CHAPTER 1

Historical Introduction

The purpose of this introductory chapter is to prepare the reader's mind for *reverse mathematics*. As its name suggests, reverse mathematics seeks not theorems but the right axioms to prove theorems already known. The criterion for an axiom to be "right" was expressed by Friedman (1975) as follows:

> When the theorem is proved from the right axioms, the axioms can be proved from the theorem.

Reverse mathematics began as a technical field of mathematical logic, but its main ideas have precedents in the ancient field of geometry and the early twentieth-century field of set theory.

In geometry, the parallel axiom is the right axiom to prove many theorems of Euclidean geometry, such as the Pythagorean theorem. To see why, we need to separate the parallel axiom from the *base theory* of Euclid's other axioms, and show that the parallel axiom is not a theorem of the base theory. This was not achieved until 1868. It is easier to see that the base theory can prove the parallel axiom *equivalent* to many other theorems, including the Pythagorean theorem. This is the hallmark of a good base theory: what it cannot prove outright it can prove equivalent to the "right axioms."

Set theory offers a more modern example: a base theory called ZF, a theorem that ZF cannot prove (the well-ordering theorem) and the "right axiom" for proving it—the axiom of choice.

From these and similar examples we can guess at a base theory for analysis, and the "right axioms" for proving some of its well-known theorems.

1.1 EUCLID AND THE PARALLEL AXIOM

The search for the "right axioms" for mathematics began with Euclid, around 300 BCE, when he proposed axioms for what we now call *Euclidean geometry*. Euclid's axioms are now known to be incomplete; nevertheless, they outline a complete system, and they distinguish between really obvious "basic" axioms and a less obvious one that is crucial for obtaining the most important theorems. For historical commentary on the axioms, see Heath (1956).

The basic axioms say, for example, that there is a unique line through two distinct points and that lines are unbounded in length. Also basic, though expressed only vaguely by Euclid, are criteria for *congruence of triangles*, such as what we call the "side angle side" or SAS criterion: if two triangles agree in two sides and the included angle then they agree in all sides and all angles. (Likewise ASA: they agree if they agree in two angles and the side between them.)

Using the basic axioms it is possible to prove many theorems of a rather unsurprising kind. An example is the *isosceles triangle theorem*: if a triangle ABC has side AB = side AC then the angles at B and C are equal. However, the basic axioms fail to prove the signature theorem of Euclidean geometry, the *Pythagorean theorem*, illustrated by figure 1.1.

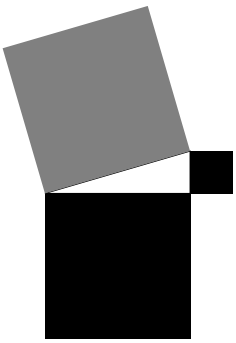

Figure 1.1 : The Pythagorean theorem

As everybody knows, the theorem says that the gray square is equal to the sum of the black squares, but the basic axioms cannot even prove the *existence* of squares. To prove the Pythagorean theorem, as Euclid realized, we need an axiom about infinity: the *parallel axiom*.

The Parallel Axiom

I call the parallel axiom an axiom about infinity because it is about lines that do not meet, *no matter how far they are extended*—and one of Euclid's basic axioms is that lines can be extended indefinitely. Thus parallelism cannot be "seen" unless we have the power to see to infinity, and Euclid preferred not to assume such a superhuman power. Instead, he gave a criterion for lines *not* to be parallel, since a meeting of lines can be "seen" a finite distance away.

Parallel axiom. If a line n falling on lines l and m (figure 1.2) makes angles α and β with $\alpha + \beta$ less than two right angles, then l and m meet on the side on which α and β occur.

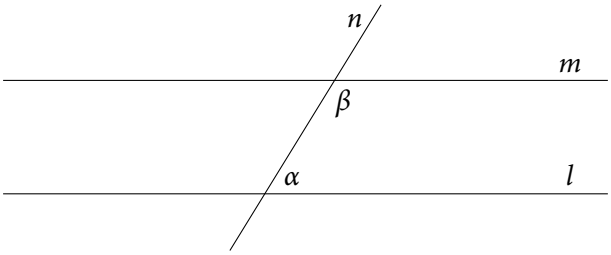

Figure 1.2 : Angles involved in the parallel axiom

It follows that if $\alpha + \beta$ equals two right angles (that is, a straight angle) then l and m do *not* meet. Because if they meet on one side (forming a triangle) they must meet on the other (forming a congruent triangle, by ASA), since there are angles α and β on both sides and one side in common (figure 1.3). This contradicts uniqueness of the line through any two points.

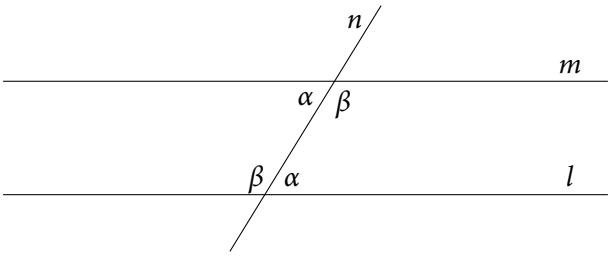

Figure 1.3 : Parallel lines

Thus Euclid's axiom about *non*-parallel lines implies that parallel lines exist. From parallel lines we quickly get the theorem that the angle sum of a triangle is a straight angle (or π, as we will write it from now on), by the construction shown in figure 1.4. From this we find in turn that an isosceles triangle with angle $\pi/2$ between its equal sides has its other angles equal to $\pi/4$, so putting two such triangles together makes a square.

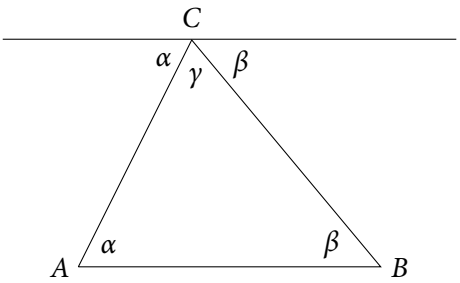

Figure 1.4 : Angle sum of a triangle

The proof of the Pythagorean theorem can now get off the ground, and there are many ways to complete it. Probably the one most easily "seen" is shown in figure 1.5, in which the gray square and the two black squares both equal the big square minus four copies of the right-angled triangle.

