TRIM 37

trim

TEXTS AND READINGS IN MATHEMATICS 37

Tao

Analysis I

Analysis I

Terence Tao

HINDUSTAN BOOK AGENCY

यथा शिखा मयूराणां नागानां मणयो यथा।
तथा वेदाङ्गशास्त्राणां गणितं मूर्धनि स्थितम्।।

As are the crests on the heads of peacocks,
as are the gems on the hoods of cobras,
so is mathematics, at the top of all sciences.

The Yajurveda, circa 600 B.C.

ISBN 81-85931-62-3

www.hindbook.com

TEXTS AND READINGS
IN MATHEMATICS

Analysis I

Analysis I

Terence Tao
University of California
Los Angeles

HINDUSTAN
BOOK AGENCY

Published by

Hindustan Book Agency (India)
P 19 Green Park Extension
New Delhi 110 016
India

email: hba@vsnl.com
http://www.hindbook.com

Produced from camera ready copy supplied by the Author.

ISBN 81-85931-62-3

To my parents, for everything

Contents

Volume 1

Preface

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 conditionally 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 persevere and obtain a rigorous proof of an intuitive fact, it solidifed 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 perspec-

tive; 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 exami-

nation 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 16 (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.

Terence Tao

Chapter 1

Introduction

1.1 What is analysis?

This text is an honours-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 rigourous 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 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

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

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

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

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

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

one gets nonsensical results:

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

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

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

we can write

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

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

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

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

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

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 $\sin(\pi/2 - 2) = \cos(z)$ 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:

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

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 book-keepers 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 i^{th} row and j^{th} 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 $\int f(x,y)\ dx$, and then we integrate the area in the y variable to obtain the volume

$$V = \int \int f(x,y) dx dy.$$

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

$$V = \int \int f(x,y) dy dx.$$

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

$$\int \int f(x,y)\ dx dy = \int \int f(x,y)\ dy dx.$$

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 $e^{-xy} - xye^{-xy}$. Suppose we believe that we can swap the integrals:

$$\int_0^\infty \int_0^1 (e^{-xy} - xye^{-xy})\, dy\, dx = \int_0^1 \int_0^\infty (e^{-xy} - xye^{-xy})\, dx\, dy.$$

Since

$$\int_0^1 (e^{-xy} - xye^{-xy})\, dy = ye^{-xy}|_{y=0}^{y=1} = e^{-x},$$

the left-hand side is $\int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$. But since

$$\int_0^\infty (e^{-xy} - xye^{-xy})\, dx = xe^{-xy}|_{x=0}^{x=\infty} = 0,$$

the right-hand side is $\int_0^1 0\, dx = 0$. 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 19.5.1 for a partial answer.)

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

$$\lim_{x\to 0} \lim_{y\to 0} \frac{x^2}{x^2+y^2} = \lim_{y\to 0} \lim_{x\to 0} \frac{x^2}{x^2+y^2}. \tag{1.1}$$

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 (1.1) 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 (1.1) 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 13.2.9 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 14.3.3 for an answer.)

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

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

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

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

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 14.6.1 for a partial answer.)

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

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

and in particular that

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

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

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

But the right-hand side is $\frac{d}{dx}x = 1$. Does this mean that it is always illegitimate to interchange limits and derivatives? (See Theorem 14.7.1 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 maneuvre 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

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

and hence

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

[1] One might object that this function is not defined at $(x,y) = (0,0)$, but if we set $f(0,0) := (0,0)$ 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)!

Thus

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

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 17.5.4 and Exercise 17.5.1 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 rigourous; see Section 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, 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 rigourously 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* $\mathbf{N}$; the *integers* $\mathbf{Z}$;

the *rationals* $\mathbf{Q}$, and the *real numbers* $\mathbf{R}$. (There are other number systems such as the *complex numbers* $\mathbf{C}$, but we will not study them until Section 15.6.) 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.)

