Thus, fh is an involution too; call it g. Finally, applying f^{-1} to both sides of $fh = g$, we get

$$h = f^{-1}g.$$

So h is the product of two involutions, $f^{-1} = f$ and g, as required. $\square$

We now consider Möbius transformations of the half plane, each of which is the unique extension of a linear fractional transformation of $\mathbb{RP}^1$. Such a function is determined by its values at three points on $\mathbb{RP}^1$, by "uniqueness of three-point maps." We use the same letter for a linear fractional transformation of $\mathbb{RP}^1$ and its extension to a Möbius transformation of the half plane, and we systematically use the fact that Möbius transformations preserve non-Euclidean lines and angles.

Three reflections theorem. *Any Möbius transformation of the half plane is the product of at most three reflections.*

The involution f in the proof above, which exchanges p, q and fixes r, necessarily maps the non-Euclidean line $\mathscr{L}$ from p to q into itself. Points of $\mathscr{L}$ near the end p are sent to points near the end q, and vice versa. It follows by continuity that some point u on $\mathscr{L}$ is fixed by f, and hence the unique non-Euclidean line $\mathscr{M}$ through u and ending at r is mapped into itself by f. Also, because any Möbius transformation preserves angles, $\mathscr{M}$ must be perpendicular to $\mathscr{L}$ (Figure 8.14). Thus, *f has the same effect on p, q, r as reflection in the line $\mathscr{M}$*, so *f is* this reflection by "uniqueness of three-point maps."

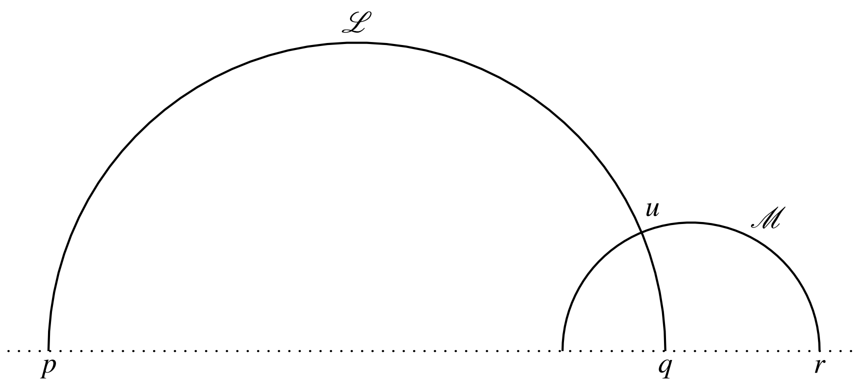

Figure 8.14: Lines involved in the involution f

Now consider the involution g, which is associated with a similar pair of lines $\mathscr{L}$ and $\mathscr{M}$. Only the names of their ends are different, but the reader is invited to draw them to keep track.

By the argument just used for f, the involution g is a reflection if it has a fixed point on $\mathbb{RP}^1$. In any case, g maps the line $\mathscr{L}$ with ends p and r into itself, exchanging the ends, so g has a fixed point u on $\mathscr{L}$ by the argument just used for f. Also, because g preserves angles, g maps the non-Euclidean line $\mathscr{M}$ through u and perpendicular to $\mathscr{L}$ into itself. Thus, if g has no fixed point on $\mathbb{RP}^1$, it necessarily exchanges the ends s and t of $\mathscr{M}$. But then *g has the same effect on the three points p, r, s as the product of reflections in $\mathscr{L}$ and $\mathscr{M}$*, so *g is* this product of reflections, by "uniqueness of three-point maps" again.

Thus, fg, which is an arbitrary Möbius transformation by the theorem above, is the product of at most three reflections. $\square$

Exercises

The argument above appeals to "continuity" to show the existence of a fixed point on a non-Euclidean line whose ends are exchanged by an involution. This argument is valid, and it may be justified by the *intermediate value theorem*, well known from real analysis courses. However, some readers may prefer an actual computation of the fixed point. One way to do it is as follows.

Suppose $f(x) = \frac{ax+b}{cx+d}$ and that $f(p) = q$, $f(q) = p$.

8.8.1 Deduce that $a = -d$ and $b = cpq - a(p+q)$, so that f has the form

$$f(x) = \frac{a(x-p-q)+cpq}{cx-a} = \frac{k(x-p-q)+pq}{x-k} \quad \text{if } c \neq 0.$$

8.8.2 Solve the equation

$$x = \frac{k(x-p-q)+pq}{x-k},$$

and hence show that the fixed points of f are

$$u = k \pm \sqrt{(k-p)(k-q)}.$$

8.8.3 Assuming that $(k-p)(k-q) < 0$, so one fixed point is in the upper half plane, show that its distance from the center $(p+q)/2$ of the semicircle with ends p and q is $|(p-q)/2|$.

