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.

S. Abbott, *Understanding Analysis*, Undergraduate Texts
in Mathematics, DOI 10.1007/978-1-4939-2712-8_1

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*

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

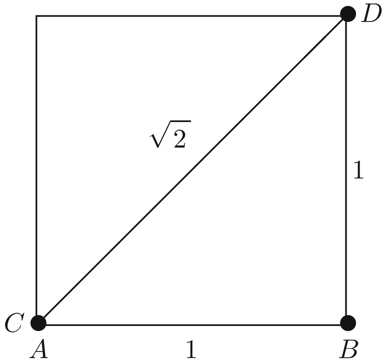


Figure 1.1: $\sqrt{2}$ EXISTS AS A GEOMETRIC LENGTH.

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

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

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.

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:

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

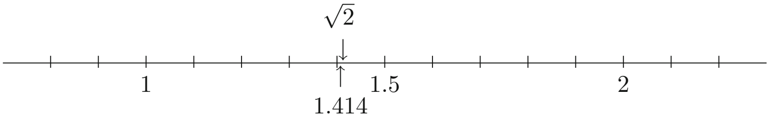


Figure 1.2: Approximating $\sqrt{2}$ with rational numbers.

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 **Z**, there are no intervals of empty space. Given any two rational numbers $r < s$, the rational number $(r+s)/2$ sits halfway in between, implying that the rational numbers are densely nestled together.

With the field properties of **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 **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* **R**.

The question of how to actually construct **R** from **Q** is rather complicated business. It is discussed in Section 1.3, and then again in more detail in Section 8.6. For the moment, it is not too inaccurate to say that **R** is obtained by filling in the gaps in **Q**. Wherever there is a hole, a new *irrational* number is defined and placed into the ordering that already exists on **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 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: $\mathbf{N} = \{1, 2, 3, \ldots\}$.

(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

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

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

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

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 1.2.5.

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 1.7.

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 8.5, 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 1.2.6) 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 form.) 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. Show that there is no rational number r satisfying $2^r = 3$.

Exercise 1.2.3. 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 $\bigcap_{n=1}^{\infty} A_n$ 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 $\bigcap_{n=1}^{\infty} A_n$ 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 1.2.4. Produce an infinite collection of sets $A_1, A_2, A_3, \ldots$ with the property that every A_i has an infinite number of elements, $A_i \cap A_j = \emptyset$ for all $i \neq j$, and $\bigcup_{i=1}^{\infty} A_i = \mathbf{N}$.

Exercise 1.2.5 (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 1.2.6. (a) Verify the triangle inequality in the special case where a and b have the same sign.

(b) Find an efficient proof for all the cases at once by first demonstrating $(a+b)^2 \leq (|a|+|b|)^2$.

(c) Prove $|a-b| \leq |a-c| + |c-d| + |d-b|$ for all a, b, c, and d.

(d) Prove $||a|-|b|| \leq |a-b|$. (The unremarkable identity $a = a - b + b$ may be useful.)

Exercise 1.2.7. 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 1.2.8. Here are two important definitions related to a function $f : A \to B$. The function f 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 $f(a) = b$.

Give an example of each or state that the request is impossible:

(a) $f : \mathbf{N} \to \mathbf{N}$ that is 1–1 but not onto.

(b) $f : \mathbf{N} \to \mathbf{N}$ that is onto but not 1–1.

(c) $f : \mathbf{N} \to \mathbf{Z}$ that is 1–1 and onto.

Exercise 1.2.9. 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 1.2.10. Decide which of the following are true statements. Provide a short justification for those that are valid and a counterexample for those that are not:

(a) Two real numbers satisfy $a < b$ if and only if $a < b + \epsilon$ for every $\epsilon > 0$.

(b) Two real numbers satisfy $a < b$ if $a < b + \epsilon$ for every $\epsilon > 0$.

(c) Two real numbers satisfy $a \leq b$ if and only if $a < b + \epsilon$ for every $\epsilon > 0$.

Exercise 1.2.11. Form the logical negation of each claim. One trivial way to do this is to simply add "It is not the case that..." in front of each assertion. To make this interesting, fashion the negation into a positive statement that avoids using the word "not" altogether. In each case, make an intuitive guess as to whether the claim or its negation is the true statement.

