1 Hyperbolic Geometry

The purpose of this chapter is to give a bare bones introduction to hyperbolic geometry. Most of material in this chapter can be found in a variety of sources, for example:

- Alan Beardon's book, *The Geometry of Discrete Groups*,
- Bill Thurston's book, *The Geometry and Topology of Three Manifolds*,
- Svetlana Katok's book, *Fuchsian Groups*,
- John Ratcliffe's book, *Hyperbolic Geometry*.

The first 2 sections of this chapter might not look like geometry at all, but they turn out to be very important for the subject.

1.1 Linear Fractional Transformations

Suppose that

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is a 2×2 matrix with complex number entries and determinant 1. The set of these matrices is denoted by $SL_2(\boldsymbol{C})$. In fact, this set forms a group under matrix multiplication.

The matrix A defines a *complex linear fractional transformation*

$$T_A(z) = \frac{az + b}{cz + d}.$$

We will sometimes omit the word *complex* from the name, though we will always have in mind a complex linear fractional transformation when we say *linear fractional transformation*. Such maps are also called *Möbius transformations*,

Note that the denominator of $T_A(z)$ is nonzero as long as $z \neq -d/c$. It is convenient to introduce an extra point ∞ and define $T_A(-d/c) = \infty$. This definition is a natural one because of the limit

$$\lim_{z \to -d/c} |T_A(z)| = \infty.$$

The determinant condition guarantees that $a(-d/c) + b \neq 0$, which explains why the above limit works. We define $T_A(\infty) = a/c$. This makes sense because of the limit

$$\lim_{|z|\to\infty} T_A(z) = a/c.$$

Exercise 1. First introduce a metric on $\boldsymbol{C} \cup \infty$ so that $\boldsymbol{C} \cup \infty$ is homeomorphic to the unit sphere $S^2 \subset \boldsymbol{R}^3$. Prove that T_A is continuous with respect to this metric. (*Hint*: Use the limit formulas above to deal with the tricky points.)

Exercise 2. Establish the general formula

$$T_{AB} = T_A \circ T_B,$$

where $A, B \in SL_2(\boldsymbol{R})$. In particular (since A^{-1} exists) the inverse map T_A^{-1} exists. By Exercise 1, this map is also a continuous map of $\boldsymbol{C} \cup \infty$. Conclude that T_A is a homeomorphism of $\boldsymbol{C} \cup \infty$.

1.2 Circle Preserving Property

A *generalized circle* in $\boldsymbol{C} \cup \infty$ is either a circle in $\boldsymbol{C}$ or a set of the form $L \cup \infty$, where L is a straight line in $\boldsymbol{C}$. Topologically, the generalized circles are all homeomorphic to circles. In this section we will prove the following well-known result.

Theorem 1.1 *Let C be a generalized circle and let T be a linear fractional transformation. Then $T(C)$ is also a generalized circle.*

One can prove this result by a direct (though tedious) calculation, and there are also proofs which go through stereographic projection. For fun, I will give a rather unconventional proof. I'll prove 4 straightforward lemmas and then give the main argument.

Lemma 1.2 *Let C be any generalized circle in $\boldsymbol{C} \cup \infty$. Then there exists a linear fractional transformation T such that $T(\boldsymbol{R} \cup \infty) = C$.*

Proof: If C is a straight line (union ∞), then a suitable translation followed by rotation will work. So, consider the case when C is a circle. The linear fractional transformation

$$T(z) = \frac{z - i}{z + i}$$

maps $\boldsymbol{R} \cup \infty$ onto the unit circle C_0 satisfying the equation $|z| = 1$. The point is that every point $z \in \boldsymbol{R}$ is the same distance from i and $-i$, so that $|T(z)| = 1$. Next, one can find a map of the form $S(z) = az + b$ that carries C_0 to C. The composition $S \circ T$ does the job. ♠

Lemma 1.3 *Suppose that L is a closed loop in $\boldsymbol{C} \cup \infty$. Then there exists a generalized circle C that intersects L in at least 3 points.*

Proof: If L is contained in a straight line (union ∞) the result is obvious. Otherwise, L has 3 noncollinear points and, like any 3 noncollinear points, these lie on a common circle. ♠