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 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 1,2,3, etc. 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 an 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 Guiseppe 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 Section 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 $\mathbf{N}$ 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] $\mathbf{N}$. However, it is not really that satisfactory, because it begs the question of what $\mathbf{N}$ is. This definition of "start at 0 and count indefinitely" seems like an intuitive enough definition of $\mathbf{N}$, 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

[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.

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. In deference to modern computer languages, we will use $n{+}{+}$ to denote the 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 **N** consists of 0 and everything which can be obtained from 0 by incrementing: **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 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 **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 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, 65536 is equal to 0 (using our definition of 65536 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 2.1,2.2,2.3. 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*[2]*, if $n++ = m++$, then we must have $n = m$.*

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 2.1 and 2.2) 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:

[2]This is an example of reformulating an implication using its *contrapositive*; see Section A.2 for more details.

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

Proof. 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 $\{0, 1, 2, 3, \ldots\}$ and the Roman number system $\{O, I, II, III, IV, V, VI, \ldots\}$, 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 3.7), 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., $\mathbf{N}$ 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, etc. 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[3]:

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. □

[3]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 rigourously in the language of set theory; see Exercise 3.5.12.

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.

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 - increment - 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 integer 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[4].

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 do not know yet that $a+b=b+a$.

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 $0+(m++)=(0+$

[4]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.

$m)++$. 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 these notes, 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 Section A.1).

Proposition 2.2.8. *If a is positive 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 anti-symmetric*) *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. □

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.)

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 $n < m_0$.)

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.)

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 (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 (Euclidean algorithm). *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 $m^0 := 1$. 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.

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 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 rigourous 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 Chapter 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* 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 3.1.10 for a more formal version of this example. However, not all objects are sets; for instance, we typically do not consider a natural number such as 3 to be a set. (The more accurate statement is that natural numbers can be the *cardinalities* of sets, rather than necessarily being sets themselves. See Section 3.6.)

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.

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 define 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 as a definition:

Definition 3.1.4 (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 3.1.5. 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.)

One can easily verify that this notion of equality is reflexive, symmetric, and transitive (Exercise 3.1.1). Observe that if $x \in A$ and $A = B$, then $x \in B$, by Definition 3.1.4. Thus 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 3.2 (Empty set). *There exists a set $\emptyset$, known as the empty set, which contains no elements, i.e., for every object x we have $x \notin \emptyset$.*

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 Definition 3.1.4 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 3.1.6 (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 3.2 we have $x \notin \emptyset$. Thus $x \in A \iff x \in \emptyset$ (both statements are equally false), and so $A = \emptyset$ by Definition 3.1.4, a contradiction. □

Remark 3.1.7. 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 3.5.12) 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 Section 8.4.

Remark 3.1.8. Note that the empty set is *not* 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 Section 3.6.

If Axiom 3.2 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 3.3 (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 3.1.9. Just as there is only one empty set, there is only one singleton set for each object a, thanks to Definition 3.1.4 (why?). Similarly, given any two objects a and b, there is only one pair set formed by a and b. Also, Definition 3.1.4 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 3.1.13). 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 3.1.10. Since $\emptyset$ is a set (and hence an object), so is singleton set $\{\emptyset\}$, i.e., the set whose only element is $\emptyset$, is a set (and it is *not* the same set as $\emptyset$, $\{\emptyset\} \neq \emptyset$ (why?). Similarly, the singleton set $\{\{\emptyset\}\}$ and the pair set $\{\emptyset, \{\emptyset\}\}$ are also sets. These three 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 3.4 (Pairwise union). *Given any two sets A, B, there exists a set $A \cup B$, called the union $A \cup B$ of A and B, whose elements 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 Section A.1.

