This result complements the constructions for the rational operations $+, -, \times$, and $\div$ we gave in Chapter 1. The constructibility of these and $\sqrt{}$ was first pointed out by Descartes in his book *Géométrie* of 1637. Rational operations and $\sqrt{}$ are in fact *precisely* what can be done with straightedge and compass. When we introduce coordinates in Chapter 3 we will see that any "constructible point" has coordinates obtainable from the unit length 1 by $+, -, \times, \div$, and $\sqrt{}$.

Exercises

Now that we know how to construct the $+, -, \times, \div$, and $\sqrt{}$ of given lengths, we can use algebra as a shortcut to decide whether certain figures are constructible by straightedge and compass. If we know that a certain figure is constructible from the length $(1+\sqrt{5})/2$, for example, then we know that the figure is constructible—period—because the length $(1+\sqrt{5})/2$ is built from the unit length by the operations $+, \times, \div$, and $\sqrt{}$.

This is precisely the case for the regular pentagon, which was constructed by Euclid in Book IV, Proposition 11, using virtually all of the geometry he had developed up to that point. We also need nearly everything we have developed up to this point, but it fills less space than four books of the *Elements*!

The following exercises refer to the regular pentagon of side 1 shown in Figure 2.20 and its diagonals of length x.

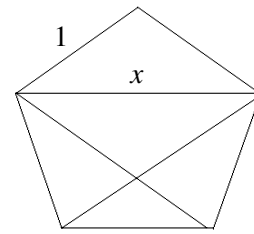

Figure 2.20: The regular pentagon

2.8.1 Use the symmetry of the regular pentagon to find similar triangles implying

$$\frac{x}{1} = \frac{1}{x-1},$$

that is, $x^2 - x - 1 = 0$.

2.8.2 By finding the positive root of this quadratic equation, show that each diagonal has length $x = (1+\sqrt{5})/2$.

2.8.3 Now show that the regular pentagon is constructible.

2.9 Discussion

Euclid found the most important axiom of geometry—the parallel axiom—and he also identified the basic theorems and traced the logical connections between them. However, his approach misses certain fine points and is not logically complete. For example, in his very first proof (the construction of the equilateral triangle), he assumes that certain circles have a point in common, but none of his axioms guarantee the existence of such a point. There are many such situations, in which Euclid assumes something is true because it *looks* true in the diagram.

Euclid's theory of area is a whole section of his geometry that seems to have no geometric support. Its concepts seem more like arithmetic—addition, subtraction, and proportion—but its concept of multiplication is not the usual one, because multiplication of more than three lengths is not allowed.

These gaps in Euclid's approach to geometry were first noticed in the 19th century, and the task of filling them was completed by David Hilbert in his *Grundlagen der Geometrie* (Foundations of Geometry) of 1899. On the one hand, Hilbert introduced axioms of *incidence* and *order*, giving the conditions under which lines (and circles) meet. These justify the belief that "geometric objects behave as the pictures suggest." On the other hand, Hilbert replaced Euclid's theory of area with a genuine arithmetic, which he called *segment arithmetic*. He defined the sum and product of segments as we did in Section 1.4 and proved that these operations on segments have the same properties as ordinary sum and product. For example,

$$a+b=b+a, \quad ab=ba, \quad a(b+c)=ab+ac, \quad \text{and so on.}$$

In the process, Hilbert discovered that the Pappus and Desargues theorems (Exercises 1.4.3 and 1.4.4) play a decisive role.

The downside of Hilbert's completion of Euclid is that it is lengthy and difficult. Nearly 20 axioms are required, and some key theorems are hard to prove. To some extent, this hardship occurs because Hilbert insists on geometric definitions of $+$ and $\times$. He wants numbers to come from "inside" geometry rather than from "outside". Thus, to prove that $ab=ba$ he needs the theorem of Pappus, and to prove that $a(bc)=(ab)c$ he needs the theorem of Desargues.

