Roads to Infinity

The Mathematics of Truth and Proof

John Stillwell

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN A K PETERS BOOK

- CHAPTER 1 -

THE DIAGONAL ARGUMENT

PREVIEW

Infinity is the lifeblood of mathematics, because there is no end to even the simplest mathematical objects—the positive integers 1, 2, 3, 4, 5, 6, 7, ... One of the oldest and best arguments about infinity is Euclid's proof that the prime numbers 2, 3, 5, 7, 11, 13, ... form an infinite sequence. Euclid succeeds despite knowing virtually nothing about the sequence, by showing instead that any *finite* sequence of primes is incomplete. That is, he shows how to find a prime p different from any given primes $p_1, p_2, \ldots, p_n$.

A set like the prime numbers is called *countably infinite* because we can order its members in a list with a first member, second member, third member, and so on. As Euclid showed, the list is infinite, but each member appears at some finite position, and hence gets "counted."

Countably infinite sets have always been with us, and indeed it is hard to grasp infinity in any way other than by counting. But in 1874 the German mathematician Georg Cantor showed that infinity is more complicated than previously thought, by showing that the set of real numbers is *uncountable*. He did this in a way reminiscent of Euclid's proof, but one level higher, by showing that any countably infinite list of real numbers is incomplete.

Cantor's method finds a real number x different from any on a given countable list $x_1, x_2, x_3, \ldots$ by what is now called the *diagonal argument*, for reasons that will become clear below. The diagonal argument (which comes in several variations) is logically the simplest way to prove the existence of uncountable sets. It is the first "road to infinity" of our title, so we devote this chapter to it. A second road—via the *ordinals*—was also discovered by Cantor, and it will be discussed in Chapter 2.

1.1 Counting and Countability

> If I should ask further how many squares there are, one might reply truly that there are as many as the corresponding number of roots, since every square has its own root and every root has its own square, while no square has more than one root and no root more than one square.
>
> —Galileo Galilei,
> *Dialogues Concerning the Two New Sciences*, First day.

The process of counting 1, 2, 3, 4, ... is the simplest and clearest example of an *infinite process*. We know that counting never ends, because there is no last number, and indeed one's first thought is that "infinite" and "neverending" mean the same thing. Yet, in a sense, the endless counting process *exhausts* the set $\{1, 2, 3, 4, \ldots\}$ of positive integers, because each positive integer is eventually reached. This distinguishes the set of positive integers from other sets—such as the set of points on a line—which seemingly cannot be "exhausted by counting." Thus it may be enlightening to dwell a little longer on the process of counting, and to survey some of the infinite sets that can be exhausted by counting their members.

First, what do we mean by "counting" a set of objects? "Counting" objects is the same as arranging them in a (possibly infinite) *list*—first object, second object, third object, and so on—so that each object in the given set appears on the list, necessarily at some positive integer position. For example, if we "count" the squares by listing them in increasing order,

$$1,\ 4,\ 9,\ 16,\ 25,\ 36,\ 49,\ 64,\ 81,\ 100,\ 121,\ 144,\ 169,\ 196,\ 225,\ \ldots,$$

then the square 900 appears at position 30 on the list. Listing a set is mathematically the same as assigning the positive integers in some way to its members, but it is often easier to visualize the list than to work out the exact integer assigned to each member.

One of the first interesting things to be noticed about infinite sets is that *counting a part may be "just as infinite" as counting the whole.* For example, the set of *positive even numbers* 2, 4, 6, 8, ... is just a part of the set of positive integers. But the positive even numbers (in increasing order) form a list that matches the list of positive integers completely, item by item. Here they are:

1	2	3	4	5	6	7	8	9	10	11	12	13	...
2	4	6	8	10	12	14	16	18	20	22	24	26	...

Thus listing the positive even numbers is a process completely parallel to the process of listing the positive integers. The reason lies in the

item-by-item matching of the two lists, which we call a *one-to-one correspondence*. The function $f(n) = 2n$ encapsulates this correspondence, because it matches each positive integer n with exactly one positive even number $2n$, and each positive even number $2n$ is matched with exactly one positive integer n.

So, to echo the example of Galileo quoted at the beginning of this section: if I should ask how many even numbers there are, one might reply truly that there are as many as the corresponding positive integers. In both Galileo's example, and my more simple-minded one, one sees a one-to-one correspondence between the set of positive integers and a part of itself. This unsettling property is the first characteristic of the world of infinite sets.

Countably Infinite Sets

A set whose members can be put in an infinite list—that is, in one-to-one correspondence with the positive integers—is called *countably infinite*. This common property of countably infinite sets was called their *cardinality* by Georg Cantor, who initiated the general study of sets in the 1870s. In the case of finite sets, two sets have the same cardinality if and only if they have the same number of elements. So the concept of cardinality is essentially the *same* as the concept of number for finite sets.

For countably infinite sets, the common cardinality can also be regarded as the "number" of elements. This "number" was called a *transfinite number* and denoted $\aleph_0$ ("aleph zero" or "aleph nought") by Cantor. One can say, for instance, that there are $\aleph_0$ positive integers. However, one has to bear in mind that $\aleph_0$ is more elastic than an ordinary number. The sets $\{1,2,3,4,\ldots\}$ and $\{2,4,6,8,\ldots\}$ both have cardinality $\aleph_0$, even though the second set is a strict subset of the first. So one can also say that there are $\aleph_0$ even numbers.

Moreover, the cardinality $\aleph_0$ stretches to cover sets that at first glance seem much larger than the set $\{1,2,3,4,\ldots\}$. Consider the set of dots shown in Figure 1.1. The grid has infinitely many infinite rows of dots, but nevertheless we can pair each dot with a different positive integer as shown in the figure. Simply view the dots along a series of finite diagonal lines, and "count" along the successive diagonals, starting in the bottom left corner.

There is a very similar proof that the set of (positive) fractions is countable, since each fraction m/n corresponds to the pair (m,n) of positive integers. It follows that the set of positive rational numbers is countable, since each positive rational number is given by a fraction. Admittedly, there are many fractions for the same number—for example the number $1/2$ is also given by the fractions $2/4$, $3/6$, $4/8$, and so on—-but we can

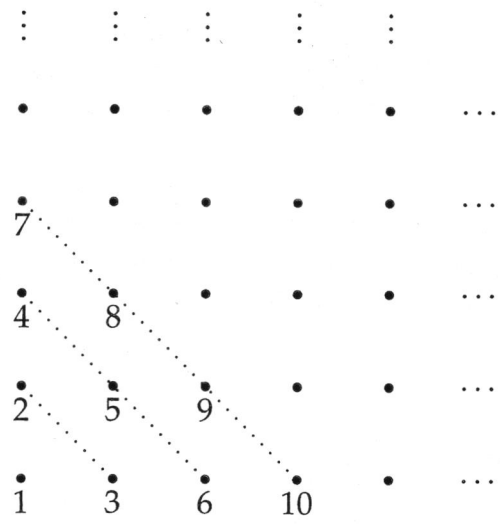

Figure 1.1. Counting the dots in an infinite grid.

list the positive rational numbers by going through the list of fractions and omitting all fractions that represent previous numbers on the list.

1.2 Does One Infinite Size Fit All?

A nice way to illustrate the elasticity of the cardinality $\aleph_0$ was introduced by the physicist George Gamow (1947) in his book *One, Two, Three, ..., Infinity*. Gamow imagines a hotel, called *Hilbert's hotel*, in which there are infinitely many rooms, numbered 1, 2, 3, 4, ... Listing the members of an infinite set is the same as accommodating the members as "guests" in Hilbert's hotel, one to each room.

