of X, and let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be functions.*

(a) If f is a constant function, i.e., there exists a real number c such that $f(x) = c$ for all $x \in X$, then f is differentiable at x_0 and $f'(x_0) = 0$.
(b) If f is the identity function, i.e., $f(x) = x$ for all $x \in X$, then f is differentiable at x_0 and $f'(x_0) = 1$.
(c) (Sum rule) If f and g are differentiable at x_0, then $f + g$ is also differentiable at x_0, and $(f + g)'(x_0) = f'(x_0) + g'(x_0)$.
(d) (Product rule) If f and g are differentiable at x_0, then fg is also differentiable at x_0, and $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$.
(e) If f is differentiable at x_0 and c is a real number, then cf is also differentiable at x_0, and $(cf)'(x_0) = cf'(x_0)$.
(f) (Difference rule) If f and g are differentiable at x_0, then $f - g$ is also differentiable at x_0, and $(f - g)'(x_0) = f'(x_0) - g'(x_0)$.
(g) If g is differentiable at x_0, and g is non-zero on X (i.e., $g(x) \neq 0$ for all $x \in X$), then $1/g$ is also differentiable at x_0, and $(\frac{1}{g})'(x_0) = -\frac{g'(x_0)}{g(x_0)^2}$.
(h) (Quotient rule) If f and g are differentiable at x_0, and g is non-zero on X, then f/g is also differentiable at x_0, and

$$\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

Remark 10.1.14 The product rule is also known as the *Leibniz rule*, after Gottfried Leibniz (1646–1716), who was the other founder of differential and integral calculus besides Newton.

Proof See Exercise 10.1.4.

As you are well aware, the above rules allow one to compute many derivatives easily. For instance, if $f : \mathbf{R}\backslash\{1\} \to \mathbf{R}$ is the function $f(x) := \frac{x-2}{x-1}$, then it is easy to use the above rules to show that $f'(x_0) = \frac{1}{(x_0-1)^2}$ for all $x_0 \in \mathbf{R}\backslash\{1\}$. (Why? Note that every point x_0 in $\mathbf{R}\backslash\{1\}$ is a limit point of $\mathbf{R}\backslash\{1\}$.)

Another fundamental property of differentiable functions is the following:

Theorem 10.1.15 *[Chain rule] Let X, Y be subsets of $\mathbf{R}$, let $x_0 \in X$ be a limit point of X, and let $y_0 \in Y$ be a limit point of Y. Let $f : X \to Y$ be a function such that $f(x_0) = y_0$, and such that f is differentiable at x_0. Suppose that $g : Y \to \mathbf{R}$ is a function which is differentiable at y_0. Then the function $g \circ f : X \to \mathbf{R}$ is differentiable at x_0, and*

$$(g \circ f)'(x_0) = g'(y_0)f'(x_0).$$

Proof See Exercise 10.1.7.

Example 10.1.16 If $f : \mathbf{R}\backslash\{1\} \to \mathbf{R}$ is the function $f(x) := \frac{x-2}{x-1}$, and $g : \mathbf{R} \to \mathbf{R}$ is the function $g(y) := y^2$, then $g \circ f(x) = (\frac{x-2}{x-1})^2$, and the chain rule gives

$$(g \circ f)'(x_0) = 2\left(\frac{x_0-2}{x_0-1}\right)\frac{1}{(x_0-1)^2}.$$

Remark 10.1.17 If one writes y for $f(x)$, and z for $g(y)$, then the chain rule can be written in the more visually appealing manner $\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}$. However, this notation can be misleading (for instance it blurs the distinction between dependent variable and independent variable, especially for y) and leads one to believe that the quantities $\mathrm{d}z$, $\mathrm{d}y$, $\mathrm{d}x$ can be manipulated like real numbers. However, these quantities are not real numbers (in fact, we have not assigned any meaning to them at all), and treating them as such can lead to problems in the future. For instance, if f depends on x_1 and x_2, which depend on t, then chain rule for several variables asserts that $\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x_1}\frac{\mathrm{d}x_1}{\mathrm{d}t} + \frac{\partial f}{\partial x_2}\frac{\mathrm{d}x_2}{\mathrm{d}t}$, but this rule might seem suspect if one treated $\mathrm{d}f$, $\mathrm{d}t$, etc. as real numbers. It is possible to think of $\mathrm{d}y$, $\mathrm{d}x$, etc. as “infinitesimal real numbers” if one knows what one is doing, but for those just starting out in analysis, I would not recommend this approach, especially if one wishes to work rigorously. (There is a way to make all of this rigorous, even for the calculus of several variables, but it requires the notion of a tangent vector and the derivative map, both of which are beyond the scope of this text.)

— Exercises —

Exercise 10.1.1 Suppose that X is a subset of $\mathbf{R}$, x_0 is a limit point of X, and $f\colon X \to \mathbf{R}$ is a function which is differentiable at x_0. Let $Y \subseteq X$ be such that $x_0 \in Y$, and x_0 is also a limit point of Y. Prove that the restricted function $f|_Y : Y \to \mathbf{R}$ is also differentiable at x_0 and has the same derivative as f at x_0. Explain why this does not contradict the discussion in Remark 10.1.2.

Exercise 10.1.2 Prove Proposition 10.1.7. (*Hint:* the cases $x = x_0$ and $x \neq x_0$ have to be treated separately.)

Exercise 10.1.3 Prove Proposition 10.1.10. (*Hint:* either use the limit laws (Proposition 9.3.14) or use Proposition 10.1.7.)

Exercise 10.1.4 Prove Theorem 10.1.13. (*Hint:* use the limit laws in Proposition 9.3.14. Use earlier parts of this theorem to prove the latter. For the product rule, use the identity

$$\begin{aligned} &f(x)g(x) - f(x_0)g(x_0)\\ &\qquad = f(x)g(x) - f(x)g(x_0) + f(x)g(x_0) - f(x_0)g(x_0)\\ &\qquad = f(x)(g(x) - g(x_0)) + (f(x) - f(x_0))g(x_0); \end{aligned}$$

this trick of adding and subtracting an intermediate term is sometimes known as the "middle-man trick" and is very useful in analysis.)

Exercise 10.1.5 Let n be a natural number, and let $f\colon \mathbf{R} \to \mathbf{R}$ be the function $f(x) := x^n$. Show that f is differentiable on $\mathbf{R}$ and $f'(x) = nx^{n-1}$ for all $x \in \mathbf{R}$, adopting the convention that nx^{n-1} is 0 when $n = 0$. (*Hint:* use Theorem 10.1.13 and induction.)

Exercise 10.1.6 Let n be a *negative* integer, and let $f\colon \mathbf{R}\backslash\{0\} \to \mathbf{R}$ be the function $f(x) := x^n$. Show that f is differentiable on $\mathbf{R}\backslash\{0\}$, and that $f'(x) = nx^{n-1}$ for all $x \in \mathbf{R}\backslash\{0\}$. (*Hint:* use Theorem 10.1.13 and Exercise 10.1.5.)

Exercise 10.1.7 Prove Theorem 10.1.15. (*Hint:* one way to do this is via Newton's approximation, Proposition 10.1.7. Another way is to use Proposition 9.3.9 and Proposition 10.1.10 to convert this problem into one involving limits of sequences; however with the latter strategy one has to treat the case $f'(x_0) = 0$ separately, as some division-by-zero subtleties can occur in that case.)

10.2 Local Maxima, Local Minima, and Derivatives

As you learnt in your basic calculus courses, one very common application of using derivatives is to locate maxima and minima. We now present this material again, but this time in a rigorous manner.

The notion of a function $f\colon X \to \mathbf{R}$ attaining a maximum or minimum at a point $x_0 \in X$ was defined in Definition 9.6.5. We now localize this definition:

Definition 10.2.1 *(Local maxima and minima)* Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ be a function, and let $x_0 \in X$. We say that f attains a *local maximum* at x_0 iff there exists a $\delta > 0$ such that the restriction $f|_{X\cap(x_0-\delta,x_0+\delta)}$ of f to $X \cap (x_0 - \delta, x_0 + \delta)$ attains a maximum at x_0. We say that f attains a *local minimum* at x_0 iff there exists a $\delta > 0$ such that the restriction $f|_{X\cap(x_0-\delta,x_0+\delta)}$ of f to $X \cap (x_0 - \delta, x_0 + \delta)$ attains a minimum at x_0.

Remark 10.2.2 If f attains a maximum at x_0, we sometimes say that f attains a *global* maximum at x_0, in order to distinguish it from the local maxima defined here. Note that if f attains a global maximum at x_0, then it certainly also attains a local maximum at this x_0, and similarly for minima.

Example 10.2.3 Let $f: \mathbf{R} \to \mathbf{R}$ denote the function $f(x) := x^2 - x^4$. This function does not attain a global minimum at 0, since for example $f(2) = -12 < 0 = f(0)$, however it does attain a local minimum, for if we choose $\delta := 1$ and restrict f to the interval $(-1, 1)$, then for all $x \in (-1, 1)$ we have $x^4 \le x^2$ and thus $f(x) = x^2 - x^4 \ge 0 = f(0)$, and so $f|_{(-1,1)}$ has a (global) minimum at 0.

Example 10.2.4 Let $f: \mathbf{Z} \to \mathbf{R}$ be the function $f(x) = x$, defined on the integers only. Then f has no global maximum or global minimum (why?), but attains both a local maximum and local minimum at every integer n (why?).

Remark 10.2.5 If $f: X \to \mathbf{R}$ attains a local maximum at a point x_0 in X, and $Y \subseteq X$ is a subset of X which contains x_0, then the restriction $f|_Y : Y \to \mathbf{R}$ also attains a local maximum at x_0 (why?). Similarly for minima.

The connection between local maxima, minima, and derivatives is the following.

Proposition 10.2.6 (Local extrema are stationary) *Let $a < b$ be real numbers, and let $f: (a, b) \to \mathbf{R}$ be a function. If $x_0 \in (a, b)$, f is differentiable at x_0, and f attains either a local maximum or a local minimum at x_0, then $f'(x_0) = 0$.*

Proof See Exercise 10.2.1.

Note that f must be differentiable for this proposition to work; see Exercise 10.2.2. Also, this proposition does not work if the open interval (a, b) is replaced by a closed interval $[a, b]$. For instance, the function $f: [1, 2] \to \mathbf{R}$ defined by $f(x) := x$ has a local maximum at $x_0 = 2$ and a local minimum $x_0 = 1$ (in fact, these local extrema are global extrema), but at both points the derivative is $f'(x_0) = 1$, not $f'(x_0) = 0$. Thus the endpoints of an interval can be local maxima or minima even if the derivative is not zero there. Finally, the converse of this proposition is false (Exercise 10.2.3).

By combining Proposition 10.2.6 with the maximum principle, one can obtain

Theorem 10.2.7 *[Rolle's theorem] Let $a < b$ be real numbers, and let $g: [a, b] \to \mathbf{R}$ be a continuous function which is differentiable on (a, b). Suppose also that $g(a) = g(b)$. Then there exists an $x \in (a, b)$ such that $g'(x) = 0$.*

Proof See Exercise 10.2.4.

Remark 10.2.8 Note that we only assume f is differentiable on the open interval (a, b), though of course the theorem also holds if we assume f is differentiable on the closed interval $[a, b]$, since this is larger than (a, b).

Rolle's theorem has an important corollary.

Corollary 10.2.9 (Mean-value theorem) *Let $a < b$ be real numbers, and let $f : [a, b] \to \mathbf{R}$ be a function which is continuous on $[a, b]$ and differentiable on (a, b). Then there exists an $x \in (a, b)$ such that $f'(x) = \frac{f(b)-f(a)}{b-a}$.*

Proof See Exercise 10.2.5.

— Exercises —

Exercise 10.2.1 Prove Proposition 10.2.6.

Exercise 10.2.2 Give an example of a function $f : (-1, 1) \to \mathbf{R}$ which is continuous and attains a global maximum at 0, but which is not differentiable at 0. Explain why this does not contradict Proposition 10.2.6.

Exercise 10.2.3 Give an example of a function $f : (-1, 1) \to \mathbf{R}$ which is differentiable, and whose derivative equals 0 at 0, but such that 0 is neither a local minimum nor a local maximum. Explain why this does not contradict Proposition 10.2.6.

Exercise 10.2.4 Prove Theorem 10.2.7. (*Hint:* use the maximum principle, Proposition 9.6.7, followed by Proposition 10.2.6. Note that the maximum principle does not tell you whether the maximum or minimum is in the open interval (a, b) or is one of the boundary points a, b, so you have to divide into cases and use the hypothesis $g(a) = g(b)$ somehow.)

Exercise 10.2.5 Use Theorem 10.2.7 to prove Corollary 10.2.9. (*Hint:* consider a function of the form $f(x) - cx$ for some carefully chosen real number c.)

Exercise 10.2.6 Let $M > 0$, and let $f : [a, b] \to \mathbf{R}$ be a function which is continuous on $[a, b]$ and differentiable on (a, b), and such that $|f'(x)| \leq M$ for all $x \in (a, b)$ (i.e., the derivative of f is bounded). Show that for any $x, y \in [a, b]$ we have the inequality $|f(x) - f(y)| \leq M|x - y|$. (*Hint:* apply the mean-value theorem (Corollary 10.2.9) to a suitable restriction of f.) Functions which obey the bound $|f(x) - f(y)| \leq M|x - y|$ are known as *Lipschitz continuous functions* with *Lipschitz constant M*; thus this exercise shows that functions with bounded derivative are Lipschitz continuous.

Exercise 10.2.7 Let $f : \mathbf{R} \to \mathbf{R}$ be a differentiable function such that f' is bounded. Show that f is uniformly continuous. (*Hint:* use the preceding exercise.)

10.3 Monotone Functions and Derivatives

In your elementary calculus courses, you may have come across the assertion that a positive derivative meant an increasing function, and a negative derivative meant a decreasing function. This statement is not completely accurate, but it is pretty close; we now give the precise version of these statements below.

Proposition 10.3.1 *Let X be a subset of* $\mathbf{R}$, *let $x_0 \in X$ be a limit point of X, and let $f : X \to \mathbf{R}$ be a function. If f is monotone increasing* and *f is differentiable at x_0, then $f'(x_0) \geq 0$. If f is monotone decreasing* and *f is differentiable at x_0, then $f'(x_0) \leq 0$.*

Proof See Exercise 10.3.1.

Remark 10.3.2 We have to assume that f is differentiable at x_0; there exist monotone functions which are not always differentiable (see Exercise 10.3.2), and of course if f is not differentiable at x_0 we cannot possibly conclude that $f'(x_0) \geq 0$ or $f'(x_0) \leq 0$.

One might naively guess that if f were *strictly* monotone increasing, and f was differentiable at x_0, then the derivative $f'(x_0)$ would be strictly positive instead of merely non-negative. Unfortunately, this is not always the case (Exercise 10.3.3).

On the other hand, we do have a converse result: if function has strictly positive derivative, then it must be strictly monotone increasing:

Proposition 10.3.3 *Let $a < b$, and let $f\colon [a,b] \to \mathbf{R}$ be a differentiable function. If $f'(x) > 0$ for all $x \in [a,b]$, then f is strictly monotone increasing. If $f'(x) < 0$ for all $x \in [a,b]$, then f is strictly monotone decreasing. If $f'(x) = 0$ for all $x \in [a,b]$, then f is a constant function.*

Proof See Exercise 10.3.4.

— Exercises —

Exercise 10.3.1 Prove Proposition 10.3.1.

Exercise 10.3.2 Give an example of a function $f\colon (-1,1) \to \mathbf{R}$ which is continuous and monotone increasing, but which is not differentiable at 0. Explain why this does not contradict Proposition 10.3.1.

Exercise 10.3.3 Give an example of a function $f\colon \mathbf{R} \to \mathbf{R}$ which is strictly monotone increasing and differentiable, but whose derivative at 0 is zero. Explain why this does not contradict Proposition 10.3.1 or Proposition 10.3.3. (*Hint:* look at Exercise 10.2.3.)

Exercise 10.3.4 Prove Proposition 10.3.3. (*Hint:* you do not have integrals or the fundamental theorem of calculus yet, so these tools cannot be used. However, one can proceed via the mean-value theorem, Corollary 10.2.9.)

Exercise 10.3.5 Give an example of a subset $X \subseteq \mathbf{R}$ and a function $f\colon X \to \mathbf{R}$ which is differentiable on X, is such that $f'(x) > 0$ for all $x \in X$, but f is not strictly monotone increasing. (*Hint:* the conditions here are subtly different from those in Proposition 10.3.3. What is the difference, and how can one exploit that difference to obtain the example?)

10.4 Inverse Functions and Derivatives

We now ask the following question: if we know that a function $f\colon X \to Y$ is differentiable, and it has an inverse $f^{-1}\colon Y \to X$, what can we say about the differentiability of f^{-1}? This will be useful for many applications, for instance if we want to differentiate the function $f(x) := x^{1/n}$.

We begin with a preliminary result.

Lemma 10.4.1 *Let X, Y be subsets of* $\mathbf{R}$, *and let $f\colon X \to Y$ be an invertible function, with inverse $f^{-1} : Y \to X$. Suppose that $x_0 \in X$ and $y_0 \in Y$ are limit points of X, Y, respectively, such that $y_0 = f(x_0)$ (which also implies that $x_0 = f^{-1}(y_0)$). If f is differentiable at x_0,* and *f^{-1} is differentiable at y_0, then*

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)}.$$

Proof From the chain rule (Theorem 10.1.15) we have

$$(f^{-1} \circ f)'(x_0) = (f^{-1})'(y_0) f'(x_0).$$

But $f^{-1} \circ f$ is the identity function on X, and hence by Theorem 10.1.13(b) $(f^{-1} \circ f)'(x_0) = 1$. The claim follows.

As a particular corollary of Lemma 10.4.1, we see that if f is differentiable at x_0 with $f'(x_0) = 0$, then f^{-1} cannot be differentiable at $y_0 = f(x_0)$, since $1/f'(x_0)$ is undefined in that case. Thus for instance, the function $g\colon [0, \infty) \to [0, \infty)$ defined by $g(y) := y^{1/3}$ cannot be differentiable at 0, since this function is the inverse $g = f^{-1}$ of the function $f\colon [0, \infty) \to [0, \infty)$ defined by $f(x) := x^3$, and this function has a derivative of 0 at $f^{-1}(0) = 0$.

If one writes $y = f(x)$, so that $x = f^{-1}(y)$, then one can write the conclusion of Lemma 10.4.1 in the more appealing form $dx/dy = 1/(dy/dx)$. However, as mentioned before, this way of writing things, while very convenient and easy to remember, can be misleading and cause errors if applied too carelessly (especially when one begins to work in the calculus of several variables).

Lemma 10.4.1 seems to answer the question of how to differentiate the inverse of a function; however it has one significant drawback: the lemma only works if one assumes *a priori* that f^{-1} is differentiable. Thus, if one does not already know that f^{-1} is differentiable, one cannot use Lemma 10.4.1 to compute the derivative of f^{-1}.

However, the following improved version of Lemma 10.4.1 will compensate for this fact, by relaxing the requirement on f^{-1} from differentiability to continuity.

Theorem 10.4.2 *[Inverse function theorem] Let X, Y be subsets of* $\mathbf{R}$, *and let $f\colon X \to Y$ be an invertible function, with inverse $f^{-1} : Y \to X$. Suppose that $x_0 \in X$ and $y_0 \in Y$ are limit points of X, Y, respectively, such that $f(x_0) = y_0$. If f is differentiable at x_0, f^{-1} is continuous at y_0, and $f'(x_0) \neq 0$, then f^{-1} is differentiable at y_0 and*

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)}.$$

Proof We have to show that

$$\lim_{y \to y_0; y \in Y \setminus \{y_0\}} \frac{f^{-1}(y) - f^{-1}(y_0)}{y - y_0} = \frac{1}{f'(x_0)}.$$

By Proposition 9.3.9, it suffices to show that

$$\lim_{n\to\infty} \frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} = \frac{1}{f'(x_0)}$$

for any sequence $(y_n)_{n=1}^{\infty}$ of elements in $Y\backslash\{y_0\}$ which converge to y_0.

To prove this, we set $x_n := f^{-1}(y_n)$. Then $(x_n)_{n=1}^{\infty}$ is a sequence of elements in $X\backslash\{x_0\}$. (Why? Note that f^{-1} is a bijection). Since f^{-1} is continuous by assumption, we know that $x_n = f^{-1}(y_n)$ converges to $f^{-1}(y_0) = x_0$ as $n \to \infty$. Thus, since f is differentiable at x_0, we have (by Proposition 9.3.9 again)

$$\lim_{n\to\infty} \frac{f(x_n) - f(x_0)}{x_n - x_0} = f'(x_0).$$

But since $x_n \neq x_0$ and f is a bijection, the fraction $\frac{f(x_n)-f(x_0)}{x_n-x_0}$ is non-zero. Also, by hypothesis $f'(x_0)$ is non-zero. So by limit laws

$$\lim_{n\to\infty} \frac{x_n - x_0}{f(x_n) - f(x_0)} = \frac{1}{f'(x_0)}.$$

But since $x_n = f^{-1}(y_n)$ and $x_0 = f^{-1}(y_0)$, we thus have

$$\lim_{n\to\infty} \frac{f^{-1}(y_n) - f^{-1}(y_0)}{y_n - y_0} = \frac{1}{f'(x_0)}$$

as desired.

We give some applications of the inverse function theorem in the exercises below.

— Exercises —

Exercise 10.4.1 Let $n \geq 1$ be a natural number, and let $g\colon (0,\infty) \to (0,\infty)$ be the function $g(x) := x^{1/n}$.

(a) Show that g is continuous on $(0,\infty)$. (*Hint:* use Proposition 9.4.11.)

(b) Show that g is differentiable on $(0,\infty)$, and that $g'(x) = \frac{1}{n}x^{\frac{1}{n}-1}$ for all $x \in (0,\infty)$. (*Hint:* use the inverse function theorem and (a).)

Exercise 10.4.2 Let q be a rational number, and let $f\colon (0,\infty) \to \mathbf{R}$ be the function $f(x) = x^q$.

(a) Show that f is differentiable on $(0,\infty)$ and that $f'(x) = qx^{q-1}$. (*Hint:* use Exercise 10.4.1 and the laws of differential calculus in Theorem 10.1.13 and Theorem 10.1.15.)

(b) Show that $\lim_{x\to 1; x\in(0,\infty)\backslash\{1\}} \frac{x^q-1}{x-1} = q$ for every rational number q. (*Hint:* use part (a) and Definition 10.1.1. An alternate route is to apply L'Hôpital's rule from the next section.)

Exercise 10.4.3 Let α be a real number, and let $f\colon (0,\infty) \to \mathbf{R}$ be the function $f(x) = x^\alpha$.

(a) Show that $\lim_{x\to 1; x\in(0,\infty)\backslash\{1\}} \frac{f(x)-f(1)}{x-1} = \alpha$. (*Hint:* use Exercise 10.4.2 and the comparison principle; you may need to consider right and left limits separately. Proposition 5.4.14 may also be helpful.)

(b) Show that f is differentiable on $(0,\infty)$ and that $f'(x) = \alpha x^{\alpha-1}$. (*Hint:* use (a), exponent laws (Proposition 6.7.3), and Definition 10.1.1.)

10.5 L'Hôpital's Rule

Finally, we present a version of a rule you are all familiar with.

Proposition 10.5.1 (L'Hôpital's rule I) *Let X be a subset of $\mathbf{R}$, let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be functions, and let $x_0 \in X$ be a limit point of X. Suppose that $f(x_0) = g(x_0) = 0$, that f and g are both differentiable at x_0, but $g'(x_0) \neq 0$. Then there exists a $\delta > 0$ such that $g(x) \neq 0$ for all $x \in (X \cap (x_0 - \delta, x_0 + \delta))\backslash\{x_0\}$, and*

$$\lim_{x \to x_0; x \in (X \cap (x_0 - \delta, x_0 + \delta))\backslash\{x_0\}} \frac{f(x)}{g(x)} = \frac{f'(x_0)}{g'(x_0)}.$$

Proof See Exercise 10.5.1.

The presence of the δ here may seem somewhat strange, but is needed because $g(x)$ might vanish at some points other than x_0, which would imply that quotient $\frac{f(x)}{g(x)}$ is not necessarily defined at all points in $X\backslash\{x_0\}$.

A more sophisticated version of L'Hôpital's rule is the following.

Proposition 10.5.2 (L'Hôpital's rule II) *Let $a < b$ be real numbers, and let $f : [a, b] \to \mathbf{R}$ and $g\colon [a, b] \to \mathbf{R}$ be functions which are continuous on $[a, b]$ and differentiable on $(a, b]$. Suppose that $f(a) = g(a) = 0$, that g' is non-zero on $(a, b]$ (i.e., $g'(x) \neq 0$ for all $x \in (a, b]$), and $\lim_{x \to a; x \in (a,b]} \frac{f'(x)}{g'(x)}$ exists and equals L. Then $g(x) \neq 0$ for all $x \in (a, b]$, and $\lim_{x \to a; x \in (a,b]} \frac{f(x)}{g(x)}$ exists and equals L.*

Remark 10.5.3 This proposition only considers limits to the right of a, but one can easily state and prove a similar proposition for limits to the left of a, or around both sides of a. Speaking very informally, the proposition states that

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)},$$

though one has to ensure all of the conditions of the proposition hold (in particular, that $f(a) = g(a) = 0$, and that the right-hand limit exists), before one can apply L'Hôpital's rule.

Proof (Optional) We first show that $g(x) \neq 0$ for all $x \in (a, b]$. Suppose for sake of contradiction that $g(x) = 0$ for some $x \in (a, b]$. But since $g(a)$ is also zero, we can apply Rolle's theorem to obtain $g'(y) = 0$ for some $a < y < x$, but this contradicts the hypothesis that g' is non-zero on $[a, b]$.

Now we show that $\lim_{x \to a; x \in (a,b]} \frac{f(x)}{g(x)} = L$. By Proposition 9.3.9, it will suffice to show that

$$\lim_{n \to \infty} \frac{f(x_n)}{g(x_n)} = L$$

for any sequence $(x_n)_{n=1}^{\infty}$ taking values in $(a, b]$ which converges to a.

Consider a single x_n, and consider the function $h_n \colon [a, x_n] \to \mathbf{R}$ defined by

$$h_n(x) := f(x)g(x_n) - g(x)f(x_n).$$

Observe that h_n is continuous on $[a, x_n]$ and equals 0 at both a and x_n and is differentiable on (a, x_n) with derivative $h_n'(x) = f'(x)g(x_n) - g'(x)f(x_n)$. (Note that $f(x_n)$ and $g(x_n)$ are constants with respect to x.) By Rolle's theorem (Theorem 10.2.7), we can thus find $y_n \in (a, x_n)$ such that $h_n'(y_n) = 0$, which implies that

$$\frac{f(x_n)}{g(x_n)} = \frac{f'(y_n)}{g'(y_n)}.$$

Since $y_n \in (a, x_n)$ for all n, and x_n converges to a as $n \to \infty$, we see from the squeeze test (Corollary 6.4.14) that y_n also converges to a as $n \to \infty$. Thus $\frac{f'(y_n)}{g'(y_n)}$ converges to L, and thus $\frac{f(x_n)}{g(x_n)}$ also converges to L, as desired.

— Exercises —

Exercise 10.5.1 Prove Proposition 10.5.1. (*Hint:* to show that $g(x) \neq 0$ near x_0, you may wish to use Newton's approximation (Proposition 10.1.7). For the rest of the proposition, use the limit laws, Proposition 9.3.14.)

Exercise 10.5.2 Explain why Example 1.2.12 does not contradict either of the propositions in this section.

Chapter 11
The Riemann Integral

In the previous chapter we reviewed *differentiation*—one of the two pillars of single variable calculus. The other pillar is, of course, *integration*, which is the focus of the current chapter. More precisely, we will turn to the *definite integral*, the integral of a function on a fixed interval, as opposed to the *indefinite integral*, otherwise known as the antiderivative. These two are of course linked by the *Fundamental theorem of calculus*, of which more will be said later.

For us, the study of the definite integral will start with an interval I which could be open, closed, or half-open, and a function $f: I \to \mathbf{R}$, and will lead us to a number $\int_I f$; we can write this integral as $\int_I f(x)\,\mathrm{d}x$ (of course, we could replace x by any other dummy variable), or if I has endpoints a and b, we shall also write this integral as $\int_a^b f$ or $\int_a^b f(x)\,\mathrm{d}x$.

To actually *define* this integral $\int_I f$ is somewhat delicate (especially if one does not want to assume any axioms concerning geometric notions such as area), and not all functions f are integrable. It turns out that there are at least two ways to define this integral: the *Riemann integral*, named after Georg Riemann (1826–1866), which we will do here and which suffices for most applications, and the *Lebesgue integral*, named after Henri Lebesgue (1875–1941), which supercedes the Riemann integral and works for a much larger class of functions. The Lebesgue integral will be constructed in Chapter 8. There is also the *Riemann–Stieltjes integral* $\int_I f(x)\,\mathrm{d}\alpha(x)$, a generalization of the Riemann integral due to Thomas Stieltjes (1856–1894), which we will discuss in Sect. 11.8.

Our strategy in defining the Riemann integral is as follows. We begin by first defining a notion of integration on a very simple class of functions—the *piecewise constant* functions. These functions are quite primitive, but their advantage is that integration is very easy for these functions, as is verifying all the usual properties. Then, we handle more general functions by approximating them by piecewise constant functions.

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4_11

11.1 Partitions

Before we can introduce the concept of an integral, we need to describe how one can partition a large interval into smaller intervals. In this chapter, all intervals will be bounded intervals (as opposed to the more general intervals defined in Definition 9.1.1).

Definition 11.1.1 Let X be a subset of $\mathbf{R}$. We say that X is *connected* iff X is non-empty and the following property is true: whenever x, y are elements in X such that $x < y$, the bounded interval $[x, y]$ is a subset of X (i.e., every number between x and y is also in X).

Remark 11.1.2 Later on, in Section 2.4 we will define a more general notion of connectedness, which applies to any metric space.

Examples 11.1.3 The set $[1, 2]$ is connected, because if $x < y$ both lie in $[1, 2]$, then $1 \leq x < y \leq 2$, and so every element between x and y also lies in $[1, 2]$. A similar argument shows that the set $(1, 2)$ is connected. However, the set $[1, 2] \cup [3, 4]$ is not connected (why?). The real line is connected (why?). All singleton sets such as $\{3\}$ are connected, but for rather trivial reasons (these sets do not contain two elements x, y for which $x < y$).

Lemma 11.1.4 *Let X be a non-empty subset of the real line. Then the following two statements are logically equivalent:*

(a) X is bounded and connected.
(b) X is a bounded interval.

Proof See Exercise 11.1.1. □

Remark 11.1.5 Recall that intervals are allowed to be singleton points (e.g., the degenerate interval $[2, 2] = \{2\}$), or even the empty set.

Corollary 11.1.6 *If I and J are bounded intervals, then the intersection $I \cap J$ is also a bounded interval.*

Proof See Exercise 11.1.2. □

Example 11.1.7 The intersection of the bounded intervals $[2, 4]$ and $[4, 6]$ is $\{4\}$, which is also a bounded interval. The intersection of $(2, 4)$ and $(4, 6)$ is $\emptyset$.

We now give each bounded interval a length.

Definition 11.1.8 *(Length of intervals)* If I is a bounded interval, we define the *length* of I, denoted $|I|$ as follows. If I is one of the intervals $[a, b]$, (a, b), $[a, b)$, or $(a, b]$ for some real numbers $a < b$, then we define $|I| := b - a$. Otherwise, if I is a point or the empty set, we define $|I| = 0$.

Example 11.1.9 For instance, the length of $[3, 5]$ is 2, as is the length of $(3, 5)$; meanwhile, the length of $\{5\}$ or the empty set is 0.

Definition 11.1.10 *(Partitions)* Let I be a bounded interval. A *partition* of I is a finite set $\mathbf{P}$ of bounded intervals contained in I, such that every x in I lies in exactly one of the bounded intervals J in $\mathbf{P}$.

Remark 11.1.11 Note that a partition is a set of intervals, while each interval is itself a set of real numbers. Thus a partition is a set consisting of other sets.

Examples 11.1.12 The set $\mathbf{P} = \{\{1\}, (1, 3), [3, 5), \{5\}, (5, 8], \emptyset\}$ of bounded intervals is a partition of $[1, 8]$, because all the intervals in $\mathbf{P}$ lie in $[1, 8]$, and each element of $[1, 8]$ lies in exactly one interval in $\mathbf{P}$. Note that one could have removed the empty set from $\mathbf{P}$ and still obtain a partition. However, the set $\{[1, 4], [3, 5]\}$ is not a partition of $[1, 5]$ because some elements of $[1, 5]$ are included in more than one interval in the set. The set $\{(1, 3), (3, 5)\}$ is not a partition of $(1, 5)$ because some elements of $(1, 5)$ are not included in any interval in the set. The set $\{(0, 3), [3, 5)\}$ is not a partition of $(1, 5)$ because some intervals in the set are not contained in $(1, 5)$.

Now we come to a basic property about length:

Theorem 11.1.13 (Length is finitely additive) *Let I be a bounded interval, n be a natural number, and let $\mathbf{P}$ be a partition of I of cardinality n. Then*

$$|I| = \sum_{J \in \mathbf{P}} |J|.$$

Proof We prove this by induction on n. More precisely, we let $P(n)$ be the property that whenever I is a bounded interval, and whenever $\mathbf{P}$ is a partition of I with cardinality n, that $|I| = \sum_{J \in \mathbf{P}} |J|$.

The base case $P(0)$ is trivial; the only way that I can be partitioned into an empty partition is if I is itself empty (why?), at which point the claim is easy. The case $P(1)$ is also very easy; the only way that I can be partitioned into a singleton set $\{J\}$ is if $J = I$ (why?), at which point the claim is again very easy.

Now suppose inductively that $P(n)$ is true for some $n \geq 1$, and now we prove $P(n + 1)$. Let I be a bounded interval, and let $\mathbf{P}$ be a partition of I of cardinality $n + 1$.

If I is the empty set or a point, then all the intervals in $\mathbf{P}$ must also be either the empty set or a point (why?), and so every interval has length zero and the claim is trivial. Thus we will assume that I is an interval of the form (a, b), $(a, b]$, $[a, b)$, or $[a, b]$.

Let us first suppose that $b \in I$, i.e., I is either $(a, b]$ or $[a, b]$. Since $b \in I$, we know that one of the intervals K in $\mathbf{P}$ contains b. Since K is contained in I, it must therefore be of the form $(c, b]$, $[c, b]$, or $\{b\}$ for some real number c, with $a \leq c \leq b$ (in the latter case of $K = \{b\}$, we set $c := b$). In particular, this means that the set $I - K$ is also an interval of the form $[a, c]$, (a, c), $(a, c]$, $[a, c)$ when $c > a$, or a point or empty set when $a = c$. Either way, we easily see that

$$|I| = |K| + |I - K|.$$

On the other hand, since $\mathbf{P}$ forms a partition of I, we see that $\mathbf{P} - \{K\}$ forms a partition of $I - K$ (why?). By the induction hypothesis, we thus have

$$|I - K| = \sum_{J \in \mathbf{P} - \{K\}} |J|.$$

Combining these two identities (and using the laws of addition for finite sets, see Proposition 7.1.11) we obtain

$$|I| = \sum_{J \in \mathbf{P}} |J|$$

as desired.

Now suppose that $b \notin I$, i.e., I is either (a, b) or $[a, b)$. Then one of the intervals K also is of the form (c, b) or $[c, b)$ (see Exercise 11.1.3). In particular, this means that the set $I - K$ is also an interval of the form $[a, c]$, (a, c), $(a, c]$, $[a, c)$ when $c > a$, or a point or empty set when $a = c$. The rest of the argument then proceeds as above. □

There are two more things we need to do with partitions. One is to say when one partition is finer than another, and the other is to talk about the common refinement of two partitions.

Definition 11.1.14 *(Finer and coarser partitions)* Let I be a bounded interval, and let $\mathbf{P}$ and $\mathbf{P}'$ be two partitions of I. We say that $\mathbf{P}'$ is *finer* than $\mathbf{P}$ (or equivalently, that $\mathbf{P}$ is *coarser* than $\mathbf{P}'$) if for every J in $\mathbf{P}'$, there exists a K in $\mathbf{P}$ such that $J \subseteq K$.

Example 11.1.15 The partition $\{[1, 2), \{2\}, (2, 3), [3, 4]\}$ is finer than $\{[1, 2], (2, 4]\}$ (why?). Both partitions are finer than $\{[1, 4]\}$, which is the coarsest possible partition of $[1, 4]$. Note that there is no such thing as a "finest" partition of $[1, 4]$. (Why? recall all partitions are assumed to be finite.) We do not compare partitions of different intervals, for instance if $\mathbf{P}$ is a partition of $[1, 4]$ and $\mathbf{P}'$ is a partition of $[2, 5]$ then we would not say that $\mathbf{P}$ is coarser or finer than $\mathbf{P}'$.

Definition 11.1.16 *(Common refinement)* Let I be a bounded interval, and let $\mathbf{P}$ and $\mathbf{P}'$ be two partitions of I. We define the *common refinement* $\mathbf{P}\#\mathbf{P}'$ of $\mathbf{P}$ and $\mathbf{P}'$ to be the set

$$\mathbf{P}\#\mathbf{P}' := \{K \cap J : K \in \mathbf{P} \text{ and } J \in \mathbf{P}'\}.$$

Example 11.1.17 Let $\mathbf{P} := \{[1, 3), [3, 4]\}$ and $\mathbf{P}' := \{[1, 2], (2, 4]\}$ be two partitions of $[1, 4]$. Then $\mathbf{P}\#\mathbf{P}'$ is the set $\{[1, 2], (2, 3), [3, 4], \emptyset\}$ (why?).

Lemma 11.1.18 *Let I be a bounded interval, and let* $\mathbf{P}$ *and* $\mathbf{P}'$ *be two partitions of I. Then* $\mathbf{P}\#\mathbf{P}'$ *is also a partition of I, and is both finer than* $\mathbf{P}$ *and finer than* $\mathbf{P}'$.

Proof See Exercise 11.1.4. □

— Exercises —

Exercise 11.1.1 Prove Lemma 11.1.4. (*Hint:* in order to show that (a) implies (b) in the case when X is non-empty, consider the supremum and infimum of X.)

Exercise 11.1.2 Prove Corollary 11.1.6. (*Hint:* use Lemma 11.1.4, and explain why the intersection of two bounded sets is automatically bounded, and why the intersection of two connected sets is automatically connected.)

Exercise 11.1.3 Let I be a bounded interval of the form $I = (a, b)$ or $I = [a, b)$ for some real numbers $a < b$. Let $I_1, \dots, I_n$ be a partition of I. Prove that one of the intervals I_j in this partition is of the form $I_j = (c, b)$ or $I_j = [c, b)$ for some $a \leq c \leq b$. (*Hint:* prove by contradiction. First show that if I_j is *not* of the form (c, b) or $[c, b)$ for any $a \leq c \leq b$, then $\sup I_j$ is *strictly* less than b.)

Exercise 11.1.4 Prove Lemma 11.1.18.

11.2 Piecewise Constant Functions

We can now describe the class of “simple” functions which we can integrate very easily.

Definition 11.2.1 *(Constant functions)* Let X be a subset of $\mathbf{R}$, and let $f : X \to \mathbf{R}$ be a function. We say that f is *constant* iff there exists a real number c such that $f(x) = c$ for all $x \in X$. If E is a subset of X, we say that f is *constant on* E if the restriction $f|_E$ of f to E is constant, in other words there exists a real number c such that $f(x) = c$ for all $x \in E$. We refer to c as the *constant value* of f on E.

Remark 11.2.2 If E is a non-empty set, then a function f which is constant on E can have only one constant value; it is not possible for a function to always equal 3 on E while simultaneously always equalling 4. However, if E is empty, every real number c is a constant value for f on E (why?).

Definition 11.2.3 *(Piecewise constant functions I)* Let I be a bounded interval, let $f : I \to \mathbf{R}$ be a function, and let $\mathbf{P}$ be a partition of I. We say that f is *piecewise constant with respect to* $\mathbf{P}$ if for every $J \in \mathbf{P}$, f is constant on J.

Example 11.2.4 The function $f : [1, 6] \to \mathbf{R}$ defined by

$$f(x) = \begin{cases} 7 & \text{if } 1 \leq x < 3 \\ 4 & \text{if } x = 3 \\ 5 & \text{if } 3 < x < 6 \\ 2 & \text{if } x = 6 \end{cases}$$

is piecewise constant with respect to the partition $\{[1, 3), \{3\}, (3, 6), \{6\}\}$ of $[1, 6]$. Note that it is also piecewise constant with respect to some other partitions as well; for instance, it is piecewise constant with respect to the partition $\{[1, 2), \{2\}, (2, 3), \{3\}, (3, 5), [5, 6), \{6\}, \emptyset\}$.

Definition 11.2.5 *(Piecewise constant functions II)* Let I be a bounded interval, and let $f: I \to \mathbf{R}$ be a function. We say that f is *piecewise constant on I* if there exists a partition $\mathbf{P}$ of I such that f is piecewise constant with respect to $\mathbf{P}$.

Example 11.2.6 The function used in the previous example is piecewise constant on $[1, 6]$. Also, every constant function on a bounded interval I is automatically piecewise constant (why?).

Lemma 11.2.7 *Let I be a bounded interval, let $\mathbf{P}$ be a partition of I, and let $f: I \to \mathbf{R}$ be a function which is piecewise constant with respect to $\mathbf{P}$. Let $\mathbf{P}'$ be a partition of I which is finer than $\mathbf{P}$. Then f is also piecewise constant with respect to $\mathbf{P}'$.*

Proof See Exercise 11.2.1. □

The space of piecewise constant functions is closed under algebraic operations:

Lemma 11.2.8 *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ and $g: I \to \mathbf{R}$ be piecewise constant functions on I. Then the functions $f + g$, $f - g$, $\max(f, g)$ and fg are also piecewise constant functions on I. Here of course $\max(f, g) : I \to \mathbf{R}$ is the function $\max(f, g)(x) := \max(f(x), g(x))$. If g does not vanish anywhere on I (i.e., $g(x) \neq 0$ for all $x \in I$), then f/g is also a piecewise constant function on I.*

Proof See Exercise 11.2.2. □

We are now ready to integrate piecewise constant functions. We begin with a temporary definition of an integral with respect to a partition.

Definition 11.2.9 *(Piecewise constant integral I)* Let I be a bounded interval, let $\mathbf{P}$ be a partition of I. Let $f: I \to \mathbf{R}$ be a function which is piecewise constant with respect to $\mathbf{P}$. Then we define the *piecewise constant integral $p.c. \int_{[\mathbf{P}]} f$* of f with respect to the partition $\mathbf{P}$ by the formula

$$p.c. \int_{[\mathbf{P}]} f := \sum_{J \in \mathbf{P}} c_J |J|,$$

where for each J in $\mathbf{P}$, we let c_J be the constant value of f on J.

Remark 11.2.10 This definition seems like it could be ill-defined, because if J is empty then every number c_J can be the constant value of f on J, but fortunately in such cases $|J|$ is zero and so the choice of c_J is irrelevant. The notation $p.c. \int_{[\mathbf{P}]} f$ is rather artificial, but we shall only need it temporarily, en route to a more useful definition. Note that since $\mathbf{P}$ is finite, the sum $\sum_{J \in \mathbf{P}} c_J |J|$ is always well-defined (it is never divergent or infinite).

Remark 11.2.11 The piecewise constant integral corresponds intuitively to one's notion of area, given that the area of a rectangle ought to be the product of the lengths of the sides. (Of course, if f is negative somewhere, then the "area" $c_J |J|$ would also be negative.)

Example 11.2.12 Let $f\colon [1,4] \to \mathbf{R}$ be the function

$$f(x) = \begin{cases} 2 & \text{if } 1 \le x < 3 \\ 4 & \text{if } x = 3 \\ 6 & \text{if } 3 < x \le 4 \end{cases}$$

and let $\mathbf{P} := \{[1,3), \{3\}, (3,4]\}$. Then

$$\begin{aligned} p.c. \int_{[\mathbf{P}]} f &= c_{[1,3)}|[1,3)| + c_{\{3\}}|\{3\}| + c_{(3,4]}|(3,4]| \\ &= 2 \times 2 + 4 \times 0 + 6 \times 1 \\ &= 10. \end{aligned}$$

Alternatively, if we let $\mathbf{P}' := \{[1,2), [2,3), \{3\}, (3,4], \emptyset\}$ then

$$\begin{aligned} p.c. \int_{[\mathbf{P}']} f &= c_{[1,2)}|[1,2)| + c_{[2,3)}|[2,3)| + c_{\{3\}}|\{3\}| \\ &\quad + c_{(3,4]}|(3,4]| + c_{\emptyset}|\emptyset| \\ &= 2 \times 1 + 2 \times 1 + 4 \times 0 + 6 \times 1 + c_{\emptyset} \times 0 \\ &= 10. \end{aligned}$$

This example suggests that this integral does not really depend on what partition you pick, so long as your function is piecewise constant with respect to that partition. That is indeed true:

Proposition 11.2.13 (Piecewise constant integral is independent of partition) *Let I be a bounded interval, and let $f\colon I \to \mathbf{R}$ be a function. Suppose that $\mathbf{P}$ and $\mathbf{P}'$ are partitions of I such that f is piecewise constant both with respect to $\mathbf{P}$ and with respect to $\mathbf{P}'$. Then $p.c. \int_{[\mathbf{P}]} f = p.c. \int_{[\mathbf{P}']} f$.*

Proof See Exercise 11.2.3. □

Because of this proposition, we can now make the following definition:

Definition 11.2.14 *(Piecewise constant integral II)* Let I be a bounded interval, and let $f\colon I \to \mathbf{R}$ be a piecewise constant function on I. We define the *piecewise constant integral $p.c. \int_I f$* by the formula

$$p.c. \int_I f := p.c. \int_{[\mathbf{P}]} f,$$

where $\mathbf{P}$ is any partition of I with respect to which f is piecewise constant. (Note that Proposition 11.2.13 tells us that the precise choice of this partition is irrelevant.)

Example 11.2.15 If f is the function given in Example 11.2.12, then $p.c. \int_{[1,4]} f = 10$.

We now give some basic properties of the piecewise constant integral. These laws will eventually be superceded by the corresponding laws for the Riemann integral (Theorem 11.4.1).

Theorem 11.2.16 (Laws of integration) *Let I be a bounded interval, and let $f\colon I \to \mathbf{R}$ and $g\colon I \to \mathbf{R}$ be piecewise constant functions on I.*

(a) We have $p.c. \int_I (f+g) = p.c. \int_I f + p.c. \int_I g$.
(b) For any real number c, we have $p.c. \int_I (cf) = c(p.c. \int_I f)$.
(c) We have $p.c. \int_I (f-g) = p.c. \int_I f - p.c. \int_I g$.
(d) If $f(x) \geq 0$ for all $x \in I$, then $p.c. \int_I f \geq 0$.
(e) If $f(x) \geq g(x)$ for all $x \in I$, then $p.c. \int_I f \geq p.c. \int_I g$.
(f) If f is the constant function $f(x) = c$ for all x in I, then $p.c. \int_I f = c|I|$.
(g) Let J be a bounded interval containing I (i.e., $I \subseteq J$), and let $F\colon J \to \mathbf{R}$ be the function

$$F(x) := \begin{cases} f(x) & \text{if } x \in I \\ 0 & \text{if } x \notin I \end{cases}$$

Then F is piecewise constant on J, and $p.c. \int_J F = p.c. \int_I f$.
(h) Suppose that $\{J, K\}$ is a partition of I into two intervals J and K. Then the functions $f|_J : J \to \mathbf{R}$ and $f|_K : K \to \mathbf{R}$ are piecewise constant on J and K respectively, and we have

$$p.c. \int_I f = p.c. \int_J f|_J + p.c. \int_K f|_K.$$

Proof See Exercise 11.2.4. □

This concludes our integration of piecewise constant functions. We now turn to the question of how to integrate bounded functions.

— Exercises —

Exercise 11.2.1 Prove Lemma 11.2.7.

Exercise 11.2.2 Prove Lemma 11.2.8. (*Hint:* use Lemmas 11.1.18 and 11.2.7 to make f and g piecewise constant with respect to the *same* partition of I.)

Exercise 11.2.3 Prove Proposition 11.2.13. (*Hint:* first use Theorem 11.1.13 to show that both integrals are equal to $p.c. \int_{[\mathbf{P}\#\mathbf{P}']} f$.)

Exercise 11.2.4 Prove Theorem 11.2.16. (*Hint:* you can use earlier parts of the theorem to prove some of the later parts of the theorem. See also the hint to Exercise 11.2.2.)

11.3 Upper and Lower Riemann Integrals

Now let $f\colon I \to \mathbf{R}$ be a bounded function defined on a bounded interval I. We want to define the Riemann integral $\int_I f$. To do this we first need to define the notion of upper and lower Riemann integrals $\overline{\int}_I f$ and $\underline{\int}_I f$. These notions are related to the Riemann integral in much the same way that the lim sup and lim inf of a sequence are related to the limit of that sequence.

Definition 11.3.1 *(Majorization of functions)* Let $f: I \to \mathbf{R}$ and $g: I \to \mathbf{R}$. We say that g *majorizes* f *on* I if we have $g(x) \geq f(x)$ for all $x \in I$, and that g *minorizes* f *on* I if $g(x) \leq f(x)$ for all $x \in I$.

The idea of the Riemann integral is to try to integrate a function by first majorizing or minorizing that function by a piecewise constant function (which we already know how to integrate).

Definition 11.3.2 *(Upper and lower Riemann integrals)* Let $f: I \to \mathbf{R}$ be a bounded function defined on a bounded interval I. We define the *upper Riemann integral* $\overline{\int}_I f$ by the formula

$$\overline{\int}_I f := \inf\{p.c. \int_I g : g \text{ is a p.c. function on } I \text{ which majorizes } f\}$$

and the *lower Riemann integral* $\underline{\int}_I f$ by the formula

$$\underline{\int}_I f := \sup\{p.c. \int_I g : g \text{ is a p.c. function on } I \text{ which minorizes } f\}.$$

We give a crude but useful bound on the lower and upper integral:

Lemma 11.3.3 *Let $f: I \to \mathbf{R}$ be a function on a bounded interval I which is bounded by some real number M, i.e., $-M \leq f(x) \leq M$ for all $x \in I$. Then we have*

$$-M|I| \leq \underline{\int}_I f \leq \overline{\int}_I f \leq M|I|.$$

In particular, both the lower and upper Riemann integrals are real numbers (i.e., they are not infinite).

Proof The function $g: I \to \mathbf{R}$ defined by $g(x) = M$ is constant, hence piecewise constant, and majorizes f; thus $\overline{\int}_I f \leq p.c. \int_I g = M|I|$ by definition of the upper Riemann integral. A similar argument gives $-M|I| \leq \underline{\int}_I f$. Finally, we have to show that $\underline{\int}_I f \leq \overline{\int}_I f$. Let g be any piecewise constant function majorizing f, and let h be any piecewise constant function minorizing f. Then g majorizes h, and hence $p.c. \int_I h \leq p.c. \int_I g$. Taking suprema in h, we obtain that $\underline{\int}_I f \leq p.c. \int_I g$. Taking infima in g, we thus obtain $\underline{\int}_I f \leq \overline{\int}_I f$, as desired. □

We now know that the upper Riemann integral is always at least as large as the lower Riemann integral. If the two integrals match, then we can define the Riemann integral:

Definition 11.3.4 *(Riemann integral)* Let $f: I \to \mathbf{R}$ be a bounded function on a bounded interval I. If $\underline{\int}_I f = \overline{\int}_I f$, then we say that f is *Riemann integrable on I* and define

$$\int_I f := \underline{\int}_I f = \overline{\int}_I f.$$

If the upper and lower Riemann integrals are unequal, we say that f is not Riemann integrable.

Remark 11.3.5 Compare this definition to the relationship between the lim sup, lim inf, and limit of a sequence a_n that was established in Proposition 6.4.12(f); the lim sup is always greater than or equal to the lim inf, but they are only equal when the sequence converges, and in this case they are both equal to the limit of the sequence. The definition given above may differ from the definition you may have encountered in your calculus courses, based on Riemann sums. However, the two definitions turn out to be equivalent; this is the purpose of the next section.

Remark 11.3.6 Note that we do not consider unbounded functions to be Riemann integrable; an integral involving such functions is known as an *improper integral*. It is possible to still evaluate such integrals using more sophisticated integration methods (such as the Lebesgue integral); we shall do this in Chap. 8.

The Riemann integral is consistent with (and supercedes) the piecewise constant integral:

Lemma 11.3.7 *Let $f: I \to \mathbf{R}$ be a piecewise constant function on a bounded interval I. Then f is Riemann integrable, and $\int_I f = p.c. \int_I f$.*

Proof See Exercise 11.3.3. □

Remark 11.3.8 Because of this lemma, we will not refer to the piecewise constant integral $p.c. \int_I$ again, and just use the Riemann integral $\int_I$ throughout (until this integral is itself superceded by the Lebesgue integral in Chapter 8). We observe one special case of Lemma 11.3.7: if I is a point or the empty set, then $\int_I f = 0$ for all functions $f: I \to \mathbf{R}$. (Note that all such functions are automatically constant.)

We have just shown that every piecewise constant function is Riemann integrable. However, the Riemann integral is more general and can integrate a wider class of functions; we shall see this shortly. For now, we connect the Riemann integral we have just defined to the concept of a *Riemann sum*, which you may have seen in other treatments of the Riemann integral.

Definition 11.3.9 *(Riemann sums)* Let $f: I \to \mathbf{R}$ be a bounded function on a bounded interval I, and let $\mathbf{P}$ be a partition of I. We define the *upper Riemann sum $U(f, \mathbf{P})$* and the *lower Riemann sum $L(f, \mathbf{P})$* by

$$U(f, \mathbf{P}) := \sum_{J \in \mathbf{P}: J \neq \emptyset} (\sup_{x \in J} f(x))|J|$$

and

$$L(f, \mathbf{P}) := \sum_{J \in \mathbf{P}: J \neq \emptyset} (\inf_{x \in J} f(x))|J|.$$

Remark 11.3.10 The restriction $J \neq \emptyset$ is required because the quantities $\inf_{x \in J} f(x)$ and $\sup_{x \in J} f(x)$ are infinite (or negative infinite) if J is empty.

We now connect these Riemann sums to the upper and lower Riemann integral.

Lemma 11.3.11 *Let $f: I \to \mathbf{R}$ be a bounded function on a bounded interval I, and let g be a function which majorizes f and which is piecewise constant with respect to some partition $\mathbf{P}$ of I. Then*

$$p.c. \int_I g \geq U(f, \mathbf{P}).$$

Similarly, if h is a function which minorizes f and is piecewise constant with respect to $\mathbf{P}$, then

$$p.c. \int_I h \leq L(f, \mathbf{P}).$$

Proof See Exercise 11.3.4. □

Proposition 11.3.12 *Let $f: I \to \mathbf{R}$ be a bounded function on a bounded interval I. Then*

$$\overline{\int}_I f = \inf\{U(f, \mathbf{P}) : \mathbf{P} \textit{ is a partition of } I\}$$

and

$$\underline{\int}_I f = \sup\{L(f, \mathbf{P}) : \mathbf{P} \textit{ is a partition of } I\}$$

Proof See Exercise 11.3.5. □

— Exercises —

Exercise 11.3.1 Let $f: I \to \mathbf{R}$, $g: I \to \mathbf{R}$, and $h: I \to \mathbf{R}$ be functions. Show that if f majorizes g and g majorizes h, then f majorizes h. Show that if f and g majorize each other, then they must be equal.

Exercise 11.3.2 Let $f: I \to \mathbf{R}$, $g: I \to \mathbf{R}$, and $h: I \to \mathbf{R}$ be functions. If f majorizes g, is it true that $f + h$ majorizes $g + h$? Is it true that $f \cdot h$ majorizes $g \cdot h$? If c is a real number, is it true that cf majorizes cg?

Exercise 11.3.3 Prove Lemma 11.3.7.

Exercise 11.3.4 Prove Lemma 11.3.11.

Exercise 11.3.5 Prove Proposition 11.3.12. (*Hint:* you will need Lemma 11.3.11, even though this Lemma will only do half of the job.)

11.4 Basic Properties of the Riemann Integral

Just as we did with limits, series, and derivatives, we now give the basic laws for manipulating the Riemann integral. These laws will eventually be superceded by the corresponding laws for the Lebesgue integral (Proposition 8.3.3).

Theorem 11.4.1 (Laws of Riemann integration) *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ and $g: I \to \mathbf{R}$ be Riemann integrable functions on I.*

(a) The function $f+g$ is Riemann integrable, and we have $\int_I (f+g) = \int_I f + \int_I g$.
(b) For any real number c, the function cf is Riemann integrable, and we have $\int_I (cf) = c(\int_I f)$.
(c) The function $f-g$ is Riemann integrable, and we have $\int_I (f-g) = \int_I f - \int_I g$.
(d) If $f(x) \geq 0$ for all $x \in I$, then $\int_I f \geq 0$.
(e) If $f(x) \geq g(x)$ for all $x \in I$, then $\int_I f \geq \int_I g$.
(f) If f is the constant function $f(x) = c$ for all x in I, then $\int_I f = c|I|$.
(g) Let J be a bounded interval containing I (i.e., $I \subseteq J$), and let $F: J \to \mathbf{R}$ be the function

$$F(x) := \begin{cases} f(x) & \text{if } x \in I \\ 0 & \text{if } x \notin I \end{cases}$$

Then F is Riemann integrable on J, and $\int_J F = \int_I f$.
(h) Suppose that $\{J, K\}$ is a partition of I into two intervals J and K. Then the functions $f|_J : J \to \mathbf{R}$ and $f|_K : K \to \mathbf{R}$ are Riemann integrable on J and K, respectively, and we have

$$\int_I f = \int_J f|_J + \int_K f|_K.$$

Proof See Exercise 11.4.1. □

Remark 11.4.2 We often abbreviate $\int_J f|_J$ as $\int_J f$, even though f is really defined on a larger domain than just J. We also observe from Theorem 11.4.1(h) and Remark 11.3.8 that if $f: [a, b] \to \mathbf{R}$ is Riemann integrable on a closed interval $[a, b]$, then $\int_{[a,b]} f = \int_{(a,b]} f = \int_{[a,b)} f = \int_{(a,b)} f$.

Theorem 11.4.1 asserts that the sum or difference of any two Riemann integrable functions is Riemann integrable, as is any scalar multiple cf of a Riemann integrable function f. We now give some further ways to create Riemann integrable functions.

Theorem 11.4.3 (Max and min preserve integrability) *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ and $g: I \to \mathbf{R}$ be a Riemann integrable function. Then the functions* $\max(f, g) : I \to \mathbf{R}$ *and* $\min(f, g) : I \to \mathbf{R}$ *defined by* $\max(f, g)(x) := \max(f(x), g(x))$ *and* $\min(f, g)(x) := \min(f(x), g(x))$ *are also Riemann integrable.*

Proof We shall just prove the claim for $\max(f, g)$, the case of $\min(f, g)$ being similar. First note that since f and g are bounded, then $\max(f, g)$ is also bounded.

Let $\varepsilon > 0$. Since $\int_I f = \underline{\int}_I f$, there exists a piecewise constant function $\underline{f} : I \to \mathbf{R}$ which minorizes f on I such that

$$\int_I \underline{f} \geq \int_I f - \varepsilon.$$

Similarly we can find a piecewise constant $\underline{g} : I \to \mathbf{R}$ which minorizes g on I such that

$$\int_I \underline{g} \geq \int_I g - \varepsilon,$$

and we can find piecewise functions $\overline{f}, \overline{g}$ which majorize f, g respectively on I such that

$$\int_I \overline{f} \leq \int_I f + \varepsilon$$

and

$$\int_I \overline{g} \leq \int_I g + \varepsilon.$$

In particular, if $h\colon I \to \mathbf{R}$ denotes the function

$$h := (\overline{f} - \underline{f}) + (\overline{g} - \underline{g})$$

we have

$$\int_I h \leq 4\varepsilon.$$

On the other hand, $\max(\underline{f}, \underline{g})$ is a piecewise constant function on I (why?) which minorizes $\max(f, g)$ (why?), while $\max(\overline{f}, \overline{g})$ is similarly a piecewise constant function on I which majorizes $\max(f, g)$. Thus

$$\int_I \max(\underline{f}, \underline{g}) \leq \underline{\int}_I \max(f, g) \leq \overline{\int}_I \max(f, g) \leq \int_I \max(\overline{f}, \overline{g}),$$

and so

$$0 \leq \overline{\int}_I \max(f, g) - \underline{\int}_I \max(f, g) \leq \int_I \max(\overline{f}, \overline{g}) - \max(\underline{f}, \underline{g}).$$

But we have

$$\overline{f}(x) = \underline{f}(x) + (\overline{f} - \underline{f})(x) \leq \underline{f}(x) + h(x)$$

and similarly

$$\overline{g}(x) = \underline{g}(x) + (\overline{g} - \underline{g})(x) \leq \underline{g}(x) + h(x)$$

and thus

$$\max(\overline{f}(x), \overline{g}(x)) \leq \max(\underline{f}(x), \underline{g}(x)) + h(x).$$

Inserting this into the previous inequality, we obtain

$$0 \leq \overline{\int}_I \max(f, g) - \underline{\int}_I \max(f, g) \leq \int_I h \leq 4\varepsilon.$$

To summarize, we have shown that

$$0 \leq \overline{\int}_I \max(f, g) - \underline{\int}_I \max(f, g) \leq 4\varepsilon$$

for every ε. Since $\overline{\int}_I \max(f, g) - \underline{\int}_I \max(f, g)$ does not depend on ε, we thus see that

$$\overline{\int}_I \max(f, g) - \underline{\int}_I \max(f, g) = 0$$

and hence that $\max(f, g)$ is Riemann integrable. □

Corollary 11.4.4 (Absolute values preserve Riemann integrability) *Let I be a bounded interval. If $f\colon I \to \mathbf{R}$ is a Riemann integrable function, then the positive part $f_+ := \max(f, 0)$ and the negative part $f_- := \min(f, 0)$ are also Riemann integrable on I. Also, the absolute value $|f|$ defined by $|f|(x) := |f(x)|$ is also Riemann integrable on I. (This latter claim follows from the observation that $|f| = f_+ - f_-$.)*

Theorem 11.4.5 (Products preserve Riemann integrability) *Let I be a bounded interval. If $f\colon I \to \mathbf{R}$ and $g\colon I \to \mathbf{R}$ are Riemann integrable, then $fg\colon I \to \mathbf{R}$ is also Riemann integrable.*

Proof This one is a little trickier. We split $f = f_+ + f_-$ and $g = g_+ + g_-$ into positive and negative parts; by Corollary 11.4.4, the functions f_+, f_-, g_+, g_- are Riemann integrable. Since

$$fg = f_+g_+ + f_+g_- + f_-g_+ + f_-g_-$$

then it suffices to show that the functions f_+g_+,f_+g_-, f_-g_+, f_-g_- are individually Riemann integrable. We will just show this for f_+g_+; the other three are similar.

Since f_+ and g_+ are bounded and positive, there are $M_1, M_2 > 0$ such that

$$0 \leq f_+(x) \leq M_1 \text{ and } 0 \leq g_+(x) \leq M_2$$

for all $x \in I$. Now let $\varepsilon > 0$ be arbitrary. Then, as in the proof of Theorem 11.4.3, we can find a piecewise constant function $\underline{f_+}$ minorizing f_+ on I, and a piecewise constant function $\overline{f_+}$ majorizing f_+ on I, such that

$$\int_I \overline{f_+} \leq \int_I f_+ + \varepsilon$$

and

$$\int_I \underline{f_+} \geq \int_I f_+ - \varepsilon.$$

Note that $\underline{f_+}$ may be negative at places, but we can fix this by replacing $\underline{f_+}$ by $\overline{\max(\underline{f_+}, 0)}$, since this still minorizes f_+ (why?) and still has integral greater than or equal to $\int_I f_+ - \varepsilon$ (why?). So without loss of generality we may assume that $\underline{f_+}(x) \geq 0$ for all $x \in I$. Similarly we may assume that $\overline{f_+}(x) \leq M_1$ for all $x \in I$; thus

$$0 \leq \underline{f_+}(x) \leq f_+(x) \leq \overline{f_+}(x) \leq M_1$$

for all $x \in I$.

Similar reasoning allows us to find piecewise constant $\underline{g_+}$ minorizing g_+, and $\overline{g_+}$ majorizing g_+, such that

$$\int_I \overline{g_+} \leq \int_I g_+ + \varepsilon$$

and

$$\int_I \underline{g_+} \geq \int_I g_+ - \varepsilon,$$

and

$$0 \leq \underline{g_+}(x) \leq g_+(x) \leq \overline{g_+}(x) \leq M_2$$

for all $x \in I$.

Notice that $\underline{f_+ g_+}$ is piecewise constant and minorizes $f_+ g_+$, while $\overline{f_+}\overline{g_+}$ is piecewise constant and majorizes $f_+ g_+$. Thus

$$0 \leq \overline{\int}_I f_+ g_+ - \underline{\int}_I f_+ g_+ \leq \int_I \overline{f_+}\overline{g_+} - \underline{f_+ g_+}.$$

However, we have

$$\overline{f_+}(x)\overline{g_+}(x) - \underline{f_+}(x)\underline{g_+}(x) = \overline{f_+}(x)(\overline{g_+} - \underline{g_+})(x) + \underline{g_+}(x)(\overline{f_+} - \underline{f_+})(x)$$

$$\leq M_1(\overline{g_+} - \underline{g_+})(x) + M_2(\overline{f_+} - \underline{f_+})(x)$$

for all $x \in I$, and thus

$$0 \le \overline{\int}_I f_+g_+ - \underline{\int}_I f_+g_+ \le M_1 \int_I (\overline{g_+} - \underline{g_+}) + M_2 \int_I (\overline{f_+} - \underline{f_+})$$

$$\le M_1(2\varepsilon) + M_2(2\varepsilon).$$

Again, since ε was arbitrary, we can conclude that f_+g_+ is Riemann integrable, as before. Similar argument show that f_+g_-, f_-g_+, f_-g_- are Riemann integrable; combining them we obtain that fg is Riemann integrable. □

— Exercises —

Exercise 11.4.1 Prove Theorem 11.4.1. (*Hint:* you may find Theorem 11.2.16 to be useful. For part (b): First do the case $c > 0$. Then do the case $c = -1$ and $c = 0$ separately. Using these cases, deduce the case of $c < 0$. You can use earlier parts of the theorem to prove later ones.)

Exercise 11.4.2 Let I be a bounded interval, let $f: I \to \mathbf{R}$ be a Riemann integrable function, and let $\mathbf{P}$ be a partition of I. Show that

$$\int_I f = \sum_{J \in \mathbf{P}} \int_J f.$$

Exercise 11.4.3 Without repeating all the computations in the above proofs, give a short explanation as to why the remaining cases of Theorem 11.4.3 and Theorem 11.4.5 follow automatically from the cases presented in the text. (*Hint:* from Theorem 11.4.1 we know that if f is Riemann integrable, then so is $-f$.)

11.5 Riemann Integrability of Continuous Functions

We have already said a lot about Riemann integrable functions so far, but we have not yet actually produced any such functions other than the piecewise constant ones. Now we rectify this by showing that a large class of useful functions are Riemann integrable. We begin with the uniformly continuous functions.

Theorem 11.5.1 *Let I be a bounded interval, and let f be a function which is uniformly continuous on I. Then f is Riemann integrable.*

Proof From Proposition 9.9.15 we see that f is bounded. Now we have to show that $\underline{\int}_I f = \overline{\int}_I f$.

If I is a point or the empty set then the theorem is trivial, so let us assume that I is one of the four intervals $[a, b]$, (a, b), $(a, b]$, or $[a, b)$ for some real numbers $a < b$.

Let $\varepsilon > 0$ be arbitrary. By uniform continuity, there exists a $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $x, y \in I$ are such that $|x - y| < \delta$. By the Archimedean principle, there exists an integer $N > 0$ such that $(b - a)/N < \delta$.

Note that we can partition I into N intervals $J_1, \ldots, J_N$, each of length $(b - a)/N$. (How? One has to treat each of the cases $[a, b]$, (a, b), $(a, b]$, $[a, b)$ slightly differently.) By Proposition 11.3.12, we thus have

$$\overline{\int}_I f \leq \sum_{k=1}^N (\sup_{x \in J_k} f(x))|J_k|$$

and

$$\underline{\int}_I f \geq \sum_{k=1}^N (\inf_{x \in J_k} f(x))|J_k|$$

so in particular

$$\overline{\int}_I f - \underline{\int}_I f \leq \sum_{k=1}^N (\sup_{x \in J_k} f(x) - \inf_{x \in J_k} f(x))|J_k|.$$

However, we have $|f(x) - f(y)| < \varepsilon$ for all $x, y \in J_k$, since $|J_k| = (b-a)/N < \delta$. In particular we have

$$f(x) < f(y) + \varepsilon \text{ for all } x, y \in J_k.$$

Taking suprema in x, we obtain

$$\sup_{x \in J_k} f(x) \leq f(y) + \varepsilon \text{ for all } y \in J_k,$$

and then taking infima in y we obtain

$$\sup_{x \in J_k} f(x) \leq \inf_{y \in J_k} f(y) + \varepsilon.$$

Inserting this bound into our previous inequality, we obtain

$$\overline{\int}_I f - \underline{\int}_I f \leq \sum_{k=1}^N \varepsilon|J_k|,$$

but by Theorem 11.1.13 we thus have

$$\overline{\int}_I f - \underline{\int}_I f \leq \varepsilon(b-a).$$

But $\varepsilon > 0$ was arbitrary, while $(b-a)$ is fixed. Thus $\overline{\int}_I f - \underline{\int}_I f$ cannot be positive. By Lemma 11.3.3 and the definition of Riemann integrability we thus have that f is Riemann integrable. □

Combining Theorem 11.5.1 with Theorem 9.9.16, we thus obtain

Corollary 11.5.2 *Let $[a, b]$ be a closed interval, and let $f: [a, b] \to \mathbf{R}$ be continuous. Then f is Riemann integrable.*

Note that this Corollary is not true if $[a, b]$ is replaced by any other sort of interval, since it is not even guaranteed then that continuous functions are bounded. For instance, the function $f: (0, 1) \to \mathbf{R}$ defined by $f(x) := 1/x$ is continuous but not Riemann integrable. However, if we assume that a function is both continuous *and* bounded, we can recover Riemann integrability:

Proposition 11.5.3 *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ be both continuous and bounded. Then f is Riemann integrable on I.*

Proof If I is a point or an empty set then the claim is trivial; if I is a closed interval the claim follows from Corollary 11.5.2. So let us assume that I is of the form $(a, b]$, (a, b), or $[a, b)$ for some $a < b$.

We have a bound M for f, so that $-M \le f(x) \le M$ for all $x \in I$. Now let $0 < \varepsilon < (b - a)/2$ be a small number. The function f when restricted to the interval $[a + \varepsilon, b - \varepsilon]$ is continuous, and hence Riemann integrable by Corollary 11.5.2. In particular, we can find a piecewise constant function $h: [a + \varepsilon, b - \varepsilon] \to \mathbf{R}$ which majorizes f on $[a + \varepsilon, b - \varepsilon]$ such that

$$\int_{[a+\varepsilon,b-\varepsilon]} h \le \int_{[a+\varepsilon,b-\varepsilon]} f + \varepsilon.$$

Define $\tilde{h}: I \to \mathbf{R}$ by

$$\tilde{h}(x) := \begin{cases} h(x) & \text{if } x \in [a + \varepsilon, b - \varepsilon] \\ M & \text{if } x \in I \backslash [a + \varepsilon, b - \varepsilon] \end{cases}$$

Clearly $\tilde{h}$ is piecewise constant on I and majorizes f; by Theorem 11.2.16 we have

$$\int_I \tilde{h} = \varepsilon M + \int_{[a+\varepsilon,b-\varepsilon]} h + \varepsilon M \le \int_{[a+\varepsilon,b-\varepsilon]} f + (2M + 1)\varepsilon.$$

In particular we have

$$\overline{\int}_I f \le \int_{[a+\varepsilon,b-\varepsilon]} f + (2M + 1)\varepsilon.$$

A similar argument gives

$$\underline{\int}_I f \ge \int_{[a+\varepsilon,b-\varepsilon]} f - (2M + 1)\varepsilon$$

and hence

$$\overline{\int}_I f - \underline{\int}_I f \leq (4M + 2)\varepsilon.$$

But ε is arbitrary, and so we can argue as in the proof of Theorem 11.5.1 to conclude Riemann integrability. □

This gives a large class of Riemann integrable functions already; the bounded continuous functions. But we can expand this class a little more, to include the bounded *piecewise* continuous functions.

Definition 11.5.4 Let I be a bounded interval, and let $f: I \to \mathbf{R}$. We say that f is *piecewise continuous on I* iff there exists a partition $\mathbf{P}$ of I such that $f|_J$ is continuous on J for all $J \in \mathbf{P}$.

Example 11.5.5 The function $f: [1,3] \to \mathbf{R}$ defined by

$$F(x) := \begin{cases} x^2 & \text{if } 1 \leq x < 2 \\ 7 & \text{if } x = 2 \\ x^3 & \text{if } 2 < x \leq 3 \end{cases}$$

is not continuous on $[1,3]$, but it is piecewise continuous on $[1,3]$ (since it is continuous when restricted to $[1,2)$ or $\{2\}$ or $(2,3]$, and those three intervals partition $[1,3]$).

Proposition 11.5.6 *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ be both piecewise continuous and bounded. Then f is Riemann integrable.*

Proof See Exercise 11.5.1. □

— Exercises —

Exercise 11.5.1 Prove Proposition 11.5.6. (*Hint:* use Theorem 11.4.1(a) and (g).)

Exercise 11.5.2 Let $a < b$ be real numbers, and let $f: [a,b] \to \mathbf{R}$ be a continuous, non-negative function (so $f(x) \geq 0$ for all $x \in [a,b]$). Suppose that $\int_{[a,b]} f = 0$. Show that $f(x) = 0$ for all $x \in [a,b]$. (*Hint:* argue by contradiction.)

11.6 Riemann Integrability of Monotone Functions

In addition to piecewise continuous functions, another wide class of functions is Riemann integrable, namely the monotone functions. We give two instances of this:

Proposition 11.6.1 *Let $[a,b]$ be a closed and bounded interval and let $f: [a,b] \to \mathbf{R}$ be a monotone function. Then f is Riemann integrable on $[a,b]$.*

Remark 11.6.2 From Exercise 9.8.5 we know that there exist monotone functions which are not piecewise continuous, so this proposition is not subsumed by Proposition 11.5.6.

Proof Without loss of generality we may take f to be monotone increasing (instead of monotone decreasing). From Exercise 9.8.1 we know that f is bounded. Now let $N > 0$ be an integer, and partition $[a, b]$ into N half-open intervals $\{[a + \frac{b-a}{N} j, a + \frac{b-a}{N}(j + 1)) : 0 \leq j \leq N - 1\}$ of length $(b - a)/N$, together with the point $\{b\}$. Then by Proposition 11.3.12 we have

$$\overline{\int}_I f \leq \sum_{j=0}^{N-1} \left(\sup_{x \in [a + \frac{b-a}{N} j, a + \frac{b-a}{N}(j+1))} f(x) \right) \frac{b-a}{N},$$

(the point $\{b\}$ clearly giving only a zero contribution). Since f is monotone increasing, we thus have

$$\overline{\int}_I f \leq \sum_{j=0}^{N-1} f\left(a + \frac{b-a}{N}(j+1)\right) \frac{b-a}{N}.$$

Similarly we have

$$\underline{\int}_I f \geq \sum_{j=0}^{N-1} f\left(a + \frac{b-a}{N} j\right) \frac{b-a}{N}.$$

Thus we have

$$\overline{\int}_I f - \underline{\int}_I f \leq \sum_{j=0}^{N-1} \left(f\left(a + \frac{b-a}{N}(j+1)\right) - f\left(a + \frac{b-a}{N} j\right) \right) \frac{b-a}{N}.$$

Using telescoping series (Lemma 7.2.14) we thus have

$$\begin{aligned} \overline{\int}_I f - \underline{\int}_I f &\leq \left(f\left(a + \frac{b-a}{N}(N)\right) - f\left(a + \frac{b-a}{N} 0\right) \right) \frac{b-a}{N} \\ &= (f(b) - f(a)) \frac{b-a}{N}. \end{aligned}$$

But N was arbitrary, so we can conclude as in the proof of Theorem 11.5.1 that f is Riemann integrable. □

Corollary 11.6.3 *Let I be a bounded interval, and let $f : I \to \mathbf{R}$ be both monotone and bounded. Then f is Riemann integrable on I.*

Proof See Exercise 11.6.1. □

We now give the famous integral test for determining convergence of monotone decreasing series.

Proposition 11.6.4 (Integral test) *Let $f: [0, \infty) \to \mathbf{R}$ be a monotone decreasing function which is non-negative (i.e., $f(x) \geq 0$ for all $x \geq 0$). Then the sum $\sum_{n=0}^{\infty} f(n)$ is convergent if and only if $\sup_{N>0} \int_{[0,N]} f$ is finite.*

Proof See Exercise 11.6.3. □

Corollary 11.6.5 *Let p be a real number. Then $\sum_{n=1}^{\infty} \frac{1}{n^p}$ converges absolutely when $p > 1$ and diverges when $p \leq 1$.*

Proof See Exercise 11.6.5. □

— Exercises —

Exercise 11.6.1 Use Proposition 11.6.1 to prove Corollary 11.6.3. (*Hint:* adapt the proof of Proposition 11.5.3.)

Exercise 11.6.2 Formulate a reasonable notion of a piecewise monotone function, and then show that all bounded piecewise monotone functions are Riemann integrable.

Exercise 11.6.3 Prove Proposition 11.6.4. (*Hint:* what is the relationship between the sum $\sum_{n=1}^{N} f(n)$, the sum $\sum_{n=0}^{N-1} f(n)$, and the integral $\int_{[0,N]} f$?)

Exercise 11.6.4 Give examples to show that both directions of the integral test break down if f is not assumed to be monotone decreasing.

Exercise 11.6.5 Use Proposition 11.6.4 to prove Corollary 11.6.5. (For this exercise, you may use the second Fundamental Theorem of Calculus (Theorem 11.9.4); there is no circularity, because Corollary 11.6.5 is not used in the proof of that theorem.)

11.7 A Non-riemann Integrable Function

We have shown that there are large classes of bounded functions which are Riemann integrable. Unfortunately, there do exist bounded functions which are not Riemann integrable:

Proposition 11.7.1 *Let $f: [0, 1] \to \mathbf{R}$ be the discontinuous function*

$$f(x) := \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q} \end{cases}$$

considered in Example 9.3.21. Then f is bounded but not Riemann integrable.

Proof It is clear that f is bounded, so let us show that it is not Riemann integrable.

Let $\mathbf{P}$ be any partition of $[0, 1]$. For any $J \in \mathbf{P}$, observe that if J is not a point or the empty set, then

$$\sup_{x \in J} f(x) = 1$$

(by Proposition 5.4.14). In particular we have

$$\left(\sup_{x\in J} f(x)\right) |J| = |J|.$$

(Note this is also true when J is a point, since both sides are zero.) In particular we see that

$$U(f, \mathbf{P}) = \sum_{J\in\mathbf{P}: J\neq\emptyset} |J| = |[0,1]| = 1$$

by Theorem 11.1.13; note that the empty set does not contribute anything to the total length. In particular we have $\overline{\int}_{[0,1]} f = 1$, by Proposition 11.3.12.

A similar argument gives that

$$\inf_{x\in J} f(x) = 0$$

for all J (other than points or the empty set), and so

$$L(f, \mathbf{P}) = \sum_{J\in\mathbf{P}: J\neq\emptyset} 0 = 0.$$

In particular we have $\underline{\int}_{[0,1]} f = 0$, by Proposition 11.3.12. Thus the upper and lower Riemann integrals do not match, and so this function is not Riemann integrable. □

Remark 11.7.2 As you can see, it is only rather "artificial" bounded functions which are not Riemann integrable. Because of this, the Riemann integral is good enough for a large majority of cases. There are ways to generalize or improve this integral, though. One of these is the *Lebesgue integral*, which we will define in Chapter 8. Another is the *Riemann–Stieltjes integral* $\int_I f\, \mathrm{d}\alpha$, where $\alpha : I \to \mathbf{R}$ is a monotone increasing function, which we define in the next section.

11.8 The Riemann–Stieltjes Integral

Let I be a bounded interval, let $\alpha : I \to \mathbf{R}$ be a monotone increasing function, and let $f : I \to \mathbf{R}$ be a function. Then there is a generalization of the Riemann integral, known as the *Riemann–Stieltjes integral*. This integral is defined just like the Riemann integral, but with one twist: instead of taking the length $|J|$ of intervals J, we take the α-length $\alpha[J]$, defined as follows.

Definition 11.8.1 *(α-length)*] Let I be a bounded interval, let X be a interval that is closed (in the sense of Definition 9.1.15) containing I, and let $\alpha : X \to \mathbf{R}$ be a monotone increasing function (i.e., $\alpha(y) \geq \alpha(x)$ whenever $x, y \in X$ are such that $y \geq x$). Then we define the *α-length* $\alpha[I]$ of I by the following rules.

(i) If I is empty, then $\alpha[I] := 0$.

(ii) If $I = \{a\}$ is a point, then $\alpha[I] := \lim_{x\to a^+: x\in X} \alpha(x) - \lim_{x\to a^-: x\in X} \alpha(x)$, with the convention that $\lim_{x\to a^+: x\in X} \alpha(x)$ (resp. $\lim_{x\to a^-: x\in X} \alpha(x)$) is equal to $\alpha(a)$ when X is the right-endpoint (resp. left-endpoint) of X).

(iii) If $I = (a, b)$, set $\alpha[I] := \lim_{x\to b^-: x\in X} \alpha(x) - \lim_{x\to b^+: x\in X} \alpha(x)$.

(iv) If I is equal to $(a, b]$, $[a, b)$, or $[a, b]$, set $\alpha[I]$ equal to $\alpha((a, b)) + \alpha(\{b\})$, $\alpha(\{a\}) + \alpha((a, b))$, or $\alpha(\{a\}) + \alpha((a, b)) + \alpha(\{b\})$, respectively.

This definition is complicated, but note that in the special case where α is continuous, we have the simpler formula

$$\alpha[I] = \alpha(b) - \alpha(a) \tag{11.1}$$

whenever $a \le b$ and I is equal to (a, b), $(a, b]$, $[a, b)$, or $[a, b]$. Using this simplified formula, one can also define $\alpha[I]$ for other continuous functions that are not necessarily monotone increasing.

Example 11.8.2 Let $\alpha\colon [0, +\infty) \to \mathbf{R}$ be the function $\alpha(x) := x^2$. Then $\alpha[[2, 3]] = \alpha(3) - \alpha(2) = 9 - 4 = 5$, $\alpha[\{2\}] = 0$ and $\alpha[\emptyset] = 0$.

Example 11.8.3 Let $\alpha\colon \mathbf{R} \to \mathbf{R}$ be the identity function $\alpha(x) := x$. Then $\alpha[I] = |I|$ for all bounded intervals I (why?) Thus the notion of length is a special case of the notion of α-length.

We sometimes write $\alpha|_a^b$ or $\alpha(x)|_{x=a}^{x=b}$ instead of $\alpha[[a, b]]$.

One of the key theorems for the theory of the Riemann integral was Theorem 11.1.13, which concerned length and partitions, and in particular showed that $|I| = \sum_{J\in\mathbf{P}} |J|$ whenever $\mathbf{P}$ was a partition of I. We now generalize this slightly.

Lemma 11.8.4 *Let I be a bounded interval, let $\alpha : X \to \mathbf{R}$ be a monotone increasing or continuous function defined on some interval X is closed and which contains I, and let $\mathbf{P}$ be a partition of I. Then we have*

$$\alpha[I] = \sum_{J\in\mathbf{P}} \alpha[J].$$

Proof See Exercise 11.8.1. □

We can now define a generalization of Definition 11.2.9.

Definition 11.8.5 *(P.c. Riemann–Stieltjes integral)* Let I be a bounded interval, and let $\mathbf{P}$ be a partition of I. Let $\alpha : X \to \mathbf{R}$ be a monotone increasing or continuous function defined on some interval X which is closed and contains I, and let $f\colon I \to \mathbf{R}$ be a function which is piecewise constant with respect to $\mathbf{P}$. Then we define

$$p.c. \int_{[\mathbf{P}]} f \, d\alpha := \sum_{J\in\mathbf{P}} c_J \alpha[J]$$

where c_J is the constant value of f on J.

Example 11.8.6 Let $f\colon [1,3] \to \mathbf{R}$ be the function

$$f(x) = \begin{cases} 4 & \text{when } x \in [1,2) \\ 2 & \text{when } x \in [2,3], \end{cases}$$

let $\alpha\colon [0,+\infty) \to \mathbf{R}$ be the function $\alpha(x) := x^2$, and let $\mathbf{P}$ be the partition $\mathbf{P} := \{[1,2), [2,3]\}$. Then

$$\begin{aligned} p.c. \int_{[\mathbf{P}]} f \, \mathrm{d}\alpha &= c_{[1,2)}\alpha[[1,2)] + c_{[2,3]}\alpha[[2,3]] \\ &= 4(\alpha(2) - \alpha(1)) + 2(\alpha(3) - \alpha(2)) = 4 \times 3 + 2 \times 5 = 22. \end{aligned}$$

Example 11.8.7 Let $\alpha\colon \mathbf{R} \to \mathbf{R}$ be the identity function $\alpha(x) := x$. Then for any bounded interval I, any partition $\mathbf{P}$ of I, and any function f that is piecewise constant with respect to $\mathbf{P}$, we have $p.c. \int_{[\mathbf{P}]} f \, \mathrm{d}\alpha = p.c. \int_{[\mathbf{P}]} f$ (why?).

We can obtain an exact analogue of Proposition 11.2.13 by replacing all the integrals $p.c. \int_{[\mathbf{P}]} f$ in the proposition with $p.c. \int_{[\mathbf{P}]} f \, \mathrm{d}\alpha$ (Exercise 11.8.2). We can thus define $p.c. \int_I f \, \mathrm{d}\alpha$ for any piecewise constant function $f\colon I \to \mathbf{R}$ and any $\alpha : X \to \mathbf{R}$ defined on an interval that is closed and contains I, in analogy to before, by the formula

$$p.c. \int_I f \, \mathrm{d}\alpha := p.c. \int_{[\mathbf{P}]} f \, \mathrm{d}\alpha$$

for any partition $\mathbf{P}$ on I with respect to which f is piecewise constant.

Let us now assume that α is monotone increasing. This implies that $\alpha(I) \geq 0$ for all intervals in X (why?). From this one can easily verify that all the results from Theorem 11.2.16 continue to hold when the integrals $p.c. \int_I f$ are replaced by $p.c. \int_I f \, \mathrm{d}\alpha$, and the lengths $|I|$ are replaced by the α-lengths $\alpha(I)$; see Exercise 11.8.3.