Even today, the construction of segment arithmetic is an admirable feat. As Hilbert pointed out, it shows that Euclid was right to believe that the

theory of proportion could be developed without new geometric axioms. Still, it is somewhat quixotic to build numbers "inside" Euclid's geometry when they are brought from "outside" into nearly every other branch of geometry. It is generally easier to build geometry on numbers than the reverse, and Euclidean geometry is no exception, as I hope to show in Chapters 3 and 4.

This is one reason for bypassing Hilbert's approach, so I will merely list his axioms here. They are thoroughly investigated in Hartshorne's *Geometry: Euclid and Beyond* or Hilbert's own book, which is available in English translation. Hartshorne's book has the clearest available derivation of ordinary geometry and segment arithmetic from the Hilbert axioms, so it should be consulted by anyone who wants to see Euclid's approach taken to its logical conclusion.

There is another reason to bypass Hilbert's axioms, apart from their difficulty. In my opinion, Hilbert's greatest geometric achievement was to build arithmetic, not in *Euclidean* geometry, but in *projective* geometry. As just mentioned, Hilbert found that the keys to segment arithmetic are the Pappus and Desargues theorems. These two theorems do not involve the concept of length, and so they really belong to a more primitive kind of geometry. This primitive geometry (projective geometry) has only a handful of axioms—*fewer than the usual axioms for arithmetic*—so it is more interesting to build arithmetic inside it. It is also less trouble, because we do not have to *prove* the Pappus and Desargues theorems. We will explain how projective geometry contains arithmetic in Chapters 5 and 6.

Hilbert's axioms

The axioms concern undefined objects called "points" and "lines," the related concepts of "line segment," "ray," and "angle," and the relations of "betweenness" and "congruence." Following Hartshorne, we simplify Hilbert's axioms slightly by stating some of them in a stronger form than necessary.

The first group of axioms is about *incidence*: conditions for points to lie on lines or for lines to pass through points.

I1. For any two points A, B, a unique line passes through A, B.

I2. Every line contains at least two points.

I3. There exist three points not all on the same line.

I4. For each line $\mathscr{L}$ and point P not on $\mathscr{L}$ there is a unique line through P not meeting $\mathscr{L}$ (parallel axiom).

The next group is about *betweenness* or *order*: a concept overlooked by Euclid, probably because it is too "obvious." The first to draw attention to betweenness was the German mathematician Moritz Pasch, in the 1880s. We write $A*B*C$ to denote that B is between A and C.

B1. If $A*B*C$, then A,B,C are three points on a line and $C*B*A$.

B2. For any two points A and B, there is a point C with $A*B*C$.

B3. Of three points on a line, exactly one is between the other two.

B4. Suppose A,B,C are three points not in a line and that $\mathscr{L}$ is a line not passing through any of A,B,C. If $\mathscr{L}$ contains a point D between A and B, then $\mathscr{L}$ contains either a point between A and C or a point between B and C, but not both (Pasch's axiom).

The next group is about *congruence of line segments* and *congruence of angles*, both denoted by $\cong$. Thus, $AB \cong CD$ means that AB and CD have equal length and $\angle ABC \cong \angle DEF$ means that $\angle ABC$ and $\angle DEF$ are equal angles. Notice that C2 and C5 contain versions of Euclid's Common Notion 1: "Things equal to the same thing are equal to each other."

C1. For any line segment AB, and any ray $\mathscr{R}$ originating at a point C, there is a unique point D on $\mathscr{R}$ with $AB \cong CD$.

C2. If $AB \cong CD$ and $AB \cong EF$, then $CD \cong EF$. For any AB, $AB \cong AB$.

C3. Suppose $A*B*C$ and $D*E*F$. If $AB \cong DE$ and $BC \cong EF$, then $AC \cong DF$. (Addition of lengths is well-defined.)

C4. For any angle $\angle BAC$, and any ray $\overrightarrow{DF}$, there is a unique ray $\overrightarrow{DE}$ on a given side of $\overrightarrow{DF}$ with $\angle BAC \cong \angle EDF$.

C5. For any angles α,β,γ, if $\alpha \cong \beta$ and $\alpha \cong \gamma$, then $\beta \cong \gamma$. Also, $\alpha \cong \alpha$.