8.8.4 Deduce from Exercises 8.8.1–8.8.3 that f has a fixed point on the non-Euclidean line with ends p and q.

8.9 Discussion

The non-Euclidean parallel hypothesis

It has often been said that the germ of non-Euclidean geometry is in Euclid's own work, because Euclid recognized the exceptional character of the parallel axiom and used it only when it was unavoidable. Later geometers noted several plausible equivalents of the parallel axiom, such as

- the equidistant curve of a line is a line,
- the angle sum of a triangle is π,
- similar figures of different sizes exist,

but no outright proof of it from Euclid's other axioms was found. On the contrary, attempts to derive a contradiction from the existence of many parallels—what we will call the *non-Euclidean parallel hypothesis*—led to a rich and apparently coherent geometry. This is the geometry we have been exploring in the half plane, now called *hyperbolic geometry*.

Hyperbolic geometry diverges from Euclidean geometry in the opposite direction from spherical geometry—for example, the angle sum of a triangle is $< \pi$, not $> \pi$—but the divergence is less extreme. The "lines" of spherical geometry violate all three of Euclid's axioms about lines, whereas the "lines" of hyperbolic geometry violate only the parallel axiom.

The first theorems of hyperbolic geometry were derived by the Italian Jesuit Girolamo Saccheri in an attempt to prove the parallel axiom. In his 1733 book, *Euclides ab omni naevo vindicatus* (Euclid cleared of every flaw), Saccheri assumed the non-Euclidean parallel hypothesis, and sought a contradiction. What he found were *asymptotic lines*: lines that do not meet but approach each other arbitrarily closely. This discovery was curious, and more curious at infinity, where Saccheri claimed that the asymptotic lines would meet *and* have a common perpendicular. Finding this "repugnant to the nature of a straight line," he declared a victory for Euclid.

But the common perpendicular at infinity is *not* a contradiction, and indeed (as we now know) it clearly holds in the half plane. There are non-Euclidean lines that approach each other arbitrarily closely in non-Euclidean distance, such as the unit semicircle and the line $x = 1$, and they have a common perpendicular at infinity—the x-axis. Saccheri had unwittingly discovered not a bug, but a key feature of hyperbolic geometry.

The non-Euclidean geometry of the hyperbolic plane began to take shape in the early 19th century. A small circle of mathematicians around Carl Friedrich Gauss (1777–1855) explored the consequences of the non-Euclidean parallel hypothesis, although Gauss did not publish on the subject through fear of ridicule. Gauss was the greatest mathematician of his time, but he was unwilling to publish "unripe" work, and he evidently felt that non-Euclidean geometry lacked a solid foundation. He knew of no concrete interpretation, or *model*, of non-Euclidean geometry, and in fact, none was discovered in his lifetime. It is a great irony that some of his own discoveries—in the geometry of curved surfaces and the geometry of complex numbers—can provide such models.

The first to publish comprehensive accounts of non-Euclidean geometry were Janos Bolyai in Hungary and Nikolai Lobachevsky in Russia. Around 1830 they discovered this geometry independently and became its first "true believers." The richness and coherence of their results convinced them that they had discovered a new geometric world, as real as the world of mainstream geometry and not needing its support. In a sense, they were right, but in their enthusiasm, they failed to notice another new geometric theory that could have been a valuable ally. Gauss's *Disquisitiones generales circa superficies curvas* (General investigations on curved surfaces) was published in 1827, but neither Bolyai, Lobachevsky, or Gauss noticed that it gives models of non-Euclidean geometry, at least in small regions.

The fundamental concept of Gauss's surface theory is the *curvature*, a quantity that is positive (and constant) for a sphere, zero for the plane and cylinder, and negative for surfaces that are "saddle-shaped" in the neighborhood of each point. In the *Disquisitiones*, Gauss investigated the relationship between the curvature of a surface and the behavior of its *geodesics*, which are its curves of shortest length and hence its "lines." He found, for example, that a geodesic triangle has

- angle sum $> \pi$ on a surface of positive curvature,
- angle sum π on a surface of zero curvature,
- angle sum $< \pi$ on a surface of negative curvature.