We can then define upper and lower Riemann–Stieltjes integrals $\overline{\int}_I f \, \mathrm{d}\alpha$ and $\underline{\int}_I f \, \mathrm{d}\alpha$ whenever $f\colon I \to \mathbf{R}$ is bounded and α is defined on an interval that is closed and contains I, by the usual formulae

$$\overline{\int}_I f \, \mathrm{d}\alpha := \inf\{p.c. \int_I g \, \mathrm{d}\alpha : g \text{ is p.c. on } I \text{ and majorizes } f\}$$

and

$$\underline{\int}_I f \, \mathrm{d}\alpha := \sup\{p.c. \int_I g \, \mathrm{d}\alpha : g \text{ is p.c. on } I \text{ and minorizes } f\}.$$

We then say that f is *Riemann–Stieltjes integrable on I with respect to α* if the upper and lower Riemann–Stieltjes integrals match, in which case we set

$$\int_I f \, \mathrm{d}\alpha := \overline{\int}_I f \, \mathrm{d}\alpha = \underline{\int}_I f \, \mathrm{d}\alpha.$$

As before, when α is the identity function $\alpha(x) := x$ then the Riemann–Stieltjes integral is identical to the Riemann integral; thus the Riemann–Stieltjes integral is a generalization of the Riemann integral. (We shall see another comparison between the two integrals a little later, in Corollary 11.10.3.) Because of this, we sometimes write $\int_I f$ as $\int_I f \, \mathrm{d}x$ or $\int_I f(x) \, \mathrm{d}x$.

Most (but not all) of the remaining theory of the Riemann integral then can be carried over without difficulty, replacing Riemann integrals with Riemann–Stieltjes integrals and lengths with α-lengths. There are a couple results which break down; Theorem 11.4.1(g), Proposition 11.5.3, and Proposition 11.5.6 are not necessarily true when α is discontinuous at key places (e.g., if f and α are both discontinuous at the same point, then $\int_I f \, \mathrm{d}\alpha$ is unlikely to be defined). However, Theorem 11.5.1 is still true (Exercise 11.8.4).

— Exercises —

Exercise 11.8.1 Prove Lemma 11.8.4. (*Hint:* modify the proof of Theorem 11.1.13.)

Exercise 11.8.2 State and prove a version of Proposition 11.2.13 for the Riemann–Stieltjes integral.

Exercise 11.8.3 State and prove a version of Theorem 11.2.16 for the Riemann–Stieltjes integral.

Exercise 11.8.4 State and prove a version of Theorem 11.5.1 for the Riemann–Stieltjes integral.

Exercise 11.8.5 Let $\mathrm{sgn}\colon \mathbf{R} \to \mathbf{R}$ be the signum function

$$\mathrm{sgn}(x) := \begin{cases} 1 & \text{when } x > 0 \\ 0 & \text{when } x = 0 \\ -1 & \text{when } x < 0. \end{cases}$$

Let $f\colon [-1, 1] \to \mathbf{R}$ be a continuous function. Show that f is Riemann–Stieltjes integrable with respect to sgn, and that

$$\int_{[-1,1]} f \, \mathrm{d}\,\mathrm{sgn} = 2f(0).$$

(*Hint:* for every $\varepsilon > 0$, find piecewise constant functions majorizing and minorizing f whose Riemann–Stieltjes integral is ε-close to $2f(0)$.)

11.9 The Two Fundamental Theorems of Calculus

We now have enough machinery to connect integration and differentiation via the familiar fundamental theorem of calculus. Actually, there are two such theorems, one involving the derivative of the integral, and the other involving the integral of the derivative.

Theorem 11.9.1 (First Fundamental Theorem of Calculus) *Let $a < b$ be real numbers, and let $f\colon [a, b] \to \mathbf{R}$ be a Riemann integrable function. Let $F\colon [a, b] \to \mathbf{R}$ be the function*

$$F(x) := \int_{[a,x]} f.$$

Then F is continuous. Furthermore, if $x_0 \in [a, b]$ and f is continuous at x_0, then F is differentiable at x_0, and $F'(x_0) = f(x_0)$.

Proof Since f is Riemann integrable, it is bounded (by Definition 11.3.4). Thus we have some real number M such that $-M \leq f(x) \leq M$ for all $x \in [a, b]$.

Now let $x < y$ be two elements of $[a, b]$. Then notice that

$$F(y) - F(x) = \int_{[a,y]} f - \int_{[a,x]} f = \int_{[x,y]} f$$

by Theorem 11.4.1(h). By Theorem 11.4.1(e) we thus have

$$\int_{[x,y]} f \leq \int_{[x,y]} M = p.c. \int_{[x,y]} M = M(y - x)$$

and

$$\int_{[x,y]} f \geq \int_{[x,y]} -M = p.c. \int_{[x,y]} -M = -M(y - x)$$

and thus

$$|F(y) - F(x)| \leq M(y - x).$$

This is for $y > x$. By interchanging x and y we thus see that

$$|F(y) - F(x)| \leq M(x - y)$$

when $x > y$. Also, we have $F(y) - F(x) = 0$ when $x = y$. Thus in all three cases we have

$$|F(y) - F(x)| \leq M|x - y|.$$

This implies that F is uniformly continuous (in fact it is Lipschitz continuous, see Exercise 10.2.6), and hence continuous.

Now suppose that $x_0 \in [a, b]$, and f is continuous at x_0. Choose any $\varepsilon > 0$. Then by continuity, we can find a $\delta > 0$ such that $|f(x) - f(x_0)| \leq \varepsilon$ for all x in the interval $I := [x_0 - \delta, x_0 + \delta] \cap [a, b]$, or in other words

$$f(x_0) - \varepsilon \leq f(x) \leq f(x_0) + \varepsilon \text{ for all } x \in I.$$

We now show that

$$|F(y) - F(x_0) - f(x_0)(y - x_0)| \le \varepsilon |y - x_0|$$

for all $y \in I$, since Proposition 10.1.7 will then imply that F is differentiable at x_0 with derivative $F'(x_0) = f(x_0)$ as desired.

Now fix $y \in I$. There are three cases. If $y = x_0$, then $F(y) - F(x_0) - f(x_0)(y - x_0) = 0$ and so the claim is obvious. If $y > x_0$, then

$$F(y) - F(x_0) = \int_{[x_0,y]} f.$$

Since $x_0, y \in I$, and I is a connected set, then $[x_0, y]$ is a subset of I, and thus we have

$$f(x_0) - \varepsilon \le f(x) \le f(x_0) + \varepsilon \text{ for all } x \in [x_0, y],$$

and thus

$$(f(x_0) - \varepsilon)(y - x_0) \le \int_{[x_0,y]} f \le (f(x_0) + \varepsilon)(y - x_0)$$

and so in particular

$$|F(y) - F(x_0) - f(x_0)(y - x_0)| \le \varepsilon |y - x_0|$$

as desired. The case $y < x_0$ is similar and is left to the reader. □

Example 11.9.2 Recall in Exercise 9.8.5 that we constructed a monotone function $f: \mathbf{R} \to \mathbf{R}$ which was discontinuous at every rational and continuous everywhere else. By Proposition 11.6.1, this monotone function is Riemann integrable on $[0, 1]$. If we define $F: [0, 1] \to \mathbf{R}$ by $F(x) := \int_{[0,x]} f$, then F is a continuous function which is differentiable at every irrational number. On the other hand, F is non-differentiable at every rational number; see Exercise 11.9.1.

Informally, the first fundamental theorem of calculus asserts that

$$\left(\int_{[a,x]} f\right)'(x) = f(x)$$

given a certain number of assumptions on f. Roughly, this means that the derivative of an integral recovers the original function. Now we show the reverse, that the integral of a derivative recovers the original function.

Definition 11.9.3 *(Antiderivatives)* Let I be a bounded interval, and let $f: I \to \mathbf{R}$ be a function. We say that a function $F: I \to \mathbf{R}$ is an *antiderivative* of f if F is differentiable on I and $F'(x) = f(x)$ for all limit points x of I.

Theorem 11.9.4 (Second Fundamental Theorem of Calculus) *Let $a \leq b$ be real numbers, and let $f\colon [a, b] \to \mathbf{R}$ be a Riemann integrable function. If $F\colon [a, b] \to \mathbf{R}$ is an antiderivative of f, then*

$$\int_{[a,b]} f = F(b) - F(a).$$

Proof The claim is trivial for $a = b$, so suppose that $a < b$; in particular every point in $[a, b]$ is now a limit point. We will use Riemann sums. The idea is to show that

$$U(f, \mathbf{P}) \geq F(b) - F(a) \geq L(f, \mathbf{P})$$

for every partition $\mathbf{P}$ of $[a, b]$. The left inequality asserts that $F(b) - F(a)$ is a lower bound for $\{U(f, \mathbf{P}) : \mathbf{P} \text{ is a partition of } [a, b]\}$, while the right inequality asserts that $F(b) - F(a)$ is an upper bound for $\{L(f, \mathbf{P}) : \mathbf{P} \text{ is a partition of } [a, b]\}$. But by Proposition 11.3.12, this means that

$$\overline{\int}_{[a,b]} f \geq F(b) - F(a) \geq \underline{\int}_{[a,b]} f,$$

but since f is assumed to be Riemann integrable, both the upper and lower Riemann integral equal $\int_{[a,b]} f$. The claim follows.

We have to show the bound $U(f, \mathbf{P}) \geq F(b) - F(a) \geq L(f, \mathbf{P})$. We shall just show the first inequality $U(f, \mathbf{P}) \geq F(b) - F(a)$; the other inequality is similar.

Let $\mathbf{P}$ be a partition of $[a, b]$. From Lemma 11.8.4 (noting from Proposition 10.1.10 that F is continuous) we have

$$F(b) - F(a) = \sum_{J \in \mathbf{P}} F[J] = \sum_{J \in \mathbf{P} : J \neq \emptyset} F[J],$$

while from definition we have

$$U(f, \mathbf{P}) = \sum_{J \in \mathbf{P} : J \neq \emptyset} \sup_{x \in J} f(x) |J|.$$

Thus it will suffice to show that

$$F[J] \leq \sup_{x \in J} f(x) |J|$$

for all $J \in \mathbf{P}$ (other than the empty set).

When J is a point then the claim is clear, since both sides are zero. Now suppose that $J = [c, d]$, $(c, d]$, $[c, d)$, or (c, d) for some $c < d$. Then the left-hand side is $F[J] = F(d) - F(c)$. (Note that as F is continuous, we may use the simplified formula (11.1) for $F[J]$.) By the mean-value theorem, this is equal to $(d - c)F'(e)$

for some $e \in J$. But since $F'(e) = f(e)$, we thus have

$$F[J] = (d - c) f(e) = f(e)|J| \leq \sup_{x \in J} f(x)|J|$$

as desired. □

Of course, as you are all aware, one can use the second fundamental theorem of calculus to compute integrals relatively easily provided that you can find an antiderivative of the integrand f. Note that the first fundamental theorem of calculus ensures that every *continuous* Riemann integrable function has an antiderivative. For discontinuous functions, the situation is more complicated and is a graduate-level real analysis topic which will not be discussed here. Also, not every function with an antiderivative is Riemann integrable; as an example, consider the function $F: [-1, 1] \to \mathbf{R}$ defined by $F(x) := x^2 \sin(1/x^3)$ when $x \neq 0$, and $F(0) := 0$. Then F is differentiable everywhere (why?), so F' has an antiderivative, but F' is unbounded (why?), and so is not Riemann integrable.

We now pause to mention the infamous "$+C$" ambiguity in antiderivatives:

Lemma 11.9.5 *Let I be a bounded interval, and let $f: I \to \mathbf{R}$ be a function. Let $F: I \to \mathbf{R}$ and $G: I \to \mathbf{R}$ be two antiderivatives of f. Then there exists a real number C such that $F(x) = G(x) + C$ for all $x \in I$.*

Proof See Exercise 11.9.2. □

— Exercises —

Exercise 11.9.1 Let $f: [0, 1] \to \mathbf{R}$ be the function in Exercise 9.8.5. Show that for every rational number $q \in \mathbf{Q} \cap (0, 1)$, the function $F: [0, 1] \to \mathbf{R}$ defined by the formula $F(x) := \int_0^x f(y)\,dy$ is not differentiable at q.

Exercise 11.9.2 Prove Lemma 11.9.5. (*Hint:* apply the mean-value theorem, Corollary 10.2.9, or Proposition 10.3.3, to the function $F - G$. One can also prove this lemma using the second Fundamental theorem of calculus (how?), but one has to be careful since we do not assume f to be Riemann integrable.)

Exercise 11.9.3 Let $a < b$ be real numbers, and let $f: [a, b] \to \mathbf{R}$ be a monotone increasing function. Let $F: [a, b] \to \mathbf{R}$ be the function $F(x) := \int_{[a,x]} f$. Let x_0 be an element of (a, b). Show that F is differentiable at x_0 if and only if f is continuous at x_0. (*Hint:* one direction is taken care of by one of the fundamental theorems of calculus. For the other, consider left and right limits of f and argue by contradiction.)

11.10 Consequences of the Fundamental Theorems

We can now give a number of useful consequences of the fundamental theorems of calculus (beyond the obvious application, that one can now compute any integral for which an antiderivative is known). The first application is the familiar integration by parts formula.

Proposition 11.10.1 (Integration by parts formula) *Let $I = [a, b]$, and let $F: [a, b] \to \mathbf{R}$ and $G : [a, b] \to \mathbf{R}$ be differentiable functions on $[a, b]$ such that F' and G' are Riemann integrable on I. Then we have*

$$\int_{[a,b]} FG' = F(b)G(b) - F(a)G(a) - \int_{[a,b]} F'G.$$

Proof See Exercise 11.10.1. □

Next, we show that under certain circumstances, one can write a Riemann–Stieltjes integral as a Riemann integral. We begin with piecewise constant functions.

Theorem 11.10.2 *Let $\alpha : [a, b] \to \mathbf{R}$ be a monotone increasing function, and suppose that α is also differentiable on $[a, b]$, with α' being Riemann integrable. Let $f: [a, b] \to \mathbf{R}$ be a piecewise constant function on $[a, b]$. Then $f\alpha'$ is Riemann integrable on $[a, b]$, and*

$$\int_{[a,b]} f\, d\alpha = \int_{[a,b]} f\alpha'.$$

Proof Since f is piecewise constant, it is Riemann integrable, and since α' is also Riemann integrable, then $f\alpha'$ is Riemann integrable by Theorem 11.4.5.

Suppose that f is piecewise constant with respect to some partition $\mathbf{P}$ of $[a, b]$; without loss of generality we may assume that $\mathbf{P}$ does not contain the empty set. Then we have

$$\int_{[a,b]} f\, \mathrm{d}\alpha = p.c. \int_{[\mathbf{P}]} f\, \mathrm{d}\alpha = \sum_{J \in \mathbf{P}} c_J \alpha[J]$$

where c_J is the constant value of f on J. On the other hand, from Theorem 11.4.1(h) (generalized to partitions of arbitrary length—why is this generalization true?) we have

$$\int_{[a,b]} f\alpha' = \sum_{J \in \mathbf{P}} \int_J f\alpha' = \sum_{J \in \mathbf{P}} \int_J c_J\alpha' = \sum_{J \in \mathbf{P}} c_J \int_J \alpha'.$$

But by the second fundamental theorem of calculus (Theorem 11.9.4), $\int_J \alpha' = \alpha[J]$, and the claim follows. □

Corollary 11.10.3 *Let $\alpha : [a, b] \to \mathbf{R}$ be a monotone increasing function, and suppose that α is also differentiable on $[a, b]$, with α' being Riemann integrable. Let $f: [a, b] \to \mathbf{R}$ be a function which is Riemann–Stieltjes integrable with respect to α on $[a, b]$. Then $f\alpha'$ is Riemann integrable on $[a, b]$, and*

$$\int_{[a,b]} f\, d\alpha = \int_{[a,b]} f\alpha'.$$

Proof Note that since f and α' are bounded, then $f\alpha'$ must also be bounded. Also, since α is monotone increasing and differentable, α' is non-negative.

Let $\varepsilon > 0$. Then, we can find a piecewise constant function $\overline{f}$ majorizing f on $[a, b]$, and a piecewise constant function $\underline{f}$ minorizing f on $[a, b]$, such that

$$\int_{[a,b]} f\,\mathrm{d}\alpha - \varepsilon \le \int_{[a,b]} \underline{f}\,\mathrm{d}\alpha \le \int_{[a,b]} \overline{f}\,\mathrm{d}\alpha \le \int_{[a,b]} f\,\mathrm{d}\alpha + \varepsilon.$$

Applying Theorem 11.10.2, we obtain

$$\int_{[a,b]} f\,\mathrm{d}\alpha - \varepsilon \le \int_{[a,b]} \underline{f}\alpha' \le \int_{[a,b]} \overline{f}\,\alpha' \le \int_{[a,b]} f\,\mathrm{d}\alpha + \varepsilon.$$

Since α' is non-negative and $\underline{f}$ minorizes f, then $\underline{f}\alpha'$ minorizes $f\alpha'$. Thus $\underline{\int}_{[a,b]}\underline{f}\alpha' \le \underline{\int}_{[a,b]} f\alpha'$ (why?). Thus

$$\int_{[a,b]} f\,\mathrm{d}\alpha - \varepsilon \le \underline{\int}_{[a,b]} f\alpha'.$$

Similarly we have

$$\overline{\int}_{[a,b]} f\alpha' \le \int_{[a,b]} f\,\mathrm{d}\alpha + \varepsilon.$$

Since these statements are true for any $\varepsilon > 0$, we must have

$$\int_{[a,b]} f\,\mathrm{d}\alpha \le \underline{\int}_{[a,b]} f\alpha' \le \overline{\int}_{[a,b]} f\alpha' \le \int_{[a,b]} f\,\mathrm{d}\alpha$$

and the claim follows. □

Remark 11.10.4 Informally, Corollary 11.10.3 asserts that $f\,\mathrm{d}\alpha$ is essentially equivalent to $f\frac{\mathrm{d}\alpha}{}\mathrm{d}x\mathrm{d}x$, when α is differentiable. However, the advantage of the Riemann–Stieltjes integral is that it still makes sense even when α is not differentiable.

We now build up to the familiar change of variables formula. We first need a preliminary lemma.

Lemma 11.10.5 *[Change of variables formula I] Let $[a, b]$ be a closed interval, and let $\phi : [a, b] \to [\phi(a), \phi(b)]$ be a continuous monotone increasing function. Let $f : [\phi(a), \phi(b)] \to \mathbf{R}$ be a piecewise constant function on $[\phi(a), \phi(b)]$. Then $f \circ \phi : [a, b] \to \mathbf{R}$ is also piecewise constant on $[a, b]$, and*

$$\int_{[a,b]} f \circ \phi\, d\phi = \int_{[\phi(a),\phi(b)]} f.$$

Proof We give a sketch of the proof, leaving the gaps to be filled in Exercise 11.10.2. Let $\mathbf{P}$ be a partition of $[\phi(a), \phi(b)]$ such that f is piecewise constant with respect to

P; we may assume that **P** does not contain the empty set. For each $J \in \mathbf{P}$, let c_J be the constant value of f on J, thus

$$\int_{[\phi(a),\phi(b)]} f = \sum_{J\in\mathbf{P}} c_J |J|.$$

For each interval J, let $\phi^{-1}(J)$ be the set $\phi^{-1}(J) := \{x \in [a,b] : \phi(x) \in J\}$. Then $\phi^{-1}(J)$ is connected (why?) and is thus an interval. Furthermore, c_J is the constant value of $f \circ \phi$ on $\phi^{-1}(J)$ (why?). Thus, if we define $\mathbf{Q} := \{\phi^{-1}(J) : J \in \mathbf{P}\}$ (ignoring the fact that **Q** has been used to represent the rational numbers), then **Q** partitions $[a,b]$ (why?), and $f \circ \phi$ is piecewise constant with respect to **Q** (why?). Thus

$$\int_{[a,b]} f \circ \phi \, \mathrm{d}\phi = \int_{[\mathbf{Q}]} f \circ \phi \, \mathrm{d}\phi = \sum_{J\in\mathbf{P}} c_J \phi[\phi^{-1}(J)].$$

But $\phi[\phi^{-1}(J)] = |J|$ (why?), and the claim follows. □

Proposition 11.10.6 (Change of variables formula II) *Let* $[a,b]$ *be a closed interval, and let* $\phi : [a,b] \to [\phi(a), \phi(b)]$ *be a continuous monotone increasing function. Let* $f : [\phi(a), \phi(b)] \to \mathbf{R}$ *be a Riemann integrable function on* $[\phi(a), \phi(b)]$. *Then* $f \circ \phi : [a,b] \to \mathbf{R}$ *is Riemann–Stieltjes integrable with respect to* ϕ *on* $[a,b]$, *and*

$$\int_{[a,b]} f \circ \phi \, d\phi = \int_{[\phi(a),\phi(b)]} f.$$

Proof This will be obtained from Lemma 11.10.5 in a similar manner to how Corollary 11.10.3 was obtained from Theorem 11.10.2. First observe that since f is Riemann integrable, it is bounded, and then $f \circ \phi$ must also be bounded (why?).

Let $\varepsilon > 0$. Then, we can find a piecewise constant function $\overline{f}$ majorizing f on $[\phi(a), \phi(b)]$, and a piecewise constant function $\underline{f}$ minorizing f on $[\phi(a), \phi(b)]$, such that

$$\int_{[\phi(a),\phi(b)]} f - \varepsilon \le \int_{[\phi(a),\phi(b)]} \underline{f} \le \int_{[\phi(a),\phi(b)]} \overline{f} \le \int_{[\phi(a),\phi(b)]} f + \varepsilon.$$

Applying Lemma 11.10.5, we obtain

$$\int_{[\phi(a),\phi(b)]} f - \varepsilon \le \int_{[a,b]} \underline{f} \circ \phi \, \mathrm{d}\phi \le \int_{[a,b]} \overline{f} \circ \phi \, \mathrm{d}\phi \le \int_{[\phi(a),\phi(b)]} f + \varepsilon.$$

Since $\underline{f} \circ \phi$ is piecewise constant and minorizes $f \circ \phi$, we have

$$\int_{[a,b]} \underline{f} \circ \phi \, \mathrm{d}\phi \le \underline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi$$

while similarly we have

$$\int_{[a,b]} \overline{f} \circ \phi \, \mathrm{d}\phi \geq \overline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi.$$

Thus

$$\int_{[\phi(a),\phi(b)]} f - \varepsilon \leq \underline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi \leq \overline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi \leq \int_{[\phi(a),\phi(b)]} f + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this implies that

$$\int_{[\phi(a),\phi(b)]} f \leq \underline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi \leq \overline{\int}_{[a,b]} f \circ \phi \, \mathrm{d}\phi \leq \int_{[\phi(a),\phi(b)]} f$$

and the claim follows. □

Combining this formula with Corollary 11.10.3, one immediately obtains the following familiar formula:

Proposition 11.10.7 (Change of variables formula III) *Let $[a, b]$ be a closed interval, and let $\phi \colon [a, b] \to [\phi(a), \phi(b)]$ be a differentiable monotone increasing function such that ϕ' is Riemann integrable. Let $f \colon [\phi(a), \phi(b)] \to \mathbf{R}$ be a Riemann integrable function on $[\phi(a), \phi(b)]$. Then $(f \circ \phi)\phi' \colon [a, b] \to \mathbf{R}$ is Riemann integrable on $[a, b]$, and*

$$\int_{[a,b]} (f \circ \phi)\phi' = \int_{[\phi(a),\phi(b)]} f.$$

— Exercises —

Exercise 11.10.1 Prove Proposition 11.10.1. (*Hint:* first use Corollary 11.5.2 and Theorem 11.4.5 to show that FG' and $F'G$ are Riemann integrable. Then use the product rule (Theorem 10.1.13(d)).)

Exercise 11.10.2 Fill in the gaps marked (why?) in the proof of Lemma 11.10.5.

Exercise 11.10.3 Let $a < b$ be real numbers, and let $f \colon [a, b] \to \mathbf{R}$ be a Riemann integrable function. Let $g \colon [-b, -a] \to \mathbf{R}$ be defined by $g(x) := f(-x)$. Show that g is also Riemann integrable, and $\int_{[-b,-a]} g = \int_{[a,b]} f$.

Exercise 11.10.4 What is the analogue of Proposition 11.10.7 when ϕ is monotone decreasing instead of monotone increasing? (When ϕ is neither monotone increasing or monotone decreasing, the situation becomes significantly more complicated.)

Appendix A
The Basics of Mathematical Logic

The purpose of this appendix is to give a quick introduction to *mathematical logic*, which is the language one uses to conduct rigorous mathematical proofs. Knowing how mathematical logic works is also very helpful for understanding the mathematical way of thinking, which once mastered allows you to approach mathematical concepts and problems in a clear and confident way—including many of the proof-type questions in this text.

Writing logically is a very useful skill. It is somewhat related to, but not the same as, writing clearly, or efficiently, or convincingly, or informatively; ideally one would want to do all of these at once, but sometimes one has to make compromises, though with practice you'll be able to achieve more of your writing objectives concurrently. Thus a logical argument may sometimes look unwieldy, excessively complicated, or otherwise appear unconvincing. The big advantage of writing logically, however, is that one can be absolutely sure that your conclusion will be correct, as long as all your hypotheses were correct and your steps were logical; using other styles of writing one can be reasonably convinced that something is true, but there is a difference between being convinced and being *sure*.

Being logical is not the only desirable trait in writing, and in fact sometimes it gets in the way; mathematicians for instance often resort to short informal arguments which are not logically rigorous when they want to convince other mathematicians of a statement without going through all of the long details, and the same is true of course for non-mathematicians as well. So saying that a statement or argument is "not logical" is not necessarily a bad thing; there are often many situations when one has good reasons not to be emphatic about being logical. However, one should be aware of the distinction between logical reasoning and more informal means of argument, and not try to pass off an illogical argument as being logically rigorous. In particular, if an exercise is asking for a proof, then it is expecting you to be logical in your answer.

Logic is a skill that needs to be learnt like any other, but this skill is also innate to all of you—indeed, you probably use the laws of logic unconsciously in your everyday speech and in your own internal (non-mathematical) reasoning. However,

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4

it does take a bit of training and practice to recognize this innate skill and to apply it to abstract situations such as those encountered in mathematical proofs. Because logic is innate, the laws of logic that you learn should *make sense*—if you find yourself having to memorize one of the principles or laws of logic here, without feeling a mental "click" or comprehending why that law should work, then you will probably *not* be able to use that law of logic correctly and effectively in practice. So, *please* don't study this appendix the way you might cram before a final—that is going to be useless. Instead, **put away your highlighter pen**, and *read* and *understand* this appendix rather than merely *studying* it!

A.1 Mathematical Statements

Any mathematical argument proceeds in a sequence of *mathematical statements*. These are precise statements concerning various mathematical objects (numbers, vectors, functions, etc.), the operations between them (addition, multiplication, differentiation, etc.), and the relations between them (equality, inequality, etc.). These objects can either be constants or variables; more on this later. Statements[1] are either true or false.

Example A.1.1 $2 + 2 = 4$ is a true statement; $2 + 2 = 5$ is a false statement.

Not every combination of mathematical symbols is a statement. For instance,

$$= 2 + + 4 = - = 2$$

is not a statement; we sometimes call it *ill-formed* or *ill-defined*. The statements in the previous example are *well-formed* or *well-defined*. Thus well-formed statements can be either true or false; ill-formed statements are considered to be neither true nor false (in fact, they are usually not considered statements at all). A more subtle example of an ill-formed statement is

$$0/0 = 1;$$

division by zero is undefined, and so the above statement is ill-formed. A logical argument should not contain any ill-formed statements, thus for instance if an argument uses a statement such as $x/y = z$, it needs to first ensure that y is not equal to zero. Many purported proofs of "$0 = 1$" or other false statements rely on overlooking this "statements must be well-formed" criterion.

Many of you have probably written ill-formed or otherwise inaccurate statements in your mathematical work, while intending to mean some other, well-formed and accurate statement. To a certain extent this is permissible—it is similar to misspelling

[1] More precisely, statements with no free variables are either true or false. We shall discuss free variables later on in this appendix.

some words in a sentence, or using a slightly inaccurate or ungrammatical word in place of a correct one ("She ran good" instead of "She ran well"). In many cases, the reader (or grader) can detect this misstep and correct for it. However, it looks unprofessional and suggests that you may not know what you are talking about. And if indeed you actually do not know what you are talking about, and are applying mathematical or logical rules blindly, then writing an ill-formed statement can quickly confuse you into writing more and more nonsense—usually of the sort which receives no credit in grading. So it is important, especially when just learning a subject, to take care in keeping statements well-formed and precise. Once you have more skill and confidence, of course you can afford once again to speak loosely, because you will know what you are doing and won't be in as much danger of veering off into nonsense.

One of the basic axioms of mathematical logic is that every well-formed statement is either true or false, but not both. (Though if there are free variables, the truth of a statement may depend on the values of these variables. More on this later.) Furthermore, the truth or falsity of a statement is intrinsic to the statement and does not depend on the opinion of the person viewing the statement (as long as all the definitions and notations are agreed upon, of course). So to prove that a statement is true, it suffices to show that it is not false, while to show that a statement is false, it suffices to show that it is not true; this is the principle underlying the powerful technique of *proof by contradiction*, which we discuss later. This axiom is viable as long as one is working with precise concepts, for which the truth or falsity can be determined (at least in principle) in an objective and consistent manner. However, if one is working in very non-mathematical situations, then this axiom becomes much more dubious, and so it can be a mistake to apply mathematical logic to non-mathematical situations. (For instance, a statement such as "this rock weighs 52 pounds" is reasonably precise and objective, and so it is fairly safe to use mathematical reasoning to manipulate it, whereas vague statements such as "this rock is heavy", "this piece of music is beautiful", or "God exists" are much more problematic. So while mathematical logic is a very useful and powerful tool, it still does have some limitations of applicability.) One can still attempt to apply logic (or principles similar to logic) in these cases (for instance, by creating a *mathematical model* of a real-life phenomenon), but this is now science or philosophy, not mathematics, and we will not discuss it further here.

Remark A.1.2 There are other models of logic which attempt to deal with statements that are not definitely true or definitely false, such as modal logic, intuitionist logic, or fuzzy logic, but these are well beyond the scope of this text.

Being true is different from being *useful* or *efficient*. For instance, the statement

$$2 = 2$$

is true but unlikely to be very useful. The statement

$$4 \leq 4$$

is also true, but not very efficient (the statement $4 = 4$ is more precise). It may also be that a statement may be false yet still be useful, for instance

$$\pi = 22/7$$

is false, but is still useful as a first approximation. In mathematical reasoning, we only concern ourselves with truth rather than usefulness or efficiency; the reason is that truth is objective (everybody can agree on it), and we can deduce true statements from precise rules, whereas usefulness and efficiency are to some extent matters of opinion and do not follow precise rules. Also, even if some of the individual steps in an argument may not seem very useful or efficient, it is still possible (indeed, quite common) for the final conclusion to be quite non-trivial (i.e., not obviously true) and useful.

Statements are different from *expressions*. Statements are true or false; expressions are a sequence of mathematical symbols which produces some mathematical object (a number, matrix, function, set, etc.) as its value. For instance

$$2 + 3 * 5$$

is an expression, not a statement; it produces a number as its value. Meanwhile,

$$2 + 3 * 5 = 17$$

is a statement, not an expression. Thus it does not make any sense to ask whether $2 + 3 * 5$ is true or false. As with statements, expressions can be well-defined or ill-defined; $2 + 3/0$, for instance, is ill-defined. More subtle examples of ill-defined expressions arise when, for instance, attempting to add a vector to a matrix or evaluating a function outside of its domain, e.g., $\sin^{-1}(2)$.

One can make statements out of expressions by using *relations* such as $=$, $<$, $\geq$, $\in$, $\subset$ or by using *properties* (such as "is prime", "is continuous", "is invertible") For instance, "$30 + 5$ is prime" is a statement, as is "$30 + 5 \leq 42 - 7$". Note that mathematical statements are allowed to contain English words.

One can make a *compound statement* from more primitive statements by using *logical connectives* such as and, or, not, if-then, if-and-only-if. We give some examples below, in decreasing order of intuitiveness.

Conjunction. If X is a statement and Y is a statement, the statement "X and Y" is true if X and Y are both true and is false otherwise. For instance, "$2 + 2 = 4$ and $3 + 3 = 6$" is true, while "$2 + 2 = 4$ and $3 + 3 = 5$" is not. Another example: "$2 + 2 = 4$ and $2 + 2 = 4$" is true, even if it is a bit redundant; logic is concerned with truth, not efficiency.

Due to the expressiveness of the English language, one can reword the statement "X and Y" in many ways, e.g., "X and also Y", or "Both X and Y are true", etc. Interestingly, the statement "X, but Y" is logically the same statement as "X and Y", but they have different connotations (both statements affirm that X and Y are both true, but the first version suggests that X and Y are in contrast to each other, while the second version suggests that X and Y support each other). Again, logic is about truth, not about connotations or suggestions.

Disjunction. If X is a statement and Y is a statement, the statement "X or Y" is true if either X or Y is true, or both. For instance, "$2 + 2 = 4$ or $3 + 3 = 5$" is true, but "$2 + 2 = 5$ or $3 + 3 = 5$" is not. Also "$2 + 2 = 4$ or $3 + 3 = 6$" is true (even if it is a bit inefficient; it would be a stronger statement to say "$2 + 2 = 4$ and $3 + 3 = 6$"). Thus by default, the word "or" in mathematical logic defaults to *inclusive or*. The reason we do this is that with inclusive or, to verify "X or Y", it suffices to verify that just one of X or Y is true; we don't need to show that the other one is false. So we know, for instance, that "$2 + 2 = 4$ or $2353 + 5931 = 7284$" is true without having to look at the second equation. As in the previous discussion, the statement "$2 + 2 = 4$ or $2 + 2 = 4$" is true, even if it is highly inefficient.

If one really does want to use exclusive or, use a statement such as "Either X or Y is true, but not both" or "Exactly one of X or Y is true". Exclusive or does come up in mathematics, but nowhere near as often as inclusive or.

Negation. The statement "X is not true" or "X is false", or "It is not the case that X", is called the *negation* of X and is true if and only if X is false, and is false if and only if X is true. For instance, the statement "It is not the case that $2 + 2 = 5$" is a true statement. Of course we could abbreviate this statement to "$2 + 2 \neq 5$".

Negations convert "and" into "or". For instance, the negation of "Jane Doe has black hair and Jane Doe has blue eyes" is "Jane Doe doesn't have black hair or doesn't have blue eyes", *not* "Jane Doe doesn't have black hair and doesn't have blue eyes" (can you see why?). Similarly, if x is an integer, the negation of "x is even and non-negative" is "x is odd or negative", not "x is odd and negative". (Note how it is important here that or is inclusive rather than exclusive.) Or the negation of "$x \geq 2$ and $x \leq 6$" (i.e., "$2 \leq x \leq 6$") is "$x < 2$ or $x > 6$", not "$x < 2$ and $x > 6$" or "$2 < x > 6$".

Similarly, negations convert "or" into "and". The negation of "John Doe has brown hair or black hair" is "John Doe does not have brown hair and does not have black hair", or equivalently "John Doe has neither brown nor black hair". If x is a real number, the negation of "$x \geq 1$ or $x \leq -1$" is "$x < 1$ and $x > -1$" (i.e., $-1 < x < 1$).

It is quite possible that a negation of a statement will produce a statement which could not possibly be true. For instance, if x is an integer, the negation of "x is either even or odd" is "x is neither even nor odd", which cannot possibly be true. Remember, though, that even if a statement is false, it is still a statement, and it is definitely possible to arrive at a true statement using an argument which at times involves false statements. (Proofs by contradiction, for instance, fall into this category. Another example is proof by dividing into cases. If one divides into three mutually exclusive cases, Case 1, Case 2, and Case 3, then at any given time two of the cases will be false and only one will be true; however this does not necessarily mean that the proof as a whole is incorrect or that the conclusion is false.)

Negations are sometimes unintuitive to work with, especially if there are multiple negations; a statement such as "It is not the case that either x is not odd, or x is not larger than or equal to 3, but not both" is not particularly pleasant to use. Fortunately, one rarely has to work with more than one or two negations at a time, since often negations cancel each other. For instance, the negation of "X is not true" is just "X is true", or more succinctly just "X". Of course one should be careful when negating more complicated expressions because of the switching of "and" and "or", and similar issues.

If and only if (iff). If X is a statement, and Y is a statement, we say that "X is true if and only if Y is true", whenever X is true, Y has to be also, and whenever Y is true, X has to be also (i.e., X and Y are "equally true"). Other ways of saying the same thing are "X and Y are logically equivalent statements", or "X is true iff Y is true", or "$X \leftrightarrow Y$". Thus for instance, if x is a real number, then the statement "$x = 3$ if and only if $2x = 6$" is true: this means that whenever $x = 3$ is true, then $2x = 6$ is true, and whenever $2x = 6$ is true, then $x = 3$ is true. On the other hand, the statement "$x = 3$ if and only if $x^2 = 9$" is false; while it is true that whenever $x = 3$ is true, $x^2 = 9$ is also true, it is not the case that whenever $x^2 = 9$ is true, that $x = 3$ is also automatically true (think of what happens when $x = -3$).

Statements that are equally true are also equally false: if X and Y are logically equivalent, and X is false, then Y has to be false also (because if Y were true, then X would also have to be true). Conversely, any two statements which are equally false will also be logically equivalent. Thus for instance $2 + 2 = 5$ if and only if $4 + 4 = 10$.

Sometimes it is of interest to show that more than two statements are logically equivalent; for instance, one might want to assert that three statements X, Y, and Z are all logically equivalent. This means whenever one of the statements is true, then all of the statements are true; and it also means that if one of the statements is false, then all of the statements are false. This may seem like a lot of logical implications to prove, but in practice, once one demonstrates enough logical implications between X, Y, and Z, one can often conclude all the others and conclude that they are all logically equivalent. See for instance Exercises A.1.5, A.1.6.

— Exercises —

Exercise A.1.1 What is the negation of the statement "either X is true, or Y is true, but not both"?

Exercise A.1.2 What is the negation of the statement "X is true if and only if Y is true"? (There may be multiple ways to phrase this negation.)

Exercise A.1.3 Suppose that you have shown that whenever X is true, then Y is true, and whenever X is false, then Y is false. Have you now demonstrated that X and Y are logically equivalent? Explain.

Exercise A.1.4 Suppose that you have shown that whenever X is true, then Y is true, and whenever Y is false, then X is false. Have you now demonstrated that X is true if and only if Y is true? Explain.

Exercise A.1.5 Suppose you know that X is true if and only if Y is true, and you know that Y is true if and only if Z is true. Is this enough to show that X, Y, Z are all logically equivalent? Explain.

Exercise A.1.6 Suppose you know that whenever X is true, then Y is true; that whenever Y is true, then Z is true; and whenever Z is true, then X is true. Is this enough to show that X, Y, Z are all logically equivalent? Explain.

A.2 Implication

Now we come to the least intuitive of the commonly used logical connectives—implication. If X is a statement, and Y is a statement, then "if X, then Y" is the implication from X to Y; it is also written "when X is true, Y is true", or "X implies Y" or "Y is true when X is" or "X is true only if Y is true" (this last one takes a bit of mental effort to see). What this statement "if X, then Y" means depends on whether X is true or false. If X is true, then "if X, then Y" is true when Y is true, and false when Y is false. If however X is false, then "if X, then Y" is *always* true, regardless of whether Y is true or false! To put it another way, when X is true, the statement "if X, then Y" implies that Y is true. But when X is false, the statement "if X, then Y" offers no information about whether Y is true or not; the statement is true, but *vacuous* (i.e., does not convey any new information beyond the fact that the hypothesis is false).

Examples A.2.1 If x is an integer, then the statement "If $x = 2$, then $x^2 = 4$" is true, regardless of whether x is actually equal to 2 or not (though this statement is only likely to be useful when x *is* equal to 2). This statement does not assert that x is equal to 2 and does not assert that x^2 is equal to 4, but it does assert that when and if x is equal to 2, then x^2 is equal to 4. If x is not equal to 2, the statement is still true but offers no conclusion on x or x^2.

Some special cases of the above implication: the implication "If $2 = 2$, then $2^2 = 4$" is true (true implies true). The implication "If $3 = 2$, then $3^2 = 4$" is true (false implies false). The implication "If $-2 = 2$, then $(-2)^2 = 4$" is true (false implies true). The latter two implications are considered vacuous—they do not offer any new information since their hypothesis is false. (Nevertheless, it is still possible to employ vacuous implications to good effect in a proof—a vacously true statement is still true. We shall see one such example shortly.)

As we see, the falsity of the hypothesis does not destroy the truth of an implication, in fact it is just the opposite! (When a hypothesis is false, the implication is automatically true.) The only way to disprove an implication is to show that the hypothesis is true while the conclusion is false. Thus "If $2 + 2 = 4$, then $4 + 4 = 2$" is a false implication. (True does not imply false.)

One can also think of the statement "if X, then Y" as "Y is *at least as true as* X"—if X is true, then Y also has to be true, but if X is false, Y could be as false as X, but it could also be true. This should be compared with "X if and only if Y", which asserts that X and Y are *equally true*.

Vacuously true implications are often used in ordinary speech, sometimes without knowing that the implication is vacuous. A somewhat frivolous example is "If wishes were wings, then pigs would fly". (The statement "hell freezes over" is also a popular choice for a false hypothesis.) A more serious one is "If John had left work at 5 pm, then he would be here by now". This kind of statement is often used in a situation in which the conclusion and hypothesis are both false; but the implication is still true regardless. This statement, by the way, can be used to illustrate the technique of proof by contradiction: if you believe that "If John had left work at 5 pm, then he would be here by now", and you also know that "John is not here by now", then you can conclude that "John did not leave work at 5 pm", because John leaving work at 5 pm would lead to a contradiction. Note how a vacuous implication can be used to derive a useful truth.

To summarize, implications are sometimes vacuous, but this is not actually a problem in logic, since these implications are still true, and vacuous implications can still be useful in logical arguments. In particular, one can safely use statements like "If X, then Y" without necessarily having to worry about whether the hypothesis X is actually true or not (i.e., whether the implication is vacuous or not).

Implications can also be true even when there is no causal link between the hypothesis and conclusion. The statement "If $1 + 1 = 2$, then Washington D.C. is the capital of the United States" is true (true implies true), although rather odd; the statement "If $2 + 2 = 3$, then New York is the capital of the United States" is similarly true (false implies false). Of course, such a statement may be unstable (the capital of the United States may one day change, while $1 + 1$ will always remain equal to 2) but it is true, at least for the moment. While it is possible to use a causal implications in a logical argument, it is not recommended as it can cause unneeded confusion. (Thus, for instance, while it is true that a false statement can be used to imply any other statement, true or false, doing so arbitrarily would probably not be helpful to the reader.)

To prove an implication "If X, then Y", the usual way to do this is to first assume that X is true, and use this (together with whatever other facts and hypotheses you have) to deduce Y. This is still a valid procedure even if X later turns out to be false; the implication does not guarantee anything about the truth of X and only guarantees the truth of Y conditionally on X first being true. For instance, the following is a valid proof of a true proposition, even though both hypothesis and conclusion of the proposition are false:

Proposition A.2.2 *If* $2 + 2 = 5$*, then* $4 = 10 - 4$.

Proof Assume $2 + 2 = 5$. Multiplying both sides by 2, we obtain $4 + 4 = 10$. Subtracting 4 from both sides, we obtain $4 = 10 - 4$ as desired. □

On the other hand, a common error is to prove an implication by first assuming the *conclusion* and then arriving at the hypothesis. For instance, the following proposition is correct, but the proof is not:

Proposition A.2.3 *Suppose that* $2x + 3 = 7$. *Show that* $x = 2$.

Proof (Incorrect) $x = 2$; so $2x = 4$; so $2x + 3 = 7$. □

When doing proofs, it is important that you are able to distinguish the hypothesis from the conclusion; there is a danger of getting hopelessly confused if this distinction is not clear.

Here is a short proof which uses implications which are possibly vacuous.

Theorem A.2.4 *Suppose that n is an integer. Then $n(n + 1)$ is an even integer.*

Proof Since n is an integer, n is even or odd. If n is even, then $n(n + 1)$ is also even, since any multiple of an even number is even. If n is odd, then $n + 1$ is even, which again implies that $n(n + 1)$ is even. Thus in either case $n(n + 1)$ is even, and we are done. □

Note that this proof relied on two implications: "if n is even, then $n(n + 1)$ is even", and "if n is odd, then $n(n + 1)$ is even". Since n cannot be both odd and even, at least one of these implications has a false hypothesis and is therefore vacuous. Nevertheless, both these implications are true, and one needs *both* of them in order to prove the theorem, because we don't know in advance whether n is even or odd. And even if we did, it might not be worth the trouble to check it. For instance, as a special case of this theorem we immediately know

Corollary A.2.5 *Let $n = (253 + 142) * 123 - (423 + 198)^{342} + 538 - 213$. Then $n(n + 1)$ is an even integer.*

In this particular case, one can work out exactly which parity n is —even or odd — and then use only one of the two implications in the above theorem, discarding the vacuous one. This may seem like it is more efficient, but it is a false economy, because one then has to determine what parity n is, and this requires a bit of effort—more effort than it would take if we had just left both implications, including the vacuous one, in the argument. So, somewhat paradoxically, the inclusion of vacuous, false, or otherwise "useless" statements in an argument can actually *save* you effort in the long run! (I'm not suggesting, of course, that you ought to pack your proofs with lots of time-wasting and irrelevant statements; all I'm saying here is that you need not be unduly concerned that some hypotheses in your argument might not be correct, as long as your argument is still structured to give the correct conclusion regardless of whether those hypotheses were true or false.)

The statement "If X, then Y" is not the same as "If Y, then X"; for instance, while "If $x = 2$, then $x^2 = 4$" is true, "If $x^2 = 4$, then $x = 2$" can be false if x is equal to -2. These two statements are called *converses* of each other; thus the converse of a true implication is not necessarily another true implication. We use the statement "X if and only if Y" to denote the statement that "If X, then Y; and if Y, then X". Thus for instance, we can say that $x = 2$ if and only if $2x = 4$, because if $x = 2$ then $2x = 4$, while if $2x = 4$ then $x = 2$. One way of thinking about an if-and-only-if statement is to view "X if and only if Y" as saying that X is just as true as Y; if one is true then so is the other, and if one is false, then so is the other. For instance, the statement "If $3 = 2$, then $6 = 4$" is true, since both hypothesis and conclusion are false. (Under this view, "If X, then Y" can be viewed as a statement that Y is at least as true as X.) Thus one could say "X and Y are equally true" instead of "X if and only if Y".

Similarly, the statement "If X is true, then Y is true" is **not** the same as "If X is false, then Y is false". Saying that "if $x = 2$, then $x^2 = 4$" does not imply that "if $x \neq 2$, then $x^2 \neq 4$", and indeed we have $x = -2$ as a counterexample in this case. If-then statements are not the same as if-and-only-if statements. (If we knew that "X is true if and only if Y is true", then we would also know that "X is false if and only if Y is false".) The statement "If X is false, then Y is false" is sometimes called the *inverse* of "If X is true, then Y is true"; thus the inverse of a true implication is not necessarily a true implication.

If you know that "If X is true, then Y is true", then it is also true that "If Y is false, then X is false" (because if Y is false, then X can't be true, since that would imply Y is true, a contradiction). For instance, if we knew that "If $x = 2$, then $x^2 = 4$", then we also know that "If $x^2 \neq 4$, then $x \neq 2$". Or if we knew "If John had left work at 5 pm, he would be here by now", then we also know "If John isn't here now, then he could not have left work at 5 pm". The statement "If Y is false, then X is false" is known as the *contrapositive* of "If X, then Y", and both statements are equally true.

In particular, if you know that X implies something which is known to be false, then X itself must be false. This is the idea behind *proof by contradiction* or *reductio ad absurdum*: to show something must be false, assume first that it is true, and show that this implies something which you know to be false (e.g., that a statement is simultaneously true and not true). For instance:

Proposition A.2.6 *Suppose that x be a positive number such that* $\sin(x) = 1$. *Then* $x \geq \pi/2$.

Proof Suppose for sake of contradiction that $x < \pi/2$. Since x is positive, we thus have $0 < x < \pi/2$. Since $\sin(x)$ is increasing for $0 \leq x \leq \pi/2$, and $\sin(0) = 0$ and $\sin(\pi/2) = 1$, we thus have $0 < \sin(x) < 1$. But this contradicts the hypothesis that $\sin(x) = 1$. Hence $x \geq \pi/2$. □

Note that one feature of proof by contradiction is that at some point in the proof you assume a hypothesis (in this case, that $x < \pi/2$) which later turns out to be false. Note however that this does not alter the fact that the argument remains valid, and that the conclusion is true; this is because the ultimate conclusion does not rely on that hypothesis being true (indeed, it relies instead on it being false!).

Proof by contradiction is particularly useful for showing "negative" statements—that X is false, that a is not equal to b, that kind of thing. But the line between positive and negative statements is sort of blurry. (Is the statement $x \geq 2$ a positive or negative statement? What about its negation, that $x < 2$?) So this is not a hard and fast rule.

Logicians often use special symbols to denote logical connectives; for instance "X implies Y" can be written "$X \implies Y$", "X is not true" can be written "$\sim X$", "$!X$", or "$\neg X$", "X and Y" can be written "$X \wedge Y$" or "$X \& Y$", and so forth. But for general-purpose mathematics, these symbols are not often used; English words are often more readable and don't take up much more space. Also, using these symbols tends to blur the line between expressions and statements; it's not as easy to understand "$((x = 3) \wedge (y = 5)) \implies (x + y = 8)$" as "If $x = 3$ and $y = 5$, then $x + y = 8$". So in general I would not recommend using these symbols (except possibly for $\implies$, which is a very intuitive symbol).

A.3 The Structure of Proofs

To prove a statement, one often starts by assuming the hypothesis and working one's way toward a conclusion; this is the *direct* approach to proving a statement. Such a proof might look something like the following:

Proposition A.3.1 *A implies B.*

Proof Assume A is true. Since A is true, C is true. Since C is true, D is true. Since D is true, B is true, as desired. □

An example of such a direct approach is

Proposition A.3.2 *If $x = \pi$, then* $\sin(x/2) + 1 = 2$.

Proof Let $x = \pi$. Since $x = \pi$, we have $x/2 = \pi/2$. Since $x/2 = \pi/2$, we have $\sin(x/2) = 1$. Since $\sin(x/2) = 1$, we have $\sin(x/2) + 1 = 2$. □

In the above proof, we started at the hypothesis and moved steadily from there toward a conclusion. It is also possible to work backward from the conclusion and seeing what it would take to imply it. For instance, a typical proof of Proposition A.3.1 of this sort might look like the following:

Proof To show B, it would suffice to show D. Since C implies D, we just need to show C. But C follows from A. □

As an example of this, we give another proof of Proposition A.3.2:

Proof To show $\sin(x/2) + 1 = 2$, it would suffice to show that $\sin(x/2) = 1$. Since $x/2 = \pi/2$ would imply $\sin(x/2) = 1$, we just need to show that $x/2 = \pi/2$. But this follows since $x = \pi$. □

Logically speaking, the above two proofs of Proposition A.3.2 are the same, just arranged differently. Note how this proof style is different from the (incorrect) approach of starting with the conclusion and seeing what it would imply (as in Proposition A.2.3); instead, we start with the conclusion and see what would imply it.

Another example of a proof written in this backward style is the following:

Proposition A.3.3 *Let $0 < r < 1$ be a real number. Then the series $\sum_{n=1}^{\infty} nr^n$ is convergent.*

Proof To show this series is convergent, it suffices by the ratio test to show that the ratio

$$\left|\frac{r^{n+1}(n+1)}{r^n n}\right| = r\frac{n+1}{n}$$

converges to something less than 1 as $n \to \infty$. Since r is already less than 1, it will be enough to show that $\frac{n+1}{n}$ converges to 1. But since $\frac{n+1}{n} = 1 + \frac{1}{n}$, it suffices to show that $\frac{1}{n} \to 0$. But this is clear since $n \to \infty$. □

One could also do any combination of moving forward from the hypothesis and backward from the conclusion. For instance, the following would be a valid proof of Proposition A.3.1:

Proof To show B, it would suffice to show D. So now let us show D. Since we have A by hypothesis, we have C. Since C implies D, we thus have D as desired. □

Again, from a logical point of view this is exactly the same proof as before. Thus there are many ways to write the same proof down; how you do so is up to you, but certain ways of writing proofs are more readable and natural than others, and different arrangements tend to emphasize different parts of the argument. (Of course, when you are just starting out doing mathematical proofs, you're generally happy to get *some* proof of a result and don't care so much about getting the "best" arrangement of that proof; but the point here is that a proof can take many different forms.)

The above proofs were pretty simple because there was just one hypothesis and one conclusion. When there are multiple hypotheses and conclusions, and the proof splits into cases, then proofs can get more complicated. For instance a proof might look as tortuous as this:

Proposition A.3.4 *Suppose that A and B are true. Then C and D are true.*

Proof Since A is true, E is true. From E and B we know that F is true. Also, in light of A, to show D it suffices to show G. There are now two cases: H and I. If H is true, then from F and H we obtain C, and from A and H we obtain G. If instead I is true, then from I we have G, and from I and G we obtain C. Thus in both cases we obtain both C and G, and hence C and D. □

Incidentally, the above proof could be rearranged into a much tidier manner, but you at least get the idea of how complicated a proof could become. To show an implication there are several ways to proceed: you can work forward from the hypothesis; you can work backward from the conclusion; or you can divide into cases in the hope to split the problem into several easier subproblems. Another is to argue by contradiction, for instance you can have an argument of the form

Proposition A.3.5 *Suppose that A is true. Then B is false.*

Proof Suppose for sake of contradiction that B is true. This would imply that C is true. But since A is true, this implies that D is true; which contradicts C. Thus B must be false. □

As you can see, there are several things to try when attempting a proof. With experience, it will become clearer which approaches are likely to work easily, which ones will probably work but require much effort, and which ones are probably going to fail. In many cases there is really only one obvious way to proceed. Of course, there may definitely be multiple ways to approach a problem, so if you see more than one way to begin a problem, you can just try whichever one looks the easiest, but be prepared to switch to another approach if it begins to look hopeless.

Also, it helps when doing a proof to keep track of which statements are *known* (either as hypotheses, or deduced from the hypotheses, or coming from other theorems and results) and which statements are *desired* (either the conclusion, or something which would imply the conclusion, or some intermediate claim or lemma which will be useful in eventually obtaining the conclusion). Mixing the two up is almost always a bad idea and can lead to one getting hopelessly lost in a proof.

A.4 Variables and Quantifiers

One can get quite far in logic just by starting with primitive statements (such as "$2 + 2 = 4$" or "John has black hair"), then forming compound statements using logical connectives, and then using various laws of logic to pass from one's hypotheses to one's conclusions; this is known as *propositional logic* or *Boolean logic*. (It is possible to list a dozen or so such laws of propositional logic, which are sufficient to do everything one wants to do, but I have deliberately chosen not to do so here, because you might then be tempted to memorize that list, and that is **not** how one should learn how to do logic, unless one happens to be a computer or some other non-thinking device. However, if you really are curious as to what the formal laws of logic are, look up "laws of propositional logic" or something similar in the library or on the internet.)

However, to do mathematics, this level of logic is insufficient, because it does not incorporate the fundamental concept of *variables*—those familiar symbols such as x or n which denote various quantities which are unknown, or set to some value, or assumed to obey some property. Indeed we have already sneaked in some of these variables in order to illustrate some of the concepts in propositional logic (mainly because it gets boring after a while to talk endlessly about variable-free statements such as $2 + 2 = 4$ or "Jane has black hair"). *Mathematical logic* is thus the same as propositional logic but with the additional ingredient of variables added.

A *variable* is a symbol, such as n or x, which denotes a certain type of mathematical object—an integer, a vector, a matrix, that kind of thing. In almost all circumstances, the type of object that the variable represents should be declared, otherwise it will be difficult to make well-formed statements using it. (There are very few true statements that one can make about variables without knowing the type of variables involved. For instance, given a variable x of any type whatsoever, it is true that $x = x$, and if we also know that $x = y$, then we can conclude that $y = x$. But one cannot say, for instance, that $x + y = y + x$, until we know what type of objects x and y are and whether they support the operation of addition; for instance, the above statement is ill-formed if x is a matrix and y is a vector. Thus if one actually wants to do some useful mathematics, then every variable should have an explicit type.)

One can form expressions and statements involving variables, for instance, if x is a real variable (i.e., a variable which is a real number), $x + 3$ is an expression, and $x + 3 = 5$ is a statement. But now the truth of a statement may depend on the value of the variables involved; for instance the statement $x + 3 = 5$ is true if x is equal to 2, but is false if x is not equal to 2. Thus the truth of a statement involving a variable may depend on the *context* of the statement—in this case, it depends on what x is supposed to be. (This is a modification of the rule for propositional logic, in which all statements have a definite truth value.)

Sometimes we do not set a variable to be anything (other than specifying its type). Thus, we could consider the statement $x + 3 = 5$ where x is an unspecified real number. In such a case we call this variable a *free variable*; thus we are considering $x + 3 = 5$ with x a free variable. Statements with free variables might not have a definite truth value, as they depend on an unspecified variable. For instance, we have already remarked that $x + 3 = 5$ does not have a definite truth value if x is a free real variable, though of course for each given value of x the statement is either true or false. On the other hand, the statement $(x + 1)^2 = x^2 + 2x + 1$ is true for every real number x, and so we can regard this as a true statement even when x is a free variable.

At other times, we *set* a variable to equal a fixed value, by using a statement such as "Let $x = 2$" or "Set x equal to 2". In this case, the variable is known as a *bound variable*, and statements involving only bound variables and no free variables do have a definite truth value. For instance, if we set $x = 342$, then the statement "$x + 135 = 477$" now has a definite truth value, whereas if x is a free real variable then "$x + 135 = 477$" could be either true or false, depending on what x is. Thus, as we have said before, the truth of a statement such as "$x + 135 = 477$" depends on the context—whether x is free or bound, and if it is bound, what it is bound to.

One can also turn a free variable into a bound variable by using the quantifiers "for all" or "for some". For instance, the statement

$$(x + 1)^2 = x^2 + 2x + 1$$

is a statement with one free variable x and need not have a definite truth value, but the statement

$$(x + 1)^2 = x^2 + 2x + 1 \text{ for all real numbers } x$$

is a statement with one bound variable x and now has a definite truth value (in this case, the statement is true). Similarly, the statement

$$x + 3 = 5$$

has one free variable and does not have a definite truth value, but the statement

$$x + 3 = 5 \text{ for some real number } x$$

is true, since it is true for $x = 2$. On the other hand, the statement

$$x + 3 = 5 \text{ for all real numbers } x$$

is false, because there are some (indeed, there are many) real numbers x for which $x + 3$ is not equal to 5.

Universal quantifiers. Let $P(x)$ be some statement depending on a free variable x. The statement "$P(x)$ is true for all x of type T" means that given any x of type T, the statement $P(x)$ is true regardless of what the exact value of x is. In other words, the statement is the same as saying "if x is of type T, then $P(x)$ is true". Thus the usual way to prove such a statement is to let x be a free variable of type T (by saying something like "Let x be any object of type T"), and then proving $P(x)$ for that object. The statement becomes false if one can produce even a single counterexample, i.e., an element x which lies in T but for which $P(x)$ is false. For instance, the statement "x^2 is greater than x for all positive x" can be shown to be false by producing a single example, such as $x = 1$ or $x = 1/2$, where x^2 is not greater than x.

On the other hand, producing a single example where $P(x)$ *is* true will not show that $P(x)$ is true for *all* x. For instance, just because the equation $x + 3 = 5$ has a solution when $x = 2$ does not imply that $x + 3 = 5$ for all real numbers x; it only shows that $x + 3 = 5$ is true for some real number x. (This is the source of the often-quoted, though somewhat inaccurate, slogan "One cannot prove a statement just by giving an example". The more precise statement is that one cannot prove a "for all" statement by examples, though one can certainly prove "for some" statements this way, and one can also *disprove* "for all" statements by a single counterexample.)

It occasionally happens that there are in fact no variables x of type T. In that case the statement "$P(x)$ is true for all x of type T" is *vacuously true*—it is true but has no content, similar to a vacuous implication. For instance, the statement

$$6 < 2x < 4 \text{ for all } 3 < x < 2$$

is true, and easily proven, but is vacuous. (Such a vacuously true statement can still be useful in an argument, though this doesn't happen very often.)

One can use phrases such as "For every" or "For each" instead of "For all", e.g., one can rephrase "$(x + 1)^2 = x^2 + 2x + 1$ for all real numbers x" as "For each real number x, $(x + 1)^2$ is equal to $x^2 + 2x + 1$". For the purposes of logic these rephrasings are equivalent. The symbol $\forall$ can be used instead of "For all", thus for instance "$\forall x \in X : P(x)$ is true" or "$P(x)$ is true $\forall x \in X$" is synonymous with "$P(x)$ is true for all $x \in X$".

Existential quantifiers. The statement "$P(x)$ is true for some x of type T" means that there exists at least one x of type T for which $P(x)$ is true, although it may be that there is more than one such x. (One would use a quantifier such as "for exactly one x” instead of “for some x” if one wanted both existence and uniqueness of such an x.) To prove such a statement it suffices to provide a single example of such an x. For instance, to show that

$$x^2 + 2x - 8 = 0 \text{ for some real number } x$$

all one needs to do is find a single real number x for which $x^2 + 2x - 8 = 0$, for instance $x = 2$ will do. (One could also use $x = -4$, but one doesn’t need to use both.) Note that one has the freedom to select x to be anything one wants when proving a for some statement; this is in contrast to proving a for all statement, where one has to let x be arbitrary. (One can compare the two statements by thinking of two games between you and an opponent. In the first game, the opponent gets to pick what x is, and then you have to prove $P(x)$; if you can always win this game, then you have proven that $P(x)$ is true for *all* x. In the second game, *you* get to choose what x is, and then you prove $P(x)$; if you can win this game, you have proven that $P(x)$ is true for *some* x.)

Usually, saying something is true for *all* x is much stronger than just saying it is true for *some* x. There is one exception though, if the condition on x is impossible to satisfy, then the for all statement is vacuously true, but the for some statement is false. For instance

$$6 < 2x < 4 \text{ for all } 3 < x < 2$$

is true, but

$$6 < 2x < 4 \text{ for some } 3 < x < 2$$

is false.

One can use phrases such as “For at least one” or “There exists …such that” instead of “For some”. For instance, one can rephrase “$x^2 + 2x - 8 = 0$ for some real number x” as “There exists a real number x such that $x^2 + 2x - 8 = 0$”. The symbol $\exists$ can be used instead of “There exists …such that”, thus for instance “$\exists x \in X : P(x)$ is true” is synonymous with “$P(x)$ is true for some $x \in X$”.

A.5 Nested Quantifiers

One can nest two or more quantifiers together. For instance, consider the statement

For every positive number x, there exists a
positive number y such that $y^2 = x$.

What does this statement mean? It means that for each positive number x, the statement

$$\text{There exists a positive number } y \text{ such that } y^2 = x$$

is true. In other words, one can find a positive square root of x for each positive number x. So the statement is saying that every positive number has a positive square root.

To continue the gaming metaphor, suppose you play a game where your opponent first picks a positive number x, and then you pick a positive number y. You win the game if $y^2 = x$. If you can always win the game regardless of what your opponent does, then you have proven that for every positive x, there exists a positive y such that $y^2 = x$.

Negating a universal statement produces an existential statement. The negation of "All swans are white" is not "All swans are not white", but rather "There is some swan which is not white". Similarly, the negation of "For every $0 < x < \pi/2$, we have $\cos(x) \geq 0$" is "For some $0 < x < \pi/2$, we have $\cos(x) < 0$, **not** "For every $0 < x < \pi/2$, we have $\cos(x) < 0$".

Negating an existential statement produces a universal statement. The negation of "There exists a black swan" is not "There exists a swan which is non-black", but rather "All swans are non-black". Similarly, the negation of "There exists a real number x such that $x^2 + x + 1 = 0$" is "For every real number x, $x^2 + x + 1 \neq 0$", **not** "There exists a real number x such that $x^2 + x + 1 \neq 0$". (The situation here is very similar to how "and" and "or" behave with respect to negations.)

If you know that a statement $P(x)$ is true for all x, then you can set x to be anything you want, and $P(x)$ will be true for that value of x; this is what "for all" means. Thus for instance if you know that

$$(x + 1)^2 = x^2 + 2x + 1 \text{ for all real numbers } x,$$

then you can conclude for instance that

$$(\pi + 1)^2 = \pi^2 + 2\pi + 1,$$

or for instance that

$$(\cos(y) + 1)^2 = \cos(y)^2 + 2\cos(y) + 1 \text{ for all real numbers } y$$

(because if y is real, then $\cos(y)$ is also real), and so forth. Thus universal statements are very versatile in their applicability—you can get $P(x)$ to hold for whatever x you wish. Existential statements, by contrast, are more limited; if you know that

$$x^2 + 2x - 8 = 0 \text{ for some real number } x$$

then you cannot simply substitute in any real number you wish, e.g., π, and conclude that $\pi^2 + 2\pi - 8 = 0$. However, you can of course still conclude that $x^2 + 2x - 8 = 0$ for some real number x, it's just that you don't get to pick which x it is. (To continue the gaming metaphor, you can make $P(x)$ hold, but your opponent gets to pick x for you, you don't get to choose for yourself.)

Remark A.5.1 In the history of logic, quantifiers were formally studied thousands of years before Boolean logic was. Indeed, *Aristotlean logic*, developed of course by Aristotle (384BC – 322BC) and his school, deals with objects, their properties, and quantifiers such as "for all" and "for some". A typical line of reasoning (or *syllogism*) in Aristotlean logic goes like this: "All men are mortal. Socrates is a man. Hence, Socrates is mortal". Aristotlean logic is a subset of mathematical logic, but is not as expressive because it lacks the concept of logical connectives such as and, or, or if-then (although "not" is allowed) and also lacks the concept of a binary relation such as $=$ or $<$.

Swapping the order of two quantifiers may or may not make a difference to the truth of a statement. Swapping two "for all" quantifiers is harmless: a statement such as

$$\text{For all real numbers } a, \text{ and for all real numbers } b,$$
$$\text{we have } (a+b)^2 = a^2 + 2ab + b^2$$

is logically equivalent to the statement

$$\text{For all real numbers } b, \text{ and for all real numbers } a,$$
$$\text{we have } (a+b)^2 = a^2 + 2ab + b^2$$

(why? The reason has nothing to do with whether the identity $(a+b)^2 = a^2 + 2ab + b^2$ is actually true or not). Similarly, swapping two "there exists" quantifiers has no effect:

$$\text{There exists a real number } a, \text{ and there exists a real number } b,$$
$$\text{such that } a^2 + b^2 = 0$$

is logically equivalent to

$$\text{There exists a real number } b, \text{ and there exists a real number } a,$$
$$\text{such that } a^2 + b^2 = 0.$$

However, swapping a "for all" with a "there exists" makes a lot of difference. Consider the following two statements:

(a) For every integer n, there exists an integer m which is larger than n.
(b) There exists an integer m such that m is larger than n for every integer n.

Statement (a) is obviously true: if your opponent hands you an integer n, you can always find an integer m which is larger than n. But Statement (b) is false: if you choose m first, then you cannot ensure that m is larger than every integer n; your opponent can easily pick a number n bigger than m to defeat that. The crucial difference between the two statements is that in Statement (a), the integer n was chosen *first*, and integer m could then be chosen in a manner depending on n; but in

Statement (b), one was forced to choose m first, without knowing in advance what n is going to be. In short, the reason why the order of quantifiers is important is that the inner variables may possibly depend on the outer variables, but not vice versa.

— Exercises —

Exercise A.5.1 What does each of the following statements mean, and which of them are true? Can you find gaming metaphors for each of these statements?

(a) For every positive number x, and every positive number y, we have $y^2 = x$.
(b) There exists a positive number x such that for every positive number y, we have $y^2 = x$.
(c) There exists a positive number x, and there exists a positive number y, such that $y^2 = x$.
(d) For every positive number y, there exists a positive number x such that $y^2 = x$.
(e) There exists a positive number y such that for every positive number x, we have $y^2 = x$.

A.6 Some Examples of Proofs and Quantifiers

Here we give some simple examples of proofs involving the "for all" and "there exists" quantifiers. The results themselves are simple, but you should pay attention instead to how the quantifiers are arranged and how the proofs are structured.

Proposition A.6.1 *For every $\varepsilon > 0$ there exists a $\delta > 0$ such that $2\delta < \varepsilon$.*

Proof Let $\varepsilon > 0$ be arbitrary. We have to show that there exists a $\delta > 0$ such that $2\delta < \varepsilon$. We only need to pick one such δ; choosing $\delta := \varepsilon/3$ will work, since one then has $2\delta = 2\varepsilon/3 < \varepsilon$. □

Notice how ε has to be arbitrary, because we are proving something for *every* ε; on the other hand, δ can be chosen as you wish, because you only need to show that there *exists* a δ which does what you want. Note also that δ can depend on ε, because the δ-quantifier is nested inside the ε-quantifier. If the quantifiers were reversed, i.e., if you were asked to prove "There exists a $\delta > 0$ such that for every $\varepsilon > 0$, $2\delta < \varepsilon$", then you would have to select δ *first* before being given ε. In this case it is impossible to prove the statement, because it is false (why?).

Normally, when one has to prove a "There exists..." statement, e.g., "Prove that there exists an $\varepsilon > 0$ such that X is true", one proceeds by selecting ε carefully, and then showing that X is true for that ε. However, this sometimes requires a lot of foresight, and it is legitimate to defer the selection of ε until later in the argument, when it becomes clearer what properties ε needs to satisfy. The only thing to watch out for is to make sure that ε does not depend on any of the bound variables nested inside X. For instance:

Proposition A.6.2 *There exists an $\varepsilon > 0$ such that $\sin(x) > x/2$ for all $0 < x < \varepsilon$.*

Proof We pick $\varepsilon > 0$ to be chosen later, and let $0 < x < \varepsilon$. Since the derivative of $\sin(x)$ is $\cos(x)$, we see from the mean-value theorem (Corollary 10.2.9) we have

$$\frac{\sin(x)}{x} = \frac{\sin(x) - \sin(0)}{x - 0} = \cos(y)$$

for some $0 < y < x$. Thus in order to ensure that $\sin(x) > x/2$, it would suffice to ensure that $\cos(y) > 1/2$. To do this, it would suffice to ensure that $0 \leq y < \pi/3$ (since the cosine function takes the value of 1 at 0, takes the value of $1/2$ at $\pi/3$, and is decreasing in between). Since $0 < y < x$ and $0 < x < \varepsilon$, we see that $0 \leq y < \varepsilon$. Thus if we pick $\varepsilon := \pi/3$, then we have $0 \leq y < \pi/3$ as desired, and so we can ensure that $\sin(x) > x/2$ for all $0 < x < \varepsilon$. □

Note that the value of ε that we picked at the end did not depend on the nested variables x and y. This makes the above argument legitimate. Indeed, we can rearrange it so that we don't have to postpone anything:

Proof We choose $\varepsilon := \pi/3$; clearly $\varepsilon > 0$. Now we have to show that for all $0 < x < \pi/3$, we have $\sin(x) > x/2$. So let $0 < x < \pi/3$ be arbitrary. By the mean-value theorem we have

$$\frac{\sin(x)}{x} = \frac{\sin(x) - \sin(0)}{x - 0} = \cos(y)$$

for some $0 \leq y \leq x$. Since $0 \leq y \leq x$ and $0 < x < \pi/3$, we have $0 \leq y < \pi/3$. Thus $\cos(y) > \cos(\pi/3) = 1/2$, since cos is decreasing on the interval $[0, \pi/3]$. Thus we have $\sin(x)/x > 1/2$ and hence $\sin(x) > x/2$ as desired. □

If we had chosen ε to depend on x and y then the argument would not be valid, because ε is the outer variable and x, y are nested inside it.

A.7 Equality

As mentioned before, one can create statements by starting with expressions (such as $2 \times 3 + 5$) and then asking whether an expression obeys a certain property, or whether two expressions are related by some sort of relation ($=$, $\leq$, $\in$, etc.). There are many relations, but the most important one is *equality*, and it is worth spending a little time reviewing this concept.

Equality is a relation linking two objects x, y of the same type T (e.g., two integers, or two matrices, or two vectors, etc.). Given two such objects x and y, the statement $x = y$ may or may not be true; it depends on the value of x and y and also on how equality is defined for the class of objects under consideration. For instance, as real numbers, the two numbers $0.9999\ldots$ and 1 are equal. In modular arithmetic with modulus 10 (in which numbers are considered equal to their remainders modulo 10), the numbers 12 and 2 are considered equal, $12 = 2$, even though this is not the case in ordinary arithmetic.

How equality is defined depends on the class T of objects under consideration, and to some extent is just a matter of definition. However, for the purposes of logic we require that equality obeys the following four *axioms of equality*:

- (Reflexive axiom). Given any object x, we have $x = x$.
- (Symmetry axiom). Given any two objects x and y of the same type, if $x = y$, then $y = x$.
- (Transitive axiom). Given any three objects x, y, z of the same type, if $x = y$ and $y = z$, then $x = z$.
- (Substitution axiom). Given any two objects x and y of the same type, if $x = y$, then $f(x) = f(y)$ for all functions or operations f. Similarly, for any property $P(x)$ depending on x, if $x = y$, then $P(x)$ and $P(y)$ are equivalent statements.

The first three axioms are clear; together, they assert that equality is an *equivalence relation*. To illustrate the substitution we give some examples.

Example A.7.1 Let x and y be real numbers. If $x = y$, then $2x = 2y$, and $\sin(x) = \sin(y)$. Furthermore, $x + z = y + z$ for any real number z.

Example A.7.2 Let n and m be integers. If n is odd and $n = m$, then m must also be odd. If we have a third integer k, and we know that $n > k$ and $n = m$, then we also know that $m > k$.

Example A.7.3 Let x, y, z be real numbers. If we know that $x = \sin(y)$ and $y = z^2$, then (by the first form of the substitution axiom) we have $\sin(y) = \sin(z^2)$, and hence (by the transitive axiom) we have $x = \sin(z^2)$. One can also obtain the conclusion $x = \sin(z^2)$ more directly by using the second form of the substitution axiom.

Thus, from the point of view of logic, we can define equality on a class of objects however we please, so long as it obeys the reflexive, symmetry, and transitive axioms, and is consistent with all other operations on the class of objects under discussion in the sense that the substitution axiom was true for all of those operations. For instance, if we decided one day to modify the integers so that 12 was now equal to 2, one could only do so if one also made sure that 2 was now equal to 12, and that $f(2) = f(12)$ for any operation f on these modified integers. For instance, we now need $2 + 5$ to be equal to $12 + 5$. (In this case, pursuing this line of reasoning will eventually lead to modular arithmetic with modulus 10.)

For most applications in analysis, one should not need to compare objects of different types: for instance, if x is a set, and y is a number, then one should not need to consider the question of whether $x = y$ is true or false. But for the purposes of doing set theory, it is convenient to adopt the convention that the statement $x = y$ is automatically false if x, y are of different types; for instance, if one is treating natural numbers and vectors as objects of different types, then a natural number would not be equal to a vector. But sometimes we override this convention by identifying objects of one type with some objects of another type, e.g., when we identified natural numbers with their counterparts in the integers, or integers with their counterparts in the rationals, and so forth. This is technically an "abuse of notation", but can be tolerated as long as one verifies that no violation of the axioms of equality occur by doing so. We will sometimes use the notation $x \equiv y$ to indicate that a mathematical object x is being identified with a mathematical object y.

— Exercises —

Exercise A.7.1 Suppose you have four real numbers a, b, c, d and you know that $a = b$ and $c = d$. Use the above four axioms to deduce that $a + d = b + c$.

Appendix B
The Decimal System

In Chaps. 2, 4, and 5 we painstakingly constructed the basic number systems of mathematics: the natural numbers, integers, rationals, and reals. Natural numbers were simply postulated to exist, and to obey five axioms; the integers then came via (formal) differences of the natural numbers; the rationals then came from (formal) quotients of the integers; and the reals then came from (formal) limits of the rationals.

This is all very well and good, but it does seem somewhat alien to one's prior experience with these numbers. In particular, very little use was made of the *decimal system*, in which the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 are combined to represent these numbers. Indeed, except for a number of examples which were not essential to the main construction, the only decimals we really used were the numbers 0, 1, and 2, and the latter two can be rewritten as 0++ and (0++)++.

The basic reason for this is that *the decimal system itself is not essential to mathematics*. It is very convenient for computations, and we have grown accustomed to it thanks to a thousand years of use, but in the history of mathematics it is actually a comparatively recent invention. Numbers have been around for about ten thousand years (starting from scratch marks on cave walls), but the modern Hindu-Arabic base 10 system for representing numbers only dates from the eleventh century or so. Some early civilizations relied on other bases; for instance the Babylonians used a base 60 system (which still survives in our time system of hours, minutes, and seconds, and in our angular system of degrees, minutes, and seconds). And the ancient Greeks were able to do quite advanced mathematics, despite the fact that the most advanced number representation system available to them was the Roman numeral system $I, II, III, IV, \ldots$, which was horrendous for computations of even two-digit numbers. And of course modern computing relies on binary, hexadecimal, or byte-based (base 256) arithmetic instead of decimal, while analog computers such as the slide rule do not really rely on any number representation system at all. In fact, now that computers can do the menial work of number-crunching, there is very little use for decimals in modern mathematics. Indeed, we rarely use any numbers other than one-digit numbers or one-digit fractions (as well as e, π, i) explicitly in modern mathematical work; any more complicated numbers usually get called more generic names such as n.

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4

Nevertheless, the subject of decimals does deserve an appendix, because it is so integral to the way we use mathematics in our everyday life, and also because we do want to use such notation as 3.14159... to refer to real numbers, as opposed to the far clunkier "$\mathrm{LIM}_{n\to\infty} a_n$, where $a_1 = 3.1, a_2 := 3.14, a_3 := 3.141, \ldots$".

We begin by reviewing how the decimal system works for the positive integers and then turn to the reals. Note that in this discussion we shall freely use all the results from earlier chapters.

B.1 The Decimal Representation of Natural Numbers

In this section we will avoid the usual convention of abbreviating $a \times b$ as ab, since this would mean that decimals such as 34 might be misconstrued as 3×4.

Definition B.1.1 *(Digits)* A *digit* is any one of the ten symbols $0, 1, 2, 3, \ldots, 9$. We equate these digits with natural numbers by the formulae $0 \equiv 0, 1 \equiv 0{+}{+}, 2 \equiv 1{+}{+}$, etc. all the way up to $9 \equiv 8{+}{+}$. We also define the number ten by the formula $\mathrm{ten} := 9{+}{+}$. (We cannot use the decimal notation 10 to denote ten yet, because that presumes knowledge of the decimal system and would be circular.)

Definition B.1.2 *(Positive integer decimals)* A *positive integer decimal* is any string $a_n a_{n-1} \ldots a_0$ of digits, where $n \geq 0$ is a natural number, and the first digit a_n is non-zero. Thus, for instance, 3049 is a positive integer decimal, but 0493 or 0 is not. We equate each positive integer decimal with a positive integer by the formula

$$a_n a_{n-1} \ldots a_0 \equiv \sum_{i=0}^{n} a_i \times \mathrm{ten}^i .$$

Remark B.1.3 Note in particular that this definition implies that

$$10 = 0 \times \mathrm{ten}^0 + 1 \times \mathrm{ten}^1 = \mathrm{ten}$$

and thus we can write ten as the more familiar 10. Also, a single-digit integer decimal is exactly equal to that digit itself, e.g., the decimal 3 by the above definition is equal to

$$3 = 3 \times \mathrm{ten}^0 = 3$$

so there is no possibility of confusion between a single digit and a single digit decimal. (This is a subtle distinction, and not one which is worth losing much sleep over.)

Now we show that this decimal system indeed represents the positive integers. It is clear from the definition that every positive decimal representation gives a positive integer, since the sum consists entirely of natural numbers, and the last term $a_n\,\mathrm{ten}^n$ is non-zero by definition.

Theorem B.1.4 (Uniqueness and existence of decimal representations) *Every positive integer m is equal to exactly one positive integer decimal.*

Proof We shall use the principle of strong induction (Proposition 2.2.14, with $m_0 := 1$). For any positive integer m, let $P(m)$ denote the statement "m is equal to exactly one positive integer decimal". Suppose we already know $P(m')$ is true for all positive integers $m' < m$; we now wish to prove $P(m)$.

First observe that either $m \geq$ ten or $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. (This is easily proved by ordinary induction.) Suppose first that $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Then m clearly is equal to a positive integer decimal consisting of a single digit, and there is only one single-digit decimal which is equal to m. Furthermore, no decimal consisting of two or more digits can equal m, since if $a_n \dots a_0$ is such a decimal (with $n > 0$) we have

$$a_n \dots a_0 = \sum_{i=0}^{n} a_i \times \text{ten}^i \geq a_n \times \text{ten}^n \geq \text{ten} > m.$$

Now suppose that $m \geq$ ten. Then by the Euclidean algorithm (Proposition 2.3.9) we can write

$$m = s \times \text{ten} + r$$

where s is a positive integer, and $r \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Since

$$s < s \times \text{ten} \leq s \times \text{ten} + r = m$$

we can use the strong induction hypothesis and conclude that $P(s)$ is true. In particular, s has a decimal representation

$$s = b_p \dots b_0 = \sum_{i=0}^{p} b_i \times \text{ten}^i .$$

Multiplying by ten, we see that

$$s \times \text{ten} = \sum_{i=0}^{p} b_i \times \text{ten}^{i+1} = b_p \dots b_0 0,$$

and then adding r we see that

$$m = s \times \text{ten} + r = \sum_{i=0}^{p} b_i \times \text{ten}^{i+1} + r = b_p \dots b_0 r.$$

Thus m has at least one decimal representation. Now we need to show that m has at most one decimal representation. Suppose for sake of contradiction that we have at least two different representations

$$m = a_n \dots a_0 = a'_{n'} \dots a'_0.$$

First observe by the previous computation that

$$a_n \dots a_0 = (a_n \dots a_1) \times \text{ten} + a_0$$

and

$$a'_{n'} \dots a'_0 = (a'_{n'} \dots a'_1) \times \text{ten} + a'_0$$

and so after some algebra we obtain

$$a'_0 - a_0 = (a_n \dots a_1 - a'_{n'} \dots a'_1) \times \text{ten}\,.$$

The right-hand side is a multiple of ten, while the left-hand side lies strictly between $-$ ten and $+$ ten. Thus both sides must be equal to 0. This means that $a_0 = a'_0$ and $a_n \dots a_1 = a'_{n'} \dots a'_1$. But by previous arguments, we know that $a_n \dots a_1$ is a smaller integer than $a_n \dots a_0$. Thus by the strong induction hypothesis, the number $a_n \dots a_1$ has only one decimal representation, which means that n' must equal n and a'_i must equal a_i for all $i = 1, \dots, n$. Thus the decimals $a_n \dots a_0$ and $a'_{n'} \dots a'_0$ are in fact identical, contradicting the assumption that they were different. □

We refer to the decimal given by the above theorem as the *decimal representation* of m. Once one has this decimal representation, one can then derive the usual laws of long addition and long multiplication to connect the decimal representation of $x + y$ or $x \times y$ to that of x or y (Exercise B.1.1).

Once one has decimal representation of positive integers, one can of course represent negative integers decimally as well by using the $-$ sign. Finally, we let 0 be a decimal as well. This gives decimal representations of all integers. Every rational is then the ratio of two decimals, e.g., $335/113$ or $-1/2$ (with the denominator required to be non-zero, of course), though there may be more than one way to represent a rational as such a ratio, e.g., $6/4 = 3/2$.

Since ten $= 10$, we will now use 10 instead of ten throughout, as is customary.

— Exercises —

Exercise B.1.1 The purpose of this exercise is to demonstrate that the procedure of long addition taught to you in elementary school is actually valid. Let $A = a_n \dots a_0$ and $B = b_m \dots b_0$ be positive integer decimals. Let us adopt the convention that $a_i = 0$ when $i > n$, and $b_i = 0$ when $i > m$; for instance, if $A = 372$, then $a_0 = 2, a_1 = 7, a_2 = 3, a_3 = 0, a_4 = 0$, and so forth. Define the numbers $c_0, c_1, \dots$ and $\varepsilon_0, \varepsilon_1, \dots$ recursively by the following *long addition algorithm*.

- We set $\varepsilon_0 := 0$.
- Now suppose that ε_i has already been defined for some $i \geq 0$. If $a_i + b_i + \varepsilon_i < 10$, we set $c_i := a_i + b_i + \varepsilon_i$ and $\varepsilon_{i+1} := 0$; otherwise, if $a_i + b_i + \varepsilon_i \geq 10$, we set $c_i := a_i + b_i + \varepsilon_i - 10$ and $\varepsilon_{i+1} = 1$. (The number ε_{i+1} is the "carry digit" from the i^{th} decimal place to the $(i+1)^{th}$ decimal place.)

Prove that the numbers $c_0, c_1, \dots$ are all digits, and that there exists an l such that $c_l \neq 0$ and $c_i = 0$ for all $i > l$. Then show that $c_l c_{l-1} \dots c_1 c_0$ is the decimal representation of $A + B$.

Note that one could in fact use this algorithm to *define* addition, but it would look extremely complicated, and to prove even such simple facts as $(a + b) + c = a + (b + c)$ would be rather difficult. This is one of the reasons why we have avoided the decimal system in our construction of the natural numbers. The procedure for long multiplication (or long subtraction, or long division) is even worse to lay out rigorously; we will not do so here.

B.2 The Decimal Representation of Real Numbers

We need a new symbol: the *decimal point* ".".

Definition B.2.1 *(Real decimals)* A *real decimal* is any sequence of digits, and a decimal point, arranged as

$$\pm a_n \dots a_0.a_{-1}a_{-2} \dots$$

which is finite to the left of the decimal point (so n is a natural number), but infinite to the right of the decimal point, where $\pm$ is either $+$ or $-$, and $a_n \dots a_0$ is a natural number decimal (i.e., either a positive integer decimal, or 0). This decimal is equated to the real number

$$\pm a_n \dots a_0.a_{-1}a_{-2} \dots \equiv \pm 1 \times \sum_{i=-\infty}^{n} a_i \times 10^i.$$

The series is always convergent (Exercise B.2.1). Next, we show that every real number has at least one decimal representation:

Theorem B.2.2 (Existence of decimal representations) *Every real number x has at least one decimal representation*

$$x = \pm a_n \dots a_0.a_{-1}a_{-2} \dots.$$

Proof We first note that $x = 0$ has the decimal representation $0.000 \dots$. Also, once we find a decimal representation for x, we automatically get a decimal representation for $-x$ by changing the sign $\pm$. Thus it suffices to prove the theorem for positive real numbers x (by Proposition 5.4.4).

