Now $f(\mathbb{R}) = \mathbb{R}$, so f has real coefficients and is in fact an isometry of H^2. $\square$

In order to obtain Proposition 4.14, we need only extend the result of this proposition to take into account the position of the third point, which determines whether the isometry preserves or reverses orientation. To this end, note that the condition $d(w_1, w_3) = d(z_1, z_3)$ implies that w_3 lies on a circle of radius $d(z_1, z_3)$ centred at w_1; similarly, it also lies on a circle of radius $d(z_2, z_3)$ centred at w_3.

Assuming z_1, z_2, z_3 do not all lie on the same geodesic, there are exactly two points which lie on both circles, each an equal distance from the geodesic connecting z_1 and z_2. One of these will necessarily be the image of z_3 under the fractional linear transformation f found above; the other one is $(r \circ f)(z_3)$ where r denotes reflection in the geodesic η.

To better describe r, pick any point $z \in H^2$ and consider the geodesic ζ which passes through z and meets η orthogonally. Denote by $d(z, \eta)$ the distance from z to the point of intersection; then the reflection $r(z)$ is the point on ζ a distance $d(z, \eta)$ beyond this point. Alternatively, we may recall that the map $R\colon z \mapsto -\bar{z}$ is reflection in the imaginary axis, which is an orientation reversing isometry. There exists a unique fractional linear transformation g taking η to the imaginary axis; then r is simply the conjugation $g^{-1} \circ R \circ g$.

Exercise 4.15. Prove that the group of orientation preserving isometries of H^2 in the unit disc model is the group of all fractional linear transformations of the form

$$z \mapsto \frac{az + \bar{c}}{cz + \bar{a}}$$

where $a, c \in \mathbb{C}$ satisfy $a\bar{a} - c\bar{c} = 1$.

Lecture 29.

a. Classification of isometries. Now we turn to the task of classifying these isometries and understanding what they look like geometrically.

a.1. *Fixed points in the extended plane.* For the time being we restrict ourselves to orientation preserving isometries. We begin by considering the fractional linear transformation f as a map on all of $\mathbb{C}$, (or, more precisely, on the Riemann sphere $\mathbb{C} \cup \{\infty\}$) and look for fixed points, given by

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

Clearing the denominator and simplifying gives the quadratic equation

$$cz^2 + (d-a)z - b = 0$$

whose roots are

$$\begin{aligned} z &= \frac{1}{2c}\left(a - d \pm \sqrt{(a-d)^2 + 4bc}\right) \\ &= \frac{1}{2c}\left(a - d \pm \sqrt{(a+d)^2 - 4(ad-bc)}\right) \\ &= \frac{1}{2c}\left(a - d \pm \sqrt{(a+d)^2 - 4}\right). \end{aligned}$$

Note that the quantity $a+d$ is just the trace of the matrix of coefficients $X = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, which we already know has unit determinant. Let λ and μ be the eigenvalues of X; then $\lambda\mu = \det X = 1$, so $\mu = 1/\lambda$, and we have

$$a + d = \operatorname{Tr} X = \lambda + \mu = \lambda + \frac{1}{\lambda}$$

There are three possibilities to consider regarding the nature of the fixed point or points $z = f(z)$:

- **(E):** $|a+d| < 2$, corresponding to $\lambda = e^{i\alpha}$ for some $\alpha \in \mathbb{R}$. In this case there are two fixed points z and $\bar{z}$, with $\operatorname{Im} z > 0$ and hence $z \in H^2$.
- **(P):** $|a+d| = 2$, corresponding to $\lambda = 1$ (since X and $-X$ give the same transformation). In this case there is exactly one fixed point $z \in \mathbb{R}$.
- **(H):** $|a+d| > 2$, corresponding to $\mu < 1 < \lambda$. In this case, there are two fixed points $z_1, z_2 \in \mathbb{R}$.