C6. Suppose that ABC and DEF are triangles with $AB \cong DE$, $AC \cong DF$, and $\angle BAC \cong \angle EDF$. Then, the two triangles are congruent, namely $BC \cong EF$, $\angle ABC \cong \angle DEF$, and $\angle ACB \cong \angle DFE$. (This is SAS.)

Then there is an axiom about the intersection of circles. It involves the concept of points *inside* the circle, which are those points whose distance from the center is less than the radius.

E. Two circles meet if one of them contains points both inside and outside the other.

Next there is the so-called *Archimedean axiom*, which says that no length can be "infinitely large" relative to another.

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

Finally, there is the so-called *Dedekind axiom*, which says that the line is *complete*, or has *no gaps*. It implies that its points correspond to real numbers. Hilbert wanted an axiom like this to force the plane of Euclidean geometry to be the same as the plane $\mathbb{R}^2$ of pairs of real numbers.

D. Suppose the points of a line $\mathscr{L}$ are divided into two nonempty subsets $\mathscr{A}$ and $\mathscr{B}$ in such a way that no point of $\mathscr{A}$ is between two points of $\mathscr{B}$ and no point of $\mathscr{B}$ is between two points of $\mathscr{A}$. Then, a unique point P, either in $\mathscr{A}$ or $\mathscr{B}$, lies between any other two points, of which one is in $\mathscr{A}$ and the other is in $\mathscr{B}$.

Axiom D is not needed to derive any of Euclid's theorems. They do not involve all real numbers but only the so-called *constructible* numbers originating from straightedge and compass constructions. However, who can be sure that we will never need nonconstructible points? One of the most important numbers in geometry, π, is nonconstructible! (Because the circle cannot be squared.) Thus, it seems prudent to use Axiom D so that the line is complete from the beginning.

In Chapter 3, we will take the real numbers as the starting point of geometry, and see what advantages this may have over the Euclid–Hilbert approach. One clear advantage is *access to algebra*, which reduces many geometric problems to simple calculations. Algebra also offers some conceptual advantages, as we will see.

3

Coordinates

PREVIEW

Around 1630, Pierre de Fermat and René Descartes independently discovered the advantages of numbers in geometry, as *coordinates*. Descartes was the first to publish a detailed account, in his book *Géométrie* of 1637. For this reason, he gets most of the credit for the idea and the coordinate approach to geometry became known as *Cartesian* (from the old way of writing his name: Des Cartes).

Descartes thought that geometry was as Euclid described it, and that numbers merely *assist* in studying geometric figures. But later mathematicians discovered objects with "non-Euclidean" properties, such as "lines" having more than one "parallel" through a given point. To clarify this situation, it became desirable to *define* points, lines, length, and so on, and to *prove* that they satisfy Euclid's axioms.

This program, carried out with the help of coordinates, is called the *arithmetization of geometry*. In the first three sections of this chapter, we do the main steps, using the set $\mathbb{R}$ of real numbers to define the *Euclidean plane* $\mathbb{R}^2$ and the points, lines, and circles in it. We also define the concepts of distance and (briefly) angle, and show how some crucial axioms and theorems follow. However, arithmetization does much more.

- It gives an algebraic description of constructibility by straightedge and compass (Section 3.4), which makes it possible to prove that certain figures are *not* constructible.
- It enables us to define what it means to "move" a geometric figure (Section 3.6), which provides justification for Euclid's proof of SAS, and raises a new kind of geometric question (Section 3.7): What kinds of "motion" exist?

3.1 The number line and the number plane

The set $\mathbb{R}$ of real numbers results from filling the gaps in the set $\mathbb{Q}$ of rational numbers with *irrational* numbers, such as $\sqrt{2}$. This innovation enables us to consider $\mathbb{R}$ as a *line*, because it has no gaps and the numbers in it are ordered just as we imagine points on a line to be. We say that $\mathbb{R}$, together with its ordering, is a *model* of the line. One of our goals in this chapter is to use $\mathbb{R}$ to build a model for all of Euclidean plane geometry: a structure containing "lines," "circles," "line segments," and so on, with all of the properties required by Euclid's or Hilbert's axioms.