Let $n \geq 0$ be any natural number. From the Archimedean property (Corollary 5.4.13) we know that there is a natural number M such that $M \times 10^{-n} > x$. Since $0 \times 10^{-n} \leq x$, we thus see that there must exist a natural number s_n such that $s_n \times 10^{-n} \leq x$ and $s_n{+}{+} \times 10^{-n} > x$. (If no such natural number existed, one could use induction to conclude that $s \times 10^{-n} \leq x$ for all natural numbers s, contradicting the Archimedean property.)

Now consider the sequence $s_0, s_1, s_2, \dots$. Since we have

$$s_n \times 10^{-n} \leq x < (s_n + 1) \times 10^{-n}$$

we thus have

$$(10 \times s_n) \times 10^{-(n{+}{+})} \leq x < (10 \times s_n + 10) \times 10^{-(n{+}{+})}.$$

On the other hand, we have

$$s_{n+1} \times 10^{-(n+1)} \leq x < (s_{n+1} + 1) \times 10^{-(n+1)}$$

and hence we have

$$10 \times s_n < s_{n+1} + 1 \text{ and } s_{n+1} < 10 \times s_n + 10.$$

From these two inequalities we see that we have

$$10 \times s_n \leq s_{n+1} \leq 10 \times s_n + 9$$

and hence we can find a digit a_{n+1} such that

$$s_{n+1} = 10 \times s_n + a_{n+1}$$

and hence

$$s_{n+1} \times 10^{-(n+1)} = s_n \times 10^{-n} + a_{n+1} \times 10^{-(n+1)}.$$

From this identity and induction, we can obtain the formula

$$s_n \times 10^{-n} = s_0 + \sum_{i=0}^{n} a_i \times 10^{-i}.$$

Now we take limits of both sides (using Exercise B.2.1) to obtain

$$\lim_{n \to \infty} s_n \times 10^{-n} = s_0 + \sum_{i=0}^{\infty} a_i \times 10^{-i}.$$

On the other hand, we have

$$x - 10^{-n} \leq s_n \times 10^{-n} \leq x$$

for all n, so by the squeeze test (Corollary 6.4.14) we have

$$\lim_{n \to \infty} s_n \times 10^{-n} = x.$$

Thus we have

$$x = s_0 + \sum_{i=0}^{\infty} a_i \times 10^{-i}.$$

Since s_0 already has a positive integer decimal representation by Theorem B.1.4, we thus see that x has a decimal representation as desired. □

There is however one slight flaw with the decimal system: it is possible for one real number to have two decimal representations.

Proposition B.2.3 (Failure of uniqueness of decimal representations) *The number 1 has two different decimal representations:* $1.000\ldots$ *and* $0.999\ldots$.

Proof The representation $1 = 1.000\ldots$ is clear. Now let's compute $0.999\ldots$. By definition, this is the limit of the Cauchy sequence

$$0.9, 0.99, 0.999, 0.9999, \ldots.$$

But this sequence has 1 as a formal limit by Proposition 5.2.8. □

It turns out that these are the only two decimal representations of 1 (Exercise B.2.2). In fact, as it turns out, all real numbers have either one or two decimal representations—two if the real is a terminating decimal, and one otherwise (Exercise B.2.3).

— Exercises —

Exercise B.2.1 If $a_n \ldots a_0.a_{-1}a_{-2}\ldots$ is a real decimal, show that the series $\sum_{i=-\infty}^{n} a_i \times 10^i$ is absolutely convergent.

Exercise B.2.2 Show that the only decimal representations

$$1 = \pm a_n \ldots a_0.a_{-1}a_{-2}\ldots$$

of 1 are $1 = 1.000\ldots$ and $1 = 0.999\ldots$.

Exercise B.2.3 A real number x is said to be a *terminating decimal* if we have $x = n/10^{-m}$ for some integers n, m. Show that if x is a terminating decimal, then x has exactly two decimal representations, while if x is not at terminating decimal, then x has exactly one decimal representation.

Exercise B.2.4 Rewrite the proof of Corollary 8.3.4 using the decimal system.

Index

Symbols

++ (increment), 14, 41
  on integers, 65
$+C$, 259
α-length, 252
ε-adherent, 120, 185
  continually ε-adherent, 120
ε-close
  eventual, 87
  functions, 191
  local, 191
  rationals, 74
  reals, 109
  sequences, 87, 110
ε-steady, 83, 110
  eventually ε-steady, 84, 110

A

Absolute convergence
  for series, 146, 168
  test, 146
Absolute value
  for rationals, 73
  for reals, 96
Absorption laws, 37
Abstraction, 18
Abuse of notation, 43, 285
Addition
  of cardinals, 61
  of functions, 190
  of integers, 65
  long, 290
  of natural numbers, 20
  of rationals, 69
  of reals, 89
Adherent point
  infinite, 215
  of sequences:, *see* limit point of sequences
  of sets, 185
Alternating series test, 146
Analysis, 1
And:, *see* conjunction
Antiderivative, 257
a posteriori, 15
a priori, 15
Archimedean property, 98
Aristotlean logic, 282
Associativity
  of addition in $mathbf N$, 21
  see also ring, field, laws of algebra
  of composition, 43, 44
  of multiplication in **N**, 24
Asymptotic discontinuity, 202
Axiom(s)
  of choice, 29, 54, 175
  of comprehension:, *see* Axiom of universal specification
  of countable choice, 175
  of the empty set, 29
  of equality, 284
  of foundation:. *see* Axiom of regularity
  of induction:, *see* principle of mathematical induction
  of infinity, 36
  in mathematics, 18
  of natural numbers:, *see* Peano axioms
  of pairwise union, 30
  of power set, 49
  of reflexivity, 285
  of regularity, 39
  of replacement, 35
  of separation, 33

T. Tao, *Analysis I*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7261-4

of set theory, 28–30, 32, 35–36, 39, 49
of singleton sets and pair sets, 30
of specification, 33
of substitution, 42, 285
of symmetry, 285
of transitivity, 285
of union, 50
of universal specification, 38

B
Bijection, 45
Binomial formula, 143
Bolzano–Weierstrass theorem, 130
Boolean algebra, 34
Boolean logic, 277
Bounded
from above and below, 203
function, 203
interval, 184
sequence, 85, 112
sequence away from zero, 92, 96
set, 187
Bound variable, 136, 278, 283

C
Cancellation law
of addition in **N**, 21
of multiplication in **N**, 25
of multiplication in **Z**, 68
of multiplication in **R**, 94
Cantor's theorem, 171
Cardinality
arithmetic of, 60
of finite sets, 60
uniqueness of, 60
Cartesian product, 52, 53
infinite, 174
Cauchy criterion, 149
Cauchy sequence, 84, 110
Chain:, *see* totally ordered set
Chain rule, 221
Change of variables formula, 261–271
Choice
arbitrary, 175
countable, 175
finite, 54
single, 29
Closed
interval, 184
set, 186
Closure, 185
Cluster point:, *see* limit point
Codomain, 40
Common refinement, 234
Commutativity
of addition in **N**, 21
of multiplication in **N**, 24
Comparison principle (or test)
for finite series, 137
for infinite series, 148
for sequences, 124
Completeness
of the reals, 125
Composition of functions, 43
Conjunction (and), 268
Connectedness, 232
Constant
function, 42, 235
sequence, 127
Continuity, 197
and convergence, 192
Continuum, 183
hypothesis, 173
Contrapositive, 274
Convergence
absolute:, *see* absolute convergence
conditional, 146
of a function at a point, 192
of sequences, 110
of series, 144
Converse, 273
Corollary, 20
Countability, 159
of the integers, 162
of the rationals, 164

D
Decimal
negative integer, 290
non-uniqueness of representation, 293
point, 291
positive integer, 290
real, 291
De Morgan laws, 34
Denumerable:, *see* countable
Derivative, 217
Difference rule, 220
Difference set, 34
Differentiability
at a point, 217
Digit, 288
Discontinuity:, *see* singularity
Disjoint sets, 34

Disjunction (or), 269
  inclusive vs. exclusive, 269
Distance
  in **Q**, 73
  in **R**, 109
Distributive law
  See also laws of algebra
  for natural numbers, 24
Divergence
  of sequences, 3
  of series, 2, 144
Divisibility, 180
Division
  formal (//), 69
  of functions, 190
  of rationals, 72
  by zero, 2
Domain, 40
Doubly infinite, 184
Dummy variable:, *see* bound variable

E
Empty
  Cartesian product, 54
  function, 43
  sequence, 54
  series, 140
  set, 29
Equality, 284
  for functions, 42
  for sets, 28
  of cardinality, 58
Equivalence
  of sequences, 87, 212
  relation, 285
Euclid's division lemma, 26
Exponentiation
  with base and exponent in **N**, 26
  with base in **Q** and exponent in **Z**, 75, 76
  with base in **R** and exponent in **Z**, 104
  with base in $\mathbf{R}^+$ and exponent in **Q**, 106
  with base in $\mathbf{R}^+$ and exponent in **R**, 132
  of cardinals, 60
Expression, 268
Extended real number system $\mathbf{R}^*$, 102, 115
Extremum:, *see* maximum, minimum

F
Factorial, 62, 143
Family, 50
Field, 71
  ordered, 73
Finite set, 60
Fixed point theorem, 208
Forward image:, *see* image
Free variable, 278
Fubini's theorem
  for finite series, 142
  for infinite series, 165
  See also interchanging integrals/sums with integrals/sums
Function, 40
  implicit definition, 42
Fundamental theorems of calculus, 256, 258

G
Geometric series, 144, 149
  formula, 149, 152
Graph, 42, 56, 190
Greatest lower bound:, *see* least upper bound

H
Half-infinite, 184
Half-open, 184
Harmonic series, 151
Hausdorff maximality principle, 182
Heine–Borel theorem
  for the real line, 187

I
Identity map (or operator), 47
If:, *see* implication
Iff (if and only if), 22
Ill-defined, 266, 268
Image
  inverse image, 47
  of sets, 47
Implication (if), 271
Improper integral, 240
Inclusion map, 47
Inconsistent, 173, 174
Index of summation:, *see* dummy variable
Index set, 50
Induction:, *see* Principle of mathematical induction
Infimum:, *see* supremum
Infinite
  interval, 184
  set, 60
Injection:, *see* one-to-one function
Integer part, 77, 99
Integers **Z**

definition, 63
identification with rationals, 70
interspersing with rationals, 76
Integral test, 251
Integration
by parts, 260–261
laws, 238, 242
piecewise constant, 236, 237
Riemann:, *see* Riemann integral
Interchanging
derivatives with derivatives, 8
finite sums with finite sums, 141, 143
integrals with integrals, 1
limits with derivatives, 8
limits with integrals, 7
limits with length, 10
limits with limits, 7
sums with sums, 4, 165
Intermediate value theorem, 206
Intersection
pairwise, 33
Interval, 184
Inverse
function theorem, 226
image, 47
in logic, 274
of functions, 46
Irrationality, 81
of $\sqrt{2}$, 77, 103
Isolated point, 186

J
Jump discontinuity, 202

L
$l^1, l^2, l^\infty, L^1, L^2, L^\infty$
See also supremum as norm
Label, 50
Laws of algebra
for integers, 66
for rationals, 71
for reals, 91
Laws of exponentiation, 75, 76, 105, 107
Least upper bound, 101
least upper bound property, 101, 118
See also supremum
Leibniz rule, 221
Lemma, 20
Length of interval, 232
L'Hôpital's rule, 9, 228
Limit
formal (LIM), 89, 112
at infinity, 216
laws, 113, 193
left and right, 200
limiting values of functions, 4, 191
of sequences, 111
uniqueness of, 111, 193
Limit inferior, *see* limit superior
Limit point
of sequences, 120
of sets, 186
Limit superior, 121
Linearity
of finite series, 140
of infinite series, 147
of integration, 237, 242
of limits, 113
Lipschitz constant, 224
Lipschitz continuous, 224
Logical connective, 268
Lower bound:, *see* upper bound

M
Majorize, 239
Maximum, 100, 177, 222
local, 222
of functions, 190, 204
principle, 204
Mean value theorem, 224
Meta-proof, 105
Metric
see also distance
Minimum, 100, 177, 222
local, 222
of a set of natural numbers, 160
of functions, 190, 204
Minorize:, *see* majorize
Monotone (increasing or decreasing)
function, 208, 254
sequence, 119
Morphism:, *see* function
Multiplication
of cardinals, 60
of functions, 191
of integers, 64
of natural numbers, 24
of rationals, 69, 70
of reals, 90

N
Natural numbers **N**

axioms:, *see* Peano axioms
informal definition, 13
Natural numbers **N**
are infinite, 60
identification with integers, 65
in set theory:, *see* Axiom of infinity
uniqueness of, 57
Negation
of extended reals, 115
of integers, 65
in logic, 269
of rationals, 69
of reals, 91
Negative:, *see* negation, positive
Newton's approximation, 219
Non-constructive, 175

O
Objects, 28
primitive, 39
One-to-one correspondence:, *see* bijection
One-to-one function, 45
Onto, 45
Open
interval, 184
Or:, *see* disjunction
Ordered n-tuple, 52
Ordered pair, 52
construction of, 55
Order ideal, 181
Ordering
lexicographical, 181
of cardinals, 173
of the extended reals, 115
of the integers, 68
of the natural numbers, 22
of orderings, 182
of partitions, 234
of the rationals, 72
of the reals, 97
of sets, 177
Oscillatory discontinuity, 202

P
Pair set, 30
Partial function, 51
Partially ordered set, 32, 176
Partial sum, 144
Partition, 233
Peano axioms, 14–17
Perfect matching:, *see* bijection
Permutation, 62
Piecewise
constant, 235
constant Riemann–Stieltjes integral, 253
continuous, 249
Pigeonhole principle, 61
Polynomial, 200
Positive
integer, 65
natural number, 22
rational, 72
real, 96
Power set, 49
Principle of infinite descent, 78
Principle of mathematical induction, 16
backwards induction, 24
strong induction, 23, 178
Product rule. *see* Leibniz rule
Proof
abstract examples, 275–277, 283–284
by contradiction, 267, 274
Proper subset, 31
Property, 268
Propositional logic, 277

Q
Quantifier, 278
existential (for some), 279
negation of, 281
nested, 280
universal (for all), 279
Quotient rule, 220
Quotient:, *see* division

R
Range, 47
Rational numbers **Q**
definition, 69
identification with reals, 91
interspersing with rationals, 76
interspersing with reals, 98
Ratio test, 157
Real numbers **R**
definition, 88
Rearrangement
of absolutely convergent series, 153
of divergent series, 154, 169
of finite series, 140
of non-negative series, 152
Reciprocal
of rationals, 71

of reals, 94

Recursive definitions, 19, 57

reductio ad absurdum:, *see* proof by contradiction

Removable discontinuity:, *see* removable singularity

Removable singularity, 195, 202

Restriction of functions, 189

Riemann hypothesis, 151

Riemann integrability, 240

of bounded continuous functions, 248

closure properties, 242–246

of continuous functions on compacta, 248

failure of, 251

of monotone functions, 249

of piecewise continuous bounded functions, 249

of uniformly continuous functions, 246

Riemann integral, 240

upper and lower, 239

Riemann–Stieltjes integral, 254

Riemann sums (upper and lower), 242

Riemann zeta function, 151

Ring, 67

commutative, 67

Rolle's theorem, 223

Root, 105

test, 155

Russell's paradox, 38

S

Scalar multiplication

of functions, 191

Schröder–Bernstein theorem, 173

Sequence, 82

finite, 54

Series

on arbitrary sets, 168

on countable sets, 165

finite, 135, 137

formal infinite, 144

laws, 147, 168

vs. sum, 136

Set

axioms:, *see* axioms of set theory

informal definition, 27

Signum function, 194

Singleton set, 30

Singularity, 203

Square root, 41

Squeeze test

for sequences, 124

Statement, 266

Strict upper bound, 178

Subsequence, 128

Subset, 31

proper, 37, 60

Substitution:, *see* rearrangement

Subtraction

formal (−−), 64

of functions, 190

of integers, 67

Successor, 14

Sum rule, 220

Supremum (and infimum)

of a set of extended reals, 117, 118

of a set of reals, 102, 104

of sequences of reals, 118

Surjection:, *see* onto

T

Tangent:, *see* trigonometric function

Telescoping series, 148

Ten, 288

Theorem, 20

Totally ordered set, 32, 177

Transformation:, *see* function

Triangle inequality

for finite series, 137, 141

in **R**, 74

Trichotomy of order

of extended reals, 116

for integers, 68

for natural numbers, 22

for rationals, 72

for reals, 96

Two-to-one function, 45

U

Unbounded set, 187

Uncountability, 159

of the reals, 171

Undecidable, 174

Uniform continuity, 211

Union, 50

pairwise, 30

Universal set, 39

Upper bound

of a partially ordered set, 178

of a set of reals, 100

see also least upper bound

V
Variable, 277
Vertical line test, 40, 56

W
Well ordering principle
  for arbitrary sets, 182
  for natural numbers, 160
Well-defined, 266
Well-ordered sets, 178

Z
Zermelo–Fraenkel(–Choice) axioms, 51
  see also axioms of set theory
Zero test
  for sequences, 125
  for series, 145
Zorn's lemma, 180

Texts and Readings in Mathematics 38

Terence Tao

Analysis II

Fourth Edition

HINDUSTAN BOOK AGENCY

Springer

Texts and Readings in Mathematics

The **Texts and Readings in Mathematics** series publishes high-quality textbooks, research-level monographs, lecture notes and contributed volumes. Undergraduate and graduate students of mathematics, research scholars and teachers would find this book series useful. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Books in this series are co-published with Hindustan Book Agency, New Delhi, India.

Terence Tao

Analysis II

Fourth Edition

HINDUSTAN
BOOK AGENCY

Springer

Terence Tao
Department of Mathematics
University of California Los Angeles
Los Angeles, CA, USA

ISSN 2366-8717 ISSN 2366-8725 (electronic)
Texts and Readings in Mathematics
ISBN ISBN 978-981-19-7284-3 (eBook)
https://doi.org/10.1007/978-981-19-7284-3

Jointly published with Hindustan Book Agency
This work is a co-publication with Hindustan Book Agency, New Delhi, licensed for sale in all countries in electronic form only. Sold and distributed in print across the world by Hindustan Book Agency, P-19 Green Park Extension, New Delhi 110016, India.
ISBN: 978-81-951961-2-8 © Hindustan Book Agency 2022

Springer has only electronic right

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To my parents, for everything

Preface to the First Edition

This text originated from the lecture notes I gave teaching the honours undergraduate-level real analysis sequence at the University of California, Los Angeles, in 2003. Among the undergraduates here, real analysis was viewed as being one of the most difficult courses to learn, not only because of the abstract concepts being introduced for the first time (e.g., topology, limits, measurability, etc.), but also because of the level of rigour and proof demanded of the course. Because of this perception of difficulty, one was often faced with the difficult choice of either reducing the level of rigour in the course in order to make it easier, or to maintain strict standards and face the prospect of many undergraduates, even many of the bright and enthusiastic ones, struggling with the course material.

Faced with this dilemma, I tried a somewhat unusual approach to the subject. Typically, an introductory sequence in real analysis assumes that the students are already familiar with the real numbers, with mathematical induction, with elementary calculus, and with the basics of set theory, and then quickly launches into the heart of the subject, for instance the concept of a limit. Normally, students entering this sequence do indeed have a fair bit of exposure to these prerequisite topics, though in most cases the material is not covered in a thorough manner. For instance, very few students were able to actually *define* a real number, or even an integer, properly, even though they could visualize these numbers intuitively and manipulate them algebraically. This seemed to me to be a missed opportunity. Real analysis is one of the first subjects (together with linear algebra and abstract algebra) that a student encounters, in which one truly has to grapple with the subtleties of a truly rigorous mathematical proof. As such, the course offered an excellent chance to go back to the foundations of mathematics, and in particular the opportunity to do a proper and thorough construction of the real numbers.

Thus the course was structured as follows. In the first week, I described some well-known "paradoxes" in analysis, in which standard laws of the subject (e.g., interchange of limits and sums, or sums and integrals) were applied in a non-rigorous way to give nonsensical results such as $0 = 1$. This motivated the need to go back to the very beginning of the subject, even to the very definition of the natural numbers, and check all the foundations from scratch. For instance, one of the first homework assignments was to check (using only the Peano axioms) that addition was associative for natural numbers (i.e., that $(a + b) + c = a + (b + c)$ for all natural numbers a, b, c: see Exercise 2.2.1). Thus even in the first week, the students had to write rigorous proofs using mathematical induction. After we had derived all the basic properties of the natural numbers, we then moved on to the integers (initially defined as formal differences of natural numbers); once the students had verified all the basic properties of the integers, we moved on to the rationals (initially defined as formal quotients of integers); and then from there we moved on (via formal limits of Cauchy sequences) to the reals. Around the same time, we covered the basics of set theory, for instance demonstrating the uncountability of the reals. Only then (after about ten lectures) did we begin what one normally considers the heart of undergraduate real analysis—limits, continuity, differentiability, and so forth.

The response to this format was quite interesting. In the first few weeks, the students found the material very easy on a conceptual level, as we were dealing only with the basic properties of the standard number systems. But on an intellectual level it was very challenging, as one was analyzing these number systems from a foundational viewpoint, in order to rigorously derive the more advanced facts about these number systems from the more primitive ones. One student told me how difficult it was to explain to his friends in the non-honours real analysis sequence (a) why he was still learning how to show why all rational numbers are either positive, negative, or zero (Exercise 4.2.4), while the non-honours sequence was already distinguishing absolutely convergent and convergent series, and (b) why, despite this, he thought his homework was significantly harder than that of his friends. Another student commented to me, quite wryly, that while she could obviously *see* why one could always divide a natural number n into a positive integer q to give a quotient a and a remainder r less than q (Exercise 2.3.5), she still had, to her frustration, much difficulty in writing down a proof of this fact. (I told her that later in the course she would have to prove statements for which it would not be as obvious to see that the statements were true; she did not seem to be particularly consoled by this.) Nevertheless, these students greatly enjoyed the homework, as when they did perservere and obtain a rigorous proof of an intuitive fact, it solidified the link in their minds between the abstract manipulations of formal mathematics and their informal intuition of mathematics (and of the real world), often in a very satisfying way. By the time they were assigned the task of giving the infamous "epsilon and delta" proofs in real analysis, they had already had so much experience with formalizing intuition, and in discerning the subtleties of mathematical logic (such as the distinction between the "for all" quantifier and the "there exists" quantifier), that the transition to these proofs was fairly smooth, and we were able to cover material both thoroughly and rapidly. By the tenth week, we had caught up with the non-honours class, and the students were verifying the change of variables formula for Riemann–Stieltjes integrals, and showing that piecewise continuous functions were Riemann integrable. By the conclusion of the sequence in the twentieth week, we had covered (both in lecture and in homework) the convergence theory of Taylor and Fourier series, the inverse and implicit function theorem for continuously differentiable functions of several variables, and established the dominated convergence theorem for the Lebesgue integral.

In order to cover this much material, many of the key foundational results were left to the student to prove as homework; indeed, this was an essential aspect of the course, as it ensured the students truly appreciated the concepts as they were being introduced. This format has been retained in this text; the majority of the exercises consist of proving lemmas, propositions and theorems in the main text. Indeed, I would strongly recommend that one do as many of these exercises as possible—and this includes those exercises proving "obvious" statements—if one wishes to use this text to learn real analysis; this is not a subject whose subtleties are easily appreciated just from passive reading. Most of the chapter sections have a number of exercises, which are listed at the end of the section.

To the expert mathematician, the pace of this book may seem somewhat slow, especially in early chapters, as there is a heavy emphasis on rigour (except for those discussions explicitly marked "Informal"), and justifying many steps that would ordinarily be quickly passed over as being self-evident. The first few chapters develop (in painful detail) many of the "obvious" properties of the standard number systems, for instance that the sum of two positive real numbers is again positive (Exercise 5.4.1), or that given any two distinct real numbers, one can find rational number between them (Exercise 5.4.5). In these foundational chapters, there is also an emphasis on *non-circularity*—not using later, more advanced results to prove earlier, more primitive ones. In particular, the usual laws of algebra are not used until they are derived (and they have to be derived separately for the natural numbers, integers, rationals, and reals). The reason for this is that it allows the students to learn the art of abstract reasoning, deducing true facts from a limited set of assumptions, in the friendly and intuitive setting of number systems; the payoff for this practice comes later, when one has to utilize the same type of reasoning techniques to grapple with more advanced concepts (e.g., the Lebesgue integral).

The text here evolved from my lecture notes on the subject, and thus is very much oriented towards a pedagogical perspective; much of the key material is contained inside exercises, and in many cases I have chosen to give a lengthy and tedious, but instructive, proof instead of a slick abstract proof. In more advanced textbooks, the student will see shorter and more conceptually coherent treatments of this material, and with more emphasis on intuition than on rigour; however, I feel it is important to know how to do analysis rigorously and "by hand" first, in order to truly appreciate the more modern, intuitive and abstract approach to analysis that one uses at the graduate level and beyond.

The exposition in this book heavily emphasizes rigour and formalism; however this does not necessarily mean that lectures based on this book have to proceed the same way. Indeed, in my own teaching I have used the lecture time to present the intuition behind the concepts (drawing many informal pictures and giving examples), thus providing a complementary viewpoint to the formal presentation in the text. The exercises assigned as homework provide an essential bridge between the two, requiring the student to combine both intuition and formal understanding together in order to locate correct proofs for a problem. This I found to be the most difficult task for the students, as it requires the subject to be genuinely *learnt*, rather than merely memorized or vaguely absorbed. Nevertheless, the feedback I received from the students was that the homework, while very demanding for this reason, was also very rewarding, as it allowed them to connect the rather abstract manipulations of formal mathematics with their innate intuition on such basic concepts as numbers, sets, and functions. Of course, the aid of a good teaching assistant is invaluable in achieving this connection.

With regard to examinations for a course based on this text, I would recommend either an open-book, open-notes examination with problems similar to the exercises given in the text (but perhaps shorter, with no unusual trickery involved), or else a take-home examination that involves problems comparable to the more intricate exercises in the text. The subject matter is too vast to force the students to memorize the definitions and theorems, so I would not recommend a closed-book examination, or an examination based on regurgitating extracts from the book. (Indeed, in my own examinations I gave a supplemental sheet listing the key definitions and theorems which were relevant to the examination problems.) Making the examinations similar to the homework assigned in the course will also help motivate the students to work through and understand their homework problems as thoroughly as possible (as opposed to, say, using flash cards or other such devices to memorize material), which is good preparation not only for examinations but for doing mathematics in general.

Some of the material in this textbook is somewhat peripheral to the main theme and may be omitted for reasons of time constraints. For instance, as set theory is not as fundamental to analysis as are the number systems, the chapters on set theory (Chapters 3, 8) can be covered more quickly and with substantially less rigour, or be given as reading assignments. The appendices on logic and the decimal system are intended as optional or supplemental reading and would probably not be covered in the main course lectures; the appendix on logic is particularly suitable for reading concurrently with the first few chapters. Also, Chapter 5 (on Fourier series) is not needed elsewhere in the text and can be omitted.

For reasons of length, this textbook has been split into two volumes. The first volume is slightly longer, but can be covered in about thirty lectures if the peripheral material is omitted or abridged. The second volume refers at times to the first, but can also be taught to students who have had a first course in analysis from other sources. It also takes about thirty lectures to cover.

I am deeply indebted to my students, who over the progression of the real analysis course corrected several errors in the lectures notes from which this text is derived, and gave other valuable feedback. I am also very grateful to the many anonymous referees who made several corrections and suggested many important improvements to the text. I also thank Adam, James Ameril, Quentin Batista, Biswaranjan Behara, José Antonio Lara Benítez, Dingjun Bian, Petrus Bianchi, Phillip Blagoveschensky, Tai-Danae Bradley, Brian, Eduardo Buscicchio, Carlos, cebismellim, Matheus Silva Costa, Gonzales Castillo Cristhian, Ck, William Deng, Kevin Doran, Lorenzo Dragani, EO, Florian, Gyao Gamm, Evangelos Georgiadis, Aditya Ghosh, Elie Goudout, Ti Gong, Ulrich Groh, Gökhan Güçlü, Yaver Gulusoy, Christian Gz., Kyle Hambrook, Minyoung Jeong, Bart Kleijngeld, Erik Koelink, Brett Lane, David Latorre, Matthis Lehmkühler, Bin Li, Percy Li, Ming Li, Mufei Li, Zijun Liu, Rami Luisto, Jason M., Manoranjan Majji, Mercedes Mata, Simon Mayer, Geoff Mess, Pieter Naaijkens, Vineet Nair, Jorge Peña-Vélez, Cristina Pereyra, Huaying Qiu, David Radnell, Tim Reijnders, Issa Rice, Eric Rodriquez, Pieter Roffelsen, Luke Rogers, Feras Saad, Gabriel Salmerón, Vijay Sarthak, Leopold Schlicht, Marc Schoolderman, SkysubO, Rainer aus dem Spring, Sundar, Rafał Szlendak, Karim Taya, Chaitanya Tappu, Winston Tsai, Kent Van Vels, Andrew Verras, Murtaza Wani, Daan Wanrooy, John Waters, Yandong Xiao, Sam Xu, Xueping, Hongjiang Ye, Luqing Ye, Muhammad Atif Zaheer, Zelin, and the students of Math 401/501 and Math 402/502 at the University of New Mexico for corrections to the first, second, and third editions.

Terence Tao

Preface to Subsequent Editions

Since the publication of the first edition, many students and lecturers have communicated a number of minor typos and other corrections to me. There was also some demand for a hardcover edition of the texts. Because of this, the publishers and I have decided to incorporate the corrections and issue a hardcover second edition of the textbooks. The layout, page numbering, and indexing of the texts have also been changed; in particular the two volumes are now numbered and indexed separately. However, the chapter and exercise numbering, as well as the mathematical content, remains the same as the first edition, and so the two editions can be used more or less interchangeably for homework and study purposes.

The third edition contains a number of corrections that were reported for the second edition, together with a few new exercises, but are otherwise essentially the same text. The fourth edition similarly incorporates a large number of additional corrections reported since the release of the third edition, as well as some additional exercises.

Los Angeles, USA Terence Tao

Contents

1 Metric Spaces 1
1.1 Definitions and Examples 1
1.2 Some Point-Set Topology of Metric Spaces 9
1.3 Relative Topology 13
1.4 Cauchy Sequences and Complete Metric Spaces 14
1.5 Compact Metric Spaces 18

2 Continuous Functions on Metric Spaces 23
2.1 Continuous Functions 23
2.2 Continuity and Product Spaces 25
2.3 Continuity and Compactness 28
2.4 Continuity and Connectedness 30
2.5 Topological Spaces (Optional) 33

3 Uniform Convergence 39
3.1 Limiting Values of Functions 40
3.2 Pointwise and Uniform Convergence 42
3.3 Uniform Convergence and Continuity 46
3.4 The Metric of Uniform Convergence 48
3.5 Series of Functions; the Weierstrass M-Test 51
3.6 Uniform Convergence and Integration 53
3.7 Uniform Convergence and Derivatives 55
3.8 Uniform Approximation by Polynomials 58

4 Power Series 65
4.1 Formal Power Series 65
4.2 Real Analytic Functions 67
4.3 Abel's Theorem 72
4.4 Multiplication of Power Series 75
4.5 The Exponential and Logarithm Functions 78
4.6 A Digression on Complex Numbers 81
4.7 Trigonometric Functions 88

5 Fourier Series . . . 93
5.1 Periodic Functions . . . 94
5.2 Inner Products on Periodic Functions . . . 96
5.3 Trigonometric Polynomials . . . 99
5.4 Periodic Convolutions . . . 102
5.5 The Fourier and Plancherel Theorems . . . 106

6 Several Variable Differential Calculus . . . 113
6.1 Linear Transformations . . . 113
6.2 Derivatives in Several Variable Calculus . . . 119
6.3 Partial and Directional Derivatives . . . 122
6.4 The Several Variable Calculus Chain Rule . . . 129
6.5 Double Derivatives and Clairaut's Theorem . . . 131
6.6 The Contraction Mapping Theorem . . . 133
6.7 The Inverse Function Theorem in Several Variable Calculus . . . 136
6.8 The Implicit Function Theorem . . . 141

7 Lebesgue Measure . . . 145
7.1 The Goal: Lebesgue Measure . . . 146
7.2 First Attempt: Outer Measure . . . 148
7.3 Outer Measure Is not Additive . . . 156
7.4 Measurable Sets . . . 158
7.5 Measurable Functions . . . 163

8 Lebesgue Integration . . . 167
8.1 Simple Functions . . . 167
8.2 Integration of Non-negative Measurable Functions . . . 172
8.3 Integration of Absolutely Integrable Functions . . . 179
8.4 Comparison with the Riemann Integral . . . 184
8.5 Fubini's Theorem . . . 185

Index . . . 191

About the Author

Terence Tao has been a professor of Mathematics at the University of California Los Angeles (UCLA), USA, since 1999, having completed his Ph.D. under Prof. Elias Stein at Princeton University, USA, in 1996. Tao's areas of research include harmonic analysis, partial differential equations, combinatorics, and number theory. He has received a number of awards, including the Salem Prize in 2000, the Bochner Prize in 2002, the Fields Medal in 2006, the MacArthur Fellowship in 2007, the Waterman Award in 2008, the Nemmers Prize in 2010, the Crafoord Prize in 2012, and the Breakthrough Prize in Mathematics in 2015. Terence Tao also currently holds the James and Carol Collins chair in Mathematics at UCLA and is a fellow of the Royal Society, the Australian Academy of Sciences (the corresponding member), the National Academy of Sciences (a foreign member), and the American Academy of Arts and Sciences. He was born in Adelaide, Australia, in 1975.

Chapter 1
Metric Spaces

1.1 Definitions and Examples

In Definition 6.1.5 we defined what it meant for a sequence $(x_n)_{n=m}^\infty$ of real numbers to converge to another real number x; indeed, this meant that for every $\varepsilon > 0$, there exists an $N \geq m$ such that $|x - x_n| \leq \varepsilon$ for all $n \geq N$. When this is the case, we write $\lim_{n\to\infty} x_n = x$.

Intuitively, when a sequence $(x_n)_{n=m}^\infty$ converges to a limit x, this means that somehow the elements x_n of that sequence will eventually be as close to x as one pleases. One way to phrase this more precisely is to introduce the *distance function* $d(x, y)$ between two real numbers by $d(x, y) := |x - y|$. (Thus for instance $d(3, 5) = 2$, $d(5, 3) = 2$, and $d(3, 3) = 0$.) Then we have

Lemma 1.1.1 *Let $(x_n)_{n=m}^\infty$ be a sequence of real numbers, and let x be another real number. Then $(x_n)_{n=m}^\infty$ converges to x if and only if* $\lim_{n\to\infty} d(x_n, x) = 0$.

Proof See Exercise 1.1.1. □

One would now like to generalize this notion of convergence, so that one can take limits not just of sequences of real numbers, but also sequences of complex numbers, or sequences of vectors, or sequences of matrices, or sequences of functions, even sequences of sequences. One way to do this is to redefine the notion of convergence each time we deal with a new type of object. As you can guess, this will quickly get tedious. A more efficient way is to work *abstractly*, defining a very general class of spaces—which includes such standard spaces as the real numbers, complex numbers, vectors, etc.—and define the notion of convergence on this entire class of spaces at once. (A *space* is just the set of all objects of a certain type—the space of all real numbers, the space of all 3×3 matrices, etc. Mathematically, there is not much distinction between a space and a set, except that spaces tend to have much more structure than what a random set would have. For instance, the space of real numbers comes with operations such as addition and multiplication, while a general set would not.)

It turns out that there are two very useful classes of spaces which do the job. The first class is that of *metric spaces*, which we will study here. There is a more general

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_1

class of spaces, called *topological spaces*, which is also very important, but we will only deal with this generalization briefly, in Sect. 2.5.

Roughly speaking, a metric space is any space X which has a concept of *distance* $d(x, y)$—and this distance should behave in a reasonable manner. More precisely, we have

Definition 1.1.2 (*Metric spaces*) A *metric space* (X, d) is a space X of objects (called *points*), together with a *distance function* or *metric* $d : X \times X \to [0, +\infty)$, which associates to each pair x, y of points in X a non-negative real number $d(x, y) \geq 0$. Furthermore, the metric must satisfy the following four axioms:

(a) For any $x \in X$, we have $d(x, x) = 0$.
(b) (Positivity) For any *distinct* $x, y \in X$, we have $d(x, y) > 0$.
(c) (Symmetry) For any $x, y \in X$, we have $d(x, y) = d(y, x)$.
(d) (Triangle inequality) For any $x, y, z \in X$, we have $d(x, z) \leq d(x, y) + d(y, z)$.

In many cases it will be clear what the metric d is, and we shall abbreviate (X, d) as just X.

Remark 1.1.3 The conditions (a) and (b) can be rephrased as follows: for any $x, y \in X$ we have $d(x, y) = 0$ if and only if $x = y$. (Why is this equivalent to (a) and (b)?)

Example 1.1.4 (The real line) Let $\mathbf{R}$ be the real numbers, and let $d : \mathbf{R} \times \mathbf{R} \to [0, \infty)$ be the metric $d(x, y) := |x - y|$ mentioned earlier. Then $(\mathbf{R}, d)$ is a metric space (Exercise 1.1.2). We refer to d as the *standard metric* on $\mathbf{R}$, and if we refer to $\mathbf{R}$ as a metric space, we assume that the metric is given by the standard metric d unless otherwise specified.

Example 1.1.5 (Induced metric spaces) Let (X, d) be any metric space, and let Y be a subset of X. Then we can restrict the metric function $d : X \times X \to [0, +\infty)$ to the subset $Y \times Y$ of $X \times X$ to create a restricted metric function $d|_{Y \times Y} : Y \times Y \to [0, +\infty)$ of Y; this is known as the metric on Y *induced* by the metric d on X. The pair $(Y, d|_{Y \times Y})$ is a metric space (Exercise 1.1.4) and is known the *subspace* of (X, d) induced by Y. Thus for instance the metric on the real line in the previous example induces a metric space structure on any subset of the reals, such as the integers $\mathbf{Z}$, or an interval $[a, b]$.

Example 1.1.6 (Euclidean spaces) Let $n \geq 1$ be a natural number, and let $\mathbf{R}^n$ be the space of n-tuples of real numbers:

$$\mathbf{R}^n = \{(x_1, x_2, \ldots, x_n) : x_1, \ldots, x_n \in \mathbf{R}\}.$$

We define the *Euclidean metric* (also called the l^2 *metric*) $d_{l^2} : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ by

$$\begin{aligned} d_{l^2}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &:= \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \\ &= \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}. \end{aligned}$$

Thus for instance, if $n = 2$, then $d_{l^2}((1, 6), (4, 2)) = \sqrt{3^2 + 4^2} = 5$. This metric corresponds to the geometric distance between the two points $(x_1, x_2, \ldots, x_n)$, $(y_1, y_2, \ldots, y_n)$ as given by Pythagoras' theorem. (We remark however that while geometry does give some very important examples of metric spaces, it is possible to have metric spaces which have no obvious geometry whatsoever. Some examples are given below.) The verification that $(\mathbf{R}^n, d)$ is indeed a metric space can be seen geometrically (for instance, the triangle inequality now asserts that the length of one side of a triangle is always less than or equal to the sum of the lengths of the other two sides), but can also be proven algebraically (see Exercise 1.1.6). We refer to $(\mathbf{R}^n, d_{l^2})$ as the *Euclidean space* of *dimension* n. Extending the convention from Example 1.1.4, if we refer to $\mathbf{R}^n$ as a metric space, we assume that the metric is given by the Euclidean metric unless otherwise specified.

Example 1.1.7 (Taxicab metric) Again let $n \geq 1$, and let $\mathbf{R}^n$ be as before. But now we use a different metric d_{l^1}, the so-called *taxicab metric* (or l^1 *metric*), defined by

$$\begin{aligned} d_{l^1}((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)) &:= |x_1 - y_1| + \cdots + |x_n - y_n| \\ &= \sum_{i=1}^{n} |x_i - y_i|. \end{aligned}$$

Thus for instance, if $n = 2$, then $d_{l^1}((1, 6), (4, 2)) = 3 + 4 = 7$. This metric is called the taxicab metric, because it models the distance a taxicab would have to traverse to get from one point to another if the cab was only allowed to move in cardinal directions (north, south, east, west) and not diagonally. As such it is always at least as large as the Euclidean metric, which measures distance "as the crow flies", as it were. We claim that the space $(\mathbf{R}^n, d_{l^1})$ is also a metric space (Exercise 1.1.7). The metrics are not quite the same, but we do have the inequalities

$$d_{l^2}(x, y) \leq d_{l^1}(x, y) \leq \sqrt{n} d_{l^2}(x, y) \tag{1.1}$$

for all x, y (see Exercise 1.1.8).

Remark 1.1.8 The taxicab metric is useful in several places, for instance in the theory of error correcting codes. A string of n binary digits can be thought of as an element of $\mathbf{R}^n$, for instance the binary string 10010 can be thought of as the point $(1, 0, 0, 1, 0)$ in $\mathbf{R}^5$. The taxicab distance between two binary strings is then the number of bits in the two strings which do not match, for instance $d_{l^1}(10010, 10101) = 3$. The goal of error-correcting codes is to encode each piece of information (e.g., a letter of the alphabet) as a binary string in such a way that all the binary strings are as far away in the taxicab metric from each other as possible; this minimizes the chance that any distortion of the bits due to random noise can accidentally change one of the coded binary strings to another and also maximizes the chance that any such distortion can be detected and correctly repaired.

Example 1.1.9 (Sup norm metric) Again let $n \geq 1$, and let $\mathbf{R}^n$ be as before. But now we use a different metric d_{l^∞}, the so-called *sup norm metric* (or l^∞ *metric*), defined by

$$d_{l^\infty}((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)) := \sup\{|x_i - y_i| : 1 \leq i \leq n\}.$$

Thus for instance, if $n = 2$, then $d_{l^\infty}((1, 6), (4, 2)) = \sup(3, 4) = 4$. The space $(\mathbf{R}^n, d_{l^\infty})$ is also a metric space (Exercise 1.1.9) and is related to the l^2 metric by the inequalities

$$\frac{1}{\sqrt{n}} d_{l^2}(x, y) \leq d_{l^\infty}(x, y) \leq d_{l^2}(x, y) \tag{1.2}$$

for all x, y (see Exercise 1.1.10).

Remark 1.1.10 The l^1, l^2, and l^∞ metrics are special cases of the more general l^p *metrics*, where $p \in [1, +\infty]$, but we will not discuss these more general metrics in this text.

Example 1.1.11 (Discrete metric) Let X be an arbitrary set (finite or infinite), and define the *discrete metric* d_{disc} by setting $d_{\text{disc}}(x, y) := 0$ when $x = y$, and $d_{\text{disc}}(x, y) := 1$ when $x \neq y$. Thus, in this metric, all points are equally far apart. The space (X, d_{disc}) is a metric space (Exercise 1.1.11). Thus every set X has at least one metric on it.

Example 1.1.12 (Geodesics) (Informal) Let X be the sphere $\{(x, y, z) \in \mathbf{R}^3 : x^2 + y^2 + z^2 = 1\}$, and let $d((x, y, z), (x', y', z'))$ be the length of the shortest curve in X which starts at (x, y, z) and ends at (x', y', z'). (This curve turns out to be an arc of a great circle; we will not prove this here, as it requires *calculus of variations*, which is beyond the scope of this text.) This makes X into a metric space; the reader should be able to verify (without using any geometry of the sphere) that the triangle inequality is more or less automatic from the definition.

Example 1.1.13 (Shortest paths) (Informal) Examples of metric spaces occur all the time in real life. For instance, X could be all the computers currently connected to the internet, and $d(x, y)$ is the shortest number of connections it would take for a packet to travel from computer x to computer y; for instance, if x and y are not directly connected, but are both connected to z, then $d(x, y) = 2$. Assuming that all computers in the internet can ultimately be connected to all other computers (so that $d(x, y)$ is always finite), then (X, d) is a metric space (why?). Games such as "six degrees of separation" are also taking place in a similar metric space (what is the space, and what is the metric, in this case?). Or, X could be a major city, and $d(x, y)$ could be the shortest time it takes to drive from x to y (although this space might not satisfy axiom (c) in real life!).

Now that we have metric spaces, we can define convergence in these spaces.

Definition 1.1.14 (*Convergence of sequences in metric spaces*) Let m be an integer, (X, d) be a metric space, and let $(x^{(n)})_{n=m}^{\infty}$ be a sequence of points in X (i.e., for every natural number $n \geq m$, we assume that $x^{(n)}$ is an element of X). Let x be a point in X. We say that $(x^{(n)})_{n=m}^{\infty}$ *converges to x with respect to the metric d*, if and only if the limit $\lim_{n\to\infty} d(x^{(n)}, x)$ exists and is equal to 0. In other words, $(x^{(n)})_{n=m}^{\infty}$ converges to x with respect to d if and only if for every $\varepsilon > 0$, there exists an $N \geq m$ such that $d(x^{(n)}, x) \leq \varepsilon$ for all $n \geq N$. (Why are these two definitions equivalent?)

Remark 1.1.15 In view of Lemma 1.1.1 we see that this definition generalizes our existing notion of convergence of sequences of real numbers. In many cases, it is obvious what the metric d is, and so we shall often just say "$(x^{(n)})_{n=m}^{\infty}$ converges to x" instead of "$(x^{(n)})_{n=m}^{\infty}$ converges to x with respect to the metric d" when there is no chance of confusion. We also sometimes write "$x^{(n)} \to x$ as $n \to \infty$" instead.

Remark 1.1.16 There is nothing special about the superscript n in the above definition; it is a dummy variable. Saying that $(x^{(n)})_{n=m}^{\infty}$ converges to x is exactly the same statement as saying that $(x^{(k)})_{k=m}^{\infty}$ converges to x, for example; and sometimes it is convenient to change superscripts, for instance if the variable n is already being used for some other purpose. Similarly, it is not necessary for the sequence $x^{(n)}$ to be denoted using the superscript (n); the above definition is also valid for sequences x_n, or functions $f(n)$, or indeed of any expression which depends on n and takes values in X. Finally, from Exercises 6.1.3 and 6.1.4 we see that the starting point m of the sequence is unimportant for the purposes of taking limits; if $(x^{(n)})_{n=m}^{\infty}$ converges to x, then $(x^{(n)})_{n=m'}^{\infty}$ also converges to x for any $m' \geq m$.

Example 1.1.17 We work in the Euclidean space $\mathbf{R}^2$ with the standard Euclidean metric d_{l^2}. Let $(x^{(n)})_{n=1}^{\infty}$ denote the sequence $x^{(n)} := (1/n, 1/n)$ in $\mathbf{R}^2$, i.e., we are considering the sequence $(1, 1), (1/2, 1/2), (1/3, 1/3), \ldots$. Then this sequence converges to $(0, 0)$ with respect to the Euclidean metric d_{l^2}, since

$$\lim_{n\to\infty} d_{l^2}(x^{(n)}, (0,0)) = \lim_{n\to\infty} \sqrt{\frac{1}{n^2} + \frac{1}{n^2}} = \lim_{n\to\infty} \frac{\sqrt{2}}{n} = 0.$$

The sequence $(x^{(n)})_{n=1}^{\infty}$ also converges to $(0, 0)$ with respect to the taxicab metric d_{l^1}, since

$$\lim_{n\to\infty} d_{l^1}(x^{(n)}, (0,0)) = \lim_{n\to\infty} \frac{1}{n} + \frac{1}{n} = \lim_{n\to\infty} \frac{2}{n} = 0.$$

Similarly the sequence converges to $(0, 0)$ in the sup norm metric d_{l^∞} (why?). However, the sequence $(x^{(n)})_{n=1}^{\infty}$ does *not* converge to $(0, 0)$ in the discrete metric d_{disc}, since

$$\lim_{n\to\infty} d_{\text{disc}}(x^{(n)}, (0,0)) = \lim_{n\to\infty} 1 = 1 \neq 0.$$

Thus the convergence of a sequence can depend on what metric one uses.[1]

[1] For a somewhat whimsical real-life example, one can give a city an "automobile metric", with $d(x, y)$ defined as the time it takes for a car to drive from x to y, or a "pedestrian metric", where $d(x, y)$ is the time it takes to walk on foot from x to y. (Let us assume for sake of argument that

In the case of the above four metrics—Euclidean, taxicab, sup norm, and discrete—it is in fact rather easy to test for convergence.

Proposition 1.1.18 (Equivalence of l^1, l^2, l^∞) *Let $\mathbf{R}^n$ be a Euclidean space, and let $(x^{(k)})_{k=m}^\infty$ be a sequence of points in $\mathbf{R}^n$. We write $x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$, i.e., for $j = 1, 2, \ldots, n$, $x_j^{(k)} \in \mathbf{R}$ is the jth co-ordinate of $x^{(k)} \in \mathbf{R}^n$. Let $x = (x_1, \ldots, x_n)$ be a point in $\mathbf{R}^n$. Then the following four statements are equivalent:*

(a) $(x^{(k)})_{k=m}^\infty$ converges to x with respect to the Euclidean metric d_{l^2}.

(b) $(x^{(k)})_{k=m}^\infty$ converges to x with respect to the taxicab metric d_{l^1}.

(c) $(x^{(k)})_{k=m}^\infty$ converges to x with respect to the sup norm metric d_{l^∞}.

(d) For every $1 \leq j \leq n$, the sequence $(x_j^{(k)})_{k=m}^\infty$ converges to x_j. (Notice that this is a sequence of real numbers, not of points in $\mathbf{R}^n$.)

Proof See Exercise 1.1.12. □

In other words, a sequence converges in the Euclidean, taxicab, or sup norm metric if and only if each of its components converges individually. Because of the equivalence of (a), (b), and (c), we say that the Euclidean, taxicab, and sup norm metrics on $\mathbf{R}^n$ are *equivalent*. (There are infinite-dimensional analogues of the Euclidean, taxicab, and sup norm metrics which are *not* equivalent, see for instance Exercise 1.1.15.)

For the discrete metric, convergence is much rarer: the sequence must be eventually constant in order to converge.

Proposition 1.1.19 (Convergence in the discrete metric) *Let X be any set, and let d_{disc} be the discrete metric on X. Let $(x^{(n)})_{n=m}^\infty$ be a sequence of points in X, and let x be a point in X. Then $(x^{(n)})_{n=m}^\infty$ converges to x with respect to the discrete metric d_{disc} if and only if there exists an $N \geq m$ such that $x^{(n)} = x$ for all $n \geq N$.*

Proof See Exercise 1.1.13. □

We now prove a basic fact about converging sequences; they can only converge to at most one point at a time.

Proposition 1.1.20 (Uniqueness of limits) *Let (X, d) be a metric space, and let $(x^{(n)})_{n=m}^\infty$ be a sequence in X. Suppose that there are two points $x, x' \in X$ such that $(x^{(n)})_{n=m}^\infty$ converges to x with respect to d, and $(x^{(n)})_{n=m}^\infty$ also converges to x' with respect to d. Then we have $x = x'$.*

Proof See Exercise 1.1.14. □

Because of the above proposition, it is safe to introduce the following notation: if $(x^{(n)})_{n=m}^\infty$ converges to x in the metric d, then we write $d - \lim_{n\to\infty} x^{(n)} = x$, or simply $\lim_{n\to\infty} x^{(n)} = x$ when there is no confusion as to what d is. For instance, in the example of $(\frac{1}{n}, \frac{1}{n})$, we have

these metrics are symmetric, though this is not always the case in real life.) One can easily imagine examples where two points are close in one metric but not another.

$$d_{l^2} - \lim_{n\to\infty} \left(\frac{1}{n}, \frac{1}{n}\right) = d_{l^1} - \lim_{n\to\infty} \left(\frac{1}{n}, \frac{1}{n}\right) = (0, 0),$$

but $d_{\text{disc}} - \lim_{n\to\infty}(\frac{1}{n}, \frac{1}{n})$ is undefined. Thus the meaning of $d - \lim_{n\to\infty} x^{(n)}$ can depend on what d is; however Proposition 1.1.20 assures us that once d is fixed, there can be at most one value of $d - \lim_{n\to\infty} x^{(n)}$. (Of course, it is still possible that this limit does not exist; some sequences are not convergent.) Note that by Lemma 1.1.1, this definition of limit generalizes the notion of limit in Definition 6.1.8.

Remark 1.1.21 It is possible for a sequence to converge to one point using one metric, and another point using a different metric, although such examples are usually quite artificial. For instance, let $X := [0, 1]$, the closed interval from 0 to 1. Using the usual metric d, we have $d - \lim_{n\to\infty} \frac{1}{n} = 0$. But now suppose we "swap" the points 0 and 1 in the following manner. Let $f\colon [0, 1] \to [0, 1]$ be the function defined by $f(0) := 1$, $f(1) := 0$, and $f(x) := x$ for all $x \in (0, 1)$, and then define $d'(x, y) := d(f(x), f(y))$. Then (X, d') is still a metric space (why?), but now $d' - \lim_{n\to\infty} \frac{1}{n} = 1$. Thus changing the metric on a space can greatly affect the nature of convergence (also called the *topology*) on that space; see Sect. 2.5 for a further discussion of topology.

— Exercises —

Exercise 1.1.1 Prove Lemma 1.1.1.

Exercise 1.1.2 Show that the real line with the metric $d(x, y) := |x - y|$ is indeed a metric space. (*Hint:* you may wish to review your proof of Proposition 4.3.3.)

Exercise 1.1.3 Let X be a set, and let $d : X \times X \to [0, \infty)$ be a function.

(a) Give an example of a pair (X, d) which obeys axioms (bcd) of Definition 1.1.2, but not (a). (*Hint:* modify the discrete metric.)
(b) Give an example of a pair (X, d) which obeys axioms (acd) of Definition 1.1.2, but not (b).
(c) Give an example of a pair (X, d) which obeys axioms (abd) of Definition 1.1.2, but not (c).
(d) Give an example of a pair (X, d) which obeys axioms (abc) of Definition 1.1.2, but not (d). (*Hint:* try examples where X is a finite set.)

Exercise 1.1.4 Show that the pair $(Y, d|_{Y\times Y})$ defined in Example 1.1.5 is indeed a metric space.

Exercise 1.1.5 Let $n \geq 1$, and let $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be real numbers. Verify the identity

$$\left(\sum_{i=1}^n a_i b_i\right)^2 + \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (a_i b_j - a_j b_i)^2 = \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{j=1}^n b_j^2\right),$$

and conclude the *Cauchy–Schwarz inequality*

$$\left|\sum_{i=1}^{n} a_i b_i\right| \leq \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \left(\sum_{j=1}^{n} b_j^2\right)^{1/2}. \tag{1.3}$$

Then use the Cauchy–Schwarz inequality to prove the *triangle inequality*

$$\left(\sum_{i=1}^{n} (a_i + b_i)^2\right)^{1/2} \leq \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} + \left(\sum_{j=1}^{n} b_j^2\right)^{1/2}.$$

Exercise 1.1.6 Show that $(\mathbf{R}^n, d_{l^2})$ in Example 1.1.6 is indeed a metric space. (*Hint:* use Exercise 1.1.5.)

Exercise 1.1.7 Show that the pair $(\mathbf{R}^n, d_{l^1})$ in Example 1.1.7 is indeed a metric space.

Exercise 1.1.8 Prove the two inequalities in (1.1). (*Hint:* For the first inequality, square both sides. For the second inequality, use Exercise (1.1.5).)

Exercise 1.1.9 Show that the pair $(\mathbf{R}^n, d_{l^\infty})$ in Example 1.1.9 is indeed a metric space.

Exercise 1.1.10 Prove the two inequalities in (1.2).

Exercise 1.1.11 Show that the discrete metric (X, d_{disc}) in Example 1.1.11 is indeed a metric space.

Exercise 1.1.12 Prove Proposition 1.1.18.

Exercise 1.1.13 Prove Proposition 1.1.19.

Exercise 1.1.14 Prove Proposition 1.1.20. (*Hint:* modify the proof of Proposition 6.1.7.)

Exercise 1.1.15 Let

$$X := \left\{(a_n)_{n=0}^{\infty} : \sum_{n=0}^{\infty} |a_n| < \infty\right\}$$

be the space of absolutely convergent sequences. Define the l^1 and l^∞ metrics on this space by

$$d_{l^1}((a_n)_{n=0}^{\infty}, (b_n)_{n=0}^{\infty}) := \sum_{n=0}^{\infty} |a_n - b_n|;$$
$$d_{l^\infty}((a_n)_{n=0}^{\infty}, (b_n)_{n=0}^{\infty}) := \sup_{n\in\mathbf{N}} |a_n - b_n|.$$

Show that these are both metrics on X, but show that there exist sequences $x^{(1)}, x^{(2)}, \ldots$ of elements of X (i.e., sequences of sequences) which are convergent with respect to the d_{l^∞} metric but not with respect to the d_{l^1} metric. Conversely, show that any sequence which converges in the d_{l^1} metric automatically converges in the d_{l^∞} metric.

Exercise 1.1.16 Let $(x_n)_{n=1}^\infty$ and $(y_n)_{n=1}^\infty$ be two sequences in a metric space (X, d). Suppose that $(x_n)_{n=1}^\infty$ converges to a point $x \in X$, and $(y_n)_{n=1}^\infty$ converges to a point $y \in X$. Show that $\lim_{n\to\infty} d(x_n, y_n) = d(x, y)$. (*Hint:* use the triangle inequality several times.)

1.2 Some Point-Set Topology of Metric Spaces

Having defined the operation of convergence on metric spaces, we now define a couple other related notions, including that of open set, closed set, interior, exterior, boundary, and adherent point. The study of such notions is known as *point-set topology*, which we shall return to in Sect. 2.5.

We first need the notion of a *metric ball*, or more simply a *ball*.

Definition 1.2.1 (*Balls*) Let (X, d) be a metric space, let x_0 be a point in X, and let $r > 0$. We define the *ball* $B_{(X,d)}(x_0, r)$ in X, centered at x_0, and with radius r, in the metric d, to be the set

$$B_{(X,d)}(x_0, r) := \{x \in X : d(x, x_0) < r\}.$$

When it is clear what the metric space (X, d) is, we shall abbreviate $B_{(X,d)}(x_0, r)$ as just $B(x_0, r)$.

Example 1.2.2 In $\mathbf{R}^2$ with the Euclidean metric d_{l^2}, the ball $B_{(\mathbf{R}^2,d_{l^2})}((0, 0), 1)$ is the open disc

$$B_{(\mathbf{R}^2,d_{l^2})}((0, 0), 1) = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 < 1\}.$$

However, if one uses the taxicab metric d_{l^1} instead, then we obtain a diamond:

$$B_{(\mathbf{R}^2,d_{l^1})}((0, 0), 1) = \{(x, y) \in \mathbf{R}^2 : |x| + |y| < 1\}.$$

If we use the discrete metric, the ball is now reduced to a single point:

$$B_{(\mathbf{R}^2,d_{\text{disc}})}((0, 0), 1) = \{(0, 0)\},$$

although if one increases the radius to be larger than 1, then the ball now encompasses all of $\mathbf{R}^2$. (Why?)

Example 1.2.3 In $\mathbf{R}$ with the usual metric d, the open interval $(3, 7)$ is also the metric ball $B_{(\mathbf{R},d)}(5, 2)$.

Remark 1.2.4 Note that the smaller the radius r, the smaller the ball $B(x_0, r)$. However, $B(x_0, r)$ always contains at least one point, namely the center x_0, as long as r stays positive, thanks to Definition 1.1.2(a). (We don't consider balls of zero radius or negative radius since they are rather boring, being just the empty set.)

Using metric balls, one can now take a set E in a metric space X and classify three types of points in X: interior, exterior, and boundary points of E.

Definition 1.2.5 (*Interior, exterior, boundary*) Let (X, d) be a metric space, let E be a subset of X, and let x_0 be a point in X. We say that x_0 is an *interior point of* E if there exists a radius $r > 0$ such that $B(x_0, r) \subseteq E$. We say that x_0 is an *exterior point of* E if there exists a radius $r > 0$ such that $B(x_0, r) \cap E = \emptyset$. We say that x_0 is a *boundary point of* E if it is neither an interior point nor an exterior point of E.

The set of all interior points of E is called the *interior* of E and is sometimes denoted $\mathrm{int}(E)$. The set of exterior points of E is called the *exterior* of E and is sometimes denoted $\mathrm{ext}(E)$. The set of boundary points of E is called the *boundary* of E and is sometimes denoted ∂E.

Remark 1.2.6 If x_0 is an interior point of E, then x_0 must actually be an element of E, since balls $B(x_0, r)$ always contain their center x_0. Conversely, if x_0 is an exterior point of E, then x_0 cannot be an element of E. In particular it is not possible for x_0 to simultaneously be an interior and an exterior point of E. If x_0 is a boundary point of E, then it could be an element of E, but it could also not lie in E; we give some examples below.

Example 1.2.7 We work on the real line $\mathbf{R}$ with the standard metric d. Let E be the half-open interval $E = [1, 2)$. The point 1.5 is an interior point of E, since one can find a ball (for instance $B(1.5, 0.1)$) centered at 1.5 which lies in E. The point 3 is an exterior point of E, since one can find a ball (for instance $B(3, 0.1)$) centered at 3 which is disjoint from E. The points 1 and 2, however, are neither interior points nor exterior points of E and are thus boundary points of E. Thus in this case $\mathrm{int}(E) = (1, 2)$, $\mathrm{ext}(E) = (-\infty, 1) \cup (2, \infty)$, and $\partial E = \{1, 2\}$. Note that in this case one of the boundary points is an element of E, while the other is not.

Example 1.2.8 When we give a set X the discrete metric d_{disc}, and E is any subset of X, then every element of E is an interior point of E, every point not contained in E is an exterior point of E, and there are no boundary points; see Exercise 1.2.1.

Definition 1.2.9 (*Closure*) Let (X, d) be a metric space, let E be a subset of X, and let x_0 be a point in X. We say that x_0 is an *adherent point* of E if for every radius $r > 0$, the ball $B(x_0, r)$ has a non-empty intersection with E. The set of all adherent points of E is called the *closure* of E and is denoted $\overline{E}$.

Note that these notions are consistent with the corresponding notions on the real line defined in Definitions 9.1.8 and 9.1.10 (why?).

The following proposition links the notions of adherent point with interior and boundary point and also to that of convergence.

Proposition 1.2.10 *Let (X, d) be a metric space, let E be a subset of X, and let x_0 be a point in X. Then the following statements are logically equivalent.*

(a) *x_0 is an adherent point of E.*
(b) *x_0 is either an interior point or a boundary point of E.*
(c) *There exists a sequence $(x_n)_{n=1}^{\infty}$ in E which converges to x_0 with respect to the metric d.*

Proof See Exercise 1.2.2. □

From the equivalence of Proposition 1.2.10(a) and (b) we obtain an immediate corollary:

Corollary 1.2.11 *Let (X, d) be a metric space, and let E be a subset of X. Then* $\overline{E} = \text{int}(E) \cup \partial E = X \backslash \text{ext}(E)$.

As remarked earlier, the boundary of a set E may or may not lie in E. Depending on how the boundary is situated, we may call a set open, closed, or neither:

Definition 1.2.12 (*Open and closed sets*) Let (X, d) be a metric space, and let E be a subset of X. We say that E is *closed* if it contains all of its boundary points, i.e., $\partial E \subseteq E$. We say that E is *open* if it contains none of its boundary points, i.e., $\partial E \cap E = \emptyset$. If E contains some of its boundary points but not others, then it is neither open nor closed.

Example 1.2.13 We work in the real line $\mathbf{R}$ with the standard metric d. The set $(1, 2)$ does not contain either of its boundary points 1, 2 and is hence open. The set $[1, 2]$ contains both of its boundary points 1, 2 and is hence closed. The set $[1, 2)$ contains one of its boundary points 1, but does not contain the other boundary point 2, so is neither open nor closed.

Remark 1.2.14 It is possible for a set to be simultaneously open and closed, if it has no boundary. For instance, in a metric space (X, d), the whole space X has no boundary (every point in X is an interior point—why?), and so X is both open and closed. The empty set $\emptyset$ also has no boundary (every point in X is an exterior point—why?), and so $\emptyset$ is both open and closed. In many cases these are the only sets that are simultaneously open and closed, but there are exceptions. For instance, using the discrete metric d_{disc}, *every* set is both open and closed! (why?)

From the above two remarks we see that the notions of being open and being closed are *not* negations of each other; there are sets that are both open and closed, and there are sets which are neither open nor closed. Thus, if one knew for instance that E was not an open set, it would be erroneous to conclude from this that E was a closed set, and similarly with the rôles of open and closed reversed. The correct relationship between open and closed sets is given by Proposition 1.2.15(e) below.

Now we list some more properties of open and closed sets.

Proposition 1.2.15 (Basic properties of open and closed sets) *Let (X, d) be a metric space.*

(a) *Let E be a subset of X. Then E is open if and only if $E = \mathrm{int}(E)$. In other words, E is open if and only if for every $x \in E$, there exists an $r > 0$ such that $B(x, r) \subseteq E$.*

(b) *Let E be a subset of X. Then E is closed if and only if E contains all its adherent points. In other words, E is closed if and only if for every convergent sequence $(x_n)_{n=m}^{\infty}$ in E, the limit $\lim_{n\to\infty} x_n$ of that sequence also lies in E.*

(c) *For any $x_0 \in X$ and $r > 0$, then the ball $B(x_0, r)$ is an open set. The set $\{x \in X : d(x, x_0) \leq r\}$ is a closed set. (This set is sometimes called the* closed ball *of radius r centered at x_0.)*

(d) *Any singleton set $\{x_0\}$, where $x_0 \in X$, is automatically closed.*

(e) *If E is a subset of X, then E is open if and only if the complement $X\backslash E := \{x \in X : x \notin E\}$ is closed.*

(f) *If $E_1, \ldots, E_n$ is a finite collection of open sets in X, then $E_1 \cap E_2 \cap \cdots \cap E_n$ is also open. If $F_1, \ldots, F_n$ is a finite collection of closed sets in X, then $F_1 \cup F_2 \cup \cdots \cup F_n$ is also closed.*

(g) *If $\{E_\alpha\}_{\alpha\in I}$ is a collection of open sets in X (where the index set I could be finite, countable, or uncountable), then the union $\bigcup_{\alpha\in I} E_\alpha := \{x \in X : x \in E_\alpha$ for some $\alpha \in I\}$ is also open. If $\{F_\alpha\}_{\alpha\in I}$ is a collection of closed sets in X, then the intersection $\bigcap_{\alpha\in I} F_\alpha := \{x \in X : x \in F_\alpha$ for all $\alpha \in I\}$ is also closed.*

(h) *If E is any subset of X, then $\mathrm{int}(E)$ is the largest open set which is contained in E; in other words, $\mathrm{int}(E)$ is open, and given any other open set $V \subseteq E$, we have $V \subseteq \mathrm{int}(E)$. Similarly $\overline{E}$ is the smallest closed set which contains E; in other words, $\overline{E}$ is closed, and given any other closed set $K \supset E$, $K \supseteq \overline{E}$.*

Proof See Exercise 1.2.3. □

— Exercises —

Exercise 1.2.1 Verify the claims in Example 1.2.8.

Exercise 1.2.2 Prove Proposition 1.2.10. (*Hint:* for some of the implications one will need the axiom of choice, as in Lemma 8.4.5.)

Exercise 1.2.3 Prove Proposition 1.2.15. (*Hint:* you can use earlier parts of the proposition to prove later ones.)

Exercise 1.2.4 Let (X, d) be a metric space, x_0 be a point in X, and $r > 0$. Let B be the open ball $B := B(x_0, r) = \{x \in X : d(x, x_0) < r\}$, and let C be the closed ball $C := \{x \in X : d(x, x_0) \leq r\}$.

(a) Show that $\overline{B} \subseteq C$.

(b) Give an example of a metric space (X, d), a point x_0, and a radius $r > 0$ such that $\overline{B}$ is *not* equal to C.

1.3 Relative Topology

When we defined notions such as open and closed sets, we mentioned that such concepts depended on the choice of metric one uses. For instance, on the real line $\mathbf{R}$, if one uses the usual metric $d(x, y) = |x - y|$, then the set $\{1\}$ is not open, however if instead one uses the discrete metric d_{disc}, then $\{1\}$ is now an open set (why?).

However, it is not just the choice of metric which determines what is open and what is not—it is also the choice of *ambient space* X. Here are some examples.

Example 1.3.1 Consider the plane $\mathbf{R}^2$ with the Euclidean metric d_{l^2}. Inside the plane, we can find the x-axis $X := \{(x, 0) : x \in \mathbf{R}\}$. The metric d_{l^2} can be restricted to X, creating a subspace $(X, d_{l^2}|_{X\times X})$ of $(\mathbf{R}^2, d_{l^2})$. (This subspace is essentially the same as the real line $(\mathbf{R}, d)$ with the usual metric; the precise way of stating this is that $(X, d_{l^2}|_{X\times X})$ is *isometric* to $(\mathbf{R}, d)$. We will not pursue this concept further in this text, however.) Now consider the set

$$E := \{(x, 0) : -1 < x < 1\}$$

which is both a subset of X and of $\mathbf{R}^2$. Viewed as a subset of $\mathbf{R}^2$, it is not open, because the point $(0, 0)$, for instance, lies in E but is not an interior point of E. (Any ball $B_{\mathbf{R}^2,d_{l^2}}(0, r)$ will contain at least one point that lies outside of the x-axis, and hence outside of E.) On the other hand, if viewed as a subset of X, it is open; every point of E is an interior point of E *with respect to the metric space* $(X, d_{l^2}|_{X\times X})$. For instance, the point $(0, 0)$ is now an interior point of E, because the ball $B_{X,d_{l^2}|_{X\times X}}(0, 1)$ is contained in E (in fact, in this case it *is* E).

Example 1.3.2 Consider the real line $\mathbf{R}$ with the standard metric d, and let X be the interval $X := (-1, 1)$ contained inside $\mathbf{R}$; we can then restrict the metric d to X, creating a subspace $(X, d|_{X\times X})$ of $(\mathbf{R}, d)$. Now consider the set $[0, 1)$. This set is not closed in $\mathbf{R}$, because the point 1 is adherent to $[0, 1)$ but is not contained in $[0, 1)$. However, when considered as a subset of X, the set $[0, 1)$ now becomes closed; the point 1 is not an element of X and so is no longer considered an adherent point of $[0, 1)$, and so now $[0, 1)$ contains all of its adherent points.

To clarify this distinction, we make a definition.

Definition 1.3.3 (*Relative topology*) Let (X, d) be a metric space, let Y be a subset of X, and let E be a subset of Y. We say that E is *relatively open with respect to* Y if it is open in the metric subspace $(Y, d|_{Y\times Y})$. Similarly, we say that E is *relatively closed with respect to* Y if it is closed in the metric space $(Y, d|_{Y\times Y})$.

The relationship between open (or closed) sets in X, and relatively open (or relatively closed) sets in Y, is the following.

Proposition 1.3.4 *Let* (X, d) *be a metric space, let* Y *be a subset of* X, *and let* E *be a subset of* Y.

(a) E is relatively open with respect to Y if and only if $E = V \cap Y$ for some set $V \subseteq X$ which is open in X.

(b) E is relatively closed with respect to Y if and only if $E = K \cap Y$ for some set $K \subseteq X$ which is closed in X.

Proof We just prove (a) and leave (b) to Exercise 1.3.1. First suppose that E is relatively open with respect to Y. Then, E is open in the metric space $(Y, d|_{Y \times Y})$. Thus, for every $x \in E$, there exists a radius $r > 0$ such that the ball $B_{(Y,d|_{Y\times Y})}(x, r)$ is contained in E. This radius r depends on x; to emphasize this we write r_x instead of r, thus for every $x \in E$ the ball $B_{(Y,d|_{Y\times Y})}(x, r_x)$ is contained in E. (Note that we have used the axiom of choice, Proposition 8.4.7, to do this.)

Now consider the set

$$V := \bigcup_{x \in E} B_{(X,d)}(x, r_x).$$

This is a subset of X. By Proposition 1.2.15(c) and (g), V is open. Now we prove that $E = V \cap Y$. Certainly any point x in E lies in $V \cap Y$, since it lies in Y and it also lies in $B_{(X,d)}(x, r_x)$, and hence in V. Now suppose that y is a point in $V \cap Y$. Then $y \in V$, which implies that there exists an $x \in E$ such that $y \in B_{(X,d)}(x, r_x)$. But since y is also in Y, this implies that $y \in B_{(Y,d|_{Y\times Y})}(x, r_x)$. But by definition of r_x, this means that $y \in E$, as desired. Thus we have found an open set V for which $E = V \cap Y$ as desired.

Now we do the converse. Suppose that $E = V \cap Y$ for some open set V; we have to show that E is relatively open with respect to Y. Let x be any point in E; we have to show that x is an interior point of E in the metric space $(Y, d|_{Y \times Y})$. Since $x \in E$, we know $x \in V$. Since V is open in X, we know that there is a radius $r > 0$ such that $B_{(X,d)}(x, r)$ is contained in V. Strictly speaking, r depends on x, and so we could write r_x instead of r, but for this argument we will only use a single choice of x (as opposed to the argument in the previous paragraph) and so we will not bother to subscript r here. Since $E = V \cap Y$, this means that $B_{(X,d)}(x, r) \cap Y$ is contained in E. But $B_{(X,d)}(x, r) \cap Y$ is exactly the same as $B_{(Y,d|_{Y\times Y})}(x, r)$ (why?), and so $B_{(Y,d|_{Y\times Y})}(x, r)$ is contained in E. Thus x is an interior point of E in the metric space $(Y, d|_{Y \times Y})$, as desired. □

— Exercises —

Exercise 1.3.1 Prove Proposition 1.3.4(b).

1.4 Cauchy Sequences and Complete Metric Spaces

We now generalize much of the theory of limits of sequences from Chap. 6 to the setting of general metric spaces. We begin by generalizing the notion of a *subsequence* from Definition 6.6.1:

Definition 1.4.1 (*Subsequences*) Suppose that $(x^{(n)})_{n=m}^{\infty}$ is a sequence of points in a metric space (X, d). Suppose that $n_1, n_2, n_3, \ldots$ is an increasing sequence of integers which are at least as large as m, thus

$$m \leq n_1 < n_2 < n_3 < \cdots .$$

Then we call the sequence $(x^{(n_j)})_{j=1}^{\infty}$ a *subsequence* of the original sequence $(x^{(n)})_{n=m}^{\infty}$.

Example 1.4.2 The sequence $\left(\left(\frac{1}{j^2}, \frac{1}{j^2}\right)\right)_{j=1}^{\infty}$ in $\mathbf{R}^2$ is a subsequence of the sequence $\left(\left(\frac{1}{n}, \frac{1}{n}\right)\right)_{n=1}^{\infty}$ (in this case, $n_j := j^2$). The sequence 1, 1, 1, 1, ... is a subsequence of 1, 0, 1, 0, 1, ...

If a sequence converges, then so do all of its subsequences:

Lemma 1.4.3 *Let $(x^{(n)})_{n=m}^{\infty}$ be a sequence in (X, d) which converges to some limit x_0. Then every subsequence $(x^{(n_j)})_{j=1}^{\infty}$ of that sequence also converges to x_0.*

Proof See Exercise 1.4.1. □

On the other hand, it is possible for a subsequence to be convergent without the sequence as a whole being convergent. For example, the sequence 1, 0, 1, 0, 1, ... is not convergent, even though certain subsequences of it (such as 1, 1, 1, ...) converge. To quantify this phenomenon, we generalize Definition 6.4.1 as follows:

Definition 1.4.4 (*Limit points*) Suppose that $(x^{(n)})_{n=m}^{\infty}$ is a sequence of points in a metric space (X, d), and let $L \in X$. We say that L is a *limit point* of $(x^{(n)})_{n=m}^{\infty}$ iff for every $N \geq m$ and $\varepsilon > 0$ there exists an $n \geq N$ such that $d(x^{(n)}, L) \leq \varepsilon$.

Proposition 1.4.5 *Let $(x^{(n)})_{n=m}^{\infty}$ be a sequence of points in a metric space (X, d), and let $L \in X$. Then the following are equivalent:*

- *L is a limit point of $(x^{(n)})_{n=m}^{\infty}$.*
- *There exists a subsequence $(x^{(n_j)})_{j=1}^{\infty}$ of the original sequence $(x^{(n)})_{n=m}^{\infty}$ which converges to L.*

Proof See Exercise 1.4.2. □

Next, we review the notion of a *Cauchy sequence* from Definition 6.1.3 (see also Definition 5.1.8).

Definition 1.4.6 (*Cauchy sequences*) Let $(x^{(n)})_{n=m}^{\infty}$ be a sequence of points in a metric space (X, d). We say that this sequence is a *Cauchy sequence* iff for every $\varepsilon > 0$, there exists an $N \geq m$ such that $d(x^{(j)}, x^{(k)}) < \varepsilon$ for all $j, k \geq N$.

Lemma 1.4.7 (Convergent sequences are Cauchy sequences) *Let $(x^{(n)})_{n=m}^{\infty}$ be a sequence in (X, d) which converges to some limit x_0. Then $(x^{(n)})_{n=m}^{\infty}$ is also a Cauchy sequence.*

Proof See Exercise 1.4.3. □

It is also easy to check that subsequence of a Cauchy sequence is also a Cauchy sequence (why?). However, not every Cauchy sequence converges:

Example 1.4.8 (Informal) Consider the sequence

$$3, 3.1, 3.14, 3.141, 3.1415, \ldots$$

in the metric space $(\mathbf{Q}, d)$ (the rationals $\mathbf{Q}$ with the usual metric $d(x, y) := |x - y|$). While this sequence is convergent in $\mathbf{R}$ (it converges to π), it does not converge in $\mathbf{Q}$ (since $\pi \notin \mathbf{Q}$, and a sequence cannot converge to two different limits).

So in certain metric spaces, Cauchy sequences do not necessarily converge. However, if even part of a Cauchy sequence converges, then the entire Cauchy sequence must converge (to the same limit):

Lemma 1.4.9 *Let $(x^{(n)})_{n=m}^{\infty}$ be a Cauchy sequence in (X, d). Suppose that there is some subsequence $(x^{(n_j)})_{j=1}^{\infty}$ of this sequence which converges to a limit x_0 in X. Then the original sequence $(x^{(n)})_{n=m}^{\infty}$ also converges to x_0.*

Proof See Exercise 1.4.4. □

In Example 1.4.8 we saw an example of a metric space which contained Cauchy sequences which did not converge. However, in Theorem 6.4.18 we saw that in the metric space $(\mathbf{R}, d)$, every Cauchy sequence did have a limit. This motivates the following definition.

Definition 1.4.10 (*Complete metric spaces*) A metric space (X, d) is said to be *complete* iff every Cauchy sequence in (X, d) is in fact convergent in (X, d).

Example 1.4.11 By Theorem 6.4.18, the reals $(\mathbf{R}, d)$ are complete; by Example 1.4.8, the rationals $(\mathbf{Q}, d)$, on the other hand, are not complete.

Complete metric spaces have some nice properties. For instance, they are *intrinsically closed*: no matter what space one places them in, they are always closed sets. More precisely:

Proposition 1.4.12 *(a) Let (X, d) be a metric space, and let $(Y, d|_{Y\times Y})$ be a subspace of (X, d). If $(Y, d|_{Y\times Y})$ is complete, then Y must be closed in X.*
(b) Conversely, suppose that (X, d) is a complete metric space, and Y is a closed subset of X. Then the subspace $(Y, d|_{Y\times Y})$ is also complete.

Proof See Exercise 1.4.7. □

In contrast, an incomplete metric space such as $(\mathbf{Q}, d)$ may be considered closed in some spaces (for instance, $\mathbf{Q}$ is closed in $\mathbf{Q}$) but not in others (for instance, $\mathbf{Q}$ is not closed in $\mathbf{R}$). Indeed, it turns out that given any incomplete metric space (X, d), there exists a *completion* $(\overline{X}, \overline{d})$, which is a larger metric space containing (X, d) which is complete, and such that X is not closed in $\overline{X}$ (indeed, the closure of X in $(\overline{X}, \overline{d})$ will be all of $\overline{X}$); see Exercise 1.4.8. For instance, one possible completion of $\mathbf{Q}$ is $\mathbf{R}$.

— Exercises —

Exercise 1.4.1 Prove Lemma 1.4.3. (*Hint:* review your proof of Proposition 6.6.5.)

Exercise 1.4.2 Prove Proposition 1.4.5. (*Hint:* review your proof of Proposition 6.6.6.)

Exercise 1.4.3 Prove Lemma 1.4.7. (*Hint:* review your proof of Proposition 6.1.12.)

Exercise 1.4.4 Prove Lemma 1.4.9.

Exercise 1.4.5 Let $(x^{(n)})_{n=m}^{\infty}$ be a sequence of points in a metric space (X, d), and let $L \in X$. Show that if L is a limit point of the sequence $(x^{(n)})_{n=m}^{\infty}$, then L is an adherent point of the set $\{x^{(n)} : n \geq m\}$. Is the converse true?

Exercise 1.4.6 Show that every Cauchy sequence can have at most one limit point.

Exercise 1.4.7 Prove Proposition 1.4.12.

Exercise 1.4.8 The following construction generalizes the construction of the reals from the rationals in Chap. 5, allowing one to view any metric space as a subspace of a complete metric space. In what follows we let (X, d) be a metric space.

(a) Given any Cauchy sequence $(x_n)_{n=1}^{\infty}$ in X, we introduce the *formal limit* $\mathrm{LIM}_{n\to\infty} x_n$. We say that two formal limits $\mathrm{LIM}_{n\to\infty} x_n$ and $\mathrm{LIM}_{n\to\infty} y_n$ are equal if $\lim_{n\to\infty} d(x_n, y_n)$ is equal to zero. Show that this equality relation obeys the reflexive, symmetry, and transitive axioms.

(b) Let $\overline{X}$ be the space of all formal limits of Cauchy sequences in X, with the above equality relation. Define a metric $d_{\overline{X}} \colon \overline{X} \times \overline{X} \to [0, +\infty)$ by setting

$$d_{\overline{X}}(\mathrm{LIM}_{n\to\infty} x_n, \mathrm{LIM}_{n\to\infty} y_n) := \lim_{n\to\infty} d(x_n, y_n).$$

Show that this function is well-defined (this means not only that the limit $\lim_{n\to\infty} d(x_n, y_n)$ exists, but also that the axiom of substitution is obeyed; cf. Lemma 5.3.7) and gives $\overline{X}$ the structure of a metric space.

(c) Show that the metric space $(\overline{X}, d_{\overline{X}})$ is complete.

(d) We identify an element $x \in X$ with the corresponding formal limit $\mathrm{LIM}_{n\to\infty} x$ in $\overline{X}$; show that this is legitimate by verifying that $x = y \iff \mathrm{LIM}_{n\to\infty} x = \mathrm{LIM}_{n\to\infty} y$. With this identification, show that $d(x, y) = d_{\overline{X}}(x, y)$, and thus (X, d) can now be thought of as a subspace of $(\overline{X}, d_{\overline{X}})$.

(e) Show that the closure of X in $\overline{X}$ is $\overline{X}$ (which explains the choice of notation $\overline{X}$).

(f) Show that the formal limit agrees with the actual limit, thus if $(x_n)_{n=1}^{\infty}$ is any Cauchy sequence in X, then we have $\lim_{n\to\infty} x_n = \mathrm{LIM}_{n\to\infty} x_n$ in $\overline{X}$.

1.5 Compact Metric Spaces

We now come to one of the most useful notions in point-set topology, that of *compactness*. Recall the Heine–Borel theorem (Theorem 9.1.24), which asserted that every sequence in a closed and bounded subset X of the real line $\mathbf{R}$ had a convergent subsequence whose limit was also in X. Conversely, only the closed and bounded sets have this property. This property turns out to be so useful that we give it a name.

Definition 1.5.1 (*Compactness*) A metric space (X, d) is said to be *compact* iff every sequence in (X, d) has at least one convergent subsequence. A subset Y of a metric space X is said to be *compact* if the subspace $(Y, d|_{Y\times Y})$ is compact.

Remark 1.5.2 The notion of a set Y being compact is *intrinsic*, in the sense that it only depends on the metric function $d|_{Y\times Y}$ restricted to Y, and not on the choice of the ambient space X. The notions of completeness in Definition 1.4.10, and of boundedness below in Definition 1.5.3, are also intrinsic, but the notions of open and closed are not (see the discussion in Sect. 1.3).

Thus, Theorem 9.1.24 shows that in the real line $\mathbf{R}$ with the usual metric, every closed and bounded set is compact, and conversely every compact set is closed and bounded.

Now we investigate how the Heine–Borel extends to other metric spaces.

Definition 1.5.3 (*Bounded sets*) Let (X, d) be a metric space, and let Y be a subset of X. We say that Y is *bounded* iff for every $x \in X$ there exists a ball $B(x, r)$ in X of some finite radius r which contains Y. We call the metric space (X, d) bounded if X is bounded.

Remark 1.5.4 This definition is compatible with the definition of a bounded set in Definition 9.1.22 (Exercise 1.5.1).

Proposition 1.5.5 *Let (X, d) be a compact metric space. Then (X, d) is both complete and bounded.*

Proof See Exercise 1.5.2. □

From this proposition and Proposition 1.4.12(a) we obtain one half of the Heine–Borel theorem for general metric spaces:

Corollary 1.5.6 (Compact sets are closed and bounded) *Let (X, d) be a metric space, and let Y be a compact subset of X. Then Y is closed and bounded.*

The other half of the Heine–Borel theorem is true in Euclidean spaces:

Theorem 1.5.7 (Heine–Borel theorem) *Let $(\mathbf{R}^n, d)$ be a Euclidean space with either the Euclidean metric, the taxicab metric, or the sup norm metric. Let E be a subset of $\mathbf{R}^n$. Then E is compact if and only if it is closed and bounded.*

Proof See Exercise 1.5.3. □

However, the Heine–Borel theorem is not true for more general metrics. For instance, the integer $\mathbf{Z}$ with the discrete metric is closed (indeed, it is complete) and bounded, but not compact, since the sequence $1, 2, 3, 4, \ldots$ is in $\mathbf{Z}$ but has no convergent subsequence (why?). Another example is in Exercise 1.5.8. However, a version of the Heine–Borel theorem is available if one is willing to replace closedness with the stronger notion of completeness and boundedness with the stronger notion of *total boundedness*; see Exercise 1.5.10.

One can characterize compactness topologically via the following, rather strange-sounding statement: every open cover of a compact set has a finite subcover.