Example 3.1.11. 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 3.1.12. 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 3.4 and Definition 3.1.4). 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 3.1.13. *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 Definition 3.1.4, 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 3.4, 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 3.4 again $x \in B \cup C$, and so by Axiom 3.4 again we have $x \in A \cup (B \cup C)$. Now suppose instead $x \in A \cup B$, then by Axiom 3.4 again $x \in A$ or $x \in B$. If $x \in A$ then $x \in A \cup (B \cup C)$ by Axiom 3.4, while if $x \in B$ then by consecutive applications of Axiom 3.4 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 3.1.14. 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 rigourously 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 rigourously. 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 3.1.15 (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 3.1.16. 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 3.1.17. 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 3.1.18 (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 3.1.19. There is a relationship between subsets and unions: see for instance Exercise 3.1.7.

Remark 3.1.20. 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 3.1.21. 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 $2 \notin \{1,2,3\}$. 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 3.5 (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., $P(x)$ is either a true statement or a false statement). Then there exists a set,*

called $\{x \in A : P(x)$ *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 $\{x \in A : P(x)$ 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 3.1.22. 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 $\{x \in A \mid P(x)\}$ 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 range and domain of a function $f : X \to Y$).

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

Definition 3.1.23 (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 3.1.24. Note that this definition is well-defined (i.e., it obeys the axiom of substitution, see Section 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 3.1.25. 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 3.1.26. 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?).

Definition 3.1.27 (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 3.1.28 (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 3.1.29. 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 3.1.30. 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; it plays a particularly important rôle in 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 3.6 (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 $\{y : P(x, y)$ is true for some $x \in A\}$, such that for any object z,*

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

Example 3.1.31. 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 $\{y : y = x{+}{+}$ for some $x \in \{3, 5, 9\}\}$ exists; in this case, it is clearly the same set as $\{4, 6, 10\}$ (why?).

Example 3.1.32. 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 $\{y : y = 1$ 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 $\{f(x) | x \in A\}$. 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 $\{f(x) : x \in A; P(x) \text{ is true}\}$ 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 3.7 (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 Section 3.6.) From the axiom of infinity we see that numbers such as 3, 5, 7, etc. 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 $\{n + 3 : n \in \mathbf{N}, 0 \leq n \leq 5\}$ is not the same thing as the expression or function $n + 3$. We emphasize this with an example:

Example 3.1.33. (Informal) This example requires the notion of subtraction, which has not yet been formally introduced. The following two sets are equal,

$$\{n + 3 : n \in \mathbf{N}, 0 \leq n \leq 5\} = \{8 - n : n \in \mathbf{N}, 0 \leq n \leq 5\}, \quad (3.1)$$

(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 counter-intuitive 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 $\{8-n : n \in \mathbf{N}, 0 \leq n \leq 5\}$ by replacing the letter n by the letter m, thus giving $\{8-m : m \in \mathbf{N}, 0 \leq m \leq 5\}$. This is exactly the same set as before (why?), so we can rewrite (3.1) as

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

Now it is easy to see (using (3.1.4)) 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 n 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!

Exercise 3.1.1. Show that the definition of equality in (3.1.4) is reflexive, symmetric, and transitive.

Exercise 3.1.2. Using only Definition 3.1.4, Axiom 3.2, and Axiom 3.3, 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 3.1.13.

Exercise 3.1.4. Prove the remaining claims in Proposition 3.1.18.

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 3.1.28. (Hint: one can use some of these claims to prove others. Some of the claims have also appeared previously in Lemma 3.1.13.)

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.

3.2 Russell's paradox (Optional)

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

Axiom 3.8 (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 $\{x : P(x)$ 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 definition 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 $P(\Omega)$ would be true, and hence $\Omega \in \Omega$. 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.

[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.

Instead, we shall simply postulate an axiom which ensures that absurdities such as Russell's paradox do not occur.

Axiom 3.9 (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.

Exercise 3.2.1. Show that the universal specification axiom, Axiom 3.8, if assumed to be true, would imply Axioms 3.2, 3.3, 3.4, 3.5, and 3.6. (If we assume that all natural numbers are objects, we also obtain Axiom 3.7.) 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 3.8 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).