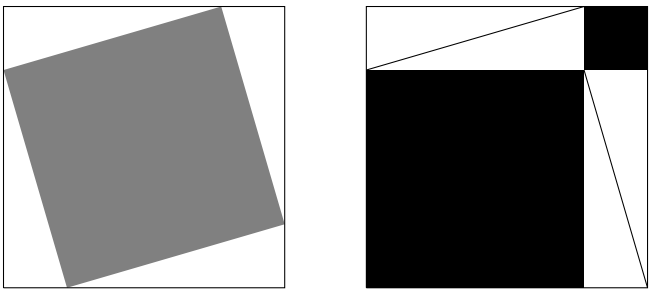

Figure 1.5 : Seeing the Pythagorean theorem

Equivalents of the Parallel Axiom

Many mathematicians considered the parallel axiom to be a "blemish" on Euclid's system—this is precisely what Saccheri (1733) called it—so they tried to show that it followed from the other axioms. Their attempts

usually took the form of deducing the parallel axiom from a seemingly more obvious statement, in the hope of reducing the problem to a simpler one. Some of the statements found to imply the parallel axiom were:

- existence of rectangles (al-Haytham, al-Tusi in medieval times),
- existence of similar triangles of different sizes (Wallis in 1693),
- angle sum of triangle = π (in Legendre's *Éléments de géométrie*, 1823),
- three noncollinear points lie on a circle (Farkas Bolyai (1832)).

All of these theorems follow from the parallel axiom, so they are equivalent to it in *strength*, in the sense that their equivalence to the parallel axiom can be proved using only the other axioms. Of course, this notion of equivalent strength is trivial if the parallel axiom itself is provable from the other axioms, but by 1830 the hopes of such a proof were fading. Farkas Bolyai's own son, János, was one of the main explorers of a hypothetical *non*-Euclidean geometry in which the parallel axiom (and hence the four theorems above) is *false*, yet Euclid's other axioms are true.

But before seeing non-Euclidean geometry, it helps to look at geometry on the sphere. Spherical geometry is clearly different from the Euclidean geometry of the plane—not only in the absence of parallels, but also in the absence of infinite lines—yet they share a common language of "points," "lines," and "angles." Seeing two different interpretations of these words will make it easier to grasp yet another interpretation, or *model*—a model of non-Euclidean geometry.

1.2 SPHERICAL AND NON-EUCLIDEAN GEOMETRY

Just as circles and lines in the plane are part of two-dimensional Euclidean geometry, spheres and planes are part of *three*-dimensional Euclidean geometry. Indeed they are mentioned, though not deeply studied, in Euclid's *Elements*, Book XI. The ancient Greeks made a serious study of spherical geometry, particularly spherical trigonometry, in their study of astronomy, because the stars appear from the earth to be fixed on a heavenly sphere. Later, navigators on the earth also took an interest in spherical geometry. For them, the natural concept of "line" is that of a *great circle*—the intersection of the sphere with a plane through its center—because a great circle gives the shortest distance between any two of its points. The concept of "angle" between any two such "lines" also makes sense, as the angle between the corresponding planes (or, what comes to

the same thing, the angle between the tangents to the great circles).

Indeed, it is often easier to describe a spherical triangle by its angles rather than the lengths of its sides. All spherical triangles with the same angles in fact have the same size, because of a famous theorem of Harriot[1] from 1603: *the angle sum of a spherical triangle, minus π, is proportional to its area*. There are several ways to tile the surface of the sphere with congruent triangles. Figure 1.6 shows one in which the sphere is divided into 48 triangles, each of which has angles $\pi/2, \pi/3, \pi/4$. Alternate triangles have been cut out of the sphere, to make it easier to see them all, and the sphere has been illuminated from the inside. This then is the standard

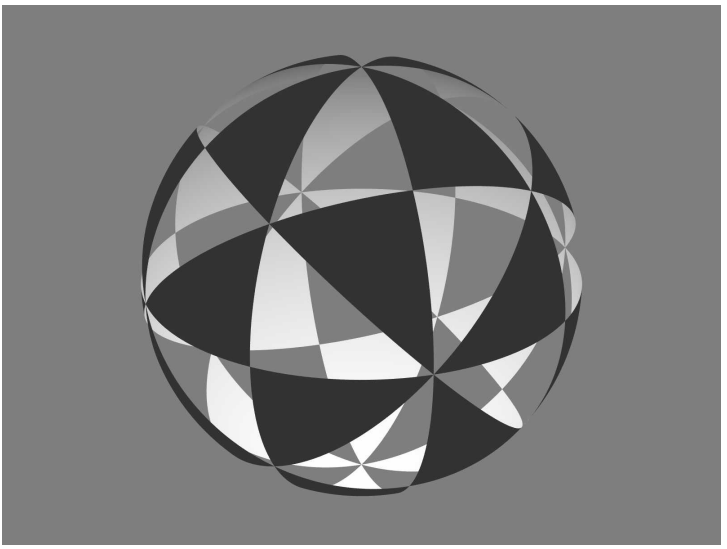

Figure 1.6 : Tiling the sphere with triangles

model of spherical geometry: "points" are ordinary points on the sphere, "lines" are great circles, and "angles" are the angles between the tangents to the great circles at their point of intersection. "Distance," if we wish to use the concept, is the distance between points on the sphere, measured along the (shorter) piece of the great circle connecting them.

Now we move to another model, by *projecting the sphere onto the plane*. Specifically, we use the light inside the sphere (at its north pole) to cast a shadow on the plane. The result is shown in figure 1.7. The pic-

[1]Thomas Harriot was mathematical consultant to Sir Walter Raleigh, and traveled with him on some of his voyages.

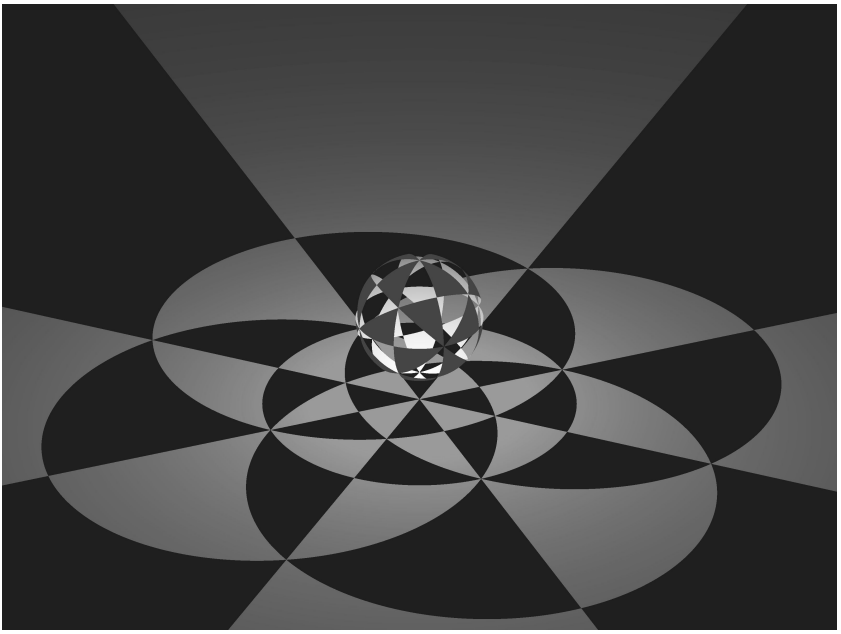

Figure 1.7 : Projecting the sphere onto the plane

ture shows two remarkable features of projection from the north pole, which is known as *stereographic* projection:

- circles map to circles (or, in exceptional cases, to straight lines, which we might call "circles of infinite radius"), and
- angles are preserved.