Moreover, if the curvature is *constant* and nonzero, then, in any geodesic triangle, (angle sum $-\pi$) is proportional to area. These results must have reminded Gauss of things he already knew in non-Euclidean geometry, so it is surprising that he failed to capitalize on them.

Close encounters between the actual and the hypothetical

The near agreement between geometry on surfaces of constant negative curvature and non-Euclidean geometry was the first of several close encounters over the next few decades. But usually the actual and hypothetical geometries passed each other like ships in a thick fog.

For example, in the late 1830s, the German mathematician Ferdinand Minding worked out the formulas of negative-curvature trigonometry. He found that they are like those of spherical trigonometry, but with hyperbolic functions in place of circular functions. At about the same time (and in the same journal!), Lobachevsky showed that the same formulas hold for triangles in his non-Euclidean plane. This would have been a nice time to introduce the name "hyperbolic geometry" for the non-Euclidean geometry of constant negative curvature, but apparently neither Minding nor Lobachevsky realized that they might have been talking about the same thing. Perhaps they were aware of a difficulty with the known surfaces of negative curvature: They are *incomplete* in the sense that their "lines" cannot be extended indefinitely. Hence, they fail to satisfy Euclid's second axiom for lines.

The simplest surface of constant negative curvature is called the *pseudosphere* (somewhat misleadingly, because constant curvature is about all it has in common with the sphere). It is more accurately known as the *tractroid*, because it is the surface of revolution of the curve known as the *tractrix*. The defining property of the tractrix is that its tangent has constant length a between the curve and the x-axis (left half of Figure 8.15).

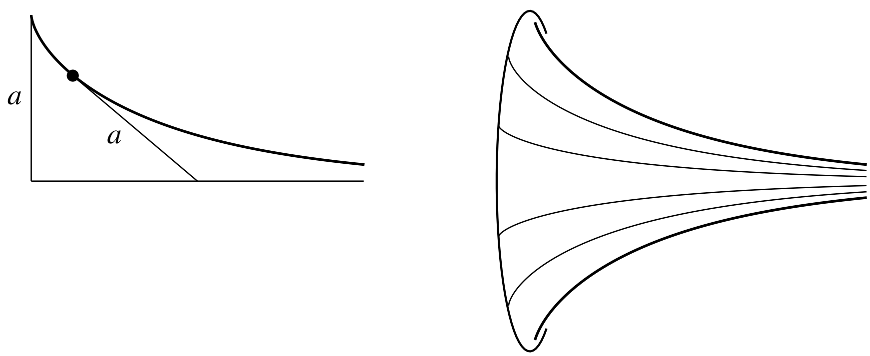

Figure 8.15: The tractrix and the tractroid

It is an unavoidable consequence of this definition that the tractrix has a singularity where the tangent becomes perpendicular to the x-axis. The tractroid likewise has an *edge* (like the rim of a trumpet), beyond which it cannot be smoothly continued. Hence, geodesics on the tractroid cannot be continued in both directions. In fact, the only geodesics on the tractroid that are infinite in even one direction are the rotated copies of the original tractrix. This problem is typical of what happens when one tries to construct a complete surface of constant negative curvature in ordinary space. The task was eventually shown to be impossible by Hilbert in 1901, but an obstacle to the construction of such surfaces was sensed much earlier.

In 1854, Gauss's student Bernhard Riemann showed a way round the obstacle by proposing an *abstract* or *intrinsic* definition of curved spaces—one that does not require a "flat" space to contain the "curved" one. This idea made it possible to define a complete surface, or indeed a complete n-dimensional space, of constant negative curvature. Riemann did exactly this, but once again non-Euclidean geometry sailed by unnoticed, as far as we know. (The elderly Gauss was very moved by Riemann's account of his discoveries. Whether he saw in them a vindication of non-Euclidean geometry, we will probably never know.)

Another close encounter occurred in 1859, when Arthur Cayley developed the concept of distance in projective geometry. He found that there is an invariant length for certain groups of projective transformations, such as those that map the circle into itself. In effect, he had discovered a model of the non-Euclidean plane, but he did not notice that his invariant length had the same properties as non-Euclidean length. Despite the efforts of Bolyai and Lobachevsky, non-Euclidean geometry remained an obscure subject until the 1860s.

Models of non-Euclidean geometry

Riemann died in 1859, and his ideas first bore fruit in Italy, where he had spent a lot of time in his final years. His most important successor was Eugenio Beltrami, who in 1868 finally brought non-Euclidean geometry and negative curvature together.