Exercise 3.2.3. Show (assuming the other axioms of set theory) that the universal specification axiom, Axiom 3.8, 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 3.8 is true, then a universal set exists, and conversely, if a universal set exists, then Axiom 3.8 is true. (This may explain why Axiom 3.8 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 range* Y to be the object which, given any input $x \in X$, assigns an output $f(x) \in Y$, defined to be the unique object $f(x)$ 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.

Example 3.3.2. 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 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 $h(2n+3) = h(2n+2)$, but $h(0)$ is undefined since 0 is not in the domain $N\backslash\{0\}$.

Example 3.3.3. (Informal) This example requires the real numbers $\mathbf{R}$, which we will define in Chapter 5. One could try to define a square root function $\surd : \mathbf{R} \to \mathbf{R}$ 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 range $\mathbf{R}$ 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 range of $\mathbf{R}$ to $[0, +\infty)$. Once one does this, then one can correctly define a square root function $\surd : [0, +\infty) \to [0, +\infty)$ 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 range, 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 3.3.2 could be defined explicitly by saying that f has domain and range 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 3.3.3 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 range of a function for brevity, and thus for instance we could refer to the function f in Example 3.3.2 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 range 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 3.3.4. 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 3.3.5. 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 $f(x)$ of a function 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 3.3.6. Strictly speaking, functions are not sets, and sets are not functions; it does not make sense to ask whether an object x is an element of a function f, and it does not make sense to apply a set A to an input x to create an output $A(x)$. On the other hand, it is possible to start with a function $f : X \to Y$ and construct its *graph* $\{(x, f(x)) : x \in X\}$, which describes the function completely: see Section 3.5.

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

Definition 3.3.7 (Equality of functions). Two functions $f : X \to Y$, $g : X \to Y$ with the same domain and range are said to be *equal*, $f = g$, if and only if $f(x) = g(x)$ for *all* $x \in X$. (If $f(x)$ and $g(x)$ agree for some values of x, but not others, then we do not consider f and g to be equal[2].)

Example 3.3.8. 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 3.3.9. A rather boring example of a function is the *empty function* $f : \emptyset \to X$ from the empty set to an arbitrary set X. Since the empty set has no elements, we do not need

[2]In Chapter 19, we shall introduce a weaker notion of equality, that of two functions being *equal almost everywhere*.

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 3.3.7 asserts that all functions from $\emptyset$ to X are equal (why?).

This notion of equality obeys the usual axioms (Exercise 3.3.1). A fundamental operation available for functions is *composition*.

Definition 3.3.10 (Composition). Let $f : X \to Y$ and $g : Y \to Z$ be two functions, such that the range of f is the same set as the domain of g. We then define the *composition* $g \circ f : X \to Z$ 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 range 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 3.3.11. Let $f : \mathbf{N} \to \mathbf{N}$ be the function $f(n) := 2n$, and let $g : \mathbf{N} \to \mathbf{N}$ 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 3.3.12 (Composition is associative). *Let $f : X \to Y$, $g : Y \to Z$, and $h : Z \to W$ be functions. Then $f \circ (g \circ h) = (f \circ g) \circ h$.*

Proof. Since $g \circ h$ is a function from Y to W, $f \circ (g \circ h)$ is a function from X to W. Similarly $f \circ g$ is a function from X to Z, 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 range. In order to check that they are equal, we see from Definition 3.3.7 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 3.3.10

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

as desired. □

Remark 3.3.13. 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 3.3.14 (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 3.3.15. (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 3.3.16. 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 3.3.17 (Onto functions). A function f is *onto* (or *surjective*) if $f(X) = Y$, i.e., every element in Y comes from applying f to some element in X:

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

Example 3.3.18. (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 range $\mathbf{Z}$ 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 3.3.19. 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 3.3.20 (Bijective functions). Functions $f : X \to Y$ which are both one-to-one and onto are also called *bijective* or *invertible*.