The positive integers can naturally be accommodated by putting each number n in room n (Figure 1.2):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	...

Figure 1.2. Standard occupancy of Hilbert's hotel.

The $\aleph_0$ positive integers fill every room in Hilbert's hotel, so we might say that $\aleph_0$ is the "size" of Hilbert's hotel, and that occupancy by more than $\aleph_0$ persons is unlawful. Nevertheless there is room for one more (say, the number 0). Each guest simply needs to move up one room, leaving the first room free (Figure 1.3):

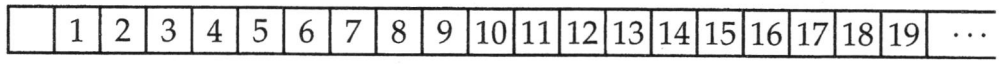

Figure 1.3. Making room for one more.

Thus $\aleph_0$ can always stretch to include one more: in symbols, $\aleph_0 + 1 = \aleph_0$. In fact, there is room for another countable infinity of "guests" (say, the negative integers $-1, -2, -3, \ldots$). The guest in room n can move to room $2n$, leaving all the odd numbered rooms free (Figure 1.4):

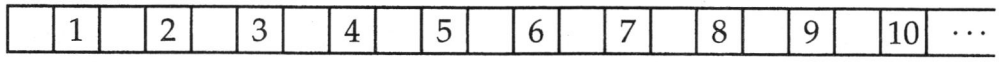

Figure 1.4. Making room for a countable infinity more.

In symbols: $\aleph_0 + \aleph_0 = \aleph_0$.

There is even room for a countable infinity of countable infinities of guests. Suppose, say, that the guests arrive on infinite buses numbered 1, 2, 3, 4, ..., and that each bus has guests numbered 1, 2, 3, 4 , ... The guests in bus 1 can be accommodated as follows:

> put guest 1 in room 1; then skip 1 room; that is,
> put guest 2 in room 3; then skip 2 rooms; that is,
> put guest 3 in room 6; then skip 3 rooms; that is,
> put guest 4 in room 10; then skip 4 rooms; ... .

Thus the first bus fills the rooms shown in Figure 1.5:

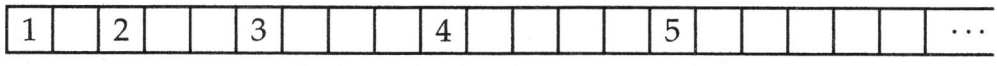

Figure 1.5. Making room for a countable infinity of countable infinities.

After the first bus has been unloaded, the unoccupied rooms are in blocks of 1, 2, 3, 4, ... rooms, so we can unload the second bus by putting its guests in the leftmost room of each block. After that, the unoccupied rooms are *again* in blocks of 1, 2, 3, 4, ... rooms, so we can repeat the process with the third bus, and so on. (You may notice that each busload occupies a sequence of rooms numbered the same as a row in Figure 1.1.)

The result is that the whole series of $\aleph_0$ busloads, each with $\aleph_0$ guests, can be packed into Hilbert's hotel—with exactly one guest per room. In symbols: $\aleph_0 \times \aleph_0 = \aleph_0$.

The equations of "cardinal arithmetic" just obtained,

$$\aleph_0 + 1 = \aleph_0,$$
$$\aleph_0 + \aleph_0 = \aleph_0,$$
$$\aleph_0 \times \aleph_0 = \aleph_0,$$

show just how elastic the transfinite number $\aleph_0$ is. So much so, one begins to suspect that cardinal arithmetic has nothing to say except that any infinite set has cardinality $\aleph_0$. And if all transfinite numbers are the same it is surely a waste of time to talk about them. But fortunately they are *not* all the same. In particular, the set of points on the line has cardinality strictly *larger* than $\aleph_0$. Cantor discovered this difference in 1874, opening a crack in the world of the infinite from which unexpected consequences have spilled ever since. There is, after all, a lot to say about infinity, and the purpose of this book is to explain why.

1.3 CANTOR'S DIAGONAL ARGUMENT

Before studying the set of points on the line, we look at a related set that is slightly easier to handle: the set of all sets of positive integers. A set S of positive integers can be described by an infinite sequence of 0s and 1s, with 1 in the nth place just in case n is a member of S. Table 1.1 shows a few examples:

subset	1	2	3	4	5	6	7	8	9	10	11	...
even numbers	0	1	0	1	0	1	0	1	0	1	0	...
squares	1	0	0	1	0	0	0	0	1	0	0	...
primes	0	1	1	0	1	0	1	0	0	0	1	...

Table 1.1. Descriptions of positive integer sets.

Now suppose that we have $\aleph_0$ sets of positive integers. That means we can form a list of the sets, $S_1, S_2, S_3, \ldots$, whose nth member S_n is the set paired with integer n. We show that such a list can never include *all* sets of positive integers by describing a set S different from each of $S_1, S_2, S_3, \ldots$.

This is easy: for each number n, put n in S just in case n is *not* in S_n. It follows that S *differs from each* S_n *with respect to the number* n: *if* n *is* S_n, *then* n *is not in* S; *if* n *is not* S_n, *then* n *is in* S. Thus S is not on the list $S_1, S_2, S_3, \ldots$, and hence no such list can include all sets of positive integers.

subset	1	2	3	4	5	6	7	8	9	10	11	...
S_1	**0**	1	0	1	0	1	0	1	0	1	0	...
S_2	1	**0**	0	1	0	0	0	0	1	0	0	...
S_3	0	1	**1**	0	1	0	1	0	0	0	1	...
S_4	1	0	1	**0**	1	0	1	0	1	0	1	...
S_5	0	0	1	0	**0**	1	0	0	1	0	0	...
S_6	1	1	0	1	1	**0**	1	1	0	1	1	...
S_7	1	1	1	1	1	1	**1**	1	1	1	1	...
S_8	0	0	0	0	0	0	0	**0**	0	0	0	...
S_9	0	0	0	0	0	0	0	0	**1**	0	0	...
S_{10}	1	0	0	1	0	0	1	0	0	**1**	0	...
S_{11}	0	1	0	0	1	0	0	1	0	0	**0**	...
$\vdots$												
S	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	...

Table 1.2. The diagonal argument.

The argument we have just made is called a *diagonal* argument because it can be presented visually as follows. Imagine an infinite table whose rows encode the sets $S_1, S_2, S_3, \ldots$ as sequences of 0s and 1s, as in the examples above. We might have, say, the sets shown in Table 1.2.

The digit (1 or 0) that signals whether or not n belongs to S_n is set in bold type, giving a diagonal sequence of bold digits

$$\mathbf{00100010110}\ldots.$$

The sequence for S is obtained by switching each digit in the diagonal sequence. Hence the sequence for S is necessarily different from the sequences for all of $S_1, S_2, S_3, \ldots$.

The cardinality of the set of all sequences of 0s and 1s is called $2^{\aleph_0}$. We use this symbol because there are two possibilities for the first digit in the sequence, two possibilities for the second digit, two possibilities for the third, and so on, for all the $\aleph_0$ digits in the sequence. Thus it is reasonable to say that there are $2 \times 2 \times 2 \times \cdots$ ($\aleph_0$ factors) possible sequences of 0s and 1s, and hence there are $2^{\aleph_0}$ sets of positive natural numbers.

The diagonal argument shows that $2^{\aleph_0}$ *is strictly greater than* $\aleph_0$ because there is a one-to-one correspondence between the positive integers and certain sets of positive integers, but not with *all* such sets. As we have just seen, if the numbers 1, 2, 3, 4, ... are assigned to sets $S_1, S_2, S_3, S_4, \ldots$ there will always be a set (such as S) that fails to be assigned a number.