Thus "points" are still points, "lines" are still circles, and "angle" is still the angle between the tangents to the circles. "Distance," alas, is not a Euclidean distance of any kind, since equal distances on the sphere can be mapped to unequal Euclidean distances in the plane. Likewise, "area" is not Euclidean area, but we can easily measure it by the angle sum minus π.

Strictly speaking, we have not projected the whole sphere onto the plane, but the sphere minus its north pole (the light source). To correct for this we add a *point at infinity* to the plane—a point approached by the shadows of points on the sphere as they approach the north pole. The point at infinity completes each straight line to a closed curve, so that they too become circles. Thus our second interpretation of spherical geometry models all "lines" by circles, and "angles" by angles between circles. In the

next subsection we will see a similar model of non-Euclidean geometry.

Models of Non-Euclidean Geometry

Beltrami (1868) discovered several models of non-Euclidean geometry; that is, of Euclid's basic axioms plus a non-Euclidean parallel axiom stating that *for any line l and a point P outside it, there is more than one line m that does not meet l*. The easiest of Beltrami's models to view in its entirety is the one shown in figure 1.8.

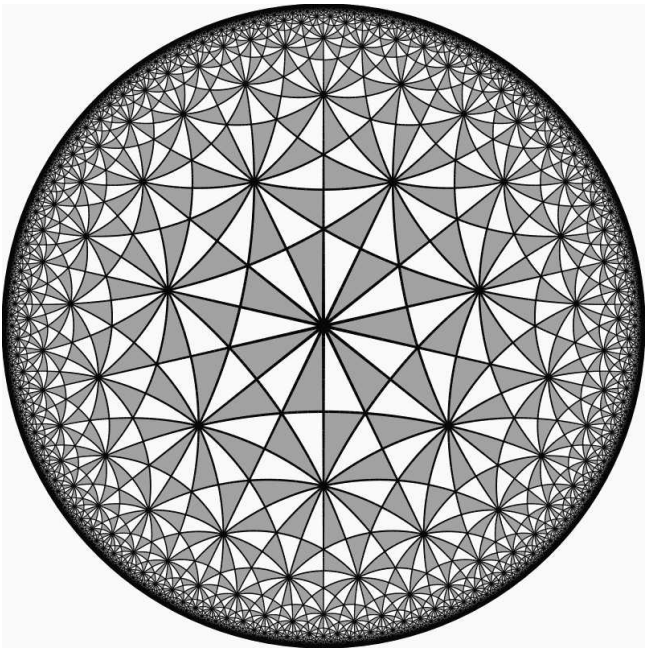

Figure 1.8 : The conformal disk model

In this model, "points" are points in the interior of the disk, "lines" are circular arcs perpendicular to the boundary circle of the disk (counting the straight line segments through the disk center as circles of infinite radius) and "angle" is the angle between circles. As in spherical geometry, triangles are congruent if they have the same angles, so in this picture the disk is filled with infinitely many congruent triangles, each with the angles $\pi/2, \pi/3, \pi/7$. These are the smallest triangles that can tile the non-Euclidean plane and, as in spherical geometry, their area is determined by their angle sum: *π minus the angle sum of a non-Euclidean triangle is*

proportional to its area.

As with the plane model of spherical geometry, the precise definition of "distance" is complicated. But here one gets a better feel for it because there are so many triangles, each of the same non-Euclidean size. One sees, for example, that infinitely many triangles lie along each "line," so each "line" is of infinite "length." It is even possible to accept that each "line" gives the least "distance" between any two points in the disk, since one counts fewer triangles when travelling on a circular arc perpendicular to the boundary than on any other route. Thus one can understand how the model satisfies the basic axioms of Euclid. But it clearly does *not* satisfy the parallel axiom. If one takes the vertical "line" l through the center of the disk and the point P, say, somewhat to its right, then there are different "lines" m and n through P that do not meet l, as is clear from figure 1.9.

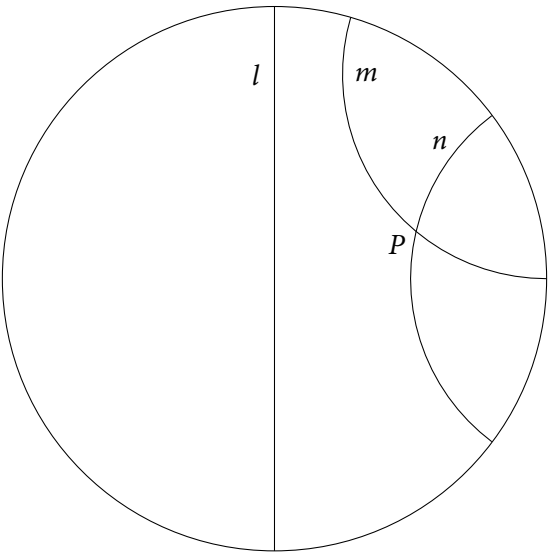

Figure 1.9 : Failure of the parallel axiom

So, when the details of Beltrami's construction are checked, one has *a model for the basic axioms of Euclid plus a counterexample to the parallel axiom*. Therefore, *the parallel axiom does not follow from the other axioms of Euclid*, and hence the theorems equivalent to the parallel axiom (such as the four mentioned in the previous section) likewise do not follow from Euclid's other axioms. However, the equivalences between

the parallel axiom and these theorems are provable from Euclid's other axioms. This situation is typical of reverse mathematics: we have a *base theory* which is too weak to prove certain desirable theorems, but strong enough to prove *equivalences* between them.

New Foundations of Geometry and Mathematics

The discovery of non-Euclidean geometry shook the foundations of mathematics, which before the nineteenth century had been implicitly based on Euclid's concepts of "line" and "plane." By creating doubts about the meaning of "line" and "plane," non-Euclidean geometry prompted a search for new foundations in *arithmetic*, since the fundamental properties of numbers were not in doubt.

In particular, the "line" was rebuilt as the system $\mathbb{R}$ of *real numbers*, which has both algebraic and geometric properties. The next few sections describe the emergence of geometry based on, or influenced by, the real number concept. In chapter 2 we will see how the real numbers also became the foundation of analysis.