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

(b) There exists a real number $x > 0$ such that $x < 1/n$ for all $n \in \mathbf{N}$.

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

Exercise 1.2.12. Let $y_1 = 6$, and for each $n \in \mathbf{N}$ define $y_{n+1} = (2y_n - 6)/3$.

(a) Use induction to prove that the sequence satisfies $y_n > -6$ for all $n \in \mathbf{N}$.

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

Exercise 1.2.13. For this exercise, assume Exercise 1.2.5 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) It is tempting to appeal to induction to conclude

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c,$$

but induction does not apply here. Induction is used to prove that a particular statement holds for every value of $n \in \mathbf{N}$, but this does not imply the validity of the infinite case. To illustrate this point, find an example of a collection of sets $B_1, B_2, B_3, \ldots$ where $\bigcap_{i=1}^{n} B_i \neq \emptyset$ is true for every $n \in \mathbf{N}$, but $\bigcap_{i=1}^{\infty} B_i \neq \emptyset$ fails.

(c) Nevertheless, the infinite version of De Morgan's Law stated in (b) is a valid statement. Provide 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 **R** of real numbers is an extension of the rational numbers **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 **R** quite similar to our own informal understanding at this point. Eventually, enough scrutiny was directed at the detailed structure of **R** so that, in the 1870s, a handful of ways to rigorously *construct* **R** from **Q** were proposed.

Following this historical model, our own rigorous construction of **R** from **Q** is postponed until Section 8.6. 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, **R** is a set containing **Q**. The operations of addition and multiplication on **Q** extend to all of **R** in such a way that every element of **R** has an additive inverse and every nonzero element of **R** has a multiplicative inverse. Echoing the discussion in Section 1.1, we assume **R** is a *field*, meaning that addition and multiplication of real numbers are 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 **Q** extend to all of **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

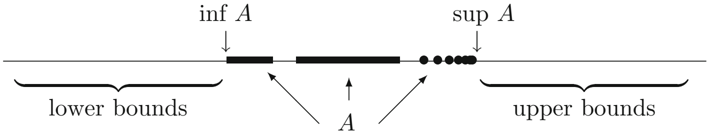


Figure 1.3: Definition of $\sup A$ and $\inf A$.

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 8.6.)

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 $s = \operatorname{lub} A$ is sometimes used, 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 1.3.1) 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.

Example 1.3.3. Let

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

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 results about how $\mathbf{Q}$ and $\mathbf{N}$ fit inside of $\mathbf{R}$. These are discussed in the next section. In the meantime, it is possible to prove some intuitive algebraic properties of least upper bounds just using the definition.

Example 1.3.7. Let $A \subseteq \mathbf{R}$ be nonempty and bounded above, and let $c \in \mathbf{R}$. Define the set $c + A$ by

$$c + A = \{c + a : a \in A\}.$$

Then $\sup(c + A) = c + \sup A$.

To properly verify this we focus separately on each part of Definition 1.3.2. Setting $s = \sup A$, we see that $a \leq s$ for all $a \in A$, which implies $c+a \leq c+s$ for all $a \in A$. Thus, $c + s$ is an upper bound for $c + A$ and condition (i) is verified.

For (ii), let b be an arbitrary upper bound for $c + A$; i.e., $c + a \leq b$ for all $a \in A$. This is equivalent to $a \leq b-c$ for all $a \in A$, from which we conclude that $b-c$ is an upper bound for A. Because s is the *least* upper bound of A, $s \leq b-c$, which can be rewritten as $c + s \leq b$. This verifies part (ii) of Definition 1.3.2, and we conclude $\sup(c + A) = c + \sup A$.

There is an equivalent and useful way of characterizing least upper bounds. As the previous example illustrates, 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 1.3.8. *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 $s \leq b$. □

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 nonempty bounded sets. One such proof is explored in Exercise 1.3.3.

Exercises

Exercise 1.3.1. (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 1.3.8 for greatest lower bounds.

Exercise 1.3.2. Give an example of each of the following, or state that the request is impossible.

(a) A set B with $\inf B \geq \sup B$.

(b) A finite set that contains its infimum but not its supremum.

(c) A bounded subset of $\mathbf{Q}$ that contains its supremum but not its infimum.

Exercise 1.3.3. (a) Let A be nonempty and 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 lower bounds exist as part of the Axiom of Completeness.