The Logic of the Diagonal Argument

> Many mathematicians aggressively maintain that there can be no doubt of the validity of this proof, whereas others do not admit it. I personally cannot see an iota of appeal in this proof ... my mind will not do the things that it is obviously expected to do if this is indeed a proof.
>
> —P .W. Bridgman (1955), p. 101

P. W. Bridgman was an experimental physicist at Harvard, and winner of the Nobel prize for physics in 1946. He was also, in all probability, one of the smartest people *not* to understand the diagonal argument. If you had any trouble with the argument above, you can rest assured that a Nobel prize winner was equally troubled. On the other hand, I do not think that any mathematically experienced reader *should* have trouble with the diagonal argument. Here is why.

The logic of the diagonal argument is really very similar to that of Euclid's proof that there are infinitely many primes. Euclid faced the difficulty that the totality of primes is hard to comprehend, since they follow no apparent pattern. So, he avoided even considering the totality of primes by arguing instead that *any finite list of primes is incomplete*...

Given a finite list of primes $p_1, p_2, \ldots, p_n$, one forms the number

$$N = p_1 p_2 \cdots p_n + 1,$$

which is obviously not divisible by any one of $p_1, p_2, \ldots, p_n$ (they each leave remainder 1). But N is divisible by *some* prime number, so the list $p_1, p_2, \ldots, p_n$ of primes is incomplete. Moreover, we can find a specific prime p not on the list by finding the smallest number ≥ 2 that divides N.

An uncountable set is likewise very hard to comprehend, so we avoid doing so and instead suppose that we are given a countable list $S_1, S_2, S_3, \ldots$ of members of the set. The word "given" may be interpreted as strictly as you like. For example, if $S_1, S_2, S_3, \ldots$ are sequences of 0s and 1s, you may demand a *rule* that gives the mth digit of S_n at stage $m + n$. The diagonal argument still works, and it gives a *completely specific* S not on the given list. (Indeed, it also leads to some interesting conclusions about rules for computing sequences, as we will see in Chapter 3.)

The Set of Points on the Line

> The goal of set theory is to answer the question of highest importance: whether one can view the line in an atomistic manner, as a set of points.
>
> —Nikolai Luzin (1930), p. 2.

By the "line" we mean the number line, whose "points" are known as the *real numbers*. Each real number has a *decimal expansion* with an infinite

sequence of decimal digits after the decimal point. For example,

$$\pi = 3.14159265358979323846\ldots.$$

If we stick to real numbers between 0 and 1, then each number is given by the sequence of digits after the decimal point, each term of which is one of 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. A countable infinity of real numbers $x_1, x_2, x_3, x_4, \ldots$ between 0 and 1 can therefore be encoded by a table very like the table we used to encode a countable infinity of sets of positive integers. Likewise, we can construct a real number x *different* from each of $x_1, x_2, x_3, x_4, \ldots$ by arranging that the nth digit of x is different from the nth digit of x_n, for each positive integer n. As before, this amounts to looking at the diagonal digits in the table, and changing each one of them.

However, there is now a slight problem with the diagonal construction. Changing each diagonal digit certainly produces a sequence of digits different from all the sequences given for $x_1, x_2, x_3, x_4, \ldots$. But this does not ensure that the new sequence represents a new number. It could happen, for example, that the sequence obtained by the diagonal construction is

$$0 \cdot 49999999999999\cdots$$

and that one of the given sequences is

$$x_1 = 0 \cdot 50000000000000\cdots$$

These are different sequences, but they represent the same number, namely $1/2$. Since two sequences can represent the same number only if one of them ends in an infinite sequence of 9s, we can avoid this problem by never changing a diagonal digit to a 0 or a 9. For example, we could use the following rule.

> If the nth digit of x_n is 1, let the nth digit of x be 2.
> If the nth digit of x_n is not 1, let the nth digit of x be 1.

With this rule, x does not merely have a sequence of digits different from those for $x_1, x_2, x_3, \ldots$. *As a number*, x is different from $x_1, x_2, x_3, \ldots$. Thus we have proved that *the set of real numbers is of greater cardinality than the set of positive natural numbers.* If we make a list of real numbers $x_1, x_2, x_3, x_4, \ldots$ there will always be a real number (such as x) not on the list.

In fact, the set of real numbers (whether between 0 and 1 or over the whole number line) has cardinality $2^{\aleph_0}$—the same as that of the set of sequences of 0s and 1s. The cardinality $2^{\aleph_0}$, like $\aleph_0$, measures the "size" among familiar sets in mathematics. The reasons for this will become clearer as we explore further examples of countable and uncountable sets.

1.4 Transcendental Numbers

> [On the question whether the real numbers can be made to correspond with the natural numbers]... if it could be answered with no, then one would have a new proof of Liouville's theorem that there are transcendental numbers.
>
> —Georg Cantor,
> Letter to Dedekind, 2 December 1873,
> in Ewald (1996), p. 844.

Cantor's discovery of uncountable sets in 1874 was one of the most unexpected events in the history of mathematics. Before 1874, infinity was not even considered a legitimate mathematical concept by most people, so the need to distinguish between countable and uncountable infinities could not have been imagined. The idea of uncountability was just *too* original to be appreciated by most mathematicians. Because of this, Cantor downplayed the idea of uncountability in the published version of his discovery, introducing it only indirectly via a "property of the algebraic numbers."

The algebraic numbers were very familiar to 19th-century mathematicians. A number x is called *algebraic* if it satisfies a polynomial equation with integer coefficients, that is, an equation of the form

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

where $a_0, a_1, \ldots, a_{n-1}, a_n$ are integers.

The algebraic numbers include all the rational numbers m/n, where m and n are integers, and also many irrational numbers, such as $\sqrt{2}$, $\sqrt{3}$, $\sqrt[3]{2}$, and so on. In fact, it is quite hard to find any numbers that are *not* algebraic. The first example was found by the French mathematician Joseph Liouville in 1844, using a result that says (roughly speaking) that an irrational algebraic number cannot be closely approximated by rationals.

Liouville then considered the number

$$x = 0 \cdot 1010010000001000000000000000000000000010\ldots,$$

where the successive blocks of 0s have lengths 1, 2, 3×2, $4 \times 3 \times 2$, ... He showed that x is irrational but that it *can* be closely approximated by rationals, hence x is not algebraic. Such numbers are now called *transcendental*, because they transcend description by algebraic means. Invariably, they have to be described by infinite processes, such as infinite decimals or infinite series. The first "well-known" number to be proved transcendental was

$$e = 1 + \frac{1}{1} + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} + \frac{1}{1 \times 2 \times 3 \times 4} + \cdots,$$

a number which makes many appearances in mathematics, perhaps most famously in the equation

$$e^{\pi\sqrt{-1}} = -1.$$

Liouville's compatriot Charles Hermite proved that e is transcendental in 1873, using some difficult calculus. In fact, Hermite was so exhausted by the effort that he gave up the idea of trying to prove that π is transcendental. The transcendance of π was first proved by the German mathematician Ferdinand Lindemann in 1882, using Hermite's methods and the above equation connecting e and π.

At any rate, in 1874 the only known approaches to transcendental numbers were those of Liouville and Hermite, using sophisticated algebra and calculus. Cantor surprised the world by showing the existence of transcendental numbers without any advanced math at all—simply by proving that *the set of algebraic numbers is countable*. Combining this with his result that real numbers are uncountable, it follows that some real numbers are transcendental. In fact, "most" real numbers must be transcendental. Not only does set theory find transcendental numbers easily, it also shows that the handful of transcendental numbers previously known actually belong to the vast, uncountable, majority.