Lemma 1.4 *Let $(z_1, z_2, z_3) = (0, 1, \infty)$. Let a_1, a_2, a_3 be a triple of distinct points in $\boldsymbol{R} \cup \infty$. Then there exists a linear fractional transformation that preserves $\boldsymbol{R} \cup \infty$ and maps a_i to z_i for $i = 1, 2, 3$.*

Proof: The map $T(z) = 1/(a_3 - z)$ carries a_3 to ∞, but does not necessarily do the right thing on the points a_1 and a_2. However, we can compose T by a suitable map of the form $z \to rz + s$ to fix the images of a_1 and a_2. ♠

Lemma 1.5 *Suppose T is a linear fractional transformation that fixes 0 and 1 and ∞. Then T is the identity map.*

Proof: Let

$$T(z) = \frac{az + b}{cz + d}.$$

The condition $T(0) = 0$ gives $b = 0$. The condition $T(\infty) = \infty$ gives $c = 0$. The condition $T(1) = 1$ gives $a = d$. Hence $T(z) = z$. ♠

Now we can give the main argument. Suppose that there is a linear fractional transformation T and a generalized circle C such that $T(C)$ is not a generalized circle. Composing T with the map from Lemma 1.2, we can assume that $C = \boldsymbol{R} \cup \infty$. By Lemma 1.3 there is a generalized circle D such that D and $T(\boldsymbol{R} \cup \infty)$ share at least 3 points. Call these 3 points c_1, c_2, c_3.

Again by Lemma 1.2, there is a linear fractional transformation S such that $S(\boldsymbol{R} \cup \infty) = D$. There are points $a_1, a_2, a_3 \in \boldsymbol{R} \cup \infty$ such that $S(a_j) = c_j$ for $j = 1, 2, 3$. Also, there are points $b_1, b_2, b_3 \in \boldsymbol{R} \cup \infty$ such that $T(b_j) = c_j$ for $j = 1, 2, 3$. By Lemma 1.4 we can find linear fractional transformations A and B, both preserving $\boldsymbol{R} \cup \infty$ such that $A(a_j) = z_j$ and $B(b_j) = z_j$ for $j = 1, 2, 3$. Here $(z_1, z_2, z_3) = (0, 1, \infty)$. The two maps

$$T \circ B^{-1}, \qquad S \circ A^{-1}$$

both map $(0, 1, \infty)$ to the same 3 points, namely (c_1, c_2, c_3). By Lemma 1.5, these maps coincide. However, note that

$$T \circ B^{-1}(\boldsymbol{R} \cup \infty) = T(\boldsymbol{R} \cup \infty)$$

is not a generalized circle and $S \circ A^{-1}(\boldsymbol{R} \cup \infty) = D$ is a generalized circle. This is a contradiction.

1.3 The Upper Half-Plane Model

Now we turn to hyperbolic geometry. Once we define the hyperbolic plane as a set of points, we will define what we mean by the lengths of curves in the hyperbolic plane.

Let $U \subset \boldsymbol{C}$ be the upper half-plane, consisting of points z with $\mathrm{Im}(z) > 0$. As a set, the hyperbolic plane is just U. However, we will describe a funny way of measuring the lengths of curves in U. Were we to use the ordinary method, we would just produce a subset of the Euclidean plane. So, given a differentiable curve $\gamma : [a, b] \to U$, we define

$$L(\gamma) = \int_a^b \frac{|\gamma'(t)|}{\mathrm{Im}(\gamma(t))}\, dt. \tag{1}$$

In words, the hyperbolic speed of the curve is the ratio of its Euclidean speed to its height above the real axis.

Here is a simple example. Consider the curve $\gamma : \boldsymbol{R} \to U$ defined by

$$\gamma(t) = i \exp(t).$$

Then the length of the portion of γ connecting $\gamma(a)$ to $\gamma(b)$, with $a < b$, is given by

$$\int_a^b \frac{\exp(t)}{\exp(t)}\, dt = \int_a^b dt = b - a.$$

The image of γ is an open vertical ray, but our formula tells us that this ray, measured hyperbolically, is infinite in both directions. Moreover, the formula tells us that γ is a unit speed curve: it accumulates $b - a$ units of length between time a and time b.