Theorem 1.5.8 *Let (X, d) be a metric space, and let Y be a compact subset of X. Let $(V_\alpha)_{\alpha \in A}$ be a collection of open sets in X, and suppose that*

$$Y \subseteq \bigcup_{\alpha \in A} V_\alpha.$$

(i.e., the collection $(V_\alpha)_{\alpha \in A}$ covers *Y). Then there exists a* finite *subset F of A such that*

$$Y \subseteq \bigcup_{\alpha \in F} V_\alpha.$$

Proof We assume for sake of contradiction that there does not exist any finite subset F of A for which $Y \subseteq \bigcup_{\alpha \in F} V_\alpha$.

Let y be any element of Y. Then y must lie in at least one of the sets V_α. Since each V_α is open, there must therefore be an $r > 0$ such that $B_{(X,d)}(y, r) \subseteq V_\alpha$. Now let $r(y)$ denote the quantity

$$r(y) := \sup\{r \in (0, \infty) : B_{(X,d)}(y, r) \subseteq V_\alpha \text{ for some } \alpha \in A\}.$$

By the above discussion, we know that $r(y) > 0$ for all $y \in Y$. Now, let r_0 denote the quantity

$$r_0 := \inf\{r(y) : y \in Y\}.$$

Since $r(y) > 0$ for all $y \in Y$, we have $r_0 \geq 0$. There are three cases: $r_0 = 0$, $0 < r_0 < \infty$, and $r_0 = \infty$.

- **Case 1:** $r_0 = 0$. Then for every integer $n \geq 1$, there is at least one point y in Y such that $r(y) < 1/n$ (why?). We thus choose, for each $n \geq 1$, a point $y^{(n)}$ in Y such that $r(y^{(n)}) < 1/n$ (we can do this because of the axiom of choice, see Proposition 8.4.7). In particular we have $\lim_{n\to\infty} r(y^{(n)}) = 0$, by the squeeze test. The sequence $(y^{(n)})_{n=1}^\infty$ is a sequence in Y; since Y is compact, we can thus find a subsequence $(y^{(n_j)})_{j=1}^\infty$ which converges to a point $y_0 \in Y$.

  As before, we know that there exists some $\alpha \in A$ such that $y_0 \in V_\alpha$, and hence (since V_α is open) there exists some $\varepsilon > 0$ such that $B(y_0, \varepsilon) \subseteq V_\alpha$. Since $y^{(n_j)}$

converges to y_0, there must exist an $N \geq 1$ such that $y^{(n_j)} \in B(y_0, \varepsilon/2)$ for all $n \geq N$. In particular, by the triangle inequality we have $B(y^{(n_j)}, \varepsilon/2) \subseteq B(y_0, \varepsilon)$, and thus $B(y^{(n_j)}, \varepsilon/2) \subseteq V_\alpha$. By definition of $r(y^{(n_j)})$, this implies that $r(y^{(n_j)}) \geq \varepsilon/2$ for all $n \geq N$. But this contradicts the fact that $\lim_{n\to\infty} r(y^{(n)}) = 0$.

- **Case 2:** $0 < r_0 < \infty$. In this case we now have $r(y) > r_0/2$ for all $y \in Y$. This implies that for every $y \in Y$ there exists an $\alpha \in A$ such that $B(y, r_0/2) \subseteq V_\alpha$ (why?).
  We now construct a sequence $y^{(1)}, y^{(2)}, \ldots$ by the following recursive procedure. We let $y^{(1)}$ be any point in Y. The ball $B(y^{(1)}, r_0/2)$ is contained in one of the V_α and thus cannot cover all of Y, since we would then obtain a finite cover, a contradiction. Thus there exists a point $y^{(2)}$ which does not lie in $B(y^{(1)}, r_0/2)$, so in particular $d(y^{(2)}, y^{(1)}) \geq r_0/2$. Choose such a point $y^{(2)}$. The set $B(y^{(1)}, r_0/2) \cup B(y^{(2)}, r_0/2)$ cannot cover all of Y, since we would then obtain two sets V_{α_1} and V_{α_2} which covered Y, a contradiction again. So we can choose a point $y^{(3)}$ which does not lie in $B(y^{(1)}, r_0/2) \cup B(y^{(2)}, r_0/2)$, so in particular $d(y^{(3)}, y^{(1)}) \geq r_0/2$ and $d(y^{(3)}, y^{(2)}) \geq r_0/2$. Continuing in this fashion we obtain a sequence $(y^{(n)})_{n=1}^\infty$ in Y with the property that $d(y^{(k)}, y^{(j)}) \geq r_0/2$ for all $k > j$. In particular the sequence $(y^{(n)})_{n=1}^\infty$ is not a Cauchy sequence, and in fact no subsequence of $(y^{(n)})_{n=1}^\infty$ can be a Cauchy sequence either. But this contradicts the assumption that Y is compact (by Lemma 1.4.7).
- **Case 3:** $r_0 = \infty$. For this case we argue as in Case 2, but replacing the role of $r_0/2$ by (say) 1.

□

It turns out that Theorem 1.5.8 has a converse: if Y has the property that every open cover has a finite subcover, then it is compact (Exercise 1.5.11). In fact, this property is often considered the more fundamental notion of compactness than the sequence-based one. (For metric spaces, the two notions, that of compactness and sequential compactness, are equivalent, but for more general *topological spaces*, the two notions are slightly different, though we will not show this here.)

Theorem 1.5.8 has an important corollary: that every nested sequence of non-empty compact sets is still non-empty.

Corollary 1.5.9 *Let (X, d) be a metric space, and let $K_1, K_2, K_3, \ldots$ be a sequence of non-empty compact subsets of X such that*

$$K_1 \supseteq K_2 \supseteq K_3 \supseteq \cdots .$$

Then the intersection $\bigcap_{n=1}^\infty K_n$ is non-empty.

Proof See Exercise 1.5.6. □

We close this section by listing some miscellaneous properties of compact sets.

Theorem 1.5.10 *Let (X, d) be a metric space.*

(a) If Y is a compact subset of X, and $Z \subseteq Y$, then Z is compact if and only if Z is closed.

(b) If $Y_1, \ldots, Y_n$ are a finite collection of compact subsets of X, then their union $Y_1 \cup \ldots \cup Y_n$ is also compact.

(c) Every finite subset of X (including the empty set) is compact.

Proof See Exercise 1.5.7. □

— Exercises —

Exercise 1.5.1 Show that Definitions 9.1.22 and 1.5.3 match when talking about subsets of the real line with the standard metric.

Exercise 1.5.2 Prove Proposition 1.5.5. (*Hint:* prove the completeness and boundedness separately. For both claims, use proof by contradiction. You will need the axiom of choice, as in Lemma 8.4.5.)

Exercise 1.5.3 Prove Theorem 1.5.7. (*Hint:* use Proposition 1.1.18 and Theorem 9.1.24.)

Exercise 1.5.4 Let $(\mathbf{R}, d)$ be the real line with the standard metric. Give an example of a continuous function $f : \mathbf{R} \to \mathbf{R}$, and an open set $V \subseteq \mathbf{R}$, such that the image $f(V) := \{f(x) : x \in V\}$ of V is *not* open.

Exercise 1.5.5 Let $(\mathbf{R}, d)$ be the real line with the standard metric. Give an example of a continuous function $f : \mathbf{R} \to \mathbf{R}$, and a closed set $F \subseteq \mathbf{R}$, such that $f(F)$ is *not* closed.

Exercise 1.5.6 Prove Corollary 1.5.9. (*Hint:* work in the compact metric space $(K_1, d|_{K_1 \times K_1})$, and consider the sets $V_n := K_1 \backslash K_n$, which are open on K_1. Assume for sake of contradiction that $\bigcap_{n=1}^{\infty} K_n = \emptyset$, and then apply Theorem 1.5.8.)

Exercise 1.5.7 Prove Theorem 1.5.10. (*Hint:* for part (c), you may wish to use (b), and first prove that every singleton set is compact.)

Exercise 1.5.8 Let (X, d_{l^1}) be the metric space from Exercise 1.1.15. For each natural number n, let $e^{(n)} = (e_j^{(n)})_{j=0}^{\infty}$ be the sequence in X such that $e_j^{(n)} := 1$ when $n = j$ and $e_j^{(n)} := 0$ when $n \neq j$. Show that the set $\{e^{(n)} : n \in \mathbf{N}\}$ is a closed and bounded subset of X, but is not compact. (This is despite the fact that (X, d_{l^1}) is even a complete metric space—a fact which we will not prove here. The problem is that not that X is incomplete, but rather that it is "infinite-dimensional", in a sense that we will not discuss here.)

Exercise 1.5.9 Show that a metric space (X, d) is compact if and only if every sequence in X has at least one limit point.

Exercise 1.5.10 A metric space (X, d) is called *totally bounded* if for every $\varepsilon > 0$, there exists a natural number n and a finite number of balls $B(x^{(1)}, \varepsilon), \ldots, B(x^{(n)}, \varepsilon)$ which cover X (i.e., $X = \bigcup_{i=1}^{n} B(x^{(i)}, \varepsilon)$.

(a) Show that every totally bounded space is bounded.

(b) Show the following stronger version of Proposition 1.5.5: if (X, d) is compact, then complete and totally bounded. (*Hint:* if X is not totally bounded, then there is some $\varepsilon > 0$ such that X cannot be covered by finitely many ε-balls. Then use Exercise 8.5.20 to find an infinite sequence of balls $B(x^{(n)}, \varepsilon/2)$ which are disjoint from each other. Use this to then construct a sequence which has no convergent subsequence.)

(c) Conversely, show that if X is complete and totally bounded, then X is compact. (*Hint:* if $(x^{(n)})_{n=1}^{\infty}$ is a sequence in X, use the total boundedness hypothesis to recursively construct a sequence of subsequences $(x^{(n;j)})_{n=1}^{\infty}$ of $(x^{(n)})_{n=1}^{\infty}$ for each positive integer j, such that for each j, the elements of the sequence $(x^{(n;j)})_{n=1}^{\infty}$ are contained in a single ball of radius $1/j$, and also that each sequence $(x^{(n;j+1)})_{n=1}^{\infty}$ is a subsequence of the previous one $(x^{(n;j)})_{n=1}^{\infty}$. Then show that the "diagonal" sequence $(x^{(n;n)})_{n=1}^{\infty}$ is a Cauchy sequence, and then use the completeness hypothesis.)

Exercise 1.5.11 Let (X, d) have the property that every open cover of X has a finite subcover. Show that X is compact. (*Hint:* if X is not compact, then by Exercise 1.5.9, there is a sequence $(x^{(n)})_{n=1}^{\infty}$ with no limit points. Then for every $x \in X$ there exists a ball $B(x, \varepsilon)$ containing x which contains at most finitely many elements of this sequence. Now use the hypothesis.)

Exercise 1.5.12 Let (X, d_{disc}) be a metric space with the discrete metric d_{disc}.

(a) Show that X is always complete.

(b) When is X compact, and when is X not compact? Prove your claim. (*Hint:* the Heine–Borel theorem will be useless here since that only applies to Euclidean spaces.)

Exercise 1.5.13 Let E and F be two compact subsets of $\mathbf{R}$ (with the standard metric $d(x, y) = |x - y|$). Show that the Cartesian product $E \times F := \{(x, y) : x \in E, y \in F\}$ is a compact subset of $\mathbf{R}^2$ (with the Euclidean metric d_{l^2}).

Exercise 1.5.14 Let (X, d) be a metric space, let E be a non-empty compact subset of X, and let x_0 be a point in X. Show that there exists a point $x \in E$ such that

$$d(x_0, x) = \inf\{d(x_0, y) : y \in E\},$$

i.e., x is the closest point in E to x_0. (*Hint:* let R be the quantity $R := \inf\{d(x_0, y) : y \in E\}$. Construct a sequence $(x^{(n)})_{n=1}^{\infty}$ in E such that $d(x_0, x^{(n)}) \leq R + \frac{1}{n}$, and then use the compactness of E.)

Exercise 1.5.15 Let (X, d) be a compact metric space. Suppose that $(K_\alpha)_{\alpha \in I}$ is a collection of closed sets in X with the property that any finite subcollection of these sets necessarily has non-empty intersection, thus $\bigcap_{\alpha \in F} K_\alpha \neq \emptyset$ for all finite $F \subseteq I$. (This property is known as the *finite intersection property*.) Show that the *entire* collection has non-empty intersection, thus $\bigcap_{\alpha \in I} K_\alpha \neq \emptyset$. Show by counterexample that this statement fails if X is not compact.

Chapter 2
Continuous Functions on Metric Spaces

2.1 Continuous Functions

In the previous chapter we studied a single metric space (X, d), and the various types of sets one could find in that space. While this is already quite a rich subject, the theory of metric spaces becomes even richer, and of more importance to analysis, when one considers not just a single metric space, but rather *pairs* (X, d_X) and (Y, d_Y) of metric spaces, as well as *continuous functions* $f : X \to Y$ between such spaces. To define this concept, we generalize Definition 9.4.1 as follows:

Definition 2.1.1 (*Continuous functions*) Let (X, d_X) be a metric space, and let (Y, d_Y) be another metric space, and let $f : X \to Y$ be a function. If $x_0 \in X$, we say that f is *continuous at* x_0 iff for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_Y(f(x), f(x_0)) < \varepsilon$ whenever $d_X(x, x_0) < \delta$. We say that f is *continuous* iff it is continuous at every point $x \in X$.

Remark 2.1.2 Continuous functions are also sometimes called *continuous maps*. Mathematically, there is no distinction between the two terminologies.

Remark 2.1.3 If $f : X \to Y$ is continuous, and K is any subset of X, then the restriction $f|_K : K \to Y$ of f to K is also continuous (why?).

We now generalize much of the discussion in Chap. 9. We first observe that continuous functions preserve convergence:

Theorem 2.1.4 (Continuity preserves convergence) *Suppose that (X, d_X) and (Y, d_Y) are metric spaces. Let $f : X \to Y$ be a function, and let $x_0 \in X$ be a point in X. Then the following three statements are logically equivalent:*

(a) *f is continuous at x_0.*
(b) *Whenever $(x^{(n)})_{n=1}^{\infty}$ is a sequence in X which converges to x_0 with respect to the metric d_X, the sequence $(f(x^{(n)}))_{n=1}^{\infty}$ converges to $f(x_0)$ with respect to the metric d_Y.*

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_2

(c) For every open set $V \subseteq Y$ that contains $f(x_0)$, there exists an open set $U \subseteq X$ containing x_0 such that $f(U) \subseteq V$.

Proof See Exercise 2.1.1. □

Another important characterization of continuous functions involves open sets.

Theorem 2.1.5 *Let (X, d_X) be a metric space, and let (Y, d_Y) be another metric space. Let $f: X \to Y$ be a function. Then the following four statements are equivalent:*

(a) f is continuous.
(b) Whenever $(x^{(n)})_{n=1}^{\infty}$ is a sequence in X which converges to some point $x_0 \in X$ with respect to the metric d_X, the sequence $(f(x^{(n)}))_{n=1}^{\infty}$ converges to $f(x_0)$ with respect to the metric d_Y.
(c) Whenever V is an open set in Y, the set $f^{-1}(V) := \{x \in X : f(x) \in V\}$ is an open set in X.
(d) Whenever F is a closed set in Y, the set $f^{-1}(F) := \{x \in X : f(x) \in F\}$ is a closed set in X.

Proof See Exercise 2.1.2. □

Remark 2.1.6 It may seem strange that continuity ensures that the *inverse* image of an open set is open. One may guess instead that the reverse should be true, that the *forward* image of an open set is open; but this is not true; see Exercises 1.5.4, 1.5.5.

As a quick corollary of the above two theorems we obtain

Corollary 2.1.7 (Continuity preserved by composition) *Let (X, d_X), (Y, d_Y), and (Z, d_Z) be metric spaces.*

(a) If $f: X \to Y$ is continuous at a point $x_0 \in X$, and $g: Y \to Z$ is continuous at $f(x_0)$, then the composition $g \circ f: X \to Z$, defined by $g \circ f(x) := g(f(x))$, is continuous at x_0.
(b) If $f: X \to Y$ is continuous, and $g: Y \to Z$ is continuous, then $g \circ f: X \to Z$ is also continuous.

Proof See Exercise 2.1.3. □

Example 2.1.8 If $f: X \to \mathbf{R}$ is a continuous function, then the function $f^2 : X \to \mathbf{R}$ defined by $f^2(x) := f(x)^2$ is automatically continuous also. This is because we have $f^2 = g \circ f$, where $g: \mathbf{R} \to \mathbf{R}$ is the squaring function $g(x) := x^2$, and g is a continuous function.

— Exercises —

Exercise 2.1.1 Prove Theorem 2.1.4. (*Hint:* review your proof of Proposition 9.4.7.)

Exercise 2.1.2 Prove Theorem 2.1.5. (*Hint:* Theorem 2.1.4 already shows that (a) and (b) are equivalent.)

Exercise 2.1.3 Use Theorem 2.1.4 and Theorem 2.1.5 to prove Corollary 2.1.7.

Exercise 2.1.4 Give an example of functions $f: \mathbf{R} \to \mathbf{R}$ and $g: \mathbf{R} \to \mathbf{R}$ such that

(a) f is not continuous, but g and $g \circ f$ are continuous.
(b) g is not continuous, but f and $g \circ f$ are continuous.
(c) f and g are not continuous, but $g \circ f$ is continuous.

Explain briefly why these examples do not contradict Corollary 2.1.7.

Exercise 2.1.5 Let (X, d) be a metric space, and let $(E, d|_{E\times E})$ be a subspace of (X, d). Let $\iota_{E\to X} : E \to X$ be the inclusion map, defined by setting $\iota_{E\to X}(x) := x$ for all $x \in E$. Show that $\iota_{E\to X}$ is continuous.

Exercise 2.1.6 Let $f: X \to Y$ be a function from one metric space (X, d_X) to another (Y, d_Y). Let E be a subset of X (which we give the induced metric $d_X|_{E\times E}$), and let $f|_E : E \to Y$ be the restriction of f to E, thus $f|_E(x) := f(x)$ when $x \in E$. If $x_0 \in E$ and f is continuous at x_0, show that $f|_E$ is also continuous at x_0. (Is the converse of this statement true? Explain.) Conclude that if f is continuous, then $f|_E$ is continuous. Thus restriction of the domain of a function does not destroy continuity. (*Hint:* use Exercise 2.1.5.)

Exercise 2.1.7 Let $f: X \to Y$ be a function from one metric space (X, d_X) to another (Y, d_Y). Suppose that the image $f(X)$ of X is contained in some subset $E \subseteq Y$ of Y. Let $g: X \to E$ be the function which is the same as f but with the codomain restricted from Y to E, thus $g(x) = f(x)$ for all $x \in X$. We give E the metric $d_Y|_{E\times E}$ induced from Y. Show that for any $x_0 \in X$, that f is continuous at x_0 if and only if g is continuous at x_0. Conclude that f is continuous if and only if g is continuous. (Thus the notion of continuity is not affected if one restricts the codomain of the function.)

2.2 Continuity and Product Spaces

Given two functions $f: X \to Y$ and $g: X \to Z$, one can define their *pairing* $(f, g): X \to Y \times Z$ defined by $(f, g)(x) := (f(x), g(x))$, i.e., this is the function taking values in the Cartesian product $Y \times Z$ whose first coordinate is $f(x)$ and whose second coordinate is $g(x)$ (cf. Exercise 3.5.7). For instance, if $f: \mathbf{R} \to \mathbf{R}$ is the function $f(x) := x^2 + 3$, and $g: \mathbf{R} \to \mathbf{R}$ is the function $g(x) = 4x$, then $(f, g): \mathbf{R} \to \mathbf{R}^2$ is the function $(f, g)(x) := (x^2 + 3, 4x)$. The pairing operation preserves continuity:

Lemma 2.2.1 *Let $f: X \to \mathbf{R}$ and $g: X \to \mathbf{R}$ be functions, and let $(f, g): X \to \mathbf{R}^2$ be their direct sum. We give $\mathbf{R}^2$ the Euclidean metric.*

(a) If $x_0 \in X$, then f and g are both continuous at x_0 if and only if (f, g) is continuous at x_0.
(b) f and g are both continuous if and only if (f, g) is continuous.

Proof See Exercise 2.2.1. □

To use this, we first need another continuity result:

Lemma 2.2.2 *The addition function* $(x, y) \mapsto x + y$*, the subtraction function* $(x, y) \mapsto x - y$*, the multiplication function* $(x, y) \mapsto xy$*, the maximum function* $(x, y) \mapsto \max(x, y)$*, and the minimum function* $(x, y) \mapsto \min(x, y)$ *are all continuous functions from* $\mathbf{R}^2$ *to* $\mathbf{R}$*. The division function* $(x, y) \mapsto x/y$ *is a continuous function from* $\mathbf{R} \times (\mathbf{R}\backslash\{0\}) = \{(x, y) \in \mathbf{R}^2 : y \neq 0\}$ *to* $\mathbf{R}$*. For any real number c, the function* $x \mapsto cx$ *is a continuous function from* $\mathbf{R}$ *to* $\mathbf{R}$*.*

Proof See Exercise 2.2.2. □

Combining these lemmas we obtain

Corollary 2.2.3 *Let* (X, d) *be a metric space, and let* $f : X \to \mathbf{R}$ *and* $g : X \to \mathbf{R}$ *be functions. Let c be a real number.*

(a) If $x_0 \in X$ *and f and g are continuous at* x_0*, then the functions* $f + g : X \to \mathbf{R}$*,* $f - g : X \to \mathbf{R}$*,* $fg : X \to \mathbf{R}$*,* $\max(f, g) : X \to \mathbf{R}$*,* $\min(f, g) : X \to \mathbf{R}$*, and* $cf : X \to \mathbf{R}$ *(see Definition 9.2.1 for definitions) are also continuous at* x_0*. If* $g(x) \neq 0$ *for all* $x \in X$*, then* $f/g : X \to \mathbf{R}$ *is also continuous.*

(b) If f and g are continuous, then the functions $f + g : X \to \mathbf{R}$*,* $f - g : X \to \mathbf{R}$*,* $fg : X \to \mathbf{R}$*,* $\max(f, g) : X \to \mathbf{R}$*,* $\min(f, g) : X \to \mathbf{R}$*, and* $cf : X \to \mathbf{R}$ *are also continuous at* x_0*. If* $g(x) \neq 0$ *for all* $x \in X$*, then* $f/g : X \to \mathbf{R}$ *is also continuous at* x_0*.*

Proof We first prove (a). Since f and g are continuous at x_0, then by Lemma 2.2.1 $(f, g) : X \to \mathbf{R}^2$ is also continuous at x_0. On the other hand, from Lemma 2.2.2 the function $(x, y) \mapsto x + y$ is continuous at every point in $\mathbf{R}^2$ and in particular is continuous at $(f, g)(x_0)$. If we then compose these two functions using Corollary 2.1.7 we conclude that $f + g : X \to \mathbf{R}$ is continuous. A similar argument gives the continuity of $f - g$, fg, $\max(f, g)$, $\min(f, g)$, and cf. To prove the claim for f/g, we first use Exercise 2.1.7 to restrict the codomain of g from $\mathbf{R}$ to $\mathbf{R}\backslash\{0\}$, and then one can argue as before. The claim (b) follows immediately from (a). □

This corollary allows us to demonstrate the continuity of a large class of functions; we give some examples below.

— Exercises —

Exercise 2.2.1 Prove Lemma 2.2.1. (*Hint:* use Proposition 1.1.18 and Theorem 2.1.4.)

Exercise 2.2.2 Prove Lemma 2.2.2. (*Hint:* use Theorem 2.1.5 and limit laws (Theorem 6.1.19).)

Exercise 2.2.3 Show that if $f : X \to \mathbf{R}$ is a continuous function, so is the function $|f| : X \to \mathbf{R}$ defined by $|f|(x) := |f(x)|$.

Exercise 2.2.4 Let $\pi_1 \colon \mathbf{R}^2 \to \mathbf{R}$ and $\pi_2 : \mathbf{R}^2 \to \mathbf{R}$ be the functions $\pi_1(x, y) := x$ and $\pi_2(x, y) := y$ (these two functions are sometimes called the *coordinate functions* on $\mathbf{R}^2$). Show that π_1 and π_2 are continuous. Conclude that if $f \colon \mathbf{R} \to X$ is any continuous function into a metric space (X, d), then the functions $g_1 \colon \mathbf{R}^2 \to X$ and $g_2 : \mathbf{R}^2 \to X$ defined by $g_1(x, y) := f(x)$ and $g_2(x, y) := f(y)$ are also continuous.

Exercise 2.2.5 Let $n, m \geq 0$ be integers. Suppose that for every $0 \leq i \leq n$ and $0 \leq j \leq m$ we have a real number c_{ij}. Form the function $P \colon \mathbf{R}^2 \to \mathbf{R}$ defined by

$$P(x, y) := \sum_{i=0}^{n} \sum_{j=0}^{m} c_{ij} x^i y^j.$$

(Such a function is known as a *polynomial of two variables*; a typical example of such a polynomial is $P(x, y) = x^3 + 2xy^2 - x^2 + 3y + 6$.) Show that P is continuous. (*Hint:* use Exercise 2.2.4 and Corollary 2.2.3.) Conclude that if $f \colon X \to \mathbf{R}$ and $g \colon X \to \mathbf{R}$ are continuous functions, then the function $P(f, g) : X \to \mathbf{R}$ defined by $P(f, g)(x) := P(f(x), g(x))$ is also continuous.

Exercise 2.2.6 Let $\mathbf{R}^m$ and $\mathbf{R}^n$ be Euclidean spaces. If $f \colon X \to \mathbf{R}^m$ and $g \colon X \to \mathbf{R}^n$ are continuous functions, show that $(f, g) \colon X \to \mathbf{R}^{m+n}$ is also continuous, where we have identified $\mathbf{R}^m \times \mathbf{R}^n$ with $\mathbf{R}^{m+n}$ in the obvious manner. Is the converse statement true?

Exercise 2.2.7 Let $k \geq 1$, let I be a finite subset of $\mathbf{N}^k$, and let $c : I \to \mathbf{R}$ be a function. Form the function $P \colon \mathbf{R}^k \to \mathbf{R}$ defined by

$$P(x_1, \dots, x_k) := \sum_{(i_1, \dots, i_k) \in I} c(i_1, \dots, i_k) x_1^{i_1} \dots x_k^{i_k}.$$

(Such a function is known as a *polynomial of k variables*; a typical example of such a polynomial is $P(x_1, x_2, x_3) = 3x_1^3 x_2 x_3^2 - x_2 x_3^2 + x_1 + 5$.) Show that P is continuous. (*Hint:* use induction on k, Exercise 2.2.6, and either Exercise 2.2.5 or Lemma 2.2.2.)

Exercise 2.2.8 Let (X, d_X) and (Y, d_Y) be metric spaces. Define the metric $d_{X \times Y} : (X \times Y) \times (X \times Y) \to [0, \infty)$ by the formula

$$d_{X \times Y}((x, y), (x', y')) := d_X(x, x') + d_Y(y, y').$$

Show that $(X \times Y, d_{X \times Y})$ is a metric space, and deduce an analogue of Proposition 1.1.18 and Lemma 2.2.1.

Exercise 2.2.9 Let $f \colon \mathbf{R}^2 \to \mathbf{R}$ be a function from $\mathbf{R}^2$ to $\mathbf{R}$. Let (x_0, y_0) be a point in $\mathbf{R}^2$. If f is continuous at (x_0, y_0), show that

$$\lim_{x \to x_0} \limsup_{y \to y_0} f(x, y) = \lim_{y \to y_0} \limsup_{x \to x_0} f(x, y) = f(x_0, y_0)$$

and

$$\lim_{x\to x_0} \liminf_{y\to y_0} f(x,y) = \lim_{y\to y_0} \liminf_{x\to x_0} f(x,y) = f(x_0, y_0).$$

(Recall that $\limsup_{x\to x_0} f(x) := \inf_{r>0} \sup_{|x-x_0|<r} f(x)$ and $\liminf_{x\to x_0} f(x) := \sup_{r>0} \inf_{|x-x_0|<r} f(x)$.) In particular, we have

$$\lim_{x\to x_0} \lim_{y\to y_0} f(x,y) = \lim_{y\to y_0} \lim_{x\to x_0} f(x,y)$$

whenever the limits on both sides exist. (Note that the limits do not necessarily exist in general; consider for instance the function $f\colon \mathbf{R}^2 \to \mathbf{R}$ such that $f(x,y) = y \sin\frac{1}{x}$ when $xy \neq 0$ and $f(x,y) = 0$ otherwise.) Discuss the comparison between this result and Example 1.2.7.

Exercise 2.2.10 Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be a continuous function. Show that for each $x \in \mathbf{R}$, the function $y \mapsto f(x,y)$ is continuous on $\mathbf{R}$, and for each $y \in \mathbf{R}$, the function $x \mapsto f(x,y)$ is continuous on $\mathbf{R}$. Thus a function $f(x,y)$ which is jointly continuous in (x,y) is also continuous in each variable x, y separately.

Exercise 2.2.11 Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be the function defined by $f(x,y) := \frac{xy}{x^2+y^2}$ when $(x,y) \neq (0,0)$, and $f(x,y) = 0$ otherwise. Show that for each fixed $x \in \mathbf{R}$, the function $y \mapsto f(x,y)$ is continuous on $\mathbf{R}$, and that for each fixed $y \in \mathbf{R}$, the function $x \mapsto f(x,y)$ is continuous on $\mathbf{R}$, but that the function $f\colon \mathbf{R}^2 \to \mathbf{R}$ is not continuous on $\mathbf{R}^2$. This shows that the converse to Exercise 2.2.10 fails; it is possible to be continuous in each variable separately without being jointly continuous.

Exercise 2.2.12 Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be the function defined by $f(x,y) := x^2/y$ when $y \neq 0$, and $f(x,y) := 0$ when $y = 0$. Show that $\lim_{t\to 0} f(tx, ty) = f(0,0)$ for every $(x,y) \in \mathbf{R}^2$, but that f is not continuous at the origin. Thus being continuous on every line through the origin is not enough to guarantee continuity at the origin!

2.3 Continuity and Compactness

Continuous functions interact well with the concept of compact sets defined in Definition 1.5.1.

Theorem 2.3.1 (Continuous maps preserve compactness) *Let $f\colon X \to Y$ be a continuous map from one metric space (X, d_X) to another (Y, d_Y). Let $K \subseteq X$ be any compact subset of X. Then the image $f(K) := \{f(x) : x \in K\}$ of K is also compact.*

Proof See Exercise 2.3.1. □

This theorem has an important consequence. Recall from Definition 9.6.5 the notion of a function $f\colon X \to \mathbf{R}$ attaining a maximum or minimum at a point. We may generalize Proposition 9.6.7 as follows:

Proposition 2.3.2 (Maximum principle) *Let (X, d) be a compact metric space, and let $f : X \to \mathbf{R}$ be a continuous function. Then f is bounded. Furthermore, if X is non-empty, then f attains its maximum at some point $x_{max} \in X$ and also attains its minimum at some point $x_{min} \in X$.*

Proof See Exercise 2.3.2. □

Remark 2.3.3 As was already noted in Exercise 9.6.1, this principle can fail if X is not compact. This proposition should be compared with Lemma 9.6.3 and Proposition 9.6.7.

Another advantage of continuous functions on compact sets is that they are *uniformly continuous*. We generalize Definition 9.9.2 as follows:

Definition 2.3.4 (*Uniform continuity*) Let $f : X \to Y$ be a map from one metric space (X, d_X) to another (Y, d_Y). We say that f is *uniformly continuous* if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_Y(f(x), f(x')) < \varepsilon$ whenever $x, x' \in X$ are such that $d_X(x, x') < \delta$.

Every uniformly continuous function is continuous, but not conversely (Exercise 2.3.3). But if the domain X is compact, then the two notions are equivalent:

Theorem 2.3.5 *Let (X, d_X) and (Y, d_Y) be metric spaces, and suppose that (X, d_X) is compact. If $f : X \to Y$ is function, then f is continuous if and only if it is uniformly continuous.*

Proof If f is uniformly continuous then it is also continuous by Exercise 2.3.3. Now suppose that f is continuous. Fix $\varepsilon > 0$. For every $x_0 \in X$, the function f is continuous at x_0. Thus there exists a $\delta(x_0) > 0$, depending on x_0, such that $d_Y(f(x), f(x_0)) < \varepsilon/2$ whenever $d_X(x, x_0) < \delta(x_0)$. In particular, by the triangle inequality this implies that $d_Y(f(x), f(x')) < \varepsilon$ whenever $x \in B_{(X,d_X)}(x_0, \delta(x_0)/2)$ and $d_X(x', x) < \delta(x_0)/2$ (why?).

Now consider the (possibly infinite) collection of balls

$$\{B_{(X,d_X)}(x_0, \delta(x_0)/2) : x_0 \in X\}.$$

Each ball in this collection is of course open, and the union of all these balls covers X, since each point x_0 in X is contained in its own ball $B_{(X,d_X)}(x_0, \delta(x_0)/2)$. Hence, by Theorem 1.5.8, there exist a finite number of points $x_1, \ldots, x_n$ such that the balls $B_{(X,d_X)}(x_j, \delta(x_j)/2)$ for $j = 1, \ldots, n$ cover X:

$$X \subseteq \bigcup_{j=1}^{n} B_{(X,d_X)}(x_j, \delta(x_j)/2).$$

Now let $\delta := \min_{j=1}^{n} \delta(x_j)/2$. Since each of the $\delta(x_j)$ is positive, and there are only a finite number of j, we see that $\delta > 0$. Now let x, x' be any two points in X such that

$d_X(x, x') < \delta$. Since the balls $B_{(X,d_X)}(x_j, \delta(x_j)/2)$ cover X, we see that there must exist $1 \le j \le n$ such that $x \in B_{(X,d_X)}(x_j, \delta(x_j)/2)$. Since $d_X(x, x') < \delta$, we have $d_X(x, x') < \delta(x_j)/2$, and so by the previous discussion we have $d_Y(f(x), f(x')) < \varepsilon$. We have thus found a δ such that $d_Y(f(x), f(x')) < \varepsilon$ whenever $d(x, x') < \delta$, and this proves uniform continuity as desired. □

— Exercises —

Exercise 2.3.1 Prove Theorem 2.3.1.

Exercise 2.3.2 Prove Proposition 2.3.2. (*Hint:* modify the proof of Proposition 9.6.7.)

Exercise 2.3.3 Show that every uniformly continuous function is continuous, but give an example that shows that not every continuous function is uniformly continuous.

Exercise 2.3.4 Let (X, d_X), (Y, d_Y), (Z, d_Z) be metric spaces, and let $f\colon X \to Y$ and $g\colon Y \to Z$ be two uniformly continuous functions. Show that $g \circ f\colon X \to Z$ is also uniformly continuous.

Exercise 2.3.5 Let (X, d_X) be a metric space, and let $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ be uniformly continuous functions. Show that the pairing $(f, g)\colon X \to \mathbf{R}^2$ defined by $(f, g)(x) := (f(x), g(x))$ is uniformly continuous.

Exercise 2.3.6 Show that the addition function $(x, y) \mapsto x + y$ and the subtraction function $(x, y) \mapsto x - y$ are uniformly continuous from $\mathbf{R}^2$ to $\mathbf{R}$, but the multiplication function $(x, y) \mapsto xy$ is not. Conclude that if $f\colon X \to \mathbf{R}$ and $g\colon X \to \mathbf{R}$ are uniformly continuous functions on a metric space (X, d), then $f + g\colon X \to \mathbf{R}$ and $f - g\colon X \to \mathbf{R}$ are also uniformly continuous. Give an example to show that $fg\colon X \to \mathbf{R}$ need not be uniformly continuous. What is the situation for $\max(f, g)$, $\min(f, g)$, f/g, and cf for a real number c?

2.4 Continuity and Connectedness

We now describe another important concept in metric spaces, that of *connectedness*.

Definition 2.4.1 (*Connected spaces*) Let (X, d) be a metric space. We say that X is *disconnected* iff there exist disjoint non-empty open sets V and W in X such that $V \cup W = X$. (Equivalently, X is disconnected if and only if X contains a non-empty proper subset which is simultaneously closed and open.) We say that X is *connected* iff it is non-empty and not disconnected.

We declare the empty set $\emptyset$ as being special—it is neither connected nor disconnected; one could think of the empty set as “unconnected”.

Example 2.4.2 Consider the set $X := [1, 2] \cup [3, 4]$, with the usual metric. This set is disconnected because the sets $[1, 2]$ and $[3, 4]$ are open relative to X (why?).

Intuitively, a disconnected set is one which can be separated into two disjoint open sets; a connected set is one which cannot be separated in this manner. We defined what it means for a metric space to be connected; we can also define what it means for a set to be connected.

Definition 2.4.3 (*Connected sets*) Let (X, d) be a metric space, and let Y be a subset of X. We say that Y is *connected* iff the metric space $(Y, d|_{Y \times Y})$ is connected, and we say that Y is *disconnected* iff the metric space $(Y, d|_{Y \times Y})$ is disconnected.

Remark 2.4.4 This definition is intrinsic; whether a set Y is connected or not depends only on what the metric is doing on Y, but not on what ambient space X one placing Y in.

On the real line, connected sets are easy to describe.

Theorem 2.4.5 *Let X be a non-empty subset of the real line* $\mathbf{R}$. *Then the following statements are equivalent.*

(a) X is connected.
(b) Whenever $x, y \in X$ and $x < y$, the interval $[x, y]$ is also contained in X.
(c) X is an interval (in the sense of Definition 9.1.1).

Proof First we show that (a) implies (b). Suppose that X is connected, and suppose for sake of contradiction that we could find points $x < y$ in X such that $[x, y]$ is *not* contained in X. Then there exists a real number $x < z < y$ such that $z \notin X$. Thus the sets $(-\infty, z) \cap X$ and $(z, \infty) \cap X$ will cover X. But these sets are non-empty (because they contain x and y, respectively) and are open relative to X, and so X is disconnected, a contradiction.

Now we show that (b) implies (a). Let X be a set obeying the property (b). Suppose for sake of contradiction that X is disconnected. Then there exist disjoint non-empty sets V, W which are open relative to X, such that $V \cup W = X$. Since V and W are non-empty, we may choose an $x \in V$ and $y \in W$. Since V and W are disjoint, we have $x \neq y$; without loss of generality we may assume $x < y$. By property (b), we know that the entire interval $[x, y]$ is contained in X.

Now consider the set $[x, y] \cap V$. This set is both bounded and non-empty (because it contains x). Thus it has a supremum

$$z := \sup([x, y] \cap V).$$

Clearly $z \in [x, y]$, and hence $z \in X$. Thus either $z \in V$ or $z \in W$. Suppose first that $z \in V$. Then $z \neq y$ (since $y \in W$ and V is disjoint from W). But V is open relative to X, which contains $[x, y]$, so there is some ball $B_{([x,y],d)}(z, r)$ which is contained in V. But this contradicts the fact that z is the supremum of $[x, y] \cap V$. Now suppose that $z \in W$. Then $z \neq x$ (since $x \in V$ and V is disjoint from W). But W is open relative to X, which contains $[x, y]$, so there is some ball $B_{([x,y],d)}(z, r)$ which is contained in W. But this again contradicts the fact that z is the supremum of $[x, y] \cap V$. Thus in either case we obtain a contradiction, which means that X cannot be disconnected and must therefore be connected.

It remains to show that (b) and (c) are equivalent; we leave this to Exercise 2.4.3. □

Continuous functions map connected sets to connected sets:

Theorem 2.4.6 (Continuity preserves connectedness) *Let $f : X \to Y$ be a continuous map from one metric space (X, d_X) to another (Y, d_Y). Let E be any connected subset of X. Then $f(E)$ is also connected.*

Proof See Exercise 2.4.4. □

An important corollary of this result is the intermediate value theorem, generalizing Theorem 9.7.1.

Corollary 2.4.7 (Intermediate value theorem) *Let $f : X \to \mathbf{R}$ be a continuous map from one metric space (X, d_X) to the real line. Let E be any connected subset of X, and let a, b be any two elements of E. Let y be a real number between $f(a)$ and $f(b)$, i.e., either $f(a) \leq y \leq f(b)$ or $f(a) \geq y \geq f(b)$. Then there exists $c \in E$ such that $f(c) = y$.*

Proof See Exercise 2.4.5. □

— Exercises —

Exercise 2.4.1 Let (X, d_{disc}) be a metric space with the discrete metric. Let E be a subset of X which contains at least two elements. Show that E is disconnected.

Exercise 2.4.2 Let $f : X \to Y$ be a function from a connected metric space (X, d) to a metric space (Y, d_{disc}) with the discrete metric. Show that f is continuous if and only if it is constant. (*Hint:* use Exercise 2.4.1.)

Exercise 2.4.3 Prove the equivalence of statements (b) and (c) in Theorem 2.4.5.

Exercise 2.4.4 Prove Theorem 2.4.6. (*Hint:* the formulation of continuity in Theorem 2.1.5(c) is the most convenient to use.)

Exercise 2.4.5 Use Theorem 2.4.6 to prove Corollary 2.4.7.

Exercise 2.4.6 Let (X, d) be a metric space, and let $(E_\alpha)_{\alpha \in I}$ be a collection of connected sets in X with I non-empty. Suppose also that $\bigcap_{\alpha \in I} E_\alpha$ is non-empty. Show that $\bigcup_{\alpha \in I} E_\alpha$ is connected.

Exercise 2.4.7 Let (X, d) be a metric space, and let E be a subset of X. We say that E is *path-connected* iff, for every $x, y \in E$, there exists a continuous function $\gamma : [0, 1] \to E$ from the unit interval $[0, 1]$ to E such that $\gamma(0) = x$ and $\gamma(1) = y$. Show that every non-empty path-connected set is connected. (The converse is false, but is a bit tricky to show and will not be detailed here.)

Exercise 2.4.8 Let (X, d) be a metric space, and let E be a subset of X. Show that if E is connected, then the closure $\overline{E}$ of E is also connected. Is the converse true?

Exercise 2.4.9 Let (X, d) be a metric space. Let us define a relation $x \sim y$ on X by declaring $x \sim y$ iff there exists a connected subset of X which contains both x and y. Show that this is an equivalence relation (i.e., it obeys the reflexive, symmetric, and transitive axioms). Also, show that the equivalence classes of this relation (i.e., the sets of the form $\{y \in X : y \sim x\}$ for some $x \in X$) are all closed and connected. (*Hint:* use Exercise 2.4.6 and Exercise 2.4.8.) These sets are known as the *connected components* of X.

Exercise 2.4.10 Combine Proposition 2.3.2 and Corollary 2.4.7 to deduce a theorem for continuous functions on a compact connected domain which generalizes Corollary 9.7.4.

2.5 Topological Spaces (Optional)

The concept of a metric space can be generalized to that of a *topological space*. The idea here is not to view the metric d as the fundamental object; indeed, in a general topological space there is no metric at all. Instead, it is the collection of *open sets* which is the fundamental concept. Thus, whereas in a metric space one introduces the metric d first, and then uses the metric to define first the concept of an open ball and then the concept of an open set, in a topological space one starts just with the notion of an open set. As it turns out, starting from the open sets, one cannot necessarily reconstruct a usable notion of a ball or metric (thus not all topological spaces will be metric spaces), but remarkably one can still define many of the concepts in the preceding sections.

We will not use topological spaces at all in this text, and so we shall be rather brief in our treatment of them here. A more complete study of these spaces can of course be found in any topology textbook or a more advanced analysis text.

Definition 2.5.1 (*Topological spaces*) A *topological space* is a pair $(X, \mathcal{F})$, where X is a set and $\mathcal{F} \subseteq 2^X$ is a collection of subsets of X, whose elements are referred to as *open sets*. Furthermore, the collection $\mathcal{F}$ must obey the following properties:

- The empty set $\emptyset$ and the whole set X are open; in other words, $\emptyset \in \mathcal{F}$ and $X \in \mathcal{F}$.
- Any finite intersection of open sets is open. In other words, if $V_1, \ldots, V_n$ is elements of $\mathcal{F}$, then $V_1 \cap \ldots \cap V_n$ is also in $\mathcal{F}$.
- Any arbitrary union of open sets is open (including infinite unions). In other words, if $(V_\alpha)_{\alpha \in I}$ is a family of sets in $\mathcal{F}$, then $\bigcup_{\alpha \in I} V_\alpha$ is also in $\mathcal{F}$.

In many cases, the collection $\mathcal{F}$ of open sets can be deduced from context, and we shall refer to the topological space $(X, \mathcal{F})$ simply as X.

From Proposition 1.2.15 we see that every metric space (X, d) is automatically also a topological space (if we set $\mathcal{F}$ equal to the collection of sets which are open in (X, d)). However, there do exist topological spaces which do not arise from metric spaces (see Exercise 2.5.1, 2.5.6).

We now develop the analogues of various notions in this chapter and the previous chapter for topological spaces. The notion of a ball must be replaced by the notion of a *neighbourhood*.

Definition 2.5.2 (*Neighborhoods*) Let $(X, \mathcal{F})$ be a topological space, and let $x \in X$. A *neighborhood of* x is defined to be any open set in $\mathcal{F}$ which contains x.

Example 2.5.3 If (X, d) is a metric space, $x \in X$, and $r > 0$, then $B(x, r)$ is a neighborhood of x.

Definition 2.5.4 (*Topological convergence*) Let m be an integer, $(X, \mathcal{F})$ be a topological space and let $(x^{(n)})_{n=m}^{\infty}$ be a sequence of points in X. Let x be a point in X. We say that $(x^{(n)})_{n=m}^{\infty}$ *converges to* x if and only if, for every neighborhood V of x, there exists an $N \geq m$ such that $x^{(n)} \in V$ for all $n \geq N$.

This notion is consistent with that of convergence in metric spaces (Exercise 2.5.2). One can then ask whether one has the basic property of uniqueness of limits (Proposition 1.1.20). The answer turns out to usually be yes—if the topological space has an additional property known as the *Hausdorff property*—but the answer can be no for other topologies; see Exercise 2.5.4.

Definition 2.5.5 (*Interior, exterior, boundary*) Let $(X, \mathcal{F})$ be a topological space, let E be a subset of X, and let x_0 be a point in X. We say that x_0 is an *interior point of* E if there exists a neighborhood V of x_0 such that $V \subseteq E$. We say that x_0 is an *exterior point of* E if there exists a neighborhood V of x_0 such that $V \cap E = \emptyset$. We say that x_0 is a *boundary point of* E if it is neither an interior point nor an exterior point of E.

This definition is consistent with the corresponding notion for metric spaces (Exercise 2.5.3).

Definition 2.5.6 (*Closure*) Let $(X, \mathcal{F})$ be a topological space, let E be a subset of X, and let x_0 be a point in X. We say that x_0 is an *adherent point* of E if every neighborhood V of x_0 has a non-empty intersection with E. The set of all adherent points of E is called the *closure* of E and is denoted $\overline{E}$.

There is a partial analogue of Theorem 1.2.10, see Exercise 2.5.9.

We define a set K in a topological space $(X, \mathcal{F})$ to be *closed* iff its complement $X \backslash K$ is open; this is consistent with the metric space definition, thanks to Proposition 1.2.15(e). Some partial analogues of that proposition are true (see Exercise 2.5.10).

To define the notion of a relative topology, we cannot use Definition 1.3.3 as this requires a metric function. However, we can instead use Proposition 1.3.4 as our starting point:

Definition 2.5.7 (*Relative topology*) Let $(X, \mathcal{F})$ be a topological space, and Y be a subset of X. Then we define $\mathcal{F}_Y := \{V \cap Y : V \in \mathcal{F}\}$ and refer this as the topology on Y *induced* by $(X, \mathcal{F})$. We call $(Y, \mathcal{F}_Y)$ a *topological subspace* of $(X, \mathcal{F})$. This is indeed a topological space, see Exercise 2.5.11.

From Proposition 1.3.4 we see that this notion is compatible with the one for metric spaces.

Next we define the notion of continuity.

Definition 2.5.8 (*Continuous functions*) Let $(X, \mathcal{F}_X)$ and $(Y, \mathcal{F}_Y)$ be topological spaces, and let $f : X \to Y$ be a function. If $x_0 \in X$, we say that f is *continuous at* x_0 iff for every neighborhood V of $f(x_0)$, there exists a neighborhood U of x_0 such that $f(U) \subseteq V$. We say that f is *continuous* iff it is continuous at every point $x \in X$.

This definition is consistent with that in Definition 2.1.1 (Exercise 2.5.14). Partial analogues of Theorems 2.1.4 and 2.1.5 are available (Exercise 2.5.15). In particular, a function is continuous iff the pre-images of every open set are open.

There is unfortunately no notion of a Cauchy sequence, a complete space, or a bounded space, for general topological spaces. However, there is certainly a notion of a compact space, as we can see by taking Theorem 1.5.8 as our starting point:

Definition 2.5.9 (*Compact topological spaces*) Let $(X, \mathcal{F})$ be a topological space. We say that this space is *compact* if every open cover of X has a finite subcover. If Y is a subset of X, we say that Y is compact if the topological space on Y induced by $(X, \mathcal{F})$ is compact.

Many basic facts about compact metric spaces continue to hold true for compact topological spaces, notably Theorem 2.3.1 and Proposition 2.3.2 (Exercise 2.5.16). However, there is no notion of uniform continuity, and so there is no analogue of Theorem 2.3.5.

We can also define the notion of connectedness by repeating Definition 2.4.1 verbatim and also repeating Definition 2.4.3 (but with Definition 2.5.7 instead of Definition 1.3.3). Many of the results and exercises in Sect. 2.4 continue to hold for topological spaces (with almost no changes to any of the proofs!).

— Exercises —

Exercise 2.5.1 Let X be an arbitrary set, and let $\mathcal{F} := \{\emptyset, X\}$. Show that $(X, \mathcal{F})$ is a topology (called the *trivial topology* on X). If X contains more than one element, show that the trivial topology cannot be obtained from by placing a metric d on X. Show that this topological space is both compact and connected.

Exercise 2.5.2 Let (X, d) be a metric space (and hence a topological space). Show that the two notions of convergence of sequences in Definition 1.1.14 and Definition 2.5.4 coincide.

Exercise 2.5.3 Let (X, d) be a metric space (and hence a topological space). Show that the two notions of interior, exterior, and boundary in Definition 1.2.5 and Definition 2.5.5 coincide.

Exercise 2.5.4 A topological space $(X, \mathcal{F})$ is said to be *Hausdorff* if given any two distinct points $x, y \in X$, there exists a neighborhood V of x and a neighborhood W of y such that $V \cap W = \emptyset$. Show that any topological space coming from a metric space is Hausdorff, and show that the trivial topology is not Hausdorff. Show that the analogue of Proposition 1.1.20 holds for Hausdorff topological spaces, but give an example of a non-Hausdorff topological space in which Proposition 1.1.20 fails. (In practice, most topological spaces one works with are Hausdorff; non-Hausdorff topological spaces tend to be so pathological that it is not very profitable to work with them.)

Exercise 2.5.5 Given any totally ordered set X with order relation $\leq$, declare a set $V \subseteq X$ to be *open* if for every $x \in V$ there exists a set I which is an interval $\{y \in X : a < y < b\}$ for some $a, b \in X$, a ray $\{y \in X : a < y\}$ for some $a \in X$, the ray $\{y \in X : y < b\}$ for some $b \in X$, or the whole space X, which contains x and is contained in V. Let $\mathcal{F}$ be the set of all open subsets of X. Show that $(X, \mathcal{F})$ is a topology (this is the *order topology* on the totally ordered set $(X, \leq)$) which is Hausdorff in the sense of Exercise 2.5.4. Show that on the real line $\mathbf{R}$ (with the standard ordering $\leq$), the order topology matches the standard topology (i.e., the topology arising from the standard metric). If instead one applies this to the extended real line $\mathbf{R}^*$, show that $\mathbf{R}$ is an open set with boundary $\{-\infty, +\infty\}$. If $(x_n)_{n=1}^{\infty}$ is a sequence of numbers in $\mathbf{R}$ (and hence in $\mathbf{R}^*$), show that x_n converges to $+\infty$ if and only if $\liminf_{n\to\infty} x_n = +\infty$, and x_n converges to $-\infty$ if and only if $\limsup_{n\to\infty} x_n = -\infty$.

Exercise 2.5.6 Let X be an uncountable set, and let $\mathcal{F}$ be the collection of all subsets E in X which are either empty or cofinite (which means that $X\backslash E$ is finite). Show that $(X, \mathcal{F})$ is a topology (this is called the *cofinite topology* on X) which is not Hausdorff in the sense of Exercise 2.5.4 and is compact and connected. Also, show that if $x \in X$ $(V_n)_{n=1}^{\infty}$ is any countable collection of open sets containing x, then $\bigcap_{n=1}^{\infty} V_n \neq \{x\}$. Use this to show that the cofinite topology cannot be obtained by placing a metric d on X. (*Hint:* what is the set $\bigcap_{n=1}^{\infty} B(x, 1/n)$ equal to in a metric space?)

Exercise 2.5.7 Let X be an uncountable set, and let $\mathcal{F}$ be the collection of all subsets E in X which are either empty or cocountable (which means that $X\backslash E$ is at most countable). Show that $(X, \mathcal{F})$ is a topology (this is called the *cocountable topology* on X) which is not Hausdorff in the sense of Exercise 2.5.4, and connected, but cannot arise from a metric space and is not compact.

Exercise 2.5.8 Let $(X, \mathcal{F})$ be a compact topological space. Assume that this space is *first countable*, which means that for every $x \in X$ there exists a countable collection $V_1, V_2, \ldots$ of neighborhoods of x, such that every neighborhood of x contains one of the V_n. Show that every sequence in X has a convergent subsequence, by modifying Exercise 1.5.11.

Exercise 2.5.9 Prove the following partial analogue of Proposition 1.2.10 for topological spaces: (c) implies both (a) and (b), which are equivalent to each other. Show that in the cocountable topology in Exercise 2.5.7, it is possible for (a) and (b) to hold without (c) holding.

Exercise 2.5.10 Let E be a subset of a topological space $(X, \mathcal{F})$. Show that E is open if and only if every element of E is an interior point, and show that E is closed if and only if E contains all of its adherent points. Prove analogues of Proposition 1.2.15(e)-(h) (some of these are automatic by definition). If we assume in addition that X is Hausdorff, prove an analogue of Proposition 1.2.15(d) also, but give an example to show that (d) can fail when X is not Hausdorff.

Exercise 2.5.11 Show that the pair $(Y, \mathcal{F}_Y)$ defined in Definition 2.5.7 is indeed a topological space.

Exercise 2.5.12 Generalize Corollary 1.5.9 to compact sets in a Hausdorff topological space.

Exercise 2.5.13 Generalize Theorem 1.5.10 to compact sets in a Hausdorff topological space.

Exercise 2.5.14 Let (X, d_X) and (Y, d_Y) be metric spaces (and hence a topological space). Show that the two notions continuity (both at a point, and on the whole domain) of a function $f: X \to Y$ in Definition 2.1.1 and Definition 2.5.8 coincide.

Exercise 2.5.15 Show that when Theorem 2.1.4 is extended to topological spaces, that (a) implies (b). (The converse is false, but constructing an example is difficult.) Show that when Theorem 2.1.5 is extended to topological spaces, that (a), (c), (d) are all equivalent to each other and imply (b). (Again, the converse implications are false, but difficult to prove.)

Exercise 2.5.16 Generalize both Theorem 2.3.1 and Proposition 2.3.2 to compact sets in a topological space.

Chapter 3
Uniform Convergence

In the previous two chapters we have seen what it means for a sequence $(x^{(n)})_{n=1}^{\infty}$ of points in a metric space (X, d_X) to converge to a limit x; it means that $\lim_{n\to\infty} d_X(x^{(n)}, x) = 0$, or equivalently that for every $\varepsilon > 0$ there exists an $N > 0$ such that $d_X(x^{(n)}, x) < \varepsilon$ for all $n > N$. (We have also generalized the notion of convergence to topological spaces $(X, \mathcal{F})$, but in this chapter we will focus on metric spaces.)

In this chapter, we consider what it means for a sequence of *functions* $(f^{(n)})_{n=1}^{\infty}$ from one metric space (X, d_X) to another (Y, d_Y) to converge. In other words, we have a sequence of functions $f^{(1)}, f^{(2)}, \ldots$, with each function $f^{(n)} : X \to Y$ being a function from X to Y, and we ask what it means for this sequence of functions to converge to some limiting function f.

It turns out that there are several different concepts of convergence of functions; here we describe the two most important ones, *pointwise convergence* and *uniform convergence*. (There are other types of convergence for functions, such as L^1 convergence, L^2 convergence, convergence in measure, almost everywhere convergence, and so forth, but these are beyond the scope of this text.) The two notions are related, but not identical; the relationship between the two is somewhat analogous to the relationship between continuity and uniform continuity.

Once we work out what convergence means for functions, and thus can make sense of such statements as $\lim_{n\to\infty} f^{(n)} = f$, we will then ask how these limits interact with other concepts. For instance, we already have a notion of limiting values of functions: $\lim_{x\to x_0; x\in X} f(x)$. Can we interchange limits, i.e.,

$$\lim_{n\to\infty} \lim_{x\to x_0; x\in X} f^{(n)}(x) = \lim_{x\to x_0; x\in X} \lim_{n\to\infty} f^{(n)}(x)?$$

As we shall see, the answer depends on what type of convergence we have for $f^{(n)}$. We will also address similar questions involving interchanging limits and integrals, or limits and sums, or sums and integrals.

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_3

3.1 Limiting Values of Functions

Before we talk about limits of sequences of functions, we should first discuss a similar, but distinct, notion, that of limiting values of functions. We shall focus on the situation for metric spaces, but there are similar notions for topological spaces (Exercise 3.1.3).

Definition 3.1.1 (*Limiting value of a function*) Let (X, d_X) and (Y, d_Y) be metric spaces, let E be a subset of X, and let $f: E \to Y$ be a function. If $x_0 \in X$ is an adherent point of E, and $L \in Y$, we say that $f(x)$ *converges to* L *in* Y *as* x *converges to* x_0 *in* E, or write $\lim_{x \to x_0; x \in E} f(x) = L$, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $d_Y(f(x), L) < \varepsilon$ for all $x \in E$ such that $d_X(x, x_0) < \delta$.

Remark 3.1.2 Some authors exclude the case $x = x_0$ from the above definition, thus requiring $0 < d_X(x, x_0) < \delta$. In our current notation, this would correspond to removing x_0 from E, thus one would consider $\lim_{x \to x_0; x \in E \backslash \{x_0\}} f(x)$ instead of $\lim_{x \to x_0; x \in E} f(x)$. See Exercise 3.1.1 for a comparison of the two concepts.

Comparing this with Definition 2.1.1, we see that f is continuous at x_0 if and only if

$$\lim_{x \to x_0; x \in X} f(x) = f(x_0).$$

Thus f is continuous on X if we have

$$\lim_{x \to x_0; x \in X} f(x) = f(x_0) \text{ for all } x_0 \in X.$$

Example 3.1.3 If $f: \mathbf{R} \to \mathbf{R}$ is the function $f(x) = x^2 - 4$, then

$$\lim_{x \to 1} f(x) = f(1) = 1 - 4 = -3$$

since f is continuous.

Remark 3.1.4 Often we shall omit the condition $x \in X$, and abbreviate $\lim_{x \to x_0; x \in X} f(x)$ as simply $\lim_{x \to x_0} f(x)$ when it is clear what space x will range in.

One can rephrase Definition 3.1.1 in terms of sequences:

Proposition 3.1.5 *Let (X, d_X) and (Y, d_Y) be metric spaces, let E be a subset of X, and let $f: E \to Y$ be a function. Let $x_0 \in X$ be an adherent point of E and $L \in Y$. Then the following four statements are logically equivalent:*

(a) $\lim_{x \to x_0; x \in E} f(x) = L$.
(b) *For every sequence $(x^{(n)})_{n=1}^{\infty}$ in E which converges to x_0 with respect to the metric d_X, the sequence $(f(x^{(n)}))_{n=1}^{\infty}$ converges to L with respect to the metric d_Y.*

(c) For every open set $V \subseteq Y$ which contains L, there exists an open set $U \subseteq X$ containing x_0 such that $f(U \cap E) \subseteq V$.

(d) If one defines the function $g\colon E \cup \{x_0\} \to Y$ by defining $g(x_0) := L$, and $g(x) := f(x)$ for $x \in E \backslash \{x_0\}$, then g is continuous at x_0. Furthermore, if $x_0 \in E$, then $f(x_0) = L$.

Proof See Exercise 3.1.2. □

Remark 3.1.6 Observe from Propositions 3.1.5(b) and 1.1.20 that a function $f(x)$ can converge to at most one limit L as x converges to x_0. In other words, if the limit

$$\lim_{x \to x_0; x \in E} f(x)$$

exists at all, then it can only take at most one value.

Remark 3.1.7 The requirement that x_0 be an adherent point of E is necessary for the concept of limiting value to be useful, otherwise x_0 will lie in the exterior of E, the notion that $f(x)$ converges to L as x converges to x_0 in E is vacuous (for δ sufficiently small, there are no points $x \in E$ so that $d(x, x_0) < \delta$).

Remark 3.1.8 Strictly speaking, we should write

$$d_Y - \lim_{x \to x_0; x \in E} f(x) \text{ instead of } \lim_{x \to x_0; x \in E} f(x),$$

since the convergence depends on the metric d_Y. However in practice it will be obvious what the metric d_Y is and so we will omit the d_Y– prefix from the notation.

— Exercises —

Exercise 3.1.1 Let (X, d_X) and (Y, d_Y) be metric spaces, let E be a subset of X, let $f\colon E \to Y$ be a function, and let x_0 be an element of E. Assume that x_0 is an adherent point of $E \backslash \{x_0\}$ (or equivalently, that x_0 is not an *isolated point* of E). Show that the limit $\lim_{x \to x_0; x \in E} f(x)$ exists if and only if the limit $\lim_{x \to x_0; x \in E \backslash \{x_0\}} f(x)$ exists and is equal to $f(x_0)$. Also, show that if the limit $\lim_{x \to x_0; x \in E} f(x)$ exists at all, then it must equal $f(x_0)$.

Exercise 3.1.2 Prove Proposition 3.1.5. (*Hint:* review your proof of Theorem 2.1.4.)

Exercise 3.1.3 Use Proposition 3.1.5(c) to define a notion of a limiting value of a function $f\colon E \to Y$ from one topological space $(X, \mathcal{F}_X)$ to another $(Y, \mathcal{F}_Y)$, with E a subset of X. If X is a topological space and Y is a Hausdorff topological space (see Exercise 2.5.4), prove the equivalence of Proposition 3.1.5(c) and (d), as well as an analogue of Remark 3.1.6. What happens to these statements if Y is not assumed to be Hausdorff?

Exercise 3.1.4 Recall from Exercise 2.5.5 that the extended real line $\mathbf{R}^*$ comes with a standard topology (the order topology). We view the natural numbers $\mathbf{N}$ as a subspace of this topological space, and $+\infty$ as an adherent point of $\mathbf{N}$ in $\mathbf{R}^*$. Let $(a_n)_{n=0}^{\infty}$ be a sequence taking values in a topological space $(Y, \mathcal{F}_Y)$, and let $L \in Y$. Show that $\lim_{n \to +\infty; n \in \mathbf{N}} a_n = L$ (in the sense of Exercise 3.1.3) if and only if $\lim_{n \to \infty} a_n = L$ (in the sense of Definition 2.5.4). This shows that the notions of limiting values of a sequence, and limiting values of a function, are compatible.

Exercise 3.1.5 Let (X, d_X), (Y, d_Y), (Z, d_Z) be metric spaces, let E be a subset of X, and let $x_0 \in X$, $y_0 \in Y$, $z_0 \in Z$. Let $f \colon E \to Y$ and $g \colon Y \to Z$ be functions, and let E be a set. If we have $\lim_{x \to x_0; x \in E} f(x) = y_0$ and $\lim_{y \to y_0; y \in f(E)} g(y) = z_0$, conclude that $\lim_{x \to x_0; x \in E} g \circ f(x) = z_0$.

Exercise 3.1.6 State and prove an analogue of the limit laws in Proposition 9.3.14 when X is now a metric space rather than a subset of $\mathbf{R}$. (*Hint:* use Corollary 2.2.3.)

3.2 Pointwise and Uniform Convergence

The most obvious notion of convergence of functions is *pointwise convergence*, or convergence at each point of the domain:

Definition 3.2.1 (*Pointwise convergence*) Let $(f^{(n)})_{n=1}^{\infty}$ be a sequence of functions from one metric space (X, d_X) to another (Y, d_Y), and let $f \colon X \to Y$ be another function. We say that $(f^{(n)})_{n=1}^{\infty}$ *converges pointwise to f on X* if we have

$$\lim_{n \to \infty} f^{(n)}(x) = f(x)$$

for all $x \in X$, i.e.,

$$\lim_{n \to \infty} d_Y(f^{(n)}(x), f(x)) = 0.$$

Or in other words, for every x and every $\varepsilon > 0$ there exists $N > 0$ such that $d_Y(f^{(n)}(x), f(x)) < \varepsilon$ for every $n > N$. We call the function f the *pointwise limit* of the functions $f^{(n)}$.

Remark 3.2.2 Note that $f^{(n)}(x)$ and $f(x)$ are points in Y, rather than functions, so we are using our prior notion of convergence in metric spaces to determine convergence of functions. Also note that we are not really using the fact that (X, d_X) is a metric space (i.e., we are not using the metric d_X); for this definition it would suffice for X to just be a plain old set with no metric structure. However, later on we shall want to restrict our attention to *continuous* functions from X to Y, and in order to do so we need a metric on X (and on Y), or at least a topological structure. Also when we introduce the concept of *uniform convergence*, then we will definitely need a metric structure on X and Y; there is no comparable notion for topological spaces.

Example 3.2.3 Consider the functions $f^{(n)}\colon \mathbf{R} \to \mathbf{R}$ defined by $f^{(n)}(x) := x/n$, while $f\colon \mathbf{R} \to \mathbf{R}$ is the zero function $f(x) := 0$. Then $f^{(n)}$ converges pointwise to f, since for each fixed real number x we have $\lim_{n\to\infty} f^{(n)}(x) = \lim_{n\to\infty} x/n = 0 = f(x)$.

From Proposition 1.1.20 we see that a sequence $(f^{(n)})_{n=1}^{\infty}$ of functions from one metric space (X, d_X) to another (Y, d_Y) can have at most one pointwise limit f (this explains why we can refer to f as *the* pointwise limit). However, it is of course possible for a sequence of functions to have no pointwise limit (can you think of an example?), just as a sequence of points in a metric space do not necessarily have a limit.

Pointwise convergence is a very natural concept, but it has a number of disadvantages: it does not preserve continuity, derivatives, limits, or integrals, as the following three examples show.

Example 3.2.4 Consider the functions $f^{(n)}\colon [0, 1] \to \mathbf{R}$ defined by $f^{(n)}(x) := x^n$, and let $f\colon [0, 1] \to \mathbf{R}$ be the function defined by setting $f(x) := 1$ when $x = 1$ and $f(x) := 0$ when $0 \le x < 1$. Then the functions $f^{(n)}$ are continuous, and converge pointwise to f on $[0, 1]$ (why? Treat the cases $x = 1$ and $0 \le x < 1$ separately), however the limiting function f is not continuous. Note that the same example shows that pointwise convergence does not preserve differentiability either.

Example 3.2.5 If $\lim_{x\to x_0; x\in E} f^{(n)}(x) = L$ for every n, and $f^{(n)}$ converges pointwise to f, we cannot always take limits conclude that $\lim_{x\to x_0; x\in E} f(x) = L$. The previous example is also a counterexample here: observe that $\lim_{x\to 1; x\in[0,1)} x^n = 1$ for every n, but x^n converges pointwise to the function f defined in the previous paragraph, and $\lim_{x\to 1; x\in[0,1)} f(x) = 0$. In particular, we see that

$$\lim_{n\to\infty} \lim_{x\to x_0; x\in X} f^{(n)}(x) \neq \lim_{x\to x_0; x\in X} \lim_{n\to\infty} f^{(n)}(x).$$

(cf. Example 1.2.8). Thus pointwise convergence does not preserve limits.

Example 3.2.6 Suppose that $f^{(n)}\colon [a, b] \to \mathbf{R}$ a sequence of Riemann-integrable functions on the interval $[a, b]$. If $\int_{[a,b]} f^{(n)} = L$ for every n, and $f^{(n)}$ converges pointwise to some new function f, this does not mean that $\int_{[a,b]} f = L$. An example comes by setting $[a, b] := [0, 1]$, and letting $f^{(n)}$ be the function $f^{(n)}(x) := 2n$ when $x \in [1/2n, 1/n]$, and $f^{(n)}(x) := 0$ for all other values of x. Then $f^{(n)}$ converges pointwise to the zero function $f(x) := 0$ (why?). On the other hand, $\int_{[0,1]} f^{(n)} = 1$ for every n, while $\int_{[0,1]} f = 0$. In particular, we have an example where

$$\lim_{n\to\infty} \int_{[a,b]} f^{(n)} \neq \int_{[a,b]} \lim_{n\to\infty} f^{(n)}.$$

One may think that this counterexample has something to do with the $f^{(n)}$ being discontinuous, but one can easily modify this counterexample to make the $f^{(n)}$ continuous (can you see how?).

Another example in the same spirit is the “moving bump” example. Let $f^{(n)}: \mathbf{R} \to \mathbf{R}$ be the function defined by $f^{(n)}(x) := 1$ if $x \in [n, n+1]$ and $f^{(n)}(x) := 0$ otherwise. Then $\int_{\mathbf{R}} f^{(n)} = 1$ for every n (where $\int_{\mathbf{R}} f$ is defined as the limit of $\int_{[-N,N]} f$ as N goes to infinity). On the other hand, $f^{(n)}$ converges pointwise to the zero function 0 (why?), and $\int_{\mathbf{R}} 0 = 0$. In both of these examples, functions of area 1 have somehow “disappeared” to produce functions of area 0 in the limit. See also Example 1.2.9.

These examples show that pointwise convergence is too weak a concept to be of much use. The problem is that while $f^{(n)}(x)$ converges to $f(x)$ for each x, the *rate* of that convergence varies substantially with x. For instance, consider the first example where $f^{(n)}: [0, 1] \to \mathbf{R}$ was the function $f^{(n)}(x) := x^n$, and $f: [0, 1] \to \mathbf{R}$ was the function such that $f(x) := 1$ when $x = 1$, and $f(x) := 0$ otherwise. Then for each x, $f^{(n)}(x)$ converges to $f(x)$ as $n \to \infty$; this is the same as saying that $\lim_{n\to\infty} x^n = 0$ when $0 \leq x < 1$, and that $\lim_{n\to\infty} x^n = 1$ when $x = 1$. But the convergence is much slower near 1 than far away from 1. For instance, consider the statement that $\lim_{n\to\infty} x^n = 0$ for all $0 \leq x < 1$. This means, for every $0 \leq x < 1$, that for every ε, there exists an $N \geq 1$ such that $|x^n| < \varepsilon$ for all $n \geq N$—or in other words, the sequence $1, x, x^2, x^3, \ldots$ will eventually get less than ε, after passing some finite number N of elements in this sequence. But the number of elements N one needs to go out to depends very much on the location of x. For instance, take $\varepsilon := 0.1$. If $x = 0.1$, then we have $|x^n| < \varepsilon$ for all $n \geq 2$—the sequence gets underneath ε after the second element. But if $x = 0.5$, then we only get $|x^n| < \varepsilon$ for $n \geq 4$—you have to wait until the fourth element to get within ε of the limit. And if $x = 0.9$, then one only has $|x^n| < \varepsilon$ when $n \geq 22$. Clearly, the closer x gets to 1, the longer one has to wait until $f^{(n)}(x)$ will get within ε of $f(x)$, although it still will get there eventually. (Curiously, however, while the convergence gets worse and worse as x approaches 1, the convergence suddenly becomes perfect when $x = 1$.)

To put things another way, the convergence of $f^{(n)}$ to f is not *uniform* in x—the N that one needs to get $f^{(n)}(x)$ within ε of f depends on x as well as on ε. This motivates a stronger notion of convergence.

Definition 3.2.7 (*Uniform convergence*) Let $(f^{(n)})_{n=1}^{\infty}$ be a sequence of functions from one metric space (X, d_X) to another (Y, d_Y), and let $f: X \to Y$ be another function. We say that $(f^{(n)})_{n=1}^{\infty}$ *converges uniformly to f on X* if for every $\varepsilon > 0$ there exists $N > 0$ such that $d_Y(f^{(n)}(x), f(x)) < \varepsilon$ for every $n > N$ and $x \in X$. We call the function f the *uniform limit* of the functions $f^{(n)}$.

Remark 3.2.8 Note that this definition is subtly different from the definition for pointwise convergence in Definition 3.2.1. In the definition of pointwise convergence, N was allowed to depend on x; now it is not. The reader should compare this distinction to the distinction between continuity and uniform continuity (i.e., between Definitions 2.1.1 and 2.3.4). A more precise formulation of this analogy is given in Exercise 3.2.1.

It is easy to see that if $f^{(n)}$ converges uniformly to f on X, then it also converges pointwise to the same function f (see Exercise 3.2.2); thus when the uniform limit and pointwise limit both exist, then they have to be equal. However, the converse is not true; for instance the functions $f^{(n)}\colon [0,1] \to \mathbf{R}$ defined earlier by $f^{(n)}(x) := x^n$ converge pointwise, but do not converge uniformly (see Exercise 3.2.2).

Example 3.2.9 Let $f^{(n)}\colon [0,1] \to \mathbf{R}$ be the functions $f^{(n)}(x) := x/n$, and let $f\colon [0,1] \to \mathbf{R}$ be the zero function $f(x) := 0$. Then it is clear that $f^{(n)}$ converges to f pointwise. Now we show that in fact $f^{(n)}$ converges to f uniformly. We have to show that for every $\varepsilon > 0$, there exists an N such that $|f^{(n)}(x) - f(x)| < \varepsilon$ for every $x \in [0,1]$ and every $n \geq N$. To show this, let us fix an $\varepsilon > 0$. Then for any $x \in [0,1]$ and $n \geq N$, we have

$$|f^{(n)}(x) - f(x)| = |x/n - 0| = x/n \leq 1/n \leq 1/N.$$

Thus if we choose N such that $N > 1/\varepsilon$ (note that this choice of N does not depend on what x is), then we have $|f^{(n)}(x) - f(x)| < \varepsilon$ for all $n \geq N$ and $x \in [0,1]$, as desired.

We make one trivial remark here: if a sequence $f^{(n)}\colon X \to Y$ of functions converges pointwise (or uniformly) to a function $f\colon X \to Y$, then the restrictions $f^{(n)}|_E\colon E \to Y$ of $f^{(n)}$ to some subset E of X will also converge pointwise (or uniformly) to $f|_E$. (Why?)

— Exercises —

Exercise 3.2.1 The purpose of this exercise is to demonstrate a concrete relationship between continuity and pointwise convergence, and between uniform continuity and uniform convergence. Let $f\colon \mathbf{R} \to \mathbf{R}$ be a function. For any $a \in \mathbf{R}$, let $f_a : \mathbf{R} \to \mathbf{R}$ be the shifted function $f_a(x) := f(x - a)$.

(a) Show that f is continuous if and only if, whenever $(a_n)_{n=0}^\infty$ is a sequence of real numbers which converges to zero, the shifted functions f_{a_n} converge pointwise to f.

(b) Show that f is uniformly continuous if and only if, whenever $(a_n)_{n=0}^\infty$ is a sequence of real numbers which converges to zero, the shifted functions f_{a_n} converge uniformly to f.

Exercise 3.2.2 (a) Let $(f^{(n)})_{n=1}^\infty$ be a sequence of functions from one metric space (X, d_X) to another (Y, d_Y), and let $f\colon X \to Y$ be another function from X to Y. Show that if $f^{(n)}$ converges uniformly to f, then $f^{(n)}$ also converges pointwise to f.

(b) For each integer $n \geq 1$, let $f^{(n)}\colon (-1,1) \to \mathbf{R}$ be the function $f^{(n)}(x) := x^n$. Prove that $f^{(n)}$ converges pointwise to the zero function 0, but does not converge uniformly to any function $f\colon (-1,1) \to \mathbf{R}$.

(c) Let $g\colon (-1,1) \to \mathbf{R}$ be the function $g(x) := x/(1-x)$. With the notation as in (b), show that the partial sums $\sum_{n=1}^N f^{(n)}$ converge pointwise as $N \to \infty$ to g, but does not converge uniformly to g, on the open interval $(-1,1)$. (*Hint:* use Lemma 7.3.3.) What would happen if we replaced the open interval $(-1,1)$ with the closed interval $[-1,1]$?

Exercise 3.2.3 Let (X, d_X) a metric space, and for every integer $n \geq 1$, let $f_n \colon X \to \mathbf{R}$ be a real-valued function. Suppose that f_n converges pointwise to another function $f \colon X \to \mathbf{R}$ on X (in this question we give $\mathbf{R}$ the standard metric $d(x, y) = |x - y|$). Let $h \colon \mathbf{R} \to \mathbf{R}$ be a continuous function. Show that the functions $h \circ f_n$ converge pointwise to $h \circ f$ on X, where $h \circ f_n \colon X \to \mathbf{R}$ is the function $h \circ f_n(x) := h(f_n(x))$, and similarly for $h \circ f$.

Exercise 3.2.4 Let $f_n \colon X \to Y$ be a sequence of bounded functions from one metric space (X, d_X) to another metric space (Y, d_Y). Suppose that f_n converges uniformly to another function $f \colon X \to Y$. Suppose that f is a bounded function; i.e., there exists a ball $B_{(Y,d_Y)}(y_0, R)$ in Y such that $f(x) \in B_{(Y,d_Y)}(y_0, R)$ for all $x \in X$. Show that the sequence f_n is *uniformly bounded*; i.e., there exists a ball $B_{(Y,d_Y)}(y_0, R)$ in Y such that $f_n(x) \in B_{(Y,d_Y)}(y_0, R)$ for all $x \in X$ and all positive integers n.

3.3 Uniform Convergence and Continuity

We now give the first demonstration that uniform convergence is significantly better than pointwise convergence. Specifically, we show that the uniform limit of continuous functions is continuous.

Theorem 3.3.1 (Uniform limits preserve continuity I) *Suppose $(f^{(n)})_{n=1}^{\infty}$ is a sequence of functions from one metric space (X, d_X) to another (Y, d_Y), and suppose that this sequence converges uniformly to another function $f \colon X \to Y$. Let x_0 be a point in X. If the functions $f^{(n)}$ are continuous at x_0 for each n, then the limiting function f is also continuous at x_0.*

Proof See Exercise 3.3.1. □

This has an immediate corollary:

Corollary 3.3.2 (Uniform limits preserve continuity II) *Let $(f^{(n)})_{n=1}^{\infty}$ be a sequence of functions from one metric space (X, d_X) to another (Y, d_Y), and suppose that this sequence converges uniformly to another function $f \colon X \to Y$. If the functions $f^{(n)}$ are continuous on X for each n, then the limiting function f is also continuous on X.*

This should be contrasted with Example 3.2.4. There is a slight variant of Theorem 3.3.1 which is also useful:

Proposition 3.3.3 (Interchange of limits and uniform limits) *Let (X, d_X) and (Y, d_Y) be metric spaces, with Y complete, and let E be a subset of X. Let $(f^{(n)})_{n=1}^{\infty}$ be a sequence of functions from E to Y, and suppose that this sequence converges uniformly in E to some function $f \colon E \to Y$. Let $x_0 \in X$ be an adherent point of E, and suppose that for each n the limit* $\lim_{x \to x_0; x \in E} f^{(n)}(x)$ *exists. Then* *the limit* $\lim_{x\to x_0;x\in E} f(x)$ *also exists, and is equal to the limit of the sequence* $(\lim_{x\to x_0;x\in E} f^{(n)}(x))_{n=1}^{\infty}$*; in other words we have the interchange of limits*

$$\lim_{n\to\infty} \lim_{x\to x_0;x\in E} f^{(n)}(x) = \lim_{x\to x_0;x\in E} \lim_{n\to\infty} f^{(n)}(x).$$

Proof See Exercise 3.3.2. □

This should be contrasted with Example 3.2.5. Finally, we have a version of these theorems for sequences:

Proposition 3.3.4 *Let* $(f^{(n)})_{n=1}^{\infty}$ *be a sequence of continuous functions from one metric space* (X, d_X) *to another* (Y, d_Y)*, and suppose that this sequence converges uniformly to another function* $f : X \to Y$*. Let* $x^{(n)}$ *be a sequence of points in* X *which converge to some limit* x*. Then* $f^{(n)}(x^{(n)})$ *converges (in* Y*) to* $f(x)$*.*

Proof See Exercise 3.3.4. □

A similar result holds for bounded functions:

Definition 3.3.5 (*Bounded functions*) A function $f : X \to Y$ from one metric space (X, d_X) to another (Y, d_Y) is *bounded* if $f(X)$ is a bounded set, i.e., there exists a ball $B_{(Y,d_Y)}(y_0, R)$ in Y such that $f(x) \in B_{(Y,d_Y)}(y_0, R)$ for all $x \in X$.

Proposition 3.3.6 (Uniform limits preserve boundedness) *Let* $(f^{(n)})_{n=1}^{\infty}$ *be a sequence of functions from one metric space* (X, d_X) *to another* (Y, d_Y)*, and suppose that this sequence converges uniformly to another function* $f : X \to Y$*. If the functions* $f^{(n)}$ *are bounded on* X *for each* n*, then the limiting function* f *is also bounded on* X*.*

Proof See Exercise 3.3.6. □

Remark 3.3.7 The above propositions sound very reasonable, but one should caution that it only works if one assumes uniform convergence; pointwise convergence is not enough. (See Exercises 3.3.3, 3.3.5 and 3.3.7.)

— Exercises —

Exercise 3.3.1 Prove Theorem 3.3.1. Explain briefly why your proof requires uniform convergence, and why pointwise convergence would not suffice. (*Hints:* it is easiest to use the "epsilon-delta" definition of continuity from Definition 2.1.1. You may find the triangle inequality

$$\begin{aligned} d_Y(f(x), f(x_0)) \le{}& d_Y(f(x), f^{(n)}(x)) + d_Y(f^{(n)}(x), f^{(n)}(x_0)) \\ &+ d_Y(f^{(n)}(x_0), f(x_0)) \end{aligned}$$

useful. Also, you may need to divide ε as $\varepsilon = \varepsilon/3 + \varepsilon/3 + \varepsilon/3$. Finally, it is possible to prove Theorem 3.3.1 from Proposition 3.3.3, but you may find it easier conceptually to prove Theorem 3.3.1 first.)

Exercise 3.3.2 Prove Proposition 3.3.3. (*Hint:* this is very similar to Theorem 3.3.1. Theorem 3.3.1 cannot be used to prove Proposition 3.3.3, however it is possible to use Proposition 3.3.3 to prove Theorem 3.3.1.)

Exercise 3.3.3 Compare Proposition 3.3.3 with Example 1.2.8. Can you now explain why the interchange of limits in Example 1.2.8 led to a false statement, whereas the interchange of limits in Proposition 3.3.3 is justified?

Exercise 3.3.4 Prove Proposition 3.3.4. (*Hint:* again, this is similar to Theorem 3.3.1 and Proposition 3.3.3, although the statements are slightly different, and one cannot deduce this directly from the other two results.)

Exercise 3.3.5 Give an example to show that Proposition 3.3.4 fails if the phrase "converges uniformly" is replaced by "converges pointwise". (*Hint:* some of the examples already given earlier will already work here.)

Exercise 3.3.6 Prove Proposition 3.3.6. Discuss how this proposition differs from Exercise 3.2.4.

Exercise 3.3.7 Give an example to show that Proposition 3.3.6 fails if the phrase "converges uniformly" is replaced by "converges pointwise". (*Hint:* some of the examples already given earlier will already work here.)

Exercise 3.3.8 Let (X, d) be a metric space, and for every positive integer n, let $f_n : X \to \mathbf{R}$ and $g_n : X \to \mathbf{R}$ be functions. Suppose that $(f_n)_{n=1}^{\infty}$ converges uniformly to another function $f : X \to \mathbf{R}$, and that $(g_n)_{n=1}^{\infty}$ converges uniformly to another function $g : X \to \mathbf{R}$. Suppose also that the functions $(f_n)_{n=1}^{\infty}$ and $(g_n)_{n=1}^{\infty}$ are uniformly bounded, i.e., there exists an $M > 0$ such that $|f_n(x)| \leq M$ and $|g_n(x)| \leq M$ for all $n \geq 1$ and $x \in X$. Prove that the functions $f_n g_n : X \to \mathbf{R}$ converge uniformly to $fg : X \to \mathbf{R}$.

3.4 The Metric of Uniform Convergence

We have now developed at least four, apparently separate, notions of limit in this text:

(a) limits $\lim_{n\to\infty} x^{(n)}$ of sequences of points in a metric space (Definition 1.1.14; see also Definition 2.5.4);
(b) limiting values $\lim_{x\to x_0; x\in E} f(x)$ of functions at a point (Definition 3.1.1);
(c) pointwise limits f of functions $f^{(n)}$ (Definition 3.2.1); and
(d) uniform limits f of functions $f^{(n)}$ (Definition 3.2.7).

This proliferation of limits may seem rather complicated. However, we can reduce the complexity slightly by observing that (d) can be viewed as a special case of (a), though in doing so it should be cautioned that because we are now dealing with functions instead of points, the convergence is not in X or in Y, but rather in a new space, the space of functions from X to Y.

Remark 3.4.1 If one is willing to work in topological spaces instead of metric spaces, we can also view (a) as a special case of (b), see Exercise 3.1.4, and (c) is also a special case of (a), see Exercise 3.4.4. Thus the notion of convergence in a topological space can be used to unify all the notions of limits we have encountered so far.

Definition 3.4.2 (*Metric space of bounded functions*) Suppose (X, d_X) and (Y, d_Y) are metric spaces. We let $B(X \to Y)$ denote the space[1] of bounded functions from X to Y:

$$B(X \to Y) := \{f \big| f : X \to Y \text{ is a bounded function}\}.$$

If X is non-empty, we define a metric $d_\infty : B(X \to Y) \times B(X \to Y) \to [0, +\infty)$ by defining

$$d_\infty(f, g) := \sup_{x \in X} d_Y(f(x), g(x)) = \sup\{d_Y(f(x), g(x)) : x \in X\}$$

for all $f, g \in B(X \to Y)$. This metric is sometimes known as the *uniform metric*, *sup norm metric* or the L^∞ *metric*. We will also use $d_{B(X\to Y)}$ as a synonym for d_∞. If X is empty, we instead define $d_\infty(f, g) = 0$.

Notice that the distance $d_\infty(f, g)$ is always finite because f and g are assumed to be bounded on X.