Since we have already seen one of Cantor's proofs that there are uncountably many reals, it remains to explain why there are only countably many algebraic numbers. To do this we come back to the equations that define algebraic numbers:

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

where $a_0, a_1, \ldots, a_{n-1}, a_n$ are integers. The equation (1) has at most n solutions, as one learns in elementary algebra, so we can list all algebraic numbers if we can list all equations of the form (1). To do this, Cantor used a quantity called the *height* of the equation,

$$h = |a_n| + |a_{n-1}| + \cdots + |a_0| + n,$$

suggested by his colleague Richard Dedekind. It is not hard to see that *there are only finitely many equations with a given height* h, since they necessarily have degree $n \leq h$ and each coefficient has absolute value less than h. Therefore, we can list all equations (and hence all algebraic numbers) by first listing the equations of height 1, then those of height 2, those of height 3, and so on.

This listing process shows that the algebraic numbers form a countable set, and hence we are done.

1.5 Other Uncountability Proofs

We have not described Cantor's first uncountability proof, from 1874, because it is more complicated than his diagonal proof, which dates from 1891. The logic of the 1874 proof is basically the same—a countable set of numbers does not include all numbers because we can always find a number outside it—but the construction of the outsider x is not obviously "diagonal." Rather, x is a *least upper bound* of a certain increasing sequence of numbers $x_1, x_2, x_3, x_4, \ldots$; that is, x is the least number greater than all of $x_1, x_2, x_3, x_4, \ldots$. However, the least upper bound begins to look "diagonal" when one studies the decimal digits of $x_1, x_2, x_3, x_4, \ldots$.

Suppose, for the sake of example, that the numbers $x_1, x_2, x_3, x_4, \ldots$ have the following decimal digits:

$$\begin{aligned} x_1 &= 1.\mathbf{4}13\ldots \\ x_2 &= 1.4\mathbf{1}41\ldots \\ x_3 &= 1.414\mathbf{2}32\ldots \\ x_4 &= 1.4142\mathbf{3}5621\ldots \\ x_5 &= 1.41423562\mathbf{3}5\ldots \\ x_6 &= 1.4142356\mathbf{2}37\ldots \\ &\vdots \end{aligned}$$

Because $x_1 < x_2 < x_3 < \cdots$, each decimal x_{i+1} agrees with its predecessor x_i up to a certain digit, and x_{i+1} has a *larger* digit than x_i at the first place where they disagree. These digits of first disagreement (shown in bold type) form a "jagged diagonal." And the least upper bound x of the sequence $x_1, x_2, x_3, x_4, \ldots$ is obtained by uniting all the decimal segments up to this "diagonal":

$$x = 1.4142356237\ldots.$$

Thus there is a sense in which Cantor's original uncountability proof involves a diagonal construction.

The same is true of a remarkable proof discovered by the German mathematician Axel Harnack in 1885, using the concept of *measure*. Suppose that $a_1, a_2, a_3, a_4, \ldots$ is any list of real numbers. Harnack observes that we can cover all of these numbers by line segments of total length as small as we please, say ε. Just take a line segment of length ε, break it in half, and use a segment of length $\varepsilon/2$ to cover the number a_1. Then break the remaining segment of length $\varepsilon/2$ in half and use a segment of length $\varepsilon/4$ to cover a_2, and so on. Thus we cover

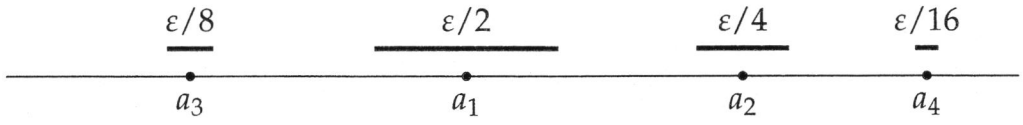

Figure 1.6. Harnack's covering of a countable set.

$$\begin{gathered} a_1 \text{ by an interval of length } \varepsilon/2, \\ a_2 \text{ by an interval of length } \varepsilon/4, \\ a_3 \text{ by an interval of length } \varepsilon/8, \\ a_4 \text{ by an interval of length } \varepsilon/16, \\ \dots, \end{gathered}$$

as shown in Figure 1.6. And *the whole infinite list* $a_1, a_2, a_3, a_4, \dots$ *is covered by line segments of total length at most* ε.

It follows that the numbers $a_1, a_2, a_3, a_4, \dots$ do not include all real numbers. In fact, far from filling the whole line, the set of numbers $a_1, a_2, a_3, a_4, \dots$ has total length zero! Thus no list exhausts the set of real numbers, or even comes close. This proof shows more dramatically why there are uncountably many real numbers, but it does not seem to yield any *particular* number not in the given list $a_1, a_2, a_3, a_4, \dots$. This defect is easily fixed, by none other than the diagonal construction. (It is also convenient to modify the lengths of the covering intervals in order to suit decimal notation.)

Suppose that the numbers $a_1, a_2, a_3, a_4, \dots$ are given as infinite decimals, say

$$\begin{aligned} a_1 &= 1 \cdot 73205\dots, \\ a_2 &= 0 \cdot 11111\dots, \\ a_3 &= 3 \cdot 14159\dots, \\ a_4 &= 0 \cdot 99999\dots, \\ &\vdots \end{aligned}$$

If we cover a_1 by an interval of length $1/10$ then numbers x that differ from a_1 in the first decimal place by at least 2 are definitely *outside* the first covering interval. We can change the first digit in a_1 to 5, for example, so $x = 0 \cdot 5\dots$ is outside the first interval. Next, if we cover a_2 by an interval of length $1/100$, then numbers x that differ from a_2 in the second decimal place by at least 2 are definitely outside the second covering interval. For example, we could change the second digit in a_2 to 3, so $x = 0 \cdot 53\dots$ is outside the first and second intervals. Similarly, if we cover a_3 by an interval of length $1/1000$, then $x = 0 \cdot 533\dots$ is outside the first, second, and third intervals.

It is clear that we can continue in this way to cover each real number a_n by an interval of length $1/10^n$, and at the same time find a number x outside all the intervals by choosing the nth digit of x to be suitably different from the nth digit of a_n. This is clearly a diagonal construction.

Thus it may be that Cantor distilled the diagonal argument from previous uncountability proofs. Indeed, a more explicitly "diagonal" construction had already been described in 1875 by another German mathematician with an interest in the infinite, Paul du Bois-Reymond. We discuss his work in the next section.

1.6 RATES OF GROWTH

From ancient times, when Archimedes tried to estimate the number of grains of sand in the universe, until today, when "exponential growth" has become a cliché, people have been fascinated by large numbers and rapid rates of growth. With modern mathematical notation it is quite easy to describe functions or sequences with extravagant growth and, with somewhat greater difficulty, to compare growth rates of different functions.

For example, the function $f(n) = n$, whose sequence of values is

$$1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6, \quad \ldots,$$

is a function that grows beyond all bounds. But it does not grow as fast as the function $g(n) = n^2$, whose values form the sequence of squares:

$$1, \quad 4, \quad 9, \quad 16, \quad 25, \quad 36, \quad \ldots .$$

We can see why by looking at $g(n)/f(n) = n$, which itself grows beyond all bounds, or *tends to infinity*, as we usually say. In general, let us agree to say that a function $G(n)$ *grows faster* than a function $F(n)$ if $G(n)/F(n)$ tends to infinity. We write this relationship symbolically as

$$G(n)/F(n) \to \infty \quad \text{as} \quad n \to \infty.$$

It follows immediately that n^3 grows faster than n^2, n^4 grows faster than n^3, and so on. Thus the infinite family of functions $\{n, n^2, n^3, n^4, \ldots\}$ represents a family of growth rates with *no greatest member*. It is called the family of *polynomial* growth rates.