The hyperbolic distance between two points $p, q \in U$ is defined to be the infimum of the lengths of all piecewise differentiable curves connecting p to q. Let us consider informally what these shortest curves ought to look like. Suppose that p and q are very near the real axis, say

$$p = 0 + i\ 10^{-100}, \qquad q = 1 + i\ 10^{-100}.$$

The most obvious way to connect these two points would be to use the path

$$\gamma(t) = t + i\ 10^{-100}.$$

This curve traces out the bottom of the (Euclidean unit) square shown in Figure 10.1. Our formula tells us that this curve has length 10^{100}.

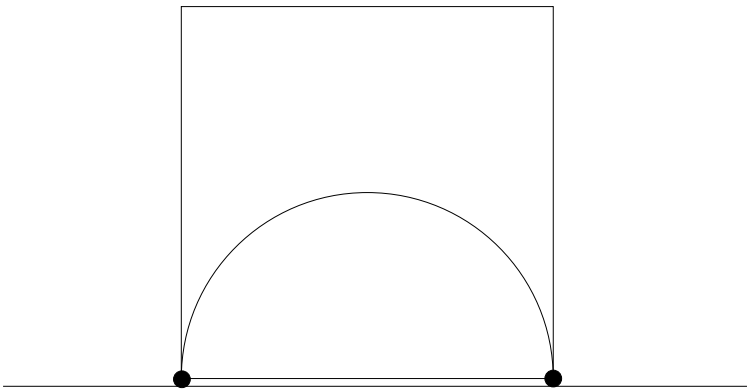

Figure 10.1. Some paths in the hyperbolic plane

Another thing we could do is go around the other three sides of the square. For the left vertical edge, we could use the path γ from our first calculation. This edge has length

$$\log(1) - \log(10^{-100}) = 100.$$

The top horizontal edge has height 1 and Euclidean length 1. So, this leg of the path has length 1. Finally, by symmetry, the length of the right vertical edge is 100. All in all, we have connected p to q by a path of length

201. This length is obviously much shorter than the first path. It pays to go upward because, so to speak, unit speed hyperbolic curves cover more ground the farther up they are. Our second path is much better than the first but certainly not the best. For openers, we could save some distance by rounding off the corners. We will show in §1.6 below that the shortest curves, or *geodesics*, in the hyperbolic plane are either arcs of vertical rays or arcs of circles that are centered on the real axis.

When U is equipped with the metric we have defined, we call U the *hyperbolic plane* and denote it by $\boldsymbol{H}^2$. So far we have talked about lengths of curves in $\boldsymbol{H}^2$, but we can also talk about angles. The angle between two differentiable and regular (i.e., nonzero speed) curves in $\boldsymbol{H}^2$ is defined simply to be the ordinary Euclidean angle between them. That is, the hyperbolic and Euclidean angle between two intersecting curves is just the Euclidean angle between the two tangent vectors at the point of intersection. So, in the upper half-plane model of hyperbolic geometry, the distances are distorted (from the Euclidean model) but the angles are not.

Now that we have talked about hyperbolic length and angles, we discuss hyperbolic area. Given how hyperbolic length relates to Euclidean length, it makes sense to say that the area of a small patch of the hyperbolic plane is the ratio of its Euclidean area to its height squared. Since the "height" of a patch varies throughout the patch, we really have something infinitesimal in mind. Thus, precisely, we define the hyperbolic area of a region $D \subset \boldsymbol{H}^2$ to be the integral

$$\int_D \frac{dx\ dy}{y^2}. \tag{2}$$

1.4 Another Point of View

An *inner product* on a real vector space V is a map $\langle\ ,\ \rangle : V \times V \to \boldsymbol{R}$ which satisfies the following properties:

- $\langle av + w, x\rangle = a\langle v, x\rangle + \langle w, x\rangle$ for all $a \in \boldsymbol{R}$ and $v, w, x \in V$.
- $\langle x, y\rangle = \langle y, x\rangle$.
- $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0$ if and only if $x = 0$.

You can remember this by noting that an inner product satisfies the same formal properties as the dot product.

