Preface

My primary goal in writing *Understanding Analysis* was to create an elementary one-semester book that exposes students to the rich rewards inherent in taking a mathematically rigorous approach to the study of functions of a real variable. The aim of a course in real analysis should be to challenge and improve mathematical intuition rather than to verify it. There is a tendency, however, to center an introductory course too closely around the familiar theorems of the standard calculus sequence. Producing a rigorous argument that polynomials are continuous is good evidence for a well-chosen definition of continuity, but it is not the reason the subject was created and certainly not the reason it should be required study. By shifting the focus to topics where an untrained intuition is severely disadvantaged (e.g., rearrangements of infinite series, nowhere-differentiable continuous functions, Fourier series), my intent is to restore an intellectual liveliness to this course by offering the beginning student access to some truly significant achievements of the subject.

The Main Objectives

In recent years, the standard undergraduate curriculum in mathematics has been subjected to steady pressure from several different sources. As computers and technology become more ubiquitous, so do the areas where mathematical thinking can be a valuable asset. Rather than preparing themselves for graduate study in pure mathematics, the present majority of mathematics majors look forward to careers in banking, medicine, law, and numerous other fields where analytical skills are desirable. Another strong influence on college mathematics is the ongoing calculus reform effort, now well over ten years old. At the core of this movement is the justifiable goal of presenting calculus in a more intuitive way, emphasizing geometric arguments over symbolic ones. Despite these various trends—or perhaps because of them—nearly every undergraduate mathematics program continues to require at least one semester of real analysis. The result is that instructors today are faced with the task of teaching a difficult, abstract course to a more diverse audience less familiar with the nature of axiomatic arguments.

The crux of the matter is that any prevailing sentiment in favor of marketing mathematics to larger groups must at some point be reconciled with the fact

that theoretical analysis is extremely challenging and even intimidating for some. One unfortunate resolution of this dilemma has been to make the course easier by making it less interesting. The omitted material is inevitably what gives analysis its true flavor. A better solution is to find a way to make the more advanced topics accessible and worth the effort.

I see three essential goals that a semester of real analysis should try to meet:

1. Students, especially those emerging from a reform approach to calculus, need to be convinced of the need for a more rigorous study of functions. The necessity of precise definitions and an axiomatic approach must be carefully motivated.

2. Having seen mainly graphical, numerical, or intuitive arguments, students need to learn what constitutes a rigorous mathematical proof and how to write one.

3. There needs to be significant reward for the difficult work of firming up the logical structure of limits. Specifically, real analysis should not be just an elaborate reworking of standard introductory calculus. Students should be exposed to the tantalizing complexities of the real line, to the subtleties of different flavors of convergence, and to the intellectual delights hidden in the paradoxes of the infinite.

The philosophy of *Understanding Analysis* is to focus attention on questions that give analysis its inherent fascination. Does the Cantor set contain any irrational numbers? Can the set of points where a function is discontinuous be arbitrary? Are derivatives continuous? Are derivatives integrable? Is an infinitely differentiable function necessarily the limit of its Taylor series? In giving these topics center stage, the hard work of a rigorous study is justified by the fact that *they are inaccessible without it.*

The Structure of the Book

This book is an introductory text. Although some fairly sophisticated topics are brought in early to advertise and motivate the upcoming material, the main body of each chapter consists of a lean and focused treatment of the core topics that make up the center of most courses in analysis. Fundamental results about completeness, compactness, sequential and functional limits, continuity, uniform convergence, differentiation, and integration are all incorporated. What is specific here is where the emphasis is placed. In the chapter on integration, for instance, the exposition revolves around deciphering the relationship between continuity and the Riemann integral. Enough properties of the integral are obtained to justify a proof of the Fundamental Theorem of Calculus, but the theme of the chapter is the pursuit of a characterization of integrable functions in terms of continuity. Whether or not Lebesgue's measure-zero criterion is treated, framing the material in this way is still valuable because it is the questions that are important. Mathematics is not a static discipline. Students

should be aware of the historical reasons for the creation of the mathematics they are learning and by extension realize that there is no last word on the subject. In the case of integration, this point is made explicitly by including some relatively recent developments on the generalized Riemann integral in the additional topics of the last chapter.

The structure of the chapters has the following distinctive features.

Discussion Sections: Each chapter begins with the discussion of some motivating examples and open questions. The tone in these discussions is intentionally informal, and full use is made of familiar functions and results from calculus. The idea is to freely explore the terrain, providing context for the upcoming definitions and theorems. A recurring theme is the resolution of the paradoxes that arise when operations that work well in finite settings are naively extended to infinite settings (e.g., differentiating an infinite series term-by-term, reversing the order of a double summation). After these exploratory introductions, the tone of the writing changes, and the treatment becomes rigorously tight but still not overly formal. With the questions in place, the need for the ensuing development of the material is well-motivated and the payoff is in sight.

Project Sections: The penultimate section of each chapter (the final section is a short epilogue) is written with the exercises incorporated into the exposition. Proofs are outlined but not completed, and additional exercises are included to elucidate the material being discussed. The point of this is to provide some flexibility. The sections are written as self-guided tutorials, but they can also be the subject of lectures. I have used them in place of a final examination, and they work especially well as collaborative assignments that can culminate in a class presentation. The body of each chapter contains the necessary tools, so there is some satisfaction in letting the students use their newly acquired skills to ferret out for themselves answers to questions that have been driving the exposition.

Building a Course

Teaching a satisfying class inevitably involves a race against time. Although this book is designed for a 12–14 week semester, there are still a few choices to make as to what to cover.

- The introductions can be discussed, assigned as reading, omitted, or substituted with something preferable. There are no theorems proved here that show up later in the text. I do develop some important examples in these introductions (the Cantor set, Dirichlet's nowhere-continuous function) that probably need to find their way into discussions at some point.

- Chapter 3, Basic Topology of $\mathbf{R}$, is much longer than it needs to be. All that is required by the ensuing chapters are fundamental results about open and closed sets and a thorough understanding of sequential compactness. The characterization of compactness using open covers as well

as the section on perfect and connected sets are included for their own intrinsic interest. They are not, however, crucial to any future proofs. The one exception to this is a presentation of the Intermediate Value Theorem (IVT) as a special case of the preservation of connected sets by continuous functions. To keep connectedness truly optional, I have included two direct proofs of IVT, one using least upper bounds and the other using nested intervals. A similar comment can be made about perfect sets. Although proofs of the Baire Category Theorem are nicely motivated by the argument that perfect sets are uncountable, it is certainly possible to do one without the other.

- All the project sections (1.5, 2.8, 3.5, 4.6, 5.4, 6.6, 7.6, 8.1–8.4) are optional in the sense that no results in later chapters depend on material in these sections. The four topics covered in Chapter 8 are also written in this project-style format, where the exercises make up a significant part of the development. The only one of these sections that might require a lecture is the unit on Fourier series, which is a bit longer than the others.

The Audience

The only prerequisite for this course is a robust understanding of the results from single-variable calculus. The theorems of linear algebra are not needed, but the exposure to abstract arguments and proof writing that usually comes with this course would be a valuable asset. Complex numbers are never used in this book.

The proofs in *Understanding Analysis* are written with the introductory student firmly in mind. Brevity and other stylistic concerns are postponed in favor of including a significant level of detail. Most proofs come with a fair amount of discussion about the context of the argument. What should the proof entail? Which definitions are relevant? What is the overall strategy? Is one particular proof similar to something already done? Whenever there is a choice, efficiency is traded for an opportunity to reinforce some previously learned technique. Especially familiar or predictable arguments are usually sketched as exercises so that students can participate directly in the development of the core material.

The search for recurring ideas exists at the proof-writing level and also on the larger expository level. I have tried to give the course a narrative tone by picking up on the unifying themes of approximation and the transition from the finite to the infinite. To paraphrase a passage from the end of the book, real numbers are approximated by rational ones; values of continuous functions are approximated by values nearby; curves are approximated by straight lines; areas are approximated by sums of rectangles; continuous functions are approximated by polynomials. In each case, the approximating objects are tangible and well-understood, and the issue is when and how well these qualities survive the limiting process. By focusing on this recurring pattern, each successive topic

builds on the intuition of the previous one. The questions seem more natural, and a method to the madness emerges from what might otherwise appear as a long list of theorems and proofs.

This book always emphasizes core ideas over generality, and it makes no effort to be a complete, deductive catalog of results. It is designed to capture the intellectual imagination. Those who become interested are then exceptionally well prepared for a second course starting from complex-valued functions on more general spaces, while those content with a single semester come away with a strong sense of the essence and purpose of real analysis. Turning once more to the concluding passages of Chapter 8, "By viewing the different infinities of mathematics through pathways crafted out of finite objects, Weierstrass and the other founders of analysis created a paradigm for how to extend the scope of mathematical exploration deep into territory previously unattainable."

This exploration has constituted the major thrill of my intellectual life. I am extremely pleased to offer this guide to what I feel are some of the most impressive highlights of the journey. Have a wonderful trip!

Acknowledgments

The genesis of this book came from an extended series of conversations with Benjamin Lotto of Vassar College. The structure of the early chapters and the book's overall thesis are in large part the result of several years of sharing classroom notes, ideas, and experiences with Ben. I am pleased with how the manuscript has turned out, and I have no doubt that it is an immeasurably better book because of Ben's early contributions.

A large part of the writing was done while I was enjoying a visiting position at the University of Virginia. Special thanks go to Nat Martin and Larry Thomas for being so generous with their time and wisdom, and especially to Loren Pitt, the scope of whose advice extends well beyond the covers of this book. I would also like to thank Julie Riddleberger for her help with many of the figures. Marian Robbins of Bellarmine College, Steve Kennedy of Carleton College, Paul Humke of Saint Olaf College, and Tom Kriete of the University of Virginia each taught from a preliminary draft of this text. I appreciate the many suggested improvements that this group provided, and I want to especially acknowledge Paul Humke for his contributions to the chapter on integration.

My department and the administration of Middlebury College have also been very supportive of this endeavor. David Guertin came to my technological rescue on numerous occasions, Priscilla Bremser read early chapter drafts, and Rick Chartrand's insightful opinions greatly improved some of the later sections. The list of students who have suffered through the long evolution of this book is now too long to present, but I would like to mention Brooke Sargent, whose meticulous class notes were the basis of the first draft, and Jesse Johnson, who has worked tirelessly to improve the presentation of the many exercises in the book. The production team at Springer has been absolutely first-rate. My sincere thanks goes to all of them with a special nod to Sheldon

Axler for encouragement and advice surely exceeding anything in his usual job description.

In a recent rereading of the completed text, I was struck by how frequently I resort to historical context to motivate an idea. This was not a conscious goal I set for myself. Instead, I feel it is a reflection of a very encouraging trend in mathematical pedagogy to humanize our subject with its history. From my own experience, a good deal of the credit for this movement in analysis should go to two books: *A Radical Approach to Real Analysis*, by David Bressoud, and *Analysis by Its History*, by E. Hairer and G. Wanner. Bressoud's book was particularly influential to the presentation of Fourier series in the last chapter. Either of these would make an excellent supplementary resource for this course. While I do my best to cite their historical origins when it seems illuminating or especially important, the present form of many of the theorems presented here belongs to the common folklore of the subject, and I have not attempted careful attribution. One exception is the material on the previously mentioned generalized Riemann integral, due independently to Jaroslav Kurzweil and Ralph Henstock. Section 8.1 closely follows the treatment laid out in Robert Bartle's article "Return to the Riemann Integral." In this paper, the author makes an italicized plea for teachers of mathematics to supplant Lebesgue's ubiquitous integral with the generalized Riemann integral. I hope that Professor Bartle will see its inclusion here as an inspired response to that request.

On a personal note, I welcome comments of any nature, and I will happily share any enlightening remarks—and any corrections—via a link on my webpage. The publication of this book comes nearly four years after the idea was first hatched. The long road to this point has required the steady support of many people but most notably that of my incredible wife, Katy. Amid the flurry of difficult decisions and hard work that go into a project of this size, the opportunity to dedicate this book to her comes as a pure and easy pleasure.

Middlebury, Vermont — Stephen Abbott
August 2000

Contents

Chapter 1

The Real Numbers

1.1 Discussion: The Irrationality of $\sqrt{2}$

Toward the end of his distinguished career, the renowned British mathematician G.H. Hardy eloquently laid out a justification for a life of studying mathematics in *A Mathematician's Apology*, an essay first published in 1940. At the center of Hardy's defense is the thesis that mathematics is an aesthetic discipline. For Hardy, the applied mathematics of engineers and economists held little charm. "Real mathematics," as he referred to it, "must be justified as art if it can be justified at all."

To help make his point, Hardy includes two theorems from classical Greek mathematics, which, in his opinion, possess an elusive kind of beauty that, although difficult to define, is easy to recognize. The first of these results is Euclid's proof that there are an infinite number of prime numbers. The second result is the discovery, attributed to the school of Pythagoras from around 500 B.C., that $\sqrt{2}$ is irrational. It is this second theorem that demands our attention. (A course in number theory would focus on the first.) The argument uses only arithmetic, but its depth and importance cannot be overstated. As Hardy says, "[It] is a 'simple' theorem, simple both in idea and execution, but there is no doubt at all about [it being] of the highest class. [It] is as fresh and significant as when it was discovered—two thousand years have not written a wrinkle on [it]."

Theorem 1.1.1. *There is no rational number whose square is 2.*

Proof. A rational number is any number that can be expressed in the form p/q, where p and q are integers. Thus, what the theorem asserts is that no matter how p and q are chosen, it is never the case that $(p/q)^2 = 2$. The line of attack is indirect, using a type of argument referred to as a proof by contradiction. The idea is to assume that there *is* a rational number whose square is 2 and then proceed along logical lines until we reach a conclusion that is unacceptable. At this point, we will be forced to retrace our steps and reject the erroneous

assumption that some rational number squared is equal to 2. In short, we will prove that the theorem is true by demonstrating that it cannot be false.

And so assume, for contradiction, that there exist integers p and q satisfying

$$\left(\frac{p}{q}\right)^2 = 2. \tag{1}$$

We may also assume that p and q have no common factor, because, if they had one, we could simply cancel it out and rewrite the fraction in lowest terms. Now, equation (1) implies

$$p^2 = 2q^2. \tag{2}$$

From this, we can see that the integer p^2 is an even number (it is divisible by 2), and hence p must be even as well because the square of an odd number is odd. This allows us to write $p = 2r$, where r is also an integer. If we substitute $2r$ for p in equation (2), then a little algebra yields the relationship

$$2r^2 = q^2.$$

But now the absurdity is at hand. This last equation implies that q^2 is even, and hence q must also be even. Thus, we have shown that p and q are both even (i.e., divisible by 2) when they were originally assumed to have no common factor. From this logical impasse, we can only conclude that equation (1) *cannot* hold for any integers p and q, and thus the theorem is proved. □

A component of Hardy's definition of beauty in a mathematical theorem is that the result have lasting and serious implications for a network of other mathematical ideas. In this case, the ideas under assault were the Greeks' understanding of the relationship between geometric *length* and arithmetic *number*. Prior to the preceding discovery, it was an assumed and commonly used fact that, given two line segments $\overline{AB}$ and $\overline{CD}$, it would always be possible to find a third line segment whose length divides evenly into the first two. In modern terminology, this is equivalent to asserting that the length of $\overline{CD}$ is a rational multiple of the length of $\overline{AB}$. Looking at the diagonal of a unit square (Fig. 1.1), it now followed (using the Pythagorean Theorem) that this was not always the case. Because the Pythagoreans implicitly interpreted number to mean rational number, they were forced to accept that number was a strictly weaker notion than length.

Rather than abandoning arithmetic in favor of geometry (as the Greeks seem to have done), our resolution to this limitation is to strengthen the concept of number by moving from the rational numbers to a larger number system. From a modern point of view, this should seem like a familiar and somewhat natural phenomenon. We begin with the *natural numbers*

$$\mathbf{N} = \{1, 2, 3, 4, 5, \dots\}.$$

The influential German mathematician Leopold Kronecker (1823–1891) once asserted that "The natural numbers are the work of God. All of the rest is

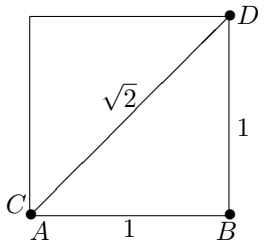


Figure 1.1: $\sqrt{2}$ EXISTS AS A GEOMETRIC LENGTH.

the work of mankind." Debating the validity of this claim is an interesting conversation for another time. For the moment, it at least provides us with a place to start. If we restrict our attention to the natural numbers $\mathbf{N}$, then we can perform addition perfectly well, but we must extend our system to the *integers*

$$\mathbf{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

if we want to have an additive identity (zero) and the additive inverses necessary to define subtraction. The next issue is multiplication and division. The number 1 acts as the multiplicative identity, but in order to define division we need to have multiplicative inverses. Thus, we extend our system again to the *rational numbers*

$$\mathbf{Q} = \left\{ \text{all fractions } \frac{p}{q} \text{ where } p \text{ and } q \text{ are integers with } q \neq 0 \right\}.$$

Taken together, the properties of $\mathbf{Q}$ discussed in the previous paragraph essentially make up the definition of what is called a *field*. More formally stated, a field is any set where addition and multiplication are well-defined operations that are commutative, associative, and obey the familiar distributive property $a(b+c) = ab+ac$. There must be an additive identity, and every element must have an additive inverse. Finally, there must be a multiplicative identity, and multiplicative inverses must exist for all nonzero elements of the field. Neither $\mathbf{Z}$ nor $\mathbf{N}$ is a field. The finite set $\{0, 1, 2, 3, 4\}$ is a field when addition and multiplication are computed modulo 5. This is not immediately obvious but makes an interesting exercise (Exercise 1.3.1).

The set $\mathbf{Q}$ also has a natural *order* defined on it. Given any two rational numbers r and s, exactly one of the following is true:

$$r < s, \quad r = s, \quad \text{or} \quad r > s.$$

This ordering is transitive in the sense that if $r < s$ and $s < t$, then $r < t$, so we are conveniently led to a mental picture of the rational numbers as being laid out from left to right along a number line. Unlike $\mathbf{Z}$, there are no intervals of empty space. Given any two rational numbers $r < s$, the rational number

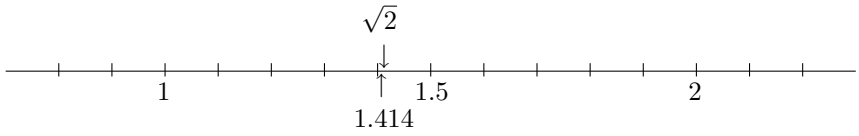


Figure 1.2: Approximating $\sqrt{2}$ with rational numbers.

$(r+s)/2$ sits halfway in between, implying that the rational numbers are densely nestled together.

With the field properties of $\mathbf{Q}$ allowing us to safely carry out the algebraic operations of addition, subtraction, multiplication, and division, let's remind ourselves just what it is that $\mathbf{Q}$ is lacking. By Theorem 1.1.1, it is apparent that we cannot always take square roots. The problem, however, is actually more fundamental than this. Using only rational numbers, it is possible to *approximate* $\sqrt{2}$ quite well (Fig. 1.2). For instance, $1.414^2 = 1.999396$. By adding more decimal places to our approximation, we can get even closer to a value for $\sqrt{2}$, but, even so, we are now well aware that there is a "hole" in the rational number line where $\sqrt{2}$ ought to be. Of course, there are quite a few other holes—at $\sqrt{3}$ and $\sqrt{5}$, for example. Returning to the dilemma of the ancient Greek mathematicians, if we want every length along the number line to correspond to an actual number, then another extension to our number system is in order. Thus, to the chain $\mathbf{N} \subseteq \mathbf{Z} \subseteq \mathbf{Q}$ we append the *real numbers* $\mathbf{R}$.