Although there is no greatest polynomial rate of growth, there is a growth rate greater than all polynomial rates. Consider, for example, the growth rate of the function 2^n. This function has exponential growth, and it grows faster than n^k for any fixed k. We shall not pause to prove

this fact, because there is another function that beats all of $n, n^2, n^3, n^4, \ldots$ more obviously; namely, the function

$$d(n) = n^n.$$

It is clear that

$$\begin{aligned} n^n &> n^2 \quad \text{for all } n > 2, \\ n^n &> n^3 \quad \text{for all } n > 3, \\ n^n &> n^4 \quad \text{for all } n > 4, \\ &\text{and so on.} \end{aligned}$$

Thus, for any k, $n^n > n^{k+1}$ for all sufficiently large n. It follows that $n^n/n^k \to \infty$ as $n \to \infty$, because we already know that $n^{k+1}/n^k \to \infty$ as $n \to \infty$.

The values of $d(n) = n^n$ are nothing but the *diagonal values* in the table of values of the functions $n, n^2, n^3, n^4, \ldots$:

$$\begin{aligned} d(1) &= 1^1 \text{ is the first value of } n, \\ d(2) &= 2^2 \text{ is the second value of } n^2, \\ d(3) &= 3^3 \text{ is the third value of } n^3, \\ &\text{and so on.} \end{aligned}$$

A similar idea applies to any sequence of functions. This is essentially what du Bois-Reymond discovered in 1875, so we name the result after him.

THEOREM OF DU BOIS-REYMOND. *If $f_1, f_2, f_3, \ldots$ is a list of positive integer functions, then there is a positive integer function that grows faster than any f_i.*

Proof: Since all function values are positive, we have

$$f_1(n) + f_2(n) + \cdots + f_n(n) > f_i(n) \quad \text{for each } i \leq n.$$

The function f defined by $f(n) = f_1(n) + f_2(n) + \cdots + f_n(n)$, therefore satisfies $f(n) \geq f_i(n)$ for all $n > i$. That is, f grows *at least as fast* as any function f_i.

Consequently, if we define a function d by

$$d(n) = nf(n),$$

then $d(n)/f(n) \to \infty$ as $n \to \infty$, so d grows faster than any function f_i.□

It follows that no list of functions grows fast enough to "overtake" every function of natural numbers. This was what interested du Bois-Reymond, and it is indeed a very pregnant discovery, to which we will

return later. But notice also that we again have a diagonal argument for uncountability. (We regard the function $d(n) = nf(n)$ as "diagonal" because $f(n)$ is the sum of entries in or above the diagonal in the table of function values for $f_1, f_2, f_3, \ldots$.)

The set of positive integer functions is uncountable, because *for any list of such functions $f_1(n), f_2(n), f_3(n), \ldots$ there is a function $d(n)$ not in the list.*

1.7 The Cardinality of the Continuum

So far we have found three uncountable sets: the set of real numbers, the set of subsets of the positive integers, and the set of functions of positive integers with positive integer values. In a certain sense these three are essentially the *same* set, so it is not so surprising that in each case their uncountability follows by a similar argument.

Like the several countable sets discussed in Section 1.1, these three uncountable sets have the same cardinality. We called it $2^{\aleph_0}$ in Section 1.3, and it is also called the *cardinality of the continuum*, since the continuum of real numbers is the most concrete set with cardinality $2^{\aleph_0}$. We can visualize the totality of all real numbers as a continuous line—the "number line"—but it is quite hard to visualize the totality of sets of positive integers, say, until these sets have been matched up with real numbers.

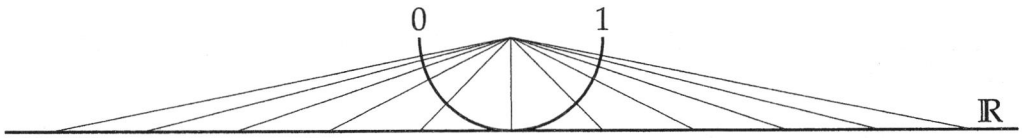

Figure 1.7. One-to-one correspondence between $\mathbb{R}$ and the interval $(0,1)$.

Before establishing a one-to-one correspondence between real numbers and sets of positive integers, we first observe that *there is a one-to-one correspondence between the set $\mathbb{R}$ of all real numbers and the interval $(0,1)$ of real numbers between 0 and 1.* This is geometrically obvious if one bends the line segment between 0 and 1 into a semicircle and then projects the semicircle onto the number line $\mathbb{R}$ as shown in Figure 1.7.

Correspondence between Real Numbers and Sets

A number in the interval $(0,1)$ and a set of natural numbers have a certain notational similarity. Namely, they both have natural descriptions as sequences of 0s and 1s. We have already seen how to encode a set of natural numbers by such a sequence in Section 1.2.

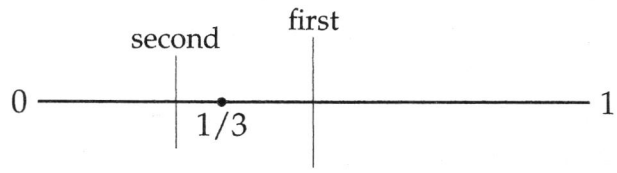

Figure 1.8. First two bisections for $1/3$.

To encode a number x in $(0,1)$ by a sequence of 0s and 1s we make successive bisections of the interval, each time choosing the subinterval in which x lies. If x lies in the left half, write down 0; if the right half, write down 1. Then bisect the subinterval in which x lies, and repeat the process. The resulting sequence of 0s and 1s is called the *binary expansion* of x. For example, Figure 1.8 shows how we find the binary expansion of $1/3$.

After the first bisection, $1/3$ lies in the left half, so we write down 0. After bisecting the left half, $1/3$ lies in the right half, so we write down 1. In the resulting quarter interval, $1/3$ is in the same position as at the beginning; namely, $1/3$ of the way from the left hand end of the interval. Therefore, if we continue the bisection process, we will write down 0 then 1 then 0 then 1, and so on, forever. Thus $1/3$ has the infinite binary expansion

$$\cdot 01010101010101010101010\ldots$$

(where the dot is a "binary point" instead of the decimal point). Ambiguity enters when x falls on a line of bisection; for example, when $x = 1/2$ or $x = 1/4$. In that case, we can assign x to either the left or right subinterval when the line of subdivision hits x, but thereafter we have no choice. If we write 0—thus assigning x to the left half—then x will lie in the right half of every subinterval constructed thereafter, and the 0 will be followed by an infinite sequence of 1s. For example, $1/2$ has the binary expansion

$$\cdot 01111111111111111111111111\ldots.$$

But if we choose 1 at the beginning, then every digit thereafter is 0, so the other binary expansion of $1/2$ is

$$\cdot 10000000000000000000000\ldots.$$

(This is analogous to the ambiguity in decimal expansions, where $1/2$ can be written as both $\cdot 499999999\ldots$ and $\cdot 500000000\ldots$.)

In general, each *binary fraction* $p/2^q$ in $(0,1)$ corresponds to two different sequences of 0s and 1s. Both sequences have the same initial segment, in one case followed by $10000\ldots$ and in the other case followed

by 01111.... Thus each binary fraction corresponds to two different sets of natural numbers. Fortunately, there are only countably many binary fractions (for example, because they form a subset of the set of rationals) so this breakdown of the one-to-one correspondence is easily fixed.