For the moment, we care mainly about inner products on $\boldsymbol{R}^2$. At the point $z = x + iy$ we introduce the inner product

$$\langle v, w \rangle_z = \frac{1}{y^2}(v \cdot w). \tag{3}$$

We mean to apply this to vectors v and w that are "based at" z. We then define the hyperbolic *norm* to be

$$\|v\|_z = \sqrt{\langle v, v \rangle_z}. \tag{4}$$

With this definition, the length of $\gamma : [a, b] \to \boldsymbol{H}^2$ is given by

$$\int_a^b \|\gamma'(t)\|_{\gamma(t)} \, dt. \tag{5}$$

With this formalism, the notion of hyperbolic length looks much closer to the Euclidean notion. This way of doing things is the beginning of Riemannian geometry.

1.5 Symmetries

The hyperbolic metric has more symmetries than you might think. Say that a *real linear fractional transformation* is a linear fractional transformation T_A based on a matrix with real entries. In this case, $T_A(z) \in \boldsymbol{C}$ provided $z \in \boldsymbol{C} - \boldsymbol{R}$.

Exercise 3. Prove that $z \notin \boldsymbol{R}$ implies that $T_A(z) \notin \boldsymbol{R}$. Prove also that T_A maps $\boldsymbol{H}^2$ into itself.

The element T_A is a homeomorphism of $\boldsymbol{C} \cup \infty$ which preserves $\boldsymbol{H}^2$.

Exercise 4. We say that a real linear fractional transformation is *basic* if it has one of three forms:

- $T(z) = z + 1$.
- $T(z) = rz$.
- $T(z) = -1/z$.

Prove that any real linear fractional transformation is the composition of basic ones.

It turns out that these maps are all hyperbolic isometries. This is pretty obvious for the map $T(z) = z + 1$. The hyperbolic metric is built so that the second map is a hyperbolic isometry, and in a moment we will give two proofs of that fact. The really surprising thing is that the third map turns out to be a hyperbolic isometry as well.

Lemma 1.6 *The map $T(z) = rz$ is a hyperbolic isometry.*

First Proof. If γ is any curve in $\boldsymbol{H}^2$, then the dilated curve $T(\gamma)$ moves r times as fast in the Euclidean sense but is r times farther from the real axis. Hence $T(\gamma)$ and γ move at the same hyperbolic speed at corresponding points. So, if we connect points p and q by some curve γ we can connect the points $T(p)$ and $T(q)$ by the curve $T(\gamma)$, which has the same length—and vice versa. This shows that the distance from p to q is the same as the distance from $T(p)$ to $T(q)$. ♠

Second Proof. Suppose that v and w are two vectors based at $z \in \boldsymbol{H}^2$. Then we think of $dT(v) = rv$ and $dT(w) = rw$ as two vectors based at $T(z)$. Here dT is linear differential of T, i.e., the matrix of first partial derivatives. Looking at the formula in equation (3), we see that

$$\langle dT(v), dT(w)\rangle_{T(z)} = \langle rv, rw\rangle_{rz} = \frac{1}{r^2y^2}(rv \cdot rw) = \frac{1}{y^2}(v \cdot w) = \langle v, w\rangle_z.$$

So, T preserves the hyperbolic inner product at each point. Since the hyperbolic metric is defined entirely in terms of this family of inner products, T is an isometry. ♠

Exercise 5. Prove that the map $T(z) = -1/z$ is a hyperbolic isometry.

Combining Exercises 4 and 5, we see that any real linear fractional transformation is a hyperbolic isometry of $\boldsymbol{H}^2$. The space $SL_2(\boldsymbol{R})$ is a 3-dimensional manifold. So, $\boldsymbol{H}^2$ has a 3-dimensional group of symmetries!

Say that a *generalized circular arc* is an arc of a generalized circle. We already know that any linear fractional transformation maps generalized circles to circles. Hence, any real linear transformation maps generalized circular

arcs to generalized circular arcs.

Exercise 6. Prove that a real linear fractional transformation T has the following property: if a and b are two smooth curves in $\boldsymbol{H}^2$ which intersect at a point x and make an angle of θ, then $T(a)$ and $T(b)$ make the same angle θ at the point $T(x)$. (*Hint*: If you don't feel like grinding out the calculation, you can assume the result is false and then deduce that the differential dT fails to map circle to circles. In any case, the result is obvious for all the basic maps except $z \to -1/z$, and so it suffices to consider this one.)