1.3 VECTOR GEOMETRY

The first major advance in geometry after the Greeks was made by Fermat and Descartes in the 1620s, and published in the *Geometry* of Descartes (1637). Their innovation was to use algebra in geometry, describing lines and curves by equations, thereby reducing many problems of geometry to routine calculations. But before they could "algebraicize" geometry they had to *arithmetize* it, a step that already took them far beyond Euclid. In fact, it was the first step towards a sweeping arithmetization of geometry and analysis that occurred in the nineteenth century.

As every mathematics student now knows, the Euclidean plane is arithmetized by assigning real number *coordinates* x and y to each point P in the plane. The numbers x and y are visualized as the horizontal and vertical distances to P from the origin O, in which case the distance $|OP|$ of P from O is $\sqrt{x^2 + y^2}$ by the Pythagorean theorem (figure 1.10). But P can be *defined* as the ordered pair[2] $\langle x, y \rangle$, and its distance from O defined as $\sqrt{x^2 + y^2}$. More generally, the distance from $P_1 = \langle x_1, y_1 \rangle$ to

[2]In this book I use $\langle a, b \rangle$ to denote the ordered pair of a and b, because (a, b) will be on duty to represent the open interval between a and b.

$P_2 = \langle x_2, y_2 \rangle$ is defined by

$$|P_1P_2| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Points $\langle x, y \rangle$ lie on a *line* if they satisfy an equation of the form $ax + by + c = 0$ (which is why we call such equations *linear*), and equations for circles are quadratic equations expressing constant distance for a point. For example, the points at distance 1 from O satisfy the equation $x^2 + y^2 = 1$.

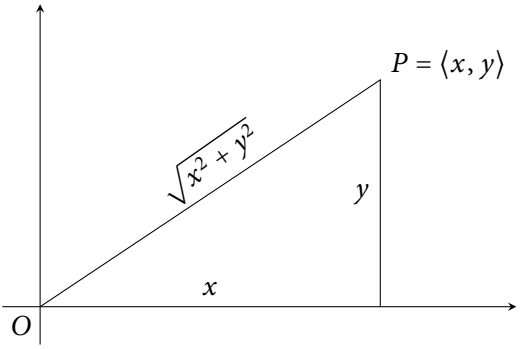

Figure 1.10 : Coordinatizing the plane

Thus one has an easy algebraic translation of all of Euclid's geometry—*and more*, since there is no obstacle, other than algebraic difficulty, to the study of curves satisfying arbitrary polynomial equations. Thus Euclidean geometry and algebraic geometry are not a perfect match. Euclidean geometry ought to be "more linear."

Grassmann's Linear Geometry

The perfect algebraic match for Euclidean geometry was found by Grassmann in the 1840s, in the concept of a *real vector space*. His first works on the subject, Grassmann (1844) and Grassmann (1847) were impenetrable to other mathematicians, and his idea started to gain traction only when Peano (1888) gave axioms for real vector spaces.

Definition. A real vector space is a set V of objects called *vectors* (denoted by boldface letters), which includes a vector $\mathbf{0}$ called the *zero vector*, and for each $\boldsymbol{u} \in V$ a vector $-\boldsymbol{u}$ called the *negative* of $\boldsymbol{u}$. V has operations of *addition* and *scalar multiplication* (by $a, b, c, \ldots \in \mathbb{R}$) satisfying the

following conditions:

$$\begin{aligned}
\boldsymbol{u} + \boldsymbol{v} &= \boldsymbol{v} + \boldsymbol{u} \\
\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) &= (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w} \\
\boldsymbol{u} + \mathbf{0} &= \boldsymbol{u} \\
\boldsymbol{u} + (-\boldsymbol{u}) &= \mathbf{0} \\
1\boldsymbol{u} &= \boldsymbol{u} \\
a(\boldsymbol{u} + \boldsymbol{v}) &= a\boldsymbol{u} + a\boldsymbol{v} \\
(a + b)\boldsymbol{u} &= a\boldsymbol{u} + b\boldsymbol{u} \\
a(b\boldsymbol{u}) &= (ab)\boldsymbol{u}
\end{aligned}$$

Typically $V = \mathbb{R}^n = \{\langle x_1, \ldots, x_n \rangle : x_1, \ldots, x_n\}$, with $\mathbf{0}$ the origin, $+$ the usual sum of n-tuples, and scalar multiplication by $a \in \mathbb{R}$ given by

$$a\langle x_1, \ldots, x_n \rangle = \langle ax_1, \ldots, ax_n \rangle.$$

This vector space is called the *real n-dimensional affine space*. It is not yet a Euclidean space because it has no concept of distance or angle, but it has considerable geometric content. $\mathbb{R}^n$ has lines, including parallel lines, and also a concept of "length in a given direction." For example, one can say that $\frac{1}{2}\boldsymbol{v} \in \mathbb{R}$ is the *midpoint* of the line from $\mathbf{0}$ to $\boldsymbol{v}$, and in general $a\boldsymbol{v}$ is a times as far from $\mathbf{0}$ as $\boldsymbol{v}$ is. Another concept that makes sense in vector geometry is that of *center of mass*. In particular, the center of mass of the triangle with vertices $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ is the point $\frac{1}{3}(\boldsymbol{u} + \boldsymbol{v} + \boldsymbol{w})$.

To promote vector geometry to Euclidean geometry one adds the concept of *inner product* of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, written $\boldsymbol{u} \cdot \boldsymbol{v}$:

Definition. If $\boldsymbol{u} = \langle u_1, \ldots, u_n \rangle$ and $\boldsymbol{v} = \langle v_1, \ldots, v_n \rangle$ then

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + \cdots + u_n v_n.$$

In particular, in $\mathbb{R}^2$ we have

$$\boldsymbol{u} \cdot \boldsymbol{u} = u_1^2 + u_2^2,$$

so the Euclidean length $|\boldsymbol{u}|$ of $\boldsymbol{u}$ is given by $|\boldsymbol{u}| = \sqrt{\boldsymbol{u} \cdot \boldsymbol{u}}$. As Grassmann (1847) remarked, the definition of inner product makes the Pythagorean theorem true almost by definition.

The Euclidean angle concept also derives from the inner product because

$$\boldsymbol{u} \cdot \boldsymbol{v} = |\boldsymbol{u}||\boldsymbol{v}| \cos\theta,$$

where θ is the angle between the lines from $\mathbf{0}$ to $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. Thus Grassmann (1847) found another way to describe Euclidean geometry as a "base theory" plus the "right axiom" to derive the Pythagorean theorem. Interestingly, his base theory (the vector space axioms) admits extension by a different axiom that gives *non*-Euclidean geometry.