In fact we have already seen, in Section 1.1, how countable sets in two-to-one correspondence are also in one-to-one correspondence (consider the set of all natural numbers and the set of all even numbers). Thus the two-to-one correspondence between sequences ending in 10000... or 01111... and binary fractions can be rearranged into a one-to-one correspondence. Combining this with the one-to-one correspondence between the remaining sequences of 0s and 1s and the remaining numbers between 0 and 1 gives the required one-to-one correspondence between sets of natural numbers and real numbers between 0 and 1.

Correspondence between Functions and Real Numbers

Each function f on the set of positive integers, with positive integer values, corresponds to a real number between 0 and 1 as follows. First write down the sequence of values of f, say,

$$4, \quad 2, \quad 6, \quad 1, \quad 1, \quad 8, \quad 3, \quad 5, \quad \ldots \quad .$$

Next, encode this sequence by a sequence of 0s and 1s, replacing each positive integer $f(n)$ by a block of $f(n) - 1$ (hence possibly zero) 1s, and each comma by a 0:

$$11101011111000111111011011110\ldots.$$

Finally, insert a binary point in front to make this sequence the binary expansion of a real number. The expansion is necessarily one with infinitely many 0s, because the function f has infinitely many values. But this is fine, because it means we omit all binary expansions ending in 1111..., and hence we get each real number between 0 and 1 at most once.

Conversely, each real between 0 and 1 has a unique binary expansion with infinitely many 0s, and there is a unique function f encoded by this expansion. For example, if the binary expansion is

$$\cdot 00101100111110110101010101\ldots.$$

then the successive values of f are

$$1, \quad 1, \quad 2, \quad 3, \quad 1, \quad 6, \quad 3, \quad 2, \quad 2, \quad 2, \quad 2, \quad \ldots \quad .$$

In general,

$f(1) =$ (number of 1s before the first 0) $+ 1$,
$f(2) =$ (number of 1s after the first 0 and before the second 0) $+ 1$,
$f(3) =$ (number of 1s after the second 0 and before the third 0) $+ 1$,
and so on.

Thus binary expansions give a one-to-one correspondence between positive integer functions and real numbers between 0 and 1.

1.8 HISTORICAL BACKGROUND

INFINITY IN ANCIENT GREECE

> Zeno's argument makes a false assumption in asserting that it is impossible for a thing to pass over or severally to meet with infinite[ly many] things in finite time.
>
> —Aristotle
> *Physics*, Book VI, Chap. 2.

Since ancient times, infinity has been a key part of mathematics, though its use has often been considered harmful. Around 500 BCE, the Pythagoreans discovered the irrationality of $\sqrt{2}$, thus beginning a long struggle to grasp (what we now call) the concept of real number. This was part of a larger struggle to reconcile the continuous magnitudes arising in geometry—length, area, volume—with the discrete natural numbers $1, 2, 3, 4, \ldots$ arising from counting.

The shock of irrationality apparently left the Greeks unwilling to treat continuous magnitudes as numbers. For example, they viewed the product of two lengths as a rectangle, the product of three lengths as a box, and the product of four lengths as having no meaning at all. Nevertheless, they made great progress in *relating* continuous magnitudes to the natural numbers. Around 350 BCE, Eudoxus introduced a "theory of proportion" (known to us from Book V of Euclid's *Elements*), which was as close as anyone came to defining real numbers before the 19th century. Euclid characterized irrational lengths l as those for which a certain process (the euclidean algorithm on l and 1) is infinite. Archimedes found areas and volumes of curved figures by infinite processes, even breaking the taboo against "actual" infinite sets in some of his methods.

Almost all mathematicians until the 19th century made a sharp distinction between what they called the *potential* and *actual* infinite. The potential infinite is typically an *unending process*, such as counting $1, 2, 3, \ldots$ or cutting a curved region into smaller and smaller triangles. The actual

infinite is the supposed *completion* of an unending process, such as the set $\{1,2,3,\ldots\}$ of all natural numbers. In principle, it is contradictory to speak of the completion of an unending process, but in practice many processes beg for completion, because they have a clear *limit*.

The famous paradoxes of Zeno, such as Achilles and the tortoise, involve processes with a clear limit. Achilles starts behind the tortoise, but runs faster, so clearly he will catch up with the tortoise at some point. The worry, at least for Zeno, is that Achilles is behind the tortoise *at infinitely many stages* (the first stage ends when Achilles reaches the tortoise's starting point, the second when Achilles completes the tortoise's first stage, the third when Achilles completes the tortoise's second stage, and so on). So, apparently, motion involves completing an infinite sequence of events. For the Greeks, this made the concept of motion seem problematic, and they looked for ways to resolve the paradox in terms of potential infinity (as Aristotle did in his *Physics*). For us, it probably supports the idea that actual infinity exists.

At any rate, in ancient Greek mathematics there are many examples where the limit of an infinite process gives interesting new knowledge. For example, Euclid and Archimedes found that both the volume of the tetrahedron and the area of a parabolic segment may be found from the infinite series

$$1+\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots,$$

which has sum $4/3$. They are able to find this sum by considering only the potential infinity of terms

$$1,\quad 1+\frac{1}{4},\quad 1+\frac{1}{4}+\frac{1}{4^2},\quad 1+\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3},\quad 1+\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4},\quad \ldots$$

and showing that

- each of these terms is less than $4/3$, and
- each number less than $4/3$ is exceeded by some term in the sequence.

Thus it is fair to call $4/3$ the sum of $1+\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots$, because we have exhausted all other possibilities. This *method of exhaustion*, also invented by Eudoxus, can be used to avoid actual infinities in many of the cases that interested the Greeks. *But not all.*

Archimedes wrote a book called *The Method*, which was lost for many centuries and did not influence the later development of mathematics. It describes the method by which he discovered some of his most famous results, such as the theorem that the volume of a sphere is $2/3$ the volume of its circumscribing cylinder. The book resurfaced around 1900, and only

then was it realized that Archimedes had used actual infinity in a way that cannot be avoided. In finding certain volumes, Archimedes views a solid body as a sum of *slices of zero thickness*. As we now know, there are uncountably many such slices—corresponding to the uncountably many points on the line—so one cannot view the sum of all slices as the "limit" of finite sums. Of course, it is unlikely that Archimedes had any inkling of uncountability, though he may have suspected that he was dealing with a new kind of infinity.[1]

The Method gives evidence that *intuition about infinity*, even about the continuum, exists and is useful in making mathematical discoveries. More fruits of this intuition appeared around 1800.

THE FIRST MODERN INTUITIONS ABOUT THE CONTINUUM

> ... we ascribe to the straight line completeness, absence of gaps, or continuity. In what then does this continuity consist?
>
> —Richard Dedekind
> *Continuity and Irrational Numbers*, Chapter III,
> in Dedekind (1901), p. 10.

The fundamental property of the continuum $[0,1]$ is that it is, well, *continuous*, in the sense that it *fills the space between its endpoints without gaps.* Around 1800, the German mathematician Carl Friedrich Gauss realized that this seemingly obvious property is the key to a difficult theorem that several mathematicians had vainly sought to prove—the so-called "fundamental theorem of algebra." This theorem states that any polynomial equation, such as

$$x^4 - 2x^2 + 3x + 7 = 0, \quad \text{or} \quad x^5 - x + 1 = 0,$$

is satisfied by some complex number x. Gauss himself had trouble proving the theorem, and all the proofs he offered are incomplete by modern standards. However, in 1816 he gave a proof that clearly identifies the difficulty: it all boils down to the absence of gaps in the continuum.