1.6 Geodesics

In this section we will describe the shortest curves connecting two points in $\boldsymbol{H}^2$. We first consider the case of points p and q that lie on the imaginary axis.

Lemma 1.7 *The portion of the imaginary axis connecting p to q is the unique shortest curve in $\boldsymbol{H}^2$ that connects p to q.*

Proof: Our proof is very similar to the proof we gave in Lemma **??** for the spherical case. Consider the map F defined by the equation $F(x+iy) = iy$; see Figure 10.2. Looking at the definition of the hyperbolic metric, we see that F is hyperbolic speed nonincreasing. That is, if γ is a curve in $\boldsymbol{H}^2$, then the hyperbolic speed of $F(\gamma)$ at any point is at most the hyperbolic speed of γ at the corresponding point. Moreover, if the velocity of γ has any x-component at all, then $F(\gamma)$ is slower at the corresponding point. The idea here is that F does not change the y-component of the hyperbolic speed, but kills the x-component. The total hyperbolic length of γ is the integral of its hyperbolic speed. Thus the hyperbolic length of $F(\gamma)$ is less than the hyperbolic length of γ, unless γ travels vertically the whole time. Our result follows immediately from this. ♠

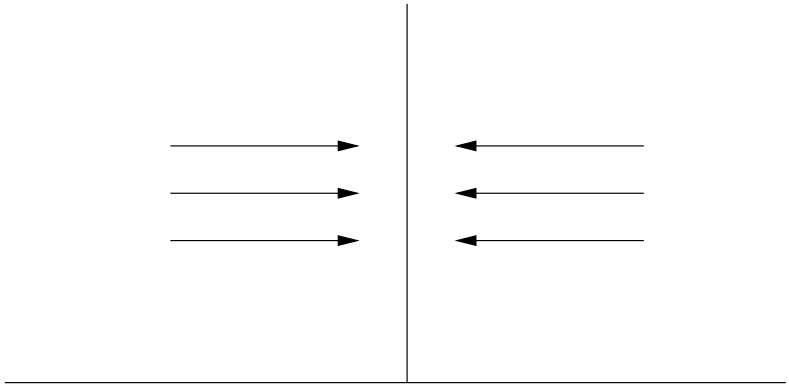

Figure 10.2. The map F

It follows from symmetry that the vertical rays in $\boldsymbol{H}^2$ are all geodesics. A vertical ray is the unique shortest path in $\boldsymbol{H}^2$ connecting any pair of points on that ray.

Exercise 7. Let p and q be two arbitary points in $\boldsymbol{H}^2$. Prove that there is a hyperbolic isometry—specifically, some linear fractional transformation—that carries p and q to points that lie on the same vertical ray.

Theorem 1.8 *Any two distinct points in $\boldsymbol{H}^2$ can be joined by a unique shortest path. This path is either a vertical line segment or else an arc of a circle that is centered on the real axis.*

Proof: We have already proved this result for points that lie on the same vertical ray. In light of Exercise 7, it suffices to prove, in general, that the image of a vertical ray under a linear fractional hyperbolic isometry is one of the two kinds of curves described in the theorem.

Let ρ be a vertical ray, and let T be a linear fractional transformation that is also a hyperbolic isometry. From the work in §1.2 we know that $T(\rho)$ is an arc of a circle. Since T preserves $\boldsymbol{R} \cup \infty$, both endpoints of this circular arc lie on $\boldsymbol{R} \cup \infty$. Finally, since T preserves angles, $T(\rho)$ meets $\boldsymbol{R}$ at right angles at any point where $T(\rho)$ intersects $\boldsymbol{R}$. If $T(\rho)$ limits on ∞, then $T(\rho)$ is another vertical ray. Otherwise, $T(\rho)$ is a semicircle, contained in a circle that is centered on the real axis. ♠

1.7 The Disk Model

Now that we have defined geodesics in the hyperbolic plane, we can go forward and define geodesic polygons. Before we do this, we would like to have another model in which to draw pictures. This other model is sometimes more convenient.