Exercise 1.3.4. Let $A_1, A_2, A_3, \ldots$ be a collection of nonempty sets, each of which is bounded above.

(a) Find a formula for $\sup(A_1 \cup A_2)$. Extend this to $\sup\left(\bigcup_{k=1}^{n} A_k\right)$.

(b) Consider $\sup\left(\bigcup_{k=1}^{\infty} A_k\right)$. Does the formula in (a) extend to the infinite case?

Exercise 1.3.5. As in Example 1.3.7, let $A \subseteq \mathbf{R}$ be nonempty and bounded above, and let $c \in \mathbf{R}$. This time define the set $cA = \{ca : a \in A\}$.

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

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

Exercise 1.3.6. Given sets A and B, define $A+B = \{a+b : a \in A \text{ and } b \in B\}$. Follow these steps to prove that if A and B are nonempty and bounded above then $\sup(A+B) = \sup A + \sup B$.

(a) Let $s = \sup A$ and $t = \sup B$. Show $s + t$ is an upper bound for $A + B$.

(b) Now let u be an arbitrary upper bound for $A + B$, and temporarily fix $a \in A$. Show $t \leq u - a$.

(c) Finally, show $\sup(A + B) = s + t$.

(d) Construct another proof of this same fact using Lemma 1.3.8.

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. Compute, without proofs, the suprema and infima (if they exist) of the following sets:

(a) $\{m/n : m, n \in \mathbf{N} \text{ with } m < n\}$.

(b) $\{(-1)^m/n : m, n \in \mathbf{N}\}$.

(c) $\{n/(3n+1) : n \in \mathbf{N}\}$.

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

Exercise 1.3.9. (a) If $\sup A < \sup B$, show that there exists an element $b \in B$ that is an upper bound for A.

(b) Give an example to show that this is not always the case if we only assume $\sup A \leq \sup B$.

Exercise 1.3.10 (Cut Property). The *Cut Property* of the real numbers is the following:

If A and B are nonempty, disjoint sets with $A \cup B = \mathbf{R}$ and $a < b$ for all $a \in A$ and $b \in B$, then there exists $c \in \mathbf{R}$ such that $x \leq c$ whenever $x \in A$ and $x \geq c$ whenever $x \in B$.

(a) Use the Axiom of Completeness to prove the Cut Property.

(b) Show that the implication goes the other way; that is, assume $\mathbf{R}$ possesses the Cut Property and let E be a nonempty set that is bounded above. Prove $\sup E$ exists.

(c) The punchline of parts (a) and (b) is that the Cut Property could be used in place of the Axiom of Completeness as the fundamental axiom that distinguishes the real numbers from the rational numbers. To drive this point home, give a concrete example showing that the Cut Property is not a valid statement when $\mathbf{R}$ is replaced by $\mathbf{Q}$.

Exercise 1.3.11. Decide if the following statements about suprema and infima are true or false. Give a short proof for those that are true. For any that are false, supply an example where the claim in question does not appear to hold.

(a) If A and B are nonempty, bounded, and satisfy $A \subseteq B$, then $\sup A \leq \sup B$.

(b) If $\sup A < \inf B$ for sets A and B, then there exists a $c \in \mathbf{R}$ satisfying $a < c < b$ for all $a \in A$ and $b \in B$.

(c) If there exists a $c \in \mathbf{R}$ satisfying $a < c < b$ for all $a \in A$ and $b \in B$, then $\sup A < \inf B$.

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*

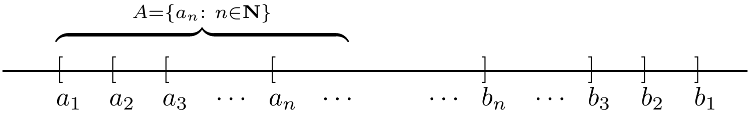
$$I_1 \supseteq I_2 \supseteq I_3 \supseteq I_4 \supseteq \cdots$$

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, especially in light of the fact that we have decided to take other familiar properties of $\mathbf{N}$, $\mathbf{Z}$, and $\mathbf{Q}$ as given.