The question of how to actually construct $\mathbf{R}$ from $\mathbf{Q}$ is rather complicated business. It is discussed in Section 1.3, and then again in more detail in Section 8.4. For the moment, it is not too inaccurate to say that $\mathbf{R}$ is obtained by filling in the gaps in $\mathbf{Q}$. Wherever there is a hole, a new *irrational* number is defined and placed into the ordering that already exists on $\mathbf{Q}$. The real numbers are then the union of these irrational numbers together with the more familiar rational ones. What properties does the set of irrational numbers have? How do the sets of rational and irrational numbers fit together? Is there a kind of symmetry between the rationals and the irrationals, or is there some sense in which we can argue that one type of real number is more common than the other? The one method we have seen so far for generating examples of irrational numbers is through square roots. Not too surprisingly, other roots such as $\sqrt[3]{2}$ or $\sqrt[5]{3}$ are most often irrational. Can all irrational numbers be expressed as algebraic combinations of nth roots and rational numbers, or are there still other irrational numbers beyond those of this form?

1.2 Some Preliminaries

The vocabulary necessary for the ensuing development comes from set theory and the theory of functions. This should be familiar territory, but a brief review

of the terminology is probably a good idea, if only to establish some agreed-upon notation.

Sets

Intuitively speaking, a *set* is any collection of objects. These objects are referred to as the *elements* of the set. For our purposes, the sets in question will most often be sets of real numbers, although we will also encounter sets of functions and, on a few rare occasions, sets whose elements are other sets.

Given a set A, we write $x \in A$ if x (whatever it may be) is an element of A. If x is not an element of A, then we write $x \notin A$. Given two sets A and B, the *union* is written $A \cup B$ and is defined by asserting that

$$x \in A \cup B \text{ provided that } x \in A \text{ or } x \in B \text{ (or potentially both).}$$

The *intersection* $A \cap B$ is the set defined by the rule

$$x \in A \cap B \text{ provided } x \in A \text{ and } x \in B.$$

Example 1.2.1. (i) There are many acceptable ways to assert the contents of a set. In the previous section, the set of natural numbers was defined by listing the elements: $\mathbf{N} = \{1, 2, 3, \dots\}$.

(ii) Sets can also be described in words. For instance, we can define the set E to be the collection of even natural numbers.

(iii) Sometimes it is more efficient to provide a kind of rule or algorithm for determining the elements of a set. As an example, let

$$S = \{r \in \mathbf{Q} : r^2 < 2\}.$$

Read aloud, the definition of S says, "Let S be the set of all rational numbers whose squares are less than 2." It follows that $1 \in S$, $4/3 \in S$, but $3/2 \notin S$ because $9/4 \geq 2$.

Using the previously defined sets to illustrate the operations of intersection and union, we observe that

$$\mathbf{N} \cup E = \mathbf{N}, \quad \mathbf{N} \cap E = E, \quad \mathbf{N} \cap S = \{1\}, \text{ and } E \cap S = \emptyset.$$

The set $\emptyset$ is called the *empty set* and is understood to be the set that contains no elements. An equivalent statement would be to say that E and S are *disjoint*.

A word about the equality of two sets is in order (since we have just used the notion). The *inclusion* relationship $A \subseteq B$ or $B \supseteq A$ is used to indicate that every element of A is also an element of B. In this case, we say A is a *subset* of B, or B *contains* A. To assert that $A = B$ means that $A \subseteq B$ and $B \subseteq A$. Put another way, A and B have exactly the same elements.

Quite frequently in the upcoming chapters, we will want to apply the union and intersection operations to infinite collections of sets.

Example 1.2.2. Let

$$\begin{aligned} A_1 &= \mathbf{N} = \{1, 2, 3, \dots\}, \\ A_2 &= \{2, 3, 4, \dots\}, \\ A_3 &= \{3, 4, 5, \dots\}, \end{aligned}$$

and, in general, for each $n \in \mathbf{N}$, define the set

$$A_n = \{n, n+1, n+2, \dots\}.$$

The result is a nested chain of sets

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq A_4 \supseteq \cdots,$$

where each successive set is a subset of all the previous ones. Notationally,

$$\bigcup_{n=1}^{\infty} A_n, \quad \bigcup_{n \in \mathbf{N}} A_n, \quad \text{or} \quad A_1 \cup A_2 \cup A_3 \cup \cdots$$

are all equivalent ways to indicate the set whose elements consist of any element that appears in at least one particular A_n. Because of the nested property of this particular collection of sets, it is not too hard to see that

$$\bigcup_{n=1}^{\infty} A_n = A_1.$$

The notion of intersection has the same kind of natural extension to infinite collections of sets. For this example, we have

$$\bigcap_{n=1}^{\infty} A_n = \emptyset.$$

Let's be sure we understand why this is the case. Suppose we had some natural number m that we thought might actually satisfy $m \in \bigcap_{n=1}^{\infty} A_n$. What this would mean is that $m \in A_n$ for *every* A_n in our collection of sets. Because m is not an element of A_{m+1}, no such m exists and the intersection is empty.

As mentioned, most of the sets we encounter will be sets of real numbers. Given $A \subseteq \mathbf{R}$, the *complement* of A, written A^c, refers to the set of all elements of $\mathbf{R}$ not in A. Thus, for $A \subseteq \mathbf{R}$,

$$A^c = \{x \in \mathbf{R} : x \notin A\}.$$

A few times in our work to come, we will refer to De Morgan's Laws, which state that

$$(A \cap B)^c = A^c \cup B^c \quad \text{and} \quad (A \cup B)^c = A^c \cap B^c.$$

Proofs of these statements are discussed in Exercise 1.2.3.

Admittedly, there is something imprecise about the definition of set presented at the beginning of this discussion. The defining sentence begins with the phrase "Intuitively speaking," which might seem an odd way to embark on a course of study that purportedly intends to supply a rigorous foundation for the theory of functions of a real variable. In some sense, however, this is unavoidable. Each repair of one level of the foundation reveals something below it in need of attention. The theory of sets has been subjected to intense scrutiny over the past century precisely because so much of modern mathematics rests on this foundation. But such a study is really only advisable once it is understood why our naive impression about the behavior of sets is insufficient. For the direction in which we are heading, this will not happen, although an indication of some potential pitfalls is given in Section 1.6.

Functions

Definition 1.2.3. Given two sets A and B, a *function* from A to B is a rule or mapping that takes each element $x \in A$ and associates with it a single element of B. In this case, we write $f : A \to B$. Given an element $x \in A$, the expression $f(x)$ is used to represent the element of B associated with x by f. The set A is called the *domain* of f. The *range* of f is not necessarily equal to B but refers to the subset of B given by $\{y \in B : y = f(x) \text{ for some } x \in A\}$.

This definition of function is more or less the one proposed by Peter Lejeune Dirichlet (1805–1859) in the 1830s. Dirichlet was a German mathematician who was one of the leaders in the development of the rigorous approach to functions that we are about to undertake. His main motivation was to unravel the issues surrounding the convergence of Fourier series. Dirichlet's contributions figure prominently in Section 8.3, where an introduction to Fourier series is presented, but we will also encounter his name in several earlier chapters along the way. What is important at the moment is that we see how Dirichlet's definition of function liberates the term from its interpretation as a type of "formula." In the years leading up to Dirichlet's time, the term "function" was generally understood to refer to algebraic entities such as $f(x) = x^2+1$ or $g(x) = \sqrt{x^4 + 4}$. Definition 1.2.3 allows for a much broader range of possibilities.

Example 1.2.4. In 1829, Dirichlet proposed the unruly function

$$g(x) = \begin{cases} 1 & \text{if } x \in \mathbf{Q} \\ 0 & \text{if } x \notin \mathbf{Q}. \end{cases}$$

The domain of g is all of $\mathbf{R}$, and the range is the set $\{0, 1\}$. There is no single formula for g in the usual sense, and it is quite difficult to graph this function (see Section 4.1 for a rough attempt), but it certainly qualifies as a function according to the criterion in Definition 1.2.3. As we study the theoretical nature of continuous, differentiable, or integrable functions, examples such as this one will provide us with an invaluable testing ground for the many conjectures we encounter.

Example 1.2.5 (Triangle Inequality). The *absolute value function* is so important that it merits the special notation $|x|$ in place of the usual $f(x)$ or $g(x)$. It is defined for every real number via the piecewise definition

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

With respect to multiplication and division, the absolute value function satisfies

(i) $|ab| = |a||b|$ and

(ii) $|a + b| \leq |a| + |b|$

for all choices of a and b. Verifying these properties (Exercise 1.2.4) is just a matter of examining the different cases that arise when a, b, and $a+b$ are positive and negative. Property (ii) is called the *triangle inequality*. This innocuous looking inequality turns out to be fantastically important and will be frequently employed in the following way. Given three real numbers a, b, and c, we certainly have

$$|a - b| = |(a - c) + (c - b)|.$$

By the triangle inequality,

$$|(a - c) + (c - b)| \leq |a - c| + |c - b|,$$

so we get

$$|a - b| \leq |a - c| + |c - b|. \tag{1}$$

Now, the expression $|a - b|$ is equal to $|b - a|$ and is best understood as the *distance* between the points a and b on the number line. With this interpretation, equation (1) makes the plausible statement that "the distance from a to b is less than or equal to the distance from a to c plus the distance from c to b." Pretending for a moment that these are points in the plane (instead of on the real line), it should be evident why this is referred to as the "triangle inequality."

Logic and Proofs

Writing rigorous mathematical proofs is a skill best learned by doing, and there is plenty of on-the-job training just ahead. As Hardy indicates, there is an artistic quality to mathematics of this type, which may or may not come easily, but that is not to say that anything especially mysterious is happening. A proof is an essay of sorts. It is a set of carefully crafted directions, which, when followed, should leave the reader absolutely convinced of the truth of the proposition in question. To achieve this, the steps in a proof must follow logically from previous steps or be justified by some other agreed-upon set of facts. In addition to being valid, these steps must also fit coherently together to form a cogent argument. Mathematics has a specialized vocabulary, to be sure,

but that does not exempt a good proof from being written in grammatically correct English.

The one proof we have seen at this point (to Theorem 1.1.1) uses an indirect strategy called *proof by contradiction*. This powerful technique will be employed a number of times in our upcoming work. Nevertheless, most proofs are direct. (It also bears mentioning that using an indirect proof when a direct proof is available is generally considered bad manners.) A direct proof begins from some valid statement, most often taken from the theorem's hypothesis, and then proceeds through rigorously logical deductions to a demonstration of the theorem's conclusion. As we saw in Theorem 1.1.1, an indirect proof always begins by negating what it is we would like to prove. This is not always as easy to do as it may sound. The argument then proceeds until (hopefully) a logical contradiction with some other accepted fact is uncovered. Many times, this accepted fact is part of the hypothesis of the theorem. When the contradiction is with the theorem's hypothesis, we technically have what is called a *contrapositive* proof.

The next proposition illustrates a number of the issues just discussed and introduces a few more.

Theorem 1.2.6. *Two real numbers a and b are equal if and only if for every real number $\epsilon > 0$ it follows that $|a - b| < \epsilon$.*

Proof. There are two key phrases in the statement of this proposition that warrant special attention. One is "for every," which will be addressed in a moment. The other is "if and only if." To say "if and only if" in mathematics is an economical way of stating that the proposition is true in two directions. In the forward direction, we must prove the statement:

($\Rightarrow$) *If $a = b$, then for every real number $\epsilon > 0$ it follows that $|a - b| < \epsilon$.*

We must also prove the converse statement:

($\Leftarrow$) *If for every real number $\epsilon > 0$ it follows that $|a - b| < \epsilon$, then we must have $a = b$.*

For the proof of the first statement, there is really not much to say. If $a = b$, then $|a - b| = 0$, and so certainly $|a - b| < \epsilon$ no matter what $\epsilon > 0$ is chosen.

For the second statement, we give a proof by contradiction. The conclusion of the proposition in this direction states that $a = b$, so we assume that $a \neq b$. Heading off in search of a contradiction brings us to a consideration of the phrase "for every $\epsilon > 0$." Some equivalent ways to state the hypothesis would be to say that "for all possible choices of $\epsilon > 0$" or "no matter how $\epsilon > 0$ is selected, it is always the case that $|a - b| < \epsilon$." But assuming $a \neq b$ (as we are doing at the moment), the choice of

$$\epsilon_0 = |a - b| > 0$$

poses a serious problem. We are assuming that $|a - b| < \epsilon$ is true for *every* $\epsilon > 0$, so this must certainly be true of the particular ϵ_0 just defined. However, the statements

$$|a - b| < \epsilon_0 \quad \text{and} \quad |a - b| = \epsilon_0$$

cannot both be true. This contradiction means that our initial assumption that $a \neq b$ is unacceptable. Therefore, $a = b$, and the indirect proof is complete. □

One of the most fundamental skills required for reading and writing analysis proofs is the ability to confidently manipulate the quantifying phrases "for all" and "there exists." Significantly more attention will be given to this issue in many upcoming discussions.

Induction

One final trick of the trade, which will arise with some frequency, is the use of *induction* arguments. Induction is used in conjunction with the natural numbers $\mathbf{N}$ (or sometimes with the set $\mathbf{N} \cup \{0\}$). The fundamental principle behind induction is that if S is some subset of $\mathbf{N}$ with the property that

(i) S contains 1 and

(ii) whenever S contains a natural number n, it also contains $n + 1$,

then it must be that $S = \mathbf{N}$. As the next example illustrates, this principle can be used to define sequences of objects as well as to prove facts about them.

Example 1.2.7. Let $x_1 = 1$, and for each $n \in \mathbf{N}$ define

$$x_{n+1} = (1/2)x_n + 1.$$

Using this rule, we can compute $x_2 = (1/2)(1) + 1 = 3/2$, $x_3 = 7/4$, and it is immediately apparent how this leads to a definition of x_n for all $n \in \mathbf{N}$.

The sequence just defined appears at the outset to be increasing. For the terms computed, we have $x_1 \leq x_2 \leq x_3$. Let's use induction to prove that this trend continues; that is, let's show

$$x_n \leq x_{n+1} \tag{2}$$

for all values of $n \in \mathbf{N}$.

For $n = 1$, $x_1 = 1$ and $x_2 = 3/2$, so that $x_1 \leq x_2$ is clear. Now, we want to show that

if we have $x_n \leq x_{n+1}$, then it follows that $x_{n+1} \leq x_{n+2}$.

Think of S as the set of natural numbers for which the claim in equation (2) is true. We have shown that $1 \in S$. We are now interested in showing that if $n \in S$, then $n+1 \in S$ as well. Starting from the induction hypothesis $x_n \leq x_{n+1}$, we can multiply across the inequality by $1/2$ and add 1 to get

$$\frac{1}{2}x_n + 1 \leq \frac{1}{2}x_{n+1} + 1,$$

which is precisely the desired conclusion $x_{n+1} \leq x_{n+2}$. By induction, the claim is proved for all $n \in \mathbf{N}$.

Any discussion about why induction is a valid argumentative technique immediately opens up a box of questions about how we understand the natural numbers. Earlier, in Section 1.1, we avoided this issue by referencing Kronecker's famous comment that the natural numbers are somehow divinely given. Although we will not improve on this explanation here, it should be pointed out that a more atheistic and mathematically satisfying approach to $\mathbf{N}$ is possible from the point of view of axiomatic set theory. This brings us back to a recurring theme of this chapter. Pedagogically speaking, the foundations of mathematics are best learned and appreciated in a kind of reverse order. A rigorous study of the natural numbers and the theory of sets is certainly recommended, but only after we have an understanding of the subtleties of the real number system. It is this latter topic that is the business of real analysis.

Exercises

Exercise 1.2.1. (a) Prove that $\sqrt{3}$ is irrational. Does a similar argument work to show $\sqrt{6}$ is irrational?

(b) Where does the proof of Theorem 1.1.1 break down if we try to use it to prove $\sqrt{4}$ is irrational?

Exercise 1.2.2. Decide which of the following represent true statements about the nature of sets. For any that are false, provide a specific example where the statement in question does not hold.

(a) If $A_1 \supseteq A_2 \supseteq A_3 \supseteq A_4 \cdots$ are all sets containing an infinite number of elements, then the intersection $\cap_{n=1}^{\infty} A_n$ is infinite as well.

(b) If $A_1 \supseteq A_2 \supseteq A_3 \supseteq A_4 \cdots$ are all finite, nonempty sets of real numbers, then the intersection $\cap_{n=1}^{\infty} A_n$ is finite and nonempty.

(c) $A \cap (B \cup C) = (A \cap B) \cup C$.

(d) $A \cap (B \cap C) = (A \cap B) \cap C$.

(e) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Exercise 1.2.3 (De Morgan's Laws). Let A and B be subsets of $\mathbf{R}$.

(a) If $x \in (A \cap B)^c$, explain why $x \in A^c \cup B^c$. This shows that $(A \cap B)^c \subseteq A^c \cup B^c$.

(b) Prove the reverse inclusion $(A \cap B)^c \supseteq A^c \cup B^c$, and conclude that $(A \cap B)^c = A^c \cup B^c$.

(c) Show $(A \cup B)^c = A^c \cap B^c$ by demonstrating inclusion both ways.

Exercise 1.2.4. Verify the triangle inequality in the special cases where

(a) a and b have the same sign;

(b) $a \geq 0$, $b < 0$, and $a + b \geq 0$.

Exercise 1.2.5. Use the triangle inequality to establish the inequalities

(a) $|a - b| \leq |a| + |b|$;

(b) $||a| - |b|| \leq |a - b|$.

Exercise 1.2.6. Given a function f and a subset A of its domain, let $f(A)$ represent the range of f over the set A; that is, $f(A) = \{f(x) : x \in A\}$.

(a) Let $f(x) = x^2$. If $A = [0, 2]$ (the closed interval $\{x \in \mathbf{R} : 0 \leq x \leq 2\}$) and $B = [1, 4]$, find $f(A)$ and $f(B)$. Does $f(A \cap B) = f(A) \cap f(B)$ in this case? Does $f(A \cup B) = f(A) \cup f(B)$?

(b) Find two sets A and B for which $f(A \cap B) \neq f(A) \cap f(B)$.

(c) Show that, for an arbitrary function $g : \mathbf{R} \to \mathbf{R}$, it is always true that $g(A \cap B) \subseteq g(A) \cap g(B)$ for all sets $A, B \subseteq \mathbf{R}$.

(d) Form and prove a conjecture about the relationship between $g(A \cup B)$ and $g(A) \cup g(B)$ for an arbitrary function g.

Exercise 1.2.7. Given a function $f : D \to \mathbf{R}$ and a subset $B \subseteq \mathbf{R}$, let $f^{-1}(B)$ be the set of all points from the domain D that get mapped into B; that is, $f^{-1}(B) = \{x \in D : f(x) \in B\}$. This set is called the *preimage* of B.

(a) Let $f(x) = x^2$. If A is the closed interval $[0, 4]$ and B is the closed interval $[-1, 1]$, find $f^{-1}(A)$ and $f^{-1}(B)$. Does $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$ in this case? Does $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$?

(b) The good behavior of preimages demonstrated in (a) is completely general. Show that for an arbitrary function $g : \mathbf{R} \to \mathbf{R}$, it is always true that $g^{-1}(A \cap B) = g^{-1}(A) \cap g^{-1}(B)$ and $g^{-1}(A \cup B) = g^{-1}(A) \cup g^{-1}(B)$ for all sets $A, B \subseteq \mathbf{R}$.

Exercise 1.2.8. Form the logical negation of each claim. One way to do this is to simply add "It is not the case that..." in front of each assertion, but for each statement, try to embed the word "not" as deeply into the resulting sentence as possible (or avoid using it altogether).