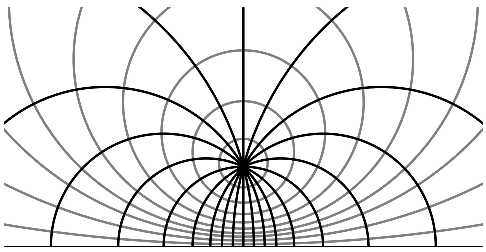

Figure 4.13. Geodesics passing through i and hyperbolic circles centred at i.

a.2. *Elliptic isometries.* Let us examine each of these in turn, beginning with **(E)**, where f fixes a unique point $z \in H^2$. Consider a geodesic γ passing through z. Then $f(\gamma)$ will also be a geodesic passing through z; let α be the angle it makes with γ at z. Then because f preserves angles, it must take any geodesic η passing through z to the unique geodesic which passes through z and makes an angle of α with η. Thus f is analogous to what we term rotation in the Euclidean context; since f preserves lengths, we can determine its action on any point in H^2 based solely on knowledge of the angle of rotation α. As our choice of notation suggests, this angle turns out to be equal to the argument of the eigenvalue λ.

As an example of a map of this form, consider

$$f\colon z \mapsto \frac{(\cos\alpha)z + \sin\alpha}{(-\sin\alpha)z + \cos\alpha}$$

which is rotation by α around the point i; the geodesics passing through i are the dark curves in Figure 4.13. The lighter curves are the circles whose (hyperbolic) centre lies at i; each of these curves intersects all of the geodesics orthogonally, and is left invariant by f.

This map does not seem terribly symmetric when viewed as a transformation of the upper half-plane; however, if we look at f in the unit disc model, we see that i is taken to the origin, and f corresponds to the rotation by α around the origin in the usual sense. Thus we associate with a rotation (as well as with the family of all rotations around a given point p) two families of curves:

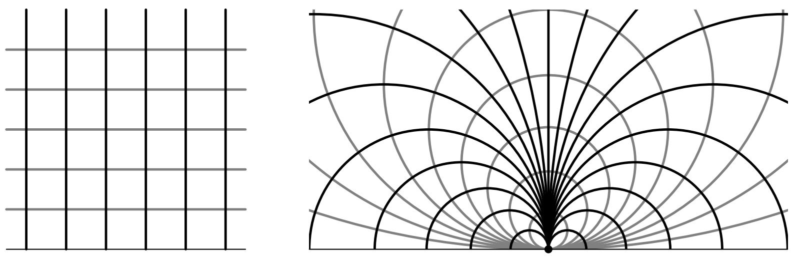

Figure 4.14. Parallel geodesics and horocycles for parabolic isometries.

(1) The *pencil* of all geodesics passing through p; each element of this family maps to another, and rotations around p act transitively on this family.

(2) The family of circles around p which are orthogonal to the geodesics from the first family. Each circle is invariant under rotations, and rotations around p act transitively on each circle.

We will discover similar pictures for the remaining two cases.

a.3. *Parabolic isometries.* Case **(P)** can be considered as a limiting case of the previous situation where the fixed point p goes to infinity. Let $t \in \mathbb{R} \cup \{\infty\}$ be the unique fixed point in the Riemann sphere, which lies on the ideal boundary. As with the family of rotations around p, we can consider the family of all orientation preserving isometries preserving t; notice that as in that case, this family is a *one-parameter group* whose members we will denote by $p_s^{(t)}$, where $s \in \mathbb{R}$. As above, one can see two invariant families of curves:

(1) The pencil of all geodesics passing through t (dark curves in Figure 4.14)—each element of this family maps to another, and the group $\{p_s^{(t)}\}$ acts transitively on this family.

(2) The family of *limit circles*, more commonly called *horocycles* (light curves in Figure 4.14), which are orthogonal to the geodesics from the first family. They are represented by circles tangent to $\mathbb{R}$ at t, or by horizontal lines if $t = \infty$.