Example 3.3.21. 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 3.3.22. The function $f : \mathbf{N} \to \mathbf{N}\backslash\{0\}$ defined by $f(n) := n{+}{+}$ is a bijection (in fact, this fact is simply restating Axioms 2.2, 2.3, 2.4). On the other hand, the function $g : \mathbf{N} \to \mathbf{N}$ 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 range (and domain) are.

Remark 3.3.23. 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 3.3.24. A common error is to say that a function $f : X \to Y$ 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 the previous example 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.

Exercise 3.3.1. Show that the definition of equality in Definition 3.3.7 is reflexive, symmetric, and transitive. Also verify the substitution property: if $f, \tilde{f} : X \to Y$ and $g, \tilde{g} : Y \to Z$ are functions such that $f = \tilde{f}$ and $g = \tilde{g}$, then $f \circ g = \tilde{f} \circ \tilde{g}$.

Exercise 3.3.2. Let $f : X \to Y$ and $g : Y \to Z$ 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 injective? surjective? bijective?

Exercise 3.3.4. In this section we give some cancellation laws for composition. Let $f : X \to Y$, $\tilde{f} : X \to Y$, $g : Y \to Z$, and $\tilde{g} : Y \to Z$ 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 $f : X \to Y$ and $g : Y \to Z$ 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 $f : X \to Y$ 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 $f : X \to Y$ and $g : Y \to Z$ 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 $f : A \to B$ is any function, then $f = f \circ \iota_{A \to A} = \iota_{B \to B} \circ f$.

(c) Show that, if $f : A \to B$ 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 $f : X \to Z$ and $g : Y \to Z$ are functions, then there is a unique function $h : X \cup Y \to Z$ such that $h \circ \iota_{X \to X \cup Y} = f$ and $h \circ \iota_{Y \to X \cup Y} = g$.

3.4 Images and inverse images

We know that a function $f : X \to Y$ 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 $f : X \to Y$ is a function from X to Y, and S is a set in X, we define $f(S)$ 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 3.6). One can also define $f(S)$ using the axiom of specification (Axiom 3.5) instead of replacement, but we leave this as a challenge to the reader.

Example 3.4.2. If $f : \mathbf{N} \to \mathbf{N}$ 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.

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 3.3.14):

Example 3.4.3. (Informal) Let $\mathbf{Z}$ be the set of integers (which we will define rigourously 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?).

Definition 3.4.4 (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 3.4.5. 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 3.4.6. (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 3.4.7. If f is a bijective function, then we have defined f^{-1} in two slightly different ways, but this is not an issue because both definitions are equivalent (Exercise 3.4.1).

As remarked earlier, functions are not 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 3.10 (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*

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

Example 3.4.8. 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 3.4.9. *Let X be a set. Then the set*

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

is a set.

Proof. See Exercise 3.4.6. □

Remark 3.4.10. The set $\{Y : Y \text{ is a subset of X}\}$ 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 Chapter 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 3.11 (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 3.4.11. 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 $\alpha \in A$. 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 3.4.12. The axioms of set theory that we have introduced (Axioms 3.1-3.11, excluding the dangerous Axiom 3.8) are known as the *Zermelo-Fraenkel axioms of set theory*[3], 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 Section 8.4), giving rise to the *Zermelo-Fraenkel-Choice (ZFC) axioms of set theory*, but we will not need this axiom for some time.

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

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

Exercise 3.4.3. 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. Prove Lemma 3.4.9. (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.

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 range 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 3.4 can be deduced from Axiom 3.3 and Axiom 3.11.

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.

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 3.1.28 (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.*

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 obeys 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

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

Remark 3.5.5. We shall simply assume that our notion of ordered pair is such that whenever X and Y are sets, the Cartesian product $X \times Y$ is also a set. This is however not a problem in practice; see Exercise 3.5.1.

Example 3.5.6. 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 $f(x,y)$ 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 $(x,y) \mapsto x + y$.

One can of course generalize the concept of ordered pairs to ordered triples, ordered quadruples, etc:

Definition 3.5.7 (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

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

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 (Exercise 3.5.2).