(a) For all real numbers satisfying $a < b$, there exists an $n \in \mathbf{N}$ such that $a + 1/n < b$.

(b) Between every two distinct real numbers, there is a rational number.

(c) For all natural numbers $n \in \mathbf{N}$, $\sqrt{n}$ is either a natural number or an irrational number.

(d) Given any real number $x \in \mathbf{R}$, there exists $n \in \mathbf{N}$ satisfying $n > x$.

Exercise 1.2.9. Show that the sequence $(x_1, x_2, x_3, \dots)$ defined in Example 1.2.7 is bounded above by 2; that is, prove that $x_n \leq 2$ for every $n \in \mathbf{N}$.

Exercise 1.2.10. Let $y_1 = 1$, and for each $n \in \mathbf{N}$ define $y_{n+1} = (3y_n + 4)/4$.

(a) Use induction to prove that the sequence satisfies $y_n < 4$ for all $n \in \mathbf{N}$.

(b) Use another induction argument to show the sequence $(y_1, y_2, y_3, \dots)$ is increasing.

Exercise 1.2.11. If a set A contains n elements, prove that the number of different subsets of A is equal to 2^n. (Keep in mind that the empty set $\emptyset$ is considered to be a subset of every set.)

Exercise 1.2.12. For this exercise, assume Exercise 1.2.3 has been successfully completed.

(a) Show how induction can be used to conclude that

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$

for any finite $n \in \mathbf{N}$.

(b) Explain why induction *cannot* be used to conclude

$$\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c.$$

It might be useful to consider part (a) of Exercise 1.2.2.

(c) Is the statement in part (b) valid? If so, write a proof that does not use induction.

1.3 The Axiom of Completeness

What exactly is a real number? In Section 1.1, we got as far as saying that the set $\mathbf{R}$ of real numbers is an extension of the rational numbers $\mathbf{Q}$ in which there are no holes or gaps. We want every length along the number line—such as $\sqrt{2}$—to correspond to a real number and vice versa.

We are going to improve on this definition, but as we do so, it is important to keep in mind our earlier acknowledgment that whatever precise statements we formulate will necessarily rest on other unproven assumptions or undefined terms. At some point, we must draw a line and confess that this is what we have decided to accept as a reasonable place to start. Naturally, there is some debate about where this line should be drawn. One way to view the mathematics of the 19th and 20th centuries is as a stalwart attempt to move this line further and further back toward some unshakable foundation. The majority of the material covered in this book is attributable to the mathematicians working in the early and middle parts of the 1800s. Augustin Louis Cauchy (1789–1857), Bernhard Bolzano (1781–1848), Niels Henrik Abel (1802–1829), Peter Lejeune Dirichlet, Karl Weierstrass (1815–1897), and Bernhard Riemann (1826–1866) all figure prominently in the discovery of the theorems that follow. But here is the interesting point. Nearly all of this work was done using intuitive assumptions about the nature of $\mathbf{R}$ quite similar to our own informal understanding at this point. Eventually, enough scrutiny was directed at the detailed structure of $\mathbf{R}$ so that, in the 1870s, a handful of ways to rigorously *construct* $\mathbf{R}$ from $\mathbf{Q}$ were proposed.

Following this historical model, our own rigorous construction of $\mathbf{R}$ from $\mathbf{Q}$ is postponed until Section 8.4. By this point, the need for such a construction will be more justified and easier to appreciate. In the meantime, we have many proofs to write, so it is important to lay down, as explicitly as possible, the assumptions that we intend to make about the real numbers.

An Initial Definition for R

First, $\mathbf{R}$ is a set containing $\mathbf{Q}$. The operations of addition and multiplication on $\mathbf{Q}$ extend to all of $\mathbf{R}$ in such a way that every element of $\mathbf{R}$ has an additive inverse and every nonzero element of $\mathbf{R}$ has a multiplicative inverse. Echoing

the discussion in Section 1.1, we assume $\mathbf{R}$ is a *field*, meaning that addition and multiplication of real numbers is commutative, associative, and the distributive property holds. This allows us to perform all of the standard algebraic manipulations that are second nature to us. We also assume that the familiar properties of the ordering on $\mathbf{Q}$ extend to all of $\mathbf{R}$. Thus, for example, such deductions as "If $a < b$ and $c > 0$, then $ac < bc$" will be carried out freely without much comment. To summarize the situation in the official terminology of the subject, we assume that $\mathbf{R}$ is an *ordered field*, which contains $\mathbf{Q}$ as a subfield. (A rigorous definition of "ordered field" is presented in Section 8.4.)

This brings us to the final, and most distinctive, assumption about the real number system. We must find some way to clearly articulate what we mean by insisting that $\mathbf{R}$ does not contain the gaps that permeate $\mathbf{Q}$. Because this is the defining difference between the rational numbers and the real numbers, we will be excessively precise about how we phrase this assumption, hereafter referred to as the *Axiom of Completeness*.

Axiom of Completeness. *Every nonempty set of real numbers that is bounded above has a least upper bound.*

Now, what exactly does this mean?

Least Upper Bounds and Greatest Lower Bounds

Let's first state the relevant definitions, and then look at some examples.

Definition 1.3.1. A set $A \subseteq \mathbf{R}$ is *bounded above* if there exists a number $b \in \mathbf{R}$ such that $a \leq b$ for all $a \in A$. The number b is called an *upper bound* for A.

Similarly, the set A is *bounded below* if there exists a *lower bound* $l \in \mathbf{R}$ satisfying $l \leq a$ for every $a \in A$.

Definition 1.3.2. A real number s is the *least upper bound* for a set $A \subseteq \mathbf{R}$ if it meets the following two criteria:

(i) s is an upper bound for A;

(ii) if b is any upper bound for A, then $s \leq b$.

The least upper bound is also frequently called the *supremum* of the set A. Although the notation $s = \text{lub}\, A$ is still common, we will always write $s = \sup A$ for the least upper bound.

The *greatest lower bound* or *infimum* for A is defined in a similar way (Exercise 1.3.2) and is denoted by $\inf A$ (Fig. 1.3).

Although a set can have a host of upper bounds, it can have only one *least* upper bound. If s_1 and s_2 are both least upper bounds for a set A, then by property (ii) in Definition 1.3.2 we can assert $s_1 \leq s_2$ and $s_2 \leq s_1$. The conclusion is that $s_1 = s_2$ and least upper bounds are unique.

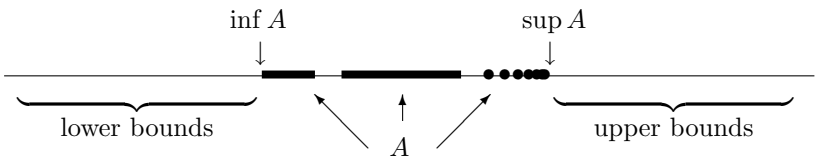


Figure 1.3: DEFINITION OF $\sup A$ AND $\inf A$.

Example 1.3.3. Let

$$A = \left\{ \frac{1}{n} : n \in N \right\} = \left\{ 1, \frac{1}{2}, \frac{1}{3}, \dots \right\}.$$

The set A is bounded above and below. Successful candidates for an upper bound include 3, 2, and 3/2. For the least upper bound, we claim $\sup A = 1$. To argue this rigorously using Definition 1.3.2, we need to verify that properties (i) and (ii) hold. For (i), we just observe that $1 \geq 1/n$ for all choices of $n \in \mathbf{N}$. To verify (ii), we begin by assuming we are in possession of some other upper bound b. Because $1 \in A$ and b is an upper bound for A, we must have $1 \leq b$. This is precisely what property (ii) asks us to show.

Although we do not quite have the tools we need for a rigorous proof (see Theorem 1.4.2), it should be somewhat apparent that $\inf A = 0$.

An important lesson to take from Example 1.3.3 is that $\sup A$ and $\inf A$ may or may not be elements of the set A. This issue is tied to understanding the crucial difference between the maximum and the supremum (or the minimum and the infimum) of a given set.

Definition 1.3.4. A real number a_0 is a *maximum* of the set A if a_0 is an element of A and $a_0 \geq a$ for all $a \in A$. Similarly, a number a_1 is a *minimum* of A if $a_1 \in A$ and $a_1 \leq a$ for every $a \in A$.

Example 1.3.5. To belabor the point, consider the open interval

$$(0, 2) = \{x \in \mathbf{R} : 0 < x < 2\},$$

and the closed interval

$$[0, 2] = \{x \in \mathbf{R} : 0 \leq x \leq 2\}.$$

Both sets are bounded above (and below), and both have the same least upper bound, namely 2. It is *not* the case, however, that both sets have a maximum. A maximum is a specific type of upper bound that is required to be an element of the set in question, and the open interval $(0, 2)$ does not possess such an element. Thus, the supremum can exist and not be a maximum, but when a maximum exists then it is also the supremum.

Let's turn our attention back to the Axiom of Completeness. Although we can see now that not every nonempty bounded set contains a maximum, the Axiom of Completeness asserts that every such set does have a least upper bound. We are not going to prove this. An *axiom* in mathematics is an accepted assumption, to be used without proof. Preferably, an axiom should be an elementary statement about the system in question that is so fundamental that it seems to need no justification. Perhaps the Axiom of Completeness fits this description, and perhaps it does not. Before deciding, let's remind ourselves why *it is not a valid statement about* $\mathbf{Q}$.

Example 1.3.6. Consider again the set

$$S = \{r \in \mathbf{Q} : r^2 < 2\},$$

and pretend for the moment that our world consists only of rational numbers. The set S is certainly bounded above. Taking $b = 2$ works, as does $b = 3/2$. But notice what happens as we go in search of the *least* upper bound. (It may be useful here to know that the decimal expansion for $\sqrt{2}$ begins $1.4142\ldots$.) We might try $b = 142/100$, which is indeed an upper bound, but then we discover that $b = 1415/1000$ is an upper bound that is smaller still. Is there a smallest one?

In the rational numbers, there is not. In the real numbers, there is. Back in $\mathbf{R}$, the Axiom of Completeness states that we may set $\alpha = \sup S$ and be confident that such a number exists. In the next section, we will prove that $\alpha^2 = 2$. But according to Theorem 1.1.1, this implies α is not a rational number. If we are restricting our attention to only rational numbers, then α is not an allowable option for $\sup S$, and the search for a least upper bound goes on indefinitely. Whatever rational upper bound is discovered, it is always possible to find one smaller.

The tools needed to carry out the computations described in Example 1.3.6 depend on some results about how $\mathbf{Q}$ and $\mathbf{N}$ fit inside of $\mathbf{R}$. These are discussed in the next section.

We now give an equivalent and useful way of characterizing least upper bounds. Recall that Definition 1.3.2 of the supremum has two parts. Part (i) says that $\sup A$ must be an upper bound, and part (ii) states that it must be the smallest one. The following lemma offers an alternative way to restate part (ii).

Lemma 1.3.7. *Assume $s \in \mathbf{R}$ is an upper bound for a set $A \subseteq \mathbf{R}$. Then, $s = \sup A$ if and only if, for every choice of $\epsilon > 0$, there exists an element $a \in A$ satisfying $s - \epsilon < a$.*

Proof. Here is a short rephrasing of the lemma: Given that s is an upper bound, s is the least upper bound if and only if any number smaller than s is not an upper bound. Putting it this way almost qualifies as a proof, but we will expand on what exactly is being said in each direction.

($\Rightarrow$) For the forward direction, we assume $s = \sup A$ and consider $s-\epsilon$, where $\epsilon > 0$ has been arbitrarily chosen. Because $s - \epsilon < s$, part (ii) of Definition 1.3.2 implies that $s - \epsilon$ is *not* an upper bound for A. If this is the case, then there must be some element $a \in A$ for which $s - \epsilon < a$ (because otherwise $s - \epsilon$ would be an upper bound). This proves the lemma in one direction.

($\Leftarrow$) Conversely, assume s is an upper bound with the property that no matter how $\epsilon > 0$ is chosen, $s - \epsilon$ is no longer an upper bound for A. Notice that what this implies is that if b is any number less than s, then b is not an upper bound. (Just let $\epsilon = s - b$.) To prove that $s = \sup A$, we must verify part (ii) of Definition 1.3.2. (Read it again.) Because we have just argued that any number smaller than s cannot be an upper bound, it follows that if b is some other upper bound for A, then $b \geq s$. □

It is certainly the case that all of our conclusions to this point about least upper bounds have analogous versions for greatest lower bounds. The Axiom of Completeness does not explicitly assert that a nonempty set bounded below has an infimum, but this is because we do not need to assume this fact as part of the axiom. Using the Axiom of Completeness, there are several ways to prove that greatest lower bounds exist for bounded sets. One such proof is explored in Exercise 1.3.3.

Exercises

Exercise 1.3.1. Let $\mathbf{Z}_5 = \{0, 1, 2, 3, 4\}$ and define addition and multiplication modulo 5. In other words, compute the integer remainder when $a + b$ and ab are divided by 5, and use this as the value for the sum and product, respectively.

(a) Show that, given any element $z \in \mathbf{Z}_5$, there exists an element y such that $z + y = 0$. The element y is called the *additive inverse* of z.

(b) Show that, given any $z \neq 0$ in $\mathbf{Z}_5$, there exists an element x such that $zx = 1$. The element x is called the *multiplicative inverse* of z.

(c) The existence of additive and multiplicative inverses is part of the definition of a *field*. Investigate the set $\mathbf{Z}_4 = \{0, 1, 2, 3\}$ (where addition and multiplication are defined modulo 4) for the existence of additive and multiplicative inverses. Make a conjecture about the values of n for which additive inverses exist in $\mathbf{Z}_n$, and then form another conjecture about the existence of multiplicative inverses.

Exercise 1.3.2. (a) Write a formal definition in the style of Definition 1.3.2 for the *infimum* or *greatest lower bound* of a set.

(b) Now, state and prove a version of Lemma 1.3.7 for greatest lower bounds.

Exercise 1.3.3. (a) Let A be bounded below, and define $B = \{b \in \mathbf{R} : b \text{ is a lower bound for } A\}$. Show that $\sup B = \inf A$.

(b) Use (a) to explain why there is no need to assert that greatest upper bounds exist as part of the Axiom of Completeness.

(c) Propose another way to use the Axiom of Completeness to prove that sets bounded below have greatest lower bounds.

Exercise 1.3.4. Assume that A and B are nonempty, bounded above, and satisfy $B \subseteq A$. Show $\sup B \leq \sup A$.

Exercise 1.3.5. Let $A \subseteq \mathbf{R}$ be bounded above, and let $c \in \mathbf{R}$. Define the sets $c + A$ and cA by $c + A = \{c + a : a \in A\}$ and $cA = \{ca : a \in A\}$.

(a) Show that $\sup(c + A) = c + \sup A$.

(b) If $c \geq 0$, show that $\sup(cA) = c \sup A$.

(c) Postulate a similar type of statement for $\sup(cA)$ for the case $c < 0$.

Exercise 1.3.6. Compute, without proofs, the suprema and infima of the following sets:

(a) $\{n \in \mathbf{N} : n^2 < 10\}$.

(b) $\{n/(m + n) : m, n \in \mathbf{N}\}$.

(c) $\{n/(2n + 1) : n \in \mathbf{N}\}$.

(d) $\{n/m : m, n \in \mathbf{N} \text{ with } m + n \leq 10\}$.

Exercise 1.3.7. Prove that if a is an upper bound for A, and if a is also an element of A, then it must be that $a = \sup A$.

Exercise 1.3.8. If $\sup A < \sup B$, then show that there exists an element $b \in B$ that is an upper bound for A.

Exercise 1.3.9. Without worrying about formal proofs for the moment, decide if the following statements about suprema and infima are true or false. For any that are false, supply an example where the claim in question does not appear to hold.

(a) A finite, nonempty set always contains its supremum.

(b) If $a < L$ for every element a in the set A, then $\sup A < L$.

(c) If A and B are sets with the property that $a < b$ for every $a \in A$ and every $b \in B$, then it follows that $\sup A < \inf B$.

(d) If $\sup A = s$ and $\sup B = t$, then $\sup(A + B) = s + t$. The set $A + B$ is defined as $A + B = \{a + b : a \in A \text{ and } b \in B\}$.

(e) If $\sup A \leq \sup B$, then there exists an element $b \in B$ that is an upper bound for A.

1.4 Consequences of Completeness

The first application of the Axiom of Completeness is a result that may look like a more natural way to mathematically express the sentiment that the real line contains no gaps.

Theorem 1.4.1 (Nested Interval Property). *For each $n \in \mathbf{N}$, assume we are given a closed interval $I_n = [a_n, b_n] = \{x \in \mathbf{R} : a_n \leq x \leq b_n\}$. Assume also that each I_n contains I_{n+1}. Then, the resulting nested sequence of closed intervals*

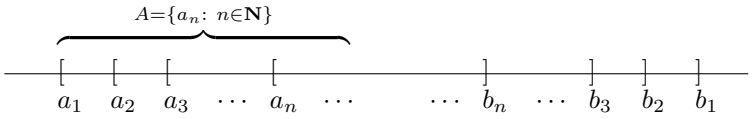
$$I_1 \supseteq I_2 \supseteq I_3 \supseteq I_4 \supseteq \cdots$$

has a nonempty intersection; that is, $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$.

Proof. In order to show that $\bigcap_{n=1}^{\infty} I_n$ is not empty, we are going to use the Axiom of Completeness (AoC) to produce a single real number x satisfying $x \in I_n$ for every $n \in \mathbf{N}$. Now, AoC is a statement about bounded sets, and the one we want to consider is the set

$$A = \{a_n : n \in \mathbf{N}\}$$

of left-hand endpoints of the intervals.



Because the intervals are nested, we see that every b_n serves as an upper bound for A. Thus, we are justified in setting

$$x = \sup A.$$

Now, consider a particular $I_n = [a_n, b_n]$. Because x is an upper bound for A, we have $a_n \leq x$. The fact that each b_n is an upper bound for A and that x is the least upper bound implies $x \leq b_n$.

Altogether then, we have $a_n \leq x \leq b_n$, which means $x \in I_n$ for every choice of $n \in \mathbf{N}$. Hence, $x \in \bigcap_{n=1}^{\infty} I_n$, and the intersection is not empty. □

The Density of Q in R

The set $\mathbf{Q}$ is an extension of $\mathbf{N}$, and $\mathbf{R}$ in turn is an extension of $\mathbf{Q}$. The next few results indicate how $\mathbf{N}$ and $\mathbf{Q}$ sit inside of $\mathbf{R}$.

Theorem 1.4.2 (Archimedean Property). (i) *Given any number $x \in \mathbf{R}$, there exists an $n \in \mathbf{N}$ satisfying $n > x$.*

(ii) *Given any real number $y > 0$, there exists an $n \in \mathbf{N}$ satisfying $1/n < y$.*

Proof. Part (i) of the proposition states that $\mathbf{N}$ is not bounded above. There has never been any doubt about the truth of this, and it could be reasonably argued that we should not have to prove it at all. This is a legitimate point of view, especially in light of the fact that we have decided to assume other familiar properties of $\mathbf{N}$, $\mathbf{Z}$, and $\mathbf{Q}$ as given.

The counterargument is that we will prove it because we can. A set can possess the Archimedean property without being complete—$\mathbf{Q}$ is a fine example—but a demonstration of this fact requires a good deal of scrutiny into the axiomatic construction of the ordered field in question. In the case of $\mathbf{R}$, the Axiom of Completeness furnishes us with a very short argument. A large number of deep results ultimately depend on this relationship between $\mathbf{R}$ and $\mathbf{N}$, so having a proof for it adds a little extra certainty to these upcoming arguments.