Making a Vector Space Non-Euclidean

The key property of Grassmann's inner product is that it is *positive definite*; that is, $|\boldsymbol{u}|^2 = \boldsymbol{u} \cdot \boldsymbol{u} > 0$ if $\boldsymbol{u} \neq \mathbf{0}$, so every nonzero vector has positive length. Einstein's theory of special relativity motivated Minkowski (1908) to introduce a *non*-positive definite inner product on the space $\mathbb{R}^4$ of spacetime vectors $\langle t, x, y, z \rangle$, namely

$$\langle t_1, x_1, y_1, z_1 \rangle \cdot \langle t_2, x_2, y_2, z_2 \rangle = -t_1 t_2 + x_1 x_2 + y_1 y_2 + z_1 z_2.$$

With the Minkowski inner product $\boldsymbol{u} = \langle t, x, y, z \rangle$ has "length" $|\boldsymbol{u}|$ given by

$$|\boldsymbol{u}|^2 = -t^2 + x^2 + y^2 + z^2,$$

which clearly is zero or negative for many vectors. To make visualization easier we consider the corresponding concept of length on the space $\mathbb{R}^3$ of vectors $\boldsymbol{u} = \langle t, x, y \rangle$, namely

$$|\boldsymbol{u}|^2 = -t^2 + x^2 + y^2.$$

This means that in $\mathbb{R}^3$ we have a "sphere[3] of radius $\sqrt{-1}$ about O," consisting of the vectors $\boldsymbol{u} = \langle t, x, y \rangle$ such that

$$-t^2 + x^2 + y^2 = -1.$$

This surface in $\mathbb{R}^3$ is the *hyperboloid* $x^2 + y^2 - t^2 = 1$.

It turns out that the Minkowski distance on the surface of the hyperboloid gives a non-Euclidean geometry—the same as that of the Beltrami model in the previous section. Figure 1.11, which is derived from a picture by Konrad Polthier of the Freel University of Berlin, shows the connection between the two. The tiling of the disk projects to a tiling of the hyperboloid by triangles that are congruent in the sense of Minkowski distance.

[3] In a remarkable prophecy, Lambert (1766) conjectured that there might be a geometry on the sphere of imaginary radius for which the angle sum of a triangle is less than π, and where the area of a triangle is proportional to π minus its angle sum. This is indeed what happens in Beltrami's non-Euclidean geometry.

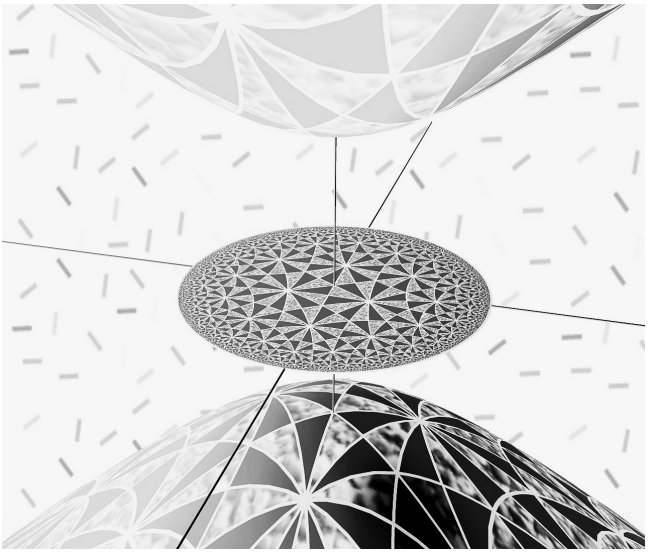

Figure 1.11 : The hyperboloid model of non-Euclidean geometry

1.4 HILBERT'S AXIOMS

Euclid's *Elements* is the first organized presentation of mathematics that survives from ancient times. It is best known for its treatment of geometry, deducing theorems from axioms in a style that became standard for mathematics until the nineteenth century. Then the discovery of non-Euclidean geometry put Euclid's geometry under the microscope, and by the late nineteenth century his axioms were found to have some gaps. But this only strengthened the movement towards axiomatization. The gaps in Euclid were filled by Hilbert (1899) and, in the meantime, axiomatic treatments of number theory and algebra were given by Dedekind, Peano, and others.

Euclid also gave a deductive treatment of numbers in the *Elements*, but it was complicated by the Greek discovery of irrationality, which was thought to disqualify some geometric quantities (such as the diagonal $\sqrt{2}$ of the unit square) from being numbers at all. Irrational quantities were not fully reconciled with whole or rational numbers until the publication of the Dedekind (1872) book on irrational numbers. Dedekind found that Euclid had been on the right track—the only new idea needed to make his theory of irrational quantities part of his theory of numbers

was acceptance of *infinite sets* of rational numbers (see section 1.5).

The two main threads of the *Elements*, geometry and the real numbers, were combined in the *Grundlagen der Geometrie* (foundations of geometry) of Hilbert (1899). Here, Hilbert not only filled the gaps in Euclid's geometric axioms, he also introduced two axioms that complete a geometric path to the real number system $\mathbb{R}$. This was a historic achievement, though Hilbert's path is not the best for all mathematical purposes. The *arithmetization* path to real numbers via the rational numbers ultimately proved more useful for analysis, and we will take it up again in chapter 2.

Hilbert (1899) found that Euclid's geometry and the arithmetic of real numbers follow from 17 axioms, described below. All but two of them are purely geometric. The exceptions are the *Archimedean axiom*, which says no line segment is "infinitely large" compared with another, and the *completeness axiom*, which says there are no "gaps" in the points on a line. (These two axioms were not needed by Euclid, who considered only points constructible by ruler and compass.) Their purpose is to prove that any line satisfying the axioms is essentially the line $\mathbb{R}$ of real numbers. It follows that any plane satisfying the axioms is essentially the plane of Descartes, so Euclid's geometry has really only one model—the plane of pairs of real numbers.

This very satisfying convergence of the geometric and arithmetic viewpoints comes about because Hilbert's geometric axioms yield not just Euclid's geometric theorems—they also yield *algebra*, which Euclid did not foresee. In fact, algebraic structure arises in stages corresponding to axiom *groups*, which Hilbert introduces one by one.

Axioms of incidence. These relate lines and points. They include Euclid's axiom that two points determine a line, and a form of the parallel axiom: for any line l and point $P \notin l$ there is exactly one line m through P not meeting l. Also (which went without saying in Euclid) each line has at least two points, and there are three points not in a line.

Axioms of order. The first three of these axioms say the obvious things about the order of three points on a line: if B is between A and C then it is also between C and A; any A and C have a point B between them; for any three points, one is between the other two. The fourth, called *Pasch's axiom*, is about the plane: a line meeting one side of a triangle at

an internal point meets exactly one of the other sides.

Axioms of congruence. The first five of these axioms are about equality of line segments or angles, and the addition of line segments. They state the existence and uniqueness of line segments or angles equal to given ones, at a given position. They also say (as Euclid put it) "things equal to the same thing are equal to each other." The last congruence axiom is the SAS criterion for congruence of triangles.