Remark 3.5.8. One can show that $\prod_{1\leq i\leq n} X_i$ is indeed a set. Indeed, from the power set axiom we can consider the set of all functions $i \mapsto x_i$ from the domain $\{1 \leq i \leq n\}$ to the range $\bigcup_{1\leq i\leq n} X_i$, and then we can restrict using the axiom of specification to restrict to those functions $i \mapsto x_i$ for which $x_i \in X_i$ for all $1 \leq i \leq n$. One can generalize this construction to infinite Cartesian products, see Definition 8.4.1.

Example 3.5.9. 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 $x_1 \in X_1$, $(x_2, x_3) \in X_3$, and so forth; we will not bother to distinguish between these different perspectives.

Remark 3.5.10. 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 Chapter 5 we shall also introduce the very useful concept of an *infinite sequence*.

Example 3.5.11. 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 $\prod_{1 \leq i \leq 1} X_i$ is just X_1 (why?). Also, the *empty Cartesian product* $\prod_{1 \leq i \leq 0} X_i$ 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 $X^n := \prod_{1 \leq i \leq n} X$. 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 the Cartesian product $X \times X$. The set X^0 is a singleton set $\{()\}$ (why?).

We can now generalize the single choice lemma (Lemma 3.1.6) to allow for multiple (but finite) number of choices.

Lemma 3.5.12 (Finite choice). *Let $n \geq 1$ be a natural number, and for each natural number $1 \leq i \leq n$, let X_i be a non-empty set. Then there exists an n-tuple $(x_i)_{1 \leq i \leq n}$ 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 3.1.6 (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 3.1.6 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 3.5.13. 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.

Exercise 3.5.1. 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 3.3). 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 indeed obeys the property (3.5), and also whenever X and Y are sets, the Cartesian product $X \times Y$ is also a set. Thus this definition can be validly used as a definition of an ordered pair. For an additional challenge, show that the alternate definition $(x, y) := \{x, \{x, y\}\}$ also verifies (3.5) and is thus also an acceptable definition of ordered pair. (For this latter task one needs the axiom of regularity, and in particular Exercise 3.2.2.)

Exercise 3.5.2. Suppose we *define* an ordered n-tuple to be a surjective function $x : \{i \in \mathbf{N} : 1 \leq i \leq n\} \to X$ whose range 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 3.5.7, 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 obey the reflexivity, symmetry, and transitivity axioms.

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.)

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 the hypotheses that the A, B, C, D are all non-empty are removed?

Exercise 3.5.7. Let X, Y be sets, and let $\pi_{X \times Y \to X} : X \times Y \to X$ and $\pi_{X \times Y \to Y} : X \times Y \to Y$ 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 $f : Z \to X$ and $g : Z \to Y$, there exists a unique function $h : Z \to X \times Y$ 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 *direct sum* of f and g and is denoted $h = f \oplus g$.

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 $(\bigcup_{\alpha \in I} A_\alpha) \cap (\bigcup_{\beta \in J} B_\beta) = \bigcup_{(\alpha, \beta) \in I \times J} (A_\alpha \cap B_\beta)$.

Exercise 3.5.10. If $f : X \to Y$ is a function, define the *graph* of f to be the subset of $X \times Y$ defined by $\{(x, f(x)) : x \in X\}$. Show that two functions $f : X \to Y$, $\tilde{f} : X \to Y$ are equal if and only if they have the same graph. 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 $f : X \to Y$ whose graph is equal to G.