The counterargument is that there is still a great deal of mystery about what the real numbers actually are. What we have said so far is that $\mathbf{R}$ is an extension of $\mathbf{Q}$ that maintains the algebraic and order properties of the rationals but also possesses the least upper bound property articulated in the Axiom of Completeness. In the absence of any other information about $\mathbf{R}$, we have to consider the possibility that in extending $\mathbf{Q}$ we unwittingly acquired some new numbers that are upper bounds for $\mathbf{N}$. In fact, as disorienting as it may sound, there *are* ordered field extensions of $\mathbf{Q}$ that include "numbers" bigger than every natural number. Theorem 1.4.2 asserts that the real numbers do not contain such exotic creatures. The Axiom of Completeness, which we adopted to patch up the holes in $\mathbf{Q}$, carries with it the implication that $\mathbf{N}$ is an unbounded subset of $\mathbf{R}$.

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 1.3.8), 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 $x = 1/y$. □

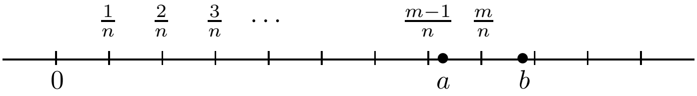
This familiar property of $\mathbf{N}$ is the key to an extremely important fact about how $\mathbf{Q}$ fits inside of $\mathbf{R}$.

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. A rational number is a quotient of integers, so we must produce $m \in \mathbf{Z}$ and $n \in \mathbf{N}$ 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 the Archimedean Property (Theorem 1.4.2), we may pick $n \in \mathbf{N}$ large enough so that

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

Inequality (1) (which we are trying to prove) is equivalent to $na < m < nb$. With n already chosen, the idea now is to choose m to be the smallest integer greater than na. In other words, pick $m \in \mathbf{Z}$ 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 1.4.5. □

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 1.4.7. □

A small modification of this proof can be made 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 $m \in \mathbf{N}$.

Exercises

Exercise 1.4.1. 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 1.4.2. Let $A \subseteq \mathbf{R}$ be nonempty and bounded above, and let $s \in \mathbf{R}$ have the property that for all $n \in \mathbf{N}$, $s + \frac{1}{n}$ is an upper bound for A and $s - \frac{1}{n}$ is not an upper bound for A. Show $s = \sup A$.

Exercise 1.4.3. 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 1.4.4. Let $a < b$ be real numbers and consider the set $T = \mathbf{Q} \cap [a, b]$. Show $\sup T = b$.

Exercise 1.4.5. Using Exercise 1.4.1, supply a proof for Corollary 1.4.4 by considering the real numbers $a - \sqrt{2}$ and $b - \sqrt{2}$.

Exercise 1.4.6. Recall that a set B is *dense* in $\mathbf{R}$ if an element of B can be found between any two real numbers $a < b$. Which of the following sets are dense in $\mathbf{R}$? Take $p \in \mathbf{Z}$ and $q \in \mathbf{N}$ in every case.

(a) The set of all rational numbers p/q with $q \leq 10$.

(b) The set of all rational numbers p/q with q a power of 2.

(c) The set of all rational numbers p/q with $10|p| \geq q$.

Exercise 1.4.7. Finish the proof of Theorem 1.4.5 by showing that the assumption $\alpha^2 > 2$ leads to a contradiction of the fact that $\alpha = \sup T$.

Exercise 1.4.8. Give an example of each or state that the request is impossible. When a request is impossible, provide a compelling argument for why this is the case.

(a) Two sets A and B with $A \cap B = \emptyset$, $\sup A = \sup B$, $\sup A \notin A$ and $\sup B \notin B$.

(b) A sequence of nested open intervals $J_1 \supseteq J_2 \supseteq J_3 \supseteq \cdots$ with $\bigcap_{n=1}^{\infty} J_n$ nonempty but containing only a finite number of elements.

(c) A sequence of nested unbounded closed intervals $L_1 \supseteq L_2 \supseteq L_3 \supseteq \cdots$ with $\bigcap_{n=1}^{\infty} L_n = \emptyset$. (An unbounded closed interval has the form $[a, \infty) = \{x \in R : x \geq a\}$.)

(d) A sequence of closed bounded (not necessarily nested) intervals $I_1, I_2, I_3, \ldots$ with the property that $\bigcap_{n=1}^{N} I_n \neq \emptyset$ for all $N \in \mathbf{N}$, but $\bigcap_{n=1}^{\infty} I_n = \emptyset$.