And so to the proof. Assume, for contradiction, that $\mathbf{N}$ *is* bounded above. By the Axiom of Completeness (AoC), $\mathbf{N}$ should then have a least upper bound,

and we can set $\alpha = \sup \mathbf{N}$. If we consider $\alpha - 1$, then we no longer have an upper bound (see Lemma 1.3.7), and therefore there exists an $n \in \mathbf{N}$ satisfying $\alpha - 1 < n$. But this is equivalent to $\alpha < n + 1$. Because $n + 1 \in \mathbf{N}$, we have a contradiction to the fact that α is supposed to be an upper bound for $\mathbf{N}$. (Notice that the contradiction here depends only on AoC and the fact that $\mathbf{N}$ is closed under addition.)

Part (ii) follows from (i) by letting $x = 1/y$. □

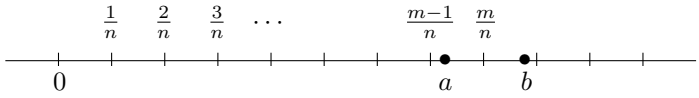
This familiar property of $\mathbf{N}$ is the key to an extremely important fact about how $\mathbf{Q}$ fits inside of $\mathbf{R}$.

Theorem 1.4.3 (Density of Q in R). *For every two real numbers a and b with $a < b$, there exists a rational number r satisfying $a < r < b$.*

Proof. To simplify matters, let's assume $0 \leq a < b$. The case where $a < 0$ follows quickly from this one (Exercise 1.4.1). A rational number is a quotient of integers, so we must produce $m, n \in \mathbf{N}$ so that

$$a < \frac{m}{n} < b. \tag{1}$$

The first step is to choose the denominator n large enough so that consecutive increments of size $1/n$ are too close together to "step over" the interval (a, b).



Using Theorem 1.4.2, we may pick $n \in \mathbf{N}$ large enough so that

$$\frac{1}{n} < b - a. \tag{2}$$

Multiplying inequality (1) by n gives $na < m < nb$. With n already chosen, the idea now is to choose m to be the smallest natural number greater than na. In other words, pick $m \in \mathbf{N}$ so that

$$m - 1 \overset{(3)}{\leq} na \overset{(4)}{<} m.$$

Now, inequality (4) immediately yields $a < m/n$, which is half of the battle. Keeping in mind that inequality (2) is equivalent to $a < b - 1/n$, we can use (3) to write

$$\begin{aligned} m &\leq na + 1 \\ &< n\left(b - \frac{1}{n}\right) + 1 \\ &= nb. \end{aligned}$$

Because $m < nb$ implies $m/n < b$, we have $a < m/n < b$, as desired. □

Theorem 1.4.3 is paraphrased by saying that $\mathbf{Q}$ is *dense* in $\mathbf{R}$. Without working too hard, we can use this result to show that the irrational numbers are dense in $\mathbf{R}$ as well.

Corollary 1.4.4. *Given any two real numbers $a < b$, there exists an irrational number t satisfying $a < t < b$.*

Proof. Exercise 1.4.3. □

The Existence of Square Roots

It is time to tend to some unfinished business left over from Example 1.3.6 and this chapter's opening discussion.

Theorem 1.4.5. *There exists a real number $\alpha \in \mathbf{R}$ satisfying $\alpha^2 = 2$.*

Proof. After reviewing Example 1.3.6, consider the set

$$T = \{t \in \mathbf{R} : t^2 < 2\}$$

and set $\alpha = \sup T$. We are going to prove $\alpha^2 = 2$ by ruling out the possibilities $\alpha^2 < 2$ and $\alpha^2 > 2$. Keep in mind that there are two parts to the definition of $\sup T$, and they will both be important. (This always happens when a supremum is used in an argument.) The strategy is to demonstrate that $\alpha^2 < 2$ violates the fact that α is an upper bound for T, and $\alpha^2 > 2$ violates the fact that it is the least upper bound.

Let's first see what happens if we assume $\alpha^2 < 2$. In search of an element of T that is larger than α, write

$$\begin{aligned}\left(\alpha + \frac{1}{n}\right)^2 &= \alpha^2 + \frac{2\alpha}{n} + \frac{1}{n^2} \\ &< \alpha^2 + \frac{2\alpha}{n} + \frac{1}{n} \\ &= \alpha^2 + \frac{2\alpha + 1}{n}.\end{aligned}$$

But now assuming $\alpha^2 < 2$ gives us a little space in which to fit the $(2\alpha + 1)/n$ term and keep the total less than 2. Specifically, choose $n_0 \in \mathbf{N}$ large enough so that

$$\frac{1}{n_0} < \frac{2 - \alpha^2}{2\alpha + 1}.$$

This implies $(2\alpha + 1)/n_0 < 2 - \alpha^2$, and consequently that

$$\left(\alpha + \frac{1}{n_0}\right)^2 < \alpha^2 + (2 - \alpha^2) = 2.$$

Thus, $\alpha + 1/n_0 \in T$, contradicting the fact that α is an upper bound for T. We conclude that $\alpha^2 < 2$ cannot happen.

Now, what about the case $\alpha^2 > 2$? This time, write

$$\begin{aligned}\left(\alpha - \frac{1}{n}\right)^2 &= \alpha^2 - \frac{2\alpha}{n} + \frac{1}{n^2} \\ &> \alpha^2 - \frac{2\alpha}{n}.\end{aligned}$$

The remainder of the argument is requested in Exercise 1.4.6. □

A small modification of this proof can be used to show that $\sqrt{x}$ exists for any $x \geq 0$. A formula for expanding $(\alpha + 1/n)^m$ called the binomial formula can be used to show that $\sqrt[m]{x}$ exists for arbitrary values of $m \in \mathbf{N}$.

Countable and Uncountable Sets

The applications of the Axiom of Completeness to this point have basically served to restore our confidence in properties we already felt we knew about the real number system. One final consequence of completeness that we are about to present is of a very different nature and, on its own, represents an astounding intellectual discovery. The traditional way that mathematics gets done is by one mathematician modifying and expanding on the work of those who came before. This model does not seem to apply to Georg Cantor (1845–1918), at least with regard to his work on the theory of infinite sets.

At the moment, we have an image of $\mathbf{R}$ as consisting of rational and irrational numbers, continuously packed together along the real line. We have seen that both $\mathbf{Q}$ and $\mathbf{I}$ (the set of irrationals) are dense in $\mathbf{R}$, meaning that in every interval (a, b) there exist rational and irrational numbers alike. Mentally, there is a temptation to think of $\mathbf{Q}$ and $\mathbf{I}$ as being intricately mixed together in equal proportions, but this turns out not to be the case. In a way that Cantor made precise, the irrational numbers far outnumber the rational numbers in making up the real line.

Cardinality

The term *cardinality* is used in mathematics to refer to the size of a set. The cardinalities of finite sets can be compared simply by attaching a natural number to each set. The set of Snow White's dwarfs is smaller than the set of United States Supreme Court Justices because 7 is less than 9. But how might we draw this same conclusion without referring to any numbers? Cantor's idea was to attempt to put the sets into a 1–1 correspondence with each other. There are fewer dwarfs than Justices because, if the dwarfs were all simultaneously appointed to the bench, there would still be two empty chairs to fill. On the other hand, the cardinality of the Supreme Court is the same as the cardinality of the set of fielders on a baseball team. This is because, when the judges take the field, it is possible to arrange them so that there is exactly one judge at every position.

The advantage of this method of comparing the sizes of sets is that it works equally well on sets that are infinite.

Definition 1.4.6. A function $f : A \to B$ is one-to-one (1–1) if $a_1 \neq a_2$ in A implies that $f(a_1) \neq f(a_2)$ in B. The function f is *onto* if, given any $b \in B$, it is possible to find an element $a \in A$ for which $f(a) = b$.

A function $f : A \to B$ that is both 1–1 and onto provides us with exactly what we mean by a 1–1 correspondence between two sets. The property of being 1–1 means that no two elements of A correspond to the same element of B (no two judges are playing the same position), and the property of being onto ensures that every element of B corresponds to something in A (there is a judge at every position).

Definition 1.4.7. Two sets A and B *have the same cardinality* if there exists $f : A \to B$ that is 1–1 and onto. In this case, we write $A \sim B$.

Example 1.4.8. (i) If we let $E = \{2, 4, 6, \dots\}$ be the set of even natural numbers, then we can show $\mathbf{N} \sim E$. To see why, let $f : \mathbf{N} \to E$ be given by $f(n) = 2n$.

$$\begin{array}{lccccccc} \mathbf{N}: & 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \cdots & \updownarrow & \\ E: & 2 & 4 & 6 & 8 & \cdots & 2n & \cdots \end{array}$$

It is certainly true that E is a proper subset of $\mathbf{N}$, and for this reason it may seem logical to say that E is a "smaller" set than $\mathbf{N}$. This is one way to look at it, but it represents a point of view that is heavily biased from an overexposure to finite sets. The definition of cardinality is quite specific, and from this point of view E and $\mathbf{N}$ are equivalent.

(ii) To make this point again, note that although $\mathbf{N}$ is contained in $\mathbf{Z}$ as a proper subset, we can show $\mathbf{N} \sim \mathbf{Z}$. This time let

$$f(n) = \begin{cases} (n-1)/2 & \text{if } n \text{ is odd} \\ -n/2 & \text{if } n \text{ is even.} \end{cases}$$

The important details to verify are that f does not map any two natural numbers to the same element of $\mathbf{Z}$ (f is 1–1) and that every element of $\mathbf{Z}$ gets "hit" by something in $\mathbf{N}$ (f is onto).

$$\begin{array}{lcccccccc} \mathbf{N}: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \cdots \\ & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\ \mathbf{Z}: & 0 & -1 & 1 & -2 & 2 & -3 & 3 & \cdots \end{array}$$

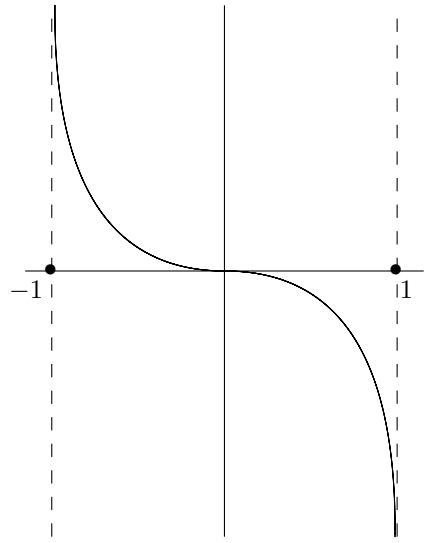


Figure 1.4: $(-1,1) \sim \mathbf{R}$ USING $f(x) = x/(x^2-1)$.

Example 1.4.9. A little calculus (which we will not supply) shows that the function $f(x) = x/(x^2-1)$ takes the interval $(-1,1)$ onto $\mathbf{R}$ in a 1–1 fashion (Fig. 1.4). Thus $(-1,1) \sim \mathbf{R}$. In fact, $(a,b) \sim \mathbf{R}$ for any interval (a,b).

Countable Sets

Definition 1.4.10. A set A is *countable* if $\mathbf{N} \sim A$. An infinite set that is not countable is called an *uncountable* set.

From Example 1.4.8, we see that both E and $\mathbf{Z}$ are countable sets. Putting a set into a 1–1 correspondence with $\mathbf{N}$, in effect, means putting all of the elements into an infinitely long list or sequence. Looking at Example 1.4.8, we can see that this was quite easy to do for E and required only a modest bit of shuffling for the set $\mathbf{Z}$. A natural question arises as to whether *all* infinite sets are countable. Given some infinite set such as $\mathbf{Q}$ or $\mathbf{R}$, it might seem as though, with enough cleverness, we should be able to fit all the elements of our set into a single list (i.e., into a correspondence with $\mathbf{N}$). After all, this list is infinitely long so there should be plenty of room. But alas, as Hardy remarks, "[The mathematician's] subject is the most curious of all—there is none in which truth plays such odd pranks."

Theorem 1.4.11. (i) *The set* $\mathbf{Q}$ *is countable.* (ii) *The set* $\mathbf{R}$ *is uncountable.*

Proof. (i) For each $n \in \mathbf{N}$, let A_n be the set given by

$$A_n = \left\{ \pm\frac{p}{q} : \text{ where } p, q \in \mathbf{N} \text{ are in lowest terms with } p+q = n \right\}.$$

The first few of these sets look like

$$A_1 = \left\{\frac{0}{1}\right\}, \quad A_2 = \left\{\frac{1}{1}, \frac{-1}{1}\right\}, \quad A_3 = \left\{\frac{1}{2}, \frac{-1}{2}, \frac{2}{1}, \frac{-2}{1}\right\},$$

$$A_4 = \left\{\frac{1}{3}, \frac{-1}{3}, \frac{3}{1}, \frac{-3}{1}\right\}, \quad \text{and} \quad A_5 = \left\{\frac{1}{4}, \frac{-1}{4}, \frac{2}{3}, \frac{-2}{3}, \frac{3}{2}, \frac{-3}{2}, \frac{4}{1}, \frac{-4}{1}\right\}.$$

The crucial observation is that each A_n is *finite* and every rational number appears in exactly one of these sets. Our 1–1 correspondence with $\mathbf{N}$ is then achieved by consecutively listing the elements in each A_n.

$$\begin{array}{lccccccccccccc}
\mathbf{N}: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & \cdots \\
 & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\
\mathbf{Q}: & \frac{0}{1} & \frac{1}{1} & -\frac{1}{1} & \frac{1}{2} & -\frac{1}{2} & \frac{2}{1} & -\frac{2}{1} & \frac{1}{3} & -\frac{1}{3} & \frac{3}{1} & -\frac{3}{1} & \frac{1}{4} & \cdots \\
 & \underbrace{}_{A_1} & \underbrace{}_{A_2} & & \underbrace{}_{A_3} & & & & \underbrace{}_{A_4} & & & & &
\end{array}$$

Admittedly, writing an explicit formula for this correspondence would be an awkward task, and attempting to do so is not the best use of time. What matters is that we see why every rational number appears in the correspondence exactly once. Given, say, $22/7$, we have that $22/7 \in A_{29}$. Because the set of elements in $A_1, \ldots, A_{28}$ is finite, we can be confident that $22/7$ eventually gets included in the sequence. The fact that this line of reasoning applies to any rational number p/q is our proof that the correspondence is onto. To verify that it is 1–1, we observe that the sets A_n were constructed to be disjoint so that no rational number appears twice. This completes the proof of (i).

(ii) The second statement of Theorem 1.4.11 is the truly unexpected part, and its proof is done by contradiction. Assume that there *does* exist a 1–1, onto function $f : \mathbf{N} \to \mathbf{R}$. Again, what this suggests is that it is possible to enumerate the elements of $\mathbf{R}$. If we let $x_1 = f(1)$, $x_2 = f(2)$, and so on, then our assumption that f is onto means that we can write

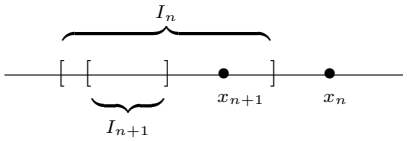
$$\mathbf{R} = \{x_1, x_2, x_3, x_4, \ldots\} \tag{1}$$

and be confident that every real number appears somewhere on the list. We will now use the Nested Interval Property (Theorem 1.4.1) to produce a real number that is not there.

Let I_1 be a closed interval that *does not* contain x_1. Next, let I_2 be a closed interval, contained in I_1, which does not contain x_2. The existence of such an I_2 is easy to verify. Certainly I_1 contains two smaller *disjoint* closed intervals, and x_2 can only be in one of these. In general, given an interval I_n, construct I_{n+1} to satisfy

(i) $I_{n+1} \subseteq I_n$ and

(ii) $x_{n+1} \notin I_{n+1}$.



We now consider the intersection $\bigcap_{n=1}^{\infty} I_n$. If x_{n_0} is some real number from the list in (1), then we have $x_{n_0} \notin I_{n_0}$, and it follows that

$$x_{n_0} \notin \bigcap_{n=1}^{\infty} I_n.$$

Now, we are assuming that the list in (1) contains every real number, and this leads to the conclusion that

$$\bigcap_{n=1}^{\infty} I_n = \emptyset.$$

However, the Nested Interval Property (NIP) asserts that $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$. By NIP, there is at least one $x \in \bigcap_{n=1}^{\infty} I_n$ that, consequently, cannot be on the list in (1). This contradiction means that such an enumeration of $\mathbf{R}$ is impossible, and we conclude that $\mathbf{R}$ is an *uncountable* set. □

What exactly should we make of this discovery? It is an important exercise to show that any subset of a countable set must be either countable or finite. This should not be too surprising. If a set can be arranged into a single list, then deleting some elements from this list results in another (shorter, and potentially terminating) list. This means that countable sets are the smallest type of infinite set. Anything smaller is either still countable or finite.

The force of Theorem 1.4.11 is that the cardinality of $\mathbf{R}$ is, informally speaking, a larger type of infinity. The real numbers so outnumber the natural numbers that there is no way to map $\mathbf{N}$ onto $\mathbf{R}$. No matter how we attempt this, there are always real numbers to spare. The set $\mathbf{Q}$, on the other hand, is countable. As far as infinite sets are concerned, this is as small as it gets. What does this imply about the set $\mathbf{I}$ of irrational numbers? By imitating the demonstration that $\mathbf{N} \sim \mathbf{Z}$, we can prove that the union of two countable sets must be countable. Because $\mathbf{R} = \mathbf{Q} \cup \mathbf{I}$, it follows that $\mathbf{I}$ cannot be countable because otherwise $\mathbf{R}$ would be. The inescapable conclusion is that, despite the fact that we have encountered so few of them, the irrational numbers form a far greater subset of $\mathbf{R}$ than $\mathbf{Q}$.

The properties of countable sets described in this discussion are useful for a few exercises in upcoming chapters. For easier reference, we state them as some final propositions and outline their proofs in the exercises that follow.

Theorem 1.4.12. *If $A \subseteq B$ and B is countable, then A is either countable, finite, or empty.*

Theorem 1.4.13. (i) *If $A_1, A_2, \ldots A_m$ are each countable sets, then the union $A_1 \cup A_2 \cup \cdots \cup A_m$ is countable.*

(ii) *If A_n is a countable set for each $n \in \mathbf{N}$, then $\bigcup_{n=1}^{\infty} A_n$ is countable.*

Exercises

Exercise 1.4.1. Without doing too much work, show how to prove Theorem 1.4.3 in the case where $a < 0$ by converting this case into the one already proven.

Exercise 1.4.2. Recall that $\mathbf{I}$ stands for the set of irrational numbers.

(a) Show that if $a, b \in \mathbf{Q}$, then ab and $a + b$ are elements of $\mathbf{Q}$ as well.

(b) Show that if $a \in \mathbf{Q}$ and $t \in \mathbf{I}$, then $a + t \in \mathbf{I}$ and $at \in \mathbf{I}$ as long as $a \neq 0$.

(c) Part (a) can be summarized by saying that $\mathbf{Q}$ is closed under addition and multiplication. Is $\mathbf{I}$ closed under addition and multiplication? Given two irrational numbers s and t, what can we say about $s + t$ and st?

Exercise 1.4.3. Using Exercise 1.4.2, supply a proof for Corollary 1.4.4 by applying Theorem 1.4.3 to the real numbers $a - \sqrt{2}$ and $b - \sqrt{2}$.

Exercise 1.4.4. Use the Archimedean Property of $\mathbf{R}$ to rigorously prove that $\inf\{1/n : n \in \mathbf{N}\} = 0$.

Exercise 1.4.5. Prove that $\bigcap_{n=1}^{\infty}(0, 1/n) = \emptyset$. Notice that this demonstrates that the intervals in the Nested Interval Property must be closed for the conclusion of the theorem to hold.