Beltrami's first discovery, in 1865, established the special role of constant curvature in geometry: *The surfaces of constant curvature are precisely those that can be mapped to the plane in such a way that geodesics go to straight lines*. The simplest example is the sphere, whose geodesics are great circles, the intersections of the sphere with the planes through its center. Great circles can be mapped to straight lines by projecting the sphere onto the plane from its center (Figure 8.16).

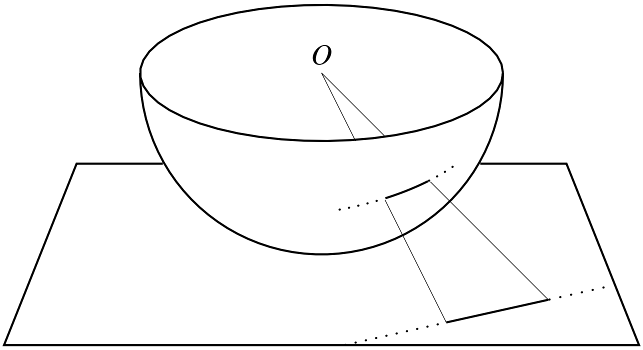

Figure 8.16: Central projection of the sphere

The geodesic-preserving map of the tractroid sends it to a wedge-shaped portion of the unit disk. The tractrix curves on the tractroid go to line segments ending at the sharp end of the wedge (Figure 8.17).

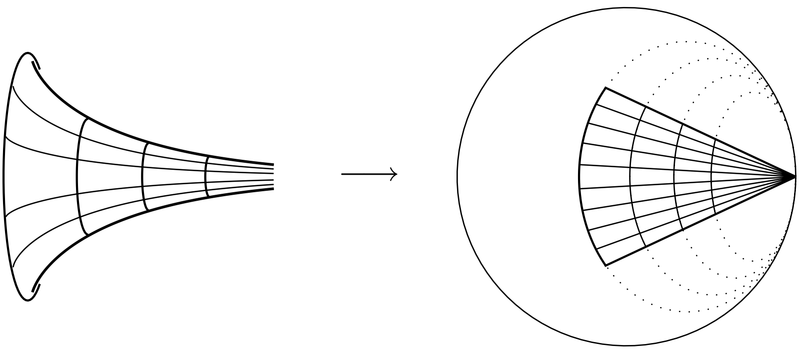

Figure 8.17: Geodesic-preserving map of the tractroid

Although this map preserves "lines," it certainly does not preserve length. Each tractrix curve has infinite length; yet it is mapped to a finite line segment in the disk. The appropriate length function for the disk assigns a "pseudodistance" to each pair of points, equal to the geodesic distance between the corresponding points on the tractroid. We do not need the formula here; the important thing is that *pseudodistance makes sense on* *the whole open disk*, that is, for all points inside the boundary circle. The curve of shortest pseudodistance between any two points in the open disk is the straight line segment between them, and the pseudodistance between any point and the boundary is infinite.

In 1868, Beltrami realized that this abstraction and extension of the tractroid is an *interpretation of the non-Euclidean plane*: a surface in which there is a unique “line” between any two points, “lines” are infinite, and the non-Euclidean parallel hypothesis is satisfied. Figure 8.18 shows why: Many “lines” through the point P do not meet the “line” $\mathscr{L}$.

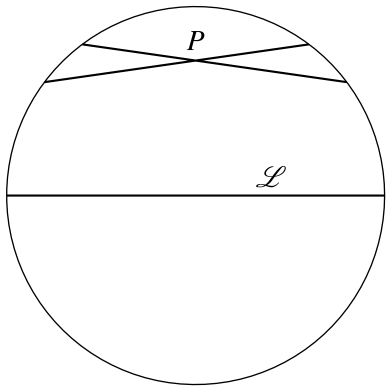

Figure 8.18: Why the non-Euclidean parallel hypothesis holds

Beltrami wrote two epic papers on models of non-Euclidean geometry in 1868, and English translations of them may be found in my book *Sources of Hyperbolic Geometry*. The first paper arrives at the non-Euclidean plane as an extension of the tractroid through the idea of “unwinding” infinitely thin sheets wrapped around it. (The dotted paths in the right half of Figure 8.17, all converging to the endpoint of the wedge, are the limit circles traced by circular sections of the tractroid as they unwind.) Beltrami was at pains to be as concrete as possible, because Riemann’s ideas were not well understood or accepted in 1868. However, at the end of the paper, Beltrami foreshadows the more abstract and general approach he intends to take in his second paper:

> where the most general principles of non-Euclidean geometry are considered independently of their possible relations with ordinary geometric entities. In the present work we have been