1.5 Cardinality

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 explore 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.

1–1 Correspondence

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 1.5.1. 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 1.5.2. The set A *has the same cardinality as* B if there exists $f : A \to B$ that is 1–1 and onto. In this case, we write $A \sim B$.

Example 1.5.3. (i) If we let $E = \{2, 4, 6, \ldots\}$ 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 $f(n) = 2n$.

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

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).

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

Example 1.5.4. 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 1.5.5. A set A is *countable* if $\mathbf{N} \sim A$. An infinite set that is not countable is called an *uncountable* set.

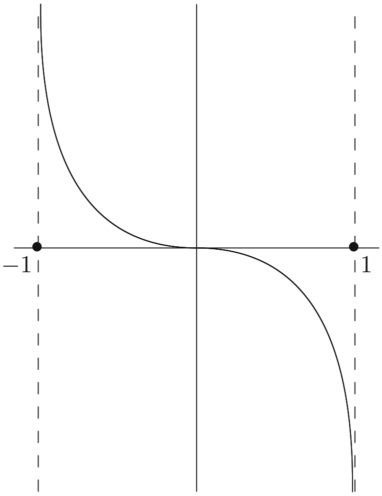


Figure 1.4: $(-1, 1) \sim \mathbf{R}$ USING $f(x) = x/(x^2 - 1)$.

From Example 1.5.3, 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 1.5.3, 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 1.5.6. (i) *The set* $\mathbf{Q}$ *is countable.* (ii) *The set* $\mathbf{R}$ *is uncountable.*

Proof. (i) Set $A_1 = \{0\}$ and for each $n \geq 2$, 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

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

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

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.

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

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 1.5.6 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

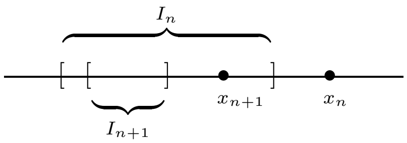
$$(1) \qquad \mathbf{R} = \{x_1, x_2, x_3, x_4, \ldots\}$$

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 1.5.6 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 1.5.7. *If $A \subseteq B$ and B is countable, then A is either countable or finite.*

Theorem 1.5.8. (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 1.5.1. Finish the following proof for Theorem 1.5.7.

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 1.5.2. Review the proof of Theorem 1.5.6, part (ii) showing that $\mathbf{R}$ is uncountable, and then find the flaw in the following erroneous proof that $\mathbf{Q}$ is uncountable:

Assume, for contradiction, that $\mathbf{Q}$ is countable. Thus we can write $\mathbf{Q} = \{r_1, r_2, r_3, \ldots\}$ and, as before, construct a nested sequence of closed intervals with $r_n \notin I_n$. Our construction implies $\bigcap_{n=1}^{\infty} I_n = \emptyset$ while NIP implies $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$. This contradiction implies $\mathbf{Q}$ must therefore be uncountable.

Exercise 1.5.3. Use the following outline to supply proofs for the statements in Theorem 1.5.8.

(a) First, prove statement (i) for two countable sets, A_1 and A_2. Example 1.5.3 (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 1.5.8 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 1.5.8 (ii).

Exercise 1.5.4. (a) Show $(a, b) \sim \mathbf{R}$ for any interval (a, b).

(b) Show that an unbounded interval like $(a, \infty) = \{x : x > a\}$ has the same cardinality as $\mathbf{R}$ as well.

(c) Using open intervals makes it more convenient to produce the required 1–1, onto functions, but it is not really necessary. Show that $[0, 1) \sim (0, 1)$ by exhibiting a 1–1 onto function between the two sets.

Exercise 1.5.5. (a) Why is $A \sim A$ for every set A?

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

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

Exercise 1.5.6. (a) Give an example of a countable collection of disjoint open intervals.

(b) Give an example of an uncountable collection of disjoint open intervals, or argue that no such collection exists.

Exercise 1.5.7. 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 . . . .)

The Schröder–Bernstein Theorem discussed in Exercise 1.5.11 can now be applied to conclude that $(0, 1) \sim S$.

Exercise 1.5.8. Let B be a set of positive real numbers with the property that adding together any finite subset of elements from B always gives a sum of 2 or less. Show B must be finite or countable.

Exercise 1.5.9. 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.