Exercise 1.4.6. (a) Finish the proof of Theorem 1.4.5 by showing that the assumption $\alpha^2 > 2$ leads to a contradiction of the fact that $\alpha = \sup T$.

(b) Modify this argument to prove the existence of $\sqrt{b}$ for any real number $b \geq 0$.

Exercise 1.4.7. Finish the following proof for Theorem 1.4.12.

Assume B is a countable set. Thus, there exists $f : \mathbf{N} \to B$, which is 1–1 and onto. Let $A \subseteq B$ be an infinite subset of B. We must show that A is countable.

Let $n_1 = \min\{n \in \mathbf{N} : f(n) \in A\}$. As a start to a definition of $g : \mathbf{N} \to A$, set $g(1) = f(n_1)$. Show how to inductively continue this process to produce a 1–1 function g from $\mathbf{N}$ onto A.

Exercise 1.4.8. Use the following outline to supply proofs for the statements in Theorem 1.4.13.

(a) First, prove statement (i) for two countable sets, A_1 and A_2. Example 1.4.8 (ii) may be a useful reference. Some technicalities can be avoided by first replacing A_2 with the set $B_2 = A_2 \backslash A_1 = \{x \in A_2 : x \notin A_1\}$. The point of this is that the union $A_1 \cup B_2$ is equal to $A_1 \cup A_2$ and the sets A_1 and B_2 are disjoint. (What happens if B_2 is finite?)

Now, explain how the more general statement in (i) follows.

(b) Explain why induction *cannot* be used to prove part (ii) of Theorem 1.4.13 from part (i).

(c) Show how arranging $\mathbf{N}$ into the two-dimensional array

$$\begin{array}{cccccc} 1 & 3 & 6 & 10 & 15 & \cdots \\ 2 & 5 & 9 & 14 & \cdots & \\ 4 & 8 & 13 & \cdots & & \\ 7 & 12 & \cdots & & & \\ 11 & \cdots & & & & \\ \vdots & & & & & \end{array}$$

leads to a proof of Theorem 1.4.13 (ii).

Exercise 1.4.9. (a) Given sets A and B, explain why $A \sim B$ is equivalent to asserting $B \sim A$.

(b) For three sets A, B, and C, show that $A \sim B$ and $B \sim C$ implies $A \sim C$. These two properties are what is meant by saying that $\sim$ is an *equivalence relation*.

Exercise 1.4.10. Show that the set of all finite subsets of $\mathbf{N}$ is a countable set. (It turns out that the set of *all* subsets of $\mathbf{N}$ is not a countable set. This is the topic of Section 1.5.)

Exercise 1.4.11. Consider the open interval (0,1), and let S be the set of points in the open unit square; that is, $S = \{(x, y) : 0 < x, y < 1\}$.

(a) Find a 1–1 function that maps $(0, 1)$ into, but not necessarily onto, S. (This is easy.)

(b) Use the fact that every real number has a decimal expansion to produce a 1–1 function that maps S into $(0, 1)$. Discuss whether the formulated function is onto. (Keep in mind that any terminating decimal expansion such as .235 represents the same real number as $.234999\ldots$.)

The Schröder–Bernstein Theorem discussed in Exercise 1.4.13 to follow can now be applied to conclude that $(0, 1) \sim S$.

Exercise 1.4.12. A real number $x \in \mathbf{R}$ is called *algebraic* if there exist integers $a_0, a_1, a_2, \ldots, a_n \in \mathbf{Z}$, not all zero, such that

$$a_n x^n + a_{n_1} x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

Said another way, a real number is algebraic if it is the root of a polynomial with integer coefficients. Real numbers that are not algebraic are called *transcendental* numbers. Reread the last paragraph of Section 1.1. The final question posed here is closely related to the question of whether or not transcendental numbers exist.

(a) Show that $\sqrt{2}$, $\sqrt[3]{2}$, and $\sqrt{3} + \sqrt{2}$ are algebraic.

(b) Fix $n \in \mathbf{N}$, and let A_n be the algebraic numbers obtained as roots of polynomials with integer coefficients that have degree n. Using the fact that every polynomial has a finite number of roots, show that A_n is countable. (For each $m \in \mathbf{N}$, consider polynomials $a_n x^n + a_{n_1} x^{n-1} + \cdots + a_1 x + a_0$ that satisfy $|a_n| + |a_{n-1}| + \cdots + |a_1| + |a_0| \leq m$.)

(c) Now, argue that the set of all algebraic numbers is countable. What may we conclude about the set of transcendental numbers?

Exercise 1.4.13 (Schröder–Bernstein Theorem). Assume there exists a 1–1 function $f : X \to Y$ and another 1–1 function $g : Y \to X$. Follow the steps to show that there exists a 1–1, onto function $h : X \to Y$ and hence $X \sim Y$.

(a) The range of f is defined by $f(X) = \{y \in Y : y = f(x) \text{ for some } x \in X\}$. Let $y \in f(X)$. (Because f is not necessarily onto, the range $f(X)$ may not be all of Y.) Explain why there exists a *unique* $x \in X$ such that $f(x) = y$. Now define $f^{-1}(y) = x$, and show that f^{-1} is a 1–1 function from $f(X)$ onto X.

In a similar way, we can also define the 1–1 function $g^{-1} : g(X) \to Y$.

(b) Let $x \in X$ be arbitrary. Let the *chain* C_x be the set consisting of all elements of the form

$$(1) \qquad \ldots, f^{-1}(g^{-1}(x)), g^{-1}(x), x, f(x), g(f(x)), f(g(f(x))), \ldots .$$

Explain why the number of elements to the left of x in the above chain may be zero, finite, or infinite.

(c) Show that any two chains are either identical or completely disjoint.

(d) Note that the terms of the chain in (1) alternate between elements of X and elements of Y. Given a chain C_x, we want to focus on $C_x \cap Y$, which is just the part of the chain that sits in Y.

Define the set A to be the union of all chains C_x satisfying $C_x \cap Y \subseteq f(X)$. Let B consist of the union of the remaining chains not in A. Show that any chain contained in B must be of the form

$$y, g(y), f(g(y)), g(f(g(y))), \ldots ,$$

where y is an element of Y that is *not* in $f(X)$.

(e) Let $X_1 = A \cap X$, $X_2 = B \cap X$, $Y_1 = A \cap Y$, and $Y_2 = B \cap Y$. Show that f maps X_1 *onto* Y_1 and that g maps Y_2 onto X_2. Use this information to prove $X \sim Y$.

1.5 Cantor's Theorem

Cantor's work into the theory of infinite sets extends far beyond the conclusions of Theorem 1.4.11. Although initially resisted, his creative and relentless assault in this area eventually produced a revolution in set theory and a paradigm shift in the way mathematicians came to understand the infinite.

Cantor's Diagonalization Method

The proof presented for Theorem 1.4.11 (ii) is different from any of the arguments that Cantor gave for this result. It was chosen because of how directly it reveals the connection between the concepts of uncountability and completeness, and because the technique of using nested intervals will be used several more times in our work ahead.

Cantor initially published his discovery that $\mathbf{R}$ is uncountable in 1874, but in 1891 he offered another proof of this same fact that is startling in its simplicity. It relies on decimal representations for real numbers, which we will accept and use without any formal definitions.

Theorem 1.5.1. *The open interval* $(0,1) = \{x \in \mathbf{R} : 0 < x < 1\}$ *is uncountable.*

Exercise 1.5.1. Show that $(0,1)$ is uncountable if and only if $\mathbf{R}$ is uncountable. This shows that Theorem 1.5.1 is equivalent to Theorem 1.4.11.

Proof. As with Theorem 1.4.11, we proceed by contradiction and assume that there does exist a function $f : \mathbf{N} \to (0,1)$ that is 1–1 and onto. For each $m \in \mathbf{N}$, $f(m)$ is a real number between 0 and 1, and we represent it using the decimal notation

$$f(m) = .a_{m1}a_{m2}a_{m3}a_{m4}a_{m5}\dots .$$

What is meant here is that for each $m, n \in \mathbf{N}$, a_{mn} is the digit from the set $\{0, 1, 2, \dots, 9\}$ that represents the nth digit in the decimal expansion of $f(m)$. The 1–1 correspondence between $\mathbf{N}$ and $(0,1)$ can be summarized in the doubly indexed array

$\mathbf{N}$		$(0,1)$								
1	$\longleftrightarrow$	$f(1)$	$=$	$.\boldsymbol{a_{11}}$	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	$\cdots$
2	$\longleftrightarrow$	$f(2)$	$=$	$.a_{21}$	$\boldsymbol{a_{22}}$	a_{23}	a_{24}	a_{25}	a_{26}	$\cdots$
3	$\longleftrightarrow$	$f(3)$	$=$	$.a_{31}$	a_{32}	$\boldsymbol{a_{33}}$	a_{34}	a_{35}	a_{36}	$\cdots$
4	$\longleftrightarrow$	$f(4)$	$=$	$.a_{41}$	a_{42}	a_{43}	$\boldsymbol{a_{44}}$	a_{45}	a_{46}	$\cdots$
5	$\longleftrightarrow$	$f(5)$	$=$	$.a_{51}$	a_{52}	a_{53}	a_{54}	$\boldsymbol{a_{55}}$	a_{56}	$\cdots$
6	$\longleftrightarrow$	$f(6)$	$=$	$.a_{61}$	a_{62}	a_{63}	a_{64}	a_{65}	$\boldsymbol{a_{66}}$	$\cdots$
$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

The key assumption about this correspondence is that *every* real number in $(0,1)$ is assumed to appear somewhere on the list.

Now for the pearl of the argument. Define a real number $x \in (0,1)$ with the decimal expansion $x = .b_1b_2b_3b_4\dots$ using the rule

$$b_n = \begin{cases} 2 & \text{if } a_{nn} \neq 2 \\ 3 & \text{if } a_{nn} = 2. \end{cases}$$

Let's be clear about this. To compute the digit b_1, we look at the digit a_{11} in the upper left-hand corner of the array. If $a_{11} = 2$, then we choose $b_1 = 3$; otherwise, we set $b_1 = 2$.

Exercise 1.5.2. (a) Explain why the real number $x = .b_1b_2b_3b_4\dots$ cannot be $f(1)$.

(b) Now, explain why $x \neq f(2)$, and in general why $x \neq f(n)$ for any $n \in \mathbf{N}$.

(c) Point out the contradiction that arises from these observations and conclude that $(0, 1)$ is uncountable.

□

Exercise 1.5.3. Supply rebuttals to the following complaints about the proof of Theorem 1.5.1.

(a) Every rational numbers has a decimal expansion so we could apply this same argument to show that the set of rational numbers between 0 and 1 is uncountable. However, because we know that any subset of $\mathbf{Q}$ must be countable, the proof of Theorem 1.5.1 must be flawed.

(b) A few numbers have *two* different decimal representations. Specifically, any decimal expansion that terminates can also be written with repeating 9's. For instance, $1/2$ can be written as .5 or as $.4999\ldots$. Doesn't this cause some problems?

Exercise 1.5.4. Let S be the set consisting of all sequences of 0's and 1's. Observe that S is not a particular sequence, but rather a large set whose elements are sequences; namely,

$$S = \{(a_1, a_2, a_3, \ldots) : a_n = 0 \text{ or } 1\}.$$

As an example, the sequence $(1, 0, 1, 0, 1, 0, 1, 0, \ldots)$ is an element of S, as is the sequence $(1, 1, 1, 1, 1, 1, \ldots)$.

Give a rigorous argument showing that S is uncountable.

Having distinguished between the countable infinity of $\mathbf{N}$ and the uncountable infinity of $\mathbf{R}$, a new question that occupied Cantor was whether or not there existed an infinity "above" that of $\mathbf{R}$. This is logically treacherous territory. The same care we gave to defining the relationship "has the same cardinality as" needs to be given to defining relationships such as "has cardinality greater than" or "has cardinality less than or equal to." Nevertheless, without getting too weighed down with formal definitions, one gets a very clear sense from the next result that there is a hierarchy of infinite sets that continues well beyond the continuum of $\mathbf{R}$.

Power Sets and Cantor's Theorem

Given a set A, the *power set* $P(A)$ refers to the collection of all subsets of A. It is important to understand that $P(A)$ is itself considered a set whose elements are the different possible subsets of A.

Exercise 1.5.5. (a) Let $A = \{a, b, c\}$. List the eight elements of $P(A)$. (Do not forget that $\emptyset$ is considered to be a subset of every set.)

(b) If A is finite with n elements, show that $P(A)$ has 2^n elements. (Constructing a particular subset of A can be interpreted as making a series of decisions about whether or not to include each element of A.)

Exercise 1.5.6. (a) Using the particular set $A = \{a, b, c\}$, exhibit two different 1–1 mappings from A into $P(A)$.

(b) Letting $B = \{1, 2, 3, 4\}$, produce an example of a 1–1 map $g : B \to P(B)$.

(c) Explain why, in parts (a) and (b), it is impossible to construct mappings that are *onto*.

Cantor's Theorem states that the phenomenon in Exercise 1.5.6 holds for infinite sets as well as finite sets. Whereas mapping A *into* $P(A)$ is quite effortless, finding an *onto* map is impossible.

Theorem 1.5.2 (Cantor's Theorem). *Given any set A, there does not exist a function $f : A \to P(A)$ that is onto.*

Proof. This proof, like the others of its kind, is indirect. Thus, assume, for contradiction, that $f : A \to P(A)$ is onto. Unlike the usual situation in which we have sets of numbers for the domain and range, f is a correspondence between a set and its power set. For each element $a \in A$, $f(a)$ is a particular *subset* of A. The assumption that f is onto means that every subset of A appears as $f(a)$ for some $a \in A$. To arrive at a contradiction, we will produce a subset $B \subseteq A$ that is not equal to $f(a)$ for any $a \in A$.

Construct B using the following rule. For each element $a \in A$, consider the subset $f(a)$. This subset of A may contain the element a or it may not. This depends on the function f. If $f(a)$ does not contain a, then we include a in our set B. More precisely, let

$$B = \{a \in A : a \notin f(a)\}.$$

Exercise 1.5.7. Return to the particular functions contructed in Exercise 1.5.6 and construct the subset B that results using the preceding rule. In each case, note that B is not in the range of the function used.

We now focus on the general argument. Because we have assumed that our function $f : A \to P(A)$ is onto, it must be that $B = f(a')$ for some $a' \in A$. The contradiction arises when we consider whether or not a' is an element of B.

Exercise 1.5.8. (a) First, show that the case $a' \in B$ leads to a contradiction.

(b) Now, finish the argument by showing that the case $a' \notin B$ is equally unacceptable.

□

Exercise 1.5.9. As a final exercise, answer each of the following by establishing a 1–1 correspondence with a set of known cardinality.

(a) Is the set of all functions from $\{0, 1\}$ to $\mathbf{N}$ countable or uncountable?

(b) Is the set of all functions from $\mathbf{N}$ to $\{0, 1\}$ countable or uncountable?

(c) Given a set B, a subset $\mathcal{A}$ of $P(B)$ is called an *antichain* if no element of $\mathcal{A}$ is a subset of any other element of $\mathcal{A}$. Does $P(\mathbf{N})$ contain an uncountable antichain?

1.6 Epilogue

The relationship of having the same cardinality is an *equivalence relation* (see Exercise 1.4.9), meaning, roughly, that all of the sets in the universe can be organized into disjoint groups according to their size. Two sets appear in the same group, or *equivalence class*, if and only if they have the same cardinality. Thus, $\mathbf{N}$, $\mathbf{Z}$, and $\mathbf{Q}$ are grouped together in one class with all of the other countable sets, whereas $\mathbf{R}$ is in another class that includes the interval $(0, 1)$ among other uncountable sets. One implication of Cantor's Theorem is that $P(\mathbf{R})$—the set of all subsets of $\mathbf{R}$—is in a different class from $\mathbf{R}$, and there is no reason to stop here. The set of subsets of $P(\mathbf{R})$—namely $P(P(\mathbf{R}))$—is in yet another class, and this process continues indefinitely.

Having divided the universe of sets into disjoint groups, it would be convenient to attach a "number" to each collection which could be used the way natural numbers are used to refer to the sizes of finite sets. Given a set X, there exists something called the *cardinal number* of X, denoted $\operatorname{card} X$, which behaves very much in this fashion. For instance, two sets X and Y satisfy $\operatorname{card} X = \operatorname{card} Y$ if and only if $X \sim Y$. (Rigorously defining $\operatorname{card} X$ requires some significant set theory. One way this is done is to define $\operatorname{card} X$ to be a very particular set that can always be uniquely found in the same equivalence class as X.)

Looking back at Cantor's Theorem, we get the strong sense that there is an *order* on the sizes of infinite sets that should be reflected in our new cardinal number system. Specifically, if it is possible to map a set X *into* Y in a 1–1 fashion, then we want $\operatorname{card} X \leq \operatorname{card} Y$. Writing the strict inequality $\operatorname{card} X < \operatorname{card} Y$ should indicate that it is possible to map X into Y but that it is impossible to show $X \sim Y$. Restated in this notation, Cantor's Theorem states that for every set A, $\operatorname{card} A < \operatorname{card} P(A)$.

There are some significant details to work out. A kind of metaphysical problem arises when we realize that an implication of Cantor's Theorem is that there can be no "largest" set. A declaration such as, "Let U be the set of all possible things," is paradoxical because we immediately get that $\operatorname{card} U < \operatorname{card} P(U)$ and thus the set U does not contain everything it was advertised to hold. Issues such as this one are ultimately resolved by imposing some restrictions on what can qualify as a set. As set theory was formalized, the axioms had to be crafted so that objects such as U are simply not allowed. A more down-to-earth problem in need of attention is demonstrating that our definition of "$\leq$" between cardinal numbers really is an ordering. This involves showing that cardinal numbers possess a property analogous to real numbers, which states that if $\operatorname{card} X \leq \operatorname{card} Y$ and $\operatorname{card} Y \leq \operatorname{card} X$, then $\operatorname{card} X = \operatorname{card} Y$. In the end, this boils down to proving that if there exists $f : X \to Y$ that is 1–1, and if there exists $g : Y \to X$ that is 1–1, then it is possible to find a function $h : X \to Y$ that is both 1–1 and onto. A proof of this fact eluded Cantor but was eventually supplied independently by Ernst Schröder (in 1896) and Felix Bernstein (in 1898). An argument for the Schröder–Bernstein Theorem is outlined in Exercise 1.4.13.

There was another deep problem stemming from the budding theory of cardinal numbers that occupied Cantor and which was not resolved during his lifetime. Because of the importance of countable sets, the symbol $\aleph_0$ ("aleph naught") is frequently used for $\operatorname{card} \mathbf{N}$. The subscript "0" is appropriate when we remember that countable sets are the smallest type of infinite set. In terms of cardinal numbers, if $\operatorname{card} X < \aleph_0$, then X is finite. Thus, $\aleph_0$ is the smallest infinite cardinal number. The cardinality of $\mathbf{R}$ is also significant enough to deserve the special designation $\boldsymbol{c} = \operatorname{card} \mathbf{R} = \operatorname{card}(0, 1)$. The content of Theorems 1.4.11 and 1.5.1 is that $\aleph_0 < \boldsymbol{c}$. The question that plagued Cantor was whether there were any cardinal numbers strictly in between these two. Put another way, does there exist a set $A \subseteq \mathbf{R}$ with $\operatorname{card} \mathbf{N} < \operatorname{card} A < \operatorname{card} \mathbf{R}$? Cantor was of the opinion that no such set existed. In the ordering of cardinal numbers, he conjectured, $\boldsymbol{c}$ was the immediate successor of $\aleph_0$.