Exercise 3.5.11. Show that Axiom 3.10 can in fact be deduced from Lemma 3.4.9 and the other axioms of set theory, and thus Lemma 3.4.9 can be used as an alternate formulation of the power set axiom. (Hint: for any two sets X and Y, use Lemma 3.4.9 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 rigourous version of Proposition 2.1.16. Let $f : \mathbf{N} \times \mathbf{N} \to \mathbf{N}$ be a function, and let c be a natural number. Show that there exists a function $a : \mathbf{N} \to \mathbf{N}$ such that

$$a(0) = c$$

and

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

and furthermore that this function is unique. (Hint: first show inductively, by a modification of the proof of Lemma 3.5.12, that for every natural number $N \in \mathbf{N}$, there exists a unique function $a_N : \{n \in \mathbf{N} : n \leq N\} \to \mathbf{N}$ such that $a_N(0) = c$ and $a_N(n{+}{+}) = f(n, a(n))$ for all $n \in \mathbf{N}$ such that $n < N$.) For an additional challenge, prove this result 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: 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 $\{8,9\}$. One way to define this is 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 rigourous 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 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.

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 $n-1$.*

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 Proposition 3.6.4, 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. □

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 $\#(\mathbf{N}) = n$. Then there is a bijection f from $\{i \in \mathbf{N} : 1 \leq i \leq n\}$ to $\mathbf{N}$. One can show that the sequence $f(1), f(2), \ldots, f(n)$ is bounded, or more precisely that there exists a natural number M such that $f(i) \leq M$ for all $1 \leq i \leq n$ (Exercise 3.6.3). But then the natural number $M+1$ is not equal to any of the $f(i)$, contradicting the hypothesis that f is a bijection. □

Remark 3.6.13. One can also use similar arguments to show that any unbounded set 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 Section 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 $\#(f(X)) \leq \#(X)$. If in addition f is one-to-one, then $\#(f(X)) = \#(X)$.*

(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 3.10) is finite and $\#(Y^X) = \#(Y)^{\#(X)}$.*

Proof. See Exercise 3.6.4. □

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 Chapter 8, once we have constructed a few more examples of infinite sets (such as the integers, rationals and reals).

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 $f : \{i \in \mathbf{N} : 1 \leq i \leq n\} \to \mathbf{N}$ 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.

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 $f : A \to B$ 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 $f : A \to B$ from A to B (i.e., A has lesser or equal cardinality to B). Show that there then exists a surjection $g : B \to A$ 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 $\#(\bigcup_{i\in}$ n. Show that there exists $i \in \{1, \ldots, n\}$ such that $\#(A_i) \geq$ known as the *pigeonhole principle.*)

Chapter 4

Integers and rationals

4.1 The integers

In Chapter 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 advan-

tage of our foreknowledge by building all this into the *definition* 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 $a—b$ to define integers, where the — is a meaningless place-holder, 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 $a—b$ 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 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.

[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): $a—b := \{(c, d) \in \mathbf{N} \times \mathbf{N} : (a, b) \sim (c, d)\}$. However, this interpretation plays no rôle 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.

We have to check that this is a legitimate notion of equality. We need to verify the reflexivity, symmetry, transitivity, and substitution axioms (see Section 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 re-verify 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

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

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 *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 operation of subtraction instead. (As remarked before, we could not use subtraction immediately because it would be circular.)

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. □

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.)

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

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

their product

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

and the negation

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

[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. However, we can eventually get a reasonable notion of dividing by a quantity which *approaches* zero - think of L'Hôpital's rule (see Section 10.5), which suffices for doing things like defining differentiation.

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 a rational number 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:

$$\begin{aligned}(a//1) + (b//1) &= (a+b)//1;\\ (a//1) \times (b//1) &= (ab//1);\\ -(a//1) &= (-a)//1.\end{aligned}$$

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 $\mathbf{Q}$ 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 $a//b$.

Proposition 4.2.4allows 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 anti-symmetric) 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.

Exercise 4.2.1. Show that the definition of equality for the rational numbers is reflexive, symmetric, and transitive. (Hint: for transitivity, use Corollary 2.3.7.)

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 real 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 real 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 $\varepsilon > 0$, and 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 $x^0 := 1$. 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) *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 $x^{-3} = 1/x^3 = 1/(x \times x \times x)$. 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. □

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-rigourous arrangement, since we have not yet defined the concept of a line, but this discussion is only intended to motivate the more rigourous 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 $\mathbf{Q}$ 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

[3] We will use the decimal system for defining terminating decimals, for instance 1.414 is defined to equal the rational number 1414/1000. We defer the formal discussion on the decimal system to an Appendix (§B).