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

Exercise 1.5.10. (a) Let $C \subseteq [0, 1]$ be uncountable. Show that there exists $a \in (0, 1)$ such that $C \cap [a, 1]$ is uncountable.

(b) Now let A be the set of all $a \in (0, 1)$ such that $C \cap [a, 1]$ is uncountable, and set $\alpha = \sup A$. Is $C \cap [\alpha, 1]$ an uncountable set?

(c) Does the statement in (a) remain true if "uncountable" is replaced by "infinite"?

Exercise 1.5.11 (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 $X \sim Y$.

The strategy is to partition X and Y into components

$$X = A \cup A' \qquad \text{and} \qquad Y = B \cup B'$$

with $A \cap A' = \emptyset$ and $B \cap B' = \emptyset$, in such a way that f maps A onto B, and g maps B' onto A'.

(a) Explain how achieving this would lead to a proof that $X \sim Y$.

(b) Set $A_1 = X \backslash g(Y) = \{x \in X : x \notin g(Y)\}$ (what happens if $A_1 = \emptyset$?) and inductively define a sequence of sets by letting $A_{n+1} = g(f(A_n))$. Show that $\{A_n : n \in \mathbf{N}\}$ is a pairwise disjoint collection of subsets of X, while $\{f(A_n) : n \in \mathbf{N}\}$ is a similar collection in Y.

(c) Let $A = \bigcup_{n=1}^{\infty} A_n$ and $B = \bigcup_{n=1}^{\infty} f(A_n)$. Show that f maps A onto B.

(d) Let $A' = X \backslash A$ and $B' = Y \backslash B$. Show g maps B' onto A'.

1.6 Cantor's Theorem

Cantor's work into the theory of infinite sets extends far beyond the conclusions of Theorem 1.5.6. 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

Cantor published his discovery that $\mathbf{R}$ is uncountable in 1874. Although it has some modern polish on it, the argument presented in Theorem 1.5.6 (ii) is actually quite similar to the one Cantor originally found. In 1891, Cantor 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 1.6.1. *The open interval* $(0, 1) = \{x \in \mathbf{R} : 0 < x < 1\}$ *is uncountable.*

Exercise 1.6.1. Show that $(0, 1)$ is uncountable if and only if $\mathbf{R}$ is uncountable. This shows that Theorem 1.6.1 is equivalent to Theorem 1.5.6.

Proof. As with Theorem 1.5.6, 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

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

What is meant here is that for each $m, n \in \mathbf{N}$, a_{mn} is the digit from the set $\{0, 1, 2, \ldots, 9\}$ 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

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

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 $x = .b_1b_2b_3b_4 \ldots$ 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 1.6.2. (a) Explain why the real number $x = .b_1b_2b_3b_4 \ldots$ 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 1.6.3. Supply rebuttals to the following complaints about the proof of Theorem 1.6.1.

(a) Every rational number 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 1.6.1 must be flawed.

(b) Some 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 . . . . Doesn't this cause some problems?

Exercise 1.6.4. 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 1.6.5. (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.

Exercise 1.6.6. (a) Using the particular set $A = \{a, b, c\}$, exhibit two different 1–1 mappings from A into $P(A)$.

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

(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 1.6.6 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 1.6.2 (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 1.6.7. Return to the particular functions constructed in Exercise 1.6.6 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 1.6.8. (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. □

To get an initial sense of its broad significance, let's apply this result to the set of natural numbers. Cantor's Theorem states that there is no onto function from $\mathbf{N}$ to $P(\mathbf{N})$; in other words, the power set of the natural numbers is uncountable. How does the cardinality of this newly discovered uncountable set compare to the uncountable set of real numbers?

Exercise 1.6.9. Using the various tools and techniques developed in the last two sections (including the exercises from Section 1.5), give a compelling argument showing that $P(\mathbf{N}) \sim \mathbf{R}$.

Exercise 1.6.10. 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?

1.7 Epilogue

The relationship of having the same cardinality is an *equivalence relation* (see Exercise 1.5.5), meaning, roughly, that all of the sets in the mathematical 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 intervals (a, b) as well as $P(\mathbf{N})$. 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 not the case that $X \sim Y$. 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 1.5.11.

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 1.5.6 and 1.6.1 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 1.6.1 and 1.6.2. 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.