Example 3.4.3 Let $X := [0, 1]$ and $Y = \mathbf{R}$. Let $f : [0, 1] \to \mathbf{R}$ and $g : [0, 1] \to \mathbf{R}$ be the functions $f(x) := 2x$ and $g(x) := 3x$. Then f and g are both bounded functions and thus live in $B([0, 1] \to \mathbf{R})$. The distance between them is

$$d_\infty(f, g) = \sup_{x \in [0,1]} |2x - 3x| = \sup_{x \in [0,1]} |x| = 1.$$

This space turns out to be a metric space (Exercise 3.4.1). Convergence in this metric turns out to be identical to uniform convergence:

Proposition 3.4.4 *Let (X, d_X) and (Y, d_Y) be metric spaces. Let $(f^{(n)})_{n=1}^\infty$ be a sequence of functions in $B(X \to Y)$, and let f be another function in $B(X \to Y)$. Then $(f^{(n)})_{n=1}^\infty$ converges to f in the metric $d_{B(X\to Y)}$ if and only if $(f^{(n)})_{n=1}^\infty$ converges uniformly to f.*

Proof See Exercise 3.4.2. □

Now let $C(X \to Y)$ be the space of bounded continuous functions from X to Y:

$$C(X \to Y) := \{f \in B(X \to Y) \big| f \text{ is continuous}\}.$$

[1] Note that this is a set, thanks to the power set axiom (Axiom 3.11) and the axiom of specification (Axiom 3.6).

This set $C(X \to Y)$ is clearly a subset of $B(X \to Y)$. Corollary 3.3.2 asserts that this space $C(X \to Y)$ is closed in $B(X \to Y)$ (why?). Actually, we can say a lot more:

Theorem 3.4.5 (The space of continuous functions is complete) *Let (X, d_X) be a metric space, and let (Y, d_Y) be a* complete *metric space. The space $(C(X \to Y), d_{B(X\to Y)}|_{C(X\to Y)\times C(X\to Y)})$ is a complete subspace of $(B(X \to Y), d_{B(X\to Y)})$. In other words, every Cauchy sequence of functions in $C(X \to Y)$ converges to a function in $C(X \to Y)$.*

Proof See Exercise 3.4.3. □

— Exercises —

Exercise 3.4.1 Let (X, d_X) and (Y, d_Y) be metric spaces. Show that the space $B(X \to Y)$ defined in Definition 3.4.2, with the metric $d_{B(X\to Y)}$, is indeed a metric space.

Exercise 3.4.2 Prove Proposition 3.4.4.

Exercise 3.4.3 Prove Theorem 3.4.5. (*Hint:* this is similar, but not identical, to the proof of Theorem 3.3.1).

Exercise 3.4.4 Let (X, d_X) and (Y, d_Y) be metric spaces, and let $Y^X := \{f \,\big|\, f : X \to Y\}$ be the space of all functions from X to Y (cf. Axiom 3.11). If $x_0 \in X$ and V is an open set in Y, let $V^{(x_0)} \subseteq Y^X$ be the set

$$V^{(x_0)} := \{f \in Y^X : f(x_0) \in V\}.$$

If E is a subset of Y^X, we say that E is *open* if for every $f \in E$, there exists a finite number of points $x_1, \ldots, x_n \in X$ and open sets $V_1, \ldots, V_n \subseteq Y$ such that

$$f \in V_1^{(x_1)} \cap \cdots \cap V_n^{(x_n)} \subseteq E.$$

(a) Show that if $\mathcal{F}$ is the collection of open sets in Y^X, then $(Y^X, \mathcal{F})$ is a topological space.
(b) For each natural number n, let $f^{(n)} : X \to Y$ be a function from X to Y, and let $f : X \to Y$ be another function from X to Y. Show that $f^{(n)}$ converges to f in the topology $\mathcal{F}$ (in the sense of Definition 2.5.4) if and only if $f^{(n)}$ converges to f pointwise (in the sense of Definition 3.2.1).

The topology $\mathcal{F}$ is known as the *topology of pointwise convergence*, for obvious reasons; it is also known as the *product topology*. It shows that the concept of pointwise convergence can be viewed as a special case of the more general concept of convergence in a topological space.

3.5 Series of Functions; the Weierstrass M-Test

Having discussed sequences of functions, we now discuss infinite series $\sum_{n=1}^{\infty} f_n$ of functions. Now we shall restrict our attention to functions $f\colon X \to \mathbf{R}$ from a metric space (X, d_X) to the real line $\mathbf{R}$ (which we of course give the standard metric); this is because we know how to add two real numbers, but don't necessarily know how to add two points in a general metric space Y. Functions whose codomain is $\mathbf{R}$ are sometimes called *real-valued* functions.

Finite summation is, of course, easy: given any finite collection $f^{(1)}, \dots, f^{(N)}$ of functions from X to $\mathbf{R}$, we can define the finite sum $\sum_{i=1}^{N} f^{(i)}\colon X \to \mathbf{R}$ by

$$\left(\sum_{i=1}^{N} f^{(i)}\right)(x) := \sum_{i=1}^{N} f^{(i)}(x).$$

Example 3.5.1 If $f^{(1)}\colon \mathbf{R} \to \mathbf{R}$ is the function $f^{(1)}(x) := x$, $f^{(2)}\colon \mathbf{R} \to \mathbf{R}$ is the function $f^{(2)}(x) := x^2$, and $f^{(3)}\colon \mathbf{R} \to \mathbf{R}$ is the function $f^{(3)}(x) := x^3$, then $f := \sum_{i=1}^{3} f^{(i)}$ is the function $f\colon \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x + x^2 + x^3$.

It is easy to show that finite sums of bounded functions are bounded, and finite sums of continuous functions are continuous (Exercise 3.5.1).

Now to add infinite series.

Definition 3.5.2 (*Infinite series*) Let (X, d_X) be a metric space. Let $(f^{(n)})_{n=1}^{\infty}$ be a sequence of functions from X to $\mathbf{R}$, and let f be another function from X to $\mathbf{R}$. If the partial sums $\sum_{n=1}^{N} f^{(n)}$ converge pointwise to f on X as $N \to \infty$, we say that the infinite series $\sum_{n=1}^{\infty} f^{(n)}$ *converges pointwise* to f, and write $f = \sum_{n=1}^{\infty} f^{(n)}$. If the partial sums $\sum_{n=1}^{N} f^{(n)}$ converge uniformly to f on X as $N \to \infty$, we say that the infinite series $\sum_{n=1}^{\infty} f^{(n)}$ *converges uniformly* to f, and again write $f = \sum_{n=1}^{\infty} f^{(n)}$. (Thus when one sees an expression such as $f = \sum_{n=1}^{\infty} f^{(n)}$, one should look at the context to see in what sense this infinite series converges.)

Remark 3.5.3 A series $\sum_{n=1}^{\infty} f^{(n)}$ converges pointwise to f on X if and only if $\sum_{n=1}^{\infty} f^{(n)}(x)$ converges to $f(x)$ for *every* $x \in X$. (Thus if $\sum_{n=1}^{\infty} f^{(n)}$ does not converge pointwise to f, this does not mean that it diverges pointwise; it may just be that it converges for some points x but diverges at other points x.)

If a series $\sum_{n=1}^{\infty} f^{(n)}$ converges uniformly to f, then it also converges pointwise to f; but not vice versa, as the following example shows.

Example 3.5.4 Let $f^{(n)}\colon (-1, 1) \to \mathbf{R}$ be the sequence of functions $f^{(n)}(x) := x^n$. Then $\sum_{n=1}^{\infty} f^{(n)}$ converges pointwise, but not uniformly, to the function $x/(1-x)$ (see Exercise 3.2.2 and Example 3.5.8).

It is not always clear when a series $\sum_{n=1}^{\infty} f^{(n)}$ converges or not. However, there is a very useful test that gives at least one test for uniform convergence.

Definition 3.5.5 (*Sup norm*) If $f: X \to \mathbf{R}$ is a bounded real-valued function, and X is non-empty, we define the *sup norm* $\|f\|_\infty$ of f to be the number

$$\|f\|_\infty := \sup\{|f(x)| : x \in X\}.$$

In other words, $\|f\|_\infty = d_\infty(f, 0)$, where $0 : X \to \mathbf{R}$ is the zero function $0(x) := 0$, and d_∞ was defined in Definition 3.4.2. (Why is this the case?) If X is empty, we instead define $\|f\|_\infty := 0$.

Example 3.5.6 Thus, for instance, if $f: (-2, 1) \to \mathbf{R}$ is the function $f(x) := 2x$, then $\|f\|_\infty = \sup\{|2x| : x \in (-2, 1)\} = 4$ (why?). Notice that when f is bounded then $\|f\|_\infty$ will always be a non-negative real number.

Theorem 3.5.7 (Weierstrass M-test) *Let (X, d) be a metric space, and let $(f^{(n)})_{n=1}^\infty$ be a sequence of bounded real-valued continuous functions on X such that the series $\sum_{n=1}^\infty \|f^{(n)}\|_\infty$ is convergent. (Note that this is a series of plain old real numbers, not of functions.) Then the series $\sum_{n=1}^\infty f^{(n)}$ converges uniformly to some function f on X, and that function f is also continuous.*

Proof See Exercise 3.5.2. □

To put the Weierstrass M-test succinctly: absolute convergence of sup norms implies uniform convergence of functions.

Example 3.5.8 Let $0 < r < 1$ be a real number, and let $f^{(n)}: [-r, r] \to \mathbf{R}$ be the series of functions $f^{(n)}(x) := x^n$. Then each $f^{(n)}$ is continuous and bounded, and $\|f^{(n)}\|_\infty = r^n$ (why?). Since the series $\sum_{n=1}^\infty r^n$ is absolutely convergent (e.g., by the root test, Theorem 7.5.1 from *Analysis I*), we thus see that $\sum_{n=1}^\infty f^{(n)}$ converges uniformly in $[-r, r]$ to some continuous function; in Exercise 3.2.2(c) we see that this function must in fact be the function $f: [-r, r] \to \mathbf{R}$ defined by $f(x) := x/(1-x)$. In other words, the series $\sum_{n=1}^\infty x^n$ is pointwise convergent, but not uniformly convergent, on $(-1, 1)$, but is uniformly convergent on the smaller interval $[-r, r]$ for any $0 < r < 1$.

The Weierstrass M-test is especially useful in relation to *power series*, which we will encounter in the next chapter.

— Exercises —

Exercise 3.5.1 Let $f^{(1)}, \ldots, f^{(N)}$ be a finite sequence of bounded functions from a metric space (X, d_X) to $\mathbf{R}$. Show that $\sum_{i=1}^N f^{(i)}$ is also bounded. Prove a similar claim when "bounded" is replaced by "continuous". What if "continuous" was replaced by "uniformly continuous"?

Exercise 3.5.2 Prove Theorem 3.5.7. (*Hint:* first show that the sequence $\sum_{i=1}^N f^{(i)}$ is a Cauchy sequence in $C(X \to \mathbf{R})$. Then use Theorem 3.4.5.)

3.6 Uniform Convergence and Integration

We now connect uniform convergence with Riemann integration (which was discussed in Chap. 11), by showing that uniform limits can be safely interchanged with integrals.

Theorem 3.6.1 *Let $[a,b]$ be an interval, and for each integer $n \geq 1$, let $f^{(n)} \colon [a,b] \to \mathbf{R}$ be a Riemann-integrable function. Suppose $f^{(n)}$ converges uniformly on $[a,b]$ to a function $f \colon [a,b] \to \mathbf{R}$. Then f is also Riemann integrable, and*

$$\lim_{n\to\infty} \int_{[a,b]} f^{(n)} = \int_{[a,b]} f.$$

Proof We first show that f is Riemann integrable on $[a,b]$. This is the same as showing that the upper and lower Riemann integrals of f match: $\underline{\int}_{[a,b]} f = \overline{\int}_{[a,b]} f$.

Let $\varepsilon > 0$. Since $f^{(n)}$ converges uniformly to f, we see that there exists an $N > 0$ such that $|f^{(n)}(x) - f(x)| < \varepsilon$ for all $n > N$ and $x \in [a,b]$. In particular we have

$$f^{(n)}(x) - \varepsilon < f(x) < f^{(n)}(x) + \varepsilon$$

for all $x \in [a,b]$. Integrating this on $[a,b]$ we obtain

$$\underline{\int}_{[a,b]} (f^{(n)} - \varepsilon) \leq \underline{\int}_{[a,b]} f \leq \overline{\int}_{[a,b]} f \leq \overline{\int}_{[a,b]} (f^{(n)} + \varepsilon).$$

Since $f^{(n)}$ is assumed to be Riemann integrable, we thus see

$$\left(\int_{[a,b]} f^{(n)}\right) - \varepsilon(b-a) \leq \underline{\int}_{[a,b]} f \leq \overline{\int}_{[a,b]} f \leq \left(\int_{[a,b]} f^{(n)}\right) + \varepsilon(b-a).$$

In particular, we see that

$$0 \leq \overline{\int}_{[a,b]} f - \underline{\int}_{[a,b]} f \leq 2\varepsilon(b-a).$$

Since this is true for every $\varepsilon > 0$, we obtain $\underline{\int}_{[a,b]} f = \overline{\int}_{[a,b]} f$ as desired.

The above argument also shows that for every $\varepsilon > 0$ there exists an $N > 0$ such that

$$\left| \int_{[a,b]} f^{(n)} - \int_{[a,b]} f \right| \leq \varepsilon(b-a)$$

for all $n \geq N$. Since ε is arbitrary, we see that $\int_{[a,b]} f^{(n)}$ converges to $\int_{[a,b]} f$ as desired. □

To rephrase Theorem 3.6.1: we can rearrange limits and integrals (on compact intervals $[a, b]$),

$$\lim_{n\to\infty} \int_{[a,b]} f^{(n)} = \int_{[a,b]} \lim_{n\to\infty} f^{(n)},$$

provided that the convergence is uniform. This should be contrasted with Examples 1.2.9 and 3.2.5.

There is an analogue of this theorem for series:

Corollary 3.6.2 *Let* $[a, b]$ *be an interval, and let* $(f^{(n)})_{n=1}^{\infty}$ *be a sequence of Riemann-integrable functions on* $[a, b]$ *such that the series* $\sum_{n=1}^{\infty} f^{(n)}$ *is uniformly convergent. Then we have*

$$\sum_{n=1}^{\infty} \int_{[a,b]} f^{(n)} = \int_{[a,b]} \sum_{n=1}^{\infty} f^{(n)}.$$

Proof See Exercise 3.6.1. □

This corollary works particularly well in conjunction with the Weierstrass M-test (Theorem 3.5.7):

Example 3.6.3 (Informal) From Lemma 7.3.3 of *Analysis I* we have the geometric series identity

$$\sum_{n=1}^{\infty} x^n = \frac{x}{1-x}$$

for $x \in (-1, 1)$, and the convergence is uniform (by the Weierstrass M-test) on $[-r, r]$ for any $0 < r < 1$. By adding 1 to both sides we obtain

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

and the converge is again uniform. We can thus integrate on $[0, r]$ and use Corollary 3.6.2 to obtain

$$\sum_{n=0}^{\infty} \int_{[0,r]} x^n \, dx = \int_{[0,r]} \frac{1}{1-x} \, dx.$$

The left-hand side is $\sum_{n=0}^{\infty} r^{n+1}/(n+1)$. If we accept for now the use of logarithms (we will justify this use in Sect. 4.5), the anti-derivative of $1/(1-x)$ is $-\log(1-x)$, and so the right-hand side is $-\log(1-r)$. We thus obtain the formula

$$-\log(1-r) = \sum_{n=0}^{\infty} r^{n+1}/(n+1)$$

for all $0 < r < 1$.

— Exercises —

Exercise 3.6.1 Use Theorem 3.6.1 to prove Corollary 3.6.2.

3.7 Uniform Convergence and Derivatives

We have already seen how uniform convergence interacts well with continuity, with limits, and with integrals. Now we investigate how it interacts with derivatives.

The first question we can ask is: if f_n converges uniformly to f, and the functions f_n are differentiable, does this imply that f is also differentiable? And does f_n' also converge to f'?

The answer to the second question is, unfortunately, no. To see a counterexample, we will use without proof some basic facts about trigonometric functions (which we will make rigorous in Sect. 4.7). Consider the functions $f_n : [0, 2\pi] \to \mathbf{R}$ defined by $f_n(x) := n^{-1/2} \sin(nx)$, and let $f : [0, 2\pi] \to \mathbf{R}$ be the zero function $f(x) := 0$. Then, since sin takes values between -1 and 1, we have $d_\infty(f_n, f) \leq n^{-1/2}$, where we use the uniform metric $d_\infty(f, g) := \sup_{x\in[0,2\pi]} |f(x) - g(x)|$ introduced in Definition 3.4.2. Since $n^{-1/2}$ converges to 0, we thus see by the squeeze test that f_n converges uniformly to f. On the other hand, $f_n'(x) = n^{1/2} \cos(nx)$, and so in particular $|f_n'(0) - f'(0)| = n^{1/2}$. Thus f_n' does not converge pointwise to f', and so in particular does not converge uniformly either. In particular we have

$$\frac{d}{dx} \lim_{n\to\infty} f_n(x) \neq \lim_{n\to\infty} \frac{d}{dx} f_n(x).$$

The answer to the first question is also no. An example is the sequence of functions $f_n : [-1, 1] \to \mathbf{R}$ defined by $f_n(x) := \sqrt{\frac{1}{n^2} + x^2}$. These functions are differentiable (why?). Also, one can easily check that

$$|x| \leq f_n(x) \leq |x| + \frac{1}{n}$$

for all $x \in [-1, 1]$ (why? square both sides), and so by the squeeze test f_n converges uniformly to the absolute value function $f(x) := |x|$. But this function is not differentiable at 0 (why?). Thus, the uniform limit of differentiable functions need not be differentiable. (See also Example 1.2.10.)

So, in summary, uniform convergence of the functions f_n says nothing about the convergence of the derivatives f_n'. However, the converse is true, as long as f_n converges at at least one point:

Theorem 3.7.1 *Let $[a, b]$ be an interval, and for every integer $n \geq 1$, let $f_n : [a, b] \to \mathbf{R}$ be a differentiable function whose derivative $f_n' : [a, b] \to \mathbf{R}$ is continuous. Suppose that the derivatives f_n' converge uniformly to a function $g : [a, b] \to \mathbf{R}$. Suppose also that there exists a point x_0 such that the limit $\lim_{n\to\infty} f_n(x_0)$ exists. Then the functions f_n converge uniformly to a differentiable function f, and the derivative of f equals g.*

Informally, the above theorem says that if f_n' converges uniformly, and $f_n(x_0)$ converges for some x_0, then f_n also converges uniformly, and $\frac{d}{dx} \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{d}{dx} f_n(x)$.

Proof We only give the beginning of the proof here; the remainder of the proof will be an exercise (Exercise 3.7.1).

Since f_n' is continuous, we see from the fundamental theorem of calculus (Theorem 11.9.4) that

$$f_n(x) - f_n(x_0) = \int_{[x_0,x]} f_n'$$

when $x \in [x_0, b]$, and

$$f_n(x) - f_n(x_0) = -\int_{[x,x_0]} f_n'$$

when $x \in [a, x_0]$.

Let L be the limit of $f_n(x_0)$ as $n \to \infty$:

$$L := \lim_{n\to\infty} f_n(x_0).$$

By hypothesis, L exists. Now, since each f_n' is continuous on $[a, b]$, and f_n' converges uniformly to g, we see by Corollary 3.3.2 that g is also continuous. Now define the function $f : [a, b] \to \mathbf{R}$ by setting

$$f(x) := L - \int_{[a,x_0]} g + \int_{[a,x]} g$$

for all $x \in [a, b]$. To finish the proof, we have to show that f_n converges uniformly to f, and that f is differentiable with derivative g; this shall be done in Exercise 3.7.1. □

Remark 3.7.2 It turns out that Theorem 3.7.1 is still true when the functions f_n' are not assumed to be continuous, but the proof is more difficult; see Exercise 3.7.2.

By combining this theorem with the Weierstrass M-test, we obtain

Corollary 3.7.3 *Let $[a, b]$ be an interval, and for every integer $n \geq 1$, let $f_n : [a, b] \to \mathbf{R}$ be a differentiable function whose derivative $f_n' : [a, b] \to \mathbf{R}$ is continuous. Suppose that the series $\sum_{n=1}^{\infty} \|f_n'\|_\infty$ is absolutely convergent, where*

$$\|f_n'\|_\infty := \sup_{x \in [a,b]} |f_n'(x)|$$

is the sup norm of f_n', as defined in Definition 3.5.5. Suppose also that the series $\sum_{n=1}^{\infty} f_n(x_0)$ is convergent for some $x_0 \in [a, b]$. Then the series $\sum_{n=1}^{\infty} f_n$ converges uniformly on $[a, b]$ to a differentiable function, and in fact

$$\frac{d}{dx} \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \frac{d}{dx} f_n(x)$$

for all $x \in [a, b]$.

Proof See Exercise 3.7.3. □

We now pause to give an example of a function which is continuous everywhere, but differentiable nowhere (this particular example was discovered by Weierstrass). Again, we will presume knowledge of the trigonometric functions, which will be covered rigorously in Sect. 4.7.

Example 3.7.4 Let $f : \mathbf{R} \to \mathbf{R}$ be the function

$$f(x) := \sum_{n=1}^{\infty} 4^{-n} \cos(32^n \pi x).$$

Note that this series is uniformly convergent, thanks to the Weierstrass M-test, and since each individual function $4^{-n} \cos(32^n \pi x)$ is continuous, the function f is also continuous. However, it is not differentiable (Exercise 4.7.10); in fact it is a *nowhere differentiable function*, one which is not differentiable at *any* point, despite being continuous everywhere!

— Exercises —

Exercise 3.7.1 Complete the proof of Theorem 3.7.1. Compare this theorem with Example 1.2.10, and explain why this example does not contradict the theorem.

Exercise 3.7.2 Prove Theorem 3.7.1 without assuming that f_n' is continuous. (This means that you cannot use the fundamental theorem of calculus. However, the mean value theorem (Corollary 10.2.9) is still available. Use this to show that if $d_\infty(f_n', f_m') \leq \varepsilon$, then $|(f_n(x) - f_m(x)) - (f_n(x_0) - f_m(x_0))| \leq \varepsilon|x - x_0|$ for all $x \in [a, b]$, and then use this to complete the proof of Theorem 3.7.1.)

Exercise 3.7.3 Prove Corollary 3.7.3.

3.8 Uniform Approximation by Polynomials

As we have just seen, continuous functions can be very badly behaved, for instance they can be nowhere differentiable (Example 3.7.4). On the other hand, functions such as polynomials are always very well behaved, in particular being always differentiable. Fortunately, while most continuous functions are not as well behaved as polynomials, they can always be *uniformly approximated* by polynomials; this important (but difficult) result is known as the *Weierstrass approximation theorem*, and is the subject of this section.

Definition 3.8.1 Let $[a, b]$ be an interval. A *polynomial on* $[a, b]$ is a function $f: [a, b] \to \mathbf{R}$ of the form $f(x) := \sum_{j=0}^{n} c_j x^j$, where $n \geq 0$ is an integer and $c_0, \ldots, c_n$ are real numbers. If $c_n \neq 0$, then n is called the *degree* of f.

Example 3.8.2 The function $f: [1, 2] \to \mathbf{R}$ defined by $f(x) := 3x^4 + 2x^3 - 4x + 5$ is a polynomial on $[1, 2]$ of degree 4.

Theorem 3.8.3 (Weierstrass approximation theorem) *If $[a, b]$ is an interval, $f: [a, b] \to \mathbf{R}$ is a continuous function, and $\varepsilon > 0$, then there exists a polynomial P on $[a, b]$ such that $d_\infty(P, f) \leq \varepsilon$ (i.e., $|P(x) - f(x)| \leq \varepsilon$ for all $x \in [a, b]$).*

Another way of stating this theorem is as follows. Recall that $C([a, b] \to \mathbf{R})$ was the space of continuous functions from $[a, b]$ to $\mathbf{R}$, with the uniform metric d_∞. Let $P([a, b] \to \mathbf{R})$ be the space of all polynomials on $[a, b]$; this is a subspace of $C([a, b] \to \mathbf{R})$, since all polynomials are continuous (Exercise 9.4.7). The Weierstrass approximation theorem then asserts that every continuous function is an adherent point of $P([a, b] \to \mathbf{R})$; or in other words, that the closure of the space of polynomials is the space of continuous functions:

$$\overline{P([a, b] \to \mathbf{R})} = C([a, b] \to \mathbf{R}).$$

In particular, every continuous function on $[a, b]$ is the uniform limit of polynomials. Another way of saying this is that the space of polynomials is *dense* in the space of continuous functions, in the *uniform topology*.

The proof of the Weierstrass approximation theorem is somewhat complicated and will be done in stages. We first need the notion of an *approximation to the identity*.

Definition 3.8.4 (*Compactly supported functions*) Let $[a, b]$ be an interval. A function $f: \mathbf{R} \to \mathbf{R}$ is said to be *supported* on $[a, b]$ if $f(x) = 0$ for all $x \notin [a, b]$. We say that f is *compactly supported* if it is supported on some interval $[a, b]$. If f is continuous and supported on $[a, b]$, we define the improper integral $\int_{-\infty}^{\infty} f$ to be $\int_{-\infty}^{\infty} f := \int_{[a,b]} f$.

Note that a function can be supported on more than one interval, for instance a function which is supported on $[3, 4]$ is also automatically supported on $[2, 5]$ (why?). In principle, this might mean that our definition of $\int_{-\infty}^{\infty} f$ is not well defined, however this is not the case:

Lemma 3.8.5 *If* $f: \mathbf{R} \to \mathbf{R}$ *is continuous and supported on an interval* $[a, b]$*, and is also supported on another interval* $[c, d]$*, then* $\int_{[a,b]} f = \int_{[c,d]} f$.

Proof See Exercise 3.8.1. □

Definition 3.8.6 *Approximation to the identity*) Let $\varepsilon > 0$ and $0 < \delta < 1$. A function $f: \mathbf{R} \to \mathbf{R}$ is said to be an (ε, δ)*-approximation to the identity* if it obeys the following three properties:

(a) f is supported on $[-1, 1]$, and $f(x) \geq 0$ for all $-1 \leq x \leq 1$.
(b) f is continuous, and $\int_{-\infty}^{\infty} f = 1$.
(c) $|f(x)| \leq \varepsilon$ for all $\delta \leq |x| \leq 1$.

Remark 3.8.7 For those of you who are familiar with the Dirac delta function, approximations to the identity are ways to approximate this (very discontinuous) delta function by a continuous function (which is easier to analyze). We will not however discuss the Dirac delta function in this text.

Our proof of the Weierstrass approximation theorem relies on three key facts. The first fact is that polynomials can be approximations to the identity:

Lemma 3.8.8 (Polynomials can approximate the identity) *For every* $\varepsilon > 0$ *and* $0 < \delta < 1$ *there exists an* (ε, δ)*-approximation to the identity which is a polynomial* P *on* $[-1, 1]$.

Proof See Exercise 3.8.2. □

We will use these polynomial approximations to the identity to approximate continuous functions by polynomials. We will need the following important notion of a *convolution.*

Definition 3.8.9 (*Convolution*) Let $f: \mathbf{R} \to \mathbf{R}$ and $g: \mathbf{R} \to \mathbf{R}$ be continuous, compactly supported functions. We define the *convolution* $f * g: \mathbf{R} \to \mathbf{R}$ of f and g to be the function

$$(f * g)(x) := \int_{-\infty}^{\infty} f(y)g(x-y)\, dy.$$

Note that if f and g are continuous and compactly supported, then for each x the function $f(y)g(x-y)$ (thought of as a function of y) is also continuous and compactly supported, so the above definition makes sense.

Remark 3.8.10 Convolutions play an important rôle in Fourier analysis and in partial differential equations (PDE), and are also important in physics, engineering, and signal processing. An in-depth study of convolution is beyond the scope of this text; only a brief treatment will be given here.

Proposition 3.8.11 (Basic properties of convolution) *Let* $f: \mathbf{R} \to \mathbf{R}$*,* $g: \mathbf{R} \to \mathbf{R}$*, and* $h: \mathbf{R} \to \mathbf{R}$ *be continuous, compactly supported functions. Then the following statements are true.*

(a) The convolution $f * g$ is also a continuous, compactly supported function.
(b) (Convolution is commutative) We have $f * g = g * f$.
(c) (Convolution is linear) We have $f * (g + h) = f * g + f * h$. Also, for any real number c, we have $f * (cg) = (cf) * g = c(f * g)$.

Proof See Exercise 3.8.4. □

Remark 3.8.12 There are many other important properties of convolution, for instance it is associative, $(f * g) * h = f * (g * h)$, and it commutes with derivatives, $(f * g)' = f' * g = f * g'$, when f and g are differentiable. The Dirac delta function δ mentioned earlier is an identity for convolution: $f * \delta = \delta * f = f$. These results are slightly harder to prove than the ones in Proposition 3.8.11, however, and we will not need them in this text.

As mentioned earlier, the proof of the Weierstrass approximation theorem relies on three facts. The second key fact is that convolution with polynomials produces another polynomial:

Lemma 3.8.13 *Let $f : \mathbf{R} \to \mathbf{R}$ be a continuous function supported on $[0, 1]$, and let $g : \mathbf{R} \to \mathbf{R}$ be a continuous function supported on $[-1, 1]$ which is a polynomial on $[-1, 1]$. Then $f * g$ is a polynomial on $[0, 1]$. (Note however that it may be non-polynomial outside of $[0, 1]$.)*

Proof Since g is polynomial on $[-1, 1]$, we may find an integer $n \geq 0$ and real numbers $c_0, c_1, \ldots, c_n$ such that

$$g(x) = \sum_{j=0}^{n} c_j x^j \text{ for all } x \in [-1, 1].$$

On the other hand, for all $x \in [0, 1]$, we have

$$f * g(x) = \int_{-\infty}^{\infty} f(y)g(x - y)\, dy = \int_{[0,1]} f(y)g(x - y)\, dy$$

since f is supported on $[0, 1]$. Since $x \in [0, 1]$ and the variable of integration y is also in $[0, 1]$, we have $x - y \in [-1, 1]$. Thus we may substitute in our formula for g to obtain

$$f * g(x) = \int_{[0,1]} f(y) \sum_{j=0}^{n} c_j (x - y)^j\, dy.$$

We expand this using the binomial formula (Exercise 7.1.4) to obtain

$$f * g(x) = \int_{[0,1]} f(y) \sum_{j=0}^{n} c_j \sum_{k=0}^{j} \frac{j!}{k!(j - k)!} x^k (-y)^{j-k}\, dy.$$

We can interchange the two summations (by Corollary 7.1.14) to obtain

$$f * g(x) = \int_{[0,1]} f(y) \sum_{k=0}^{n} \sum_{j=k}^{n} c_j \frac{j!}{k!(j-k)!} x^k (-y)^{j-k}\, dy$$

(why did the limits of summation change? It may help to plot j and k on a graph). Now we interchange the k summation with the integral, and observe that x^k is independent of y, to obtain

$$f * g(x) = \sum_{k=0}^{n} x^k \int_{[0,1]} f(y) \sum_{j=k}^{n} c_j \frac{j!}{k!(j-k)!} (-y)^{j-k}\, dy.$$

If we thus define

$$C_k := \int_{[0,1]} f(y) \sum_{j=k}^{n} c_j \frac{j!}{k!(j-k)!} (-y)^{j-k}\, dy$$

for each $k = 0, \ldots, n$, then C_k is a number which is independent of x, and we have

$$f * g(x) = \sum_{k=0}^{n} C_k x^k$$

for all $x \in [0, 1]$. Thus $f * g$ is a polynomial on $[0, 1]$. □

The third key fact is that if one convolves a uniformly continuous function with an approximation to the identity, we obtain a new function which is close to the original function (which explains the terminology "approximation to the identity"):

Lemma 3.8.14 *Let* $f : \mathbf{R} \to \mathbf{R}$ *be a continuous function supported on* $[0, 1]$*, which is bounded by some* $M > 0$ *(i.e.,* $|f(x)| \leq M$ *for all* $x \in \mathbf{R}$*), and let* $\varepsilon > 0$ *and* $0 < \delta < 1$ *be such that one has* $|f(x) - f(y)| < \varepsilon$ *whenever* $x, y \in \mathbf{R}$ *and* $|x - y| < \delta$*. Let* g *be any* (ε, δ)*-approximation to the identity. Then we have*

$$|f * g(x) - f(x)| \leq (1 + 4M)\varepsilon$$

for all $x \in [0, 1]$.

Proof See Exercise 3.8.6. □

Combining these together, we obtain a preliminary version of the Weierstrass approximation theorem:

Corollary 3.8.15 (Weierstrass approximation theorem I) *Let* $f : \mathbf{R} \to \mathbf{R}$ *be a continuous function supported on* $[0, 1]$*. Then for every* $\varepsilon > 0$*, there exists a function*

$P : \mathbf{R} \to \mathbf{R}$ *which is polynomial on* $[0, 1]$ *and such that* $|P(x) - f(x)| \leq \varepsilon$ *for all* $x \in [0, 1]$.

Proof See Exercise 3.8.7. □

Now we perform a series of modifications to convert Corollary 3.8.15 into the actual Weierstrass approximation theorem. We first need a simple lemma.

Lemma 3.8.16 *Let* $f : [0, 1] \to \mathbf{R}$ *be a continuous function which equals 0 on the boundary of* $[0, 1]$, *i.e.,* $f(0) = f(1) = 0$. *Let* $F : \mathbf{R} \to \mathbf{R}$ *be the function defined by setting* $F(x) := f(x)$ *for* $x \in [0, 1]$ *and* $F(x) := 0$ *for* $x \notin [0, 1]$. *Then* F *is also continuous.*

Proof See Exercise 3.8.9. □

Remark 3.8.17 The function F obtained in Lemma 3.8.16 is sometimes known as the *extension of* f *by zero*.

From Corollary 3.8.15 and Lemma 3.8.16 we immediately obtain

Corollary 3.8.18 (Weierstrass approximation theorem II) *Let* $f : [0, 1] \to \mathbf{R}$ *be a continuous function such that* $f(0) = f(1) = 0$. *Then for every* $\varepsilon > 0$ *there exists a polynomial* $P : [0, 1] \to \mathbf{R}$ *such that* $|P(x) - f(x)| \leq \varepsilon$ *for all* $x \in [0, 1]$.

Now we strengthen Corollary 3.8.18 by removing the assumption that $f(0) = f(1) = 0$.

Corollary 3.8.19 (Weierstrass approximation theorem III) *Let* $f : [0, 1] \to \mathbf{R}$ *be a continuous function. Then for every* $\varepsilon > 0$ *there exists a polynomial* $P : [0, 1] \to \mathbf{R}$ *such that* $|P(x) - f(x)| \leq \varepsilon$ *for all* $x \in [0, 1]$.

Proof Let $F : [0, 1] \to \mathbf{R}$ denote the function

$$F(x) := f(x) - f(0) - x(f(1) - f(0)).$$

Observe that F is also continuous (why?), and that $F(0) = F(1) = 0$. By Corollary 3.8.18, we can thus find a polynomial $Q : [0, 1] \to \mathbf{R}$ such that $|Q(x) - F(x)| \leq \varepsilon$ for all $x \in [0, 1]$. But

$$Q(x) - F(x) = Q(x) + f(0) + x(f(1) - f(0)) - f(x),$$

so the claim follows by setting P to be the polynomial $P(x) := Q(x) + f(0) + x(f(1) - f(0))$. □

Finally, we can prove the full Weierstrass approximation theorem.

Proof of Theorem 3.8.3 Let $f : [a, b] \to \mathbf{R}$ be a continuous function on $[a, b]$. Let $g : [0, 1] \to \mathbf{R}$ denote the function

$$g(x) := f(a + (b - a)x) \text{ for all } x \in [0, 1]$$

Observe then that

$$f(y) = g((y - a)/(b - a)) \text{ for all } y \in [a, b].$$

The function g is continuous on $[0, 1]$ (why?), and so by Corollary 3.8.19 we may find a polynomial $Q : [0, 1] \to \mathbf{R}$ such that $|Q(x) - g(x)| \leq \varepsilon$ for all $x \in [0, 1]$. In particular, for any $y \in [a, b]$, we have

$$|Q((y - a)/(b - a)) - g((y - a)/(b - a))| \leq \varepsilon.$$

If we thus set $P(y) := Q((y - a)/(b - a))$, then we observe that P is also a polynomial (why?), and so we have $|P(y) - f(y)| \leq \varepsilon$ for all $y \in [a, b]$, as desired.

Remark 3.8.20 Note that the Weierstrass approximation theorem only works on bounded intervals $[a, b]$; continuous functions on $\mathbf{R}$ cannot be uniformly approximated by polynomials. For instance, the exponential function $f : \mathbf{R} \to \mathbf{R}$ defined by $f(x) := e^x$ (which we shall study rigorously in Sect. 4.5) cannot be approximated by any polynomial, because exponential functions grow faster than any polynomial (Exercise 4.5.9) and so there is no way one can even make the sup metric between f and a polynomial finite.

Remark 3.8.21 There is a generalization of the Weierstrass approximation theorem to higher dimensions: if K is any compact subset of $\mathbf{R}^n$ (with the Euclidean metric d_{l^2}), and $f : K \to \mathbf{R}$ is a continuous function, then for every $\varepsilon > 0$ there exists a polynomial $P : K \to \mathbf{R}$ of n variables $x_1, \ldots, x_n$ such that $d_\infty(f, P) < \varepsilon$. This general theorem can be proven by a more complicated variant of the arguments here, but we will not do so. (There is in fact an even more general version of this theorem applicable to an arbitrary metric space, known as the *Stone-Weierstrass theorem*, but this is beyond the scope of this text.)

— Exercises —

Exercise 3.8.1 Prove Lemma 3.8.5.

Exercise 3.8.2 (a) Prove that for any real number $0 \leq y \leq 1$ and any natural number $n \geq 0$, that $(1 - y)^n \geq 1 - ny$. (*Hint:* induct on n. Alternatively, differentiate with respect to y.)

(b) Show that $\int_{-1}^{1} (1 - x^2)^n \, dx \geq \frac{1}{\sqrt{n}}$. (*Hint:* for $|x| \leq 1/\sqrt{n}$, use part (a); for $|x| \geq 1/\sqrt{n}$, just use the fact that $(1 - x^2)$ is positive. It is also possible to proceed via trigonometric substitution, but I would not recommend this unless you know what you are doing.)

(c) Prove Lemma 3.8.8. (*Hint:* choose $f(x)$ to equal $c(1-x^2)^N$ for $x \in [-1, 1]$ and to equal zero for $x \notin [-1, 1]$, where N is a large number N, where c is chosen so that f has integral 1, and use (b).)

Exercise 3.8.3 Let $f\colon \mathbf{R} \to \mathbf{R}$ be a compactly supported, continuous function. Show that f is bounded and uniformly continuous. (*Hint:* the idea is to use Proposition 2.3.2 and Theorem 2.3.5, but one must first deal with the issue that the domain $\mathbf{R}$ of f is non-compact.)

Exercise 3.8.4 Prove Proposition 3.8.11. (*Hint:* to show that $f * g$ is continuous, use Exercise 3.8.3.)

Exercise 3.8.5 Let $f\colon \mathbf{R} \to \mathbf{R}$ and $g\colon \mathbf{R} \to \mathbf{R}$ be continuous, compactly supported functions. Suppose that f is supported on the interval $[0, 1]$, and g is constant on the interval $[0, 2]$ (i.e., there is a real number c such that $g(x) = c$ for all $x \in [0, 2]$). Show that the convolution $f * g$ is constant on the interval $[1, 2]$.

Exercise 3.8.6 (a) Let g be an (ε, δ) approximation to the identity. Show that $1 - 2\varepsilon \le \int_{[-\delta,\delta]} g \le 1$.

(b) Prove Lemma 3.8.14. (*Hint:* begin with the identity

$$f * g(x) = \int f(x-y)g(y)\,dy = \int_{[-\delta,\delta]} f(x-y)g(y)\,dy + \int_{[\delta,1]} f(x-y)g(y)\,dy + \int_{[-1,-\delta]} f(x-y)g(y)\,dy.$$

The idea is to show that the first integral is close to $f(x)$, and that the second and third integrals are very small. To achieve the former task, use (a) and the fact that $f(x)$ and $f(x-y)$ are within ε of each other; to achieve the latter task, use property (c) of the approximation to the identity and the fact that f is bounded.)

Exercise 3.8.7 Prove Corollary 3.8.15. (*Hint:* combine Exercise 3.8.3 and Lemmas 3.8.8, 3.8.13, 3.8.14.)

Exercise 3.8.8 Let $f\colon [0, 1] \to \mathbf{R}$ be a continuous function, and suppose that $\int_{[0,1]} f(x)x^n\,dx = 0$ for all non-negative integers $n = 0, 1, 2, \ldots$. Show that f must be the zero function $f \equiv 0$. (*Hint:* first show that $\int_{[0,1]} f(x)P(x)\,dx = 0$ for all polynomials P. Then, using the Weierstrass approximation theorem, show that $\int_{[0,1]} f(x)f(x)\,dx = 0$.)

Exercise 3.8.9 Prove Lemma 3.8.16.

Chapter 4
Power Series

4.1 Formal Power Series

We now discuss an important subclass of series of functions, that of *power series*. As in earlier chapters, we begin by introducing the notion of a formal power series and then focus in later sections on when the series converges to a meaningful function and what one can say about the function obtained in this manner.

Definition 4.1.1 (*Formal power series*) Let a be a real number. A *formal power series centered at a* is any series of the form

$$\sum_{n=0}^{\infty} c_n (x-a)^n$$

where $c_0, c_1, \ldots$ is a sequence of real numbers (not depending on x); we refer to c_n as the n^{th} *coefficient* of this series. Note that each term $c_n(x-a)^n$ in this series is a function of a real variable x.

Example 4.1.2 The series $\sum_{n=0}^{\infty} n!(x-2)^n$ is a formal power series centered at 2. The series $\sum_{n=0}^{\infty} 2^x (x-3)^n$ is not a formal power series, since the coefficients 2^x depend on x.

We call these power series *formal* because we do not yet assume that these series converge for any x. However, these series are automatically guaranteed to converge when $x = a$ (why?). In general, the closer x gets to a, the easier it is for this series to converge. To make this more precise, we need the following definition.

Definition 4.1.3 (*Radius of convergence*) Let $\sum_{n=0}^{\infty} c_n(x-a)^n$ be a formal power series. We define the *radius of convergence* R of this series to be the quantity

$$R := \frac{1}{\limsup_{n\to\infty} |c_n|^{1/n}}$$

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_4

where we adopt the convention that $\frac{1}{0} = +\infty$ and $\frac{1}{+\infty} = 0$.

Remark 4.1.4 Each number $|c_n|^{1/n}$ is non-negative, so the limit $\limsup_{n\to\infty} |c_n|^{1/n}$ can take on any value from 0 to $+\infty$, inclusive. Thus R can also take on any value between 0 and $+\infty$ inclusive (in particular it is not necessarily a real number). Note that the radius of convergence always exists, even if the sequence $|c_n|^{1/n}$ is not convergent, because the lim sup of any sequence always exists (though it might be $+\infty$ or $-\infty$).

Example 4.1.5 The series $\sum_{n=0}^{\infty} n(-2)^n(x-3)^n$ has radius of convergence

$$\frac{1}{\limsup_{n\to\infty} |n(-2)^n|^{1/n}} = \frac{1}{\limsup_{n\to\infty} 2n^{1/n}} = \frac{1}{2}.$$

The series $\sum_{n=0}^{\infty} 2^{n^2}(x+2)^n$ has radius of convergence

$$\frac{1}{\limsup_{n\to\infty} |2^{n^2}|^{1/n}} = \frac{1}{\limsup_{n\to\infty} 2^n} = \frac{1}{+\infty} = 0.$$

The series $\sum_{n=0}^{\infty} 2^{-n^2}(x+2)^n$ has radius of convergence

$$\frac{1}{\limsup_{n\to\infty} |2^{-n^2}|^{1/n}} = \frac{1}{\limsup_{n\to\infty} 2^{-n}} = \frac{1}{0} = +\infty.$$

The significance of the radius of convergence is the following.

Theorem 4.1.6 *Let $\sum_{n=0}^{\infty} c_n(x-a)^n$ be a formal power series, and let R be its radius of convergence.*

(a) (Divergence outside of the radius of convergence) If $x \in \mathbf{R}$ is such that $|x-a| > R$, then the series $\sum_{n=0}^{\infty} c_n(x-a)^n$ is divergent for that value of x.

(b) (Convergence inside the radius of convergence) If $x \in \mathbf{R}$ is such that $|x-a| < R$, then the series $\sum_{n=0}^{\infty} c_n(x-a)^n$ is absolutely convergent for that value of x.

For parts (c)-(e) we assume that $R > 0$ (i.e., the series converges at at least one other point than $x = a$). Let $f\colon (a-R, a+R) \to \mathbf{R}$ be the function $f(x) := \sum_{n=0}^{\infty} c_n(x-a)^n$; this function is guaranteed to exist by (b).

(c) (Uniform convergence on compact sets) For any $0 < r < R$, the series $\sum_{n=0}^{\infty} c_n$ $(x-a)^n$ converges uniformly to f on the compact interval $[a-r, a+r]$. In particular, f is continuous on $(a-R, a+R)$.

(d) (Differentiation of power series) The function f is differentiable on $(a-R, a+R)$, and for any $0 < r < R$, the series $\sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$ converges uniformly to f' on the interval $[a-r, a+r]$.

(e) (Integration of power series) For any closed interval $[y, z]$ contained in $(a-R, a+R)$, we have

$$\int_{[y,z]} f = \sum_{n=0}^{\infty} c_n \frac{(z-a)^{n+1} - (y-a)^{n+1}}{n+1}.$$

Proof See Exercise 4.1.1. □

Theorem 4.1.6 (a) and (b) of the above theorem give another way to find the radius of convergence, by using your favorite convergence test to work out the range of x for which the power series converges:

Example 4.1.7 Consider the power series $\sum_{n=0}^{\infty} n(x-1)^n$. The ratio test shows that this series converges when $|x-1| < 1$ and diverges when $|x-1| > 1$ (why?). Thus the only possible value for the radius of convergence is $R = 1$ (if $R < 1$, then we have contradicted Theorem 4.1.6(a); if $R > 1$, then we have contradicted Theorem 4.1.6(b)).

Remark 4.1.8 Theorem 4.1.6 is silent on what happens when $|x-a| = R$, i.e., at the points $a - R$ and $a + R$. Indeed, one can have either convergence or divergence at those points; see Exercise 4.1.2.

Remark 4.1.9 Note that while Theorem 4.1.6(b) assures us that the power series $\sum_{n=0}^{\infty} c_n(x-a)^n$ will converge pointwise on the interval $(a-R, a+R)$, it need not converge uniformly on that interval (see Exercise 4.1.2(e)). On the other hand, Theorem 4.1.6(c) assures us that the power series will converge on any smaller interval $[a-r, a+r]$. In particular, being uniformly convergent on every closed subinterval of $(a-R, a+R)$ is not enough to guarantee being uniformly convergent on all of $(a-R, a+R)$.

—Exercise—

Exercise 4.1.1 Prove Theorem 4.1.6. (*Hints:* for (a) and (b), use the root test (Theorem 7.5.1). For (c), use the Weierstrass M-test (Theorem 3.5.7). For (d), use Theorem 3.7.1. For (e), use Corollary 3.6.2.)

Exercise 4.1.2 Give examples of a formal power series $\sum_{n=0}^{\infty} c_n x^n$ centered at 0 with radius of convergence 1, which

(a) diverges at both $x = 1$ and $x = -1$;
(b) diverges at $x = 1$ but converges at $x = -1$;
(c) converges at $x = 1$ but diverges at $x = -1$;
(d) converges at both $x = 1$ and $x = -1$.
(e) converges pointwise on $(-1, 1)$, but does not converge uniformly on $(-1, 1)$.

4.2 Real Analytic Functions

A function $f(x)$ which is lucky enough to be representable as a power series has a special name; it is a *real analytic* function.

Definition 4.2.1 (*Real analytic functions*) Let E be a subset of $\mathbf{R}$, and let $f: E \to \mathbf{R}$ be a function. If a is an interior point of E, we say that f is *real analytic at a* if there exists an open interval $(a-r, a+r)$ in E for some $r > 0$ such that there exists a power series $\sum_{n=0}^{\infty} c_n (x-a)^n$ centered at a which has a radius of convergence greater than or equal to r and which converges to f on $(a-r, a+r)$. If E is an open set, and f is real analytic at every point a of E, we say that f is *real analytic on E*.

Example 4.2.2 Consider the function $f: \mathbf{R}\backslash\{1\} \to \mathbf{R}$ defined by $f(x) := 1/(1-x)$. This function is real analytic at 0 because we have a power series $\sum_{n=0}^{\infty} x^n$ centered at 0 which converges to $1/(1-x) = f(x)$ on the interval $(-1, 1)$. This function is also real analytic at 2 because we have a power series $\sum_{n=0}^{\infty} (-1)^{n+1} (x-2)^n$ which converges to $\frac{-1}{1-(-(x-2))} = \frac{1}{1-x} = f(x)$ on the interval $(1, 3)$ (why? Use Lemma 7.3.3). In fact this function is real analytic on all of $\mathbf{R}\backslash\{1\}$; see Exercise 4.2.2.

Remark 4.2.3 The notion of being real analytic is closely related to another notion, that of being *complex analytic*, but this is a topic for complex analysis, and will not be discussed here.

We now discuss which functions are real analytic. From Theorem 4.1.6(c) and (d) we see that if f is real analytic at a point a, then f is both continuous and differentiable on $(a-r, a+r)$ for some $r > 0$. We can in fact say more:

Definition 4.2.4 (*k-times differentiability*) Let E be a subset of $\mathbf{R}$ with the property that every element of E is a limit point of E. We say a function $f: E \to \mathbf{R}$ is *once differentiable on E* iff it is differentiable (so in particular $f': E \to \mathbf{R}$ is also a function on E. More generally, for any $k \geq 2$ we say that $f: E \to \mathbf{R}$ is *k times differentiable on E*, or just *k times differentiable*, iff f is differentiable, and f' is $k-1$ times differentiable. If f is k times differentiable, we define the k^{th} derivative $f^{(k)}: E \to \mathbf{R}$ by the recursive rule $f^{(1)} := f'$, and $f^{(k)} := (f^{(k-1)})'$ for all $k \geq 2$. We also define $f^{(0)} := f$ (this is f differentiated 0 times), and we allow every function to be zero times differentiable (since clearly $f^{(0)}$ exists for every f). A function is said to be *infinitely differentiable* (or *smooth*) iff it is k times differentiable for every $k \geq 0$.

Example 4.2.5 The function $f(x) := |x|^3$ is twice differentiable on $\mathbf{R}$, but not three times differentiable (why?). Indeed, $f^{(2)} = f'' = 6|x|$, which is not differentiable, at 0.

Proposition 4.2.6 (Real analytic functions are k-times differentiable) *Let E be a subset of $\mathbf{R}$, let a be an interior point of E, and and let f be a function which is real analytic at a, thus there is an $r > 0$ for which we have the power series expansion*

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n$$

for all $x \in (a-r, a+r)$. Then for every $k \geq 0$, the function f is k-times differentiable on $(a-r, a+r)$, and for each $k \geq 0$ the k^{th} derivative is given by

$$f^{(k)}(x) = \sum_{n=0}^{\infty} c_{n+k}(n+1)(n+2)\dots(n+k)(x-a)^n$$
$$= \sum_{n=0}^{\infty} c_{n+k}\frac{(n+k)!}{n!}(x-a)^n$$

for all $x \in (a-r, a+r)$.

Proof See Exercise 4.2.3. □

Corollary 4.2.7 (Real analytic functions are infinitely differentiable) *Let* E *be an open subset of* $\mathbf{R}$*, and let* $f: E \to \mathbf{R}$ *be a real analytic function on* E*. Then* f *is infinitely differentiable on* E*. Also, all derivatives of* f *are also real analytic on* E*.*

Proof For every point $a \in E$ and $k \geq 0$, we know from Proposition 4.2.6 that f is k-times differentiable at a (we will have to apply Exercise 10.1.1 k times here, why?). Thus f is k-times differentiable on E for every $k \geq 0$ and is hence infinitely differentiable. Also, from Proposition 4.2.6 we see that each derivative $f^{(k)}$ of f has a convergent power series expansion at every $x \in E$ and thus $f^{(k)}$ is real analytic. □

Example 4.2.8 Consider the function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := |x|$. This function is not differentiable at $x = 0$ and hence cannot be real analytic at $x = 0$. It is however real analytic at every other point $x \in \mathbf{R}\backslash\{0\}$ (why?).

Remark 4.2.9 The converse statement to Corollary 4.2.7 is not true; there are infinitely differentiable functions which are not real analytic. See Exercise 4.5.4.

Proposition 4.2.6 has an important corollary, due to Brook Taylor (1685–1731).

Corollary 4.2.10 (Taylor's formula) *Let* E *be a subset of* $\mathbf{R}$*, let* a *be an interior point of* E*, and let* $f: E \to \mathbf{R}$ *be a function which is real analytic at* a *and has the power series expansion*

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$$

for all $x \in (a-r, a+r)$ *and some* $r > 0$*. Then for any integer* $k \geq 0$*, we have*

$$f^{(k)}(a) = k!c_k,$$

where $k! := 1 \times 2 \times \dots \times k$ *(and we adopt the convention that* $0! = 1$*). In particular, we have* Taylor's formula

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

for all x *in* $(a-r, a+r)$.

Proof See Exercise 4.2.4. □

The power series $\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$ is sometimes called the *Taylor series* of f around a. Taylor's formula thus asserts that if a function is real analytic, then it is equal to its Taylor series.

Remark 4.2.11 Note that Taylor's formula only works for functions which are real analytic; there are examples of functions which are infinitely differentiable but for which Taylor's theorem fails (see Exercise 4.5.4).

Another important corollary of Taylor's formula is that a real analytic function can have at most one power series at a point:

Corollary 4.2.12 (Uniqueness of power series) *Let E be a subset of* $\mathbf{R}$, *let a be an interior point of E, and let $f: E \to \mathbf{R}$ be a function which is real analytic at a. Suppose that f has two power series expansions*

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n$$

and

$$f(x) = \sum_{n=0}^{\infty} d_n (x-a)^n$$

centered at a, each with a nonzero radius of convergence. Then $c_n = d_n$ for all $n \geq 0$.

Proof By Corollary 4.2.10, we have $f^{(k)}(a) = k!c_k$ for all $k \geq 0$. But we also have $f^{(k)}(a) = k!d_k$, by similar reasoning. Since $k!$ is never zero, we can cancel it and obtain $c_k = d_k$ for all $k \geq 0$, as desired. □

Remark 4.2.13 While a real analytic function has a unique power series around any given point, it can certainly have different power series at different points. For instance, the function $f(x) := \frac{1}{1-x}$, defined on $\mathbf{R} - \{1\}$, has the power series

$$f(x) := \sum_{n=0}^{\infty} x^n$$

around 0, on the interval $(-1, 1)$, but also has the power series

$$\begin{aligned} f(x) &= \frac{1}{1-x} = \frac{2}{1-2(x-\frac{1}{2})} \\ &= \sum_{n=0}^{\infty} 2\left(2\left(x-\frac{1}{2}\right)\right)^n = \sum_{n=0}^{\infty} 2^{n+1}\left(x-\frac{1}{2}\right)^n \end{aligned}$$

around 1/2, on the interval $(0, 1)$ (note that the above power series has a radius of convergence of $1/2$, thanks to the root test; see also Exercise 4.2.8).

—Exercise—

Exercise 4.2.1 Let $n \geq 0$ be an integer, let c, a be real numbers, and let f be the function $f(x) := c(x-a)^n$. Show that f is infinitely differentiable, and that $f^{(k)}(x) = c\frac{n!}{(n-k)!}(x-a)^{n-k}$ for all integers $0 \leq k \leq n$. What happens when $k > n$?

Exercise 4.2.2 Show that the function f defined in Example 4.2.2 is real analytic on all of $\mathbf{R}\backslash\{1\}$.

Exercise 4.2.3 Prove Proposition 4.2.6. (*Hint:* induct on k and use Theorem 4.1.6(d).)

Exercise 4.2.4 Use Proposition 4.2.6 and Exercise 4.2.1 to prove Corollary 4.2.10.

Exercise 4.2.5 Let a, b be real numbers, and let $n \geq 0$ be an integer. Prove the identity

$$(x-a)^n = \sum_{m=0}^{n} \frac{n!}{m!(n-m)!}(b-a)^{n-m}(x-b)^m$$

for any real number x. (*Hint:* use the binomial formula, Exercise 7.1.4.) Explain why this identity is consistent with Taylor's theorem and Exercise 4.2.1. (Note however that Taylor's theorem cannot be rigorously applied until one verifies Exercise 4.2.6 below.)

Exercise 4.2.6 Using Exercise 4.2.5, show that every polynomial $P(x)$ of one variable is real analytic on $\mathbf{R}$.

Exercise 4.2.7 Let $m \geq 0$ be a positive integer, and let $0 < x < r$ be real numbers. Use Lemma 7.3.3 to establish the identity

$$\frac{r}{r-x} = \sum_{n=0}^{\infty} x^n r^{-n}$$

for all $x \in (-r, r)$. Using Proposition 4.2.6, conclude the identity

$$\frac{r}{(r-x)^{m+1}} = \sum_{n=m}^{\infty} \frac{n!}{m!(n-m)!} x^{n-m} r^{-n}$$

for all integers $m \geq 0$ and $x \in (-r, r)$. Also explain why the series on the right-hand side is absolutely convergent.

Exercise 4.2.8 Let E be a subset of $\mathbf{R}$, let a be an interior point of E, and let $f: E \to \mathbf{R}$ be a function which is real analytic at a and has a power series expansion

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n$$

at a which converges on the interval $(a-r, a+r)$. Let $(b-s, b+s)$ be any subinterval of $(a-r, a+r)$ for some $s > 0$.

(a) Prove that $|a-b| \leq r-s$, so in particular $|a-b|<r$.

(b) Show that for every $0<\varepsilon<r$, there exists a $C>0$ such that $|c_n| \leq C(r-\varepsilon)^{-n}$ for all integers $n \geq 0$. (*Hint:* what do we know about the radius of convergence of the series $\sum_{n=0}^{\infty} c_n(x-a)^n$?)

(c) Show that the numbers $d_0, d_1, \ldots$ given by the formula

$$d_m := \sum_{n=m}^{\infty} \frac{n!}{m!(n-m)!}(b-a)^{n-m} c_n \text{ for all integers } m \geq 0$$

are well-defined, in the sense that the above series is absolutely convergent. (*Hint:* use (b) and the comparison test, Corollary 7.3.2, followed by Exercise 4.2.7.)

(d) Show that for every $0<\varepsilon<s$ there exists a $C>0$ such that

$$|d_m| \leq C(s-\varepsilon)^{-m}$$

for all integers $m \geq 0$. (*Hint:* use the comparison test, and Exercise 4.2.7.)

(e) Show that the power series $\sum_{m=0}^{\infty} d_m(x-b)^m$ is absolutely convergent for $x \in (b-s, b+s)$ and converges to $f(x)$. (You may need Fubini's theorem for infinite series, Theorem 8.2.2 of *Analysis I*, as well as Exercise 4.2.5. One may also need to study a variant of the d_m in which the c_n are replaced by $|c_n|$.)

(f) Conclude that f is real analytic at every point in $(a-r, a+r)$.

4.3 Abel's Theorem

Let $f(x)=\sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series centered at a with a radius of convergence $0<R<\infty$ strictly between 0 and infinity. From Theorem 4.1.6 we know that the power series converges absolutely whenever $|x-a|<R$ and diverges when $|x-a|>R$. However, at the boundary $|x-a|=R$ the situation is more complicated; the series may either converge or diverge (see Exercise 4.1.2). However, if the series does converge at the boundary point, then it is reasonably well-behaved; in particular, it is continuous at that boundary point.

Theorem 4.3.1 (Abel's theorem) *Let $f(x)=\sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series centered at a with radius of convergence $0<R<\infty$. If the power series converges at $a+R$, then f is continuous at $a+R$, i.e.,*

$$\lim_{x \to a+R: x \in(a-R, a+R)} \sum_{n=0}^{\infty} c_n(x-a)^n=\sum_{n=0}^{\infty} c_n R^n .$$

Similarly, if the power series converges at $a-R$, then f is continuous at $a-R$, i.e.,

$$\lim_{x \to a-R : x \in (a-R, a+R)} \sum_{n=0}^{\infty} c_n (x-a)^n = \sum_{n=0}^{\infty} c_n (-R)^n.$$

Before we prove Abel's theorem, we need the following lemma.

Lemma 4.3.2 (Summation by parts formula) *Let* $(a_n)_{n=0}^{\infty}$ *and* $(b_n)_{n=0}^{\infty}$ *be sequences of real numbers which converge to limits A and B, respectively, i.e.,* $\lim_{n\to\infty} a_n = A$ *and* $\lim_{n\to\infty} b_n = B$. *Suppose that the sum* $\sum_{n=0}^{\infty}(a_{n+1} - a_n)b_n$ *is convergent. Then the sum* $\sum_{n=0}^{\infty} a_{n+1}(b_{n+1} - b_n)$ *is also convergent, and*

$$\sum_{n=0}^{\infty}(a_{n+1} - a_n)b_n = AB - a_0 b_0 - \sum_{n=0}^{\infty} a_{n+1}(b_{n+1} - b_n).$$

Proof See Exercise 4.3.1. □

Remark 4.3.3 One should compare this formula with the more well-known *integration by parts formula*

$$\int_0^{\infty} f'(x)g(x)\,dx = f(x)g(x)|_0^{\infty} - \int_0^{\infty} f(x)g'(x)\,dx,$$

see Proposition 11.10.1.

Proof of Abel's theorem It will suffice to prove the first claim, i.e., that

$$\lim_{x \to a+R : x \in (a-R, a+R)} \sum_{n=0}^{\infty} c_n (x-a)^n = \sum_{n=0}^{\infty} c_n R^n$$

whenever the sum $\sum_{n=0}^{\infty} c_n R^n$ converges; the second claim will then follow (why?) by replacing c_n by $(-1)^n c_n$ in the above claim. If we make the substitutions $d_n := c_n R^n$ and $y := \frac{x-a}{R}$, then the above claim can be rewritten as

$$\lim_{y \to 1 : y \in (-1,1)} \sum_{n=0}^{\infty} d_n y^n = \sum_{n=0}^{\infty} d_n$$

whenever the sum $\sum_{n=0}^{\infty} d_n$ converges. (Why is this equivalent to the previous claim?)

Write $D := \sum_{n=0}^{\infty} d_n$, and for every $N \geq 0$ write

$$S_N := \left(\sum_{n=0}^{N-1} d_n\right) - D$$

so in particular $S_0 = -D$. Then observe that $\lim_{N\to\infty} S_N = 0$, and that $d_n = S_{n+1} - S_n$. Thus for any $y \in (-1, 1)$ we have

$$\sum_{n=0}^{\infty} d_n y^n = \sum_{n=0}^{\infty} (S_{n+1} - S_n) y^n.$$

Applying the summation by parts formula (Lemma 4.3.2), and noting that $\lim_{n\to\infty} y^n = 0$, we obtain

$$\sum_{n=0}^{\infty} d_n y^n = -S_0 y^0 - \sum_{n=0}^{\infty} S_{n+1}(y^{n+1} - y^n).$$

Observe that $-S_0 y^0 = +D$. Thus to finish the proof of Abel's theorem, it will suffice to show that

$$\lim_{y\to 1: y\in(-1,1)} \sum_{n=0}^{\infty} S_{n+1}(y^{n+1} - y^n) = 0.$$

Since y converges to 1, we may as well restrict y to $[0, 1)$ instead of $(-1, 1)$; in particular we may take y to be positive.

From the triangle inequality for series (Proposition 7.2.9), we have

$$\begin{aligned}\left|\sum_{n=0}^{\infty} S_{n+1}(y^{n+1} - y^n)\right| &\leq \sum_{n=0}^{\infty} |S_{n+1}(y^{n+1} - y^n)| \\ &= \sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}),\end{aligned}$$

so by the squeeze test (Corollary 6.4.14) it suffices to show that

$$\lim_{y\to 1: y\in[0,1)} \sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) = 0.$$

The expression $\sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1})$ is clearly non-negative, so it will suffice to show that

$$\limsup_{y\to 1: y\in[0,1)} \sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) = 0.$$

Let $\varepsilon > 0$. Since S_n converges to 0, there exists an N such that $|S_n| \leq \varepsilon$ for all $n > N$. Thus we have

$$\sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) \leq \sum_{n=0}^{N} |S_{n+1}|(y^n - y^{n+1}) + \sum_{n=N+1}^{\infty} \varepsilon(y^n - y^{n+1}).$$

The last summation is a telescoping series, which sums to εy^{N+1} (See Lemma 7.2.14, recalling from Lemma 6.5.2 that $y^n \to 0$ as $n \to \infty$), and thus

$$\sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) \leq \sum_{n=0}^{N} |S_{n+1}|(y^n - y^{n+1}) + \varepsilon y^{N+1}.$$

Now take limits as $y \to 1$. Observe that $y^n - y^{n+1} \to 0$ as $y \to 1$ for every $n \in 0, 1, \dots, N$. Since we can interchange limits and *finite* sums (Exercise 7.1.5), we thus have

$$\limsup_{y \to 1: y \in [0,1)} \sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) \leq \varepsilon.$$

But $\varepsilon > 0$ was arbitrary, and thus we must have

$$\limsup_{y \to 1: y \in [0,1)} \sum_{n=0}^{\infty} |S_{n+1}|(y^n - y^{n+1}) = 0$$

since the left-hand side must be non-negative. The claim follows. □

—Exercise—

Exercise 4.3.1 Prove Lemma 4.3.2. (*Hint:* first work out the relationship between the partial sums $\sum_{n=0}^{N}(a_{n+1} - a_n)b_n$ and $\sum_{n=0}^{N} a_{n+1}(b_{n+1} - b_n)$.)

4.4 Multiplication of Power Series

We now show that the product of two real analytic functions is again real analytic.

Theorem 4.4.1 *Let $f\colon (a-r, a+r) \to \mathbf{R}$ and $g\colon (a-r, a+r) \to \mathbf{R}$ be functions analytic on $(a-r, a+r)$, with power series expansions*

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n$$

and

$$g(x) = \sum_{n=0}^{\infty} d_n (x-a)^n$$

, respectively. Then $fg\colon (a-r, a+r) \to \mathbf{R}$ is also analytic on $(a-r, a+r)$, with power series expansion

$$f(x)g(x) = \sum_{n=0}^{\infty} e_n (x-a)^n$$

where $e_n := \sum_{m=0}^{n} c_m d_{n-m}$.

Remark 4.4.2 The sequence $(e_n)_{n=0}^{\infty}$ is sometimes referred to as the *convolution* of the sequences $(c_n)_{n=0}^{\infty}$ and $(d_n)_{n=0}^{\infty}$; it is closely related (though not identical) to the notion of convolution introduced in Definition 3.8.9.

Proof We have to show that the series $\sum_{n=0}^{\infty} e_n(x-a)^n$ converges to $f(x)g(x)$ for all $x \in (a-r, a+r)$. Now fix x to be any point in $(a-r, a+r)$. By Theorem 4.1.6, we see that both f and g have radii of convergence at least r. In particular, the series $\sum_{n=0}^{\infty} c_n(x-a)^n$ and $\sum_{n=0}^{\infty} d_n(x-a)^n$ are absolutely convergent. Thus if we define

$$C := \sum_{n=0}^{\infty} |c_n(x-a)^n|$$

and

$$D := \sum_{n=0}^{\infty} |d_n(x-a)^n|$$

then C and D are both finite.

For any $N \geq 0$, consider the partial sum

$$\sum_{n=0}^{N} \sum_{m=0}^{\infty} |c_m(x-a)^m d_n(x-a)^n|.$$

We can rewrite this as

$$\sum_{n=0}^{N} |d_n(x-a)^n| \sum_{m=0}^{\infty} |c_m(x-a)^m|,$$

which by definition of C is equal to

$$\sum_{n=0}^{N} |d_n(x-a)^n| C,$$

which by definition of D is less than or equal to DC. Thus the above partial sums are bounded by DC for every N. In particular, the series

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} |c_m(x-a)^m d_n(x-a)^n|$$

is convergent, which means that the sum

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_m(x-a)^m d_n(x-a)^n$$

is absolutely convergent.

Let us now compute this sum in two ways. First of all, we can pull the $d_n(x-a)^n$ factor out of the $\sum_{m=0}^\infty$ summation, to obtain

$$\sum_{n=0}^\infty d_n(x-a)^n \sum_{m=0}^\infty c_m(x-a)^m.$$

By our formula for $f(x)$, this is equal to

$$\sum_{n=0}^\infty d_n(x-a)^n f(x);$$

by our formula for $g(x)$, this is equal to $f(x)g(x)$. Thus

$$f(x)g(x) = \sum_{n=0}^\infty \sum_{m=0}^\infty c_m(x-a)^m d_n(x-a)^n.$$

Now we compute this sum in a different way. We rewrite it as

$$f(x)g(x) = \sum_{n=0}^\infty \sum_{m=0}^\infty c_m d_n(x-a)^{n+m}.$$

By Fubini's theorem for series (Theorem 8.2.2), because the series was absolutely convergent, we may rewrite it as

$$f(x)g(x) = \sum_{m=0}^\infty \sum_{n=0}^\infty c_m d_n(x-a)^{n+m}.$$

Now make the substitution $n' := n+m$, to rewrite this as

$$f(x)g(x) = \sum_{m=0}^\infty \sum_{n'=m}^\infty c_m d_{n'-m}(x-a)^{n'}.$$

If we adopt the convention that $d_j = 0$ for all negative j, then this is equal to

$$f(x)g(x) = \sum_{m=0}^\infty \sum_{n'=0}^\infty c_m d_{n'-m}(x-a)^{n'}.$$

Applying Fubini's theorem again, we obtain

$$f(x)g(x) = \sum_{n'=0}^{\infty} \sum_{m=0}^{\infty} c_m d_{n'-m} (x-a)^{n'},$$

which we can rewrite as

$$f(x)g(x) = \sum_{n'=0}^{\infty} (x-a)^{n'} \sum_{m=0}^{\infty} c_m d_{n'-m}.$$

Since d_j was 0 when j is negative, we can rewrite this as

$$f(x)g(x) = \sum_{n'=0}^{\infty} (x-a)^{n'} \sum_{m=0}^{n'} c_m d_{n'-m},$$

which by definition of e is

$$f(x)g(x) = \sum_{n'=0}^{\infty} e_{n'} (x-a)^{n'},$$

as desired. □

4.5 The Exponential and Logarithm Functions

We can now use the machinery developed in the last few sections to develop a rigorous foundation for many standard functions used in mathematics. We begin with the exponential function.

Definition 4.5.1 (*Exponential function*) For every real number x, we define the *exponential function* $\exp(x)$ to be the real number

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Theorem 4.5.2 (Basic properties of exponential)

(a) For every real number x, the series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ is absolutely convergent. In particular, $\exp(x)$ exists and is real for every $x \in \mathbf{R}$, the power series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ has an infinite radius of convergence, and $\exp$ is a real analytic function on $(-\infty, \infty)$.
(b) $\exp$ is differentiable on $\mathbf{R}$, and for every $x \in \mathbf{R}$, $\exp'(x) = \exp(x)$.
(c) $\exp$ is continuous on $\mathbf{R}$, and for every interval $[a, b]$, we have $\int_{[a,b]} \exp(x)\, dx = \exp(b) - \exp(a)$.
(d) For every $x, y \in \mathbf{R}$, we have $\exp(x + y) = \exp(x) \exp(y)$.

(e) *We have* $\exp(0) = 1$. *Also, for every* $x \in \mathbf{R}$, $\exp(x)$ *is positive, and* $\exp(-x) = 1/\exp(x)$.

(f) $\exp$ *is strictly monotone increasing: in other words, if* x, y *are real numbers, then we have* $\exp(y) > \exp(x)$ *if and only if* $y > x$.

Proof See Exercise 4.5.1. □

One can write the exponential function in a more compact form, introducing famous *Euler's number* $e = 2.71828183\ldots$, also known as the *base of the natural logarithm*:

Definition 4.5.3 (*Euler's number*) The number e is defined to be

$$e := \exp(1) = \sum_{n=0}^{\infty} \frac{1}{n!} = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots.$$

Proposition 4.5.4 *For every real number* x, *we have* $\exp(x) = e^x$.

Proof See Exercise 4.5.3. □

In light of this proposition we can and will use e^x and $\exp(x)$ interchangeably.

Since $e > 1$ (why?), we see that $e^x \to +\infty$ as $x \to +\infty$, and $e^x \to 0$ as $x \to -\infty$. From this and the intermediate value theorem (Theorem 9.7.1) we see that the range of the function exp is $(0, \infty)$. Since exp is strictly increasing, it is injective, and hence exp is a bijection from $\mathbf{R}$ to $(0, \infty)$ and thus has an inverse from $(0, \infty) \to \mathbf{R}$. This inverse has a name:

Definition 4.5.5 (*Logarithm*) We define the *natural logarithm function* $\log\colon (0, \infty) \to \mathbf{R}$ (also called ln) to be the inverse of the exponential function. Thus $\exp(\log(x)) = x$ and $\log(\exp(x)) = x$.

Since exp is continuous and strictly monotone increasing, we see that log is also continuous and strictly monotone increasing (see Proposition 9.8.3). Since exp is also differentiable, and the derivative is never zero, we see from the inverse function theorem (Theorem 10.4.2) that log is also differentiable. We list some other properties of the natural logarithm below.

Theorem 4.5.6 (Logarithm properties)

(a) *For every* $x \in (0, \infty)$, *we have* $\ln'(x) = \frac{1}{x}$. *In particular, by the fundamental theorem of calculus, we have* $\int_{[a,b]} \frac{1}{x}\, dx = \ln(b) - \ln(a)$ *for any interval* $[a, b]$ *in* $(0, \infty)$.

(b) *We have* $\ln(xy) = \ln(x) + \ln(y)$ *for all* $x, y \in (0, \infty)$.

(c) *We have* $\ln(1) = 0$ *and* $\ln(1/x) = -\ln(x)$ *for all* $x \in (0, \infty)$.

(d) *For any* $x \in (0, \infty)$ *and* $y \in \mathbf{R}$, *we have* $\ln(x^y) = y \ln(x)$.

(e) For any $x \in (-1, 1)$*, we have*

$$\ln(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n}.$$

In particular, $\ln$ *is analytic at* 1*, with the power series expansion*

$$\ln(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}(x-1)^n$$

for $x \in (0, 2)$*, with radius of convergence 1.*

Proof See Exercise 4.5.5. □

Example 4.5.7 We now give a modest application of Abel's theorem (Theorem 4.3.1): from the alternating series test we see that $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$ is convergent. By Abel's theorem we thus see that

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \lim_{x \to 2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}(x-1)^n$$
$$= \lim_{x \to 2} \ln(x) = \ln(2),$$

thus we have the formula

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \dots.$$

—Exercise—

Exercise 4.5.1 Prove Theorem 4.5.2. (*Hints:* for part (a), use the ratio test. For parts (bc), use Theorem 4.1.6. For part (d), use Theorem 4.4.1. For part (e), use part (d). For part (f), use part (d), and prove that $\exp(x) > 1$ when x is positive. You may find the binomial formula from Exercise 7.1.4 to be useful.)

Exercise 4.5.2 Show that for every integer $n \geq 3$, we have

$$0 < \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \dots < \frac{1}{n!}.$$

(*Hint:* first show that $(n+k)! > 2^k n!$ for all $k = 1, 2, 3, \dots$.) Conclude that $n!e$ is not an integer for every $n \geq 3$. Deduce from this that e is irrational. (*Hint:* prove by contradiction.)

Exercise 4.5.3 Prove Proposition 4.5.4. (*Hint:* first prove the claim when x is a natural number. Then prove it when x is an integer. Then prove it when x is a rational number. Then use the fact that real numbers are the limits of rational numbers to prove it for all real numbers. You may find the exponent laws (Proposition 6.7.3) to be useful.)

Exercise 4.5.4 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function defined by setting $f(x) := \exp(-1/x)$ when $x > 0$, and $f(x) := 0$ when $x \leq 0$. Prove that f is infinitely differentiable, and $f^{(k)}(0) = 0$ for every integer $k \geq 0$, but that f is not real analytic at 0.

Exercise 4.5.5 Prove Theorem 4.5.6. (*Hints:* for part (a), use the inverse function theorem (Theorem 10.4.2) or the chain rule (Theorem 10.1.15). For parts (bcd), use Theorem 4.5.2 and the exponent laws (Proposition 6.7.3). For part (e), start with the geometric series formula (Lemma 7.3.3) and integrate using Theorem 4.1.6).

Exercise 4.5.6 Prove that the natural logarithm function is real analytic on $(0, +\infty)$.

Exercise 4.5.7 Let $f\colon \mathbf{R} \to (0, \infty)$ be a positive, real analytic function such that $f'(x) = f(x)$ for all $x \in \mathbf{R}$. Show that $f(x) = Ce^x$ for some positive constant C; justify your reasoning. (*Hint:* there are basically three different proofs available. One proof uses the logarithm function, another proof uses the function e^{-x}, and a third proof uses power series. Of course, you only need to supply one proof.)

Exercise 4.5.8 Let $m > 0$ be an integer. Show that

$$\lim_{x \to +\infty} \frac{e^x}{x^m} = +\infty.$$

(*Hint:* what happens to the ratio between $e^{x+1}/(x+1)^m$ and e^x/x^m as $x \to +\infty$?)

Exercise 4.5.9 Let $P(x)$ be a polynomial, and let $c > 0$. Show that there exists a real number $N > 0$ such that $e^{cx} > |P(x)|$ for all $x > N$; thus an exponentially growing function, no matter how small the growth rate c, will eventually overtake any given polynomial $P(x)$, no matter how large. (*Hint:* use Exercise 4.5.8.)

Exercise 4.5.10 Let $f\colon (0, +\infty) \times \mathbf{R} \to \mathbf{R}$ be the exponential function $f(x, y) := x^y$. Show that f is continuous. (*Hint:* note that Propositions 9.4.10, 9.4.11 only show that f is continuous in each variable, which is insufficient, as Exercise 2.2.11 shows. The easiest way to proceed is to write $f(x, y) = \exp(y \ln x)$ and use the continuity of exp() and ln(). For an extra challenge, try proving this exercise without using the logarithm function.)

4.6 A Digression on Complex Numbers

To proceed further we need the complex number system $\mathbf{C}$, which is an extension of the real number system $\mathbf{R}$. A full discussion of this important number system (and in particular the branch of mathematics known as *complex analysis*) is beyond the scope of this text; here, we need the system primarily because of a very useful mathematical operation, the *complex exponential function* $z \mapsto \exp(z)$, which generalizes the real exponential function $x \mapsto \exp(x)$ introduced in the previous section.

Informally, we could define the complex numbers as

Definition 4.6.1 (*Informal definition of complex numbers*) The complex numbers $\mathbf{C}$ are the set of all numbers of the form $a + bi$, where a, b are real numbers and i is a square root of -1, $i^2 = -1$.

However, this definition is a little unsatisfactory as it does not explain how to add, multiply, or compare two complex numbers. To construct the complex numbers rigorously we will first introduce a *formal* version of the complex number $a + bi$, which we shall temporarily denote as (a, b); this is similar to how in Chap. 4, when constructing the integers $\mathbf{Z}$, we needed a formal notion of subtraction a——b before the actual notion of subtraction $a - b$ could be introduced, or how when constructing the rational numbers, a formal notion of division $a//b$ was needed before it was superceded by the actual notion a/b of division. It is also similar to how, in the construction of the real numbers, we defined a formal limit $\mathrm{LIM}_{n\to\infty} a_n$ before we defined a genuine limit $\lim_{n\to\infty} a_n$.

Definition 4.6.2 (*Formal definition of complex numbers*) A *complex number* is any pair of the form (a, b), where a, b are real numbers, thus for instance $(2, 4)$ is a complex number. Two complex numbers (a, b), (c, d) are said to be equal iff $a = c$ and $b = d$, thus for instance $(2 + 1, 3 + 4) = (3, 7)$, but $(2, 1) \neq (1, 2)$ and $(2, 4) \neq (2, -4)$. The set of all complex numbers is denoted $\mathbf{C}$.

At this stage the complex numbers $\mathbf{C}$ are indistinguishable from the Cartesian product $\mathbf{R}^2 = \mathbf{R} \times \mathbf{R}$ (also known as the *Cartesian plane*). However, we will introduce a number of operations on the complex numbers, notably that of *complex multiplication*, which are not normally placed on the Cartesian plane $\mathbf{R}^2$. Thus one can think of the complex number system $\mathbf{C}$ as the Cartesian plane $\mathbf{R}^2$ equipped with a number of additional structures. We begin with the notion of addition and negation. Using the informal definition of the complex numbers, we expect

$$(a, b) + (c, d) = (a + bi) + (c + di) = (a + c) + (b + d)i = (a + c, b + d)$$

and similarly

$$-(a, b) = -(a + bi) = (-a) + (-b)i = (-a, -b).$$

As these derivations used the informal definition of the complex numbers, these identities have not yet been rigorously proven. However we shall simply *encode* these identities into our complex number system by defining the notion of addition and negation by the above rules:

Definition 4.6.3 (*Complex addition, negation, and zero*) If $z = (a, b)$ and $w = (c, d)$ are two complex numbers, we define their *sum* $z + w$ to be the complex number $z + w := (a + c, b + d)$. Thus for instance $(2, 4) + (3, -1) = (5, 3)$. We also define the *negation* $-z$ of z to be the complex number $-z := (-a, -b)$, thus for instance $-(3, -1) = (-3, 1)$. We also define the *complex zero* $0_{\mathbf{C}}$ to be the complex number $0_{\mathbf{C}} = (0, 0)$.

It is easy to see that notion of addition is well-defined in the sense that if $z = z'$ and $w = w'$ then $z + w = z' + w'$. Similarly for negation. The complex addition, negation, and zero operations obey the usual laws of arithmetic:

Lemma 4.6.4 (The complex numbers are an additive group) *If z_1, z_2, z_3 are complex numbers, then we have the commutative property $z_1 + z_2 = z_2 + z_1$, the associative property $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$, the identity property $z_1 + 0_{\mathbf{C}} = 0_{\mathbf{C}} + z_1 = z_1$, and the inverse property $z_1 + (-z_1) = (-z_1) + z_1 = 0_{\mathbf{C}}$.*

Proof See Exercise 4.6.1. □

Next, we define the notion of complex multiplication and reciprocal. The informal justification of the complex multiplication rule is

$$\begin{aligned}(a, b) \cdot (c, d) &= (a + bi)(c + di)\\ &= ac + adi + bic + bidi\\ &= (ac - bd) + (ad + bc)i\\ &= (ac - bd, ad + bc)\end{aligned}$$

since i^2 is supposed to equal -1. Thus we define

Definition 4.6.5 (*Complex multiplication*) If $z = (a, b)$ and $w = (c, d)$ are complex numbers, then we define their *product* zw to be the complex number $zw := (ac - bd, ad + bc)$. We also introduce the *complex identity* $1_{\mathbf{C}} := (1, 0)$.

This operation is easily seen to be well-defined, and also obeys the usual laws of arithmetic:

Lemma 4.6.6 *If z_1, z_2, z_3 are complex numbers, then we have the commutative property $z_1 z_2 = z_2 z_1$, the associative property $(z_1 z_2) z_3 = z_1 (z_2 z_3)$, the identity property $z_1 1_{\mathbf{C}} = 1_{\mathbf{C}} z_1 = z_1$, and the distributivity properties $z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3$ and $(z_2 + z_3) z_1 = z_2 z_1 + z_3 z_1$.*

Proof See Exercise 4.6.2. □

The above lemma can also be stated more succinctly, as the assertion that $\mathbf{C}$ is a commutative ring. As is usual, we now write $z - w$ as shorthand for $z + (-w)$.

We now identify the real numbers $\mathbf{R}$ with a subset of the complex numbers $\mathbf{C}$ by identifying any real number x with the complex number $(x, 0)$, thus $x \equiv (x, 0)$. Note that this identification is consistent with equality (thus $x = y$ iff $(x, 0) = (y, 0)$), with addition ($x_1 + x_2 = x_3$ iff $(x_1, 0) + (x_2, 0) = (x_3, 0)$), with negation ($x = -y$ iff $(x, 0) = -(y, 0)$), and multiplication ($x_1 x_2 = x_3$ iff $(x_1, 0)(x_2, 0) = (x_3, 0)$), so we will no longer need to distinguish between "real addition" and "complex addition", and similarly for equality, negation, and multiplication. For instance, we can compute $3(2, 4)$ by identifying the real number 3 with the complex number $(3, 0)$ and then computing $(3, 0)(2, 4) = (3 \times 2 - 0 \times 4, 3 \times 4 + 0 \times 2) = (6, 12)$. Note also that $0 \equiv 0_{\mathbf{C}}$ and $1 \equiv 1_{\mathbf{C}}$, so we can now drop the $\mathbf{C}$ subscripts from the zero 0 and the identity 1.

We now define i to be the complex number $i := (0, 1)$. We can now reconstruct the informal definition of the complex numbers as a lemma:

Lemma 4.6.7 *Every complex number $z \in \mathbf{C}$ can be written as $z = a + bi$ for exactly one pair a, b of real numbers. Also, we have $i^2 = -1$, and $-z = (-1)z$.*

Proof See Exercise 4.6.3. □

Because of this lemma, we will now refer to complex numbers in the more usual notation $a + bi$ and discard the formal notation (a, b) henceforth.

Definition 4.6.8 (*Real and imaginary parts*) If z is a complex number with the representation $z = a + bi$ for some real numbers a, b, we shall call a the *real part* of z and denote $\Re(z) := a$, and call b the *imaginary part* of z and denote $\Im(z) := b$, thus for instance $\Re(3 + 4i) = 3$ and $\Im(3 + 4i) = 4$, and in general $z = \Re(z) + i\Im(z)$. Note that z is real iff $\Im(z) = 0$. We say that z is *imaginary* iff $\Re(z) = 0$, thus for instance $4i$ is imaginary, while $3 + 4i$ is neither real nor imaginary, and 0 is both real and imaginary. We define the *complex conjugate* $\overline{z}$ of z to be the complex number $\overline{z} := \Re(z) - i\Im(z)$, thus for instance $\overline{3 + 4i} = 3 - 4i$, $\overline{i} = -i$, and $\overline{3} = 3$.

The operation of complex conjugation has several nice properties:

Lemma 4.6.9 (Complex conjugation is an involution) *Let z, w be complex numbers, then $\overline{z + w} = \overline{z} + \overline{w}$, $\overline{-z} = -\overline{z}$, and $\overline{zw} = \overline{z}\,\overline{w}$. Also $\overline{\overline{z}} = z$. Finally, we have $\overline{z} = \overline{w}$ if and only if $z = w$, and $\overline{z} = z$ if and only if z is real.*

Proof See Exercise 4.6.4. □

The notion of absolute value $|x|$ was defined for rational numbers x in Definition 4.3.1, and this definition extends to real numbers in the obvious manner. However, we cannot extend this definition directly to the complex numbers, as most complex numbers are neither positive nor negative. (For instance, we do not classify i as either a positive or negative number; see Exercise 4.6.11 for some reasons why.) However, we can still define absolute value by generalizing the formula $|x| = \sqrt{x^2}$ from Exercise 5.6.4:

Definition 4.6.10 (*Complex absolute value*) If $z = a + bi$ is a complex number, we define the *absolute value* $|z|$ of z to be the real number $|z| := \sqrt{a^2 + b^2} = (a^2 + b^2)^{1/2}$.