Let Δ be the open unit disk. There is a (complex) linear fractional map $M : \boldsymbol{H}^2 \to \Delta$ given by

$$M(z) = \frac{z - i}{z + i}. \tag{6}$$

This map does the right thing because $z \in \boldsymbol{H}^2$ is always closer to i than to $-i$ and so $|M(z)| < 1$. Since M maps circles to circles and preserves angles, M maps geodesics in $\boldsymbol{H}^2$ to circular arcs in Δ that meet the unit circle at right angles.

Sometimes it is convenient to draw pictures of geodesics in the unit disk rather than in the hyperbolic plane. So, when it comes time to draw pictures, we will be drawing circular arcs that meet the unit circle at right angles. The geodesics that go through the Euclidean center of Δ are just unit line segments. The rest of them "bend inward" toward the origin.

Exercise 8. Draw pictures of 10 geodesics in the disk model.

Rather than just think of the open unit disk Δ as a convenient place to draw pictures, we can also think of Δ as another model of $\boldsymbol{H}^2$. The cheapest way to do this is to say that the distance between the two points $p, q \in \Delta$ is defined to be the hyperbolic distance between the points $M^{-1}(p)$ and $M^{-1}(q)$ in $\boldsymbol{H}^2$.

A more direct approach is to define a new inner product at each point $z \in \Delta$. The formula is given by

$$\langle v, w \rangle_z = \frac{4v \cdot w}{(1 - |z|^2)^2}. \tag{7}$$

Once we have this inner product, we can directly define lengths of curves in Δ as in equation (5). Then we can define distances in Δ as in the upper half-plane model. It turns out that this new method produces the same result as the cheap method. The proof is a calculation similar to our second proof of Lemma 1.6. We just prove that M is an isometry relative to the inner product on $\boldsymbol{H}^2$ and the inner product on Δ.

Exercise 9. Prove that the map M is an isometry from $\boldsymbol{H}^2$ and Δ, when lengths are defined in terms of the inner product in equation (7). That is, prove that

$$\langle v, w \rangle_z = \langle dM(v), dM(w) \rangle_{M(z)}$$

for any pair of vectors v and w based at $z \in \boldsymbol{H}^2$.

The open unit disk Δ, equipped with its metric, is known as the *Poincaré disk model* of the hyperbolic plane. When T is a real linear fractional transformation, the map $M \circ T \circ M^{-1}$ is an isometry of Δ. Since M preserves angles, the hyperbolic angle between two curves in Δ is the same as the Euclidean angle between them. Thus, in both our models, Euclidean and hyperbolic angles coincide.

Before we continue, we mention one more piece of terminology. The *ideal boundary* of $\boldsymbol{H}^2$ is defined to be $\boldsymbol{R} \cup \infty$ in the upper half-plane model and the unit circle in the disk model. Points on the ideal boundary are called *ideal points*. The ideal points are not points in $\boldsymbol{H}^2$. They are considered "limit points" of geodesics in $\boldsymbol{H}^2$.

1.8 Geodesic Polygons

Now that we have our two models of the hyperbolic plane, and we know that the geodesics are, we are ready to consider geodesic polygons in the hyperbolic plane. To save words, we will use the term $\boldsymbol{H}^2$ rather loosely to refer to either of our two models of the hyperbolic plane. Since there is an isometry, namely M, carrying one model to the other, there doesn't seem to be much harm in doing this.

Say that a *geodesic polygon* in $\boldsymbol{H}^2$ is a simple closed path made from geodesic segments. Here, "simple" means that the path does not intersect itself. Say that a *solid* geodesic polygon is the region in $\boldsymbol{H}^2$ bounded by a geodesic polygon. It is convenient to allow some of the "vertices" of the polygon to be ideal points. We call such "vertices" by the name *ideal vertices*. The interior angle of a polygon at an ideal vertex is 0: the two geodesics both meet the ideal point perpendicular to the ideal boundary.

We point out a special geodesic triangle, called an *ideal triangle*. An ideal triangle is a geodesic triangle having 3 infinite geodesic sides and 3 ideal vertices; see Figure 10.3 below. The main result in this section, the

Gauss–Bonnet formula for hyperbolic geodesic triangles.

Theorem 1.9 *Let T be a geodesic triangle in the hyperbolic plane. The area of T equals π minus the sum of the interior angles of T. In particular, the sum of these interior angles is less than π.*

We will give the same kind of proof that we gave for the analogous result in §**??**.