Gauss's 1816 proof takes any polynomial equation and reduces it, by purely algebraic manipulations, to an equation of *odd degree*, say $p(x) = 0$. This means that the highest-degree term in $p(x)$, such as x^5 in the second example above, is an odd power of x. Now, for large values of x, the polynomial $p(x)$ is dominated by its highest-degree term, and so $p(x)$ *changes sign* between large negative values of x and large positive values of x, since this happens to any odd power of x.

Thus the graph $y = p(x)$ is a continuous curve which moves from negative values of y to positive values. Figure 1.9 shows the curve for

[1] For an up-to-date report on *The Method*, with some interesting mathematical speculations, see Netz and Noel (2007).

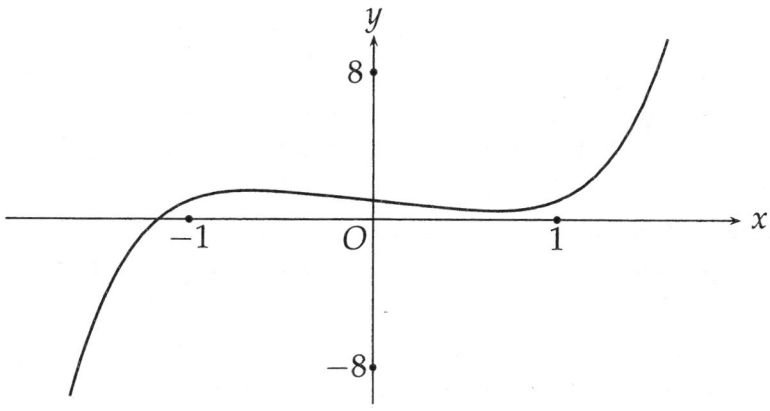

Figure 1.9. Graph of the polynomial $y = x^5 - x + 1$.

the actual example $y = x^5 - x + 1$. Since the curve passes from below to above the x-axis, and the x-axis has no gaps, the curve necessarily *meets* the x-axis.

That is, *there is a real value of x for which* $x^5 - x + 1 = 0$. Similarly, any odd-degree polynomial equation has a real solution, and the fundamental theorem of algebra follows. (Gauss's reduction to odd degree also involves solving quadratic equations, which we know have solutions, sometimes in the complex numbers.)

Gauss's intuition had led him to assume what we now call the *intermediate value theorem*: any continuous function $f(x)$ that takes both positive and negative values between $x = a$ and $x = b$ takes the value zero for some $x = c$ between a and b. The first to identify this assumption, and to attempt to prove it, was the Czech mathematician Bernard Bolzano in 1816. Bolzano was ahead of his time, not only in noticing a property of continuous functions in a theorem previously thought to belong to algebra, but also in realizing that the intermediate value property depends on the nature of the continuum.

Bolzano's attempted proof was incomplete, because a definition of the continuum was completely lacking in his time. However, he correctly identified a *completeness condition* that any reasonable concept of continuum must satisfy. This is the *least upper bound property* : if S is a set of real numbers with an upper bound, then S has a least upper bound. (That is, among the numbers greater than or equal to all members of S, there is a least.)

In 1858, Richard Dedekind brought this train of thought to a satisfying conclusion, defining real numbers by what he called *cuts* in the rationals. A cut is intuitively a separation of the rational numbers into a *lower set* L and an *upper set* U, as if by an infinitely sharp knife. Formally, a cut is a pair (L, U) where L and U are sets that together are all the rationals and

such that every member of L is less than every member of U. Cuts (L, U) represent both rational and irrational numbers as follows:

- If L has a greatest member, or U has a least member, say r, then (L, U) represents the rational number r.
- If L has no greatest member and U has no least member, then (L, U) represents an irrational number. (This happens, for example, when U consists of the positive rationals with square > 2, and L consists of the remaining rationals. The pair (L, U) then represents the irrational number we call $\sqrt{2}$.)

The latter type of cut represents a *gap* in the rationals and, at the same time, provides an object to fill the gap—the cut (L, U). With breathtaking nerve, Dedekind created a gapless continuum by filling each gap in the rationals, *taking the object that fills each gap to be essentially the gap itself.*[2]

INFINITE DECIMALS REVISITED

It should be mentioned that the infinite decimals, used to model real numbers earlier in this chapter, are essentially a more readable version of Dedekind cuts. An infinite decimal, such as 3.14159..., represents a cut in the less crowded set of *decimal fractions*, separating the nearest neighbors less than 3.14159...,

$$3, \quad 3.1, \quad 3.14, \quad 3.141, \quad 3.1415, \quad 3.14159, \quad \ldots,$$

from the nearest neighbors greater than 3.14159... ,

$$4, \quad 3.2, \quad 3.15, \quad 3.142, \quad 3.1416, \quad 3.14160, \quad \ldots.$$

Infinite decimals are easy to read and understand, but try defining their sum and product! You will probably fall back on adding and multiplying their neighboring decimal fractions, much as with Dedekind cuts.

Of course, it is not enough that there are no gaps in the set of cuts. We also need to know that *cuts are entities that behave like numbers.* This is true. One can define the "sum" and "product" of two cuts in terms of the rational numbers in them, and this "sum" and "product" have the usual algebraic properties. For example, the cut for $\sqrt{2} + \sqrt{3}$ has lower

[2]Dedekind slightly tarnished the purity and boldness of his idea by insisting on his right to *create a new object* to fill each gap. It is perfectly valid, and more economical, to insist that the gap itself is a genuine mathematical object, which we can take to be the pair (L, U).

set consisting of all the numbers $r + s$, where r is in the lower set for $\sqrt{2}$ and s is in the lower set for $\sqrt{3}$. This is the same as the cut for $\sqrt{3} + \sqrt{2}$ because $r + s = s + r$. Best of all, *any bounded set S of cuts has a least upper bound.* The least upper bound of S has a lower set obtained by uniting the lower sets for all members of S.

Thus, with Dedekind's definition of real numbers, it was finally possible to prove the intermediate value theorem, and hence the fundamental theorem of algebra. At the same time, the proof initiated a new direction in mathematical thought. Previously undefined mathematical objects became defined in terms of sets, and every set became a legitimate mathematical object—even the uncountable set of real numbers.

Indeed, the set of real numbers was welcomed by many mathematicians as a mathematical model of the line. Today, the "number line" seems like a simple idea, but it is not! A "point" is a whole universe to someone who knows that it is actually a cut in the infinite set of rational numbers. Nonetheless, in the 1870s, many mathematicians saw this *arithmetization of geometry* as the best way to build the foundations of mathematics. Arithmetization resolved the ancient conflict between numbers and geometric magnitudes; it also provided a common foundation for geometry and calculus. And arithmetization was timely because Cantor had just started exploring the set concept itself.

However, further exploration of the set concept led to some surprises.

The Paradoxes of Set Theory

> ... a mere practical joke played on mankind by the goddess of wisdom.
>
> —Azriel Levy (1979), p. 7.

The function-based diagonal argument of Paul du Bois-Reymond (Section 1.6) had far-reaching consequences, as we will see in later chapters of this book. The set-based diagonal argument of Cantor (Section 1.3) had consequences that were more immediate and dramatic.