Cantor's "continuum hypothesis," as it came to be called, was one of the most famous mathematical challenges of the past century. Its unexpected resolution came in two parts. In 1940, the German logician and mathematician Kurt Gödel demonstrated that, using only the agreed-upon set of axioms of set theory, there was no way to disprove the continuum hypothesis. In 1963, Paul Cohen successfully showed that, under the same rules, it was also impossible to prove this conjecture. Taken together, what these two discoveries imply is that the continuum hypothesis is undecidable. It can be accepted or rejected as a statement about the nature of infinite sets, and in neither case will any logical contradictions arise.

The mention of Kurt Gödel brings to mind a final comment about the significance of Cantor's work. Gödel is best known for his "Incompleteness Theorems," which pertain to the strength of axiomatic systems in general. What Gödel showed was that any consistent axiomatic system created to study arithmetic was necessarily destined to be "incomplete" in the sense that there would always be true statements that the system of axioms would be too weak to prove. At the heart of Gödel's very complicated proof is a type of manipulation closely related to what is happening in the proofs of Theorems 1.5.1 and 1.5.2. Variations of Cantor's proof methods can also be found in the limitative results of computer science. The "halting problem" asks, loosely, whether some general algorithm exists that can look at every program and decide if that program eventually terminates. The proof that no such algorithm exists uses a diagonalization-type construction at the core of the argument. The main point to make is that not only are the implications of Cantor's theorems profound but the argumentative techniques are as well. As a more immediate example of this phenomenon, the diagonalization method is used again in Chapter 6—in a constructive way—as a crucial step in the proof of the Arzela–Ascoli Theorem.

Chapter 2

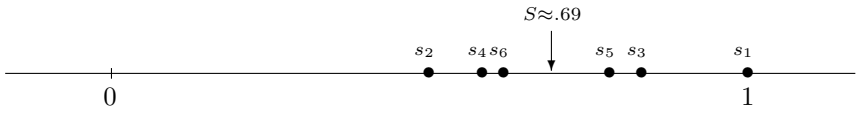
Sequences and Series

2.1 Discussion: Rearrangements of Infinite Series

Consider the infinite series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots .$$

If we naively begin adding from the left-hand side, we get a sequence of what are called *partial sums*. In other words, let s_n equal the sum of the first n terms of the series, so that $s_1 = 1$, $s_2 = 1/2$, $s_3 = 5/6$, $s_4 = 7/12$, and so on. One immediate observation is that the successive sums oscillate in a progressively narrower space. The odd sums decrease ($s_1 > s_3 > s_5 > \dots$) while the even sums increase ($s_2 < s_4 < s_6 < \dots$).



$$s_2 < s_4 < s_6 < \cdots S \cdots < s_5 < s_3 < s_1$$

It seems reasonable—and we will soon prove—that the sequence (s_n) eventually hones in on a value, call it S, where the odd and even partial sums "meet". At this moment, we cannot compute S precisely, but we know it falls somewhere between 7/12 and 5/6. Summing a few hundred terms reveals that $S \approx .69$. Whatever its value, there is now an overwhelming temptation to write

$$S = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots \tag{1}$$

meaning, perhaps, that if we could indeed add up *all* infinitely many of these numbers, then the sum would *equal* S. A more familiar example of an equation

of this type might be

$$2 = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \cdots,$$

the only difference being that in the second equation we have a more recognizable value for the sum.

But now for the crux of the matter. The symbols $+$, $-$, and $=$ in the preceding equations are deceptively familiar notions being used in a very unfamiliar way. The crucial question is whether or not properties of addition and equality that are well understood for finite sums remain valid when applied to infinite objects such as equation (1). The answer, as we are about to witness, is somewhat ambiguous.

Treating equation (1) in a standard algebraic way, let's multiply through by $1/2$ and add it back to equation (1):

$$\begin{array}{rl} \frac{1}{2}S = & \quad \frac{1}{2} \quad - \frac{1}{4} \quad + \frac{1}{6} \quad - \frac{1}{8} \quad + \frac{1}{10} \quad - \frac{1}{12} \quad + \cdots \\ + \; S = & \; 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \frac{1}{11} - \frac{1}{12} + \frac{1}{13} - \cdots \\ \hline \end{array}$$

$$(2) \qquad \frac{3}{2}S = 1 \quad + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} \quad + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} \quad + \frac{1}{11} \; - \frac{1}{6} + \frac{1}{13} \quad \cdots$$

Now, look carefully at the result. The sum in equation (2) consists *precisely* of the same terms as those in the original equation (1), only in a different order. Specifically, the series in (2) is a rearrangement of (1) where we list the first two positive terms $(1 + \frac{1}{3})$ followed by the first negative term $(-\frac{1}{2})$, followed by the next two positive terms $(\frac{1}{5} + \frac{1}{7})$ and then the next negative term $(-\frac{1}{4})$. Continuing this, it is apparent that every term in (2) appears in (1) and vice versa. The rub comes when we realize that equation (2) asserts that the sum of these rearranged, but otherwise unaltered, numbers is equal to $3/2$ its original value. Indeed, adding a few hundred terms of equation (2) produces partial sums in the neighborhood of 1.03. Addition, in this infinite setting, is not commutative!

Let's look at a similar rearrangement of the series

$$\sum_{n=0}^{\infty} (-1/2)^n.$$

This series is geometric with first term 1 and common ratio $r = -1/2$. Using the formula $1/(1-r)$ for the sum of a geometric series (Example 2.7.5), we get

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \frac{1}{32} + \frac{1}{64} - \frac{1}{128} + \frac{1}{256} \cdots = \frac{1}{1 - (-\frac{1}{2})} = \frac{2}{3}.$$

This time, some computational experimentation with the "two positives, one negative" rearrangement

$$1 + \frac{1}{4} - \frac{1}{2} + \frac{1}{16} + \frac{1}{64} - \frac{1}{8} + \frac{1}{256} + \frac{1}{1024} - \frac{1}{32} \cdots$$

yields partial sums quite close to 2/3. The sum of the first 30 terms, for instance, equals .666667. Infinite addition is commutative in some instances but not in others.

Far from being a charming theoretical oddity of infinite series, this phenomenon can be the source of great consternation in many applied situations. How, for instance, should a double summation over two index variables be defined? Let's say we are given a *grid* of real numbers $\{a_{ij} : i, j \in \mathbf{N}\}$, where $a_{ij} = 1/2^{j-i}$ if $j > i$, $a_{ij} = -1$ if $j = i$, and $a_{ij} = 0$ if $j < i$.

$$\begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \cdots \\ 0 & -1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \cdots \\ 0 & 0 & -1 & \frac{1}{2} & \frac{1}{4} & \cdots \\ 0 & 0 & 0 & -1 & \frac{1}{2} & \cdots \\ 0 & 0 & 0 & 0 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

We would like to attach a mathematical meaning to the summation

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

whereby we intend to include every term in the preceding array in the total. One natural idea is to temporarily fix i and sum across each row. A moment's reflection (and a fact about geometric series) shows that each row sums to 0. Summing the sums of the rows, we get

$$\sum_{i,j=1}^{\infty} a_{ij} = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} a_{ij} \right) = \sum_{i=1}^{\infty} (0) = 0.$$

We could just as easily have decided to fix j and sum down each column first. In this case, we have

$$\sum_{i,j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \left(\sum_{i=1}^{\infty} a_{ij} \right) = \sum_{j=1}^{\infty} \left(\frac{-1}{2^{j-1}} \right) = -2.$$

Changing the order of the summation changes the value of the sum. One common way that double sums arise (although not this particular one) is from the multiplication of two series. There is a natural desire to write

$$\left(\sum a_i \right) \left(\sum b_j \right) = \sum_{i,j} a_i b_j,$$

except that the expression on the right-hand side makes no sense at the moment.

It is the pathologies that give rise to the need for rigor. A satisfying resolution to the questions raised will require that we be absolutely precise about what we mean as we manipulate these infinite objects. It may seem that progress is slow at first, but that is because we do not want to fall into the trap of letting the biases of our intuition corrupt our arguments. Rigorous proofs are meant to be a check on intuition, and in the end we will see that they actually improve our mental picture of the mathematical infinite. As a final example, consider something as intuitively fundamental as the associative property of addition applied to the series $\sum_{n=1}^{\infty}(-1)^n$. Grouping the terms one way gives

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

whereas grouping in another yields

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

Manipulations that are legitimate in finite settings do not always extend to infinite settings. Deciding when they do and why they do not is one of the central themes of analysis.

2.2 The Limit of a Sequence

An understanding of infinite series depends heavily on a clear understanding of the theory of sequences. In fact, most of the concepts in analysis can be reduced to statements about the behavior of sequences. Thus, we will spend a significant amount of time investigating sequences before taking on infinite series.

Definition 2.2.1. A *sequence* is a function whose domain is $\mathbf{N}$.

This formal definition leads immediately to the familiar depiction of a sequence as an ordered list of real numbers. Given a function $f : \mathbf{N} \to \mathbf{R}$, $f(n)$ is just the nth term on the list. The notation for sequences reinforces this familiar understanding.

Example 2.2.2. Each of the following are common ways to describe a sequence.

(i) $(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots)$,

(ii) $(\frac{1+n}{n})_{n=1}^{\infty} = (\frac{2}{1}, \frac{3}{2}, \frac{4}{3}, \cdots)$,

(iii) (a_n), where $a_n = 2^n$ for each $n \in \mathbf{N}$,

(iv) (x_n), where $x_1 = 2$ and $x_{n+1} = \frac{x_n+1}{2}$.

On occasion, it will be more convenient to index a sequence beginning with $n = 0$ or $n = n_0$ for some natural number n_0 different from 1. These minor variations should cause no confusion. What is essential is that a sequence be an *infinite* list of real numbers. What happens at the beginning of such a list is of

little importance in most cases. The business of analysis is concerned with the behavior of the infinite "tail" of a given sequence.

We now present what is arguably the most important definition in the book.

Definition 2.2.3 (Convergence of a Sequence). A sequence (a_n) *converges* to a real number a if, for every positive number ϵ, there exists an $N \in \mathbf{N}$ such that whenever $n \geq N$ it follows that $|a_n - a| < \epsilon$.

To indicate that (a_n) converges to a, we write either $\lim a_n = a$ or $(a_n) \to a$.

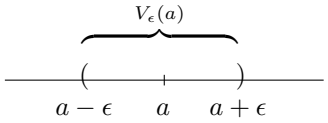
In an effort to decipher this complicated definition, it helps first to consider the ending phrase "$|a_n - a| < \epsilon$," and think about the points that satisfy an inequality of this type.

Definition 2.2.4. Given a real number $a \in \mathbf{R}$ and a positive number $\epsilon > 0$, the set

$$V_\epsilon(a) = \{x \in \mathbf{R} : |x - a| < \epsilon\}$$

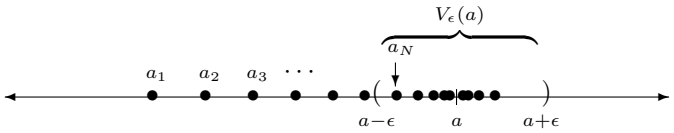
is called the ϵ-*neighborhood of* a.

Notice that $V_\epsilon(a)$ consists of all of those points whose distance from a is less than ϵ. Said another way, $V_\epsilon(a)$ is an interval, centered at a, with radius ϵ.



Recasting the definition of convergence in terms of ϵ-neighborhoods gives a more geometric impression of what is being described.

Definition 2.2.3B (Convergence of a Sequence: Topological Version). A sequence (a_n) converges to a if, given any ϵ-neighborhood $V_\epsilon(a)$ of a, there exists a point in the sequence after which all of the terms are in $V_\epsilon(a)$. In other words, every ϵ-neighborhood contains all but a finite number of the terms of (a_n).



Definition 2.2.3 and Definition 2.2.3B say precisely the same thing; the natural number N in the original version of the definition is the point where the sequence (a_n) enters $V_\epsilon(a)$, never to leave. It should be apparent that *the value of* N *depends on the choice of* ϵ. The smaller the ϵ-neighborhood, the larger N may have to be.

Example 2.2.5. Consider the sequence (a_n), where $a_n = 1/\sqrt{n}$.

Our intuitive understanding of limits points confidently to the conclusion that

$$\lim\left(\frac{1}{\sqrt{n}}\right) = 0.$$

Before trying to prove this not too impressive fact, let's first explore the relationship between ϵ and N in the definition of convergence. For the moment, take ϵ to be $1/10$. This defines a sort of "target zone" for the terms in the sequence. By claiming that the limit of (a_n) is 0, we are saying that the terms in this sequence eventually get arbitrarily close to 0. How close? What do we mean by "eventually"? We have set $\epsilon = 1/10$ as our standard for closeness, which leads to the ϵ-neighborhood $(-1/10, 1/10)$ centered around the limit 0. How far out into the sequence must we look before the terms fall into this interval? The 100th term $a_{100} = 1/10$ puts us right on the boundary, and a little thought reveals that

$$\text{if} \quad n > 100, \quad \text{then} \quad a_n \in \left(-\frac{1}{10}, \frac{1}{10}\right).$$

Thus, for $\epsilon = 1/10$ we choose $N = 101$ (or anything larger) as our response.

Now, our choice of $\epsilon = 1/10$ was rather whimsical, and we can do this again, letting $\epsilon = 1/50$. In this case, our target neighborhood shrinks to $(-1/50, 1/50)$, and it is apparent that we must travel farther out into the sequence before a_n falls into this interval. How far? Essentially, we require that

$$\frac{1}{\sqrt{n}} < \frac{1}{50} \quad \text{which occurs as long as} \quad n > 50^2 = 2500.$$

Thus, $N = 2501$ is a suitable response to the challenge of $\epsilon = 1/50$.

It may seem as though this duel could continue forever, with different ϵ challenges being handed to us one after another, each one requiring a suitable value of N in response. In a sense, this is correct, except that the game is effectively over the instant we recognize a *rule* for how to choose N given an *arbitrary* $\epsilon > 0$. For this problem, the desired algorithm is implicit in the algebra carried out to compute the previous response of $N = 2501$. Whatever ϵ happens to be, we want

$$\frac{1}{\sqrt{n}} < \epsilon \quad \text{which is equivalent to insisting that} \quad n > \frac{1}{\epsilon^2}.$$

With this observation, we are ready to write the formal argument.

We claim that

$$\lim\left(\frac{1}{\sqrt{n}}\right) = 0.$$

Proof. Let $\epsilon > 0$ be an arbitrary positive number. Choose a natural number N satisfying

$$N > \frac{1}{\epsilon^2}.$$

We now verify that this choice of N has the desired property. Let $n \geq N$. Then,

$$n > \frac{1}{\epsilon^2} \quad \text{implies} \quad \frac{1}{\sqrt{n}} < \epsilon \quad \text{and hence} \quad |a_n - 0| < \epsilon.$$

□

Quantifiers

The definition of convergence given ealier is the result of hundreds of years of refining the intuitive notion of limit into a mathematically rigorous statement. The logic involved is complicated and is intimately tied to the use of the quantifiers "for all" and "there exists." Learning to write a grammatically correct convergence proof goes hand in hand with a deep understanding of why the quantifiers appear in the order that they do.

The definition begins with the phrase,

"*For all* ϵ, *there exists* $N \in \mathbf{N}$ such that ..."

Looking back at our first example, we see that our formal proof begins with, "Let $\epsilon > 0$ be an arbitrary positive number." This is followed by a construction of N and then a demonstration that this choice of N has the desired property. This, in fact, is a basic outline for how every convergence proof should be presented.

Template for a proof that $(x_n) \to x$:

- "Let $\epsilon > 0$ be arbitrary."
- Demonstrate a choice for $N \in \mathbf{N}$. This step usually requires the most work, almost all of which is done prior to actually writing the formal proof.
- Now, show that N actually works.
- "Assume $n \geq N$."
- With N well chosen, it should be possible to derive the inequality $|x_n - x| < \epsilon$.

Example 2.2.6. Show

$$\lim \left(\frac{n+1}{n} \right) = 1.$$

As mentioned, before attempting a formal proof, we first need to do some preliminary scratch work. In the first example, we experimented by assigning specific values to ϵ (and it is not a bad idea to do this again), but let us skip straight to the algebraic punch line. The last line of our proof should be that for suitably large values of n,

$$\left| \frac{n+1}{n} - 1 \right| < \epsilon.$$

Because

$$\left|\frac{n+1}{n} - 1\right| = \frac{1}{n},$$

this is equivalent to the inequality $1/n < \epsilon$ or $n > 1/\epsilon$. Thus, choosing N to be an integer greater than $1/\epsilon$ will suffice.

With the work of the proof done, all that remains is the formal writeup.

Proof. Let $\epsilon > 0$ be arbitrary. Choose $N \in \mathbf{N}$ with $N > 1/\epsilon$. To verify that this choice of N is appropriate, let $n \in \mathbf{N}$ satisfy $n \geq N$. Then, $n \geq N$ implies $n > 1/\epsilon$, which is the same as saying $1/n < \epsilon$. Finally, this means

$$\left|\frac{n+1}{n} - 1\right| < \epsilon,$$

as desired. □

Divergence

Significant insight into the role of the quantifiers in the definition of convergence can be gained by studying an example of a sequence that does not have a limit.

Example 2.2.7. Consider the sequence

$$\left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \cdots\right).$$

How can we argue that this sequence does not converge to zero? Looking at the first few terms, it seems the initial evidence actually supports such a conclusion. Given a challenge of $\epsilon = 1/2$, a little reflection reveals that after $N = 3$ all the terms fall into the neighborhood $(-1/2, 1/2)$. We could also handle $\epsilon = 1/4$. (What is the smallest possible N in this case?)

But the definition of convergence says "*For all* $\epsilon > 0$...," and it should be apparent that there is no response to a choice of $\epsilon = 1/10$, for instance. This leads us to an important observation about the logical negation of the definition of convergence of a sequence. To prove that a particular number x is *not* the limit of a sequence (x_n), we must produce a single value of ϵ for which no $N \in \mathbf{N}$ works. More generally speaking, the negation of a statement that begins "For all P, there exists Q..." is the statement, "For at least one P, no Q is possible..." For instance, how could we disprove the spurious claim that "At every college in the United States, there is a student who is at least seven feet tall"?

We have argued that the preceding sequence does not converge to 0. Let's argue against the claim that it converges to 1/5. Choosing $\epsilon = 1/10$ produces the neighborhood $(1/10, 3/10)$. Although the sequence continually revisits this neighborhood, there is no point at which it enters and never leaves as the definition requires. Thus, no N exists for $\epsilon = 1/10$, so the sequence does not converge to 1/5.

Of course, this sequence does not converge to any other real number, and it would be more satisfying to simply say that this sequence does not converge.

Definition 2.2.8. A sequence that does not converge is said to *diverge*.

Although it is not too difficult, we will postpone arguing for divergence in general until we develop a more economical divergence criterion later in Section 2.5.

Exercises

Exercise 2.2.1. Verify, using the definition of convergence of a sequence, that the following sequences converge to the proposed limit.

(a) $\lim \frac{1}{(6n^2+1)} = 0$.

(b) $\lim \frac{3n+1)}{(2n+5)} = \frac{3}{2}$.

(c) $\lim \frac{2}{\sqrt{n+3}} = 0$.

Exercise 2.2.2. What happens if we reverse the order of the quantifiers in Definition 2.2.3?

Definition: A sequence (x_n) *verconges* to x if *there exists* an $\epsilon > 0$ such that *for all* $N \in \mathbf{N}$ it is true that $n \geq N$ implies $|x_n - x| < \epsilon$.

Give an example of a vercongent sequence. Can you give an example of a vergonent sequence that is divergent? What exactly is being described in this strange definition?