Lemma 1.10 *Theorem 1.9 holds for ideal triangles.*

Proof: We are trying to prove that any ideal triangle has area π. By emma 1.4, we can move any one ideal triangle to any other using an isometry of $\boldsymbol{H}^2$. So, it suffices to prove this result for a single triangle. Let us prove this for the triangle T, in the upper half-plane model, with vertices -1 and 1 and ∞. We first observe that

$$\int_{y=y_0}^{\infty} \frac{1}{y^2} dy = 1/y_0.$$

Now we compute our area, using equation (2). Integrating in the y direction, we have

$$\text{area}(T) = \int_{x=-1}^{1} \int_{y=\sqrt{1-x^2}}^{\infty} \frac{1}{y^2} dy \; dx = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} dx = \pi.$$

The last integral is most easily done making the trigonometric substitution $x = \sin(t)$ and $dx = \cos(t)dt$. ♠

Let $T(\theta)$ denote a geodesic triangle having two vertices on the ideal boundary of $\boldsymbol{H}^2$ and one interior vertex having interior angle θ.

Lemma 1.11 *Theorem 1.9 holds for $T(\theta)$.*

Proof: Any two such triangles are isometric to each other. We first match up the interior vertices and then suitably rotate one triangle so that the sides emanating from the common vertex match. In particular, any incarnation of $T(\theta)$ has the same area. Let

$$f(\theta) = \pi - \text{area}(T(\theta)).$$

We want to show that $f(\theta) = \theta$ for all $\theta \in [0, \pi)$. We already know that $f(0) = 0$, by the previous result.

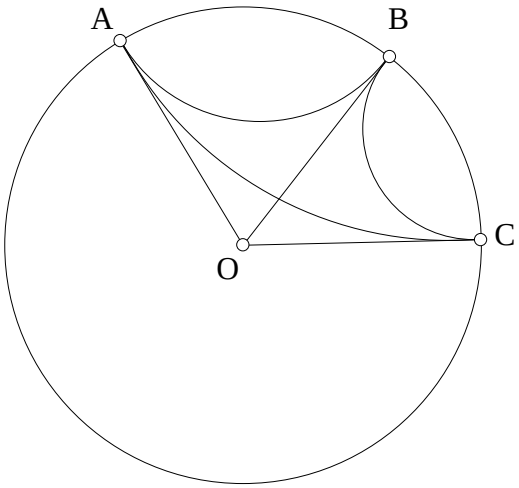

Figure 10.3. Two dissections

To analyze the general situation, we work in the disk model and choose $T(\theta)$ so that it has an interior vertex O at 0. Figure 10.3 shows a dissection proof that

$$f(\theta_1 + \theta_2) = f(\theta_1) + f(\theta_2),$$

as long as $\theta_1 + \theta_2 \leq \pi$. Just to make the picture clear, we point out the following:

- The triangle $T(\theta_1)$ has vertices O, A, B.
- The triangle $T(\theta_2)$ has vertices O, B, C.
- The triangle $T(\theta_1 + \theta_2)$ has vertices O, A, C.
- The triangle with vertices A, B, C is an ideal triangle.

To make this formula work even when $\theta_1 + \theta_2 = \pi$, we set $f(\pi) = \pi$. The quadrilateral we have drawn can be dissected in two ways. One way gives $A_1 + A_2$. The other way gives $A + \pi$. Here A_k is the area of $T(\theta_k)$ and A is the area of $T(\theta_1 + \theta_2)$.

Since $f(\pi) = \pi$, we can use our formula inductively to show $f(r\pi) = r\pi$ for any rational $r \in (0, 1)$. But the function f is pretty clearly continuous. Since f is the identity on a dense set, f is the identity everywhere. ♠

Now we take an arbitrary geodesic triangle and extend the sides so that they hit the ideal boundary of $\boldsymbol{H}^2$. Then we consider the dissection of the

ideal triangle defined by the (ideal) endpoints of these sides, as shown in Figure 10.4.

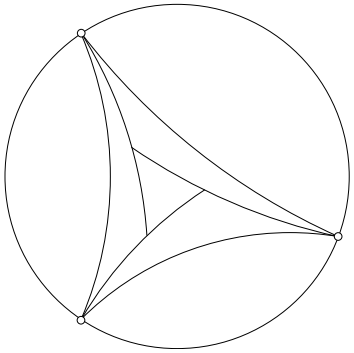

Figure 10.4. A Dissected ideal triangle.