Circle intersection axiom. Two circles meet if one of them contains points both inside and outside the other. (Euclid overlooked this axiom, even though he assumed it in his very first proposition, constructing an equilateral triangle.) Note that the points "inside" a circle of radius r are those at distance $< r$ from its center.

Archimedean axiom. For any nonzero line segments AB and CD there is a natural number n such that n copies of AB are together greater than CD.

Completeness axiom. Suppose the points of a line l are divided into two nonempty subsets $\mathcal{A}$ and $\mathcal{B}$ such that no point of $\mathcal{A}$ is between two points of $\mathcal{B}$ and no point of $\mathcal{B}$ is between two points of $\mathcal{A}$. Then there is a unique point P, in either $\mathcal{A}$ or $\mathcal{B}$, that lies between any other two points, of which one is in $\mathcal{A}$ and the other is in $\mathcal{B}$. (Thus, there is no "gap" between $\mathcal{A}$ and $\mathcal{B}$.)

These axioms give precise meaning to the idea of a theorem being *equivalent* to the parallel axiom: namely, the equivalence is provable in the *base theory* of Hilbert's axioms *minus* the parallel axiom. All theorems previously thought to be equivalent to the parallel axiom (such as those mentioned in section 1.1) are equivalent to it in this sense. As suggested at the end of section 1.2, proving equivalences in a weaker system is the hallmark of *reverse mathematics*. We will see further historical examples in the later sections of this chapter. Today, the idea has been most fully developed in systems of analysis, and we will see some its main results in chapters 6 and 7.

Algebraic Content of Hilbert's Axioms

The incidence axioms allow us to define sum and product of points on a line by means of the constructions shown in figures 1.12 and 1.13.

The sum construction chooses a point 0 on the line then, for any points a and b on the line, constructs a point $a + b$ with the help of the parallels shown. In effect, the parallels allow the point b to be "translated" along the line by the distance between 0 and a.

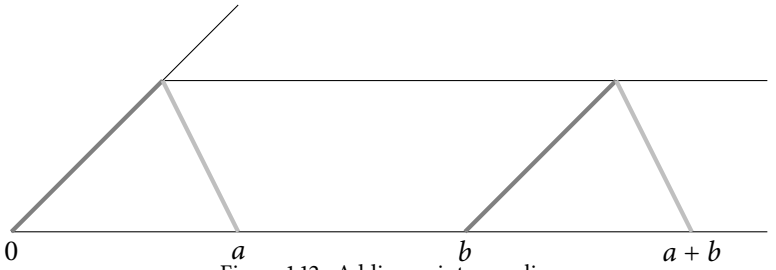

Figure 1.12 : Adding points on a line

The product construction also requires a point 1 on the line (the "unit of length"), and various parallels now allow us to "magnify" the distance from 0 to b by the distance from 0 to a, producing the point ab.

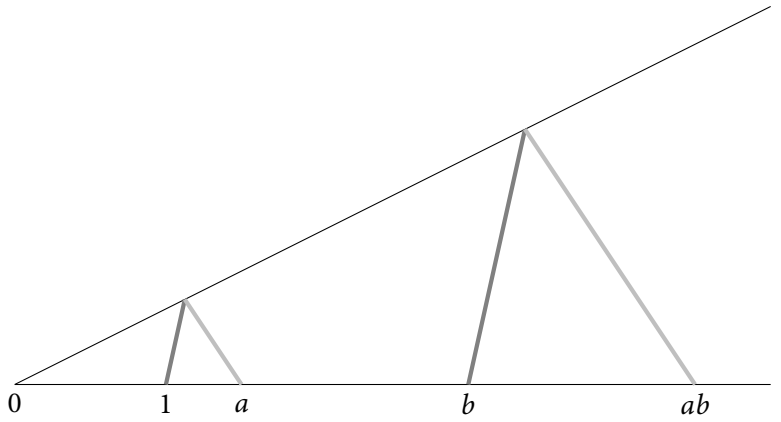

Figure 1.13 : Multiplying points on a line

With the help of the congruence axioms one can prove that the sum and product operations just defined have the following algebraic prop-

erties, the *field properties* (also used as the *axioms* that define a field):

$$a + b = b + a \qquad a \cdot b = b \cdot a \qquad \text{(commutativity)}$$

$$a + (b + c) = (a + b) + c \qquad a \cdot (b \cdot c) = (a \cdot b) \cdot c \qquad \text{(associativity)}$$

$$a + 0 = a \qquad a \cdot 1 = a \qquad \text{(identity)}$$

$$a + (-a) = 0 \qquad a \cdot a^{-1} = 1 \text{ for } a \neq 0 \qquad \text{(inverse)}$$

$$a \cdot (b + c) = a \cdot b + a \cdot c \qquad \text{(distributivity)}$$

It is easiest to deduce the field properties from the congruence axioms, but there is in fact a pure incidence axiom—the so-called *Pappus theorem*—from which all the field properties follow with the help of the other incidence axioms.[4] Thus the algebraic structure of a field emerges from axioms that Euclid almost completely overlooked: the incidence axioms describing how points and lines interact.

The order axioms give the points on a line an ordering, $\leq$, with the properties that, for any a, b, c:

- $a \leq a$,
- if $a \neq b$ then either $a < b$ or $b < a$, but not both,
- if $a \leq b$ and $b \leq c$ then $a \leq c$.

The order relation meshes with the field properties to produce an *ordered field*. Its defining properties, beyond the field properties above, are that:

- if $a \leq b$ then $a + c \leq b + c$,
- if $0 \leq a$ and $0 \leq b$ then $0 \leq ab$.

Finally, the Archimedean and completeness axioms say that the order relation is *Archimedean* and *complete* in the sense described by those axioms. It can be proved that a *complete Archimedean ordered field is isomorphic to the field* $\mathbb{R}$ *of real numbers*. Given such a field $\mathbb{F}$, the idea of the proof is to build a copy of $\mathbb{R}$ inside $\mathbb{F}$ in the following stages. (Readers not yet familiar with the construction of the real numbers as Dedekind cuts may wish to take these steps on faith and confirm them later when reading chapter 2.)

[4] Incidentally, the field properties can be proved in the setting of *projective geometry*, where all axioms are incidence axioms and the parallel axiom is replaced by the axiom that any two lines meet in a unique point. The above constructions can be carried out in this setting when we call one line a "line at infinity" and call lines "parallel" when they meet on the line at infinity.

