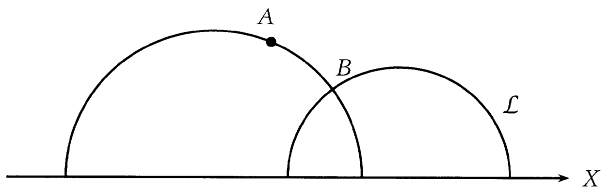

FIGURE 3.17 Perpendicular "lines."

Now we construct the reflection C of A in $\mathcal{L}$. This is the point opposite A on the non-Euclidean circle $\mathcal{K}$ with radius BA. $\mathcal{K}$ itself is therefore a Euclidean circle through A and C perpendicular to the semicircle through A and C (Figure 3.18).

3.9.4. Describe a ruler and compass construction of the Euclidean center of this circle.

This completes the proof that non-Euclidean "ruler and compass" constructions can be done by Euclidean ruler and compass. It follows that if we take, say, the line segment from $(0, 1)$ to $(0, 2)$ as the non-Euclidean unit of length, then all non-Euclidean constructible points have coordinates expressible by rational operations and square roots (by Section 3.2). In particular, it is impossible to construct the point $(0, \sqrt[3]{2})$, by Exercises 3.2.5* to 3.2.8*. This leads to a surprising conclusion.

3.9.5. The line segment from $(0, 1)$ to $(0, \sqrt[3]{2})$ has $1/3$ the non-Euclidean length of the line segment from $(0, 1)$ to $(0, 2)$. Why? Deduce that trisection of a "line" segment by "ruler and compass" is not always possible in the non-Euclidean plane.

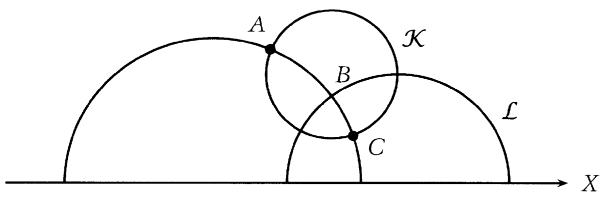

FIGURE 3.18 Non-Euclidean circle and diameters.

3.10 Discussion

Algebra and Geometry

The main message of the last two chapters is that geometry has two sides. First there is the visual, self-contained, synthetic side, which seems intuitively natural; then the algebraic, analytic side, which takes over when intuition fails and integrates geometry into the larger world of mathematics. Visualization will surely continue to inspire new discoveries in geometry, but it is equally likely that algebra will rule geometry as long as it is more efficient and more conducive to mathematical unity. A rigorous synthetic development of geometry requires too many complicated axioms that, unlike those of algebra, are not easily used in other parts of mathematics.

A second important message is that there is an apparent conflict between these two sides, and that this conflict has been very fruitful for the development of mathematics. As we know, the conflict began with the discovery of irrational lengths, such as $\sqrt{2}$. At the time, numbers were rational by definition, so irrational lengths could *not* be numbers, and hence geometry could not be based on arithmetic. Doing geometry without arithmetic turned out to be fruitful, however, because it led to Euclid's *Elements*, the most influential mathematics book of all time. In fact, it was only when Euclid's influence began to wane, in the 19th century, that mathematicians finally considered resolving the conflict between arithmetic and geometry, by extending the concept of number.

The latter development was also extraordinarily fruitful, as completion of the number concept not only clarifies the nature of points and lines, but also of curves and other objects too complicated to be grasped by Euclid's methods. For example, we shall see in Chapters 5 and 9 how completeness of the real numbers enables us to define lengths of curves and areas of curved regions.

The 19th century also saw a great enlargement in the scope of algebra, which allowed the operations $+$, $-$, $\times$, and $\div$ to be applied to objects that are not necessarily numbers. Of course, it is helpful to use the symbol $+$ only for a function that behaves like ordinary addition on numbers, so first it was necessary to find the characteristic properties of ordinary $+$, $-$, $\times$, and $\div$, and to describe them as simply as possible. This led to the definitions of *ring* and *field*, whose characteristic properties were listed in Section 1.4.