The first step is to build the "plane," and in this we are guided by the properties of parallels in Euclid's geometry. We imagine a pair of perpendicular lines, called the *x-axis* and the *y-axis*, intersecting at a point O called the *origin* (Figure 3.1). We interpret the axes as number lines, with O the number 0 on each, and we assume that the positive direction on the x-axis is to the right and that the positive direction on the y-axis is upward.

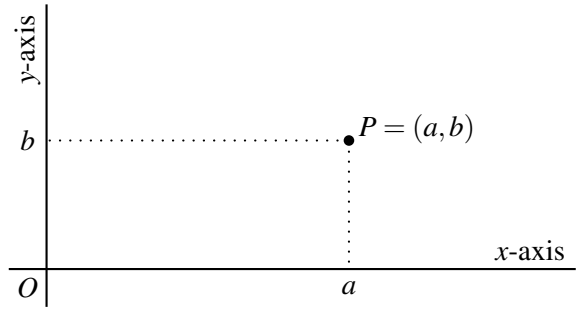

Figure 3.1: Axes and coordinates

Through any point P, there is (by the parallel axiom) a unique line parallel to the y-axis and a unique line parallel to the x-axis. These two lines meet the x-axis and y-axis at numbers a and b called the x- and y-*coordinates* of P, respectively. It is important to remember which number is on the x-axis and which is on the y-axis, because obviously the point with x-coordinate $= 3$ and y-coordinate $= 4$ is different from the point with x-coordinate $= 4$ and y-coordinate $= 3$ (just as the intersection of 3rd Street and 4th Avenue is different from the intersection of 4th Street and 3rd Avenue).

To keep the x-coordinate a and the y-coordinate b in their places, we use the *ordered pair* (a,b). For example, $(3,4)$ is the point with x-coordinate $=3$ and y-coordinate $=4$, whereas $(4,3)$ is the point with x-coordinate $=4$ and y-coordinate $=3$. The ordered pair (a,b) specifies P uniquely because any other point will have at least one different parallel passing through it and hence will differ from P in either the x- or y-coordinate.

Thus, given the existence of a *number line* $\mathbb{R}$ whose points are real numbers, we also have a *number plane* whose points are ordered pairs of real numbers. We often write this number plane as $\mathbb{R}\times\mathbb{R}$ or $\mathbb{R}^2$.

3.2 Lines and their equations

As mentioned in Chapter 2, one of the most important consequences of the parallel axiom is the Thales theorem and hence the proportionality of similar triangles. When coordinates are introduced, this allows us to define the property of straight lines known as *slope*. You know from high-school mathematics that slope is the quotient "rise over run" and, more importantly, that the value of the slope does not depend on which two points of the line define the rise and the run. Figure 3.2 shows why.

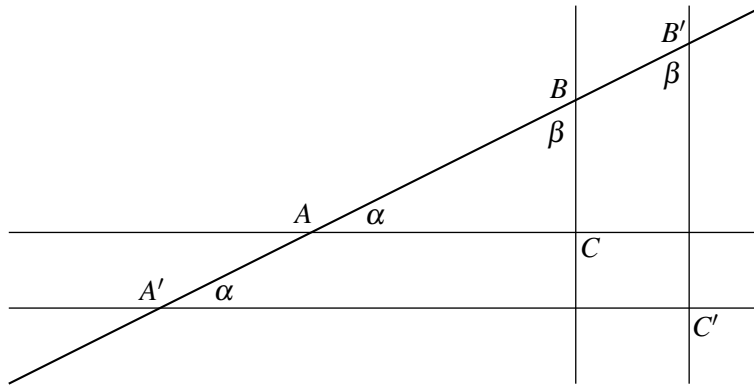

Figure 3.2: Why the slope of a line is constant

In this figure, we have two segments of the same line:

- AB, for which the rise is $|BC|$ and the run is $|AC|$, and
- $A'B'$, for which the rise is $|B'C'|$ and the run is $|A'C'|$.

The angles marked α are equal because AC and $A'C'$ are parallel, and the angles marked β are equal because the BC and $B'C'$ are parallel. Also, the angles at C and C' are both right angles.