From Exercise 5.6.4 we see that this notion of absolute value generalizes the notion of real absolute value. The absolute value has a number of other good properties:

Lemma 4.6.11 (Properties of complex absolute value) *Let z, w be complex numbers. Then $|z|$ is a non-negative real number, and $|z| = 0$ if and only if $z = 0$. Also we have the identity $z\overline{z} = |z|^2$, and so $|z| = \sqrt{z\overline{z}}$. As a consequence we have $|zw| = |z||w|$ and $|\overline{z}| = |z|$. Finally, we have the inequalities*

$$-|z| \leq \Re(z) \leq |z|; \quad -|z| \leq \Im(z) \leq |z|; \quad |z| \leq |\Re(z)| + |\Im(z)|$$

as well as the triangle inequality $|z + w| \leq |z| + |w|$.

Proof See Exercise 4.6.6. □

Using the notion of absolute value, we can define a notion of reciprocal:

Definition 4.6.12 (*Complex reciprocal*) If z is a nonzero complex number, we define the *reciprocal* z^{-1} of z to be the complex number $z^{-1} := |z|^{-2}\overline{z}$ (note that $|z|^{-2}$ is well-defined as a positive real number because $|z|$ is positive real, thanks to Lemma 4.6.11). Thus for instance $(1 + 2i)^{-1} = |1 + 2i|^{-2}(1 - 2i) = (1^2 + 2^2)^{-1}(1 - 2i) = \frac{1}{5} - \frac{2}{5}i$. If z is zero, $z = 0$, we leave the reciprocal 0^{-1} undefined.

From the definition and Lemma 4.6.11, we see that

$$zz^{-1} = z^{-1}z = |z|^{-2}\overline{z}z = |z|^{-2}|z|^2 = 1,$$

and so z^{-1} is indeed the reciprocal of z. We can thus define a notion of quotient z/w for any two complex numbers z, w with $w \neq 0$ in the usual manner by the formula $z/w := zw^{-1}$.

The complex numbers can be given a distance by defining $d(z, w) = |z - w|$.

Lemma 4.6.13 *The complex numbers* **C** *with the distance d form a metric space. If $(z_n)_{n=1}^{\infty}$ is a sequence of complex numbers, and z is another complex number, then we have* $\lim_{n\to\infty} z_n = z$ *in this metric space if and only if* $\lim_{n\to\infty} \Re(z_n) = \Re(z)$ *and* $\lim_{n\to\infty} \Im(z_n) = \Im(z)$.

Proof See Exercise 4.6.9. □

Observe that with our choice of definitions, the space **C** of complex numbers is identical (as a metric space) to the Euclidean plane $\mathbf{R}^2$, since the complex distance between two complex numbers (a, b), (a', b') is exactly the same as the Euclidean distance $\sqrt{(a - a')^2 + (b - b')^2}$ between these points. Thus, every metric property that $\mathbf{R}^2$ satisfies is also obeyed by **C**; for instance, **C** is complete and connected, but not compact.

We also have the usual limit laws:

Lemma 4.6.14 (Complex limit laws) *Let $(z_n)_{n=1}^{\infty}$ and $(w_n)_{n=1}^{\infty}$ be convergent sequences of complex numbers, and let c be a complex number. Then the sequences $(z_n + w_n)_{n=1}^{\infty}$, $(z_n - w_n)_{n=1}^{\infty}$, $(cz_n)_{n=1}^{\infty}$, $(z_n w_n)_{n=1}^{\infty}$, and $(\overline{z_n})_{n=1}^{\infty}$ are also convergent, with*

$$\lim_{n\to\infty} z_n + w_n = \lim_{n\to\infty} z_n + \lim_{n\to\infty} w_n$$
$$\lim_{n\to\infty} z_n - w_n = \lim_{n\to\infty} z_n - \lim_{n\to\infty} w_n$$
$$\lim_{n\to\infty} cz_n = c \lim_{n\to\infty} z_n$$
$$\lim_{n\to\infty} z_n w_n = \left(\lim_{n\to\infty} z_n\right)\left(\lim_{n\to\infty} w_n\right)$$
$$\lim_{n\to\infty} \overline{z_n} = \overline{\lim_{n\to\infty} z_n}$$

Also, if the w_n are all nonzero and $\lim_{n\to\infty} w_n$ *is also nonzero, then* $(z_n/w_n)_{n=1}^{\infty}$ *is also a convergent sequence, with*

$$\lim_{n\to\infty} z_n/w_n = \left(\lim_{n\to\infty} z_n)/(\lim_{n\to\infty} w_n\right).$$

Proof See Exercise 4.6.10. □

Observe that the real and complex number systems are in fact quite similar; they both obey similar laws of arithmetic, and they have similar structure as metric spaces. Indeed many of the results in this textbook that were proven for real-valued functions are also valid for complex-valued functions, simply by replacing "real" with "complex" in the proofs but otherwise leaving all the other details of the proof unchanged. Alternatively, one can always split a complex-valued function f into real and imaginary parts $\Re(f)$, $\Im(f)$, thus $f = \Re(f) + i\Im(f)$, and then deduce results for the complex-valued function f from the corresponding results for the real-valued functions $\Re(f)$, $\Im(f)$. For instance, the theory of pointwise and uniform convergence from Chapter 3, or the theory of power series from this chapter, extends without any difficulty to complex-valued functions. In particular, we can define the complex exponential function in exactly the same manner as for real numbers:

Definition 4.6.15 (*Complex exponential*) If z is a complex number, we define the function $\exp(z)$ by the formula

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Inspired by Proposition 4.5.4, we shall use $\exp(z)$ and e^z interchangeably. It is also possible to define a^z for complex z and other real numbers $a > 0$, but we will not need to do so in this text.

One can state and prove the ratio test for complex series and use it to show that $\exp(z)$ converges for every z. It turns out that many of the properties from Theorem 4.5.2 still hold: we have that $\exp(z + w) = \exp(z)\exp(w)$, for instance; see Exercise 4.6.12. (The other properties require complex differentiation and complex integration, but these topics are beyond the scope of this text.) Another useful observation is that $\overline{\exp(z)} = \exp(\overline{z})$; this can be seen by conjugating the partial sums $\sum_{n=0}^{N} \frac{z^n}{n!}$ and taking limits as $N \to \infty$.

The complex logarithm turns out to be somewhat more subtle, mainly because exp is no longer invertible, and also because the various power series for the logarithm only have a finite radius of convergence (unlike exp, which has an infinite radius of convergence). This rather delicate issue is beyond the scope of this text and will not be discussed here.

—Exercise—

Exercise 4.6.1 Prove Lemma 4.6.4.

Exercise 4.6.2 Prove Lemma 4.6.6.

Exercise 4.6.3 Prove Lemma 4.6.7.

Exercise 4.6.4 Prove Lemma 4.6.9.

Exercise 4.6.5 If z is a complex number, show that $\Re(z) = \frac{z+\overline{z}}{2}$ and $\Im(z) = \frac{z-\overline{z}}{2i}$.

Exercise 4.6.6 Prove Lemma 4.6.11. (*Hint:* to prove the triangle inequality, first prove that $\Re(z\overline{w}) \leq |z||w|$, and hence (from Exercise 4.6.5) that $z\overline{w} + \overline{z}w \leq 2|z||w|$. Then add $|z|^2 + |w|^2$ to both sides of this inequality.)

Exercise 4.6.7 Show that if z, w are complex numbers with $w \neq 0$, then $|z/w| = |z|/|w|$.

Exercise 4.6.8 Let z, w be nonzero complex numbers. Show that $|z + w| = |z| + |w|$ if and only if there exists a positive real number $c > 0$ such that $z = cw$.

Exercise 4.6.9 Prove Lemma 4.6.13.

Exercise 4.6.10 Prove Lemma 4.6.14. (*Hint:* split z_n and w_n into real and imaginary parts and use the usual limit laws, Lemma 6.1.19, combined with Lemma 4.6.13.)

Exercise 4.6.11 The purpose of this exercise is to explain why we do not try to organize the complex numbers into positive and negative parts. Suppose that there was a notion of a "positive complex number" and a "negative complex number" which obeyed the following reasonable axioms (cf. Proposition 4.2.9):

- (Trichotomy) For every complex number z, exactly one of the following statements is true: z is positive, z is negative, z is zero.
- (Negation) If z is a positive complex number, then $-z$ is negative. If z is a negative complex number, then $-z$ is positive.
- (Additivity) If z and w are positive complex numbers, then $z + w$ is also positive.
- (Multiplicativity) If z and w are positive complex numbers, then zw is also positive.

Show that these four axioms are inconsistent,, i.e., one can use these axioms to deduce a contradiction. (*Hints:* first use the axioms to deduce that 1 is positive, and then conclude that -1 is negative. Then apply the Trichotomy axiom to $z = i$ and obtain a contradiction in any one of the three cases.)

Exercise 4.6.12 Prove the ratio test for complex series, and use it to show that the series used to define the complex exponential is absolutely convergent. Then prove that $\exp(z + w) = \exp(z)\exp(w)$ for all complex numbers z, w.

4.7 Trigonometric Functions

We now discuss the next most important class of special functions, after the exponential and logarithmic functions, namely the trigonometric functions. (There are several other useful special functions in mathematics, such as the hyperbolic trigonometric functions and hypergeometric functions, the gamma and zeta functions, and elliptic functions, but they occur more rarely and will not be discussed here.)

Trigonometric functions are often defined using geometric concepts, notably those of circles, triangles, and angles. However, it is also possible to define them using more analytic concepts and in particular the (complex) exponential function.

Definition 4.7.1 (*Trigonometric functions*) If z is a complex number, then we define

$$\cos(z) := \frac{e^{iz} + e^{-iz}}{2}$$

and

$$\sin(z) := \frac{e^{iz} - e^{-iz}}{2i}.$$

We refer to cos and sin as the *cosine* and *sine* functions, respectively.

These formulae were discovered by Leonhard Euler (1707–1783) in 1748, who recognized the link between the complex exponential and the trigonometric functions. Note that since we have defined the sine and cosine for complex numbers z, we automatically have defined them also for real numbers x. In fact in most applications one is only interested in the trigonometric functions when applied to real numbers.

From the power series definition of exp, we have

$$e^{iz} = 1 + iz - \frac{z^2}{2!} - \frac{iz^3}{3!} + \frac{z^4}{4!} + \dots$$

and

$$e^{-iz} = 1 - iz - \frac{z^2}{2!} + \frac{iz^3}{3!} + \frac{z^4}{4!} - \dots$$

and so from the above formulae we have

$$\cos(z) = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}$$

and

$$\sin(z) = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}.$$

In particular, $\cos(x)$ and $\sin(x)$ are always real whenever x is real. From the ratio test we see that the two power series $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$, $\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$ are absolutely convergent for every x, thus $\sin(x)$ and $\cos(x)$ are real analytic at 0 with an infinite radius of convergence. From Exercise 4.2.8 we thus see that the sine and cosine functions are real analytic on all of **R**. (They are also complex analytic on all of **C**, but we will not pursue this matter in this text.) In particular the sine and cosine functions are continuous and differentiable.

We list some basic properties of the sine and cosine functions below.

Theorem 4.7.2 (Trigonometric identities) *Let x, y be real numbers.*

(a) *We have* $\sin(x)^2 + \cos(x)^2 = 1$. *In particular, we have* $\sin(x) \in [-1, 1]$ *and* $\cos(x) \in [-1, 1]$ *for all* $x \in \mathbf{R}$.
(b) *We have* $\sin'(x) = \cos(x)$ *and* $\cos'(x) = -\sin(x)$.
(c) *We have* $\sin(-x) = -\sin(x)$ *and* $\cos(-x) = \cos(x)$.
(d) *We have* $\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$ *and* $\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y)$.
(e) *We have* $\sin(0) = 0$ *and* $\cos(0) = 1$.
(f) *We have* $e^{ix} = \cos(x) + i\sin(x)$ *and* $e^{-ix} = \cos(x) - i\sin(x)$. *In particular* $\cos(x) = \Re(e^{ix})$ *and* $\sin(x) = \Im(e^{ix})$.

Proof See Exercise 4.7.1. □

Now we describe some other properties of sin and cos.

Lemma 4.7.3 *There exists a positive number x such that* $\sin(x)$ *is equal to* 0.

Proof Suppose for sake of contradiction that $\sin(x) \neq 0$ for all $x \in (0, \infty)$. Observe that this would also imply that $\cos(x) \neq 0$ for all $x \in (0, \infty)$, since if $\cos(x) = 0$ then $\sin(2x) = 0$ by Theorem 4.7.2(d) (why?). Since $\cos(0) = 1$, this implies by the intermediate value theorem (Theorem 9.7.1) that $\cos(x) > 0$ for all $x > 0$ (why?). Also, since $\sin(0) = 0$ and $\sin'(0) = 1 > 0$, we see that sin increasing near 0, hence is positive to the right of 0. By the intermediate value theorem again we conclude that $\sin(x) > 0$ for all $x > 0$ (otherwise sin would have a zero on $(0, \infty)$).

In particular if we define the cotangent function $\cot(x) := \cos(x)/\sin(x)$, then $\cot(x)$ would be positive and differentiable on all of $(0, \infty)$. From the quotient rule (Theorem 10.1.13(h)) and Theorem 4.7.2 we see that the derivative of $\cot(x)$ is $-1/\sin(x)^2$ (why?). In particular, we have $\cot'(x) \leq -1$ for all $x > 0$. By the fundamental theorem of calculus (Theorem 11.9.1) this implies that $\cot(x + s) \leq \cot(x) - s$ for all $x > 0$ and $s > 0$. But letting $s \to \infty$ we see that this contradicts our assertion that cot is positive on $(0, \infty)$ (why?). □

Let E be the set $E := \{x \in (0, +\infty) : \sin(x) = 0\}$, i.e., E is the set of roots of sin on $(0, +\infty)$. By Lemma 4.7.3, E is non-empty. Since $\sin'(0) > 0$, there exists a $c > 0$ such that $E \subseteq [c, +\infty)$ (see Exercise 4.7.2). Also, since sin is continuous in $[c, +\infty)$, E is closed in $[c, +\infty)$ (why? Use Theorem 2.1.5(d)). Since $[c, +\infty)$ is closed in **R**, we conclude that E is closed in **R**. Thus E contains all its adherent points, and thus contains $\inf(E)$. Thus if we make the definition

Definition 4.7.4 We define π to be the number

$$\pi := \inf\{x \in (0, \infty) : \sin(x) = 0\}$$

then we have $\pi \in E \subseteq [c, +\infty)$ (so in particular $\pi > 0$) and $\sin(\pi) = 0$. By definition of π, sin cannot have any zeroes in $(0, \pi)$, and so in particular must be positive on $(0, \pi)$, (cf. the arguments in Lemma 4.7.3 using the intermediate value theorem). Since $\cos'(x) = -\sin(x)$, we thus conclude that $\cos(x)$ is strictly decreasing on $(0, \pi)$. Since $\cos(0) = 1$, this implies in particular that $\cos(\pi) < 1$; since $\sin^2(\pi) + \cos^2(\pi) = 1$ and $\sin(\pi) = 0$, we thus conclude that $\cos(\pi) = -1$.

In particular we have Euler's famous formula

$$e^{\pi i} = \cos(\pi) + i \sin(\pi) = -1.$$

We now conclude with some other properties of sine and cosine.

Theorem 4.7.5 (Periodicity of trigonometric functions) *Let x be a real number.*

(a) We have $\cos(x + \pi) = -\cos(x)$ and $\sin(x + \pi) = -\sin(x)$. In particular we have $\cos(x + 2\pi) = \cos(x)$ and $\sin(x + 2\pi) = \sin(x)$, i.e., sin *and* cos *are periodic with period 2π.*
(b) We have $\sin(x) = 0$ if and only if x/π is an integer.
(c) We have $\cos(x) = 0$ if and only if x/π is an integer plus 1/2.

Proof See Exercise 4.7.3. □

We can of course define all the other trigonometric functions: tangent, cotangent, secant, and cosecant, and develop all the familiar identities of trigonometry; some examples of this are given in the exercises.

—Exercise—

Exercise 4.7.1 Prove Theorem 4.7.2. (*Hint:* write everything in terms of exponentials whenever possible.)

Exercise 4.7.2 Let $f : \mathbf{R} \to \mathbf{R}$ be a function which is differentiable at x_0, with $f(x_0) = 0$ and $f'(x_0) \neq 0$. Show that there exists a $c > 0$ such that $f(y)$ is nonzero whenever $0 < |x_0 - y| < c$. Conclude in particular that there exists a $c > 0$ such that $\sin(x) \neq 0$ for all $0 < x < c$.

Exercise 4.7.3 Prove Theorem 4.7.5. (*Hint:* for (c), you may wish to first compute $\sin(\pi/2)$ and $\cos(\pi/2)$, and then link $\cos(x)$ to $\sin(x + \pi/2)$.)

Exercise 4.7.4 Let x, y be real numbers such that $x^2 + y^2 = 1$. Show that there is exactly one real number $\theta \in (-\pi, \pi]$ such that $x = \sin(\theta)$ and $y = \cos(\theta)$. (*Hint:* you may need to divide into cases depending on whether x, y are positive, negative, or zero.)

Exercise 4.7.5 Show that if $r, s > 0$ are positive real numbers, and θ, α are real numbers such that $re^{i\theta} = se^{i\alpha}$, then $r = s$ and $\theta = \alpha + 2\pi k$ for some integer k.

Exercise 4.7.6 Let z be a nonzero complex number. Using Exercise 4.7.4, show that there is exactly one pair of real numbers r, θ such that $r > 0, \theta \in (-\pi, \pi]$, and $z = re^{i\theta}$. (This is sometimes known as the *standard polar representation* of z.)

Exercise 4.7.7 For any real number θ and integer n, prove the *de Moivre identities*

$$\cos(n\theta) = \Re((\cos\theta + i\sin\theta)^n); \quad \sin(n\theta) = \Im((\cos\theta + i\sin\theta)^n).$$

Exercise 4.7.8 Let $\tan\colon (-\pi/2, \pi/2) \to \mathbf{R}$ be the tangent function $\tan(x) := \sin(x)/\cos(x)$. Show that tan is differentiable and monotone increasing, with $\frac{d}{dx}\tan(x) = 1 + \tan(x)^2$, and that $\lim_{x\to\pi/2}\tan(x) = +\infty$ and $\lim_{x\to-\pi/2}\tan(x) = -\infty$. Conclude that tan is in fact a bijection from $(-\pi/2, \pi/2) \to \mathbf{R}$, and thus has an inverse function $\tan^{-1}\colon \mathbf{R} \to (-\pi/2, \pi/2)$ (this function is called the *arctangent function*). Show that $\tan^{-1}$ is differentiable and $\frac{d}{dx}\tan^{-1}(x) = \frac{1}{1+x^2}$.

Exercise 4.7.9 Recall the arctangent function $\tan^{-1}$ from Exercise 4.7.8. By modifying the proof of Theorem 4.5.6(e), establish the identity

$$\tan^{-1}(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{2n+1}$$

for all $x \in (-1, 1)$. Using Abel's theorem (Theorem 4.3.1) to extend this identity to the case $x = 1$, conclude in particular the identity

$$\pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \ldots = 4\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}.$$

(Note that the series converges by the alternating series test, Proposition 7.2.11.) Conclude in particular that $4 - \frac{4}{3} < \pi < 4$. (One can of course compute $\pi = 3.1415926\ldots$ to much higher accuracy, though if one wishes to do so it is advisable to use a different formula than the one above, which converges very slowly.)

Exercise 4.7.10 Let $f\colon \mathbf{R} \to \mathbf{R}$ be the function

$$f(x) := \sum_{n=1}^{\infty} 4^{-n}\cos(32^n \pi x).$$

(a) Show that this series is uniformly convergent, and that f is continuous.

(b) Show that for every integer j and every integer $m \geq 1$, we have

$$|f\left(\frac{j+1}{32^m}\right) - f\left(\frac{j}{32^m}\right)| \geq 4^{-m}.$$

(*Hint:* use the identity

$$\sum_{n=1}^{\infty} a_n = \left(\sum_{n=1}^{m-1} a_n \right) + a_m + \sum_{n=m+1}^{\infty} a_n$$

for certain sequences a_n. Also, use the fact that the cosine function is periodic with period 2π, as well as the geometric series formula $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$ for any $|r| < 1$. Finally, you will need the inequality $|\cos(x) - \cos(y)| \leq |x - y|$ for any real numbers x and y; this can be proven by using the mean value theorem (Corollary 10.2.9), or the fundamental theorem of calculus (Theorem 11.9.4).)

(c) Using (b), show that for every real number x_0, the function f is not differentiable at x_0. (*Hint:* for every x_0 and every $m \geq 1$, there exists an integer j such that $j \leq 32^m x_0 \leq j + 1$, thanks to Exercise 5.4.3.)

(d) Explain briefly why the result in (c) does not contradict Corollary 3.7.3.

Chapter 5
Fourier Series

In the previous two chapters, we discussed the issue of how certain functions (for instance, compactly supported continuous functions) could be approximated by polynomials. Later, we showed how a different class of functions (real analytic functions) could be written exactly (not approximately) as an infinite polynomial, or more precisely a power series.

Power series are already immensely useful, especially when dealing with special functions such as the exponential and trigonometric functions discussed earlier. However, there are some circumstances where power series are not so useful, because one has to deal with functions (e.g., $\sqrt{x}$) which are not real analytic, and so do not have power series.

Fortunately, there is another type of series expansion, known as *Fourier series*, which is also a very powerful tool in analysis (though used for slightly different purposes). Instead of analyzing compactly supported functions, it instead analyzes *periodic functions*; instead of decomposing into polynomials, it decomposes into *trigonometric polynomials*. Roughly speaking, the theory of Fourier series asserts that just about every periodic function can be decomposed as an (infinite) sum of sines and cosines.

Remark 5.0.1 Jean-Baptiste Fourier (1768–1830) was, among other things, an administrator accompanying Napoleon on his invasion of Egypt, and then a Prefect in France during Napoleon's reign. After the Napoleonic wars, he returned to mathematics. He introduced Fourier series in an important 1807 paper in which he used them to solve what is now known as the heat equation. At the time, the claim that every periodic function could be expressed as a sum of sines and cosines was extremely controversial, even such leading mathematicians as Euler declared that it was impossible. Nevertheless, Fourier managed to show that this was indeed the case, although the proof was not completely rigorous and was not totally accepted for almost another hundred years.

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_5

There will be some similarities between the theory of Fourier series and that of power series, but there are also some major differences. For instance, the convergence of Fourier series is usually not uniform (i.e., not in the L^∞ metric), but instead we have convergence in a different metric, the L^2-metric. Also, we will need to use complex numbers heavily in our theory, while they played only a tangential rôle in power series.

The theory of Fourier series (and of related topics such as Fourier integrals and the Laplace transform) is vast, and deserves an entire course in itself. It has many, many applications, most directly to differential equations, signal processing, electrical engineering, physics, and analysis, but also to algebra and number theory. We will only give the barest bones of the theory here, however, and almost no applications.

5.1 Periodic Functions

The theory of Fourier series has to do with the analysis of *periodic functions*, which we now define. It turns out to be convenient to work with complex-valued functions rather than real-valued ones.

Definition 5.1.1 Let $L > 0$ be a real number. A function $f: \mathbf{R} \to \mathbf{C}$ is *periodic with period* L, or L*-periodic*, if we have $f(x + L) = f(x)$ for every real number x.

Example 5.1.2 The real-valued functions $f(x) = \sin(x)$ and $f(x) = \cos(x)$ are 2π-periodic, as is the complex-valued function $f(x) = e^{ix}$. These functions are also 4π-periodic, 6π-periodic, etc. (why?). The function $f(x) = x$, however, is not periodic. The constant function $f(x) = 1$ is L-periodic for every L.

Remark 5.1.3 If a function f is L-periodic, then we have $f(x + kL) = f(x)$ for every integer k (why? Use induction for the positive k, and then use a substitution to convert the positive k result to a negative k result. The $k = 0$ case is of course trivial). In particular, if a function f is 1-periodic, then we have $f(x + k) = f(x)$ for every $k \in \mathbf{Z}$. Because of this, 1-periodic functions are sometimes also called $\mathbf{Z}$*-periodic* (and L-periodic functions called $L\mathbf{Z}$-periodic).

Example 5.1.4 For any integer n, the functions $\cos(2\pi nx)$, $\sin(2\pi nx)$, and $e^{2\pi inx}$ are all $\mathbf{Z}$-periodic. (What happens when n is not an integer?) Another example of a $\mathbf{Z}$-periodic function is the function $f: \mathbf{R} \to \mathbf{C}$ defined by $f(x):=1$ when $x \in [n, n + \frac{1}{2})$ for some integer n, and $f(x):=0$ when $x \in [n + \frac{1}{2}, n + 1)$ for some integer n. This function is an example of a *square wave*.

Henceforth, for simplicity, we shall only deal with functions which are $\mathbf{Z}$-periodic (for the Fourier theory of L-periodic functions, see Exercise 5.5.6). Note that in order to completely specify a $\mathbf{Z}$-periodic function $f: \mathbf{R} \to \mathbf{C}$, one only needs to specify its values on the interval $[0, 1)$, since this will determine the values of f everywhere else. This is because every real number x can be written in the form $x = k + y$

where k is an integer (called the *integer part* of x, and sometimes denoted $[x]$) and $y \in [0, 1)$ (this is called the *fractional part* of x, and sometimes denoted $\{x\}$); see Exercise 5.1.1. Because of this, sometimes when we wish to describe a $\mathbf{Z}$-periodic function f we just describe what it does on the interval $[0, 1)$, and then say that it is *extended periodically* to all of $\mathbf{R}$. This means that we define $f(x)$ for any real number x by setting $f(x) := f(y)$, where we have decomposed $x = k + y$ as discussed above. (One can in fact replace the interval $[0, 1)$ by any other half-open interval of length 1, but we will not do so here.)

The space of complex-valued continuous $\mathbf{Z}$-periodic functions is denoted $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$. (The notation $\mathbf{R}/\mathbf{Z}$ comes from algebra, and denotes the quotient group of the additive group $\mathbf{R}$ by the additive group $\mathbf{Z}$; more information in this can be found in any algebra text.) By "continuous" we mean continuous at all points on $\mathbf{R}$; merely being continuous on an interval such as $[0, 1]$ will not suffice, as there may be a discontinuity between the left and right limits at 1 (or at any other integer). Thus for instance, the functions $\sin(2\pi nx)$, $\cos(2\pi nx)$, and $e^{2\pi inx}$ are all elements of $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, as are the constant functions, however the square wave function described earlier is not in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ because it is not continuous. Also the function $\sin(x)$ would also not qualify to be in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ since it is not $\mathbf{Z}$-periodic.

Lemma 5.1.5 (Basic properties of $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$)

(a) (Boundedness) If $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, then f is bounded (i.e., there exists a real number $M > 0$ such that $|f(x)| \leq M$ for all $x \in \mathbf{R}$).

(b) (Vector space and algebra properties) If $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, then the functions $f + g$, $f - g$, and fg are also in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$. Also, if c is any complex number, then the function cf is also in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.

(c) (Closure under uniform limits) If $(f_n)_{n=1}^{\infty}$ is a sequence of functions in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ which converges uniformly to another function $f: \mathbf{R} \to \mathbf{C}$, then f is also in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.

Proof See Exercise 5.1.2. □

One can make $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ into a metric space by re-introducing the now familiar sup norm metric

$$d_\infty(f, g) = \sup_{x \in \mathbf{R}} |f(x) - g(x)| = \sup_{x \in [0,1)} |f(x) - g(x)|$$

of uniform convergence. (Why is the first supremum the same as the second?) See Exercise 5.1.3.

— Exercise —

Exercise 5.1.1 Show that every real number x can be written in exactly one way in the form $x = k + y$, where k is an integer and $y \in [0, 1)$. (Hint: to prove existence of such a representation, set $k := \sup\{l \in \mathbf{Z} : l \leq x\}$.)

Exercise 5.1.2 Prove Lemma 5.1.5. (Hint: for (a), first show that f is bounded on $[0, 1]$.)

Exercise 5.1.3 Show that $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ with the sup norm metric d_∞ is a metric space. Furthermore, show that this metric space is complete.

5.2 Inner Products on Periodic Functions

From Lemma 5.1.5 we know that we can add, subtract, multiply, and take limits of continuous periodic functions. We will need a couple more operations on the space $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, though. The first one is that of *inner product*.

Definition 5.2.1 (*Inner product*) If $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, we define the *inner product* $\langle f, g\rangle$ to be the quantity

$$\langle f, g\rangle = \int_{[0,1]} f(x)\overline{g(x)}\, \mathrm{d}x.$$

Remark 5.2.2 In order to integrate a complex-valued function, $f(x) = g(x) + ih(x)$, we use the definition that $\int_{[a,b]} f := \int_{[a,b]} g + i \int_{[a,b]} h$; i.e., we integrate the real and imaginary parts of the function separately. Thus for instance $\int_{[1,2]}(1 + ix)\, \mathrm{d}x = \int_{[1,2]} 1\, \mathrm{d}x + i \int_{[1,2]} x\, \mathrm{d}x = 1 + \frac{3}{2}i$. It is easy to verify that all the standard rules of calculus (integration by parts, fundamental theorem of calculus, substitution, etc.) still hold when the functions are complex-valued instead of real-valued.

Example 5.2.3 Let f be the constant function $f(x):=1$, and let $g(x)$ be the function $g(x):=e^{2\pi ix}$. Then we have

$$\begin{aligned}\langle f, g\rangle &= \int_{[0,1]} 1\overline{e^{2\pi ix}}\, \mathrm{d}x \\ &= \int_{[0,1]} e^{-2\pi ix}\, \mathrm{d}x \\ &= \frac{e^{-2\pi ix}}{-2\pi i}|_{x=0}^{x=1} \\ &= \frac{e^{-2\pi i} - e^0}{-2\pi i} \\ &= \frac{1-1}{-2\pi i} \\ &= 0.\end{aligned}$$

Remark 5.2.4 In general, the inner product $\langle f, g\rangle$ will be a complex number. (Note that $f(x)\overline{g(x)}$ will be Riemann integrable since both functions are bounded and continuous.)

Roughly speaking, the inner product $\langle f, g\rangle$ is to the space $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ what the dot product $x \cdot y$ is to Euclidean spaces such as $\mathbf{R}^n$. We list some basic properties of the inner product below; a more in-depth study of inner products on vector spaces can be found in any linear algebra text but is beyond the scope of this text.

Lemma 5.2.5 *Let $f, g, h \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.*

(a) (Hermitian property) We have $\langle g, f\rangle = \overline{\langle f, g\rangle}$.
(b) (Positivity) We have $\langle f, f\rangle \geq 0$. Furthermore, we have $\langle f, f\rangle = 0$ if and only if $f = 0$ (i.e., $f(x) = 0$ for all $x \in \mathbf{R}$).
(c) (Linearity in the first variable) We have $\langle f + g, h\rangle = \langle f, h\rangle + \langle g, h\rangle$. For any complex number c, we have $\langle cf, g\rangle = c\langle f, g\rangle$.
(d) (Antilinearity in the second variable) We have $\langle f, g + h\rangle = \langle f, g\rangle + \langle f, h\rangle$. For any complex number c, we have $\langle f, cg\rangle = \overline{c}\langle f, g\rangle$.

Proof See Exercise 5.2.1. □

From the positivity property, it makes sense to define the L^2 *norm* $\|f\|_2$ of a function $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ by the formula

$$\|f\|_2 := \sqrt{\langle f, f\rangle} = \left(\int_{[0,1]} f(x)\overline{f(x)}\, dx\right)^{1/2} = \left(\int_{[0,1]} |f(x)|^2\, dx\right)^{1/2}.$$

Thus $\|f\|_2 \geq 0$ for all f. The norm $\|f\|_2$ is sometimes called the *root mean square* of f.

Example 5.2.6 If $f(x)$ is the function $e^{2\pi i x}$, then

$$\|f\|_2 = \left(\int_{[0,1]} e^{2\pi i x}e^{-2\pi i x}\, dx\right)^{1/2} = \left(\int_{[0,1]} 1\, dx\right)^{1/2} = 1^{1/2} = 1.$$

This L^2 norm is related to, but is distinct from, the L^∞ norm $\|f\|_\infty := \sup_{x\in\mathbf{R}} |f(x)|$. For instance, if $f(x) = \sin(2\pi x)$, then $\|f\|_\infty = 1$ but $\|f\|_2 = \frac{1}{\sqrt{2}}$. In general, the best one can say is that $0 \leq \|f\|_2 \leq \|f\|_\infty$; see Exercise 5.2.3.

Some basic properties of the L^2 norm are given below.

Lemma 5.2.7 *Let $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.*

(a) (Non-degeneracy) We have $\|f\|_2 = 0$ if and only if $f = 0$.
(b) (Cauchy–Schwarz inequality) We have $|\langle f, g\rangle| \leq \|f\|_2\|g\|_2$.
(c) (Triangle inequality) We have $\|f + g\|_2 \leq \|f\|_2 + \|g\|_2$.

(d) (Pythagoras' theorem) If $\langle f, g\rangle = 0$*, then* $\|f+g\|_2^2 = \|f\|_2^2 + \|g\|_2^2$.
(e) (Homogeneity) We have $\|cf\|_2 = |c|\|f\|_2$ *for all* $c \in \mathbf{C}$.

Proof See Exercise 5.2.2. □

In light of Pythagoras' theorem, we sometimes say that f and g are *orthogonal* iff $\langle f, g\rangle = 0$.

We can now define the L^2 *metric* d_{L^2} on $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ by defining

$$d_{L^2}(f,g) := \|f-g\|_2 = \left(\int_{[0,1]} |f(x)-g(x)|^2 \, \mathrm{d}x\right)^{1/2}.$$

Remark 5.2.8 One can verify that d_{L^2} is indeed a metric (Exercise 5.2.4). Indeed, the L^2 metric is very similar to the l^2 metric on Euclidean spaces $\mathbf{R}^n$, which is why the notation is deliberately chosen to be similar; you should compare the two metrics yourself to see the analogy.

Note that a sequence f_n of functions in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ will *converge in the* L^2 *metric* to $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ if $d_{L^2}(f_n, f) \to 0$ as $n \to \infty$, or in other words that

$$\lim_{n\to\infty} \int_{[0,1]} |f_n(x) - f(x)|^2 \, \mathrm{d}x = 0.$$

Remark 5.2.9 The notion of convergence in L^2 metric is different from that of uniform or pointwise convergence; see Exercise 5.2.6.

Remark 5.2.10 The L^2 metric is not as well-behaved as the L^∞ metric. For instance, it turns out the space $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ is not complete in the L^2 metric, despite being complete in the L^∞ metric; see Exercise 5.2.5.

— Exercise —

Exercise 5.2.1 Prove Lemma 5.2.5. (Hint: the last part of (b) is a little tricky. You may need to prove by contradiction, assuming that f is not the zero function, and then show that $\int_{[0,1]} |f(x)|^2$ is strictly positive. You will need to use the fact that f, and hence $|f|$, is continuous, to do this.)

Exercise 5.2.2 Prove Lemma 5.2.7. (Hint: use Lemma 5.2.5 frequently. For the Cauchy–Schwarz inequality, begin with the positivity property $\langle f, f\rangle \geq 0$, but with f replaced by the function $f\|g\|_2^2 - \langle f, g\rangle g$, and then simplify using Lemma 5.2.5. You may have to treat the case $\|g\|_2 = 0$ separately. Use the Cauchy–Schwarz inequality to prove the triangle inequality.)

Exercise 5.2.3 If $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ is a non-zero function, show that $0 < \|f\|_2 \leq \|f\|_{L^\infty}$. Conversely, if $0 < A \leq B$ are real numbers, show that there exists a non-zero function $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ such that $\|f\|_2 = A$ and $\|f\|_\infty = B$. (Hint: let g

be a non-constant non-negative real-valued function in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, and consider functions f of the form $f = (c + dg)^{1/2}$ for some constant real numbers $c, d > 0$.)

Exercise 5.2.4 Prove that the L^2 metric d_{L^2} on $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ does indeed turn $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ into a metric space. (cf. Exercise 1.1.6).

Exercise 5.2.5 Find a sequence of continuous periodic functions which converge in L^2 to a discontinuous periodic function. (Hint: try converging to the square wave function.)

Exercise 5.2.6 Let $f \in C(\mathbf{R}/\mathbf{Z}, \mathbf{C})$, and let $(f_n)_{n=1}^{\infty}$ be a sequence of functions in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.

(a) Show that if f_n converges uniformly to f, then f_n also converges to f in the L^2 metric.
(b) Give an example where f_n converges to f in the L^2 metric, but does not converge to f uniformly. (Hint: take $f = 0$. Try to make the functions f_n large in sup norm.)
(c) Give an example where f_n converges to f in the L^2 metric, but does not converge to f pointwise. (Hint: take $f = 0$. Try to make the functions f_n large at one point.)
(d) Give an example where f_n converges to f pointwise, but does not converge to f in the L^2 metric. (Hint: take $f = 0$. Try to make the functions f_n large in L^2 norm.)

5.3 Trigonometric Polynomials

We now define the concept of a *trigonometric polynomial*. Just as polynomials are combinations of the functions x^n (sometimes called *monomials*), trigonometric polynomials are combinations of the functions $e^{2\pi inx}$ (sometimes called *characters*).

Definition 5.3.1 (*Characters*) For every integer n, we let $e_n \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ denote the function

$$e_n(x) := e^{2\pi inx}.$$

This is sometimes referred to as the *character with frequency n*.

Definition 5.3.2 (*Trigonometric polynomials*) A function f in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ is said to be a *trigonometric polynomial* if we can write $f = \sum_{n=-N}^{N} c_n e_n$ for some integer $N \geq 0$ and some complex numbers $(c_n)_{n=-N}^{N}$.

Example 5.3.3 The function $f = 4e_{-2} + ie_{-1} - 2e_0 + 0e_1 - 3e_2$ is a trigonometric polynomial; it can be written more explicitly as

$$f(x) = 4e^{-4\pi ix} + ie^{-2\pi ix} - 2 - 3e^{4\pi ix}.$$

Example 5.3.4 For any integer n, the function $\cos(2\pi nx)$ is a trigonometric polynomial, since

$$\cos(2\pi nx) = \frac{e^{2\pi inx} + e^{-2\pi inx}}{2} = \frac{1}{2}e_{-n} + \frac{1}{2}e_n.$$

Similarly the function $\sin(2\pi nx) = \frac{-1}{2i}e_{-n} + \frac{1}{2i}e_n$ is a trigonometric polynomial. In fact, any linear combination of sines and cosines is also a trigonometric polynomial, for instance $3 + i\cos(2\pi x) + 4i\sin(4\pi x)$ is a trigonometric polynomial.

The Fourier theorem will allow us to write any function in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ as a Fourier series, which is to trigonometric polynomials what power series is to polynomials. To do this we will use the inner product structure from the previous section. The key computation is

Lemma 5.3.5 (Characters are an orthonormal system) *For any integers n and m, we have $\langle e_n, e_m\rangle = 1$ when $n = m$ and $\langle e_n, e_m\rangle = 0$ when $n \neq m$. Also, we have $\|e_n\| = 1$.*

Proof See Exercise 5.3.2. □

As a consequence, we have a formula for the coefficients of a trigonometric polynomial.

Corollary 5.3.6 *Let $f = \sum_{n=-N}^{N} c_n e_n$ be a trigonometric polynomial. Then we have the formula*

$$c_n = \langle f, e_n\rangle$$

for all integers $-N \leq n \leq N$. Also, we have $0 = \langle f, e_n\rangle$ whenever $n > N$ or $n < -N$. Also, we have the identity

$$\|f\|_2^2 = \sum_{n=-N}^{N} |c_n|^2.$$

Proof See Exercise 5.3.3. □

We rewrite the conclusion of this corollary in a different way.

Definition 5.3.7 (*Fourier transform*) For any function $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{R})$, and any integer $n \in \mathbf{Z}$, we define the n^{th} *Fourier coefficient of f*, denoted $\hat{f}(n)$, by the formula

$$\hat{f}(n) := \langle f, e_n\rangle = \int_{[0,1]} f(x)e^{-2\pi inx}\, \mathrm{d}x.$$

The function $\hat{f}\colon \mathbf{Z} \to \mathbf{C}$ is called the *Fourier transform* of f.

From Corollary 5.3.6, we see that whenever $f = \sum_{n=-N}^{N} c_n e_n$ is a trigonometric polynomial, we have

$$f = \sum_{n=-N}^{N} \langle f, e_n \rangle e_n = \sum_{n=-\infty}^{\infty} \langle f, e_n \rangle e_n$$

and in particular we have the *Fourier inversion formula*

$$f = \sum_{n=-\infty}^{\infty} \hat{f}(n) e_n$$

or in other words

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{2\pi i n x}.$$

The right-hand side is referred to as the *Fourier series* of f. Also, from the second identity of Corollary 5.3.6 we have the *Plancherel formula*

$$\|f\|_2^2 = \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2.$$

Remark 5.3.8 We stress that at present we have only proven the Fourier inversion and Plancherel formulae in the case when f is a trigonometric polynomial. Note that in this case that the Fourier coefficients $\hat{f}(n)$ are mostly zero (indeed, they can only be non-zero when $-N \leq n \leq N$), and so this infinite sum is really just a finite sum in disguise. In particular there are no issues about what sense the above series converge in; they both converge pointwise, uniformly, and in L^2 metric, since they are just finite sums.

In the next few sections we will extend the Fourier inversion and Plancherel formulae to general functions in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, not just trigonometric polynomials. (It is also possible to extend the formula to discontinuous functions such as the square wave, but we will not do so here.) To do this we will need a version of the Weierstrass approximation theorem, this time requiring that a continuous periodic function be approximated uniformly by *trigonometric* polynomials. Just as convolutions were used in the proof of the polynomial Weierstrass approximation theorem, we will also need a notion of convolution tailored for periodic functions.

— Exercise —

Exercise 5.3.1 Show that the sum or product of any two trigonometric polynomials is again a trigonometric polynomial.

Exercise 5.3.2 Prove Lemma 5.3.5.

Exercise 5.3.3 Prove Corollary 5.3.6. (Hint: use Lemma 5.3.5. For the second identity, either use Pythagoras' theorem and induction, or substitute $f = \sum_{n=-N}^{N} c_n e_n$ and expand everything out.)

5.4 Periodic Convolutions

The goal of this section is to prove the Weierstrass approximation theorem for trigonometric polynomials:

Theorem 5.4.1 *Let $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, and let $\varepsilon > 0$. Then there exists a trigonometric polynomial P such that $\|f - P\|_\infty \leq \varepsilon$.*

This theorem asserts that any continuous periodic function can be uniformly approximated by trigonometric polynomials. To put it another way, if we let $P(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ denote the space of all trigonometric polynomials, then the closure of $P(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ in the L^∞ metric is $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.

It is possible to prove this theorem directly from the Weierstrass approximation theorem for polynomials (Theorem 3.8.3), and both theorems are a special case of a much more general theorem known as the *Stone-Weierstrass theorem*, which we will not discuss here. However we shall instead prove this theorem from scratch, in order to introduce a couple of interesting notions, notably that of periodic convolution. The proof here, though, should strongly remind you of the arguments used to prove Theorem 3.8.3.

Definition 5.4.2 (*Periodic convolution*) Let $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$. Then we define the *periodic convolution* $f * g \colon \mathbf{R} \to \mathbf{C}$ of f and g by the formula

$$f * g(x) := \int_{[0,1]} f(y) g(x - y)\, dy.$$

Remark 5.4.3 Note that this formula is slightly different from the convolution for compactly supported functions defined in Definition 3.8.9, because we are only integrating over $[0, 1]$ and not on all of $\mathbf{R}$. Thus, in principle we have given the symbol $f * g$ two conflicting meanings. However, in practice there will be no confusion, because it is not possible for a non-zero function to both be periodic and compactly supported (Exercise 5.4.1).

Lemma 5.4.4 (Basic properties of periodic convolution) *Let $f, g, h \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.*

(a) (Closure) The convolution $f * g$ is continuous and $\mathbf{Z}$-periodic. In other words, $f * g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$.

(b) (Commutativity) We have $f * g = g * f$.

(c) (Bilinearity) We have $f * (g + h) = f * g + f * h$ and $(f + g) * h = f * h + g * h$. For any complex number c, we have $c(f * g) = (cf) * g = f * (cg)$.

Proof See Exercise 5.4.2. □

Now we observe an interesting identity: for any $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ and any integer n, we have

$$f * e_n = \hat{f}(n) e_n.$$

To prove this, we compute

$$f * e_n(x) = \int_{[0,1]} f(y) e^{2\pi i n(x-y)}\, \mathrm{d}y$$

$$= e^{2\pi i n x} \int_{[0,1]} f(y) e^{-2\pi i n y}\, \mathrm{d}y = \hat{f}(n) e^{2\pi i n x} = \hat{f}(n) e_n$$

as desired.

More generally, we see from Lemma 5.4.4(iii) that for any trigonometric polynomial $P = \sum_{n=-N}^{n=N} c_n e_n$, we have

$$f * P = \sum_{n=-N}^{n=N} c_n (f * e_n) = \sum_{n=-N}^{n=N} \hat{f}(n) c_n e_n.$$

Thus the periodic convolution of any function in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ with a trigonometric polynomial, is again a trigonometric polynomial. (Compare with Lemma 3.8.13.)

Next, we introduce the periodic analogue of an approximation to the identity.

Definition 5.4.5 (*Periodic approximation to the identity*) Let $\varepsilon > 0$ and $0 < \delta < 1/2$. A function $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ is said to be a *periodic* (ε, δ) *approximation to the identity* if the following properties are true:

(a) $f(x) \geq 0$ for all $x \in \mathbf{R}$, and $\int_{[0,1]} f = 1$.
(b) We have $f(x) < \varepsilon$ for all $\delta \leq |x| \leq 1 - \delta$.

Now we have an analogue of Lemma 3.8.8:

Lemma 5.4.6 *For every $\varepsilon > 0$ and $0 < \delta < 1/2$, there exists a trigonometric polynomial P which is an (ε, δ) approximation to the identity.*

Proof We sketch the proof of this Lemma here, and leave the completion of it to Exercise 5.4.3. Let $N \geq 1$ be an integer. We define the *Fejér kernel* F_N to be the function

$$F_N = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) e_n.$$

Clearly F_N is a trigonometric polynomial. We observe the identity

$$F_N = \frac{1}{N}\left|\sum_{n=0}^{N-1} e_n\right|^2$$

(why?). But from the geometric series formula (Lemma 7.3.3) we have

$$\sum_{n=0}^{N-1} e_n(x) = \frac{e_N - e_0}{e_1 - e_0} = \frac{e^{\pi i(N-1)x}\sin(\pi N x)}{\sin(\pi x)}$$

when x is not an integer, (why?) and hence we have the formula

$$F_N(x) = \frac{\sin(\pi N x)^2}{N\sin(\pi x)^2}.$$

When x is an integer, the geometric series formula does not apply, but one has $F_N(x) = N$ in that case, as one can see by direct computation. In either case we see that $F_N(x) \geq 0$ for any x. Also, we have

$$\int_{[0,1]} F_N(x)\, \mathrm{d}x = \sum_{n=-N}^{N}\left(1 - \frac{|n|}{N}\right)\int_{[0,1]} e_n = \left(1 - \frac{|0|}{N}\right)1 = 1$$

(why?). Finally, since $\sin(\pi N x) \leq 1$, we have

$$F_N(x) \leq \frac{1}{N\sin(\pi x)^2} \leq \frac{1}{N\sin(\pi\delta)^2}$$

whenever $\delta < |x| < 1 - \delta$ (this is because sin is increasing on $[0, \pi/2]$ and decreasing on $[\pi/2, \pi]$). Thus by choosing N large enough, we can make $F_N(x) \leq \varepsilon$ for all $\delta < |x| < 1 - \delta$. □

Proof of Theorem 5.4.1 Let f be any element of $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$; we know that f is bounded, so that we have some $M > 0$ such that $|f(x)| \leq M$ for all $x \in \mathbf{R}$.

Let $\varepsilon > 0$ be arbitrary. Since f is uniformly continuous, there exists a $\delta > 0$ such that $|f(x) - f(y)| \leq \varepsilon$ whenever $|x - y| \leq \delta$. Now use Lemma 5.4.6 to find a trigonometric polynomial P which is a (ε, δ) approximation to the identity. Then $f * P$ is also a trigonometric polynomial. We now estimate $\|f - f * P\|_\infty$.

Let x be any real number. We have

$$\begin{aligned}
|f(x) - f * P(x)| &= |f(x) - P * f(x)| \\
&= \left| f(x) - \int_{[0,1]} f(x-y)P(y)\,\mathrm{d}y \right| \\
&= \left| \int_{[0,1]} f(x)P(y)\,\mathrm{d}y - \int_{[0,1]} f(x-y)P(y)\,\mathrm{d}y \right| \\
&= \left| \int_{[0,1]} (f(x) - f(x-y))P(y)\,\mathrm{d}y \right| \\
&\leq \int_{[0,1]} |f(x) - f(x-y)|P(y)\,\mathrm{d}y.
\end{aligned}$$

The right-hand side can be split as

$$\begin{aligned}
\int_{[0,\delta]} |f(x) - f(x-y)|P(y)\,\mathrm{d}y &+ \int_{[\delta,1-\delta]} |f(x) - f(x-y)|P(y)\,\mathrm{d}y \\
&+ \int_{[1-\delta,1]} |f(x) - f(x-y)|P(y)\,\mathrm{d}y
\end{aligned}$$

which we can bound from above by

$$\begin{aligned}
&\leq \int_{[0,\delta]} \varepsilon P(y)\,\mathrm{d}y + \int_{[\delta,1-\delta]} 2M\varepsilon\,\mathrm{d}y \\
&\quad + \int_{[1-\delta,1]} |f(x-1) - f(x-y)|P(y)\,\mathrm{d}y \\
&\leq \int_{[0,\delta]} \varepsilon P(y)\,\mathrm{d}y + \int_{[\delta,1-\delta]} 2M\varepsilon\,\mathrm{d}y + \int_{[1-\delta,1]} \varepsilon P(y)\,\mathrm{d}y \\
&\leq \varepsilon + 2M\varepsilon + \varepsilon \\
&= (2M+2)\varepsilon.
\end{aligned}$$

Thus we have $\|f - f * P\|_\infty \leq (2M+2)\varepsilon$. Since M is fixed and ε is arbitrary, we can thus make $f * P$ arbitrarily close to f in sup norm, which proves the periodic Weierstrass approximation theorem. □

— Exercise —

Exercise 5.4.1 Show that if $f\colon \mathbf{R} \to \mathbf{C}$ is both compactly supported and $\mathbf{Z}$-periodic, then it is identically zero.

Exercise 5.4.2 Prove Lemma 5.4.4. (Hint: to prove that $f * g$ is continuous, you will have to do something like use the fact that f is bounded, and g is uniformly continuous, or vice versa. To prove that $f * g = g * f$, you will need to use the periodicity to "cut and paste" the interval [0, 1].)

Exercise 5.4.3 Fill in the gaps marked (why?) in Lemma 5.4.6. (Hint: for the first identity, use the identities $|z|^2 = z\overline{z}$, $\overline{e_n} = e_{-n}$, and $e_n e_m = e_{n+m}$.)

5.5 The Fourier and Plancherel Theorems

Using the Weierstrass approximation theorem (Theorem 5.4.1), we can now generalize the Fourier and Plancherel identities to arbitrary continuous periodic functions.

Theorem 5.5.1 (Fourier theorem) *For any $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, the series $\sum_{n=-\infty}^{\infty} \hat{f}(n)e_n$ converges in L^2 metric to f. In other words, we have*

$$\lim_{N\to\infty} \left\| f - \sum_{n=-N}^{N} \hat{f}(n)e_n \right\|_2 = 0.$$

Proof Let $\varepsilon > 0$. We have to show that there exists an N_0 such that $\|f - \sum_{n=-N}^{N} \hat{f}(n)e_n\|_2 \leq \varepsilon$ for all sufficiently large N.

By the Weierstrass approximation theorem (Theorem 5.4.1), we can find a trigonometric polynomial $P = \sum_{n=-N_0}^{N_0} c_n e_n$ such that $\|f - P\|_\infty \leq \varepsilon$, for some $N_0 > 0$. In particular we have $\|f - P\|_2 \leq \varepsilon$.

Now let $N > N_0$, and let $F_N := \sum_{n=-N}^{n=N} \hat{f}(n)e_n$. We claim that $\|f - F_N\|_2 \leq \varepsilon$. First observe that for any $|m| \leq N$, we have

$$\langle f - F_N, e_m\rangle = \langle f, e_m\rangle - \sum_{n=-N}^{N} \hat{f}(n)\langle e_n, e_m\rangle = \hat{f}(m) - \hat{f}(m) = 0,$$

where we have used Lemma 5.3.5. In particular we have

$$\langle f - F_N, F_N - P\rangle = 0$$

since we can write $F_N - P$ as a linear combination of the e_m for which $|m| \leq N$. By Pythagoras' theorem we therefore have

$$\|f - P\|_2^2 = \|f - F_N\|_2^2 + \|F_N - P\|_2^2$$

and in particular

$$\|f - F_N\|_2 \leq \|f - P\|_2 \leq \varepsilon$$

as desired. □

Remark 5.5.2 Note that we have only obtained convergence of the Fourier series $\sum_{n=-\infty}^{\infty} \hat{f}(n) e_n$ to f in the L^2 metric. One may ask whether one has convergence in the uniform or pointwise sense as well, but it turns out (perhaps somewhat surprisingly) that the answer is no to both of those questions. However, if one assumes that the function f is not only continuous, but is also differentiable, then one can recover pointwise convergence; if one assumes continuously differentiable, then one gets uniform convergence as well. These results are beyond the scope of this text and will not be proven here. However, we will prove one theorem about when one can improve the L^2 convergence to uniform convergence.

Theorem 5.5.3 *Let $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, and suppose that the series $\sum_{n=-\infty}^{\infty} |\hat{f}(n)|$ is absolutely convergent. Then the series $\sum_{n=-\infty}^{\infty} \hat{f}(n) e_n$ converges uniformly to f. In other words, we have*

$$\lim_{N\to\infty} \left\| f - \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_{\infty} = 0.$$

Proof By the Weierstrass M-test (Theorem 3.5.7), we see that $\sum_{n=-\infty}^{\infty} \hat{f}(n) e_n$ converges to *some* function F, which by Lemma 5.1.5(iii) is also continuous and $\mathbf{Z}$-periodic. (Strictly speaking, the Weierstrass M test was phrased for series from $n = 1$ to $n = \infty$, but also works for series from $n = -\infty$ to $n = +\infty$; this can be seen by splitting the doubly infinite series into two pieces.) Thus

$$\lim_{N\to\infty} \left\| F - \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_{\infty} = 0$$

which implies that

$$\lim_{N\to\infty} \left\| F - \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_{2} = 0$$

since the L^2 norm is always less than or equal to the L^∞ norm. But the sequence $\sum_{n=-N}^{N} \hat{f}(n) e_n$ is already converging in L^2 metric to f by the Fourier theorem, so can only converge in L^2 metric to F if $F = f$ (cf. Proposition 1.1.20). Thus $F = f$, and so we have

$$\lim_{N\to\infty} \left\| f - \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_{\infty} = 0$$

as desired. □

As a corollary of the Fourier theorem, we obtain

Theorem 5.5.4 (Plancherel theorem) *For any $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, the series $\sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2$ is absolutely convergent, and*

$$\|f\|_2^2 = \sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2.$$

This theorem is also known as *Parseval's theorem.*

Proof Let $\varepsilon > 0$. By the Fourier theorem we know that

$$\left\| f - \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_2 \leq \varepsilon$$

if N is large enough (depending on ε). In particular, by the triangle inequality this implies that

$$\|f\|_2 - \varepsilon \leq \left\| \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_2 \leq \|f\|_2 + \varepsilon.$$

On the other hand, by Corollary 5.3.6 we have

$$\left\| \sum_{n=-N}^{N} \hat{f}(n) e_n \right\|_2 = \left(\sum_{n=-N}^{N} |\hat{f}(n)|^2 \right)^{1/2}$$

and hence

$$(\|f\|_2 - \varepsilon)^2 \leq \sum_{n=-N}^{N} |\hat{f}(n)|^2 \leq (\|f\|_2 + \varepsilon)^2.$$

Taking lim sup, we obtain

$$(\|f\|_2 - \varepsilon)^2 \leq \limsup_{N \to \infty} \sum_{n=-N}^{N} |\hat{f}(n)|^2 \leq (\|f\|_2 + \varepsilon)^2.$$

Since ε is arbitrary, we thus obtain by the squeeze test that

$$\limsup_{N \to \infty} \sum_{n=-N}^{N} |\hat{f}(n)|^2 = \|f\|_2^2$$

and the claim follows. □

There are many other properties of the Fourier transform, but we will not develop them here. In the exercises you will see a small number of applications of the Fourier and Plancherel theorems.

— Exercise —

Exercise 5.5.1 Let f be a function in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, and define the *trigonometric Fourier coefficients* a_n, b_n for $n = 0, 1, 2, 3, \ldots$ by

$$a_n{:=}2 \int_{[0,1]} f(x)\cos(2\pi nx)\,\mathrm{d}x; \quad b_n{:=}2 \int_{[0,1]} f(x)\sin(2\pi nx)\,\mathrm{d}x.$$

(a) Show that the series

$$\frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n\cos(2\pi nx) + b_n\sin(2\pi nx))$$

converges in L^2 metric to f. (*Hint:* use the Fourier theorem, and break up the exponentials into sines and cosines. Combine the positive n terms with the negative n terms.)

(b) Show that if $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ are absolutely convergent, then the above series actually converges uniformly to f, and not just in L^2 metric. (*Hint:* use Theorem 5.5.3.)

Exercise 5.5.2 Let $f(x)$ be the function defined by $f(x) = (1-2x)^2$ when $x \in [0, 1)$, and extended to be $\mathbf{Z}$-periodic for the rest of the real line.

(a) Using Exercise 5.5.1, show that the series

$$\frac{1}{3} + \sum_{n=1}^{\infty}\frac{4}{\pi^2 n^2}\cos(2\pi nx)$$

converges uniformly to f.

(b) Conclude that $\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$. (Hint: evaluate the above series at $x = 0$.)

(c) Conclude that $\sum_{n=1}^{\infty}\frac{1}{n^4} = \frac{\pi^4}{90}$. (Hint: expand the cosines in terms of exponentials, and use Plancherel's theorem.)

Exercise 5.5.3 If $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ and P is a trigonometric polynomial, show that

$$\widehat{f * P}(n) = \hat{f}(n)c_n = \hat{f}(n)\hat{P}(n)$$

for all integers n. More generally, if $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, show that

$$\widehat{f * g}(n) = \hat{f}(n)\hat{g}(n)$$

for all integers n. (A fancy way of saying this is that the Fourier transform *intertwines* convolution and multiplication.)

Exercise 5.5.4 Let $f \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$ be a function which is differentiable, and whose derivative f' is also continuous (where we define derivatives of complex-valued functions in exactly the same way as for their real-valued counterparts). Show that f' also lies in $C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$, and that $\widehat{f'}(n) = 2\pi in\hat{f}(n)$ for all integers n.

Exercise 5.5.5 Let $f, g \in C(\mathbf{R}/\mathbf{Z}; \mathbf{C})$. Prove the *Parseval identity*

$$\Re \int_0^1 f(x)\overline{g(x)}\, \mathrm{d}x = \Re \sum_{n\in\mathbf{Z}} \hat{f}(n)\overline{\hat{g}(n)}.$$

(Hint: apply the Plancherel theorem to $f + g$ and $f - g$, and subtract the two.) Then conclude that the real parts can be removed, thus

$$\int_0^1 f(x)\overline{g(x)}\, \mathrm{d}x = \sum_{n\in\mathbf{Z}} \hat{f}(n)\overline{\hat{g}(n)}.$$

(Hint: apply the first identity with f replaced by if.)

Exercise 5.5.6 In this exercise we shall develop the theory of Fourier series for functions of any fixed period L.

Let $L > 0$, and let $f : \mathbf{R} \to \mathbf{C}$ be a complex-valued function which is continuous and L-periodic. Define the numbers c_n for every integer n by

$$c_n := \frac{1}{L} \int_{[0,L]} f(x)e^{-2\pi inx/L}\, \mathrm{d}x.$$

(a) Show that the series

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi inx/L}$$

converges in L^2 metric to f. More precisely, show that

$$\lim_{N\to\infty} \int_{[0,L]} |f(x) - \sum_{n=-N}^{N} c_n e^{2\pi inx/L}|^2\, \mathrm{d}x = 0.$$

(Hint: apply the Fourier theorem to the function $f(Lx)$.)

(b) If the series $\sum_{n=-\infty}^{\infty} |c_n|$ is absolutely convergent, show that

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi inx/L}$$

converges uniformly to f.

(c) Show that

$$\frac{1}{L}\int_{[0,L]} |f(x)|^2 \, \mathrm{d}x = \sum_{n=-\infty}^{\infty} |c_n|^2.$$

(Hint: apply the Plancherel theorem to the function $f(Lx)$.)

Chapter 6
Several Variable Differential Calculus

6.1 Linear Transformations

We shall now switch to a different topic, namely that of differentiation in several variable calculus. More precisely, we shall be dealing with maps $f : \mathbf{R}^n \to \mathbf{R}^m$ from one Euclidean space to another, and trying to understand what the derivative of such a map is.

Before we do so, however, we need to recall some notions from linear algebra, most importantly that of a linear transformation and a matrix. We shall be rather brief here; a more thorough treatment of this material can be found in any linear algebra text.

Definition 6.1.1 (*Row vectors*) Let $n \geq 1$ be an integer. We refer to elements of $\mathbf{R}^n$ as *n-dimensional row vectors*. A typical n-dimensional row vector may take the form $x = (x_1, x_2, \ldots, x_n)$, which we abbreviate as $(x_i)_{1 \leq i \leq n}$; the quantities $x_1, x_2, \ldots, x_n$ are of course real numbers. If $(x_i)_{1 \leq i \leq n}$ and $(y_i)_{1 \leq i \leq n}$ are n-dimensional row vectors, we can define their vector sum by

$$(x_i)_{1 \leq i \leq n} + (y_i)_{1 \leq i \leq n} = (x_i + y_i)_{1 \leq i \leq n},$$

and also if $c \in \mathbf{R}$ is any scalar, we can define the scalar product $c(x_i)_{1 \leq i \leq n}$ by

$$c(x_i)_{1 \leq i \leq n} := (cx_i)_{1 \leq i \leq n}.$$

Of course one has similar operations on $\mathbf{R}^m$ as well. However, if $n \neq m$, then we do not define any operation of vector addition between vectors in $\mathbf{R}^n$ and vectors in $\mathbf{R}^m$ (e.g., $(2, 3, 4) + (5, 6)$ is undefined). We also refer to the vector $(0, \ldots, 0)$ in $\mathbf{R}^n$ as the *zero vector* and also denote it by 0. (Strictly speaking, we should denote the zero vector of $\mathbf{R}^n$ by $0_{\mathbf{R}^n}$, as they are technically distinct from each other and from the number zero, but we shall not take care to make this distinction.) We abbreviate $(-1)x$ as $-x$.

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_6

The operations of vector addition and scalar multiplication obey a number of basic properties:

Lemma 6.1.2 ($\mathbf{R}^n$ is a vector space) *Let x, y, z be vectors in $\mathbf{R}^n$, and let c, d be real numbers. Then we have the commutativity property $x + y = y + x$, the additive associativity property $(x + y) + z = x + (y + z)$, the additive identity property $x + 0 = 0 + x = x$, the additive inverse property $x + (-x) = (-x) + x = 0$, the multiplicative associativity property $(cd)x = c(dx)$, the distributivity properties $c(x + y) = cx + cy$ and $(c + d)x = cx + dx$, and the multiplicative identity property $1x = x$.*

Proof See Exercise 6.1.1. □

Definition 6.1.3 (*Transpose*) If $(x_i)_{1\leq i\leq n} = (x_1, x_2, \ldots, x_n)$ is an n-dimensional row vector, we can define its *transpose* $(x_i)^T_{1\leq i\leq n}$ by

$$(x_i)^T_{1\leq i\leq n} = (x_1, x_2, \ldots, x_n)^T := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

We refer to objects such as $(x_i)^T_{1\leq i\leq n}$ as *n-dimensional column vectors.*

Remark 6.1.4 There is no functional difference between a row vector and a column vector (e.g., one can add and scalar multiply column vectors just as well as we can row vectors); however we shall (rather annoyingly) need to transpose our row vectors into column vectors in order to be consistent with the conventions of matrix multiplication, which we will see later. Note that we view row vectors and column vectors as residing in different spaces; thus for instance we will not define the sum of a row vector with a column vector, even when they have the same number of elements.

Definition 6.1.5 (*Standard basis row vectors*) We identify n special vectors in $\mathbf{R}^n$, the *standard basis row vectors* $e_1, \ldots, e_n$. For each $1 \leq j \leq n$, e_j is the vector which has 0 in all entries except for the j-th entry, which is equal to 1.

For instance, in $\mathbf{R}^3$, we have $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. Note that if $x = (x_j)_{1\leq j\leq n}$ is a vector in $\mathbf{R}^n$, then

$$x = x_1e_1 + x_2e_2 + \ldots + x_ne_n = \sum_{j=1}^{n} x_je_j,$$

or in other words every vector in $\mathbf{R}^n$ is a *linear combination* of the standard basis vectors $e_1, \ldots, e_n$. (The notation $\sum_{j=1}^n x_je_j$ is unambiguous because the operation of vector addition is both commutative and associative). Of course, just as every row vector is a linear combination of standard basis row vectors, every column vector is a linear combination of standard basis column vectors:

$$x^T = x_1 e_1^T + x_2 e_2^T + \ldots + x_n e_n^T = \sum_{j=1}^{n} x_j e_j^T .$$

There are (many) other ways to create a basis for $\mathbf{R}^n$, but this is a topic for a linear algebra text and will not be discussed here.

Definition 6.1.6 (*Linear transformations*) A *linear transformation* $T\colon \mathbf{R}^n \to \mathbf{R}^m$ is any function from one Euclidean space $\mathbf{R}^n$ to another $\mathbf{R}^m$ which obeys the following two axioms:

(a) (Additivity) For every $x, x' \in \mathbf{R}^n$, we have $T(x + x') = Tx + Tx'$.
(b) (Homogeneity) For every $x \in \mathbf{R}^n$ and every $c \in \mathbf{R}$, we have $T(cx) = cTx$.

Example 6.1.7 The *dilation operator* $T_1\colon \mathbf{R}^3 \to \mathbf{R}^3$ defined by $T_1 x{:=}5x$ (i.e., it dilates each vector x by a factor of 5) is a linear transformation, since $5(x + x') = 5x + 5x'$ for all $x, x' \in \mathbf{R}^3$ and $5(cx) = c(5x)$ for all $x \in \mathbf{R}^3$ and $c \in \mathbf{R}$.

Example 6.1.8 The *rotation operator* $T_2\colon \mathbf{R}^2 \to \mathbf{R}^2$ defined by a counterclockwise rotation by $\pi/2$ radians around the origin (so that $T_2(1, 0) = (0, 1)$, $T_2(0, 1) = (-1, 0)$, etc.) is a linear transformation; this can best be seen geometrically rather than analytically.

Example 6.1.9 The *projection operator* $T_3 : \mathbf{R}^3 \to \mathbf{R}^2$ defined by $T_3(x, y, z){:=}(x, y)$ is a linear transformation (why?). The *inclusion operator* $T_4 : \mathbf{R}^2 \to \mathbf{R}^3$ defined by $T_4(x, y){:=}(x, y, 0)$ is also a linear transformation (why?). Finally, the *identity operator* $I_n : \mathbf{R}^n \to \mathbf{R}^n$, defined for any n by $I_n x{:=}x$ is also a linear transformation (why?).

As we shall shortly see, there is a connection between linear transformations and matrices.

Definition 6.1.10 (*Matrices*) An *$m \times n$ matrix* is an object A of the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix};$$

we shall abbreviate this as

$$A = (a_{ij})_{1 \le i \le m; 1 \le j \le n}.$$

In particular, n-dimensional row vectors are $1 \times n$ matrices, while n-dimensional column vectors are $n \times 1$ matrices.

Definition 6.1.11 (*Matrix product*) Given an $m \times n$ matrix A and an $n \times p$ matrix B, we can define the *matrix product AB* to be the $m \times p$ matrix defined as

$$(a_{ij})_{1 \leq i \leq m; 1 \leq j \leq n}(b_{jk})_{1 \leq j \leq n; 1 \leq k \leq p} := \left(\sum_{j=1}^{n} a_{ij} b_{jk} \right)_{1 \leq i \leq m; 1 \leq k \leq p}.$$

In particular, if $x^T = (x_j)^T_{1 \leq j \leq n}$ is an n-dimensional column vector, and $A = (a_{ij})_{1 \leq i \leq m; 1 \leq j \leq n}$ is an $m \times n$ matrix, then Ax^T is an m-dimensional column vector:

$$Ax^T = \left(\sum_{j=1}^{n} a_{ij} x_j \right)^T_{1 \leq i \leq m}.$$

We now relate matrices to linear transformations. If A is an $m \times n$ matrix, we can define the transformation $L_A : \mathbf{R}^n \to \mathbf{R}^m$ by the formula

$$(L_A x)^T := Ax^T.$$

Example 6.1.12 If A is the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

and $x = (x_1, x_2, x_3)$ is a 3-dimensional row vector, then $L_A x$ is the 2-dimensional row vector defined by

$$(L_A x)^T = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 + 3x_3 \\ 4x_1 + 5x_2 + 6x_3 \end{pmatrix}$$

or in other words

$$L_A(x_1, x_2, x_3) = (x_1 + 2x_2 + 3x_3, 4x_1 + 5x_2 + 6x_3).$$

More generally, if

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

then we have

$$L_A(x_j)_{1\leq j\leq n} = \left(\sum_{j=1}^{n} a_{ij}x_j\right)_{1\leq i\leq m}.$$

For any $m \times n$ matrix A, the transformation L_A is automatically linear; one can easily verify that $L_A(x+y) = L_A x + L_A y$ and $L_A(cx) = c(L_A x)$ for any n-dimensional row vectors x, y and any scalar c. (Why?)

Perhaps surprisingly, the converse is also true, i.e., every linear transformation from $\mathbf{R}^n$ to $\mathbf{R}^m$ is given by a matrix:

Lemma 6.1.13 *Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. Then there exists exactly one $m \times n$ matrix A such that $T = L_A$.*

Proof Suppose $T\colon \mathbf{R}^n \to \mathbf{R}^m$ is a linear transformation. Let $e_1, e_2, \ldots, e_n$ be the standard basis row vectors of $\mathbf{R}^n$. Then $Te_1, Te_2, \ldots, Te_n$ are vectors in $\mathbf{R}^m$. For each $1 \leq j \leq n$, we write Te_j in co-ordinates as

$$Te_j = (a_{1j}, a_{2j}, \ldots, a_{mj}) = (a_{ij})_{1\leq i\leq m},$$

i.e., we define a_{ij} to be the i^{th} component of Te_j. Then for any n-dimensional row vector $x = (x_1, \ldots, x_n)$, we have

$$Tx = T\left(\sum_{j=1}^{n} x_j e_j\right),$$

which (since T is linear) is equal to

$$\begin{aligned}
&= \sum_{j=1}^{n} T(x_j e_j)\\
&= \sum_{j=1}^{n} x_j Te_j\\
&= \sum_{j=1}^{n} x_j (a_{ij})_{1\leq i\leq m}\\
&= \sum_{j=1}^{n} (a_{ij}x_j)_{1\leq i\leq m}\\
&= \left(\sum_{j=1}^{n} a_{ij}x_j\right)_{1\leq i\leq m}.
\end{aligned}$$

But if we let A be the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

then the previous vector is precisely $L_A x$. Thus $Tx = L_A x$ for all n-dimensional vectors x, and thus $T = L_A$.

Now we show that A is unique, i.e., there does not exist any other matrix

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}$$

for which T is equal to L_B. Suppose for sake of contradiction that we could find such a matrix B which was different from A. Then we would have $L_A = L_B$. In particular, we have $L_A e_j = L_B e_j$ for every $1 \leq j \leq n$. But from the definition of L_A we see that

$$L_A e_j = (a_{ij})_{1 \leq i \leq m}$$

and

$$L_B e_j = (b_{ij})_{1 \leq i \leq m}$$

and thus we have $a_{ij} = b_{ij}$ for every $1 \leq i \leq m$ and $1 \leq j \leq n$, thus A and B are equal, a contradiction. □

Remark 6.1.14 Lemma 6.1.13 establishes a one-to-one correspondence between linear transformations and matrices, and is one of the fundamental reasons why matrices are so important in linear algebra. One may ask then why we bother dealing with linear transformations at all, and why we don't just work with matrices all the time. The reason is that sometimes one does not want to work with the standard basis $e_1, \dots, e_n$, but instead wants to use some other basis. In that case, the correspondence between linear transformations and matrices changes, and so it is still important to keep the notions of linear transformation and matrix distinct. More discussion on this somewhat subtle issue can be found in any linear algebra text.

Remark 6.1.15 If $T = L_A$, then A is sometimes called the *matrix representation of* T and is sometimes denoted $A = [T]$. We shall avoid this notation here, however.

The composition $T \circ S$ of two linear transformations T, S is again a linear transformation (Exercise 6.1.2). It is customary in linear algebra to abbreviate such compositions $T \circ S$ simply as TS. The next lemma shows that the operation of composing linear transformations is connected to that of matrix multiplication.

Lemma 6.1.16 *Let A be an $m \times n$ matrix, and let B be an $n \times p$ matrix. Then $L_A L_B = L_{AB}$.*

Proof See Exercise 6.1.3. □

— Exercise —

Exercise 6.1.1 Prove Lemma 6.1.2.

Exercise 6.1.2 If $T : \mathbf{R}^n \to \mathbf{R}^m$ is a linear transformation, and $S : \mathbf{R}^p \to \mathbf{R}^n$ is a linear transformation, show that the composition $TS : \mathbf{R}^p \to \mathbf{R}^m$ of the two transforms, defined by $TS(x) := T(S(x))$, is also a linear transformation. (*Hint:* expand $TS(x + y)$ and $TS(cx)$ carefully, using plenty of parentheses.)

Exercise 6.1.3 Prove Lemma 6.1.16.

Exercise 6.1.4 Let $T : \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. Show that there exists a number $M > 0$ such that $\|Tx\| \leq M \|x\|$ for all $x \in \mathbf{R}^n$. (*Hint:* use Lemma 6.1.13 to write T in terms of a matrix A, and then set M to be the sum of the absolute values of all the entries in A. Use the triangle inequality often—it's easier than messing around with square roots, etc.) Conclude in particular that every linear transformation from $\mathbf{R}^n$ to $\mathbf{R}^m$ is continuous.

6.2 Derivatives in Several Variable Calculus

Now that we've reviewed some linear algebra, we turn now to our main topic of this chapter, which is that of understanding differentiation of functions of the form $f : \mathbf{R}^n \to \mathbf{R}^m$, i.e., functions from one Euclidean space to another. For instance, one might want to differentiate the function $f : \mathbf{R}^3 \to \mathbf{R}^4$ defined by

$$f(x, y, z) = (xy, yz, xz, xyz).$$

In single-variable calculus, when one wants to differentiate a function $f : E \to \mathbf{R}$ at a point x_0, where E is a subset of $\mathbf{R}$ that contains x_0, this is given by

$$f'(x_0) := \lim_{x \to x_0; x \in E \setminus \{x_0\}} \frac{f(x) - f(x_0)}{x - x_0}.$$

One could try to mimic this definition in the several variable case $f : E \to \mathbf{R}^m$, where E is now a subset of $\mathbf{R}^n$; however we encounter a difficulty in this case: the quantity $f(x) - f(x_0)$ will live in $\mathbf{R}^m$, and $x - x_0$ lives in $\mathbf{R}^n$, and we do not know how to divide an m-dimensional vector by an n-dimensional vector.

To get around this problem, we first rewrite the concept of derivative (in one dimension) in a way which does not involve division of vectors. Instead, we view differentiability at a point x_0 as an assertion that a function f is "approximately linear" near x_0.

Lemma 6.2.1 *Let E be a subset of $\mathbf{R}$, $f : E \to \mathbf{R}$ be a function, and $L \in \mathbf{R}$. Let x_0 be a limit point of E. Then the following two statements are equivalent.*

(a) f is differentiable at x_0, and $f'(x_0) = L$.
(b) We have $\lim_{x \to x_0; x \in E - \{x_0\}} \frac{|f(x) - (f(x_0) + L(x - x_0))|}{|x - x_0|} = 0$.

Proof See Exercise 6.2.1. □

In light of the above lemma, we see that the derivative $f'(x_0)$ can be interpreted as the number L for which $|f(x) - (f(x_0) + L(x - x_0))|$ is small, in the sense that it tends to zero as x tends to x_0, even if we divide out by the very small number $|x - x_0|$. More informally, the derivative is the quantity L such that we have the approximation $f(x) - f(x_0) \approx L(x - x_0)$.

This does not seem too different from the usual notion of differentiation, but the point is that we are no longer explicitly dividing by $x - x_0$. (We are still dividing by $|x - x_0|$, but this will turn out to be OK.) When we move to the several variable case $f : E \to \mathbf{R}^m$, where $E \subseteq \mathbf{R}^n$, we shall still want the derivative to be some quantity L such that $f(x) - f(x_0) \approx L(x - x_0)$. However, since $f(x) - f(x_0)$ is now an m-dimensional vector and $x - x_0$ is an n-dimensional vector, we no longer want L to be a scalar; we want it to be a linear transformation. More precisely:

Definition 6.2.2 (*Differentiability*) Let E be a subset of $\mathbf{R}^n$, $f : E \to \mathbf{R}^m$ be a function, $x_0 \in E$ be a limit point of E, and let $L : \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. We say that f is *differentiable at x_0 with derivative L* if we have

$$\lim_{x \to x_0; x \in E - \{x_0\}} \frac{\|f(x) - (f(x_0) + L(x - x_0))\|}{\|x - x_0\|} = 0.$$

Here $\|x\|$ is the length of x (as measured in the l^2 metric):

$$\|(x_1, x_2, \ldots, x_n)\| = (x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2}.$$

Example 6.2.3 Let $f : \mathbf{R}^2 \to \mathbf{R}^2$ be the map $f(x, y) := (x^2, y^2)$, let x_0 be the point $x_0 := (1, 2)$, and let $L : \mathbf{R}^2 \to \mathbf{R}^2$ be the map $L(x, y) := (2x, 4y)$. We claim that f is differentiable at x_0 with derivative L. To see this, we compute

$$\lim_{(x,y) \to (1,2) : (x,y) \neq (1,2)} \frac{\|f(x, y) - (f(1, 2) + L((x, y) - (1, 2)))\|}{\|(x, y) - (1, 2)\|}.$$

Making the change of variables $(x, y) = (1, 2) + (a, b)$, this becomes

$$\lim_{(a,b) \to (0,0) : (a,b) \neq (0,0)} \frac{\|f(1 + a, 2 + b) - (f(1, 2) + L(a, b))\|}{\|(a, b)\|}.$$

Substituting the formula for f and for L, this becomes

$$\lim_{(a,b) \to (0,0) : (a,b) \neq (0,0)} \frac{\|((1 + a)^2, (2 + b)^2) - (1, 4) - (2a, 4b))\|}{\|(a, b)\|},$$

which simplifies to

$$\lim_{(a,b)\to(0,0):(a,b)\neq(0,0)} \frac{\|(a^2,b^2)\|}{\|(a,b)\|}.$$

We use the squeeze test. The expression $\frac{\|(a^2,b^2)\|}{\|(a,b)\|}$ is clearly non-negative. On the other hand, we have by the triangle inequality

$$\|(a^2,b^2)\| \leq \|(a^2,0)\| + \|(0,b^2)\| = a^2 + b^2$$

and hence

$$\frac{\|(a^2,b^2)\|}{\|(a,b)\|} \leq \sqrt{a^2+b^2}.$$

Since $\sqrt{a^2+b^2} \to 0$ as $(a,b) \to 0$, we thus see from the squeeze test that the above limit exists and is equal to 0. Thus f is differentiable at x_0 with derivative L.

As you can see, verifying that a function is differentiable from first principles can be somewhat tedious. Later on we shall find better ways to verify differentiability, and to compute derivatives.

Before we proceed further, we have to check a basic fact, which is that a function can have at most one derivative at any *interior* point of its domain:

Lemma 6.2.4 (Uniqueness of derivatives) *Let E be a subset of $\mathbf{R}^n$, $f : E \to \mathbf{R}^m$ be a function, $x_0 \in E$ be an* interior point *of E, and let $L_1 : \mathbf{R}^n \to \mathbf{R}^m$ and $L_2 : \mathbf{R}^n \to \mathbf{R}^m$ be linear transformations. Suppose that f is differentiable at x_0 with derivative L_1, and also differentiable at x_0 with derivative L_2. Then $L_1 = L_2$.*

Proof See Exercise 6.2.2. □

Because of Lemma 6.2.4, we can now talk about *the* derivative of f at interior points x_0, and we will denote this derivative by $f'(x_0)$. Thus $f'(x_0)$ is the unique linear transformation from $\mathbf{R}^n$ to $\mathbf{R}^m$ such that

$$\lim_{x\to x_0; x\neq x_0} \frac{\|f(x) - (f(x_0) + f'(x_0)(x-x_0))\|}{\|x-x_0\|} = 0.$$

Informally, this means that the derivative $f'(x_0)$ is the linear transformation such that we have

$$f(x) - f(x_0) \approx f'(x_0)(x-x_0)$$

or equivalently

$$f(x) \approx f(x_0) + f'(x_0)(x-x_0)$$

(this is known as *Newton's approximation*; compare with Proposition 10.1.7).

Another consequence of Lemma 6.2.4 is that if you know that $f(x) = g(x)$ for all $x \in E$, and f, g are differentiable at x_0, then you also know that $f'(x_0) = g'(x_0)$ at every *interior* point of E. However, this is not necessarily true if x_0 is a boundary point of E; for instance, if E is just a single point $E = \{x_0\}$, merely knowing that $f(x_0) = g(x_0)$ does not imply that $f'(x_0) = g'(x_0)$. We will not deal with these boundary issues here and only compute derivatives on the interior of the domain.

We will sometimes refer to f' as the *total derivative* of f, to distinguish this concept from that of partial and directional derivatives below. The total derivative f is also closely related to the *derivative matrix* Df, which we shall define in the next section.

— Exercise —

Exercise 6.2.1 Prove Lemma 6.2.1.

Exercise 6.2.2 Prove Lemma 6.2.4. (*Hint:* prove by contradiction. If $L_1 \neq L_2$, then there exists a vector v such that $L_1 v \neq L_2 v$; this vector must be nonzero (why?). Now apply the definition of derivative, and try to specialize to the case where $x = x_0 + tv$ for some scalar t, to obtain a contradiction.)

6.3 Partial and Directional Derivatives

We now connect the notion of differentiability with that of partial and directional derivatives, which we now introduce.

Definition 6.3.1 (*Directional derivative*) Let E be a subset of $\mathbf{R}^n$, $f : E \to \mathbf{R}^m$ be a function, let x_0 be an interior point of E, and let v be a vector in $\mathbf{R}^n$. If the limit

$$\lim_{t\to 0; t>0, x_0+tv\in E} \frac{f(x_0 + tv) - f(x_0)}{t}$$

exists, we say that f is *differentiable in the direction v at x_0*, and we denote the above limit by $D_v f(x_0)$:

$$D_v f(x_0) := \lim_{t\to 0; t>0} \frac{f(x_0 + tv) - f(x_0)}{t}.$$

Remark 6.3.2 One should compare this definition with Definition 6.2.2. Note that we are dividing by a scalar t, rather than a vector, so this definition makes sense, and $D_v f(x_0)$ will be a vector in $\mathbf{R}^m$. It is sometimes possible to also define directional derivatives on the boundary of E, if the vector v is pointing in an “inward” direction (this generalizes the notion of left derivatives and right derivatives from single-variable calculus); but we will not pursue these matters here.

Example 6.3.3 If $f : \mathbf{R} \to \mathbf{R}$ is a function, then $D_{+1} f(x)$ is the same as the right derivative of $f(x)$ (if it exists), and similarly $D_{-1} f(x)$ is the same as the negative of the left derivative of $f(x)$ (if it exists).

Example 6.3.4 We use the function $f : \mathbf{R}^2 \to \mathbf{R}^2$ defined by $f(x, y) := (x^2, y^2)$ from before, and let $x_0 := (1, 2)$ and $v := (3, 4)$. Then

$$\begin{aligned} D_v f(x_0) &= \lim_{t \to 0; t>0} \frac{f(1+3t, 2+4t) - f(1, 2)}{t} \\ &= \lim_{t \to 0; t>0} \frac{(1+6t+9t^2, 4+16t+16t^2) - (1, 4)}{t} \\ &= \lim_{t \to 0; t>0} (6+9t, 16+16t) = (6, 16). \end{aligned}$$

Directional derivatives are connected with total derivatives as follows:

Lemma 6.3.5 *Let E be a subset of $\mathbf{R}^n$, $f : E \to \mathbf{R}^m$ be a function, x_0 be an interior point of E, and let v be a vector in $\mathbf{R}^n$. If f is differentiable at x_0, then f is also differentiable in the direction v at x_0, and*

$$D_v f(x_0) = f'(x_0)v.$$

Proof See Exercise 6.3.1. □

Remark 6.3.6 One consequence of this lemma is that total differentiability implies directional differentiability. However, the converse is not true; see Exercise 6.3.3.

Closely related to the concept of directional derivative is that of *partial derivative*:

Definition 6.3.7 (*Partial derivative*) Let E be a subset of $\mathbf{R}^n$, let $f : E \to \mathbf{R}^m$ be a function, let x_0 be an interior point of E, and let $1 \le j \le n$. Then the *partial derivative of f with respect to the x_j variable* at x_0, denoted $\frac{\partial f}{\partial x_j}(x_0)$, is defined by

$$\frac{\partial f}{\partial x_j}(x_0) := \lim_{t \to 0; t \neq 0, x_0 + te_j \in E} \frac{f(x_0 + te_j) - f(x_0)}{t} = \frac{\mathrm{d}}{\mathrm{d}t} f(x_0 + te_j)|_{t=0}$$

provided of course that the limit exists. (If the limit does not exist, we leave $\frac{\partial f}{\partial x_j}(x_0)$ undefined.)

We say that f is *continuously differentiable* if the partial derivatives $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$ exist and are continuous on E.