In 1891, Cantor realized that the diagonal argument applies to *any* set, showing that *any set X has more subsets than members.* Of course, we cannot generally visualize a tabulation of subsets of X, as we did in Section 1.3 for subsets of the natural numbers. It is sufficient to consider any *one-to-one correspondence* between members x of X and subsets of X. Let S_x be the subset corresponding to x. From the sets S_x we define the "diagonal set" S, whose members are the x such that x does *not* belong to S_x.

It is clear that S *differs from each set* S_x with respect to the element x: if x is in S_x then x is not in S, and if x is not S_x then x is in S. Thus, any pairing of subsets of X with members of X fails to include all subsets. This is what we mean by saying that X has more subsets than members.

It follows, in turn, that *there is no largest set*. And, therefore, *there is no such thing as the "set of all sets"*—because the set of all sets, if it exists, is necessarily the largest set. When Cantor noticed this, in 1895, it seems to have given him some pause. What exactly does the word "set" mean if there is no set of all sets? Cantor did not have a precise answer to this question, but neither was he greatly perturbed. He had no commitment to the "set of all sets" and was content to consider only sets obtained by clear operations on given sets, such as collecting all subsets.

The question was more troubling for philosophers of mathematics, such as Gottlob Frege and Bertrand Russell, who believed that any property $\mathcal{P}$ should determine a set—the set of all objects with property $\mathcal{P}$. When the property in question is "being a set" then this belief leads to the "set of all sets." Indeed, Russell rediscovered the contradiction in the "set of all sets" in 1901, in a form that became famous as the *Russell paradox*. Russell's contribution was to distill the contradiction into "the set of all sets that are not members of themselves." The latter "set" R is immediately self-contradictory, since R belongs to R if and only if R does *not* belong to R.

Russell's argument convinced mathematicians that the set concept needs clarification, and it reinforced the idea from the 1870s that mathematics needs secure foundations. This "crisis in foundations" (and other "crises" we will meet later) had profound consequences, as we will see in the rest of the book. The problem facing set theory was described by the German mathematician Ernst Zermelo (1908) as follows.

> ...the very existence of this discipline seems to be threatened by certain contradictions ... In particular, in view of the "Russell antinomy" of the set of all sets that do not contain themselves as elements, it no longer seems admissible today to assign to an arbitrary logically defined notion a set, or class, as its extension.

Zermelo believed that set theory could be saved by *axioms for sets*, formalizing Cantor's intuition that all sets arise from given sets (such as the set of natural numbers) by well-defined operations (such as collecting all subsets).

The axioms for sets most commonly used today are due to Zermelo (1908), with an important supplement from his compatriot (who later moved to Israel) Abraham Fraenkel in 1922. Because of this they are called the *ZF axioms*. They are written in a formal language for set theory that I have not introduced, but most of them can be clearly expressed in ordinary language with the help of the concept of "membership." We write the set with members $a, b, c, \ldots$ as $\{a, b, c, \ldots\}$. Thus the brackets "comprehend" the objects $a, b, c, \ldots$ (which may themselves be sets) as members of a set.

AXIOM 1. Two sets are equal if and only if they have the same members.

AXIOM 2. There is a set with no members, called the *empty set*.

AXIOM 3. For any sets X and Y, there is a set whose only members are X and Y. (This set, $\{X, Y\}$, is called the *unordered pair* of X and Y. Note that, when $Y = Z$, Axiom 1 gives $\{Y, Z\} = \{Y\}$. Thus the pairing axiom also gives us the "singleton" set $\{Y\}$ whose single member is Y.) The ZF axioms are the following.

AXIOM 4. For any set X there is a set whose members are the members of members of X. (In the case where $X = \{Y, Z\}$, the members of members of X form what is called the *union of Y and Z*, denoted by $Y \cup Z$. In all cases, the set of members of members of X is called the *union* of the members of X.)

AXIOM 5. For any set X, there is a set whose members are the *subsets* of X, where a subset of X is a set whose members are members of X. (The set of subsets of X is called the *power set* of X.)

AXIOM 6. For any function definition f, and set X, the values $f(x)$, where x is a member of X, form a set. (This set is called the *range* of the function f on the *domain* X, and the axiom is called *replacement*.)

AXIOM 7. Any nonempty set X has a member Y with no members in X. (A more enlightening version of this axiom—though harder to express in formal language—is the following. *There is no infinite descending sequence for set membership.* That is, if one takes a member X_1 of X, then a member X_2 of X_1, and so on, then this process can continue for only finitely many steps.)

AXIOM 8. There is an infinite set, in fact a nonempty set which, along with any member X, also has the member $X \cup \{X\}$.

Axioms 1–6 say that sets are built from the empty set by operations of pairing, union, power, and replacement (taking the range of a function). To say that a function f is "defined" means that f is expressed by a formula in the formal language of ZF, which basically consists of logic symbols and symbols for membership and equality. We study formal languages for mathematics in Chapters 3, 4, and 5.

The mysterious Axiom 7, called the *axiom of foundation*, enables us to prove that *every* set arises from the empty set by the operations above. It may seem that we have a severely limited menu of sets, but in fact there are sets that can play the roles of all the objects normally needed for mathematics. First, following an idea of the Hungarian (later American)

mathematician John von Neumann from the 1920s, we can define the natural numbers as follows. Let 0 be the empty set (what else?), then let

$$\begin{aligned} 1 &= \{0\}, \\ 2 &= \{0, 1\}, \\ 3 &= \{0, 1, 2\}, \end{aligned}$$

and so on. Notice that we have $n + 1 = n \cup \{n\}$ and $m < n$ if and only if m is a member of n. Thus set theory gives us definitions of the successor function and the "less than" relation for free, and we can develop arithmetic.

Next, Axiom 8 (the *axiom of infinity*), together with Axiom 6, tells us that there is a set whose members are $0, 1, 2, 3, \ldots$, so we have the set of natural numbers. Taking its power set, we are well on the way to the real numbers, the number line, geometry, calculus, and virtually everything else.

Who knew that the empty set could be so fruitful?

- CHAPTER 2 -

ORDINALS

PREVIEW

previous chapter we studied sets that can be "listed" completely, very member in a positive integer position. We found that some nnot be listed in this way, even though the list is infinite. The al argument always gives a new member. Can we continue the list finity?

org Cantor counted past infinity using the concept of *ordinal num-*ne natural numbers 0, 1, 2, 3, ... are the finite ordinal numbers, but ave a least upper bound ω—the first *transfinite* ordinal number. A ay to formalize this is to let ω be the set $\{0, 1, 2, 3, \ldots\}$. Then ω is t term beyond the finite ordinals on the list $0, 1, 2, 3, \ldots, \omega$, ordered mbership.

e the finite ordinals, each transfinite ordinal has a successor, so re ordinals $\omega + 1, \omega + 2, \omega + 3$, and so on.

dinals keep pace with the production of new objects by the diag-rgument, because any countable set of ordinals has a least upper . Beyond $\omega + 1, \omega + 2, \omega + 3, \ldots$ there is $\omega \cdot 2$. Beyond $\omega \cdot 2, \omega \cdot$ $, \ldots$ there is ω^2. Beyond $\omega^2, \omega^3, \omega^4, \ldots$ there is ω^ω. And these are a few of what we call the *countable ordinals.*

e least upper bound of the countable ordinals is the *first uncountable* , called ω_1. Thus ordinals offer a different road to uncountable y. The ordinal road is slower, but more—shall we say—orderly.

t only are ordinals ordered, they are *well*-ordered. Any nonempty them has a least member; equivalently, there is no infinite descend-quence of ordinals. Thus, *even though the road to an ordinal may be ny return trip is short* (because it is finite). This fact has surprising quences in the world of finite objects, as we show in Sections 2.7 8.