Thus, triangles ABC and $A'B'C'$ are similar, and so their corresponding sides are proportional. In particular,

$$\frac{|BC|}{|AC|} = \frac{|B'C'|}{|A'C'|},$$

that is, slope = constant.

Now suppose we are given a line of slope a that crosses the y-axis at the point Q where $y = c$ (Figure 3.3). If $P = (x, y)$ is any point on this line, then the rise from Q to P is $y - c$ and the run is x. Hence

$$\text{slope} = a = \frac{y-c}{x}$$

and therefore, multiplying both sides by x, $y - c = ax$, that is,

$$y = ax + c.$$

This equation is satisfied by all points on the line, and only by them, so we call it the *equation of the line*.

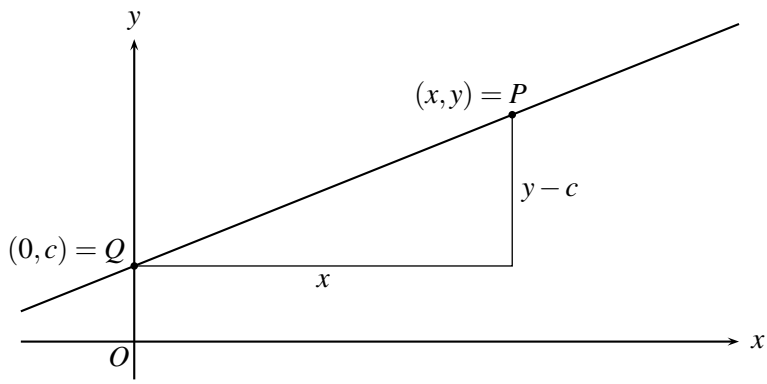

Figure 3.3: Typical point on the line

Almost all lines have equations of this form; the only exceptions are lines that do not cross the y-axis. These are the vertical lines, which also do not have a slope as we have defined it, although we could say they have *infinite slope*. Such a line has an equation of the form

$$x = c, \quad \text{for some constant } c.$$

Thus, all lines have equations of the form

$$ax + by + c = 0, \quad \text{for some constants } a, b, \text{ and } c,$$

called a *linear* equation in the variables x and y.

Up to this point we have been following the steps of Descartes, who viewed equations of lines as *information deduced from Euclid's axioms* (in particular, from the parallel axiom). It is true that Euclid's axioms prompt us to describe lines by linear equations, but we can also take the opposite view: Equations *define* what lines and curves are, and they provide a *model* of Euclid's axioms—showing that geometry follows from properties of the real numbers.

In particular, if a line is defined to be the set of points (x, y) in the number plane satisfying a linear equation then we can prove the following statements that Euclid took as axioms:

- there is a unique line through any two distinct points,
- for any line $\mathscr{L}$ and point P outside $\mathscr{L}$, there is a unique line through P not meeting $\mathscr{L}$.

Because these statements are easy to prove, we leave them to the exercises.

Exercises

Given distinct points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$, suppose that $P = (x, y)$ is any point on a line through P_1 and P_2.

3.2.1 By equating slopes, show that x and y satisfy the equation

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y - y_1}{x - x_1} \quad \text{if} \quad x_2 \neq x_1.$$

3.2.2 Explain why the equation found in Exercise 3.2.1 is the equation of a straight line.

3.2.3 What happens if $x_2 = x_1$?

Parallel lines, not surprisingly, turn out to be lines with the *same slope*.

3.2.4 Show that distinct lines $y = ax + c$ and $y = a'x + c'$ have a common point unless they have the same slope ($a = a'$). Show that this is also the case when one line has infinite slope.

3.2.5 Deduce from Exercise 3.2.4 that the parallel to a line $\mathscr{L}$ is the unique line through P with the same slope as $\mathscr{L}$.

3.2.6 If $\mathscr{L}$ has equation $y = 3x$, what is the equation of the parallel to $\mathscr{L}$ through $P = (2, 2)$?

3.3 Distance

We introduce the concept of *distance* or *length* into the number plane $\mathbb{R}^2$ much as we introduce lines. First we see what Euclid's geometry *suggests* distance should mean; then we turn around and take the suggested meaning as a definition.