The ideal triangle and also the three outer triangles are of the kind we have already considered. Theorem 1.9 holds true for these. The ideal triangle has area π, and the three outer triangles have areas α, β, and γ, the three interior angles of the inner triangle. Hence, the inner triangle has area $\pi - \alpha - \beta - \gamma$, as desired. This completes the proof.

A solid geodesic polygon P is *convex* if it has the following propery: if $p, q \in P$ are two points then the geodesic segment joining p and q is also contained in P. It is easy to prove, inductively, that any convex geodesic polygon can be decomposed into geodesic triangles.

Lemma 1.12 *The area of a convex geodesic n-gon is $(n-2)\pi$ minus the sum of the interior angles.*

Proof: Just decompose into triangles and then apply the triangle theorem multiple times. ♠

Exercise 10 (Challenge). Suppose that $\theta_1, \theta_2, \theta_3$ are three numbers whose sum is less than π. Prove that there is a hyperbolic geodesic triangle with angles $\theta_1, \theta_2, \theta_3$.

Exercise 11 (Challenge). Say that a geodesic triangle is δ-*thin* if every point in the interior of the (solid version of) triangle is within δ of a point on the boundary. Note that there is no universal δ so that all Euclidean triangles are δ-thin. Prove that all hyperbolic geodesic triangles are 10-thin. (The value $\delta = 10$ is far from optimal.)

1.9 Classification of Isometries

Let T be a real linear fractional transformation. If $T(\infty) = \infty$, then we have $T(z) = az + b$. If $T(\infty) \neq \infty$, then the equation $T(z) = z$ leads to a quadratic equation $az^2 + bz + c = 0$, with $a, b, c \in \boldsymbol{R}$. If T is not the identity, then there are 3 possibilities:

- T fixes one point in $\boldsymbol{H}^2$ and no other points.
- T fixes no points in $\boldsymbol{H}^2$ and one point in $\boldsymbol{R} \cup \infty$.
- T fixes no points in $\boldsymbol{H}^2$ and two points in $\boldsymbol{R} \cup \infty$.

T is called *elliptic*, *parabolic*, or *hyperbolic*, according to which possibility occurs. We will discuss these three cases in turn. Before we start, we mention a helpful construction. Given isometries g and T, we call $S = gTg^{-1}$ a *conjugate* of T. Note that g maps the fixed points of T to the fixed points of S.

Suppose T is elliptic. Working in the disk model, we can conjugate T so that the result S fixes the origin. In this case, S maps each geodesic through the origin to another geodesic through the origin. Moreover, S preserves the distances along these geodesics. From here, we see that S must be a rotation. So, in the disk model, all the elliptic isometries are conjugate to ordinary rotations.

Suppose that T is parabolic. Working in the upper half-plane model, we can conjugate T so that the result S fixes ∞. In this case $S(z) = az + b$. If $a \neq 1$, then S fixes an additional point in $\boldsymbol{R}$. Since this does not happen, $a = 1$. Hence $S(z) = z + b$. So, in the upper half-plane model, all parabolic isometries are conjugate to a translation.

Suppose that T is hyperbolic. Working in the upper half-plane model, we can conjugate T so that the result S fixes 0 and ∞. But then $S(z) = rz$ for some $r \neq 0$. So, in the upper half-plane model, all hyperbolic isometries are conjugate to dilations (or contractions).

Neither the parabolic elements nor the hyperbolic elements have fixed points in $\boldsymbol{H}^2$, but they still behave in a qualitatively different way. Considering the parabolic map $S(z) = z + b$, we see that there is no $\epsilon > 0$ such that S moves all points of $\boldsymbol{H}^2$ more than ϵ. For example, the hyperbolic distance between iy and $S(iy)$ tends to 0 as $y \to \infty$. On the other hand, if we examine the map $S(z) = rz$, we see that there is some $\epsilon > 0$ such that S moves all points of $\boldsymbol{H}^2$ by at least ϵ. Indeed, $\epsilon = |\log(r)|$.