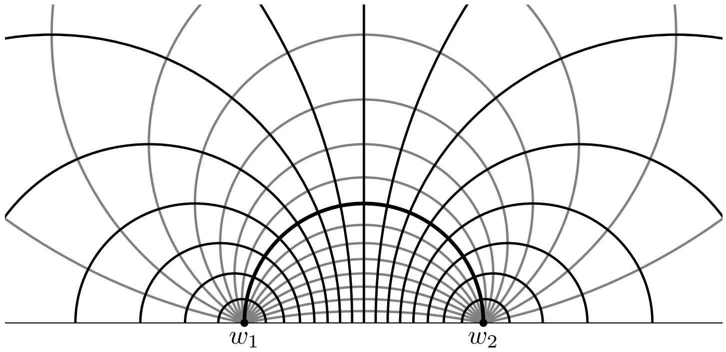

Figure 4.15. Orthogonal geodesics and equidistant curves for the geodesic connecting w_1 and w_2.

Each horocycle is invariant under $p_s^{(t)}$, and the group acts transitively on each horocycle.

A useful (but visually somewhat misleading) example is given by the case $t = \infty$ with

$$p_s^{(\infty)} z = z + s.$$

We will see later in the lecture that for the parabolic case, the 'angle' s does not have properties similar to the rotation angle α. In particular, it is not an invariant of the isometry.

Exercise 4.16. Show that given two points $z_1, z_2 \in H^2$, there are exactly two different horocycles which pass through z_1 and z_2.

a.4. *Hyperbolic isometries.* Finally, consider the case **(H)**, in which we have two real fixed points $w_1 < w_2$. Since f takes geodesics to geodesics and fixes w_1 and w_2, the half-circle γ which intersects $\mathbb{R}$ at w_1 and w_2 is mapped to itself by f, and so f acts as translation along this curve by a fixed distance. The geodesic γ is the only geodesic invariant under the transformation; in a sense, it plays the same role as the centre of rotation in the elliptic case, a role for which there is no counterpart in the parabolic case.

To see what the action of f is on the rest of H^2, consider as above two invariant families of curves:

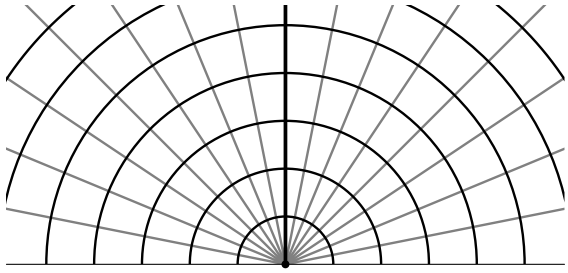

Figure 4.16. Orthogonal geodesics and equidistant curves for the imaginary axis.

(1) The family of geodesics which intersect γ orthogonally (the dark curves in Figure 4.15). If η is a member of this family, then f will carry η to another member of the family; which member is determined by the effect of f on the point where η intersects γ.

(2) The family of curves orthogonal to these geodesics (the light curves in Figure 4.15)—these are the *equidistant curves* (or *hypercircles*). Such a curve ζ is defined as the locus of points which lie a fixed distance from the geodesic γ; in Euclidean geometry this condition defines a geodesic, but this is no longer the case in the hyperbolic plane. Each equidistant curve ζ is carried into itself by the action of f.

A good example of maps f falling into the case **(H)** are the maps which fix 0 and ∞:

$$f\colon z \mapsto \lambda^2 z$$

In this case the geodesic γ connecting the fixed points is the imaginary axis (the vertical line in Figure 4.16), the geodesics intersecting γ orthogonally are the (Euclidean) circles centred at the origin (the dark curves), and the equidistant curves are the (Euclidean) lines emanating from the origin (the lighter curves).

To be precise, given any geodesic γ in the hyperbolic plane, we define an *r-equidistant curve* as one of the two connected components of the locus of points at a distance r from γ.

Exercise 4.17. For any given $r > 0$, show that there are exactly two different r-equidistant curves (for some geodesics) which pass through two given points in the hyperbolic plane.