1. From the element $1 \in \mathbb{F}$ build the "positive integers" of $\mathbb{F}$, namely

$$1, \quad 1+1, \quad 1+1+1, \quad 1+1+1+1, \ldots,$$

   using the + operation of $\mathbb{F}$.
2. Build "integers" of $\mathbb{F}$ using 0 and the − operation.
3. Build "rational numbers" of $\mathbb{F}$ using inverse and product operations.
4. Use the order and completeness properties of $\mathbb{F}$ to build "real numbers" of $\mathbb{F}$ as Dedekind cuts in the "rational numbers" of $\mathbb{F}$.
5. Check that the "real numbers" of $\mathbb{F}$ exhaust the members of $\mathbb{F}$ and have the same properties as the actual real numbers.

This proof shows that any complete Archimedean ordered field is essentially the "same" as $\mathbb{R}$, so every line in Hilbert's geometry is essentially the real number line. The next question is: how well do we understand $\mathbb{R}$?

1.5 WELL-ORDERING AND THE AXIOM OF CHOICE

In Book V of the *Elements*, Euclid gave a very sophisticated treatment of the geometric line and its relationship to the rational numbers. He stopped short of declaring irrational points to *be* numbers, but he essentially showed that each point is approximated arbitrarily closely by rational numbers. This means that each point is *determined by* rational numbers (those to the left of the point, for example), so *we need only accept infinite sets as mathematical objects* in order to view points as arithmetical objects.

However, until the mid-nineteenth century, most mathematicians rejected the idea of infinite sets as mathematical objects. They were influenced by the ancient Greek distinction between "potential" and "actual" infinity. For example, it was permissible to view the natural numbers as an open-ended process—start with 0 and keep adding 1—but not as a completed or "actual" entity $\mathbb{N} = \{0, 1, 2, \ldots\}$. Today, this seems a rather hair-splitting distinction, because—as far as anyone knew in the mid-nineteenth century—all infinite sets could be viewed as "potential" infinities.

For example, the integers, $\mathbb{Z}$, can be viewed as a potential infinity by listing them in the order

$$0, \quad 1, \quad -1, \quad 2, \quad -2, \quad 3, \quad -3, \quad \ldots .$$

The positive rationals can likewise be viewed as a potential infinity by listing them in the order

$$\frac{1}{1}, \quad \frac{2}{1}, \quad \frac{1}{2}, \quad \frac{3}{1}, \quad \frac{1}{3}, \quad \frac{4}{1}, \quad \frac{3}{2}, \quad \frac{2}{3}, \quad \frac{1}{4}, \quad \dots .$$

(The rule is to list fractions m/n in order of the sums $m + n$: first those with $m+n = 2$, then those with $m+n = 3$, then those with $m+n = 4$, and so on.) And then we can view *all* the rationals, $\mathbb{Q}$, as a potential infinity by listing positive and negative elements alternately as we did with $\mathbb{Z}$:

$$0, \quad \frac{1}{1}, \quad -\frac{1}{1}, \quad \frac{2}{1}, \quad -\frac{2}{1}, \quad \frac{1}{2}, \quad -\frac{1}{2}, \quad \frac{3}{1}, \quad -\frac{3}{1}, \quad \frac{1}{3}, \quad -\frac{1}{3}, \quad \dots .$$

Thus $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$, which we now regard as sets, could all be finessed as "potential" infinities by mathematicians who were fastidious about the distinction between potential and actual.

A much more serious problem arose in 1874, when Cantor showed that $\mathbb{R}$ is not by any means a potential infinity.

Uncountability

The means by which we showed the sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ to be potential infinities was by counting, or *ordering* their members in a sequence:

$$\text{1st member}, \quad \text{2nd member}, \quad \text{3rd member}, \quad \dots$$

—with an implied process for counting members that reaches each member at some finite stage. Cantor (1874) showed that $\mathbb{R}$ is *uncountable* in the sense that *no such ordering of* $\mathbb{R}$ *exists*.

He showed that any sequence $x_1, x_2, x_3, \dots$ of real numbers fails to include some real number x. In fact, given the decimal expansions of $x_1, x_2, x_3, \dots$ we can *compute* the decimal expansion of x. For example, we can use the rule:

$$n\text{th decimal digit of } x = \begin{cases} 1 & \text{if } n\text{th decimal digit of } x_n \neq 1 \\ 2 & \text{if } n\text{th decimal digit of } x_n = 1. \end{cases}$$

Then $x \neq$ each x_n because x differs from x_n in the nth decimal place.

Thus if we accept $\mathbb{R}$ we have to accept it as an *actual* infinity. The proof given here is essentially one given by Cantor (1891). It is, incidentally, a harbinger of many proofs about $\mathbb{R}$ that we will see later in this book.

Given an arbitrary object, such as a sequence or a function, we prove existence of some other object by *computing it from* the given object. The computation of one object relative to others is seldom noticed in classical analysis—in fact, many mathematicians have thought that Cantor's proof is *nonconstructive*—but it is important, as we will see in later chapters.

Well-ordering

Cantor's theorem shows that $\mathbb{R}$ cannot be ordered in the simple way that $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ can: 1st member, 2nd member, 3rd member, … . Nevertheless Cantor (1883) stated his belief in a more general kind of order:

> In a later article I shall discuss the law of thought that says that it is always possible to bring any *well-defined* set into the *form* of a *well-ordered* set—a law which seems to me fundamental and momentous and quite astonishing. (Ewald (1996), vol. II, p. 886)

Cantor called a set S well-ordered if the ordering is such that every nonempty subset T of S has a *least* member. This is clearly the case for the orderings of $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ above, where each member is labeled with a positive integer (take the member of T whose integer is least). It is also the case for the following ordering of $\mathbb{Z}$,

$$0, \quad 1, \quad 2, \quad 3, \quad \ldots, \quad -1, \quad -2, \quad -3, \quad -4, \quad \ldots,$$

in which 0 and the positive integers precede all the negative integers. If T is a nonempty subset of $\mathbb{Z}$ then the least member of T in the above ordering is

the least non-negative integer in T, if T has any non-negative members,

or

the greatest negative integer in T, if T has only negative members.

When it comes to $\mathbb{R}$, however, all human ingenuity fails to find a well-ordering of the real numbers. The usual ordering $<$ fails dismally, because subsets such as $\{x \in \mathbb{R} : 0 < x\}$ have no least member. Thus Cantor was very bold to assume that well-orderings exist for all "well-defined" sets—which surely include $\mathbb{R}$.

The Well-ordering Theorem and Zermelo's Axioms

Cantor perhaps thought that his "fundamental law of thought" should be an axiom of set theory. But he did not suggest a set of axioms for set

theory, so it remained unclear whether well-ordering should be an axiom or a theorem. The picture became clearer when Zermelo (1904) *proved* well-ordering from an intuitively simpler assumption, now known as the *axiom of choice*.