At this stage (around 1900) it became clear that the concept of *field* was the appropriate algebraic setting for geometry. We have already seen (in Sections 3.1 to 3.5) how to build Euclid's geometry using the field $\mathbb{R}$ of real numbers. Conversely, it is possible to build a field from Euclid's geometric concepts. The field consists of "positive" elements l and their additive inverses $-l$, and each l is a "length." Lengths are added, multiplied, and divided using the constructions of Exercises 3.2.1 and 3.2.2, and it follows from the axioms of geometry that the lengths and their additive inverses indeed form a field. In fact, if we also assume completeness, the field turns out to be nothing but $\mathbb{R}$.

Thus doing geometry analytically, using real number coordinates, is *almost* equivalent to doing geometry synthetically. The only extra ingredient in analytic geometry is completeness, which amounts to assigning a number to each point on the line. This of course is precisely the step the Greeks refused to take. Does this mean they missed analytic geometry only by a whisker, because of their scruples over irrationals, or was the concept of $\mathbb{R}$ really remote from Greek mathematics?

The Jump from $\mathbb{Q}$ to $\mathbb{R}$

Dedekind's definition of real numbers as cuts in $\mathbb{Q}$ is exquisitely simple, but deceptive in a way, because it hides the fact that $\mathbb{R}$ is a far less comprehensible set than $\mathbb{Q}$. The set $\mathbb{Q}$ is simpler than it looks, being similar in character to the set $\mathbb{N}$. Of course, $\mathbb{N}$ is an infinite set, but we can comprehend it as the result of the process of starting with 1 and repeatedly adding 1. It is not necessary to imagine this process actually *completed*, only continued indefinitely, to grasp the meaning of $\mathbb{N}$, because any *particular* member of $\mathbb{N}$ is reached after a finite amount of time. The infinite set $\mathbb{Z}$ can be grasped in the same way, as the result of a process that starts with 0 and alternately adds 1 and changes sign. This is why it makes sense to write

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

The process that generates the list $0, 1, -1, 2, -2, 3, -3, \ldots$ is clear, and it is also clear that any member of $\mathbb{Z}$ will eventually appear.

To capture the rationals by a list-generating process is only slightly more complicated. First we generate the positive rationals, by the process that produces them in the following order:

$\frac{1}{1}$, (reduced fractions whose top and bottom line sum to 2)
$\frac{2}{1}, \frac{1}{2}$, (reduced fractions whose top and bottom line sum to 3)
$\frac{3}{1}, \frac{1}{3}$, (reduced fractions whose top and bottom line sum to 4)
$\frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}$, (reduced fractions whose top and bottom line sum to 5)
$\frac{5}{1}, \frac{1}{5}$, (reduced fractions whose top and bottom line sum to 6)
$\vdots$

Then $\mathbb{Q}$ itself can be listed by alternating positive and negative numbers the way we did with $\mathbb{Z}$. Thus, when viewed merely as an infinite set, $\mathbb{Q}$ is just as comprehensible as $\mathbb{N}$.

The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are called *countable* because the listing processes give each of them a first member, second member, third member, and so on. And because the members of $\mathbb{Z}$ or $\mathbb{Q}$ are thereby paired with members of $\mathbb{N}$ (first member with 1, second member with 2 and so on), all three sets are reckoned to be the "same size" or, as we say, the same *cardinality*.

It was a great surprise to mathematicians when Cantor discovered in 1874 that *not* all infinite sets have the same cardinality. In particular, $\mathbb{R}$ is *uncountable*, and hence of greater cardinality than $\mathbb{N}$, because there is no way to list its members. This means that $\mathbb{R}$ can only be comprehended (if at all) as a *completed whole*. To understand $\mathbb{R}$, we not only have to grasp the individual cuts in $\mathbb{Q}$, we have to grasp them all at once!

This is what makes the concept of $\mathbb{R}$ really remote from Greek mathematics. The Greeks were willing to accept a "potential" infinity, such as the process for generating $\mathbb{N}$, but not the "actual" infinity of the set $\mathbb{N}$ itself. This was no obstacle to elementary number theory, because they could speak of an "arbitrary natural number" instead of the set $\mathbb{N}$. But they could not speak of an arbitrary *real* number, because this presupposes infinite subsets L and U of $\mathbb{Q}$, given without generating processes, and hence actually infinite.