Exercise 2.2.3. Describe what we would have to demonstrate in order to disprove each of the following statements.

(a) At every college in the United States, there is a student who is at least seven feet tall.

(b) For all colleges in the United States, there exists a professor who gives every student a grade of either A or B.

(c) There exists a college in the United States where every student is at least six feet tall.

Exercise 2.2.4. Argue that the sequence

$$1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, (5 \text{ zeros}), 1, \ldots$$

does not converge to zero. For what values of $\epsilon > 0$ does there exist a response N. For which values of $\epsilon > 0$ is there no suitable response?

Exercise 2.2.5. Let $[[x]]$ be the greatest integer less than or equal to x. For example, $[[\pi]] = 3$ and $[[3]] = 3$. Find $\lim a_n$ and supply proofs for each conclusion if

(a) $a_n = [[1/n]]$,

(b) $a_n = [[(10 + n)/2n]]$.

Reflecting on these examples, comment on the statement following Definition 2.2.3 that "the smaller the ϵ-neighborhood, the larger N may have to be."

Exercise 2.2.6. Suppose that for a particular $\epsilon > 0$ we have found a suitable value of N that "works" for a given sequence in the sense of Definition 2.2.3.

(a) Then, any larger/smaller (*pick one*) N will also work for the same $\epsilon > 0$.

(b) Then, this same N will also work for any larger/smaller value of ϵ.

Exercise 2.2.7. Informally speaking, the sequence $\sqrt{n}$ "converges to infinity."

(a) Imitate the logical structure of Definition 2.2.3 to create a rigorous definition for the mathematical statement $\lim x_n = \infty$. Use this definition to prove $\lim \sqrt{n} = \infty$.

(b) What does your definition in (a) say about the particular sequence $(1, 0, 2, 0, 3, 0, 4, 0, 5, 0, \dots)$?

Exercise 2.2.8. Here are two useful definitions:

(i) A sequence (a_n) is *eventually* in a set $A \subseteq \mathbf{R}$ if there exists an $N \in \mathbf{N}$ such that $a_n \in A$ for all $n \geq N$.

(ii) A sequence (a_n) is *frequently* in a set $A \subseteq \mathbf{R}$ if, for every $N \in \mathbf{N}$, there exists an $n \geq N$ such that $a_n \in A$.

(a) Is the sequence $(-1)^n$ eventually or frequently in the set $\{1\}$?

(b) Which definition is stronger? Does frequently imply eventually or does eventually imply frequently?

(c) Give an alternate rephrasing of Definition 2.2.3B using either frequently or eventually. Which is the term we want?

(d) Suppose an infinite number of terms of a sequence (x_n) are equal to 2. Is (x_n) necessarily eventually in the interval $(1.9, 2.1)$? Is it frequently in $(1.9, 2.1)$?

2.3 The Algebraic and Order Limit Theorems

The real purpose of creating a rigorous definition for convergence of a sequence is *not* to have a tool to verify computational statements such as $\lim 2n/(n+2) = 2$. Historically, a definition of the limit like Definition 2.2.3 came 150 years after the founders of calculus began working with intuitive notions of convergence. The point of having such a logically tight description of convergence is so that we can confidently state and prove statements *about convergence sequences in general*. We are ultimately trying to resolve arguments about what is and is not true regarding the behavior of limits with respect to the mathematical manipulations we intend to inflict on them.

As a first example, let us prove that convergent sequences are bounded. The term"bounded" has a rather familiar connotation but, like everything else, we need to be explicit about what it means in this context.

Definition 2.3.1. A sequence (x_n) is *bounded* if there exists a number $M > 0$ such that $|x_n| \leq M$ for all $n \in \mathbf{N}$.

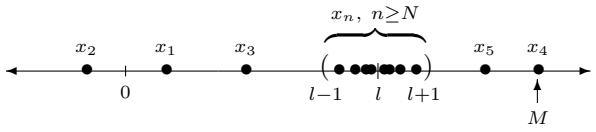
Geometrically, this means that we can find an interval $[-M, M]$ that contains every term in the sequence (x_n).

Theorem 2.3.2. *Every convergent sequence is bounded.*

Proof. Assume (x_n) converges to a limit l. This means that given a particular value of ϵ, say $\epsilon = 1$, we know there must exist an $N \in \mathbf{N}$ such that if $n \geq N$, then x_n is in the interval $(l-1, l+1)$. Not knowing whether l is positive or negative, we can certainly conclude that

$$|x_n| < |l| + 1$$

for all $n \geq N$.



We still need to worry (slightly) about the the terms in the sequence that come before the Nth term. Because there are only a finite number of these, we let

$$M = \max\{|x_1|, |x_2|, |x_3|, \ldots, |x_{N-1}|, |l| + 1\}.$$

It follows that $|x_n| \leq M$ for all $n \in \mathbf{N}$, as desired. □

This chapter began with a demonstration of how applying familiar algebraic properties (commutativity of addition) to infinite objects (series) can lead to paradoxical results. These examples are meant to instill in us a sense of caution and justify the extreme care we are taking in drawing our conclusions. The following theorems illustrate that sequences behave extremely well with respect to the operations of addition, multiplication, division, and order.

Theorem 2.3.3 (Algebraic Limit Theorem). *Let* $\lim a_n = a$, *and* $\lim b_n = b$. *Then,*

(i) $\lim(ca_n) = ca$, *for all* $c \in \mathbf{R}$;

(ii) $\lim(a_n + b_n) = a + b$;

(iii) $\lim(a_n b_n) = ab$;

(iv) $\lim(a_n/b_n) = a/b$, *provided* $b \neq 0$.

Proof. (i) Consider the case where $c \neq 0$. We want to show that the sequence (ca_n) converges to ca, so the structure of the proof follows the template we described in Section 2.2. First, we let ϵ be some arbitrary positive number. Our goal is to find some point in the sequence (ca_n) after which we have

$$|ca_n - ca| < \epsilon.$$

Now,

$$|ca_n - ca| = |c||a_n - a|.$$

We are given that $(a_n) \to a$, so we know we can make $|a_n - a|$ as small as we like. In particular, we can choose an N such that

$$|a_n - a| < \frac{\epsilon}{|c|}$$

whenever $n \geq N$. To see that this N indeed works, observe that, for all $n \geq N$,

$$|ca_n - ca| = |c||a_n - a| < |c|\frac{\epsilon}{|c|} = \epsilon.$$

The case $c = 0$ reduces to showing that the constant sequence $(0, 0, 0, \dots)$ converges to 0. This is addressed in Exercise 2.3.1.

Before continuing with parts (ii), (iii), and (iv), we should point out that the proof of (i), while somewhat short, is extremely typical for a convergence proof. Before embarking on a formal argument, it is a good idea to take an inventory of what we *want* to make less than ϵ, and what we are *given* can be made small for suitable choices of n. For the previous proof, we wanted to make $|ca_n - ca| < \epsilon$, and we were given $|a_n - a| <$ *anything we like* (for large values of n). Notice that in (i), and all of the ensuing arguments, the strategy each time is to bound the quantity we want to be less than ϵ, which in each case is

$$|(\text{terms of sequence}) - (\text{proposed limit})|,$$

with some algebraic combination of quantities over which we have control.

(ii) To prove this statement, we need to argue that the quantity

$$|(a_n + b_n) - (a + b)|$$

can be made less than an arbitrary ϵ using the assumptions that $|a_n - a|$ and $|b_n - b|$ can be made as small as we like for large n. The first step is to use the triangle inequality (Example 1.2.5) to say

$$|(a_n + b_n) - (a + b)| = |(a_n - a) + (b_n - b)| \leq |a_n - a| + |b_n - b|.$$

Again, we let $\epsilon > 0$ be arbitrary. The technique this time is to divide the ϵ between the two expressions on the right-hand side in the preceding inequality. Using the hypothesis that $(a_n) \to a$, we know there exists an N_1 such that

$$|a_n - a| < \frac{\epsilon}{2} \quad \text{whenever} \quad n \geq N_1.$$

Likewise, the assumption that $(b_n) \to b$ means that we can choose an N_2 so that

$$|b_n - b| < \frac{\epsilon}{2} \quad \text{whenever} \quad n \geq N_2.$$

The question now arises as to which of N_1 or N_2 we should take to be our choice of N. By choosing $N = \max\{N_1, N_2\}$, we ensure that if $n \geq N$, then $n \geq N_1$ and $n \geq N_2$. This allows us to conclude that

$$\begin{aligned} |(a_n + b_n) - (a + b)| &\leq |a_n - a| + |b_n - b| \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

for all $n \geq N$, as desired.

(iii) To show that $(a_n b_n) \to ab$, we begin by observing that

$$\begin{aligned} |a_n b_n - ab| &= |a_n b_n - ab_n + ab_n - ab| \\ &\leq |a_n b_n - ab_n| + |ab_n - ab| \\ &= |b_n||a_n - a| + |a||b_n - b|. \end{aligned}$$

In the initial step, we subtracted and then added ab_n, which created an opportunity to use the triangle inequality. Essentially, we have broken up the distance from $a_n b_n$ to ab with a midway point and are using the sum of the two distances to overestimate the original distance. This clever trick will become a familiar technique in arguments to come.

Letting $\epsilon > 0$ be arbitrary, we again proceed with the strategy of making each piece in the preceding inequality less than $\epsilon/2$. For the piece on the right-hand side ($|a||b_n - b|$), if $a \neq 0$ we can choose N_1 so that

$$n \geq N_1 \quad \text{implies} \quad |b_n - b| < \frac{1}{|a|}\frac{\epsilon}{2}.$$

(The case when $a = 0$ is handled in Exercise 2.3.7.) Getting the term on the left-hand side ($|b_n||a_n - a|$) to be less than $\epsilon/2$ is complicated by the fact that we have a variable quantity $|b_n|$ to contend with as opposed to the constant $|a|$ we encountered in the right-hand term. The idea is to replace $|b_n|$ with a worst-case estimate. Using the fact that convergent sequences are bounded (Theorem 2.3.2), we know there exists a bound $M > 0$ satisfying $|b_n| \leq M$ for all $n \in \mathbf{N}$. Now, we can choose N_2 so that

$$|a_n - a| < \frac{1}{M}\frac{\epsilon}{2} \quad \text{whenever} \quad n \geq N_2.$$

To finish the argument, pick $N = \max\{N_1, N_2\}$, and observe that if $n \geq N$, then

$$\begin{aligned} |a_n b_n - ab| &\leq |a_n b_n - ab_n| + |ab_n - ab| \\ &= |b_n||a_n - a| + |a||b_n - b| \\ &\leq M|a_n - a| + |a||b_n - b| \\ &< M\left(\frac{\epsilon}{M2}\right) + |a|\left(\frac{\epsilon}{|a|2}\right) = \epsilon. \end{aligned}$$

(iv) This final statement will follow from (iii) if we can prove that

$$(b_n) \to b \quad \text{implies} \quad \left(\frac{1}{b_n}\right) \to \frac{1}{b}$$

whenever $b \neq 0$. We begin by observing that

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \frac{|b - b_n|}{|b||b_n|}.$$

Because $(b_n) \to b$, we can make the preceding numerator as small as we like by choosing n large. The problem comes in that we need a worst-case estimate on the size of $1/(|b||b_n|)$. Because the b_n terms are in the denominator, we are no longer interested in an upper bound on $|b_n|$ but rather in an inequality of the form $|b_n| \geq \delta > 0$. This will then lead to a bound on the size of $1/(|b||b_n|)$.

The trick is to look far enough out into the sequence (b_n) so that the terms are closer to b than they are to 0. Consider the particular value $\epsilon_0 = |b|/2$. Because $(b_n) \to b$, there exists an N_1 such that $|b_n - b| < |b|/2$ for all $n \geq N_1$. This implies $|b_n| > |b|/2$.

Next, choose N_2 so that $n \geq N_2$ implies

$$|b_n - b| < \frac{\epsilon |b|^2}{2}.$$

Finally, if we let $N = \max\{N_1, N_2\}$, then $n \geq N$ implies

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = |b - b_n| \frac{1}{|b||b_n|} < \frac{\epsilon |b|^2}{2} \frac{1}{|b| \frac{|b|}{2}} = \epsilon.$$

□

Limits and Order

Although there are a few dangers to avoid (see Exercise 2.3.8), the Algebraic Limit Theorem verifies that the relationship between algebraic combinations of sequences and the limiting process is as trouble-free as we could hope for. Limits can be computed from the individual component sequences provided that each component limit exists. The limiting process is also well-behaved with respect to the order operation.

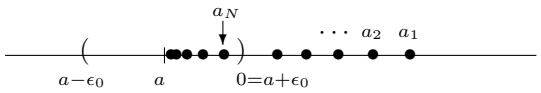
Theorem 2.3.4 (Order Limit Theorem). *Assume* $\lim a_n = a$ *and* $\lim b_n = b$.

(i) *If* $a_n \geq 0$ *for all* $n \in \mathbf{N}$, *then* $a \geq 0$.

(ii) *If* $a_n \leq b_n$ *for all* $n \in \mathbf{N}$, *then* $a \leq b$.

(iii) *If there exists* $c \in \mathbf{R}$ *for which* $c \leq b_n$ *for all* $n \in \mathbf{N}$, *then* $c \leq b$. *Similarly, if* $a_n \leq c$ *for all* $n \in \mathbf{N}$, *then* $a \leq c$.

Proof. (i) We will prove this by contradiction; thus, let's assume $a < 0$. The idea is to produce a term in the sequence (a_n) that is also less than zero. To do this, we consider the particular value $\epsilon_0 = |a|$. The definition of convergence guarantees that we can find an N such that $|a_n - a| < |a|$ for all $n \geq N$. In particular, this would mean that $|a_N - a| < |a|$, which implies $a_N < 0$. This contradicts our hypothesis that $a_N \geq 0$. We therefore conclude that $a \geq 0$.



(ii) The Algebraic Limit Theorem ensures that the sequence $(b_n - a_n)$ converges to $b-a$. Because $b_n - a_n \geq 0$, we can apply part (i) to get that $b-a \geq 0$.

(iii) Take $a_n = c$ (or $b_n = c$) for all $n \in \mathbf{N}$, and apply (ii). □

A word about the idea of "tails" is in order. Loosely speaking, limits and their properties do not depend at all on what happens at the beginning of the sequence but are strictly determined by what happens when n gets large. Changing the value of the first ten—or ten thousand—terms in a particular sequence has no effect on the limit. Theorem 2.3.4, part (i), for instance, assumes that $a_n \geq 0$ *for all* $n \in \mathbf{N}$. However, the hypothesis could be weakened by assuming only that there exists some point N_1 where $a_n \geq 0$ for all $n \geq N_1$. The theorem remains true, and in fact the same proof is valid with the provision that when N is chosen it be at least as large as N_1.

In the language of analysis, when a property (such as non-negativity) is not necessarily true about some finite number of initial terms but is true for all terms in the sequence after some point N, we say that the sequence *eventually* has this property. (See Exercise 2.2.8.) Theorem 2.3.4, part (i), could be restated, "Convergent sequences that are eventually nonnegative converge to nonnegative limits." Parts (ii) and (iii) have similar modifications, as will many other upcoming results.

Exercises

Exercise 2.3.1. Show that the constant sequence $(a, a, a, a, \dots)$ converges to a.

Exercise 2.3.2. Let $x_n \geq 0$ for all $n \in \mathbf{N}$.
(a) If $(x_n) \to 0$, show that $(\sqrt{x_n}) \to 0$.
(b) If $(x_n) \to x$, show that $(\sqrt{x_n}) \to \sqrt{x}$.

Exercise 2.3.3 (Squeeze Theorem). Show that if $x_n \leq y_n \leq z_n$ for all $n \in \mathbf{N}$, and if $\lim x_n = \lim z_n = l$, then $\lim y_n = l$ as well.

Exercise 2.3.4. Show that limits, if they exist, must be unique. In other words, assume $\lim a_n = l_1$ and $\lim a_n = l_2$, and prove that $l_1 = l_2$.

Exercise 2.3.5. Let (x_n) and (y_n) be given, and define (z_n) to be the "shuffled" sequence $(x_1, y_1, x_2, y_2, x_3, y_3, \dots, x_n, y_n, \dots)$. Prove that (z_n) is convergent if and only if (x_n) and (y_n) are both convergent with $\lim x_n = \lim y_n$.

Exercise 2.3.6. (a) Show that if $(b_n) \to b$, then the sequence of absolute values $|b_n|$ converges to $|b|$.

(b) Is the converse of part (a) true? If we know that $|b_n| \to |b|$, can we deduce that $(b_n) \to b$?

Exercise 2.3.7. (a) Let (a_n) be a bounded (not necessarily convergent) sequence, and assume $\lim b_n = 0$. Show that $\lim(a_n b_n) = 0$. Why are we not allowed to use the Algebraic Limit Theorem to prove this?

(b) Can we conclude anything about the convergence of $(a_n b_n)$ if we assume that (b_n) converges to some nonzero limit b?

(c) Use (a) to prove Theorem 2.3.3, part (iii), for the case when $a = 0$.

Exercise 2.3.8. Give an example of each of the following, or state that such a request is impossible by referencing the proper theorem(s):

(a) sequences (x_n) and (y_n), which both diverge, but whose sum $(x_n + y_n)$ converges;

(b) sequences (x_n) and (y_n), where (x_n) converges, (y_n) diverges, and $(x_n + y_n)$ converges;

(c) a convergent sequence (b_n) with $b_n \neq 0$ for all n such that $(1/b_n)$ diverges;

(d) an unbounded sequence (a_n) and a convergent sequence (b_n) with $(a_n - b_n)$ bounded;

(e) two sequences (a_n) and (b_n), where $(a_n b_n)$ and (a_n) converge but (b_n) does not.

Exercise 2.3.9. Does Theorem 2.3.4 remain true if all of the inequalities are assumed to be strict? If we assume, for instance, that a convergent sequence (x_n) satisfies $x_n > 0$ for all $n \in \mathbf{N}$, what may we conclude about the limit?

Exercise 2.3.10. If $(a_n) \to 0$ and $|b_n - b| \leq a_n$, then show that $(b_n) \to b$.

Exercise 2.3.11 (Cesaro Means). Show that if (x_n) is a convergent sequence, then the sequence given by the averages

$$y_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

also converges to the same limit.

Give an example to show that it is possible for the sequence (y_n) of averages to converge even if (x_n) does not.

Exercise 2.3.12. Consider the doubly indexed array $a_{m,n} = m/(m+n)$.

(a) Intuitively speaking, what should $\lim_{m,n\to\infty} a_{m,n}$ represent? Compute the "iterated" limits

$$\lim_{n\to\infty} \lim_{m\to\infty} a_{m,n} \quad \text{and} \quad \lim_{m\to\infty} \lim_{n\to\infty} a_{m,n}.$$

(b) Formulate a rigorous definition in the style of Definition 2.2.3 for the statement

$$\lim_{m,n\to\infty} a_{m,n} = l.$$

2.4 The Monotone Convergence Theorem and a First Look at Infinite Series

We showed in Theorem 2.3.2 that convergent sequences are bounded. The converse statement is certainly not true. It is not too difficult to produce an example of a bounded sequence that does not converge. On the other hand, if a bounded sequence is *monotone*, then in fact it does converge.

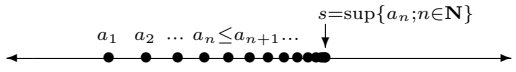
Definition 2.4.1. A sequence (a_n) is *increasing* if $a_n \leq a_{n+1}$ for all $n \in \mathbf{N}$ and *decreasing* if $a_n \geq a_{n+1}$ for all $n \in \mathbf{N}$. A sequence is *monotone* if it is either increasing or decreasing.