Informally, the partial derivative can be obtained by holding all the variables other than x_j fixed and then applying the single-variable calculus derivative in the x_j variable. Note that if f takes values in $\mathbf{R}^m$, then so will $\frac{\partial f}{\partial x_j}$. Indeed, if we write f in components as $f = (f_1, \ldots, f_m)$, it is easy to see (why?) that

$$\frac{\partial f}{\partial x_j}(x_0) = \left(\frac{\partial f_1}{\partial x_j}(x_0), \ldots, \frac{\partial f_m}{\partial x_j}(x_0) \right),$$

i.e., to differentiate a vector-valued function one just has to differentiate each of the components separately.

We sometimes replace the variables x_j in $\frac{\partial f}{\partial x_j}$ with other symbols. For instance, if we are dealing with the function $f(x, y) = (x^2, y^2)$, then we might refer to $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ instead of $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$. (In this case, $\frac{\partial f}{\partial x}(x, y) = (2x, 0)$ and $\frac{\partial f}{\partial y}(x, y) = (0, 2y)$.) One should caution however that one should only relabel the variables if it is absolutely clear which symbol refers to the first variable, which symbol refers to the second variable, etc.; otherwise one may become unintentionally confused. For instance, in the above example, the expression $\frac{\partial f}{\partial x}(x, x)$ is just $(2x, 0)$; however one may mistakenly compute

$$\frac{\partial f}{\partial x}(x, x) = \frac{\partial}{\partial x}(x^2, x^2) = (2x, 2x);$$

the problem here is that the symbol x is being used for more than just the first variable of f. (On the other hand, it is true that $\frac{\mathrm{d}}{\mathrm{d}x}f(x, x)$ is equal to $(2x, 2x)$; thus the operation of total differentiation $\frac{\mathrm{d}}{\mathrm{d}x}$ is not the same as that of partial differentiation $\frac{\partial}{\partial x}$.)

From Lemma 6.3.5 (and Proposition 9.5.3 from *Analysis I*), we know that if a function is differentiable at a point x_0, then all the partial derivatives $\frac{\partial f}{\partial x_j}$ exist at x_0, and that

$$\frac{\partial f}{\partial x_j}(x_0) = D_{e_j}f(x_0) = -D_{-e_j}f(x_0) = f'(x_0)e_j.$$

Also, if $v = (v_1, \dots, v_n) = \sum_j v_j e_j$, then we have

$$D_v f(x_0) = f'(x_0)\sum_j v_j e_j = \sum_j v_j f'(x_0)e_j$$

(since $f'(x_0)$ is linear) and thus

$$D_v f(x_0) = \sum_j v_j \frac{\partial f}{\partial x_j}(x_0).$$

Thus one can write directional derivatives in terms of partial derivatives, *provided that* the function is actually differentiable at that point.

Just because the partial derivatives exist at a point x_0, we cannot conclude that the function is differentiable there (Exercise 6.3.3). However, if we know that the partial derivatives not only exist, but are continuous, then we can in fact conclude differentiability, thanks to the following handy theorem:

Theorem 6.3.8 *Let E be a subset of $\mathbf{R}^n$, $f : E \to \mathbf{R}^m$ be a function, F be a subset of E, and x_0 be an interior point of F. If all the partial derivatives $\frac{\partial f}{\partial x_j}$ exist on F and are continuous at x_0, then f is differentiable at x_0, and the linear transformation $f'(x_0) : \mathbf{R}^n \to \mathbf{R}^m$ is defined by*

$$f'(x_0)(v_j)_{1\le j\le n} = \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(x_0).$$

Proof Let $L : \mathbf{R}^n \to \mathbf{R}^m$ be the linear transformation

$$L(v_j)_{1\leq j\leq n} := \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(x_0).$$

We have to prove that

$$\lim_{x\to x_0; x\in E-\{x_0\}} \frac{\|f(x) - (f(x_0) + L(x - x_0))\|}{\|x - x_0\|} = 0.$$

Let $\varepsilon > 0$. It will suffice to find a radius $\delta > 0$ such that

$$\frac{\|f(x) - (f(x_0) + L(x - x_0))\|}{\|x - x_0\|} \leq \varepsilon$$

for all $x \in B(x_0, \delta)\backslash\{x_0\}$. Equivalently, we wish to show that

$$\|f(x) - f(x_0) - L(x - x_0)\| \leq \varepsilon\|x - x_0\|$$

for all $x \in B(x_0, \delta)\backslash\{x_0\}$.

Because x_0 is an interior point of F, there exists a ball $B(x_0, r)$ which is contained inside F. Because each partial derivative $\frac{\partial f}{\partial x_j}$ exists on F and is continuous at x_0, there thus exists an $0 < \delta_j < r$ such that $\|\frac{\partial f}{\partial x_j}(x) - \frac{\partial f}{\partial x_j}(x_0)\| \leq \varepsilon/nm$ for every $x \in B(x_0, \delta_j)$. If we take $\delta = \min(\delta_1, \ldots, \delta_n)$, then we thus have $\|\frac{\partial f}{\partial x_j}(x) - \frac{\partial f}{\partial x_j}(x_0)\| \leq \varepsilon/nm$ for every $x \in B(x_0, \delta)$ and every $1 \leq j \leq n$.

Let $x \in B(x_0, \delta)$. We write $x = x_0 + v_1e_1 + v_2e_2 + \ldots + v_ne_n$ for some scalars $v_1, \ldots, v_n$. Note that

$$\|x - x_0\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$$

and in particular we have $|v_j| \leq \|x - x_0\|$ for all $1 \leq j \leq n$. Our task is to show that

$$\left\| f(x_0 + v_1e_1 + \ldots + v_ne_n) - f(x_0) - \sum_{j=1}^{n} v_j \frac{\partial f}{\partial x_j}(x_0) \right\| \leq \varepsilon\|x - x_0\|.$$

Write f in components as $f = (f_1, f_2, \ldots, f_m)$ (so each f_i is a function from E to $\mathbf{R}$). From the mean value theorem in the x_1 variable, we see that

$$f_i(x_0 + v_1e_1) - f_i(x_0) = \frac{\partial f_i}{\partial x_1}(x_0 + t_ie_1)v_1$$

for some t_i between 0 and v_1. But we have

$$\left|\frac{\partial f_i}{\partial x_j}(x_0+t_ie_1)-\frac{\partial f_i}{\partial x_j}(x_0)\right| \le \left\|\frac{\partial f}{\partial x_j}(x_0+t_ie_1)-\frac{\partial f}{\partial x_j}(x_0)\right\| \le \varepsilon/nm$$

and hence

$$\left|f_i(x_0+v_1e_1)-f_i(x_0)-\frac{\partial f_i}{\partial x_1}(x_0)v_1\right| \le \varepsilon|v_1|/nm.$$

Summing this over all $1 \le i \le m$ (and noting that $\|(y_1,\ldots,y_m)\| \le |y_1|+\ldots+|y_m|$ from the triangle inequality) we obtain

$$\left\|f(x_0+v_1e_1)-f(x_0)-\frac{\partial f}{\partial x_1}(x_0)v_1\right\| \le \varepsilon|v_1|/n;$$

since $|v_1| \le \|x-x_0\|$, we thus have

$$\left\|f(x_0+v_1e_1)-f(x_0)-\frac{\partial f}{\partial x_1}(x_0)v_1\right\| \le \varepsilon\|x-x_0\|/n.$$

A similar argument gives

$$\left\|f(x_0+v_1e_1+v_2e_2)-f(x_0+v_1e_1)-\frac{\partial f}{\partial x_2}(x_0)v_2\right\| \le \varepsilon\|x-x_0\|/n$$

and so forth up to

$$\|f(x_0+v_1e_1+\cdots+v_ne_n)-f(x_0+v_1e_1+\cdots+v_{n-1}e_{n-1})$$
$$-\frac{\partial f}{\partial x_n}(x_0)v_n\Big\| \le \varepsilon\,\|x-x_0\|\,/n.$$

If we sum these n inequalities and use the triangle inequality $\|x+y\| \le \|x\|+\|y\|$, we obtain a telescoping series which simplifies to

$$\left\|f(x_0+v_1e_1+\ldots+v_ne_n)-f(x_0)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(x_0)v_j\right\| \le \varepsilon\|x-x_0\|$$

as desired. □

From Theorem 6.3.8 and Lemma 6.3.5 we see that if the partial derivatives of a function $f : E \to \mathbf{R}^m$ exist and are continuous on some set F, then all the directional derivatives also exist at every interior point x_0 of F, and we have the formula

$$D_{(v_1,\ldots,v_n)}f(x_0)=\sum_{j=1}^{n}v_j\frac{\partial f}{\partial x_j}(x_0).$$

In particular, if $f: E \to \mathbf{R}$ is a real-valued function, and we define the *gradient* $\nabla f(x_0)$ of f at x_0 to be the n-dimensional row vector $\nabla f(x_0) := (\frac{\partial f}{\partial x_1}(x_0), \ldots, \frac{\partial f}{\partial x_n}(x_0))$, then we have the familiar formula

$$D_v f(x_0) = v \cdot \nabla f(x_0)$$

whenever x_0 is in the interior of the region where the gradient exists and is continuous.

More generally, if $f: E \to \mathbf{R}^m$ is a function taking values in $\mathbf{R}^m$, with $f = (f_1, \ldots, f_m)$, and x_0 is in the interior of the region where the partial derivatives of f exist and are continuous, then we have from Theorem 6.3.8 that

$$f'(x_0)(v_j)_{1 \le j \le n} = \sum_{j=1}^n v_j \frac{\partial f}{\partial x_j}(x_0)$$

$$= \left(\sum_{j=1}^n v_j \frac{\partial f_i}{\partial x_j}(x_0) \right)_{1 \le i \le m},$$

which we can rewrite as

$$L_{Df(x_0)}(v_j)_{1 \le j \le n}$$

where $Df(x_0)$ is the $m \times n$ matrix

$$Df(x_0) := \left(\frac{\partial f_i}{\partial x_j}(x_0) \right)_{1 \le i \le m; 1 \le j \le n}$$

$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \frac{\partial f_1}{\partial x_2}(x_0) & \ldots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \frac{\partial f_2}{\partial x_1}(x_0) & \frac{\partial f_2}{\partial x_2}(x_0) & \ldots & \frac{\partial f_2}{\partial x_n}(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \frac{\partial f_m}{\partial x_2}(x_0) & \ldots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix}.$$

Thus we have

$$(D_v f(x_0))^T = (f'(x_0)v)^T = Df(x_0)v^T.$$

The matrix $Df(x_0)$ is sometimes also called the *derivative matrix* or *differential matrix* of f at x_0 and is closely related to the total derivative $f'(x_0)$. One can also write Df as

$$Df(x_0) = \left(\frac{\partial f}{\partial x_1}(x_0)^T, \frac{\partial f}{\partial x_2}(x_0)^T, \ldots, \frac{\partial f}{\partial x_n}(x_0)^T \right),$$

i.e., each of the columns of $Df(x_0)$ is one of the partial derivatives of f, expressed as a column vector. Or one could write

$$Df(x_0) = \begin{pmatrix} \nabla f_1(x_0) \\ \nabla f_2(x_0) \\ \vdots \\ \nabla f_m(x_0) \end{pmatrix}$$

i.e., the rows of $Df(x_0)$ are the gradient of various components of f. In particular, if f is scalar-valued (i.e., $m = 1$), then Df is the same as ∇f.

Example 6.3.9 Let $f : \mathbf{R}^2 \to \mathbf{R}^2$ be the function $f(x, y) = (x^2 + xy, y^2)$. Then $\frac{\partial f}{\partial x} = (2x + y, 0)$ and $\frac{\partial f}{\partial y} = (x, 2y)$. Since these partial derivatives are continuous on $\mathbf{R}^2$, we see that f is differentiable on all of $\mathbf{R}^2$, and

$$Df(x, y) = \begin{pmatrix} 2x + y & x \\ 0 & 2y \end{pmatrix}.$$

Thus for instance, the directional derivative in the direction (v, w) is

$$D_{(v,w)}f(x, y) = ((2x + y)v + xw, 2yw).$$

— Exercise —

Exercise 6.3.1 Prove Lemma 6.3.5. (This will be similar to Exercise 6.2.1).

Exercise 6.3.2 Let E be a subset of $\mathbf{R}^n$, let $f : E \to \mathbf{R}^m$ be a function, let x_0 be an interior point of E, and let $1 \leq j \leq n$. Show that $\frac{\partial f}{\partial x_j}(x_0)$ exists if and only if $D_{e_j}f(x_0)$ and $D_{-e_j}f(x_0)$ exist and are negatives of each other (thus $D_{e_j}f(x_0) = -D_{-e_j}f(x_0)$); furthermore, one has $\frac{\partial f}{\partial x_j}(x_0) = D_{e_j}f(x_0)$ in this case.

Exercise 6.3.3 Let $f : \mathbf{R}^2 \to \mathbf{R}$ be the function defined by $f(x, y) := \frac{x^3}{x^2+y^2}$ when $(x, y) \neq (0, 0)$, and $f(0, 0) := 0$. Show that f is not differentiable at $(0, 0)$, despite being differentiable in every direction $v \in \mathbf{R}^2$ at $(0, 0)$. Explain why this does not contradict Theorem 6.3.8.

Exercise 6.3.4 Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a differentiable function such that $f'(x) = 0$ for all $x \in \mathbf{R}^n$. Show that f is constant. (*Hint*: you may use the mean value theorem or fundamental theorem of calculus for one-dimensional functions, but bear in mind that there is no direct analogue of these theorems for several variable functions. I would not advise proceeding via first principles.) For a tougher challenge, replace the domain $\mathbf{R}^n$ by an open connected subset Ω of $\mathbf{R}^n$.

6.4 The Several Variable Calculus Chain Rule

We are now ready to state the several variable calculus chain rule. Recall that if $f : X \to Y$ and $g : Y \to Z$ are two functions, then the composition $g \circ f : X \to Z$ is defined by $g \circ f(x) := g(f(x))$ for all $x \in X$.

Theorem 6.4.1 (Several variable calculus chain rule) *Let E be a subset of $\mathbf{R}^n$, and let F be a subset of $\mathbf{R}^m$. Let $f : E \to F$ be a function, and let $g : F \to \mathbf{R}^p$ be another function. Let x_0 be a point in the interior of E. Suppose that f is differentiable at x_0, and that $f(x_0)$ is in the interior of F. Suppose also that g is differentiable at $f(x_0)$. Then $g \circ f : E \to \mathbf{R}^p$ is also differentiable at x_0, and we have the formula*

$$(g \circ f)'(x_0) = g'(f(x_0))f'(x_0).$$

Proof See Exercise 6.4.3. □

One should compare this theorem with the single-variable chain rule, Theorem 10.1.15; indeed one can easily deduce the single-variable rule as a consequence of the several variable rule.

Intuitively, one can think of the several variable chain rule as follows. Let x be close to x_0. Then Newton's approximation asserts that

$$f(x) - f(x_0) \approx f'(x_0)(x - x_0)$$

and in particular $f(x)$ is close to $f(x_0)$. Since g is differentiable at $f(x_0)$, we see from Newton's approximation again that

$$g(f(x)) - g(f(x_0)) \approx g'(f(x_0))(f(x) - f(x_0)).$$

Combining the two, we obtain

$$g \circ f(x) - g \circ f(x_0) \approx g'(f(x_0))f'(x_0)(x - x_0)$$

which then should give $(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$. This argument however is rather imprecise; to make it more precise one needs to manipulate limits rigorously; see Exercise 6.4.3.

As a corollary of the chain rule and Lemma 6.1.16 (and Lemma 6.1.13), we see that

$$D(g \circ f)(x_0) = Dg(f(x_0))Df(x_0);$$

i.e., we can write the chain rule in terms of matrices and matrix multiplication, instead of in terms of linear transformations and composition.

Example 6.4.2 Let $f : \mathbf{R}^n \to \mathbf{R}$ and $g : \mathbf{R}^n \to \mathbf{R}$ be differentiable functions. We form the combined function $h : \mathbf{R}^n \to \mathbf{R}^2$ by defining $h(x) := (f(x), g(x))$. Now let $k : \mathbf{R}^2 \to \mathbf{R}$ be the multiplication function $k(a, b) := ab$. Note that

$$Dh(x_0) = \begin{pmatrix} \nabla f(x_0) \\ \nabla g(x_0) \end{pmatrix}$$

while

$$Dk(a, b) = (b, a)$$

(why?). By the chain rule, we thus see that

$$D(k \circ h)(x_0) = (g(x_0), f(x_0)) \begin{pmatrix} \nabla f(x_0) \\ \nabla g(x_0) \end{pmatrix} = g(x_0)\nabla f(x_0) + f(x_0)\nabla g(x_0).$$

But $k \circ h = fg$ (why?), and $D(fg) = \nabla(fg)$. We have thus proven the *product rule*

$$\nabla(fg) = g\nabla f + f\nabla g.$$

A similar argument gives the sum rule $\nabla(f + g) = \nabla f + \nabla g$, or the difference rule $\nabla(f - g) = \nabla f - \nabla g$, as well as the quotient rule (Exercise 6.4.4). As you can see, the several variable chain rule is quite powerful and can be used to deduce many other rules of differentiation.

We record one further useful application of the chain rule. Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. From Exercise 6.4.1 we observe that T is continuously differentiable at every point, and in fact $T'(x) = T$ for every x. (This equation may look a little strange, but perhaps it is easier to swallow if you view it in the form $\frac{\mathrm{d}}{\mathrm{d}x}(Tx) = T$.) Thus, for any differentiable function $f : E \to \mathbf{R}^n$, we see that $Tf : E \to \mathbf{R}^m$ is also differentiable, and hence by the chain rule

$$(Tf)'(x_0) = T(f'(x_0)).$$

This is a generalization of the single-variable calculus rule $(cf)' = c(f')$ for constant scalars c.

Another special case of the chain rule which is quite useful is the following: if $f : \mathbf{R}^n \to \mathbf{R}^m$ is some differentiable function, and $x_j : \mathbf{R} \to \mathbf{R}$ are differentiable functions for each $j = 1, \ldots n$, then

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x_1(t), x_2(t), \ldots, x_n(t)) = \sum_{j=1}^{n} x_j'(t) \frac{\partial f}{\partial x_j}(x_1(t), x_2(t), \ldots, x_n(t)).$$

(Why is this a special case of the chain rule?).

— Exercise —

Exercise 6.4.1 Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. Show that T is continuously differentiable at every point, and in fact $T'(x) = T$ for every x. What is DT?

Exercise 6.4.2 Let E be a subset of $\mathbf{R}^n$. Prove that if a function $f: E \to \mathbf{R}^m$ is differentiable at an interior point x_0 of E, then it is also continuous at x_0. (*Hint:* use Exercise 6.1.4.)

Exercise 6.4.3 Prove Theorem 6.4.1. (*Hint:* you may wish to review the proof of the ordinary chain rule in single-variable calculus, Theorem 10.1.15. The easiest way to proceed is by using the sequence-based definition of limit (see Proposition 3.1.5(b)), and use Exercise 6.1.4.)

Exercise 6.4.4 State and prove some version of the quotient rule for functions of several variables (i.e., functions of the form $f : E \to \mathbf{R}$ for some subset E of $\mathbf{R}^n$). In other words, state a rule which gives a formula for the gradient of f/g; compare your answer with Theorem 10.1.13(h). Be sure to make clear what all your assumptions are.

Exercise 6.4.5 Let $\mathbf{x}: \mathbf{R} \to \mathbf{R}^3$ be a differentiable function, and let $r : \mathbf{R} \to \mathbf{R}$ be the function $r(t) := \|\mathbf{x}(t)\|$, where $\|\mathbf{x}\|$ denotes the length of $\mathbf{x}$ as measured in the usual l^2 metric. Let t_0 be a real number. Show that if $r(t_0) \neq 0$, then r is differentiable at t_0, and

$$r'(t_0) = \frac{\mathbf{x}'(t_0) \cdot \mathbf{x}(t_0)}{r(t_0)}.$$

(*Hint*: use Theorem 6.4.1.)

6.5 Double Derivatives and Clairaut's Theorem

We now investigate what happens if one differentiates a function twice.

Definition 6.5.1 (*Twice continuous differentiability*) Let E be an open subset of $\mathbf{R}^n$, and let $f : E \to \mathbf{R}^m$ be a function. We say that f is *twice continuously differentiable* if it is continuously differentiable, and the partial derivatives $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$ are themselves continuously differentiable.

Remark 6.5.2 Continuously differentiable functions are sometimes called C^1 functions; twice continuously differentiable functions are sometimes called C^2 functions. One can also define C^3, C^4, etc., but we shall not do so here.

Example 6.5.3 Let $f: \mathbf{R}^2 \to \mathbf{R}^2$ be the function $f(x, y) = (x^2 + xy, y^2)$. Then f is continuously differentiable because the partial derivatives $\frac{\partial f}{\partial x}(x, y) = (2x + y, 0)$ and $\frac{\partial f}{\partial y}(x, y) = (x, 2y)$ exist and are continuous on all of $\mathbf{R}^2$. It is also twice continuously differentiable, because the double partial derivatives $\frac{\partial}{\partial x}\frac{\partial f}{\partial x}(x, y) = (2, 0)$, $\frac{\partial}{\partial y}\frac{\partial f}{\partial x}(x, y) = (1, 0)$, $\frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x, y) = (1, 0)$, $\frac{\partial}{\partial y}\frac{\partial f}{\partial y}(x, y) = (0, 2)$ all exist and are continuous.

Observe in the above example that the double derivatives $\frac{\partial}{\partial y}\frac{\partial f}{\partial x}$ and $\frac{\partial}{\partial x}\frac{\partial f}{\partial y}$ are the same. This is in fact a general phenomenon:

Theorem 6.5.4 (Clairaut's theorem) *Let E be an open subset of $\mathbf{R}^n$, and let $f : E \to \mathbf{R}^m$ be a twice continuously differentiable function on E. Then we have $\frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(x_0) = \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}(x_0)$ for all $1 \leq i, j \leq n$.*

Proof By working with one component of f at a time we can assume that $m = 1$. The claim is trivial if $i = j$, so we shall assume that $i \neq j$. We shall prove the theorem for $x_0 = 0$; the general case is similar. (Actually, once one proves Clairaut's theorem for $x_0 = 0$, one can immediately obtain it for general x_0 by applying the theorem with $f(x)$ replaced by $f(x + x_0)$.)

Let a be the number $a := \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(0)$, and a' denote the quantity $a' := \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}(0)$. Our task is to show that $a' = a$.

Let $\varepsilon > 0$. Because the double derivatives of f are continuous, we can find a $\delta > 0$ such that

$$\left|\frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(x) - a\right| \leq \varepsilon$$

and

$$\left|\frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}(x) - a'\right| \leq \varepsilon$$

whenever $\|x\| \leq 2\delta$.

Now we consider the quantity

$$X := f(\delta e_i + \delta e_j) - f(\delta e_i) - f(\delta e_j) + f(0).$$

From the fundamental theorem of calculus in the e_i variable, we have

$$f(\delta e_i + \delta e_j) - f(\delta e_j) = \int_0^\delta \frac{\partial f}{\partial x_i}(x_i e_i + \delta e_j)\, \mathrm{d}x_i$$

and

$$f(\delta e_i) - f(0) = \int_0^\delta \frac{\partial f}{\partial x_i}(x_i e_i)\, \mathrm{d}x_i$$

and hence

$$X = \int_0^\delta \left(\frac{\partial f}{\partial x_i}(x_i e_i + \delta e_j) - \frac{\partial f}{\partial x_i}(x_i e_i)\right) \mathrm{d}x_i.$$

But by the mean value theorem, for each x_i we have

$$\frac{\partial f}{\partial x_i}(x_i e_i + \delta e_j) - \frac{\partial f}{\partial x_i}(x_i e_i) = \delta \frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(x_i e_i + x_j e_j)$$

for some $0 \le x_j \le \delta$. By our construction of δ, we thus have

$$\left| \frac{\partial f}{\partial x_i}(x_i e_i + \delta e_j) - \frac{\partial f}{\partial x_i}(x_i e_i) - \delta a \right| \le \varepsilon\delta.$$

Integrating this from 0 to δ, we thus obtain

$$|X - \delta^2 a| \le \varepsilon\delta^2.$$

We can run the same argument with the rôle of i and j reversed (note that X is symmetric in i and j), to obtain

$$|X - \delta^2 a'| \le \varepsilon\delta^2.$$

From the triangle inequality we thus obtain

$$|\delta^2 a - \delta^2 a'| \le 2\varepsilon\delta^2,$$

and thus

$$|a - a'| \le 2\varepsilon.$$

But this is true for all $\varepsilon > 0$, and a and a' do not depend on ε, and so we must have $a = a'$, as desired. □

One should caution that Clairaut's theorem fails if we do not assume the double derivatives to be continuous; see Exercise 6.5.1.

— Exercise —

Exercise 6.5.1 Let $f : \mathbf{R}^2 \to \mathbf{R}$ be the function defined by $f(x,y) := \frac{xy^3}{x^2+y^2}$ when $(x,y) \neq (0,0)$, and $f(0,0) := 0$. Show that f is continuously differentiable, and the double derivatives $\frac{\partial}{\partial y}\frac{\partial f}{\partial x}$ and $\frac{\partial}{\partial x}\frac{\partial f}{\partial y}$ exist, but are not equal to each other at $(0,0)$. Explain why this does not contradict Clairaut's theorem.

6.6 The Contraction Mapping Theorem

Before we turn to the next topic—namely the inverse function theorem—we need to develop a useful fact from the theory of complete metric spaces, namely the contraction mapping theorem.

Definition 6.6.1 (*Contraction*) Let (X, d) be a metric space, and let $f : X \to X$ be a map. We say that f is a *contraction* if we have $d(f(x), f(y)) \le d(x,y)$ for all

$x, y \in X$. We say that f is a *strict contraction* if there exists a constant $0 < c < 1$ such that $d(f(x), f(y)) \leq cd(x, y)$ for all $x, y \in X$; we call c the *contraction constant* of f.

Examples 6.6.2 The map $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x + 1$ is a contraction but not a strict contraction. The map $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) := x/2$ is a strict contraction. The map $f: [0, 1] \to [0, 1]$ defined by $f(x) := x - x^2$ is a contraction but not a strict contraction. (For justifications of these statements, see Exercise 6.6.5.)

Definition 6.6.3 (*Fixed points*) Let $f: X \to X$ be a map, and $x \in X$. We say that x is a *fixed point* of f if $f(x) = x$.

Contractions do not necessarily have any fixed points; for instance, the map $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) = x + 1$ does not. However, it turns out that *strict* contractions always do, at least when X is complete:

Theorem 6.6.4 (Contraction mapping theorem) *Let (X, d) be a metric space, and let $f: X \to X$ be a strict contraction. Then f can have at most one fixed point. Moreover, if we also assume that X is non-empty and complete, then f has* exactly *one fixed point.*

Proof See Exercise 6.6.7. □

Remark 6.6.5 The contraction mapping theorem is one example of a *fixed point theorem*—a theorem which guarantees, assuming certain conditions, that a map will have a fixed point. There are a number of other fixed point theorems which are also useful. One amusing one is the so-called *hairy ball theorem*, which (among other things) states that any continuous map $f: S^2 \to S^2$ from the sphere $S^2 := \{(x, y, z) \in \mathbf{R}^3 : x^2 + y^2 + z^2 = 1\}$ to itself, must contain either a fixed point, or an anti-fixed point (a point $x \in S^2$ such that $f(x) = -x$). A proof of this theorem can be found in any topology text; it is beyond the scope of this text.

We shall give one consequence of the contraction mapping theorem which is important for our application to the inverse function theorem. Basically, this says that any map f on a ball which is a "small" perturbation of the identity map, remains one-to-one and cannot create any internal holes in the ball.

Lemma 6.6.6 *Let $B(0, r)$ be a ball in $\mathbf{R}^n$ centered at the origin, and let $g: B(0, r) \to \mathbf{R}^n$ be a map such that $g(0) = 0$ and*

$$\|g(x) - g(y)\| \leq \frac{1}{2}\|x - y\|$$

for all $x, y \in B(0, r)$ (here $\|x\|$ denotes the length of x in $\mathbf{R}^n$). Then the function $f: B(0, r) \to \mathbf{R}^n$ defined by $f(x) := x + g(x)$ is one-to-one, and furthermore the image $f(B(0, r))$ of this map contains the ball $B(0, r/2)$.

Proof We first show that f is one-to-one. Suppose for sake of contradiction that we had two different points $x, y \in B(0, r)$ such that $f(x) = f(y)$. But then we would have $x + g(x) = y + g(y)$, and hence

$$\|g(x) - g(y)\| = \|x - y\|.$$

The only way this can be consistent with our hypothesis $\|g(x) - g(y)\| \leq \frac{1}{2}\|x - y\|$ is if $\|x - y\| = 0$, i.e., if $x = y$, a contradiction. Thus f is one-to-one.

Now we show that $f(B(0, r))$ contains $B(0, r/2)$. Let y be any point in $B(0, r/2)$; our objective is to find a point $x \in B(0, r)$ such that $f(x) = y$, or in other words that $x = y - g(x)$. So the problem is now to find a fixed point of the map $x \mapsto y - g(x)$.

Let $F \colon B(0, r) \to B(0, r)$ denote the function $F(x) := y - g(x)$. Observe that if $x \in B(0, r)$, then

$$\|F(x)\| \leq \|y\| + \|g(x)\| \leq \frac{r}{2} + \|g(x) - g(0)\| \leq \frac{r}{2} + \frac{1}{2}\|x - 0\| < \frac{r}{2} + \frac{r}{2} = r,$$

so F does indeed map $B(0, r)$ to itself. The same argument shows that for a sufficiently small $\varepsilon > 0$, F maps the closed ball $\overline{B(0, r - \varepsilon)}$ to itself. Also, for any x, x' in $B(0, r)$ we have

$$\|F(x) - F(x')\| = \|g(x') - g(x)\| \leq \frac{1}{2}\|x' - x\|$$

so F is a strict contraction on $B(0, r)$, and hence on the complete space $\overline{B(0, r - \varepsilon)}$. By the contraction mapping theorem, F has a fixed point, i.e., there exists an x such that $x = y - g(x)$. But this means that $f(x) = y$, as desired. □

— Exercise —

Exercise 6.6.1 Let $f \colon [a, b] \to [a, b]$ be a differentiable function of one variable such that $|f'(x)| \leq 1$ for all $x \in [a, b]$. Prove that f is a contraction. (*Hint:* use the mean value theorem, Corollary 10.2.9.) If in addition $|f'(x)| < 1$ for all $x \in [a, b]$ and f' is continuous, show that f is a strict contraction.

Exercise 6.6.2 Show that if $f \colon [a, b] \to \mathbf{R}$ is differentiable and is a contraction, then $|f'(x)| \leq 1$.

Exercise 6.6.3 Give an example of a function $f \colon [a, b] \to \mathbf{R}$ which is continuously differentiable and such that $|f(x) - f(y)| < |x - y|$ for all distinct $x, y \in [a, b]$, but such that $|f'(x)| = 1$ for at least one value of $x \in [a, b]$.

Exercise 6.6.4 Given an example of a function $f \colon [a, b] \to \mathbf{R}$ which is a strict contraction but which is not differentiable for at least one point x in $[a, b]$.

Exercise 6.6.5 Verify the claims in Examples 6.6.2.

Exercise 6.6.6 Show that every contraction on a metric space X is necessarily continuous.

Exercise 6.6.7 Prove Theorem 6.6.4. (*Hint*: to prove that there is at most one fixed point, argue by contradiction. To prove that there is at least one fixed point, pick any $x_0 \in X$ and define recursively $x_1 = f(x_0)$, $x_2 = f(x_1)$, $x_3 = f(x_2)$, etc. Prove inductively that $d(x_{n+1}, x_n) \leq c^n d(x_1, x_0)$, and conclude (using the geometric series formula, Lemma 7.3.3) that the sequence $(x_n)_{n=0}^\infty$ is a Cauchy sequence. Then prove that the limit of this sequence is a fixed point of f.)

Exercise 6.6.8 Let (X, d) be a complete metric space, and let $f : X \to X$ and $g : X \to X$ be two strict contractions on X with contraction coefficients c and c', respectively. From Theorem 6.6.4 we know that f has some fixed point x_0, and g has some fixed point y_0. Suppose we know that there is an $\varepsilon > 0$ such that $d(f(x), g(x)) \leq \varepsilon$ for all $x \in X$ (i.e., f and g are within ε of each other in the uniform metric). Show that $d(x_0, y_0) \leq \varepsilon/(1 - \min(c, c'))$. Thus nearby contractions have nearby fixed points.

6.7 The Inverse Function Theorem in Several Variable Calculus

We recall the inverse function theorem in single-variable calculus (Theorem 10.4.2), which asserts that if a function $f : \mathbf{R} \to \mathbf{R}$ is invertible, differentiable, and $f'(x_0)$ is nonzero, then f^{-1} is differentiable at $f(x_0)$, and

$$(f^{-1})'(f(x_0)) = \frac{1}{f'(x_0)}.$$

In fact, one can say something even when f' is not invertible, as long as we know that f is *continuously* differentiable. If $f'(x_0)$ is nonzero, then $f'(x_0)$ must be either strictly positive or strictly negative, which implies (since we are assuming f' to be continuous) that $f'(x)$ is either strictly positive for x near x_0, or strictly negative for x near x_0. In particular, f must be either strictly increasing near x_0, or strictly decreasing near x_0. In either case, f will become invertible if we restrict the domain and codomain of f to be sufficiently close to x_0 and to $f(x_0)$, respectively. (The technical terminology for this is that f is *locally invertible near* x_0.)

The requirement that f be continuously differentiable is important; see Exercise 6.7.1.

It turns out that a similar theorem is true for functions $f : \mathbf{R}^n \to \mathbf{R}^n$ from one Euclidean space to the same space. However, the condition that $f'(x_0)$ is nonzero must be replaced with a slightly different one, namely that $f'(x_0)$ is *invertible*. We first remark that the inverse of a linear transformation is also linear:

Lemma 6.7.1 *Let $T : \mathbf{R}^n \to \mathbf{R}^n$ be a linear transformation which is also invertible. Then the inverse transformation $T^{-1} : \mathbf{R}^n \to \mathbf{R}^n$ is also linear.*

Proof See Exercise 6.7.2. □

We can now prove an important and useful theorem, arguably one of the most important theorems in several variable differential calculus.

Theorem 6.7.2 (Inverse function theorem) *Let E be an open subset of $\mathbf{R}^n$, and let $f : E \to \mathbf{R}^n$ be a function which is continuously differentiable on E. Suppose $x_0 \in E$ is such that the linear transformation $f'(x_0) : \mathbf{R}^n \to \mathbf{R}^n$ is invertible. Then there exists an open set U in E containing x_0, and an open set V in $\mathbf{R}^n$ containing $f(x_0)$, such that f is a bijection from U to V. In particular, there is an inverse map $f^{-1} : V \to U$. Furthermore, this inverse map is differentiable at $f(x_0)$, and*

$$(f^{-1})'(f(x_0)) = (f'(x_0))^{-1}.$$

Proof We first observe that once we know the inverse map f^{-1} is differentiable, the formula $(f^{-1})'(f(x_0)) = (f'(x_0))^{-1}$ is automatic. This comes from starting with the identity

$$I = f^{-1} \circ f$$

on U, where $I : \mathbf{R}^n \to \mathbf{R}^n$ is the identity map $Ix := x$, and then differentiating both sides using the chain rule at x_0 to obtain

$$I'(x_0) = (f^{-1})'(f(x_0))f'(x_0).$$

Since $I'(x_0) = I$, we thus have $(f^{-1})'(f(x_0)) = (f'(x_0))^{-1}$ as desired.

We remark that this argument shows that if $f'(x_0)$ is *not* invertible, then there is no way that an inverse f^{-1} can exist and be differentiable at $f(x_0)$.

Next, we observe that it suffices to prove the theorem under the additional assumption $f(x_0) = 0$. The general case then follows from the special case by replacing f by a new function $\tilde{f}(x) := f(x) - f(x_0)$ and then applying the special case to $\tilde{f}$ (note that V will have to shift by $f(x_0)$). Note that $f^{-1}(y) = \tilde{f}^{-1}(y - f(x_0))$ (why?). Henceforth we will always assume $f(x_0) = 0$.

In a similar manner, one can make the assumption $x_0 = 0$. The general case then follows from this case by replacing f by a new function $\tilde{f}(x) := f(x + x_0)$ and applying the special case to $\tilde{f}$ (note that E and U will have to shift by x_0). Note that $f^{-1}(y) = \tilde{f}^{-1}(y) + x_0$ - why? Henceforth we will always assume $x_0 = 0$. Thus we now have that $f(0) = 0$ and that $f'(0)$ is invertible.

Finally, one can assume that $f'(0) = I$, where $I : \mathbf{R}^n \to \mathbf{R}^n$ is the identity transformation $Ix = x$. The general case then follows from this case by replacing f with a new function $\tilde{f} : E \to \mathbf{R}^n$ defined by $\tilde{f}(x) := f'(0)^{-1} f(x)$, and applying the special case to this case. Note from Lemma 6.7.1 that $f'(0)^{-1}$ is a linear transformation. In particular, we note that $\tilde{f}(0) = 0$ and that

$$\tilde{f}'(0) = f'(0)^{-1} f'(0) = I,$$

so by the special case of the inverse function theorem we know that there exists an open set U' containing 0, and an open set V' containing 0, such that $\tilde{f}$ is a bijection from U' to V', and that $\tilde{f}^{-1} : V' \to U'$ is differentiable at 0 with derivative I. But we have $f(x) = f'(0)\tilde{f}(x)$, and hence f is a bijection from U' to $f'(0)(V')$ (note that $f'(0)$ is also a bijection). Since $f'(0)$ and its inverse are both continuous, $f'(0)(V')$ is open, and it certainly contains 0. Now consider the inverse function $f^{-1} : f'(0)(V') \to U'$. Since $f(x) = f'(0)\tilde{f}(x)$, we see that $f^{-1}(y) = \tilde{f}^{-1}(f'(0)^{-1}y)$ for all $y \in f'(0)(V')$ (why? use the fact that $\tilde{f}$ is a bijection from U' to V'). In particular we see that f^{-1} is differentiable at 0.

So all we have to do now is prove the inverse function theorem in the special case, when $x_0 = 0, f(x_0) = 0$, and $f'(x_0) = I$. Let $g\colon E \to \mathbf{R}^n$ denote the function $g(x) := f(x) - x$. Then $g(0) = 0$ and $g'(0) = 0$. In particular

$$\frac{\partial g}{\partial x_j}(0) = 0$$

for $j = 1, \ldots, n$. Since g is continuously differentiable, there thus exists a ball $B(0, r)$ in E such that

$$\left\| \frac{\partial g}{\partial x_j}(x) \right\| \le \frac{1}{2n^2}$$

for all $x \in B(0, r)$. (There is nothing particularly special about $\frac{1}{2n^2}$, we just need a nice small number here.) In particular, for any $x \in B(0, r)$ and $v = (v_1, \ldots, v_n)$ we have

$$\|D_v g(x)\| = \left\| \sum_{j=1}^{n} v_j \frac{\partial g}{\partial x_j}(x) \right\|$$

$$\le \sum_{j=1}^{n} |v_j| \left\| \frac{\partial g}{\partial x_j}(x) \right\|$$

$$\le \sum_{j=1}^{n} \|v\| \frac{1}{2n^2} \le \frac{1}{2n}\|v\|.$$

But now for any $x, y \in B(0, r)$, we have by the fundamental theorem of calculus

$$g(y) - g(x) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} g(x + t(y - x))\, \mathrm{d}t$$

$$= \int_0^1 D_{y-x} g(x + t(y - x))\, \mathrm{d}t,$$

where the integral of a vector-valued function is defined by integrating each component separately. By the previous remark, the vectors $D_{y-x}g(x+t(y-x))$ have a magnitude of at most $\frac{1}{2n}\|y-x\|$. Thus every component of these vectors has magnitude at most $\frac{1}{2n}\|y-x\|$. Thus every component of $g(y)-g(x)$ has magnitude at most $\frac{1}{2n}\|y-x\|$, and hence $g(y)-g(x)$ itself has magnitude at most $\frac{1}{2}\|y-x\|$ (actually, it will be substantially less than this, but this bound will be enough for our purposes). In other words, g is a contraction. By Lemma 6.6.6, the map $f=g+I$ is thus one-to-one on $B(0,r)$, and the image $f(B(0,r))$ contains $B(0,r/2)$. In particular we have an inverse map $f^{-1}:B(0,r/2)\to B(0,r)$ defined on $B(0,r/2)$.

Applying the contraction bound with $y=0$ we obtain in particular that

$$\|g(x)\|\le\frac{1}{2}\|x\|$$

for all $x\in B(0,r)$, and so by the triangle inequality

$$\frac{1}{2}\|x\|\le\|f(x)\|\le\frac{3}{2}\|x\|$$

for all $x\in B(0,r)$.

Now we set $V{:=}B(0,r/2)$ and $U{:=}f^{-1}(V)\cap B(0,r)$. Then by construction f is a bijection from U to V. V is clearly open, and U is also open since f is continuous. (Notice that if a set is open relative to $B(0,r)$, then it is open in $\mathbf{R}^n$ as well.) Now we want to show that $f^{-1}:V\to U$ is differentiable at 0 with derivative $I^{-1}=I$. In other words, we wish to show that

$$\lim_{x\to 0;x\in V\backslash\{0\}}\frac{\|f^{-1}(x)-f^{-1}(0)-I(x-0)\|}{\|x\|}=0.$$

Since $f(0)=0$, we have $f^{-1}(0)=0$, and the above simplifies to

$$\lim_{x\to 0;x\in V\backslash\{0\}}\frac{\|f^{-1}(x)-x\|}{\|x\|}=0.$$

Let $(x_n)_{n=1}^{\infty}$ be any sequence in $V\backslash\{0\}$ that converges to 0. By Proposition 3.1.5(b), it suffices to show that

$$\lim_{n\to\infty}\frac{\|f^{-1}(x_n)-x_n\|}{\|x_n\|}=0.$$

Write $y_n{:=}f^{-1}(x_n)$. Then $y_n\in B(0,r)$ and $x_n=f(y_n)$. In particular we have

$$\frac{1}{2}\|y_n\|\le\|x_n\|\le\frac{3}{2}\|y_n\|$$

and so since $\|x_n\|$ goes to 0, $\|y_n\|$ goes to zero also, and their ratio remains bounded. It will thus suffice to show that

$$\lim_{n\to\infty} \frac{\|y_n - f(y_n)\|}{\|y_n\|} = 0.$$

But since y_n is going to 0, and f is differentiable at 0, we have

$$\lim_{n\to\infty} \frac{\|f(y_n) - f(0) - f'(0)(y_n - 0)\|}{\|y_n\|} = 0$$

as desired (since $f(0) = 0$ and $f'(0) = I$). □

The inverse function theorem gives a useful criterion for when a function is (locally) invertible at a point x_0 - all we need is for its derivative $f'(x_0)$ to be invertible (and then we even get further information, for instance we can compute the derivative of f^{-1} at $f(x_0)$). Of course, this begs the question of how one can tell whether the linear transformation $f'(x_0)$ is invertible or not. Recall that we have $f'(x_0) = L_{Df(x_0)}$, so by Lemmas 6.1.13 and 6.1.16 we see that the linear transformation $f'(x_0)$ is invertible if and only if the matrix $Df(x_0)$ is. There are many ways to check whether a matrix such as $Df(x_0)$ is invertible; for instance, one can use determinants, or alternatively Gaussian elimination methods. We will not pursue this matter here, but refer the reader to any linear algebra text.

If $f'(x_0)$ exists but is non-invertible, then the inverse function theorem does not apply. In such a situation it is not possible for f^{-1} to exist and be differentiable at $f(x_0)$; this was remarked in the above proof. But it is still possible for f to be invertible. For instance, the single-variable function $f: \mathbf{R} \to \mathbf{R}$ defined by $f(x) = x^3$ is invertible despite $f'(0)$ not being invertible.

— Exercise —

Exercise 6.7.1 Let $f: \mathbf{R} \to \mathbf{R}$ be the function defined by $f(x):=x + x^2 \sin(1/x^4)$ for $x \neq 0$ and $f(0):=0$. Show that f is differentiable and $f'(0) = 1$, but f is not increasing on any open set containing 0 (*Hint:* show that the derivative of f can turn negative arbitrarily close to 0. Drawing a graph of f may aid your intuition.)

Exercise 6.7.2 Prove Lemma 6.7.1.

Exercise 6.7.3 Let $f: \mathbf{R}^n \to \mathbf{R}^n$ be a continuously differentiable function such that $f'(x)$ is an invertible linear transformation for every $x \in \mathbf{R}^n$. Show that whenever V is an open set in $\mathbf{R}^n$, that $f(V)$ is also open. (*Hint*: use the inverse function theorem.)

Exercise 6.7.4 Let the notation and hypotheses be as in Theorem 6.7.2. Show that after shrinking the open sets U, V as necessarily (while still keeping x_0 in U and $f(x_0)$ in V), the derivative map $f'(x)$ is invertible for all $x \in U$, and that the inverse map f^{-1} is differentiable at every point of V with $(f^{-1})'(f(x) = (f'(x))^{-1}$ for all $x \in U$. Finally, show that f^{-1} is continuously differentiable on V.

6.8 The Implicit Function Theorem

Recall (from Exercise 3.5.10) that a function $f : \mathbf{R} \to \mathbf{R}$ gives rise to a *graph*

$$\{(x, f(x)) : x \in \mathbf{R}\}$$

which is a subset of $\mathbf{R}^2$, usually looking like a curve. However, not all curves are graphs, they must obey the *vertical line test*, that for every x there is exactly one y such that (x, y) is in the curve. For instance, the circle $\{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = 1\}$ is not a graph, although if one restricts to a semicircle such as $\{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = 1, y > 0\}$ then one again obtains a graph. Thus while the entire circle is not a graph, certain local portions of it are. (The portions of the circle near $(1, 0)$ and $(-1, 0)$ are not graphs over the variable x, but they are graphs over the variable y).

Similarly, any function $g : \mathbf{R}^n \to \mathbf{R}$ gives rise to a graph $\{(x, g(x)) : x \in \mathbf{R}^n\}$ in $\mathbf{R}^{n+1}$, which in general looks like some sort of n-dimensional surface in $\mathbf{R}^{n+1}$ (the technical term for this is a *hypersurface*). Conversely, one may ask which hypersurfaces are actually graphs of some function, and whether that function is continuous or differentiable.

If the hypersurface is given geometrically, then one can again invoke the vertical line test to work out whether it is a graph or not. But what if the hypersurface is given algebraically, for instance the surface $\{(x, y, z) \in \mathbf{R}^3 : xy + yz + zx = -1\}$? Or more generally, a hypersurface of the form $\{x \in \mathbf{R}^n : g(x) = 0\}$, where $g : \mathbf{R}^n \to \mathbf{R}$ is some function? In this case, it is still possible to say whether the hypersurface is a graph, locally at least, by means of the *implicit function theorem*.

Theorem 6.8.1 (Implicit function theorem) *Let E be an open subset of $\mathbf{R}^n$, let $f : E \to \mathbf{R}$ be continuously differentiable, and let $y = (y_1, \ldots, y_n)$ be a point in E such that $f(y) = 0$ and $\frac{\partial f}{\partial x_n}(y) \neq 0$. Then there exists an open subset U of $\mathbf{R}^{n-1}$ containing $(y_1, \ldots, y_{n-1})$, an open subset V of E containing y, and a function $g : U \to \mathbf{R}$ such that $g(y_1, \ldots, y_{n-1}) = y_n$, and*

$$\{(x_1, \ldots, x_n) \in V : f(x_1, \ldots, x_n) = 0\}$$

$$= \{(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) : (x_1, \ldots, x_{n-1}) \in U\}.$$

In other words, the set $\{x \in V : f(x) = 0\}$ is a graph of a function over U. Moreover, g is differentiable at $(y_1, \ldots, y_{n-1})$, and we have

$$\frac{\partial g}{\partial x_j}(y_1, \ldots, y_{n-1}) = -\frac{\partial f}{\partial x_j}(y) / \frac{\partial f}{\partial x_n}(y) \tag{6.1}$$

for all $1 \leq j \leq n - 1$.

Remark 6.8.2 Equation (6.1) is sometimes derived using *implicit differentiation*. Basically, the point is that if you know that

$$f(x_1, \ldots, x_n) = 0$$

then (as long as $\frac{\partial f}{\partial x_n} \neq 0$) the variable x_n is “implicitly” defined in terms of the other $n-1$ variables, and one can differentiate the above identity in, say, the x_j direction using the chain rule to obtain

$$\frac{\partial f}{\partial x_j} + \frac{\partial f}{\partial x_n}\frac{\partial x_n}{\partial x_j} = 0$$

which is (6.1) in disguise (we are using g to represent the implicit function defining x_n in terms of $x_1, \ldots, x_n$). Thus, the implicit function theorem allows one to define a dependence implicitly, by means of a constraint rather than by a direct formula of the form $x_n = g(x_1, \ldots, x_{n-1})$.

Proof This theorem looks somewhat fearsome, but actually it is a fairly quick consequence of the inverse function theorem. Let $F\colon E \to \mathbf{R}^n$ be the function

$$F(x_1, \ldots, x_n) := (x_1, \ldots, x_{n-1}, f(x_1, \ldots, x_n)).$$

This function is continuously differentiable. Also note that

$$F(y) = (y_1, \ldots, y_{n-1}, 0)$$

and

$$DF(y) = \left(\frac{\partial F}{\partial x_1}(y)^T, \frac{\partial F}{\partial x_2}(y)^T, \ldots, \frac{\partial F}{\partial x_n}(y)^T\right)$$

$$= \begin{pmatrix} 1 & 0 & \ldots 0 & 0 \\ 0 & 1 & \ldots 0 & 0 \\ \vdots & \vdots & \ddots \vdots & \vdots \\ 0 & 0 & \ldots 1 & 0 \\ \frac{\partial f}{\partial x_1}(y) & \frac{\partial f}{\partial x_2}(y) & \ldots \frac{\partial f}{\partial x_{n-1}}(y) & \frac{\partial f}{\partial x_n}(y) \end{pmatrix}.$$

Since $\frac{\partial f}{\partial x_n}(y)$ is assumed by hypothesis to be nonzero, this matrix is invertible; this can be seen either by computing the determinant, or using row reduction, or by computing the inverse explicitly, which is

$$DF(y)^{-1} = \begin{pmatrix} 1 & 0 & \ldots 0 & 0 \\ 0 & 1 & \ldots 0 & 0 \\ \vdots & \vdots & \ddots \vdots & \vdots \\ 0 & 0 & \ldots 1 & 0 \\ -\frac{\partial f}{\partial x_1}(y)/a & -\frac{\partial f}{\partial x_2}(y)/a & \ldots -\frac{\partial f}{\partial x_{n-1}}(y)/a & 1/a \end{pmatrix},$$

where we have written $a = \frac{\partial f}{\partial x_n}(y)$ for short. Thus the inverse function theorem applies, and we can find an open set V in E containing y, and an open set W in $\mathbf{R}^n$ containing $F(y) = (y_1, \ldots, y_{n-1}, 0)$, such that F is a bijection from V to W, and that F^{-1} is differentiable at $(y_1, \ldots, y_{n-1}, 0)$.

Let us write F^{-1} in co-ordinates as

$$F^{-1}(x) = (h_1(x), h_2(x), \ldots, h_n(x))$$

where $x \in W$. Since $F(F^{-1}(x)) = x$, we have $h_j(x_1, \ldots, x_n) = x_j$ for all $1 \leq j \leq n-1$ and $x \in W$, and

$$f(x_1, \ldots, x_{n-1}, h_n(x_1, \ldots, x_n)) = x_n.$$

Also, h_n is differentiable at $(y_1, \ldots, y_{n-1}, 0)$ since F^{-1} is.

Now we set $U := \{(x_1, \ldots, x_{n-1}) \in \mathbf{R}^{n-1} : (x_1, \ldots, x_{n-1}, 0) \in W\}$. Note that U is open and contains $(y_1, \ldots, y_{n-1})$. Now we define $g : U \to \mathbf{R}$ by $g(x_1, \ldots, x_{n-1}) := h_n(x_1, \ldots, x_{n-1}, 0)$. Then g is differentiable at $(y_1, \ldots, y_{n-1})$. Now we prove that

$$\{(x_1, \ldots, x_n) \in V : f(x_1, \ldots, x_n) = 0\}$$

$$= \{(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) : (x_1, \ldots, x_{n-1}) \in U\}.$$

First suppose that $(x_1, \ldots, x_n) \in V$ and $f(x_1, \ldots, x_n) = 0$. Then we have $F(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-1}, 0)$, which lies in W. Thus $(x_1, \ldots, x_{n-1})$ lies in U. Applying F^{-1}, we see that $(x_1, \ldots, x_n) = F^{-1}(x_1, \ldots, x_{n-1}, 0)$. In particular $x_n = h_n(x_1, \ldots, x_{n-1}, 0)$, and hence $x_n = g(x_1, \ldots, x_{n-1})$. Thus every element of the left-hand set lies in the right-hand set. The reverse inclusion comes by reversing all the above steps and is left to the reader.

Finally, we show the formula for the partial derivatives of g. From the preceding discussion we have

$$f(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) = 0$$

for all $(x_1, \ldots, x_{n-1}) \in U$. Since g is differentiable at $(y_1, \ldots, y_{n-1})$, and f is differentiable at $(y_1, \ldots, y_{n-1}, g(y_1, \ldots, y_{n-1})) = y$, we may use the chain rule, differentiating in x_j, to obtain

$$\frac{\partial f}{\partial x_j}(y) + \frac{\partial f}{\partial x_n}(y)\frac{\partial g}{\partial x_j}(y_1, \ldots, y_{n-1}) = 0$$

and the claim follows by simple algebra. □

Example 6.8.3 Consider the surface $S := \{(x, y, z) \in \mathbf{R}^3 : xy + yz + zx = -1\}$, which we rewrite as $\{(x, y, z) \in \mathbf{R}^3 : f(x, y, z) = 0\}$, where $f : \mathbf{R}^3 \to \mathbf{R}$ is the function $f(x, y, z) := xy + yz + zx + 1$. Clearly f is continuously differentiable, and $\frac{\partial f}{\partial z} = y + x$. Thus for any (x_0, y_0, z_0) in S with $y_0 + x_0 \neq 0$, one can write this surface (near (x_0, y_0, z_0)) as a graph of the form $\{(x, y, g(x, y)) : (x, y) \in U\}$ for some open set U containing (x_0, y_0), and some function g which is differentiable at (x_0, y_0). Indeed one can implicitly differentiate to obtain that

$$\frac{\partial g}{\partial x}(x_0, y_0) = -\frac{y_0 + z_0}{y_0 + x_0} \text{ and } \frac{\partial g}{\partial y}(x_0, y_0) = -\frac{x_0 + z_0}{y_0 + x_0}.$$

In the implicit function theorem, if the derivative $\frac{\partial f}{\partial x_n}$ equals zero at some point, then it is unlikely that the set $\{x \in \mathbf{R}^n : f(x) = 0\}$ can be written as a graph of the x_n variable in terms of the other $n - 1$ variables near that point. However, if some other derivative $\frac{\partial f}{\partial x_j}$ is nonzero, then it would be possible to write the x_j variable in terms of the other $n - 1$ variables, by a variant of the implicit function theorem. Thus as long as the gradient ∇f is not entirely zero, one can write this set $\{x \in \mathbf{R}^n : f(x) = 0\}$ as a graph of *some* variable x_j in terms of the other $n - 1$ variables. (The circle $\{(x, y) \in \mathbf{R}^2 : x^2 + y^2 - 1 = 0\}$ is a good example of this; it is not a graph of y in terms of x, or x in terms of y, but near every point it is one of the two. And this is because the gradient of $x^2 + y^2 - 1$ is never zero on the circle.) However, if ∇f does vanish at some point x_0, then we say that f has a *critical point* at x_0 and the behavior there is much more complicated. For instance, the set $\{(x, y) \in \mathbf{R}^2 : x^2 - y^2 = 0\}$ has a critical point at $(0, 0)$ and there the set does not look like a graph of any sort (it is the union of two lines).

Remark 6.8.4 Sets which look like graphs of continuous functions at every point have a name, they are called *manifolds*. Thus $\{x \in \mathbf{R}^n : f(x) = 0\}$ will be a manifold if it contains no critical points of f. The theory of manifolds is very important in modern geometry (especially differential geometry and algebraic geometry), but we will not discuss it here as it is a graduate level topic.

— Exercise —

Exercise 6.8.1 Let the notation and hypotheses be as in Theorem 6.8.1. Show that, after shrinking the open sets U, V as necessary, that the function g becomes continuously differentiable on all of U, and the Eq. (6.1) holds at all points of U.

Chapter 7
Lebesgue Measure

In the previous chapter we discussed differentiation in several variable calculus. It is now only natural to consider the question of integration in several variable calculus. The general question we wish to answer is this: given some subset Ω of $\mathbf{R}^n$, and some real-valued function $f : \Omega \to \mathbf{R}$, is it possible to integrate f on Ω to obtain some number $\int_\Omega f$? (It is possible to consider other types of functions, such as complex-valued or vector-valued functions, but this turns out not to be too difficult once one knows how to integrate real-valued functions, since one can integrate a complex or vector-valued function, by integrating each real-valued component of that function separately.)

In one dimension we already have developed (in Chap. 11) the notion of a *Riemann integral* $\int_{[a,b]} f$, which answers this question when Ω is an interval $\Omega = [a, b]$, and f is *Riemann integrable*. Exactly what Riemann integrability means is not important here, but let us just remark that every piecewise continuous function is Riemann integrable, and in particular every piecewise constant function is Riemann integrable. However, not all functions are Riemann integrable. It is possible to extend this notion of a Riemann integral to higher dimensions, but it requires quite a bit of effort and one can still only integrate "Riemann integrable" functions, which turn out to be a rather unsatisfactorily small class of functions. (For instance, the pointwise limit of Riemann integrable functions need not be Riemann integrable, and the same goes for an L^2 limit, although we have already seen that uniform limits of Riemann integrable functions remain Riemann integrable.)

Because of this, we must look beyond the Riemann integral to obtain a truly satisfactory notion of integration, one that can handle even very discontinuous functions. This leads to the notion of the *Lebesgue integral*, which we shall spend this chapter and the next constructing. The Lebesgue integral can handle a very large class of functions, including all the Riemann integrable functions but also many others as well; in fact, it is safe to say that it can integrate virtually any function that one actually needs in mathematics, at least if one works on Euclidean spaces and everything is absolutely integrable. (If one assumes the axiom of choice, then there are still some

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_7

pathological functions one can construct which cannot be integrated by the Lebesgue integral, but these functions will not come up in real-life applications.)

Before we turn to the details, we begin with an informal discussion. In order to understand how to compute an integral $\int_\Omega f$, we must first understand a more basic and fundamental question: how does one compute the *length/area/volume* of Ω? To see why this question is connected to that of integration, observe that if one integrates the function 1 on the set Ω, then one should obtain the length of Ω (if Ω is one-dimensional), the area of Ω (if Ω is two-dimensional), or the volume of Ω (if Ω is three-dimensional). To avoid splitting into cases depending on the dimension, we shall refer to the *measure* of Ω as either the length, area, volume, (or hypervolume, etc.) of Ω, depending on what Euclidean space $\mathbf{R}^n$ we are working in.

Ideally, to every subset Ω of $\mathbf{R}^n$ we would like to associate a non-negative number $m(\Omega)$, which will be the measure of Ω (i.e., the length, area, volume, etc.). We allow the possibility for $m(\Omega)$ to be zero (e.g., if Ω is just a single point or the empty set) or for $m(\Omega)$ to be infinite (e.g., if Ω is all of $\mathbf{R}^n$). This measure should obey certain reasonable properties; for instance, the measure of the unit cube $(0, 1)^n := \{(x_1, \dots, x_n) : 0 < x_i < 1\}$ should equal 1, we should have $m(A \cup B) = m(A) + m(B)$ if A and B are disjoint (and similarly that $m\left(\bigcup_{n=1}^\infty A_n\right) = \sum_{n=1}^\infty m(A_n)$ when the A_n are disjoint), we should have $m(A) \leq m(B)$ whenever $A \subseteq B$, and we should have $m(x + A) = m(A)$ for any $x \in \mathbf{R}^n$ (i.e., if we shift A by the vector x the measure should be the same).

Remarkably, it turns out that such a measure *does not exist*; one cannot assign a non-negative number to *every* subset of $\mathbf{R}^n$ which has the above properties. This is quite a surprising fact, as it goes against one's intuitive concept of volume; we shall prove it later in these notes. (An even more dramatic example of this failure of intuition is the *Banach-Tarski paradox*, in which a unit ball in $\mathbf{R}^3$ is decomposed into five pieces, and then the five pieces are reassembled via translations and rotations to form two complete and disjoint unit balls, thus violating any concept of conservation of volume; however we will not discuss this paradox here.)

What these paradoxes mean is that it is impossible to find a reasonable way to assign a measure to every single subset of $\mathbf{R}^n$. However, we can salvage matters by only measuring a certain class of sets in $\mathbf{R}^n$—the *measurable sets*. These are the only sets Ω for which we will define the measure $m(\Omega)$, and once one restricts one's attention to measurable sets, one recovers all the above properties again. Furthermore, almost all the sets one encounters in real life are measurable (e.g., all open and closed sets will be measurable), and so this turns out to be good enough to do analysis.

7.1 The Goal: Lebesgue Measure

Let $\mathbf{R}^n$ be a Euclidean space. Our goal in this chapter is to define a concept of *measurable set*, which will be a special kind of subset of $\mathbf{R}^n$, and for every such measurable set $\Omega \subseteq \mathbf{R}^n$, we will define the *Lebesgue measure* $m(\Omega)$ to be a certain number in $[0, \infty]$. The concept of measurable set will obey the following properties:

(i) (Borel property) Every open set in $\mathbf{R}^n$ is measurable, as is every closed set.
(ii) (Complementarity) If Ω is measurable, then $\mathbf{R}^n \backslash \Omega$ is also measurable.
(iii) (Boolean algebra property) If $(\Omega_j)_{j \in J}$ is any finite collection of measurable sets (so J is finite), then the union $\bigcup_{j \in J} \Omega_j$ and intersection $\bigcap_{j \in J} \Omega_j$ are also measurable.
(iv) (σ-algebra property) If $(\Omega_j)_{j \in J}$ are any countable collection of measurable sets (so J is countable), then the union $\bigcup_{j \in J} \Omega_j$ and intersection $\bigcap_{j \in J} \Omega_j$ are also measurable.

Note that some of these properties are redundant; for instance, (iv) will imply (iii), and once one knows all open sets are measurable, (ii) will imply that all closed sets are measurable also. The properties (i–iv) will ensure that virtually every set one cares about is measurable; though as indicated in the introduction, there do exist non-measurable sets.

To every measurable set Ω, we associate the *Lebesgue measure* $m(\Omega)$ of Ω, which will obey the following properties:

(v) (Empty set) The empty set $\emptyset$ has measure $m(\emptyset) = 0$.
(vi) (Positivity) We have $0 \leq m(\Omega) \leq +\infty$ for every measurable set Ω.
(vii) (Monotonicity) If $A \subseteq B$, and A and B are both measurable, then $m(A) \leq m(B)$.
(viii) (Finite sub-additivity) If $(A_j)_{j \in J}$ are a finite collection of measurable sets, then $m\left(\bigcup_{j \in J} A_j\right) \leq \sum_{j \in J} m(A_j)$.
(ix) (Finite additivity) If $(A_j)_{j \in J}$ are a finite collection of *disjoint* measurable sets, then $m(\bigcup_{j \in J} A_j) = \sum_{j \in J} m(A_j)$.
(x) (Countable sub-additivity) If $(A_j)_{j \in J}$ are a countable collection of measurable sets, then $m\left(\bigcup_{j \in J} A_j\right) \leq \sum_{j \in J} m(A_j)$.
(xi) (Countable additivity) If $(A_j)_{j \in J}$ are a countable collection of *disjoint* measurable sets, then $m\left(\bigcup_{j \in J} A_j\right) = \sum_{j \in J} m(A_j)$.
(xii) (Normalization) The unit cube $[0, 1]^n = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : 0 \leq x_j \leq 1$ for all $1 \leq j \leq n\}$ has measure $m([0, 1]^n) = 1$.
(xiii) (Translation invariance) If Ω is a measurable set, and $x \in \mathbf{R}^n$, then $x + \Omega := \{x + y : y \in \Omega\}$ is also measurable, and $m(x + \Omega) = m(\Omega)$.

Again, many of these properties are redundant; for instance the countable additivity property can be used to deduce the finite additivity property, which in turn can be used to derive monotonicity (when combined with the positivity property). One can also obtain the sub-additivity properties from the additivity ones. Note that $m(\Omega)$ can be $+\infty$, and so in particular some of the sums in the above properties may also equal $+\infty$; in this chapter we adopt the convention that an infinite sum $\sum_{j \in J} a_j$ of non-negative quantities a_j is equal to $+\infty$ if the sum is not absolutely convergent. (Since everything is non-negative we will never have to deal with indeterminate forms such as $-\infty + +\infty$.)

Our goal for this chapter can then be stated thus:

Theorem 7.1.1 (Existence of Lebesgue measure). *There exists a concept of a measurable set, and a way to assign a number $m(\Omega)$ to every measurable subset $\Omega \subseteq \mathbf{R}^n$, which obeys all of the properties (i)–(xiii).*

It turns out that Lebesgue measure is pretty much unique; any other concept of measurability and measure which obeys axioms (i)–(xiii) will largely coincide with the construction we give. However there are other measures which obey only some of the above axioms; also, we may be interested in concepts of measure for other domains than Euclidean spaces $\mathbf{R}^n$. This leads to *measure theory*, which is an entire subject in itself and will not be pursued here; however we do remark that the concept of measures is very important in modern probability, and in the finer points of analysis (e.g., in the theory of distributions).

7.2 First Attempt: Outer Measure

Before we construct Lebesgue measure, we first discuss a somewhat naive approach to finding the measure of a set—namely, we try to cover the set by boxes, and then add up the volume of each box. This approach will almost work, giving us a concept called *outer measure* which can be applied to every set and obeys all of the properties (v)–(xiii) except for the additivity properties (ix), (xi). Later we will have to modify outer measure slightly to recover the additivity property.

We begin by starting with the notion of an open box.

Definition 7.2.1 (*Open box*) An *open box* (or *box* for short) B in $\mathbf{R}^n$ is any set of the form

$$B = \prod_{i=1}^{n} (a_i, b_i) := \{(x_1, \dots, x_n) \in \mathbf{R}^n : x_i \in (a_i, b_i) \text{ for all } 1 \le i \le n\},$$

where $b_i \geq a_i$ are real numbers. We define the *volume* $\mathrm{vol}(B)$ of this box to be the number

$$\mathrm{vol}(B) := \prod_{i=1}^{n} (b_i - a_i) = (b_1 - a_1)(b_2 - a_2) \dots (b_n - a_n).$$

For instance, the unit cube $(0, 1)^n$ is a box, and has volume 1. In one dimension $n = 1$, boxes are the same as open intervals. One can easily check that in general dimension that open boxes are indeed open. Note that if we have $b_i = a_i$ for some i, then the box becomes empty, and has volume 0, but we still consider this to be a box (albeit a rather silly one). Sometimes we will use $\mathrm{vol}_n(B)$ instead of $\mathrm{vol}(B)$ to emphasize that we are dealing with n-dimensional volume, thus for instance $\mathrm{vol}_1(B)$ would be the length of a one-dimensional box B, $\mathrm{vol}_2(B)$ would be the area of a two-dimensional box B, etc.

Remark 7.2.2 We of course expect the measure $m(B)$ of a box to be the same as the volume $\mathrm{vol}(B)$ of that box. This is in fact an inevitable consequence of the axioms (i)–(xiii) (see Exercise 7.2.5).

Definition 7.2.3 (*Covering by boxes*) Let $\Omega \subseteq \mathbf{R}^n$ be a subset of $\mathbf{R}^n$. We say that a collection $(B_j)_{j \in J}$ of boxes *cover* Ω iff $\Omega \subseteq \bigcup_{j \in J} B_j$.

Suppose $\Omega \subseteq \mathbf{R}^n$ can be covered by a finite or countable collection of boxes $(B_j)_{j \in J}$. If we wish Ω to be measurable, and if we wish to have a measure obeying the monotonicity and sub-additivity properties (vii), (viii), (x) and if we wish $m(B_j) = \mathrm{vol}(B_j)$ for every box j, then we must have

$$m(\Omega) \leq m\left(\bigcup_{j \in J} B_j\right) \leq \sum_{j \in J} m(B_j) = \sum_{j \in J} \mathrm{vol}(B_j).$$

We thus conclude

$$m(\Omega) \leq \inf\left\{\sum_{j \in J} \mathrm{vol}(B_j) : (B_j)_{j \in J} \text{ covers } \Omega;\ J \text{ at most countable}\right\}.$$

Inspired by this, we define

Definition 7.2.4 (*Outer measure*) If Ω is a set, we define the *outer measure* $m^*(\Omega)$ of Ω to be the quantity

$$m^*(\Omega) := \inf\left\{\sum_{j \in J} \mathrm{vol}(B_j) : (B_j)_{j \in J} \text{ covers } \Omega;\ J \text{ at most countable}\right\}.$$

Since $\sum_{j=1}^{\infty} \mathrm{vol}(B_j)$ is non-negative, we know that $m^*(\Omega) \geq 0$ for all Ω. However, it is quite possible that $m^*(\Omega)$ could equal $+\infty$. Note that because we are allowing ourselves to use a countable number of boxes, that every subset of $\mathbf{R}^n$ has at least one countable cover by boxes; in fact $\mathbf{R}^n$ itself can be covered by countably many translates of the unit cube $(0, 1)^n$ (how?). We will sometimes write $m_n^*(\Omega)$ instead of $m^*(\Omega)$ to emphasize the fact that we are using n-dimensional outer measure.

Note that outer measure can be defined for every single set (not just the measurable ones), because we can take the infimum of any non-empty set. It obeys several of the desired properties of a measure:

Lemma 7.2.5 (Properties of outer measure) *Outer measure has the following six properties:*

(v) (Empty set) The empty set $\emptyset$ *has outer measure* $m^*(\emptyset) = 0$.

(vi) (Positivity) We have $0 \leq m^*(\Omega) \leq +\infty$ *for every measurable set* Ω.

(vii) (Monotonicity) If $A \subseteq B \subseteq \mathbf{R}^n$, *then* $m^*(A) \leq m^*(B)$.

(viii) *(Finite sub-additivity) If $(A_j)_{j\in J}$ are a finite collection of subsets of $\mathbf{R}^n$, then $m^*\left(\bigcup_{j\in J} A_j\right) \le \sum_{j\in J} m^*(A_j)$.*

(x) *(Countable sub-additivity) If $(A_j)_{j\in J}$ are a countable collection of subsets of $\mathbf{R}^n$, then $m^*\left(\bigcup_{j\in J} A_j\right) \le \sum_{j\in J} m^*(A_j)$.*

(xiii) *(Translation invariance) If Ω is a subset of $\mathbf{R}^n$, and $x \in \mathbf{R}^n$, then $m^*(x+\Omega) = m^*(\Omega)$.*

Proof See Exercise 7.2.1. □

The outer measure of a closed box is also what we expect:

Proposition 7.2.6 (Outer measure of closed box) *For any closed box*

$$B = \prod_{i=1}^n [a_i, b_i] := \{(x_1, \dots, x_n) \in \mathbf{R}^n : x_i \in [a_i, b_i] \textit{ for all } 1 \le i \le n\},$$

we have

$$m^*(B) = \prod_{i=1}^n (b_i - a_i).$$

Proof Clearly, we can cover the closed box $B = \prod_{i=1}^n [a_i, b_i]$ by the open box $\prod_{i=1}^n (a_i - \varepsilon, b_i + \varepsilon)$ for every $\varepsilon > 0$. Thus we have

$$m^*(B) \le \text{vol}\left(\prod_{i=1}^n (a_i - \varepsilon, b_i + \varepsilon)\right) = \prod_{i=1}^n (b_i - a_i + 2\varepsilon)$$

for every $\varepsilon > 0$. Taking limits as $\varepsilon \to 0$, we obtain

$$m^*(B) \le \prod_{i=1}^n (b_i - a_i).$$

To finish the proof, we need to show that

$$m^*(B) \ge \prod_{i=1}^n (b_i - a_i).$$

By the definition of $m^*(B)$, it suffices to show that

$$\sum_{j\in J} \text{vol}(B_j) \ge \prod_{i=1}^n (b_i - a_i)$$

whenever $(B_j)_{j\in J}$ is a finite or countable cover of B.

Since B is closed and bounded, it is compact (by the Heine–Borel theorem, Theorem 1.5.7), and in particular every open cover has a finite subcover (Theorem 1.5.8). Thus to prove the above inequality for countable covers, it suffices to do it for finite covers (since if $(B_j)_{j\in J'}$ is a finite subcover of $(B_j)_{j\in J}$ then $\sum_{j\in J}\mathrm{vol}(B_j)$ will be greater than or equal to $\sum_{j\in J'}\mathrm{vol}(B_j)$).

To summarize, our goal is now to prove that

$$\sum_{j\in J}\mathrm{vol}(B^{(j)}) \geq \prod_{i=1}^{n}(b_i - a_i) \tag{7.1}$$

whenever $(B^{(j)})_{j\in J}$ is a finite cover of $\prod_{i=1}^{n}[a_i, b_i]$; we have changed the subscript B_j to superscript $B^{(j)}$ because we will need the subscripts to denote components.

To prove the inequality (7.1), we shall use induction on the dimension n. First we consider the base case $n = 1$. Here B is just a closed interval $B = [a, b]$, and each box $B^{(j)}$ is just an open interval $B^{(j)} = (a_j, b_j)$. We have to show that

$$\sum_{j\in J}(b_j - a_j) \geq (b - a).$$

To do this we use the Riemann integral. For each $j \in J$, let $f^{(j)} : \mathbf{R} \to \mathbf{R}$ be the function such that $f^{(j)}(x) = 1$ when $x \in (a_j, b_j)$ and $f^{(j)}(x) = 0$ otherwise. Then we have that $f^{(j)}$ is Riemann integrable (because it is piecewise constant, and compactly supported) and

$$\int_{-\infty}^{\infty} f^{(j)} = b_j - a_j.$$

Summing this over all $j \in J$, and interchanging the integral with the finite sum, we have

$$\int_{-\infty}^{\infty}\sum_{j\in J} f^{(j)} = \sum_{j\in J} b_j - a_j.$$

But since the intervals (a_j, b_j) cover $[a, b]$, we have $\sum_{j\in J} f^{(j)}(x) \geq 1$ for all $x \in [a, b]$ (why?). For all other values of x, we have $\sum_{j\in J} f^{(j)}(x) \geq 0$. Thus

$$\int_{-\infty}^{\infty}\sum_{j\in J} f^{(j)} \geq \int_{[a,b]} 1 = b - a$$

and the claim follows by combining this inequality with the previous equality. This proves (7.1) when $n = 1$.

Now assume inductively that $n > 1$, and we have already proven the inequality (7.1) for dimensions $n-1$. We shall use a similar argument to the preceding one. Each box $B^{(j)}$ is now of the form

$$B^{(j)} = \prod_{i=1}^{n} (a_i^{(j)}, b_i^{(j)}).$$

We can write this as

$$B^{(j)} = A^{(j)} \times (a_n^{(j)}, b_n^{(j)})$$

where $A^{(j)}$ is the $n-1$-dimensional box $A^{(j)} := \prod_{i=1}^{n-1} (a_i^{(j)}, b_i^{(j)})$. Note that

$$\mathrm{vol}(B^{(j)}) = \mathrm{vol}_{n-1}(A^{(j)})(b_n^{(j)} - a_n^{(j)})$$

where we have subscripted vol_{n-1} by $n-1$ to emphasize that this is $n-1$-dimensional volume being referred to here. We similarly write

$$B = A \times [a_n, b_n]$$

where $A := \prod_{i=1}^{n-1} [a_i, b_i]$, and again note that

$$\mathrm{vol}(B) = \mathrm{vol}_{n-1}(A)(b_n - a_n).$$

For each $j \in J$, let $f^{(j)}$ be the function such that $f^{(j)}(x_n) = \mathrm{vol}_{n-1}(A^{(j)})$ for all $x_n \in (a_n^{(j)}, b_n^{(j)})$, and $f^{(j)}(x_n) = 0$ for all other x_n. Then $f^{(j)}$ is Riemann integrable and

$$\int_{-\infty}^{\infty} f^{(j)} = \mathrm{vol}_{n-1}(A^{(j)})(b_n^{(j)} - a_n^{(j)}) = \mathrm{vol}(B^{(j)})$$

and hence

$$\sum_{j \in J} \mathrm{vol}(B^{(j)}) = \int_{-\infty}^{\infty} \sum_{j \in J} f^{(j)}.$$

Now let $x_n \in [a_n, b_n]$ and $(x_1, \ldots, x_{n-1}) \in A$. Then $(x_1, \ldots, x_n)$ lies in B, and hence lies in one of the $B^{(j)}$. Clearly we have $x_n \in (a_n^{(j)}, b_n^{(j)})$, and $(x_1, \ldots, x_{n-1}) \in A^{(j)}$. In particular, we see that for each $x_n \in [a_n, b_n]$, the set

$$\{A^{(j)} : j \in J; x_n \in (a_n^{(j)}, b_n^{(j)})\}$$

of $n-1$-dimensional boxes covers A. Applying the inductive hypothesis (7.1) at dimension $n-1$ we thus see that

$$\sum_{j\in J: x_n \in (a_n^{(j)}, b_n^{(j)})} \mathrm{vol}_{n-1}(A^{(j)}) \geq \mathrm{vol}_{n-1}(A),$$

or in other words

$$\sum_{j\in J} f^{(j)}(x_n) \geq \mathrm{vol}_{n-1}(A).$$

Integrating this over $[a_n, b_n]$, we obtain

$$\int_{[a_n,b_n]} \sum_{j\in J} f^{(j)} \geq \mathrm{vol}_{n-1}(A)(b_n - a_n) = \mathrm{vol}(B)$$

and in particular

$$\int_{-\infty}^{\infty} \sum_{j\in J} f^{(j)} \geq \mathrm{vol}_{n-1}(A)(b_n - a_n) = \mathrm{vol}(B)$$

since $\sum_{j\in J} f^{(j)}$ is always non-negative. Combining this with our previous identity for $\int_{-\infty}^{\infty} \sum_{j\in J} f^{(j)}$ we obtain (7.1), and the induction is complete. □

Once we obtain the measure of a closed box, the corresponding result for an open box is easy:

Corollary 7.2.7 *For any open box*

$$B = \prod_{i=1}^{n}(a_i, b_i) := \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_i \in (a_i, b_i)\, \textit{for all}\ 1 \leq i \leq n\},$$

we have

$$m^*(B) = \prod_{i=1}^{n}(b_i - a_i).$$

In particular, outer measure obeys the normalization (xii).

Proof We may assume that $b_i > a_i$ for all i, since if $b_i = a_i$ this follows from Lemma 7.2.5(v). Now observe that

$$\prod_{i=1}^{n}[a_i + \varepsilon, b_i - \varepsilon] \subseteq \prod_{i=1}^{n}(a_i, b_i) \subseteq \prod_{i=1}^{n}[a_i, b_i]$$

for all $\varepsilon > 0$, assuming that ε is small enough that $b_i - \varepsilon > a_i + \varepsilon$ for all i. Applying Proposition 7.2.6 and Lemma 7.2.5(vii) we obtain

$$\prod_{i=1}^{n}(b_i - a_i - 2\varepsilon) \leq m^*\left(\prod_{i=1}^{n}(a_i, b_i)\right) \leq \prod_{i=1}^{n}(b_i - a_i).$$

Sending $\varepsilon \to 0$ and using the squeeze test (Corollary 6.4.14), one obtains the result. □

We now compute some examples of outer measure on the real line $\mathbf{R}$.

Example 7.2.8 Let us compute the one-dimensional measure of $\mathbf{R}$. Since $(-R, R) \subseteq \mathbf{R}$ for all $R > 0$, we have

$$m^*(\mathbf{R}) \geq m^*((-R, R)) = 2R$$

by Corollary 7.2.7. Letting $R \to +\infty$ we thus see that $m^*(\mathbf{R}) = +\infty$.

Example 7.2.9 Now let us compute the one-dimensional measure of $\mathbf{Q}$. From Proposition 7.2.6 we see that for each rational number $\mathbf{Q}$, the point $\{q\}$ has outer measure $m^*(\{q\}) = 0$. Since $\mathbf{Q}$ is clearly the union $\mathbf{Q} = \bigcup_{q \in \mathbf{Q}}\{q\}$ of all these rational points q, and $\mathbf{Q}$ is countable, we have

$$m^*(\mathbf{Q}) \leq \sum_{q \in \mathbf{Q}} m^*(\{q\}) = \sum_{q \in \mathbf{Q}} 0 = 0,$$

and so $m^*(Q)$ must equal zero. In fact, the same argument shows that every countable set has measure zero. (This, incidentally, gives another proof that the real numbers are uncountable, Corollary 8.3.4.)

Remark 7.2.10 One consequence of the fact that $m^*(\mathbf{Q}) = 0$ is that given any $\varepsilon > 0$, it is possible to cover the rationals $\mathbf{Q}$ by a countable number of intervals whose total length is less than ε. This fact is somewhat un-intuitive; can you find a more explicit way to construct such a countable covering of $\mathbf{Q}$ by short intervals?

Example 7.2.11 Now let us compute the one-dimensional measure of the irrationals $\mathbf{R}\backslash\mathbf{Q}$. From finite sub-additivity we have

$$m^*(\mathbf{R}) \leq m^*(\mathbf{R}\backslash\mathbf{Q}) + m^*(\mathbf{Q}).$$

Since $\mathbf{Q}$ has outer measure 0, and $m^*(\mathbf{R})$ has outer measure $+\infty$, we thus see that the irrationals $\mathbf{R}\backslash\mathbf{Q}$ have outer measure $+\infty$. A similar argument shows that $[0, 1]\backslash\mathbf{Q}$, the irrationals in $[0, 1]$, have outer measure 1 (why?).

Example 7.2.12 By Proposition 7.2.6, the unit interval $[0, 1]$ in $\mathbf{R}$ has one-dimensional outer measure 1, but the unit interval $\{(x, 0) : 0 \leq x \leq 1\}$ in $\mathbf{R}^2$ has two-dimensional outer measure 0. Thus one-dimensional outer measure and two-dimensional outer measure are quite different. Note that the above remarks and countable sub-additivity imply that the entire x-axis of $\mathbf{R}^2$ has two-dimensional outer measure 0, despite the fact that $\mathbf{R}$ has infinite one-dimensional measure.

— Exercises —

Exercise 7.2.1 Prove Lemma 7.2.5. (*Hint:* you will have to use the definition of inf, and probably introduce a parameter ε. You may have to treat separately the cases when certain outer measures are equal to $+\infty$. (viii) can be deduced from (x) and (v). For (x), label the index set J as $J = \{j_1, j_2, j_3, \ldots\}$, and for each A_j, pick a covering of A_j by boxes whose total volume is no larger than $m^*(A_j) + \varepsilon/2^j$.)

Exercise 7.2.2 Let A be a subset of $\mathbf{R}^n$, and let B be a subset of $\mathbf{R}^m$. Note that the Cartesian product $\{(a, b) : a \in A, b \in B\}$ is then a subset of $\mathbf{R}^{n+m}$. Show that $m^*_{n+m}(A \times B) \leq m^*_n(A)m^*_m(B)$. Here we adopt the convention that $c \times +\infty = +\infty \times c$ is equal to $+\infty$ for any $0 < c \leq +\infty <$ and equal to zero for $c = 0$. (It is in fact true that $m^*_{n+m}(A \times B) = m^*_n(A)m^*_m(B)$, but this is substantially harder to prove.)

In Exercises 7.2.3–7.2.5, we assume that $\mathbf{R}^n$ is a Euclidean space, and we have a notion of measurable set in $\mathbf{R}^n$ (which may or may not coincide with the notion of Lebesgue measurable set) and a notion of measure (which may or may not coincide with Lebesgue measure) which obeys axioms (i)–(xiii).

Exercise 7.2.3 (a) Show that if $A_1 \subseteq A_2 \subseteq A_3 \ldots$ is an increasing sequence of measurable sets (so $A_j \subseteq A_{j+1}$ for every positive integer j), then we have $m\left(\bigcup_{j=1}^{\infty} A_j\right) = \lim_{j\to\infty} m(A_j)$.

(b) Show that if $A_1 \supseteq A_2 \supseteq A_3 \ldots$ is a decreasing sequence of measurable sets (so $A_j \supseteq A_{j+1}$ for every positive integer j), and $m(A_1) < +\infty$, then we have $m\left(\bigcap_{j=1}^{\infty} A_j\right) = \lim_{j\to\infty} m(A_j)$.

Exercise 7.2.4 Show that for any positive integer $q > 1$, that the open box

$$(0, 1/q)^n := \{(x_1, \ldots, x_n) \in \mathbf{R}^n : 0 < x_j < 1/q \text{ for all } 1 \leq j \leq n\}$$

and the closed box

$$[0, 1/q]^n := \{(x_1, \ldots, x_n) \in \mathbf{R}^n : 0 \leq x_j \leq 1/q \text{ for all } 1 \leq j \leq n\}$$

both measure q^{-n}. (*Hint:* first show that $m((0, 1/q)^n) \leq q^{-n}$ for every $q \geq 1$ by covering $(0, 1)^n$ by some translates of $(0, 1/q)^n$. Using a similar argument, show that $m([0, 1/q]^n) \geq q^{-n}$. Then show that $m([0, 1/q]^n \backslash (0, 1/q)^n) \leq \varepsilon$ for every $\varepsilon > 0$, by covering the boundary of $[0, 1/q]^n$ with some very small boxes.)

Exercise 7.2.5 Show that for any box B, that $m(B) = \text{vol}(B)$. (*Hint:* first prove this when the co-ordinates a_j, b_j are rational, using Exercise 7.2.4. Then take limits somehow (perhaps using Q1) to obtain the general case when the co-ordinates are real.)

Exercise 7.2.6 Use Lemma 7.2.5 and Proposition 7.2.6 to furnish another proof that the reals are uncountable (i.e., reprove Corollary 8.3.4 from *Analysis I*).

7.3 Outer Measure Is not Additive

In light of Lemma 7.2.5, it would seem now that all we need to do is to verify the additivity properties (ix), (xi), and we have everything we need to have a usable measure. Unfortunately, these properties fail for outer measure, even in one dimension.

Proposition 7.3.1 (Failure of countable additivity) *There exists a countable collection $(A_j)_{j\in J}$ of disjoint subsets of* $\mathbf{R}$, *such that* $m^*(\bigcup_{j\in J} A_j) \neq \sum_{j\in J} m^*(A_j)$.