Interestingly, the reason $\mathbb{R}$ can be comprehended only as a completed whole is precisely its completeness in the mathematical sense. We can, in fact, prove that any countable set A of numbers contains a gap, and hence is not complete. The idea is to use a list of members of A to find sequences $b_1, b_2, b_3, \ldots$ and $c_1, c_2, c_3, \ldots$ of members with

$$b_1 < b_2 < b_3 < \cdots < c_3 < c_2 < c_1,$$

and no member of A between the b_is and c_js. Then the separation of A into the set B of numbers $\leq$ some b_i and the set C of numbers $\geq$ some c_j gives a gap. Either the sequences of b_is and c_js are finite and there is gap between the greatest b_i and the least c_j; or else there is no greatest b_i and no least c_j, so (B, C) itself is a gap in $\mathbb{R}$. Given the job of separating A, the sequences $b_1, b_2, b_3, \ldots$ and $c_1, c_2, c_3, \ldots$ virtually define themselves.

Let $a_1, a_2, a_3, \ldots$ be a list of members of A, and let

$$\begin{aligned} b_1 &= \text{first number on the list that is not the} \\ &\quad\ \text{maximum member of } A; \\ c_1 &= \text{first number on the list that is} > b_1. \end{aligned}$$

Notice that this means b_1 is one of a_1 or a_2, and c_1 is the other. The remaining numbers $b_2, c_2, b_3, c_3, \ldots$ are chosen in that order by looking at $a_3, a_4, a_5, \ldots$ in turn and letting

$$\begin{aligned} b_{n+1} &= \text{first } a_k > b_1, b_2, \ldots, b_n \text{ and } < c_1, c_2, \ldots, c_n; \\ c_{n+1} &= \text{first } a_k > b_1, b_2, \ldots, b_{n+1} \text{ and } < c_1, c_2, \ldots, c_n. \end{aligned}$$

It follows immediately that there is no a_k between the sequences $b_1, b_2, b_3, \ldots$ and $c_1, c_2, c_3, \ldots$. If there were, we would look at it at some stage and *make it a member* of one of these sequences, which is a contradiction.

This argument is essentially the one given by Cantor himself in 1874. He later gave a more popular argument, called the *diagonal argument*, which is based on decimal expansions of the real numbers. The current argument is more elementary, however, and better suited to the point we wish to make—that the completeness of $\mathbb{R}$ implies its uncountability.

Incidentally, Cantor's argument gives another way to show that irrational numbers exist. If we take A to be the set of rationals, then (B, C) is a cut defining an irrational.

A Different Definition of Euclidean Geometry

One of the characteristic features of Euclidean plane geometry is the existence of *similarities*: mappings of $\mathbb{R} \times \mathbb{R}$ that multiply all lengths by a constant. A typical similarity is the *dilatation* dil_c (for $c \neq 0$) that sends each point (x, y) to the point (cx, cy) and consequently multiplies all lengths by c. We take the existence of similarities for granted in real life, in assuming that scale models, maps, and so on are faithful representations of real objects.

At the same time, the existence of scale models means that length is not really an essential concept in Euclidean geometry. The important properties of a triangle, for example, are not the lengths of its sides, but the *ratios* of the lengths (which determine the angles of the triangle). In fact, the theorems of Euclid's geometry are really about ratios of lengths, not about lengths themselves.

For this reason, we might very well define Euclidean geometry by using the group of similarities of $\mathbb{R} \times \mathbb{R}$ rather than the group of isometries. Similarities do indeed form a group, because each similarity is the composite of a dilatation dil_c with an isometry f, and its inverse is the composite of f^{-1} with the inverse $\mathrm{dil}_{c^{-1}}$ of dil_c. Because similarities preserve the ratio of any two lengths, they preserve all angles. A very remarkable theorem shows, conversely, that similarities are the *only* maps of $\mathbb{R} \times \mathbb{R}$ that preserve all angles. The only proofs of this theorem I know of use complex analysis, which is well beyond the scope of this book. See, for example, Jones and Singerman (1987), p.200.

There is an equally remarkable theorem about the non-Euclidean plane that says that its only angle-preserving maps are the isometries. This means, in particular, that there are *no* maps of the non-Euclidean plane that multiply all lengths by a constant $\neq 1$. Hence beings in a non-Euclidean world would not enjoy the benefits of scale models and maps. On the other hand, they would be able to determine the size of a figure from its shape alone, because each shape (of a triangle, say) exists in only one size.