Theorem 2.4.2 (Monotone Convergence Theorem). *If a sequence is monotone and bounded, then it converges.*

Proof. Let (a_n) be monotone and bounded. To prove (a_n) converges using the definition of convergence, we are going to need a candidate for the limit. Let's assume the sequence is increasing (the decreasing case is handled similarly), and consider the *set* of points $\{a_n : n \in \mathbf{N}\}$. By assumption, this set is bounded, so we can let

$$s = \sup\{a_n : n \in \mathbf{N}\}.$$

It seems reasonable to claim that $\lim(a_n) = s$.



To prove this, let $\epsilon > 0$. Because s is the least upper bound of $\{a_n : n \in \mathbf{N}\}$, $s - \epsilon$ is not an upper bound, so there exists a point in the sequence a_N such that $s - \epsilon < a_N$. Now, the fact that (a_n) is increasing implies that if $n \geq N$, then $a_N \leq a_n$. Hence,

$$s - \epsilon < a_N \leq a_n \leq s < s + \epsilon,$$

which implies $|a_n - s| < \epsilon$, as desired. □

The Monotone Convergence Theorem is extremely useful for the study of infinite series, largely because it asserts the convergence of a sequence without explicit mention of the actual limit. This is a good moment to do some preliminary investigations, so it is time to formalize the relationship between sequences and series.

Definition 2.4.3. Let (b_n) be a sequence. An *infinite series* is a formal expression of the form

$$\sum_{n=1}^{\infty} b_n = b_1 + b_2 + b_3 + b_4 + b_5 + \cdots .$$

We define the corresponding *sequence of partial sums* (s_m) by

$$s_m = b_1 + b_2 + b_3 + \cdots + b_m,$$

and say that the series $\sum_{n=1}^{\infty} b_n$ *converges to* B if the sequence (s_m) converges to B. In this case, we write $\sum_{n=1}^{\infty} b_n = B$.

Example 2.4.4. Consider

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

Because the terms in the sum are all positive, the sequence of partial sums given by

$$s_m = 1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{m^2}$$

is increasing. The question is whether or not we can find some upper bound on (s_m). To this end, observe

$$\begin{aligned} s_m &= 1 + \frac{1}{2 \cdot 2} + \frac{1}{3 \cdot 3} + \frac{1}{4 \cdot 4} + \cdots + \frac{1}{m^2} \\ &< 1 + \frac{1}{2 \cdot 1} + \frac{1}{3 \cdot 2} + \frac{1}{4 \cdot 3} + \cdots + \frac{1}{m(m-1)} \\ &= 1 + \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{(m-1)} - \frac{1}{m}\right) \\ &= 1 + 1 - \frac{1}{m} \\ &< 2. \end{aligned}$$

Thus, 2 is an upper bound for the sequence of partial sums, so by the Monotone Convergence Theorem, $\sum_{n=1}^{\infty} 1/n^2$ converges to some (presently unknown) limit less than 2.

Example 2.4.5 (Harmonic Series). This time, consider the so-called *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n}.$$

Again, we have an increasing sequence of partial sums,

$$s_m = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m},$$

that upon naive inspection appears as though it may be bounded. However, 2 is no longer an upper bound because

$$s_4 = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) > 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) = 2.$$

A similar calculation shows that $s_8 > 2\frac{1}{2}$, and we can see that in general

$$\begin{aligned} s_{2^k} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \cdots + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^{k-1}+1} + \cdots + \frac{1}{2^k}\right) \\ &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \cdots + \frac{1}{8}\right) + \cdots + \left(\frac{1}{2^k} + \cdots + \frac{1}{2^k}\right) \\ &= 1 + \frac{1}{2} + 2\left(\frac{1}{4}\right) + 4\left(\frac{1}{8}\right) + \cdots + 2^{k-1}\left(\frac{1}{2^k}\right) \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} \\ &= 1 + k\left(\frac{1}{2}\right), \end{aligned}$$

which is unbounded. Thus, despite the incredibly slow pace, the sequence of partial sums of $\sum_{n=1}^{\infty} 1/n$ eventually surpasses every number on the positive real line. Because convergent sequences are bounded, the harmonic series diverges.

The previous example is a special case of a general argument that can be used to determine the convergence or divergence of a large class of infinite series.

Theorem 2.4.6 (Cauchy Condensation Test). *Suppose (b_n) is decreasing and satisfies $b_n \geq 0$ for all $n \in \mathbf{N}$. Then, the series $\sum_{n=1}^{\infty} b_n$ converges if and only if the series*

$$\sum_{n=0}^{\infty} 2^n b_{2^n} = b_1 + 2b_2 + 4b_4 + 8b_8 + 16b_{16} + \cdots$$

converges.

Proof. First, assume that $\sum_{n=0}^{\infty} 2^n b_{2^n}$ converges. Theorem 2.3.2 guarantees that the partial sums

$$t_k = b_1 + 2b_2 + 4b_4 + \cdots + 2^k b_{2^k}$$

are bounded; that is, there exists an $M > 0$ such that $t_k \leq M$ for all $k \in \mathbf{N}$. We want to prove that $\sum_{n=1}^{\infty} b_n$ converges. Because $b_n \geq 0$, we know that the partial sums are increasing, so we only need to show that

$$s_m = b_1 + b_2 + b_3 + \cdots + b_m$$

is bounded.

Fix m and let k be large enough to ensure $m \leq 2^{k+1} - 1$. Then, $s_m \leq s_{2^{k+1}-1}$ and

$$\begin{aligned} s_{2^{k+1}-1} &= b_1 + (b_2 + b_3) + (b_4 + b_5 + b_6 + b_7) + \cdots + (b_{2^k} + \cdots + b_{2^{k+1}-1}) \\ &\leq b_1 + (b_2 + b_2) + (b_4 + b_4 + b_4 + b_4) + \cdots + (b_{2^k} + \cdots + b_{2^k}) \\ &= b_1 + 2b_2 + 4b_4 + \cdots + 2^k b_{2^k} = t_k. \end{aligned}$$

Thus, $s_m \leq t_k \leq M$, and the sequence (s_m) is bounded. By the Monotone Convergence Theorem, we can conclude that $\sum_{n=1}^{\infty} b_n$ converges.

The proof that $\sum_{n=0}^{\infty} 2^n b_{2^n}$ diverges implies $\sum_{n=1}^{\infty} b_n$ diverges is similar to Example 2.4.5. The details are requested in Exercise 2.4.1. □

Corollary 2.4.7. *The series $\sum_{n=1}^{\infty} 1/n^p$ converges if and only if $p > 1$.*

A rigorous argument for this corollary requires a few basic facts about geometric series. The proof is requested in Exercise 2.7.7 at the end of Section 2.7 where geometric series are discussed.

Exercises

Exercise 2.4.1. Complete the proof of Theorem 2.4.6 by showing that if the series $\sum_{n=0}^{\infty} 2^n b_{2^n}$ diverges, then so does $\sum_{n=1}^{\infty} b_n$. Example 2.4.5 may be a useful reference.

Exercise 2.4.2. (a) Prove that the sequence defined by $x_1 = 3$ and

$$x_{n+1} = \frac{1}{4 - x_n}$$

converges.

(b) Now that we know $\lim x_n$ exists, explain why $\lim x_{n+1}$ must also exist and equal the same value.

(c) Take the limit of each side of the recursive equation in part (a) of this exercise to explicitly compute $\lim x_n$.

Exercise 2.4.3. Following the model of Exercise 2.4.2, show that the sequence defined by $y_1 = 1$ and $y_{n+1} = 4 - 1/y_n$ converges and find the limit.

Exercise 2.4.4. Show that

$$\sqrt{2}, \sqrt{2\sqrt{2}}, \sqrt{2\sqrt{2\sqrt{2}}}, \ldots$$

converges and find the limit.

Exercise 2.4.5 (Calculating Square Roots). Let $x_1 = 2$, and define

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right).$$

(a) Show that x_n^2 is always greater than 2, and then use this to prove that $x_n - x_{n+1} \geq 0$. Conclude that $\lim x_n = \sqrt{2}$.

(b) Modify the sequence (x_n) so that it converges to $\sqrt{c}$.

Exercise 2.4.6 (Limit Superior). Let (a_n) be a bounded sequence.

(a) Prove that the sequence defined by $y_n = \sup\{a_k : k \geq n\}$ converges.

(b) The *limit superior* of (a_n), or $\limsup a_n$, is defined by

$$\limsup a_n = \lim y_n,$$

where y_n is the sequence from part (a) of this exercise. Provide a reasonable definition for $\liminf a_n$ and briefly explain why it always exists for any bounded sequence.

(c) Prove that $\liminf a_n \leq \limsup a_n$ for every bounded sequence, and give an example of a sequence for which the inequality is strict.

(d) Show that $\liminf a_n = \limsup a_n$ if and only if $\lim a_n$ exists. In this case, all three share the same value.

2.5 Subsequences and the Bolzano–Weierstrass Theorem

In Example 2.4.5, we showed that the sequence of partial sums (s_m) of the harmonic series does not converge by focusing our attention on a particular *subsequence* (s_{2^k}) of the original sequence. For the moment, we will put the topic of infinite series aside and more fully develop the important concept of subsequences.

Definition 2.5.1. Let (a_n) be a sequence of real numbers, and let $n_1 < n_2 < n_3 < n_4 < n_5 < \cdots$ be an increasing sequence of natural numbers. Then the sequence

$$a_{n_1}, a_{n_2}, a_{n_3}, a_{n_4}, a_{n_5}, \cdots$$

is called a *subsequence* of (a_n) and is denoted by (a_{n_j}), where $j \in \mathbf{N}$ indexes the subsequence.

Notice that the order of the terms in a subsequence is the same as in the original sequence, and repetitions are not allowed. Thus if

$$(a_n) = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \cdots\right),$$

then

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \cdots\right) \quad \text{and} \quad \left(\frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \frac{1}{10000}, \cdots\right)$$

are examples of legitimate subsequences, whereas

$$\left(\frac{1}{10}, \frac{1}{5}, \frac{1}{100}, \frac{1}{50}, \frac{1}{1000}, \frac{1}{500}, \cdots\right) \quad \text{and} \quad \left(1, 1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}, \cdots\right)$$

are not.

Theorem 2.5.2. *Subsequences of a convergent sequence converge to the same limit as the original sequence.*

Proof. Exercise 2.5.1 □

This not too surprising result has several somewhat surprising applications. It is the key ingredient for understanding when infinite sums are associative (Exercise 2.5.2). We can also use it in the following clever way to compute values of some familiar limits.

Example 2.5.3. Let $0 < b < 1$. Because

$$b > b^2 > b^3 > b^4 > \cdots > 0,$$

the sequence (b^n) is decreasing and bounded below. The Monotone Convergence Theorem allows us to conclude that (b^n) converges to some l satisfying $b > l \geq 0$. To compute l, notice that (b^{2n}) is a subsequence, so $(b^{2n}) \to l$ by Theorem 2.5.2. But $(b^{2n}) = (b^n)(b^n)$, so by the Algebraic Limit Theorem, $(b^{2n}) \to l \cdot l = l^2$. Because limits are unique, $l^2 = l$, and thus $l = 0$.

Without much trouble (Exercise 2.5.5), we can generalize this example to conclude $(b^n) \to 0$ whenever $-1 < b < 1$.

Example 2.5.4 (Divergence Criterion). Theorem 2.5.2 is also useful for providing economical proofs for divergence. In Example 2.2.7, we were quite sure that

$$\left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \frac{1}{5}, -\frac{1}{5}, \dots\right)$$

did not converge to any proposed limit. Notice that

$$\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \dots\right)$$

is a subsequence that converges to $1/5$. Also,

$$\left(-\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, -\frac{1}{5}, \dots\right)$$

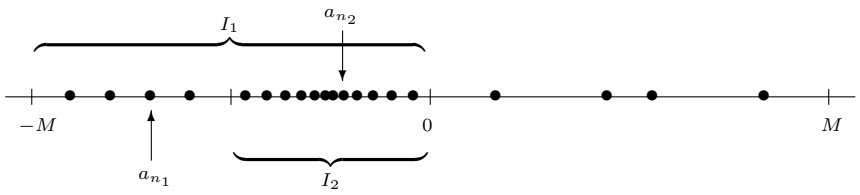
is a different subsequence of the original sequence that converges to $-1/5$. Because we have two subsequences converging to two different limits, we can rigorously conclude that the original sequence diverges.

The Bolzano–Weierstrass Theorem

In the previous example, it was rather easy to spot a convergent subsequence (or two) hiding in the original sequence. For *bounded* sequences, it turns out that it is always possible to find at least one such convergent subsequence.

Theorem 2.5.5 (Bolzano–Weierstrass Theorem). *Every bounded sequence contains a convergent subsequence.*

Proof. Let (a_n) be a bounded sequence so that there exists $M > 0$ satisfying $|a_n| \leq M$ for all $n \in \mathbf{N}$. Bisect the closed interval $[-M, M]$ into the two closed intervals $[-M, 0]$ and $[0, M]$. (The midpoint is included in both halves.) Now, it must be that at least one of these closed intervals contains an infinite number of the points in the sequence (a_n). Select a half for which this is the case and label that interval as I_1. Then, let a_{n_1} be some point in the sequence (a_n) satisfying $a_{n_1} \in I_1$.



Next, we bisect I_1 into closed intervals of equal length, and let I_2 be a half that again contains an infinite number of points of the original sequence. Because there are an infinite number of points from (a_n) to choose from, we can select an a_{n_2} from the original sequence with $n_2 > n_1$ and $a_{n_2} \in I_2$. In general, we construct the closed interval I_k by taking a half of I_{k-1} containing an infinite number of points of (a_n) and then select $n_k > n_{k-1} > \cdots > n_2 > n_1$ so that $a_{n_k} \in I_k$.

We want to argue that (a_{n_k}) is a convergent subsequence, but we need a candidate for the limit. The sets

$$I_1 \supseteq I_2 \supseteq I_3 \supseteq \cdots$$

form a nested sequence of closed intervals, and by the Nested Interval Property there exists at least one point $x \in \mathbf{R}$ contained in every I_k. This provides us with the candidate we were looking for. It just remains to show that $(a_{n_k}) \to x$.

Let $\epsilon > 0$. By construction, the length of I_k is $M(1/2)^{k-1}$ which converges to zero. (This follows from Example 2.5.3 and the Algebraic Limit Theorem.) Choose N so that $k \geq N$ implies that the length of I_k is less than ϵ. Because x and a_{n_k} are both in I_k, it follows that $|a_{n_k} - x| < \epsilon$. □

Exercises

Exercise 2.5.1. Prove Theorem 2.5.2.

Exercise 2.5.2. (a) Prove that if an infinite series converges, then the associative property holds. Assume $a_1 + a_2 + a_3 + a_4 + a_5 + \cdots$ converges to a limit L (i.e., the sequence of partial sums $(s_n) \to L$). Show that any regrouping of the terms

$$(a_1 + a_2 + \cdots + a_{n_1}) + (a_{n_1+1} + \cdots + a_{n_2}) + (a_{n_2+1} + \cdots + a_{n_3}) + \cdots$$

leads to a series that also converges to L.

(b) Compare this result to the example discussed at the end of Section 2.1 where infinite addition was shown not to be associative. Why doesn't our proof in (a) apply to this example?

Exercise 2.5.3. Give an example of each of the following, or argue that such a request is impossible.

(a) A sequence that does not contain 0 or 1 as a term but contains subsequences converging to each of these values.

(b) A monotone sequence that diverges but has a convergent subsequence.

(c) A sequence that contains subsequences converging to every point in the infinite set $\{1, 1/2, 1/3, 1/4, 1/5, \dots\}$.

(d) An unbounded sequence with a convergent subsequence.

(e) A sequence that has a subsequence that is bounded but contains no subsequence that converges.

Exercise 2.5.4. Assume (a_n) is a bounded sequence with the property that every convergent subsequence of (a_n) converges to the same limit $a \in \mathbf{R}$. Show that (a_n) must converge to a.

Exercise 2.5.5. Extend the result proved in Example 2.5.3 to the case $|b| < 1$. Show $\lim(b^n) = 0$ whenever $-1 < b < 1$.

Exercise 2.5.6. Let (a_n) be a bounded sequence, and define the set

$$S = \{x \in \mathbf{R} : x < a_n \text{ for infinitely many terms } a_n\}.$$

Show that there exists a subsequence (a_{n_k}) converging to $s = \sup S$. (This is a direct proof of the Bolzano–Weierstrass Theorem using the Axiom of Completeness.)

2.6 The Cauchy Criterion

The following definition bears a striking resemblance to the definition of convergence for a sequence.

Definition 2.6.1. A sequence (a_n) is called a *Cauchy sequence* if, for every $\epsilon > 0$, there exists an $N \in \mathbf{N}$ such that whenever $m, n \geq N$ it follows that $|a_n - a_m| < \epsilon$.

To make the comparison easier, let's restate the definition of convergence.

Definition 2.2.3 (Convergence of a Sequence). A sequence (a_n) *converges* to a real number a if, for every $\epsilon > 0$, there exists an $N \in \mathbf{N}$ such that whenever $n \geq N$ it follows that $|a_n - a| < \epsilon$.

As we have discussed, the definition of convergence asserts that, given an arbitrary positive ϵ, it is possible to find a point in the sequence after which the terms of the sequence are all closer to the limit a than the given ϵ. On

the other hand, a sequence is a Cauchy sequence if, for every ϵ, there is a point in the sequence after which the terms are all closer *to each other* than the given ϵ. To spoil the surprise, we will argue in this section that in fact these two definitions are equivalent: Convergent sequences are Cauchy sequences, and Cauchy sequences converge. The significance of the definition of a Cauchy sequence is that there is no mention of a limit. This is somewhat like the situation with the Monotone Convergence Theorem in that we will have another way of proving that sequences converge without having any explicit knowledge of what the limit might be.

Theorem 2.6.2. *Every convergent sequence is a Cauchy sequence.*

Proof. Assume (x_n) converges to x. To prove that (x_n) is Cauchy, we must find a point in the sequence after which we have $|x_n - x_m| < \epsilon$. This can be done using an application of the triangle inequality. The details are requested in Exercise 2.6.2. □

The converse is a bit more difficult to prove, mainly because, in order to prove that a sequence converges, we must have a proposed limit for the sequence to approach. We have been in this situation before in the proofs of the Monotone Convergence Theorem and the Bolzano–Weierstrass Theorem. Our strategy here will be to use the Bolzano–Weierstrass Theorem. This is the reason for the next lemma. (Compare this with Theorem 2.3.2.)

Lemma 2.6.3. *Cauchy sequences are bounded.*

Proof. Given $\epsilon = 1$, there exists an N such that $|x_m - x_n| < 1$ for all $m, n \geq N$. Thus, we must have $|x_n| < |x_N| + 1$ for all $n \geq N$. It follows that

$$M = \max\{|x_1|, |x_2|, |x_3|, \ldots, |x_{N-1}|, |x_N| + 1\}$$

is a bound for the sequence (x_n). □

Theorem 2.6.4 (Cauchy Criterion). *A sequence converges if and only if it is a Cauchy sequence.*

Proof. ($\Rightarrow$) This direction is Theorem 2.6.2.

($\Leftarrow$) For this direction, we start with a Cauchy sequence (x_n). Lemma 2.6.3 guarantees that (x_n) is bounded, so we may use the Bolzano–Weierstrass Theorem to produce a convergent subsequence (x_{n_k}). Set

$$x = \lim x_{n_k}.$$

The idea is to show that the original sequence (x_n) converges to this same limit. Once again, we will use a triangle inequality argument. We know the terms in the subsequence are getting close to the limit x, and the assumption that (x_n) is Cauchy implies the terms in the "tail" of the sequence are close to each other. Thus, we want to make each of these distances less than half of the prescribed ϵ.