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 the Appendix §B.)

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.)

(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.

Chapter 5

The real numbers

To review our progress to date, we have rigourously constructed three fundamental number systems: the natural number system $\mathbf{N}$, the integers $\mathbf{Z}$, and the rationals $\mathbf{Q}$[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, $a—b$. 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

[1]The symbols $\mathbf{N}$, $\mathbf{Q}$, and $\mathbf{R}$ stand for "natural", "quotient", and "real" respectively. $\mathbf{Z}$ stands for "Zahlen", the German word for number. There is also the *complex numbers* $\mathbf{C}$, which obviously stands for "complex".

[2]*Formal* means "having the form of"; at the beginning of our construction the expression $a—b$ 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.

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. Things get even worse when one starts to deal with the sub-field 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 rigourous 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 rigourously turns out to be quite difficult. (Even such great mathematicians as Euler and Newton had diffi-

culty with this concept. It was only in the nineteenth century that mathematicians such as Cauchy and Dedekind figured out how to deal with limits rigourously.)

In Section 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 12.4.8.

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,... 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 $9, 16, 25, \ldots$ 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 se-

quences 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 $d(a_j, a_k) \leq \varepsilon$ 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 $d(a_j, a_k) \leq \varepsilon$ 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 $a_N, a_{N+1}, a_{N+2}, \ldots$ is ε-steady for some natural number $N \geq 0$. In other words, the sequence $a_0, a_1, a_2, \ldots$ is eventually ε-steady iff there exists an $N \geq 0$ such that $d(a_j, a_k) \leq \varepsilon$ for all $j, k \geq N$.

Example 5.1.7. The sequence $a_1, a_2, \ldots$ defined by $a_n := 1/n$, (i.e., the sequence $1, 1/2, 1/3, 1/4, \ldots$) is not 0.1-steady, but is eventually 0.1-steady, because the sequence $a_{10}, a_{11}, a_{12}, \ldots$ (i.e., $1/10, 1/11, 1/12, \ldots$) is 0.1-steady. The sequence $10, 0, 0, 0, 0, \ldots$ 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 $a_0, a_1, a_2, \ldots$ 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 (Proposition 6.1.4). 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

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

mentioned earlier. This sequence is already 1-steady. If one discards the first element 1.4, then the remaining sequence

$$1.41, 1.414, 1.4142, \ldots$$

is now 0.1-steady, which means that the original sequence was eventually 0.1-steady. Discarding the next element gives the 0.01-steady sequence $1.414, 1.4142, \ldots$; thus the original sequence was eventually 0.01-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 rigourous for several reasons, for instance we have not precisely defined what this sequence $1.4, 1.41, 1.414, \ldots$ really is. An example of a rigourous 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. □

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.)

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 rigourous 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 $|a_n - b_n| \leq \varepsilon$ for all $n = 0, 1, 2, \ldots$.

Example 5.2.2. The two sequences

$$1, -1, 1, -1, 1, \ldots$$

and

$$1.1, -1.1, 1.1, -1.1, 1.1, \ldots$$

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, $a_0, a_1, a_2, \ldots$ is eventually ε-close to $b_0, b_1, b_2, \ldots$ 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 notations for ε-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

$$1.1, 1.01, 1.001, 1.0001, \ldots$$

and

$$0.9, 0.99, 0.999, 0.9999, \ldots$$

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, $a_0, a_1, a_2, \ldots$ and $b_0, b_1, b_2, \ldots$ 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 rigourously.

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. □

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 $\text{LIM}_{n\to\infty} a_n$, where $(a_n)_{n=1}^{\infty}$ is a Cauchy sequence of rational numbers. Two real numbers $\text{LIM}_{n\to\infty} a_n$ and $\text{LIM}_{n\to\infty} b_n$ 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 **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 laws 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 law 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