Proof We shall need some notation. Let $\mathbf{Q}$ be the rationals, and $\mathbf{R}$ be the reals. We say that a set $A \subseteq \mathbf{R}$ is a *coset* of $\mathbf{Q}$ if it is of the form $A = x + \mathbf{Q}$ for some real number x. For instance, $\sqrt{2} + \mathbf{Q}$ is a coset of $\mathbf{Q}$, as is $\mathbf{Q}$ itself, since $\mathbf{Q} = 0 + \mathbf{Q}$. Note that a coset A can correspond to several values of x; for instance $2 + \mathbf{Q}$ is exactly the same coset as $0 + \mathbf{Q}$. Also observe that it is not possible for two cosets to partially overlap; if $x + \mathbf{Q}$ intersects $y + \mathbf{Q}$ in even just a single point z, then $x - y$ must be rational (why? Use the identity $x - y = (x - z) - (y - z)$), and thus $x + \mathbf{Q}$ and $y + \mathbf{Q}$ must be equal (why?). So any two cosets are either identical or disjoint.

We observe that every coset A of the rationals $\mathbf{Q}$ has a non-empty intersection with $[0, 1]$. Indeed, if A is a coset, then $A = x + \mathbf{Q}$ for some real number x. If we then pick a rational number q in $[-x, 1 - x]$ then we see that $x + q \in [0, 1]$, and thus $A \cap [0, 1]$ contains $x + q$.

Let $\mathbf{R}/\mathbf{Q}$ denote the set of all cosets of $\mathbf{Q}$; note that this is a set whose elements are themselves sets (of real numbers). For each coset A in $\mathbf{R}/\mathbf{Q}$, let us pick an element x_A of $A \cap [0, 1]$. (This requires us to make an infinite number of choices, and thus requires the axiom of choice, see Sect. 8.4.) Let E be the set of all such x_A, i.e., $E := \{x_A : A \in \mathbf{R}/\mathbf{Q}\}$. Note that $E \subseteq [0, 1]$ by construction.

Now consider the set

$$X = \bigcup_{q\in\mathbf{Q}\cap[-1,1]} (q + E).$$

Clearly this set is contained in $[-1, 2]$ (since $q + x \in [-1, 2]$ whenever $q \in [-1, 1]$ and $x \in E \subseteq [0, 1]$). We claim that this set contains the interval $[0, 1]$. Indeed, for any $y \in [0, 1]$, we know that y must belong to some coset A (for instance, it belongs to the coset $y + \mathbf{Q}$). But we also have x_A belonging to the same coset, and thus $y - x_A$ is equal to some rational q. Since y and x_A both live in $[0, 1]$, then q lives in $[-1, 1]$. Since $y = q + x_A$, we have $y \in q + E$, and hence $y \in X$ as desired.

Note that the translates $q + E$ for $q \in \mathbf{Q}$ are all disjoint. For, if there were two distinct $q, q' \in \mathbf{Q}$ with $q + E$ intersecting $q' + E$, then there would be $A, A' \in \mathbf{R}/\mathbf{Q}$ such that $q + x_A = q' + x_{A'}$. But then $A = x_A + \mathbf{Q} = x_{A'} + \mathbf{Q} = A'$ and thus $x_A = x_{A'}$ which implies that $q = q'$, contradicting the hypothesis.

We claim that

$$m^*(X) \neq \sum_{q\in\mathbf{Q}\cap[-1,1]} m^*(q + E),$$

which would prove the claim. To see why this is true, observe that since $[0, 1] \subseteq X \subseteq [-1, 2]$, that we have $1 \leq m^*(X) \leq 3$ by monotonicity and Proposition 7.2.6.

For the right-hand side, observe from translation invariance that

$$\sum_{q \in \mathbf{Q} \cap [-1,1]} m^*(q+E) = \sum_{q \in \mathbf{Q} \cap [-1,1]} m^*(E).$$

The set $\mathbf{Q} \cap [-1, 1]$ is countably infinite (why?). Thus the right-hand side is either 0 (if $m^*(E) = 0$) or $+\infty$ (if $m^*(E) > 0$). Either way, it cannot be between 1 and 3, and the claim follows. □

Remark 7.3.2 The above proof used the axiom of choice. This turns out to be absolutely necessary; one can prove using some advanced techniques in mathematical logic that if one does not assume the axiom of choice, then it is possible to have a mathematical model where outer measure is countably additive.

One can refine the above argument, and show in fact that m^* is not finitely additive either:

Proposition 7.3.3 (Failure of finite additivity) *There exists a finite collection* $(A_j)_{j \in J}$ *of disjoint subsets of* $\mathbf{R}$, *such that*

$$m^*\left(\bigcup_{j \in J} A_j\right) \neq \sum_{j \in J} m^*(A_j).$$

Proof This is accomplished by an indirect argument. Suppose for sake of contradiction that m^* *was* finitely additive. Let E and X be the sets introduced in Proposition 7.3.1. From countable sub-additivity and translation invariance we have

$$m^*(X) \leq \sum_{q \in \mathbf{Q} \cap [-1,1]} m^*(q+E) = \sum_{q \in \mathbf{Q} \cap [-1,1]} m^*(E).$$

Since we know that $1 \leq m^*(X) \leq 3$, we thus have $m^*(E) \neq 0$, since otherwise we would have $m^*(X) \leq 0$, a contradiction.

Since $m^*(E) \neq 0$, there exists a finite integer $n > 0$ such that $m^*(E) > 1/n$. Now let J be a finite subset of $\mathbf{Q} \cap [-1, 1]$ of cardinality 3n. If m^* were finitely additive, then we would have

$$m^*\left(\bigcup_{q \in J} q + E\right) = \sum_{q \in J} m^*(q+E) = \sum_{q \in J} m^*(E) > 3n\frac{1}{n} = 3.$$

But we know that $\bigcup_{q \in J} q + E$ is a subset of X, which has outer measure at most 3. This contradicts monotonicity. Hence m^* cannot be finitely additive. □

Remark 7.3.4 The examples here are related to the *Banach-Tarski paradox*, which demonstrates (using the axiom of choice) that one can partition the unit ball in $\mathbf{R}^3$

into a finite number of pieces which, when rotated and translated, can be reassembled to form *two* complete unit balls! Of course, this partition involves non-measurable sets. We will not present this paradox here as it requires some group theory which is beyond the scope of this text.

7.4 Measurable Sets

In the previous section we saw that certain sets were badly behaved with respect to outer measure, in particular they could be used to contradict finite or countable additivity. However, those sets were rather pathological, being constructed using the axiom of choice and looking rather artificial. One would hope to be able to exclude them and then somehow recover finite and countable additivity. Fortunately, this can be done, thanks to a clever definition of Constantin Carathéodory (1873–1950):

Definition 7.4.1 (*Lebesgue measurability*) Let E be a subset of $\mathbf{R}^n$. We say that E is *Lebesgue measurable*, or *measurable* for short, iff we have the identity

$$m^*(A) = m^*(A \cap E) + m^*(A \backslash E)$$

for every subset A of $\mathbf{R}^n$. If E is measurable, we define the *Lebesgue measure* of E to be $m(E) = m^*(E)$; if E is not measurable, we leave $m(E)$ undefined.

In other words, E being measurable means that if we use the set E to divide up an arbitrary set A into two parts, we keep the additivity property. Of course, if m^* were finitely additive then every set E would be measurable; but we know from Proposition 7.3.3 that not every set is finitely additive. One can think of the measurable sets as the sets for which finite additivity works. We sometimes subscript $m(E)$ as $m_n(E)$ to emphasize the fact that we are using n-dimensional Lebesgue measure.

The above definition is somewhat hard to work with, and in practice one does not verify a set is measurable directly from this definition. Instead, we will use this definition to prove various useful properties of measurable sets (Lemmas 7.4.2–7.4.11), and after that we will rely more or less exclusively on the properties in those lemmas, and no longer need to refer to the above definition.

We begin by showing that a large number of sets are indeed measurable. The empty set $E = \emptyset$ and the whole space $E = \mathbf{R}^n$ are clearly measurable (why?). Here is another example of a measurable set:

Lemma 7.4.2 (Half-spaces are measurable) *The half-space*

$$\{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_n > 0\}$$

is measurable.

Proof See Exercise 7.4.3. □

Remark 7.4.3 A similar argument will also show that any half-space of the form $\{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_j > 0\}$ or $\{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_j < 0\}$ for some $1 \leq j \leq n$ is measurable.

Now for some more properties of measurable sets.

Lemma 7.4.4 (Properties of measurable sets)

(a) If E is measurable, then $\mathbf{R}^n \backslash E$ is also measurable.
(b) (Translation invariance) If E is measurable, and $x \in \mathbf{R}^n$, then $x + E$ is also measurable, and $m(x + E) = m(E)$.
(c) If E_1 and E_2 are measurable, then $E_1 \cap E_2$ and $E_1 \cup E_2$ are measurable.
(d) (Boolean algebra property) If $E_1, E_2, \ldots, E_N$ are measurable, then $\bigcup_{j=1}^N E_j$ and $\bigcap_{j=1}^N E_j$ are measurable.
(e) Every open box, and every closed box, is measurable.
(f) Any set E of outer measure zero (i.e., $m^*(E) = 0$) is measurable.

Proof See Exercise 7.4.4. □

From Lemma 7.4.4, we have proven properties (ii), (iii), (xiii) on our wish list of measurable sets, and we are making progress toward (i). We also have finite additivity (property (ix) on our wish list):

Lemma 7.4.5 (Finite additivity) *If $(E_j)_{j \in J}$ are a finite collection of* disjoint *measurable sets, then for any set A (not necessarily measurable), we have*

$$m^*\left(A \cap \bigcup_{j \in J} E_j\right) = \sum_{j \in J} m^*(A \cap E_j).$$

Furthermore, we have $m\left(\bigcup_{j \in J} E_j\right) = \sum_{j \in J} m(E_j)$.

Proof See Exercise 7.4.6. □

Remark 7.4.6 Lemma 7.4.5 and Proposition 7.3.3, when combined, imply that there exist non-measurable sets: see Exercise 7.4.5.

Corollary 7.4.7 *If $A \subseteq B$ are two measurable sets, then $B \backslash A$ is also measurable, and*

$$m(B \backslash A) + m(A) = m(B).$$

Proof See Exercise 7.4.7. □

Now we show countable additivity.

Lemma 7.4.8 (Countable additivity) *If $(E_j)_{j \in J}$ are a countable collection of* disjoint *measurable sets, then $\bigcup_{j \in J} E_j$ is measurable, and $m\left(\bigcup_{j \in J} E_j\right) = \sum_{j \in J} m(E_j)$.*

Proof Let $E := \bigcup_{j \in J} E_j$. Our first task will be to show that E is measurable. Thus, let A be an arbitrary set (not necessarily measurable); we need to show that

$$m^*(A) = m^*(A \cap E) + m^*(A \backslash E).$$

Since J is countable, we may write $J = \{j_1, j_2, j_3, \ldots\}$. Note that

$$A \cap E = \bigcup_{k=1}^{\infty} (A \cap E_{j_k})$$

(why?) and hence by countable sub-additivity

$$m^*(A \cap E) \leq \sum_{k=1}^{\infty} m^*(A \cap E_{j_k}).$$

We rewrite this as

$$m^*(A \cap E) \leq \sup_{N \geq 1} \sum_{k=1}^{N} m^*(A \cap E_{j_k}).$$

Let F_N be the set $F_N := \bigcup_{k=1}^{N} E_{j_k}$. Since the $A \cap E_{j_k}$ are all disjoint, and their union is $A \cap F_N$, we see from Lemma 7.4.5 that

$$\sum_{k=1}^{N} m^*(A \cap E_{j_k}) = m^*(A \cap F_N)$$

and hence

$$m^*(A \cap E) \leq \sup_{N \geq 1} m^*(A \cap F_N).$$

Now we look at $A \backslash E$. Since $F_N \subseteq E$ (why?), we have $A \backslash E \subseteq A \backslash F_N$ (why?). By monotonicity, we thus have

$$m^*(A \backslash E) \leq m^*(A \backslash F_N)$$

for all N. In particular, we see that

$$\begin{aligned} m^*(A \cap E) + m^*(A \backslash E) &\leq \sup_{N \geq 1} \left(m^*(A \cap F_N) + m^*(A \backslash E) \right) \\ &\leq \sup_{N \geq 1} \left(m^*(A \cap F_N) + m^*(A \backslash F_N) \right). \end{aligned}$$

But from Lemma 7.4.4(d) we know that F_N is measurable, and hence

$$m^*(A \cap F_N) + m^*(A \backslash F_N) = m^*(A).$$

Putting this all together we obtain

$$m^*(A \cap E) + m^*(A \backslash E) \leq m^*(A).$$

But from finite sub-additivity we have

$$m^*(A \cap E) + m^*(A \backslash E) \geq m^*(A)$$

and the claim follows. This shows that E is measurable.

To finish the lemma, we need to show that $m(E)$ is equal to $\sum_{j \in J} m(E_j)$. We first observe from countable sub-additivity that

$$m(E) \leq \sum_{j \in J} m(E_j) = \sum_{k=1}^{\infty} m(E_{j_k}).$$

On the other hand, by finite additivity and monotonicity we have

$$m(E) \geq m(F_N) = \sum_{k=1}^{N} m(E_{j_k}).$$

Taking limits as $N \to \infty$ we obtain

$$m(E) \geq \sum_{k=1}^{\infty} m(E_{j_k})$$

and thus we have

$$m(E) = \sum_{k=1}^{\infty} m(E_{j_k}) = \sum_{j \in J} m(E_j)$$

as desired. □

This proves property (xi) on our wish list. Next, we do countable unions and intersections.

Lemma 7.4.9 (σ-algebra property) *If $(\Omega_j)_{j \in J}$ are any countable collection of measurable sets (so J is countable), then the union $\bigcup_{j \in J} \Omega_j$ and the intersection $\bigcap_{j \in J} \Omega_j$ are also measurable.*

Proof See Exercise 7.4.8. □

The final property left to verify on our wish list is (a). We first need a preliminary lemma.

Lemma 7.4.10 *Every open set can be written as a countable or finite union of open boxes.*

Proof We first need some notation. Call a box $B = \prod_{i=1}^n (a_i, b_i)$ *rational* if all of its components a_i, b_i are rational numbers. Observe that there are only a countable number of rational boxes (this is since a rational box is described by $2n$ rational numbers, and so has the same cardinality as $\mathbf{Q}^{2n}$. But $\mathbf{Q}$ is countable, and the Cartesian product of any finite number of countable sets is countable; see Corollaries 8.1.14, 8.1.15).

We make the following claim: given any open ball $B(x, r)$, there exists a rational box B which is contained in $B(x, r)$ and which contains x. To prove this claim, write $x = (x_1, \ldots, x_n)$. For each $1 \leq i \leq n$, let a_i and b_i be rational numbers such that

$$x_i - \frac{r}{n} < a_i < x_i < b_i < x_i + \frac{r}{n}.$$

Then it is clear that the box $\prod_{i=1}^n (a_i, b_i)$ is rational and contains x. A simple computation using Pythagoras' theorem (or the triangle inequality) also shows that this box is contained in $B(x, r)$; we leave this to the reader.

Now let E be an open set, and let Σ be the set of all rational boxes B which are subsets of E, and consider the union $\bigcup_{B \in \Sigma} B$ of all those boxes. Clearly, this union is contained in E, since every box in Σ is contained in E by construction. On the other hand, since E is open, we see that for every $x \in E$ there is a ball $B(x, r)$ contained in E, and by the previous claim this ball contains a rational box which contains x. In particular, x is contained in $\bigcup_{B \in \Sigma} B$. Thus we have

$$E = \bigcup_{B \in \Sigma} B$$

as desired; note that Σ is countable or finite because it is a subset of the set of all rational boxes, which is countable. □

Lemma 7.4.11 (Borel property) *Every open set, and every closed set, is Lebesgue measurable.*

Proof It suffices to do this for open sets, since the claim for closed sets then follows by Lemma 7.4.4(a) (i.e., property (ii)). Let E be an open set. By Lemma 7.4.10, E is the countable union of boxes. Since we already know that boxes are measurable, and that the countable union of measurable sets is measurable, the claim follows. □

The construction of Lebesgue measure and its basic properties are now complete. Now we make the next step in constructing the Lebesgue integral—describing the class of functions we can integrate.

— Exercises —

Exercise 7.4.1 If A is an open interval in $\mathbf{R}$, show that $m^*(A) = m^*(A \cap (0, \infty)) + m^*(A \backslash (0, \infty))$.

Exercise 7.4.2 If A is an open box in $\mathbf{R}^n$, and E is the half-plane $E := \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_n > 0\}$, show that $m^*(A) = m^*(A \cap E) + m^*(A \backslash E)$. (*Hint:* use Exercise 7.4.1.)

Exercise 7.4.3 Prove Lemma 7.4.2. (*Hint:* use Exercise 7.4.2.)

Exercise 7.4.4 Prove Lemma 7.4.4. (*Hints:* for (c), first prove that

$$m^*(A) = m^*(A \cap E_1 \cap E_2) + m^*(A \cap E_1 \backslash E_2) + m^*(A \cap E_2 \backslash E_1) + m^*(A \backslash (E_1 \cup E_2)).$$

A Venn diagram may be helpful. Also you may need the finite sub-additivity property. Use (c) to prove (d), and use (bd) and the various versions of Lemma 7.4.2 to prove (e)).

Exercise 7.4.5 Show that the set E used in the proof of Propositions 7.3.1 and 7.3.3 is non-measurable.

Exercise 7.4.6 Prove Lemma 7.4.5.

Exercise 7.4.7 Use Lemma 7.4.5 to prove Corollary 7.4.7.

Exercise 7.4.8 Prove Lemma 7.4.9. (*Hint:* for the countable union problem, write $J = \{j_1, j_2, \ldots\}$, write $F_N := \bigcup_{k=1}^N \Omega_{j_k}$, and write $E_N := F_N \backslash F_{N-1}$, with the understanding that F_0 is the empty set. Then apply Lemma 7.4.8. For the countable intersection problem, use what you just did and Lemma 7.4.4(a).)

Exercise 7.4.9 Let $A \subseteq \mathbf{R}^2$ be the set $A := [0, 1]^2 \backslash \mathbf{Q}^2$; i.e., A consists of all the points (x, y) in $[0, 1]^2$ such that x and y are not both rational. Show that A is measurable and $m(A) = 1$, but that A has no interior points. (*Hint:* it's easier to use the properties of outer measure and measure, including those in the exercises above, than to try to do this problem from first principles.)

Exercise 7.4.10 Let $A \subseteq B \subseteq \mathbf{R}^n$. Show that if B is Lebesgue measurable with measure zero, then A is also Lebesgue measurable with measure zero.

7.5 Measurable Functions

In the theory of the Riemann integral, we are only able to integrate a certain class of functions—the Riemann integrable functions. We will now be able to integrate a much larger range of functions—the *measurable functions*. More precisely, we can only integrate those measurable functions which are absolutely integrable—but more on that later.

Definition 7.5.1 (*Measurable functions*) Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}^m$ be a function. A function f is *measurable* iff $f^{-1}(V)$ is measurable for every open set $V \subseteq \mathbf{R}^m$.

As discussed earlier, most sets that we deal with in real life are measurable, so it is only natural to learn that most functions we deal with in real life are also measurable. For instance, continuous functions are automatically measurable:

Lemma 7.5.2 (Continuous functions are measurable) *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f: \Omega \to \mathbf{R}^m$ be continuous. Then f is also measurable.*

Proof Let V be any open subset of $\mathbf{R}^m$. Then since f is continuous, $f^{-1}(V)$ is open relative to Ω (see Theorem 2.1.5(c)), i.e., $f^{-1}(V) = W \cap \Omega$ for some open set $W \subseteq \mathbf{R}^n$ (see Proposition 1.3.4(a)). Since W is open, it is measurable; since Ω is measurable, $W \cap \Omega$ is also measurable. □

Because of Lemma 7.4.10, we have an easy criterion to test whether a function is measurable or not:

Lemma 7.5.3 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f: \Omega \to \mathbf{R}^m$ be a function. Then f is measurable if and only if $f^{-1}(B)$ is measurable for every open box B.*

Proof See Exercise 7.5.1. □

Corollary 7.5.4 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f: \Omega \to \mathbf{R}^m$ be a function. Suppose that $f = (f_1, \ldots, f_m)$, where $f_j: \Omega \to \mathbf{R}$ is the jth co-ordinate of f. Then f is measurable if and only if all of the f_j are individually measurable.*

Proof See Exercise 7.5.2. □

Unfortunately, it is not true that the composition of two measurable functions is automatically measurable; however we can do the next best thing: a continuous function applied to a measurable function is measurable.

Lemma 7.5.5 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let W be an open subset of $\mathbf{R}^m$. If $f: \Omega \to W$ is measurable, and $g: W \to \mathbf{R}^p$ is continuous, then $g \circ f: \Omega \to \mathbf{R}^p$ is measurable.*

Proof See Exercise 7.5.3. □

This has an immediate corollary:

Corollary 7.5.6 *Let Ω be a measurable subset of $\mathbf{R}^n$. If $f: \Omega \to \mathbf{R}$ is a measurable function, then so is $|f|$,* $\max(f, 0)$*, and* $\min(f, 0)$.

Proof Apply Lemma 7.5.5 with $g(x) := |x|$, $g(x) := \max(x, 0)$, and $g(x) := \min(x, 0)$. □

A slightly less immediate corollary:

Corollary 7.5.7 *Let Ω be a measurable subset of $\mathbf{R}^n$. If $f: \Omega \to \mathbf{R}$ and $g: \Omega \to \mathbf{R}$ are measurable functions, then so is $f + g$, $f - g$, fg,* $\max(f, g)$*, and* $\min(f, g)$. *If $g(x) \neq 0$ for all $x \in \Omega$, then f/g is also measurable.*

Proof Consider $f + g$. We can write this as $k \circ h$, where $h \colon \Omega \to \mathbf{R}^2$ is the function $h(x) = (f(x), g(x))$, and $k : \mathbf{R}^2 \to \mathbf{R}$ is the function $k(a, b) := a + b$. Since f, g are measurable, then h is also measurable by Corollary 7.5.4. Since k is continuous, we thus see from Lemma 7.5.5 that $k \circ h$ is measurable, as desired. A similar argument deals with all the other cases; the only thing concerning the f/g case is that the space $\mathbf{R}^2$ must be replaced with $\{(a, b) \in \mathbf{R}^2 : b \neq 0\}$ in order to keep the map $(a, b) \mapsto a/b$ continuous and well-defined. □

Another characterization of measurable functions is given by

Lemma 7.5.8 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f \colon \Omega \to \mathbf{R}$ be a function. Then f is measurable if and only if $f^{-1}((a, \infty))$ is measurable for every real number a.*

Proof See Exercise 7.5.4. □

Inspired by this lemma, we extend the notion of a measurable function to the extended real number system $\mathbf{R}^* := \mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$:

Definition 7.5.9 (*Measurable functions in the extended reals*) Let Ω be a measurable subset of $\mathbf{R}^n$. A function $f \colon \Omega \to \mathbf{R}^*$ is said to be *measurable* iff $f^{-1}((a, +\infty])$ is measurable for every real number a.

Note that Lemma 7.5.8 ensures that the notion of measurability for functions taking values in the extended reals $\mathbf{R}^*$ is compatible with that for functions taking values in just the reals $\mathbf{R}$.

Measurability behaves well with respect to limits:

Lemma 7.5.10 (Limits of measurable functions are measurable) *Let Ω be a measurable subset of $\mathbf{R}^n$. For each positive integer n, let $f_n : \Omega \to \mathbf{R}^*$ be a measurable function. Then the functions* $\sup_{n \geq 1} f_n$, $\inf_{n \geq 1} f_n$, $\limsup_{n \to \infty} f_n$, *and* $\liminf_{n \to \infty} f_n$ *are also measurable. In particular, if the f_n converge pointwise to another function $f \colon \Omega \to \mathbf{R}^*$, then f is also measurable.*

Proof We first prove the claim about $\sup_{n \geq 1} f_n$. Call this function g. We have to prove that $g^{-1}((a, +\infty])$ is measurable for every a. But by the definition of supremum, we have

$$g^{-1}((a, +\infty]) = \bigcup_{n \geq 1} f_n^{-1}((a, +\infty])$$

(why?), and the claim follows since the countable union of measurable sets is again measurable.

A similar argument works for $\inf_{n \geq 1} f_n$. The claim for lim sup and lim inf then follow from the identities

$$\limsup_{n \to \infty} f_n = \inf_{N \geq 1} \sup_{n \geq N} f_n$$

and

$$\liminf_{n\to\infty} f_n = \sup_{N\geq 1} \inf_{n\geq N} f_n$$

(see Definition 6.4.6). □

As you can see, just about anything one does to a measurable function will produce another measurable function. This is basically why almost every function one deals with in mathematics is measurable. (Indeed, the only way to construct non-measurable functions is via artificial means such as invoking the axiom of choice.)

— Exercises —

Exercise 7.5.1 Prove Lemma 7.5.3. (*Hint:* use Lemma 7.4.10 and the σ-algebra property.)

Exercise 7.5.2 Use Lemma 7.5.3 to deduce Corollary 7.5.4.

Exercise 7.5.3 Prove Lemma 7.5.5.

Exercise 7.5.4 Prove Lemma 7.5.8. (*Hint:* use Lemma 7.5.3. As a preliminary step, you may need to show that if $f^{-1}((a,\infty))$ is measurable for all a, then $f^{-1}([a,\infty))$ is also measurable for all a.)

Exercise 7.5.5 Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be Lebesgue measurable, and let $g\colon \mathbf{R}^n \to \mathbf{R}$ be a function which agrees with f outside of a set of measure zero, thus there exists a set $A \subseteq \mathbf{R}^n$ of measure zero such that $f(x) = g(x)$ for all $x \in \mathbf{R}^n \backslash A$. Show that g is also Lebesgue measurable. (*Hint:* use Exercise 7.4.10.)

Chapter 8
Lebesgue Integration

In Chap. 11, we approached the Riemann integral by first integrating a particularly simple class of functions, namely the *piecewise constant* functions. Among other things, piecewise constant functions only attain a finite number of values (as opposed to most functions in real life, which can take an infinite number of values). Once one learns how to integrate piecewise constant functions, one can then integrate other Riemann integrable functions by a similar procedure.

We shall use a similar philosophy to construct the Lebesgue integral. We shall begin by considering a special subclass of measurable functions—the *simple* functions. Then we will show how to integrate simple functions, and then from there we will integrate all measurable functions (or at least the absolutely integrable ones).

8.1 Simple Functions

Definition 8.1.1 (*Simple functions*) Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f\colon \Omega \to \mathbf{R}$ be a measurable function. We say that f is a *simple function* if the image $f(\Omega)$ is finite. In other words, there exists a finite number of real numbers $c_1, c_2, \ldots, c_N$ such that for every $x \in \Omega$, we have $f(x) = c_j$ for some $1 \leq j \leq N$.

Example 8.1.2 Let Ω be a measurable subset of $\mathbf{R}^n$, and let E be a measurable subset of Ω. We define the *characteristic function* $\chi_E : \Omega \to \mathbf{R}$ by setting $\chi_E(x){:=}1$ if $x \in E$, and $\chi_E(x){:=}0$ if $x \notin E$. (In some texts, χ_E is also written 1_E and is referred to as an *indicator function*.) Then χ_E is a measurable function (why?) and is a simple function, because the image $\chi_E(\Omega)$ is $\{0, 1\}$ (or $\{0\}$ if E is empty, or $\{1\}$ if $E = \Omega$).

We remark on three basic properties of simple functions: that they form a vector space, that they are linear combinations of characteristic functions, and that they approximate measurable functions. More precisely, we have the following three lemmas:

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3_8

Lemma 8.1.3 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}$ and $g : \Omega \to \mathbf{R}$ be simple functions. Then $f + g$ is also a simple function. Also, for any scalar $c \in \mathbf{R}$, the function cf is also a simple function.*

Proof See Exercise 8.1.1. □

Lemma 8.1.4 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}$ be a simple function. Then there exists a finite number of real numbers $c_1, \ldots, c_N$, and a finite number of disjoint measurable sets $E_1, E_2, \ldots, E_N$ in Ω, such that $f = \sum_{i=1}^{N} c_i \chi_{E_i}$.*

Proof See Exercise 8.1.2. □

Lemma 8.1.5 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to [0, +\infty]$ be a measurable function. Then there exists a sequence $f_1, f_2, f_3, \ldots$ of simple functions, $f_n : \Omega \to \mathbf{R}$, such that the f_n are non-negative and increasing,*

$$0 \leq f_1(x) \leq f_2(x) \leq f_3(x) \leq \ldots \textit{ for all } x \in \Omega$$

and converge pointwise to f:

$$\lim_{n\to\infty} f_n(x) = f(x) \textit{ for all } x \in \Omega.$$

Proof See Exercise 8.1.3. □

We now show how to compute the integral of simple functions.

Definition 8.1.6 (*Lebesgue integral of simple functions*) Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}$ be a simple function which is non-negative; thus f is measurable and the image $f(\Omega)$ is finite and contained in $[0, \infty)$. We then define the *Lebesgue integral* $\int_\Omega f$ of f on Ω by

$$\int_\Omega f := \sum_{\lambda \in f(\Omega); \lambda > 0} \lambda m(\{x \in \Omega : f(x) = \lambda\}).$$

We will also sometimes write $\int_\Omega f$ as $\int_\Omega f\, dm$ (to emphasize the rôle of Lebesgue measure m) or use a dummy variable such as x, e.g., $\int_\Omega f(x)\, \mathrm{d}x$.

Example 8.1.7 Let $f : \mathbf{R} \to \mathbf{R}$ be the function which equals 3 on the interval $[1, 2]$, equals 4 on the interval $(2, 4)$, and is zero everywhere else. Then

$$\int_\Omega f := 3 \times m([1, 2]) + 4 \times m((2, 4)) = 3 \times 1 + 4 \times 2 = 11.$$

Or if $g : \mathbf{R} \to \mathbf{R}$ is the function which equals 1 on $[0, \infty)$ and is zero everywhere else, then

$$\int_\Omega g = 1 \times m([0,\infty)) = 1 \times +\infty = +\infty.$$

Thus the simple integral of a simple function can equal $+\infty$. (The reason why we restrict this integral to non-negative functions is to avoid ever encountering the indefinite form $+\infty + (-\infty)$.)

Remark 8.1.8 Note that this definition of integral corresponds to one's intuitive notion of integration (at least of non-negative functions) as the area under the graph of the function (or volume, if one is in higher dimensions).

Another formulation of the integral for non-negative simple functions is as follows.

Lemma 8.1.9 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $E_1, \ldots, E_N$ br a finite number of disjoint measurable subsets in Ω. Let $c_1, \ldots, c_N$ be non-negative numbers (not necessarily distinct). Then we have*

$$\int_\Omega \sum_{j=1}^N c_j \chi_{E_j} = \sum_{j=1}^N c_j m(E_j).$$

Proof We can assume that none of the c_j are zero, since we can just remove them from the sum on both sides of the equation. Let $f := \sum_{j=1}^N c_j \chi_{E_j}$. Then $f(x)$ is either equal to one of the c_j (if $x \in E_j$) or equal to 0 (if $x \notin \bigcup_{j=1}^N E_j$). Thus f is a simple function, and $f(\Omega) \subseteq \{0\} \cup \{c_j : 1 \leq j \leq N\}$. Thus, by the definition,

$$\begin{aligned}\int_\Omega f &= \sum_{\lambda \in \{c_j : 1 \leq j \leq N\}} \lambda m(\{x \in \Omega : f(x) = \lambda\}) \\ &= \sum_{\lambda \in \{c_j : 1 \leq j \leq N\}} \lambda m\left(\bigcup_{1 \leq j \leq N : c_j = \lambda} E_j\right).\end{aligned}$$

But by the finite additivity property of Lebesgue measure, this is equal to

$$\begin{aligned}&\sum_{\lambda \in \{c_j : 1 \leq j \leq N\}} \lambda \sum_{1 \leq j \leq N : c_j = \lambda} m(E_j) \\ &= \sum_{\lambda \in \{c_j : 1 \leq j \leq N\}} \sum_{1 \leq j \leq N : c_j = \lambda} c_j m(E_j).\end{aligned}$$

Each j appears exactly once in this sum, since c_j is only equal to exactly one value of λ. So the above expression is equal to $\sum_{1 \leq j \leq N} c_j m(E_j)$ as desired. □

Some basic properties of Lebesgue integration of non-negative simple functions:

Proposition 8.1.10 *Let Ω be a measurable set, and let $f : \Omega \to \mathbf{R}$ and $g : \Omega \to \mathbf{R}$ be non-negative simple functions.*

(a) We have $0 \leq \int_\Omega f \leq \infty$*. Furthermore, we have* $\int_\Omega f = 0$ *if and only if* $m(\{x \in \Omega : f(x) \neq 0\}) = 0$*.*
(b) We have $\int_\Omega (f+g) = \int_\Omega f + \int_\Omega g$*.*
(c) For any positive number c, we have $\int_\Omega cf = c\int_\Omega f$*.*
(d) If $f(x) \leq g(x)$ *for all* $x \in \Omega$*, then we have* $\int_\Omega f \leq \int_\Omega g$*.*

We make a very convenient notational convention: if a property $P(x)$ holds for all points in Ω, except for a set of measure zero, then we say that P holds for *almost every* point in Ω. Thus (a) asserts that $\int_\Omega f = 0$ if and only if f is zero for almost every point in Ω.

Proof From Lemma 8.1.4 or from the formula

$$f = \sum_{\lambda \in f(\Omega)\backslash\{0\}} \lambda \chi_{\{x \in \Omega : f(x) = \lambda\}}$$

we can write f as a combination of characteristic functions, say

$$f = \sum_{j=1}^{N} c_j \chi_{E_j},$$

where $E_1, \ldots, E_N$ are disjoint subsets of Ω and the c_j are positive. Similarly we can write

$$g = \sum_{k=1}^{M} d_k \chi_{F_k}$$

where $F_1, \ldots, F_M$ are disjoint subsets of Ω and the d_k are positive.

(a) Since $\int_\Omega f = \sum_{j=1}^N c_j m(E_j)$ it is clear that the integral is between 0 and infinity. If f is zero almost everywhere, then all of the E_j must have measure zero (why?) and so $\int_\Omega f = 0$. Conversely, if $\int_\Omega f = 0$, then $\sum_{j=1}^N c_j m(E_j) = 0$, which can only happen when all of the $m(E_j)$ are zero (since all the c_j are positive). But then $\bigcup_{j=1}^N E_j$ has measure zero, and hence f is zero almost everywhere in Ω.

(b) Write $E_0 := \Omega \backslash \bigcup_{j=1}^N E_j$ and $c_0 := 0$, then we have $\Omega = E_0 \cup E_1 \cup \ldots \cup E_N$ and

$$f = \sum_{j=0}^{N} c_j \chi_{E_j}.$$

Similarly if we write $F_0 := \Omega \backslash \bigcup_{k=1}^M F_k$ and $d_0 := 0$ then

$$g = \sum_{k=0}^{M} d_k \chi_{F_k}.$$

Since $\Omega = E_0 \cup \ldots \cup E_N = F_0 \cup \ldots \cup F_M$, we have

$$f = \sum_{j=0}^{N} \sum_{k=0}^{M} c_j \chi_{E_j \cap F_k}$$

and

$$g = \sum_{k=0}^{M} \sum_{j=0}^{N} d_k \chi_{E_j \cap F_k}$$

and hence

$$f + g = \sum_{0 \leq j \leq N; 0 \leq k \leq M} (c_j + d_k) \chi_{E_j \cap F_k}.$$

By Lemma 8.1.9, we thus have

$$\int_{\Omega} (f + g) = \sum_{0 \leq j \leq N; 0 \leq k \leq M} (c_j + d_k) m(E_j \cap F_k).$$

On the other hand, we have

$$\int_{\Omega} f = \sum_{0 \leq j \leq N} c_j m(E_j) = \sum_{0 \leq j \leq N; 0 \leq k \leq M} c_j m(E_j \cap F_k)$$

and similarly

$$\int_{\Omega} g = \sum_{0 \leq k \leq M} d_k m(F_k) = \sum_{0 \leq j \leq N; 0 \leq k \leq M} d_k m(E_j \cap F_k)$$

and the claim (b) follows.

(c) Since $cf = \sum_{j=1}^{N} c c_j \chi_{E_j}$, we have $\int_{\Omega} cf = \sum_{j=1}^{N} c c_j m(E_j)$. Since $\int_{\Omega} f = \sum_{j=1}^{N} c_j m(E_j)$, the claim follows.

(d) Write $h := g - f$. Then h is simple and non-negative and $g = f + h$, hence by (b) we have $\int_{\Omega} g = \int_{\Omega} f + \int_{\Omega} h$. But by (a) we have $\int_{\Omega} h \geq 0$, and the claim follows.

□

— Exercise —

Exercise 8.1.1 Prove Lemma 8.1.3.

Exercise 8.1.2 Prove Lemma 8.1.4.

Exercise 8.1.3 Prove Lemma 8.1.5. (*Hint*: set

$$f_n(x) := \sup\{\frac{j}{2^n} : j \in \mathbf{Z}, \frac{j}{2^n} \leq \min(f(x), 2^n)\},$$

, i.e., $f_n(x)$ is the greatest integer multiple of 2^{-n} which does not exceed either $f(x)$ or 2^n. You may wish to draw a picture to see how f_1, f_2, f_3, etc., works. Then prove that f_n obeys all the required properties.)

8.2 Integration of Non-negative Measurable Functions

We now pass from the integration of non-negative simple functions to the integration of non-negative measurable functions. We will allow our measurable functions to take the value of $+\infty$ sometimes.

Definition 8.2.1 (*Majorization*) Let $f : \Omega \to \mathbf{R}$ and $g : \Omega \to \mathbf{R}$ be functions. We say that f *majorizes* g, or g *minorizes* f, if we have $f(x) \geq g(x)$ for all $x \in \Omega$.

We sometimes use the phrase "f dominates g" instead of "f majorizes g".

Definition 8.2.2 (*Lebesgue integral for non-negative functions*) Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to [0, \infty]$ be measurable and non-negative. Then we define the *Lebesgue integral* $\int_\Omega f$ of f on Ω to be

$$\int_\Omega f := \sup\left\{\int_\Omega s : s \text{ is simple and non-negative, and minorizes } f\right\}.$$

Remark 8.2.3 The reader should compare this notion to that of a lower Riemann integral from Definition 11.3.2. Interestingly, we will not need to match this lower integral with an upper integral here.

Remark 8.2.4 Note that if Ω' is any measurable subset of Ω, then we can define $\int_{\Omega'} f$ as well by restricting f to Ω', thus $\int_{\Omega'} f := \int_{\Omega'} f|_{\Omega'}$.

We have to check that this definition is consistent with our previous notion of Lebesgue integral for non-negative simple functions; in other words, if $f : \Omega \to \mathbf{R}$ is a non-negative simple function, then the value of $\int_\Omega f$ given by this definition should be the same as the one given in the previous definition. But this is clear because f certainly minorizes itself, and any other non-negative simple function s which minorizes f will have an integral $\int_\Omega s$ less than or equal to $\int_\Omega f$, thanks to Proposition 8.1.10(d).

Remark 8.2.5 Note that $\int_\Omega f$ is always at least 0, since 0 is simple, non-negative, and minorizes f. Of course, $\int_\Omega f$ could equal $+\infty$.

Some basic properties of the Lebesgue integral on non-negative measurable functions (which supercede Proposition 8.1.10):

Proposition 8.2.6 *Let Ω be a measurable set, and let $f\colon \Omega \to [0,\infty]$ and $g\colon \Omega \to [0,\infty]$ be non-negative measurable functions.*

(a) *We have $0 \leq \int_\Omega f \leq \infty$. Furthermore, we have $\int_\Omega f = 0$ if and only if $f(x) = 0$ for almost every $x \in \Omega$.*
(b) *For any positive number c, we have $\int_\Omega cf = c \int_\Omega f$.*
(c) *If $f(x) \leq g(x)$ for all $x \in \Omega$, then we have $\int_\Omega f \leq \int_\Omega g$.*
(d) *If $f(x) = g(x)$ for almost every $x \in \Omega$, then $\int_\Omega f = \int_\Omega g$.*
(e) *If $\Omega' \subseteq \Omega$ is measurable, then $\int_{\Omega'} f = \int_\Omega f\chi_{\Omega'} \leq \int_\Omega f$.*

Proof See Exercise 8.2.1. □

Remark 8.2.7 Proposition 8.2.6(d) is quite interesting; it says that one can modify the values of a function on any measure zero set (e.g., you can modify a function on every rational number), and not affect its integral at all. It is as if no individual point, or even a measure zero collection of points, has any "vote" in what the integral of a function should be; only the collective set of points has an influence on an integral.

Remark 8.2.8 Note that we do not yet try to interchange sums and integrals. From the definition it is fairly easy to prove that $\int_\Omega (f+g) \geq \int_\Omega f + \int_\Omega g$ (Exercise 8.2.2), but to prove equality requires more work and will be done later.

As we have seen in previous chapters, we cannot always interchange an integral with a limit (or with limit-like concepts such as supremum). However, with the Lebesgue integral it is possible to do so if the functions are increasing:

Theorem 8.2.9 (Lebesgue monotone convergence theorem) *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $(f_n)_{n=1}^\infty$ be a sequence of non-negative measurable functions from Ω to $[0,+\infty]$ which are increasing in the sense that*

$$0 \leq f_1(x) \leq f_2(x) \leq f_3(x) \leq \ldots \textit{ for all } x \in \Omega.$$

(Note we are assuming that $f_n(x)$ is increasing with respect to n; this is a different notion from $f_n(x)$ increasing with respect to x.) Then we have

$$0 \leq \int_\Omega f_1 \leq \int_\Omega f_2 \leq \int_\Omega f_3 \leq \ldots$$

and

$$\int_\Omega \sup_n f_n = \sup_n \int_\Omega f_n.$$

Proof The first conclusion is clear from Proposition 8.2.6(c). Now we prove the second conclusion. From Proposition 8.2.6(c) again we have

$$\int_\Omega \sup_m f_m \geq \int_\Omega f_n$$

for every n; taking suprema in n we obtain

$$\int_{\Omega} \sup_m f_m \geq \sup_n \int_{\Omega} f_n$$

which is one half of the desired conclusion. To finish the proof we have to show

$$\int_{\Omega} \sup_m f_m \leq \sup_n \int_{\Omega} f_n.$$

From the definition of $\int_{\Omega} \sup_m f_m$, it will suffice to show that

$$\int_{\Omega} s \leq \sup_n \int_{\Omega} f_n$$

for all simple non-negative functions which minorize $\sup_m f_m$.

Fix s. We will show that

$$(1-\varepsilon)\int_{\Omega} s \leq \sup_n \int_{\Omega} f_n$$

for every $0 < \varepsilon < 1$; the claim then follows by taking limits as $\varepsilon \to 0$.

Fix ε. By construction of s, we have

$$s(x) \leq \sup_n f_n(x)$$

for every $x \in \Omega$. Hence, for every $x \in \Omega$ there exists an N (depending on x) such that

$$f_N(x) \geq (1-\varepsilon)s(x).$$

Since the f_n are increasing, this will imply that $f_n(x) \geq (1-\varepsilon)s(x)$ for all $n \geq N$. Thus, if we define the sets E_n by

$$E_n := \{x \in \Omega : f_n(x) \geq (1-\varepsilon)s(x)\}$$

then we have $E_1 \subseteq E_2 \subseteq E_3 \subseteq \ldots$ and $\bigcup_{n=1}^{\infty} E_n = \Omega$.

It is not difficult to check that all the E_n are measurable. From Proposition 8.2.6(bce) we have

$$(1-\varepsilon)\int_{E_n} s = \int_{E_n} (1-\varepsilon)s \leq \int_{E_n} f_n \leq \int_{\Omega} f_n$$

so to finish the argument it will suffice to show that

$$\sup_n \int_{E_n} s = \int_{\Omega} s.$$

Since s is a simple function, we may write $s = \sum_{j=1}^{N} c_j \chi_{F_j}$ for some measurable F_j and positive c_j. Since

$$\int_{\Omega} s = \sum_{j=1}^{N} c_j m(F_j)$$

and

$$\int_{E_n} s = \int_{E_n} \sum_{j=1}^{N} c_j \chi_{F_j \cap E_n} = \sum_{j=1}^{N} c_j m(F_j \cap E_n)$$

it thus suffices to show that

$$\sup_n m(F_j \cap E_n) = m(F_j)$$

for each j. But this follows from Exercise 7.2.3(a). □

This theorem is extremely useful. For instance, we can now interchange addition and integration:

Lemma 8.2.10 (Interchange of addition and integration) *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f: \Omega \to [0, \infty]$ and $g: \Omega \to [0, \infty]$ be measurable functions. Then $\int_{\Omega}(f+g) = \int_{\Omega} f + \int_{\Omega} g$.*

Proof By Lemma 8.1.5, there exists a sequence $0 \le s_1 \le s_2 \le \cdots \le f$ of simple functions such that $\sup_n s_n = f$, and similarly a sequence $0 \le t_1 \le t_2 \le \ldots \le g$ of simple functions such that $\sup_n t_n = g$. Since the s_n are increasing and the t_n are increasing, it is then easy to check that $s_n + t_n$ is also increasing and $\sup_n (s_n + t_n) = f + g$ (why?). By the monotone convergence theorem (Theorem 8.2.9) we thus have

$$\int_{\Omega} f = \sup_n \int_{\Omega} s_n$$

$$\int_{\Omega} g = \sup_n \int_{\Omega} t_n$$

$$\int_{\Omega} (f+g) = \sup_n \int_{\Omega} (s_n + t_n).$$

But by Proposition 8.1.10(db) we have $\int_\Omega (s_n + t_n) = \int_\Omega s_n + \int_\Omega t_n$. By Proposition 8.1.9(d), $\int_\Omega s_n$ and $\int_\Omega t_n$ are both increasing in n, so

$$\sup_n \left(\int_\Omega s_n + \int_\Omega t_n \right) = \left(\sup_n \int_\Omega s_n \right) + \left(\sup_n \int_\Omega t_n \right)$$

and the claim follows. □

Of course, once one can interchange an integral with a sum of two functions, one can handle an integral and any finite number of functions by induction. More surprisingly, one can handle infinite sums as well of *non-negative* functions:

Corollary 8.2.11 *If Ω is a measurable subset of $\mathbf{R}^n$, and $g_1, g_2, \ldots$ are a sequence of non-negative measurable functions from Ω to $[0, \infty]$, then*

$$\int_\Omega \sum_{n=1}^{\infty} g_n = \sum_{n=1}^{\infty} \int_\Omega g_n.$$

Proof See Exercise 8.2.3. □

Remark 8.2.12 Note that we do not need to assume anything about the convergence of the above sums; it may well happen that both sides are equal to $+\infty$. However, we *do* need to assume non-negativity; see Exercise 8.3.4.

One could similarly ask whether we could interchange limits and integrals; in other words, is it true that

$$\int_\Omega \lim_{n\to\infty} f_n = \lim_{n\to\infty} \int_\Omega f_n.$$

Unfortunately, this is not true, as the following "moving bump" example shows. For each $n = 1, 2, 3 \ldots$, let $f_n : \mathbf{R} \to \mathbf{R}$ be the function $f_n = \chi_{[n,n+1)}$. Then $\lim_{n\to\infty} f_n (x) = 0$ for every x, but $\int_\mathbf{R} f_n = 1$ for every n, and hence $\lim_{n\to\infty} \int_\mathbf{R} f_n = 1 \neq 0$. In other words, the limiting function $\lim_{n\to\infty} f_n$ can end up having significantly smaller integral than any of the original integrals. However, the following very useful lemma of Fatou shows that the reverse cannot happen—there is no way the limiting function has larger integral than the (limit of the) original integrals:

Lemma 8.2.13 (Fatou's lemma) *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f_1, f_2, \ldots$ be a sequence of non-negative functions from Ω to $[0, \infty]$. Then*

$$\int_\Omega \liminf_{n\to\infty} f_n \leq \liminf_{n\to\infty} \int_\Omega f_n.$$

Proof Recall that

$$\liminf_{n\to\infty} f_n = \sup_n \left(\inf_{m\geq n} f_m \right)$$

and hence by the monotone convergence theorem

$$\int_\Omega \liminf_{n\to\infty} f_n = \sup_n \int_\Omega \left(\inf_{m\geq n} f_m \right).$$

By Proposition 8.2.6(c) we have

$$\int_\Omega \left(\inf_{m\geq n} f_m \right) \leq \int_\Omega f_j$$

for every $j \geq n$; taking infima in j we obtain

$$\int_\Omega \left(\inf_{m\geq n} f_m \right) \leq \inf_{j\geq n} \int_\Omega f_j.$$

Thus

$$\int_\Omega \liminf_{n\to\infty} f_n \leq \sup_n \inf_{j\geq n} \int_\Omega f_j = \liminf_{n\to\infty} \int_\Omega f_n$$

as desired. □

Note that we are allowing our functions to take the value $+\infty$ at some points. It is even possible for a function to take the value $+\infty$ but still have a finite integral; for instance, if E is a measure zero set, and $f : \Omega \to \mathbf{R}$ is equal to $+\infty$ on E but equals 0 everywhere else, then $\int_\Omega f = 0$ by Proposition 8.2.6(a). However, if the integral is finite, the function must be finite almost everywhere:

Lemma 8.2.14 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to [0, \infty]$ be a non-negative measurable function such that $\int_\Omega f$ is finite. Then f is finite almost everywhere (i.e., the set $\{x \in \Omega : f(x) = +\infty\}$ has measure zero).*

Proof See Exercise 8.2.4. □

Form Corollary 8.2.11 and Lemma 8.2.14 one has a useful lemma:

Lemma 8.2.15 (Borel–Cantelli lemma) *Let $\Omega_1, \Omega_2, \ldots$ be measurable subsets of $\mathbf{R}^n$ such that $\sum_{n=1}^\infty m(\Omega_n)$ is finite. Then the set*

$$\{x \in \mathbf{R}^n : x \in \Omega_n \textit{ for infinitely many } n\}$$

is a set of measure zero. In other words, almost every point belongs to only finitely many Ω_n.

Proof See Exercise 8.2.5. □

— Exercise —

Exercise 8.2.1 Prove Proposition 8.2.6. (*Hint:* do not attempt to mimic the proof of Proposition 8.1.10; rather, try to use Proposition 8.1.10 and Definition 8.2.2. For one direction of part (a), start with $\int_\Omega f = 0$ and conclude that $m(\{x \in \Omega : f(x) > 1/n\}) = 0$ for every $n = 1, 2, 3, \ldots$, and then use the countable subadditivity. To prove (e), first prove it for simple functions.)

Exercise 8.2.2 Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f\colon \Omega \to [0, +\infty]$ and $g\colon \Omega \to [0, +\infty]$ be measurable functions. Without using Theorem 8.2.9 or Lemma 8.2.10, prove that $\int_\Omega (f + g) \geq \int_\Omega f + \int_\Omega g$.

Exercise 8.2.3 Prove Corollary 8.2.11. (*Hint*: use the monotone convergence theorem with $f_N := \sum_{n=1}^N g_n$.)

Exercise 8.2.4 Prove Lemma 8.2.14.

Exercise 8.2.5 Use Corollary 8.2.11 and Lemma 8.2.14 to prove Lemma 8.2.15. (*Hint:* use the indicator functions χ_{Ω_n}.)

Exercise 8.2.6 Let $p > 2$ and $c > 0$. Using the Borel–Cantelli lemma, show that the set

$$\left\{x \in [0, 1] : |x - \frac{a}{q}| \leq \frac{c}{q^p} \text{ for infinitely many positive integers } a, q\right\}$$

has measure zero. (*Hint*: one only has to consider those integers a in the range $0 \leq a \leq q$ (why?). Use Corollary 11.6.5 to show that the sum $\sum_{q=1}^\infty \frac{c(q+1)}{q^p}$ is finite.)

Exercise 8.2.7 Call a real number $x \in \mathbf{R}$ *diophantine* if there exist real numbers $p, C > 0$ such that $|x - \frac{a}{q}| > C/|q|^p$ for all nonzero integers q and all integers a. Using Exercise 8.2.6, show that almost every real number is diophantine. (*Hint:* first work in the interval $[0, 1]$. Show that one can take p and C to be rational and one can also take $p > 2$. Then use the fact that the countable union of measure zero sets has measure zero.)

Exercise 8.2.8 For every positive integer n, let $f_n\colon \mathbf{R} \to [0, \infty)$ be a non-negative measurable function such that

$$\int_{\mathbf{R}} f_n \leq \frac{1}{4^n}.$$

Show that for every $\varepsilon > 0$, there exists a set E of Lebesgue measure $m(E) \leq \varepsilon$ such that $f_n(x)$ converges pointwise to zero for all $x \in \mathbf{R} \backslash E$. (*Hint*: first prove that $m(\{x \in \mathbf{R} : f_n(x) > \frac{1}{\varepsilon 2^n}\}) \leq \frac{\varepsilon}{2^n}$ for all $n = 1, 2, 3, \ldots$, and then consider the union of all the sets $\{x \in \mathbf{R} : f_n(x) > \frac{1}{\varepsilon 2^n}\}$.)

Exercise 8.2.9 For every positive integer n, let $f_n\colon [0,1] \to [0,\infty)$ be a non-negative measurable function such that f_n converges pointwise to zero. Show that for every $\varepsilon > 0$, there exists a set E of Lebesgue measure $m(E) \leq \varepsilon$ such that $f_n(x)$ converges *uniformly* to zero for all $x \in [0,1]\backslash E$. (This is a special case of *Egoroff's theorem*. To prove it, first show that for any positive integer m, we can find an $N > 0$ such that $m(\{x \in [0,1] : f_n(x) > 1/m$ for all $n \geq N\}) \leq \varepsilon/2^m$.) Is the claim still true if $[0,1]$ is replaced by $\mathbf{R}$?

Exercise 8.2.10 Give an example of a bounded non-negative function $f\colon \mathbf{N} \times \mathbf{N} \to \mathbf{R}^+$ such that $\sum_{m=1}^\infty f(n,m)$ converges for every n, and such that $\lim_{n\to\infty} f(n,m)$ exists for every m, but such that

$$\lim_{n\to\infty} \sum_{m=1}^\infty f(n,m) \neq \sum_{m=1}^\infty \lim_{n\to\infty} f(n,m).$$

(*Hint*: modify the moving bump example. It is even possible to use a function f which only takes the values 0 and 1.) This shows that interchanging limits and infinite sums can be dangerous.

8.3 Integration of Absolutely Integrable Functions

We have now completed the theory of the Lebesgue integral for non-negative functions. Now we consider how to integrate functions which can be both positive and negative. However, we do wish to avoid the indefinite expression $+\infty + (-\infty)$, so we will restrict our attention to a subclass of measurable functions—the *absolutely integrable functions*.

Definition 8.3.1 (*Absolutely integrable functions*) Let Ω be a measurable subset of $\mathbf{R}^n$. A measurable function $f\colon \Omega \to \mathbf{R}^*$ is said to be *absolutely integrable* if the integral $\int_\Omega |f|$ is finite.

Of course, $|f|$ is always non-negative, so this definition makes sense even if f changes sign. Absolutely integrable functions are also known as $L^1(\Omega)$ functions.

If $f\colon \Omega \to \mathbf{R}^*$ is a function, we define the *positive part* $f^+\colon \Omega \to [0,\infty]$ and *negative part* $f^-\colon \Omega \to [0,\infty]$ by the formulae

$$f^+ := \max(f,0); \quad f^- := -\min(f,0).$$

From Corollary 7.5.6 (which can be extended to $\mathbf{R}^*$-valued functions without difficulty) we know that f^+ and f^- are measurable. Observe also that f^+ and f^- are non-negative, that $f = f^+ - f^-$, and $|f| = f^+ + f^-$. (Why?).

Definition 8.3.2 (*Lebesgue integral*) Let $f\colon \Omega \to \mathbf{R}^*$ be an absolutely integrable function. We define the *Lebesgue integral* $\int_\Omega f$ of f to be the quantity

$$\int_\Omega f := \int_\Omega f^+ - \int_\Omega f^-.$$

Note that since f is absolutely integrable, $\int_\Omega f^+$ and $\int_\Omega f^-$ are less than or equal to $\int_\Omega |f|$ and hence are finite. Thus $\int_\Omega f$ is always finite; we are never encountering the indeterminate form $+\infty - (+\infty)$.

Note that this definition is consistent with our previous definition of the Lebesgue integral for non-negative functions, since if f is non-negative then $f^+ = f$ and $f^- = 0$. We also have the useful *triangle inequality*

$$\left|\int_\Omega f\right| \leq \int_\Omega f^+ + \int_\Omega f^- = \int_\Omega |f| \tag{8.1}$$

(Exercise 8.3.1).

Some other properties of the Lebesgue integral:

Proposition 8.3.3 *Let Ω be a measurable set, and let $f\colon \Omega \to \mathbf{R}$ and $g\colon \Omega \to \mathbf{R}$ be absolutely integrable functions.*

(a) For any real number c (positive, zero, or negative), we have that cf is absolutely integrable and $\int_\Omega cf = c\int_\Omega f$.
(b) The function $f + g$ is absolutely integrable, and $\int_\Omega (f+g) = \int_\Omega f + \int_\Omega g$.
(c) If $f(x) \leq g(x)$ for all $x \in \Omega$, then we have $\int_\Omega f \leq \int_\Omega g$.
(d) If $f(x) = g(x)$ for almost every $x \in \Omega$, then $\int_\Omega f = \int_\Omega g$.

Proof See Exercise 8.3.2. □

As mentioned in the previous section, one cannot necessarily interchange limits and integrals, $\lim \int f_n = \int \lim f_n$, as the "moving bump example" showed. However, it is possible to exclude the moving bump example and successfully interchange limits and integrals, if we know that the functions f_n are all majorized by a single absolutely integrable function. This important theorem is known as the *Lebesgue dominated convergence theorem* and is extremely useful:

Theorem 8.3.4 (Lebesgue dominated convergence thm) *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f_1, f_2, \ldots$ be a sequence of measurable functions from Ω to $\mathbf{R}^*$ which converge pointwise. Suppose also that there is an absolutely integrable function $F\colon \Omega \to [0, \infty]$ such that $|f_n(x)| \leq F(x)$ for all $x \in \Omega$ and all $n = 1, 2, 3, \ldots$. Then*

$$\int_\Omega \lim_{n\to\infty} f_n = \lim_{n\to\infty} \int_\Omega f_n.$$

Proof If F was infinite on a set of positive measure then F would not be absolutely integrable; thus the set where F is infinite has zero measure. We may delete this set from Ω (this does not affect any of the integrals) and thus assume without loss of generality that $F(x)$ is finite for every $x \in \Omega$, which implies the same assertion for the $f_n(x)$.

Let $f : \Omega \to \mathbf{R}^*$ be the function $f(x) := \lim_{n\to\infty} f_n(x)$; this function exists by hypothesis. By Lemma 7.5.10, f is measurable. Also, since $|f_n(x)| \le F(x)$ for all n and all $x \in \Omega$, we see that each f_n is absolutely integrable, and by taking limits we obtain $|f(x)| \le F(x)$ for all $x \in \Omega$, so f is also absolutely integrable. Our task is to show that $\lim_{n\to\infty} \int_\Omega f_n = \int_\Omega f$.

The functions $F + f_n$ are non-negative and converge pointwise to $F + f$. So by Fatou's lemma (Lemma 8.2.13)

$$\int\limits_\Omega F + f \le \liminf_{n\to\infty} \int\limits_\Omega F + f_n$$

and thus

$$\int\limits_\Omega f \le \liminf_{n\to\infty} \int\limits_\Omega f_n.$$

But the functions $F - f_n$ are also non-negative and converge pointwise to $F - f$. So by Fatou's lemma again

$$\int\limits_\Omega F - f \le \liminf_{n\to\infty} \int\limits_\Omega F - f_n.$$

Since the right-hand side is $\int_\Omega F - \limsup_{n\to\infty} \int_\Omega f_n$ (why did the lim inf become a lim sup?), we thus have

$$\int\limits_\Omega f \ge \limsup_{n\to\infty} \int\limits_\Omega f_n.$$

Thus the lim inf and lim sup of $\int_\Omega f_n$ are both equal to $\int_\Omega f$, as desired. □

Finally, we record a lemma which is not particularly interesting in itself, but will have some useful consequences later in these notes.

Definition 8.3.5 (*(Upper and lower Lebesgue integral)*) Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}$ be a function (not necessarily measurable). We define the *upper Lebesgue integral* $\overline{\int}_\Omega f$ to be

$$\overline{\int\limits_\Omega} f := \inf\Big\{ \int\limits_\Omega g : g \text{ is an absolutely integrable function}$$
$$\text{from } \Omega \text{ to } \mathbf{R} \text{ that majorizes } f \Big\}$$

and the *lower Lebesgue integral* $\underline{\int}_\Omega f$ to be

$$\underline{\int}_{\Omega} f := \sup\left\{ \int_{\Omega} g : g \text{ is an absolutely integrable function} \right.$$
$$\left. \text{from } \Omega \text{ to } \mathbf{R} \text{ that minorizes } f \right\}.$$

It is easy to see that $\underline{\int}_{\Omega} f \leq \overline{\int}_{\Omega} f$ (why? Use Proposition 8.3.3(c)). When f is absolutely integrable then equality occurs (why?). The converse is also true:

Lemma 8.3.6 *Let Ω be a measurable subset of $\mathbf{R}^n$, and let $f : \Omega \to \mathbf{R}$ be a function (not necessarily measurable). Let A be a real number, and suppose $\overline{\int}_{\Omega} f = \underline{\int}_{\Omega} f = A$. Then f is absolutely integrable, and*

$$\int_{\Omega} f = \overline{\int}_{\Omega} f = \underline{\int}_{\Omega} f = A.$$

Proof By definition of upper Lebesgue integral, for every integer $n \geq 1$ we may find an absolutely integrable function $f_n^+ : \Omega \to \mathbf{R}$ which majorizes f such that

$$\int_{\Omega} f_n^+ \leq A + \frac{1}{n}.$$

Similarly we may find an absolutely integrable function $f_n^- : \Omega \to \mathbf{R}$ which minorizes f such that

$$\int_{\Omega} f_n^- \geq A - \frac{1}{n}.$$

Let $F^+ := \inf_n f_n^+$ and $F^- := \sup_n f_n^-$. Then F^+ and F^- are measurable (by Lemma 7.5.10) and absolutely integrable (because they are squeezed between the absolutely integrable functions f_1^+ and f_1^-, for instance). Also, F^+ majorizes f and F^- minorizes f. Finally, we have

$$\int_{\Omega} F^+ \leq \int_{\Omega} f_n^+ \leq A + \frac{1}{n}$$

for every n, and hence

$$\int_{\Omega} F^+ \leq A.$$

Similarly we have

$$\int_{\Omega} F^- \geq A.$$

but F^+ majorizes F^-, and hence $\int_\Omega F^+ \geq \int_\Omega F^-$. Hence we must have

$$\int_\Omega F^+ = \int_\Omega F^- = A.$$

In particular

$$\int_\Omega F^+ - F^- = 0.$$

By Proposition 8.2.6(a), we thus have $F^+(x) = F^-(x)$ for almost every x. But since f is squeezed between F^- and F^+, we thus have $f(x) = F^+(x) = F^-(x)$ for almost every x. In particular, f differs from the absolutely integrable function F^+ only on a set of measure zero and is thus measurable (see Exercise 7.5.5) and absolutely integrable, with

$$\int_\Omega f = \int_\Omega F^+ = \int_\Omega F^- = A$$

as desired. □

— Exercise —

Exercise 8.3.1 Prove (8.1) whenever Ω is a measurable subset of $\mathbf{R}^n$ and f is an absolutely integrable function.

Exercise 8.3.2 Prove Proposition 8.3.3. (*Hint:* for (b), break f, g, and $f+g$ up into positive and negative parts, and try to write everything in terms of integrals of non-negative functions only, using Lemma 8.2.10.)

Exercise 8.3.3 Let $f: \mathbf{R} \to \mathbf{R}$ and $g: \mathbf{R} \to \mathbf{R}$ be absolutely integrable, measurable functions such that $f(x) \leq g(x)$ for all $x \in \mathbf{R}$, and that $\int_{\mathbf{R}} f = \int_R g$. Show that $f(x) = g(x)$ for almost every $x \in \mathbf{R}$ (i.e., that $f(x) = g(x)$ for all $x \in \mathbf{R}$ except possibly for a set of measure zero).

Exercise 8.3.4 For each $n = 1, 2, 3, \ldots$, let $f_n: \mathbf{R} \to \mathbf{R}$ be the function $f_n = \chi_{[n,n+1)} - \chi_{[n+1,n+2)}$; i.e., let $f_n(x)$ equal $+1$ when $x \in [n, n+1)$, equal -1 when $x \in [n+1, n+2)$, and 0 everywhere else. Show that

$$\int_{\mathbf{R}} \sum_{n=1}^{\infty} f_n \neq \sum_{n=1}^{\infty} \int_{\mathbf{R}} f_n.$$

Explain why this does not contradict Corollary 8.2.11.

8.4 Comparison with the Riemann Integral

We have spent a lot of effort constructing the Lebesgue integral, but have not yet addressed the question of how to actually compute any Lebesgue integrals, and whether Lebesgue integration is any different from the Riemann integral (say for integrals in one dimension). Now we show that the Lebesgue integral is a generalization of the Riemann integral. To clarify the following discussion, we shall temporarily distinguish the Riemann integral from the Lebesgue integral by writing the Riemann integral $\int_I f$ as $R. \int_I f$.

Our objective here is to prove

Proposition 8.4.1 *Let $I \subseteq \mathbf{R}$ be a bounded interval, and let $f : I \to \mathbf{R}$ be a Riemann integrable function. Then f is also absolutely integrable, and $\int_I f = R. \int_I f$.*

Proof Write $A := R. \int_I f$. Since f is Riemann integrable, we know that the upper and lower Riemann integrals are equal to A. Thus, for every $\varepsilon > 0$, there exists a partition $\mathbf{P}$ of I into smaller intervals J such that

$$A - \varepsilon \leq \sum_{J \in \mathbf{P}} |J| \inf_{x \in J} f(x) \leq A \leq \sum_{J \in \mathbf{P}} |J| \sup_{x \in J} f(x) \leq A + \varepsilon,$$

where $|J|$ denotes the length of J. Note that $|J|$ is the same as $m(J)$, since J is a box.

Let $f_\varepsilon^- : I \to \mathbf{R}$ and $f_\varepsilon^+ : I \to \mathbf{R}$ be the functions

$$f_\varepsilon^-(x) = \sum_{J \in \mathbf{P}} \inf_{x \in J} f(x) \chi_J(x)$$

and

$$f_\varepsilon^+(x) = \sum_{J \in \mathbf{P}} \sup_{x \in J} f(x) \chi_J(x);$$

these are simple functions and hence measurable and absolutely integrable. By Lemma 8.1.9 we have

$$\int_I f_\varepsilon^- = \sum_{J \in \mathbf{P}} |J| \inf_{x \in J} f(x)$$

and

$$\int_I f_\varepsilon^+ = \sum_{J \in \mathbf{P}} |J| \sup_{x \in J} f(x)$$

and hence

$$A - \varepsilon \leq \int_I f_\varepsilon^- \leq A \leq \int_I f_\varepsilon^+ \leq A + \varepsilon.$$

Since f_ε^+ majorizes f, and f_ε^- minorizes f, we thus have

$$A - \varepsilon \leq \underline{\int}_I f \leq \overline{\int}_I f \leq A + \varepsilon$$

for every ε, and thus

$$\underline{\int}_I f = \overline{\int}_I f = A$$

and hence by Lemma 8.3.6, f is absolutely integrable with $\int_I f = A$, as desired. □

Thus every Riemann integrable function is also Lebesgue integrable, at least on bounded intervals, and we no longer need the $R. \int_I f$ notation. However, the converse is not true. Take for instance the function $f\colon [0, 1] \to \mathbf{R}$ defined by $f(x){:=}1$ when x is rational, and $f(x){:=}0$ when x is irrational. Then from Proposition 11.7.1 we know that f is not Riemann integrable. On the other hand, f is the characteristic function of the set $\mathbf{Q} \cap [0, 1]$, which is countable and hence measure zero. Thus f is Lebesgue integrable and $\int_{[0,1]} f = 0$. Thus the Lebesgue integral can handle more functions than the Riemann integral; this is one of the primary reasons why we use the Lebesgue integral in analysis. (The other reason is that the Lebesgue integral interacts well with limits, as the Lebesgue monotone convergence theorem, Fatou's lemma, and Lebesgue dominated convergence theorem already attest. There are no comparable theorems for the Riemann integral.)

8.5 Fubini's Theorem

In one dimension we have shown that the Lebesgue integral is connected to the Riemann integral. Now we will try to understand the connection in higher dimensions. To simplify the discussion we shall just study two-dimensional integrals, although the arguments we present here can easily be extended to higher dimensions.

We shall study integrals of the form $\int_{\mathbf{R}^2} f$. Note that once we know how to integrate on $\mathbf{R}^2$, we can integrate on measurable subsets Ω of $\mathbf{R}^2$, since $\int_\Omega f$ can be rewritten as $\int_{\mathbf{R}^2} f\chi_\Omega$.

Let $f(x, y)$ be a function of two variables. In principle, we have three different ways to integrate f on $\mathbf{R}^2$. First of all, we can use the two-dimensional Lebesgue integral, to obtain $\int_{\mathbf{R}^2} f$. Secondly, we can fix x and compute a one-dimensional integral in y, and then take that quantity and integrate in x, thus obtaining $\int_{\mathbf{R}}(\int_{\mathbf{R}} f(x, y)\, \mathrm{d}y)\, \mathrm{d}x$. Thirdly, we could fix y and integrate in x, and then integrate in y, thus obtaining $\int_{\mathbf{R}}(\int_{\mathbf{R}} f(x, y)\, \mathrm{d}x)\, \mathrm{d}y$.

Fortunately, if the function f is absolutely integrable on f, then all three integrals are equal:

Theorem 8.5.1 (Fubini's theorem) *Let $f\colon \mathbf{R}^2 \to \mathbf{R}$ be an absolutely integrable function. Then there exists absolutely integrable functions $F\colon \mathbf{R} \to \mathbf{R}$ and $G\colon \mathbf{R} \to \mathbf{R}$ such that for almost every x, $f(x, y)$ is absolutely integrable in y with*

$$F(x) = \int_{\mathbf{R}} f(x, y)\, dy,$$

and for almost every y, $f(x, y)$ is absolutely integrable in x with

$$G(y) = \int_{\mathbf{R}} f(x, y)\, dx.$$

Finally, we have

$$\int_{\mathbf{R}} F(x)\, dx = \int_{\mathbf{R}^2} f = \int_{\mathbf{R}} G(y)\, dy.$$

Remark 8.5.2 Very roughly speaking, Fubini's theorem says that

$$\int_{\mathbf{R}} \left(\int_{\mathbf{R}} f(x, y)\, \mathrm{d}y \right) \mathrm{d}x = \int_{\mathbf{R}^2} f = \int_{\mathbf{R}} \left(\int_{\mathbf{R}} f(x, y)\, \mathrm{d}x \right) \mathrm{d}y.$$

This allows us to compute two-dimensional integrals by splitting them into two one-dimensional integrals. The reason why we do not write Fubini's theorem this way, though, is that it is possible that the integral $\int_{\mathbf{R}} f(x, y)\, \mathrm{d}y$ does not actually exist for every x, and similarly $\int_{\mathbf{R}} f(x, y)\, \mathrm{d}x$ does not exist for every y; Fubini's theorem only asserts that these integrals only exist for *almost every* x and y. For instance, if $f(x, y)$ is the function which equals 1 when $y > 0$ and $x = 0$, equals -1 when $y < 0$ and $x = 0$, and is zero otherwise, then f is absolutely integrable on $\mathbf{R}^2$ and $\int_{\mathbf{R}^2} f = 0$ (since f equals zero almost everywhere in $\mathbf{R}^2$), but $\int_{\mathbf{R}} f(x, y)\, \mathrm{d}y$ is not absolutely integrable when $x = 0$ (though it is absolutely integrable for every other x).

Proof The proof of Fubini's theorem is quite complicated, and we will only give a sketch here. We begin with a series of reductions.

Roughly speaking (ignoring issues relating to sets of measure zero), we have to show that

$$\int_{\mathbf{R}} \left(\int_{\mathbf{R}} f(x, y)\, \mathrm{d}y \right) \mathrm{d}x = \int_{\mathbf{R}^2} f$$

together with a similar equality with x and y reversed. We shall just prove the above equality, as the other one is very similar.

First of all, it suffices to prove the theorem for non-negative functions, since the general case then follows by writing a general function f as a difference $f^+ - f^-$ of two non-negative functions, and applying Fubini's theorem to f^+ and f^- separately (and using Proposition 8.3.3(a) and (b)). Thus we will henceforth assume that f is non-negative.

Next, it suffices to prove the theorem for non-negative functions f supported on a bounded set such as $[-N, N] \times [-N, N]$ for some positive integer N. Indeed, once one obtains Fubini's theorem for such functions, one can then write a general function f as the supremum of such compactly supported functions as

$$f = \sup_{N>0} f\chi_{[-N,N]\times[-N,N]},$$

apply Fubini's theorem to each function $f\chi_{[-N,N]\times[-N,N]}$ separately, and then take suprema using the monotone convergence theorem. Thus we will henceforth assume that f is supported on $[-N, N] \times [-N, N]$.

By another similar argument, it suffices to prove the theorem for non-negative simple functions supported on $[-N, N] \times [-N, N]$, since one can use Lemma 8.1.5 to write f as the supremum of simple functions (which must also be supported on $[-N, N]$), apply Fubini's theorem to each simple function, and then take suprema using the monotone convergence theorem. Thus we may assume that f is a non-negative simple function supported on $[-N, N] \times [-N, N]$.

Next, we see that it suffices to prove the theorem for characteristic functions supported in $[-N, N] \times [-N, N]$. This is because every simple function is a linear combination of characteristic functions, and so we can deduce Fubini's theorem for simple functions from Fubini's theorem for characteristic functions. Thus we may take $f = \chi_E$ for some measurable $E \subseteq [-N, N] \times [-N, N]$. Our task is then to show (ignoring sets of measure zero) that

$$\int_{[-N,N]} \left(\int_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x = m(E).$$

It will suffice to show the upper Lebesgue integral estimate

$$\overline{\int_{[-N,N]}} \left(\overline{\int_{[-N,N]}} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x \le m(E). \tag{8.2}$$

We will prove this estimate later. Once we show this for every set E, we may substitute E with $[-N, N] \times [-N, N]\backslash E$ and obtain

$$\overline{\int_{[-N,N]}} \left(\overline{\int_{[-N,N]}} (1 - \chi_E(x, y))\, \mathrm{d}y \right) \mathrm{d}x \le 4N^2 - m(E).$$

But the left-hand side is equal to

$$\overline{\int}_{[-N,N]} (2N - \underline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y)\, \mathrm{d}x$$

which is in turn equal to

$$4N^2 - \underline{\int}_{[-N,N]} \left(\underline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x$$

and thus we have

$$\underline{\int}_{[-N,N]} \left(\underline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x \geq m(E).$$

In particular we have

$$\underline{\int}_{[-N,N]} \left(\overline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x \geq m(E)$$

and hence by Lemma 8.3.6 we see that $\overline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y$ is absolutely integrable and

$$\int_{[-N,N]} \left(\overline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x = m(E).$$

A similar argument shows that

$$\int_{[-N,N]} \left(\underline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y \right) \mathrm{d}x = m(E)$$

and hence

$$\int_{[-N,N]} \left(\overline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y - \underline{\int}_{[-N,N]} \chi_E(x, y) \right) \mathrm{d}x = 0.$$

Thus by Proposition 8.2.6(a) we have

$$\underline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y = \overline{\int}_{[-N,N]} \chi_E(x, y)\, \mathrm{d}y$$

for almost every $x \in [-N, N]$. Thus $\chi_E(x, y)$ is absolutely integrable in y for almost every x, and $\int_{[-N,N]} \chi_E(x, y)$ is thus equal (almost everywhere) to a function $F(x)$ such that

$$\int\limits_{[-N,N]} F(x)\, \mathrm{d}x = m(E)$$

as desired.

It remains to prove the bound (8.2). Let $\varepsilon > 0$ be arbitrary. Since $m(E)$ is the same as the outer measure $m^*(E)$, we know that there exists an at most countable collection $(B_j)_{j\in J}$ of boxes such that $E \subseteq \bigcup_{j\in J} B_j$ and

$$\sum_{j\in J} m(B_j) \leq m(E) + \varepsilon.$$

Each box B_j can be written as $B_j = I_j \times I'_j$ for some intervals I_j and I'_j. Observe that

$$m(B_j) = |I_j||I'_j| = \int\limits_{I_j} |I'_j|\, \mathrm{d}x = \int\limits_{I_j} \left(\int\limits_{I'_j} \mathrm{d}y \right) \mathrm{d}x$$

$$= \int\limits_{[-N,N]} \left(\int\limits_{[-N,N]} \chi_{I_j \times I'_j}(x, y)\, \mathrm{d}x \right) \mathrm{d}y$$

$$= \int\limits_{[-N,N]} \left(\int\limits_{[-N,N]} \chi_{B_j}(x, y)\, \mathrm{d}x \right) \mathrm{d}y.$$

Adding this over all $j \in J$ (using Corollary 8.2.11) we obtain

$$\sum_{j\in J} m(B_j) = \int\limits_{[-N,N]} \left(\int\limits_{[-N,N]} \sum_{j\in J} \chi_{B_j}(x, y)\, \mathrm{d}x \right) \mathrm{d}y.$$

In particular we have

$$\overline{\int\limits_{[-N,N]}} \left(\overline{\int\limits_{[-N,N]}} \sum_{j\in J} \chi_{B_j}(x, y)\, \mathrm{d}x \right) \mathrm{d}y \leq m(E) + \varepsilon.$$

But $\sum_{j\in J}\chi_{B_j}$ majorizes χ_E (why?) and thus

$$\overline{\int}_{[-N,N]}\left(\overline{\int}_{[-N,N]}\chi_E(x,y)\,\mathrm{d}x\right)\mathrm{d}y \le m(E)+\varepsilon.$$

But ε is arbitrary, and so we have (8.2) as desired. This completes the proof of Fubini's theorem. □

Index

A
Abel's theorem, 72
Absolutely integrable, 179
Absolute value
  for complex numbers, 84
Abstraction, 1
Addition
  of complex numbers, 82
(countably) Additive measure, 147, 159
Adherent point
  of sets, 10
Ambient space, 13
Approximation to the identity, 59, 61, 103
Arctangent, *see* trigonometric functions
Associativity
  of addition in **C**, 82
  of scalar multiplication, 117
  of vector addition, 114

B
Ball, 9
Banach-Tarski paradox, 146, 157
Base of the natural logarithm, *see e*
Basis
  standard basis of row vectors, 114
Boolean algebra, 147, 159
Borel–Cantelli lemma, 177
Borel property, 147, 162
Boundary (point), 10, 34
Bounded
  function, 47
  set, 18

C
C, C^0, C^1, C^2, C^k, 131
Cauchy–Schwarz inequality, 8, 97
Cauchy sequence, 15
Chain rule
  in higher dimensions, 128, 130
Character, 99
Characteristic function, 167
Clairaut's theorem, *see* interchanging derivatives with derivatives
Closed
  box, 150
  set, 11, 34
Closure, 10, 34
Cocountable topology, 36
Coefficient, 65
Cofinite topology, 36
Column vector, 114
Commutativity, 102
  of addition in **C**, 82
  of addition in vector spaces, 114
  of convolution, 60, 102
Compactness, 18, 35
  compact support, 58
Completeness
  of metric spaces, 16
  of the space of continuous functions, 50
Completion of a metric space, 17
Complex numbers **C**, 82
  complex conjugation, 84
Connectedness, 30
  connected component, 33
Continuity, 23, 35
  and compactness, 28
  and connectedness, 32
  and convergence, 23
Contraction, 133
  mapping theorem, 134
Convergence

T. Tao, *Analysis II*, Texts and Readings in Mathematics,
https://doi.org/10.1007/978-981-19-7284-3

in L^2, 517, 98
of a function at a point, 40
of sequences, 4, 34
pointwise, *see* pointwise convergence
uniform, *see* uniform convergence
Convolution, 59, 76, 102
Coset, 156
Cosine, *see* trigonometric functions
Cotangent, *see* trigonometric functions
Cover, 149
see also open cover, 151
Critical point, 144

D
Degree, 58
De Moivre identities, 91
Dense, 58
Derivative
directional, 122
in higher dimensions, 120, 123, 127, 130
matrix, 127
partial, 123
total, 122, 130
uniqueness of, 121
Differentiability
continuous, 131
directional, 122
infinite, 68
in higher dimensions, 120
k-fold, 68, 131
Differential matrix, *see* derivative matrix
Dilation, 115
Diophantine, 178
Dirac delta function, 59
Discrete metric, 4
Distance
in **C**, 85
in **R**, 2
Distributive law
for complex numbers, 83

E
e, 79
Egoroff's theorem, 179
Empty
set, 147, 149
Error-correcting codes, 3
Euclidean metric, 2
Euclidean space, 2
Euler's formula, 88, 90
Euler's number, *see e*

Exponential function, 78, 86
Exterior (point), 10, 34

F
Fatou's lemma, 176
Fejér kernel, 103
Finite intersection property, 22
Fixed point theorem, 134
Fourier
coefficients, 100
inversion formula, 101
series, 101
series for arbitrary periods, 110
theorem, 106
transform, 100
Fractional part, 95
Frequency, 99
Fubini's theorem, 185

G
Geodesic, 4
Geometric series
formula, 54
Gradient, 127
Graph, 141

H
Hairy ball theorem, 134
Half-space, 158
Hausdorff space, 36
Heine-Borel theorem, 18
Hermitian form, 97
Homogeneity, 98, 115
Hypersurface, 141

I
Identity map (or operator), 115
Imaginary, 84
Implicit differentiation, 141
Implicit function theorem, 141
Inconsistent, 87
Indicator function, *see* characteristic function
Induced
metric, 2, 13
topology, 13, 35
Inner product, 96
Integer part, 95
Integration
by parts, 73

Interchanging
derivatives with derivatives, 132
integrals with integrals, 176, 185
limits with derivatives, 56
limits with integrals, 56, 173, 180
limits with limits, 46
limits with sums, 179
sums with derivatives, 57, 66
sums with integrals, 54, 66, 176, 178
Interior (point), 10, 34
Intermediate value theorem, 32
Intrinsic, 18
Inverse
function theorem, 137
Invertible function, *see* bijection
local, 136
Involution, 84
Isolated point, 41
Isometry, 13

L
$l^1, l^2, l^\infty, L^1, L^2, L^\infty$, 2–4, 97, 179
equivalence of in finite dimensions, 6
Laws of algebra
for complex numbers, 82, 83
Laws of arithmetic, *see* laws of algebra
Laws of exponentiation, 78
Lebesgue dominated convergence theorem, 180
Lebesgue integral
of absolutely integrable functions, 179
of simple functions, 168
of nonnegative functions, 172
upper and lower, 181
vs. the Riemann integral, 184
Lebesgue measurable, 158
Lebesgue measure, 148
motivation of, 146–148
Lebesgue monotone convergence theorem, 173
Leibniz rule, 130
Limit
formal (LIM), 17
laws, 85
limiting values of functions, 40
pointwise, 42
uniform, *see* uniform limit
uniqueness of, 6, 41
Limit point
of sequences, 15
Linear combination, 114
Linearity
approximate, 119
of convolution, 62, 102
of inner product, 97
of integration, 170, 175
of transformations, 115
Logarithm (natural), 79
power series of, 54, 80

M
Majorize, 172
Manifold, 144
Map, *see* function
Matrix, 115
identification with linear transformations, 116–119
Maximum
principle, 29
Measurability
for functions, 163, 164
for sets, 158
motivation of, 146
see also Lebesgue measure, outer measure, 148
Metric, 2
ball, *see* ball
on **C**, 85
on **R**, 2
space, 2
Monomial, 99
Monotone (increasing or decreasing)
convergence, *see* Lebesgue monotone convergence theorem
measure, 147, 150
Moving bump example, 44, 176
Multiplication
of complex numbers, 83
of matrices, 115, 118

N
Negation
of complex numbers, 83
Neighbourhood, 34
Newton's approximation, 121
Non-degenerate, 97
Nowhere differentiable function, 57, 91

O
Of sets
adherent point, 34
Open
box, 148

cover, 19
set, 11
Order topology, 36
Orthogonality, 98
Orthonormal, 100
non-additivity of, 157
Outer measure, 149
non-Additivity of, 156

P
Parseval identity, 101, 110
Path-connected, 32
Periodic, 94
extension, 95
π, 90, 91
Plancherel formula (or theorem), 101, 107
Pointwise convergence, 42
of series, 51
topology of, 50
Polar representation, 91
Polynomial, 58
and convolution, 60
approximation by, 58, 61
Positive
complex number, 83, 87
inner product, 97
measure, 147, 149
Power series, 65
formal, 65
multiplication of, 75
uniqueness of, 70
Pre-image, *see* inverse image
Product topology, 50
Projection, 115
Pythagoras' theorem, 98

Q
Quotient rule, 131

R
Radius of convergence, 65
Real analytic, 68
Real numbers **R**
are uncountable, *see* uncountability of the reals
Real part, 84
Real-valued, 51
Reciprocal
of complex numbers, 85
Relative topology, *see* induced topology
Ring
commutative, 83
Root
mean square, *see* L^2
Row vector, 113

S
Scalar multiplication, 113
Series
of functions, 54
σ-algebra, 147, 161
Simple function, 167
Sine, *see* trigonometric functions
Space, 2
Square wave, 94, 99
Stone-Weierstrass theorem, 63, 102
Sub-additive measure, 147, 150
Subsequence, 15
Summation by parts, 73
Sup norm, *see* supremum as norm
Support, 58
Supremum (and infimum)
as metric, 4
as norm, 4, 52

T
Taxi-cab metric, 3
Taylor series, 69
Taylor's formula, *see* Taylor series
Topological space, 33
Totally bounded, 21
Translation invariance, 147, 150, 159
Transpose, 114
Triangle inequality
for integrals, 180
in **C**, 85
in Euclidean space, 8
in inner product spaces, 97
in metric spaces, 2
Trigonometric functions, 88, 92
and Fourier series, 109
power series, 88, 91
trigonometric polynomials, 99
Trivial topology, 35

U
Uniform continuity, 29
Uniform convergence, 44
and anti-derivatives, 55
and derivatives, 47
and integrals, 54
and limits, 46

and radius of convergence, 65
as a metric, 49, 95
of series, 51
Uniform limit
and Riemann integration, 53
of bounded functions, 47
of continuous functions, 46

V
Vector space, 114
Vertical line test, 59
Volume, 148

W
Weierstrass approximation theorem, 58, 61–62, 101
Weierstrass example, *see* nowhere differentiable function
Weierstrass M-test, 52