Suppose that $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ are any two points in $\mathbb{R}^2$. Then it follows from the meaning of coordinates that there is a right-angled triangle as shown in Figure 3.4, and that $|P_1P_2|$ is the length of its hypotenuse.

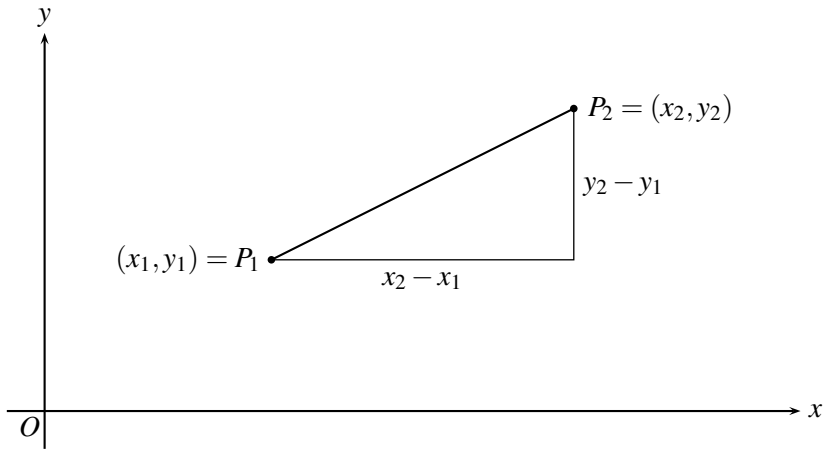

Figure 3.4: The triangle that defines distance

The vertical side of the triangle has length $y_2 - y_1$, and the horizontal side has length $x_2 - x_1$. Then it follows from the Pythagorean theorem that

$$|P_1P_2|^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2,$$

and therefore,

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

Thus, it is sensible to *define* the distance $|P_1P_2|$ between any two points P_1 and P_2 by the formula (*). If we do this, the Pythagorean theorem is virtually "true by definition." It is certainly true when the right-angled triangle has a vertical side and a horizontal side, as in Figure 3.4. And we will see later how to rotate any right-angled triangle to such a position (without changing the lengths of its sides).

The equation of a circle

The distance formula (*) leads immediately to the equation of a circle, as follows. Suppose we have a circle with radius r and center at the point $P = (a,b)$. Then any point $Q = (x,y)$ on the circle is at distance r from P, and hence formula (*) gives:

$$r = |PQ| = \sqrt{(x-a)^2 + (y-b)^2}.$$

Squaring both sides, we get

$$(x-a)^2 + (y-b)^2 = r^2.$$

We call this the *equation of the circle* because it is satisfied by any point (x,y) on the circle, and only by such points.

The equidistant line of two points

A circle is the set of points equidistant from a point—its center. It is also natural to ask: What is the set of points equidistant from *two* points in $\mathbb{R}^2$? Answer: *The set of points equidistant from two points is a line.*

To see why, let the two points be $P_1 = (a_1,b_1)$ and $P_2 = (a_2,b_2)$. Then a point $P = (x,y)$ is equidistant from P_1 and P_2 if $|PP_1| = |PP_2|$, that is, if x and y satisfy the equation

$$\sqrt{(x-a_1)^2 + (y-b_1)^2} = \sqrt{(x-a_2)^2 + (y-b_2)^2}.$$

Squaring both sides of this equation, we get

$$(x-a_1)^2 + (y-b_1)^2 = (x-a_2)^2 + (y-b_2)^2.$$

Expanding the squares gives

$$x^2 - 2a_1x + a_1^2 + y^2 - 2b_1y + b_1^2 = x^2 - 2a_2x + a_2^2 + y^2 - 2b_2y + b_2^2.$$

The important thing is that the x^2 and y^2 terms now cancel, which leaves the *linear* equation

$$2(a_2 - a_1)x + 2(b_2 - b_1)y + (b_1^2 - b_2^2) = 0.$$

Thus, the points $P = (x,y)$ equidistant from P_1 and P_2 form a line. □