Axiom of choice (AC). For any set X of nonempty sets x there is a *choice function*; that is, a function f such that $f(x) \in x$ for each $x \in X$.

To provide a precise framework for his proof (and at the same time to clear up some doubts about the foundations of set theory) Zermelo (1908) gave the first set of axioms for set theory. Within his system, now called Z, it was possible to prove that AC is *equivalent* to the well-ordering theorem. Fraenkel (1922) strengthened one of Zermelo's axioms, obtaining a system now known as ZF set theory.

The ZF axioms of set theory have remained stable since 1922 and have become generally accepted as a foundation for all of mainstream mathematics, at least when supplemented by AC. Indeed, it was proved in ZF that AC is equivalent to many sought-after theorems, including the well-ordering theorem, that were apparently not provable outright in ZF.

This put AC in a position, relative to ZF, like that of the parallel axiom relative to the other axioms of Euclid (or, more precisely, relative to the other axioms of Hilbert). Theorems proved equivalent to AC in ZF were not of clear interest until it was known that AC is *not* provable in ZF. This was done by Cohen (1963). As Beltrami in 1868 did for the parallel axiom, Cohen showed the unprovability of AC by constructing a *model* of ZF in which AC is false. His construction, like Beltrami's, was a breakthrough that completely changed the face of the subject. It is unfortunately too technical to be described in this book, but we can describe some of its consequences.

A Mathematical Equivalent of the Axiom of Choice

Like the parallel axiom in geometry, AC in set theory occupies an important position "above" the basic (ZF) axioms. ZF cannot prove AC, but ZF is a good base theory because it can prove that AC is equivalent to many other interesting statements of set theory and general mathematics. In this sense, AC is the "right axiom" to prove these statements. As we know, one such statement is the well-ordering theorem. Another is the following property of *vector spaces* over an arbitrary field $\mathbb{F}$. (We

defined a *real* vector space in section 1.3. The definition of an arbitrary vector space is the same, except with $\mathbb{F}$ in place of $\mathbb{R}$.)

Existence of a vector space basis. *Any vector space V has a* basis*; that is, a subset U of vectors $\boldsymbol{u}$ such that:*

(i) For any $\boldsymbol{v} \in V$ there are $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in U$ and $f_1, \ldots, f_k \in \mathbb{F}$ such that $\boldsymbol{v} = f_1\boldsymbol{u}_1 + \cdots + f_k\boldsymbol{u}_k$. ("U spans V.")

(ii) For any distinct $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in U$ and $f_1, \ldots, f_k \in \mathbb{F}$, $\mathbf{0} = f_1\boldsymbol{u}_1 + \cdots + f_k\boldsymbol{u}_k$ if and only if $f_1 = \cdots = f_k = 0$. ("U is an independent set.")

The existence of a basis is clear for finite-dimensional real vector spaces, where we can take the basis vectors $\boldsymbol{u}$ to be the unit points on the coordinate axes. The first case in which a basis is hard to find—in fact utterly mysterious—is when $\mathbb{R}$ is viewed as a vector space over $\mathbb{Q}$. Hamel (1905) showed existence of a basis with the help of a well-ordering of $\mathbb{R}$, but the so-called *Hamel basis* is no easier to define than a well-ordering of $\mathbb{R}$ itself.

Thus it is no surprise that all proofs of the existence of a basis for an arbitrary vector space depend on AC. We now know that AC is unavoidable because Blass (1984) showed that existence of such a basis can be proved *equivalent* to AC in ZF.

1.6 LOGIC AND COMPUTABILITY

The previous sections of this chapter suggest that the real number system $\mathbb{R}$ is an essential part of the foundations of mathematics. When we turn to analysis, in the next chapter, the unavoidability of $\mathbb{R}$ will become even more obvious. At the same time, we have seen that our understanding of $\mathbb{R}$ can never be complete, if only because of the uncountability of $\mathbb{R}$.

Since we cannot list all real numbers we certainly cannot list all *facts* about real numbers, let alone set up an axiom system for proving them. This observation is the first step on the road towards the profound theorems of Gödel (1931) and Turing (1936) about unprovable theorems and unsolvable algorithmic problems—a road we will describe in more detail in chapter 4.

Gödel's theorem rules out any possibility of a complete axiom system for analysis. Yet it also presents an opportunity. If we are lucky we may be able to find a *base theory* for analysis, in which we can prove that sought-after theorems are equivalent to certain axioms—axioms that play a role,

like that of the parallel axiom in geometry or AC in set theory, of attracting desirable theorems into their "orbit" of equivalent theorems.

This indeed is what happens. We now know a good base theory, called RCA_0, and at least four *set existence axioms* that play this role for theorems of analysis. Moreover, the axioms are of increasing strength, in the sense that each implies the one before, so they classify theorems of analysis by increasing strength. The crucial axioms state "set existence" rather than "real number existence" because it is technically convenient to encode real numbers by sets of natural numbers (see next chapter for details). The axioms in question state there is a set of natural numbers n corresponding to each property $\varphi(n)$ in a certain class.

For RCA_0 we assert set existence for the class of *computable* properties $\varphi(n)$. These are the properties for which there is an algorithm that decides, for each n, whether $\varphi(n)$ holds. It turns out, because *non*computable properties exist, that RCA_0 is too weak to prove many important theorems of analysis. But RCA_0 can prove many equivalences, since these often involve computing an object (such as a sequence or a function) from a given object. For example, RCA_0 cannot prove the Bolzano-Weierstrass theorem, but it can prove that Bolzano-Weierstrass is equivalent to an axiom stating the existence of sets realizing each *arithmetically definable* property $\varphi(n)$. Thus, if we add the latter axiom to RCA_0, we obtain a stronger system in which Bolzano-Weierstrass is provable.

In this way we find, rather surprisingly, that most of the well-known theorems of analysis can be assigned a precise level of "strength." They are either at the lowest level—provable in RCA_0—or at a higher level represented by one of four set existence axioms. In this book we focus mainly on the lower three levels, where most of the well-known theorems of analysis are known to reside (see chapters 6 and 7).

Arithmetization

From the discussion above we can see that a study of arithmetic and computation will be needed before we can define the system RCA_0. Arithmetic itself is axiomatized in a fairly standard way that goes back to the *Peano axioms* of Peano (1889). But before that we have to talk about *arithmetization*—both in the nineteenth-century sense of making analysis "arithmetical," and in the 1930s sense of making logic and computation "arithmetical."

The remarkable convergence of analysis and computation to a common source in arithmetic is what makes the reverse mathematics of analysis possible. The arithmetization of analysis is discussed in chapter 2, computation is discussed in chapter 4, and its arithmetization in chapter 5. We also give a refresher course on the real numbers and continuity in chapter 3, including classical proofs of the best-known theorems.