Thus we have answered the question about the significance of (Euclidean) circles tangent to the real lines and arcs which intersect it. The former (along with horizontal lines) are horocycles, and the latter (along with rays intersecting the real line) are equidistant curves. Notice that all horocycles are isometric to each other (they can be viewed as circles of infinite radius), whereas for equidistant curves there is an isometry invariant, namely the angle between the curve and the real line. One can see that this angle uniquely determines the distance r between an equidistant curve and its geodesic, and vice versa. The correspondence between the two can be easily calculated in the particular case shown in Figure 4.16.

Exercise 4.18. The arc of the circle $|z - 2i|^2 = 8$ in the upper half-plane represents an r-equidistant curve. Find r.

a.5. *Canonical form for elliptic, parabolic, and hyperbolic isometries.* The technique of understanding an isometry by showing that it is conjugate to a particular standard transformation has great utility in our classification of isometries of H^2. Recall that we have a one-to-one correspondence between 2×2 real matrices with unit determinant (up to a choice of sign) and fractional linear transformations preserving $\mathbb{R}$, which are the isometries of H^2 that preserves orientation.

$$PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\pm \mathrm{Id} \longleftrightarrow \mathrm{Isom}^+(H^2)$$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \longleftrightarrow f_A \colon z \mapsto \frac{az+b}{cz+d}$$

Composition of isometries corresponds to matrix multiplication:

$$f_A \circ f_B = f_{AB}$$

We may easily verify that two maps f_A and f_B corresponding to conjugate matrices are themselves conjugate; that is, if $A = CBC^{-1}$ for some $C \in GL(2,\mathbb{R})$, we may assume without loss of generality that $C \in SL(2,\mathbb{R})$ by scaling C by its determinant. Then we have

$$f_A = f_C \circ f_B \circ f_C^{-1}$$

It follows that f_A and f_B have the same geometric properties; fixed points, actions on geodesics, etc. Conjugation by f_C has the effect of changing coordinates by an isometry, and so the intrinsic geometric properties of an isometry are conjugacy invariants. For example, in the Euclidean plane, any two rotations by an angle α around different fixed points x and y are conjugated by the translation taking x to y, and any two translations by vectors of equal length are conjugated by any rotation by the angle between those vectors. Thus, in the Euclidean plane, the conjugacy invariants are the angle of rotation and the length of the translation.

In order to classify orientation preserving isometries of H^2, it suffices to understand certain canonical examples. We begin by recalling the following result from linear algebra:

Proposition 4.16. *Every matrix in $SL(2,\mathbb{R})$ is conjugate to one of the following (up to sign):*

(E): *An* elliptic *matrix of the form*

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

for some $\alpha \in \mathbb{R}$.

(P): *The* parabolic *matrix*

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

(H): *The* hyperbolic *matrix*

$$\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$$

for some $t \in (0,\infty)$.

The three cases **(E)**, **(P)**, and **(H)** for the matrix A correspond to the three cases discussed above for the fractional linear transformation f_A. Recall that the isometries corresponding to the elliptic case **(E)** have one fixed point in H^2, those corresponding to the parabolic case **(P)** have one fixed point on the *ideal boundary* $\mathbb{R} \cup \{\infty\}$, and those corresponding to the hyperbolic case **(H)** have two fixed points on the ideal boundary.

The only invariants under conjugation are the parameters α (up to a sign) and t, which correspond to the angle of rotation and the distance of translation, respectively. Thus two orientation preserving isometries of H^2 are conjugate *in the full isometry group of* H^2 iff they fall into the same category **(E)**, **(P)**, or **(H)** and have the same value of the invariant α or t, if applicable.

Notice that if we consider only conjugacy by orientation preserving isometries, then α itself (rather than its absolute value) is an invariant in the elliptic case, and the two parabolic matrices $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ are not conjugate. In contrast, the conjugacy classes in the hyperbolic case do not change.

Thus we see that there are both similarities and differences between the structure of the group of orientation preserving isometries in the Euclidean and hyperbolic planes. Among the similarities is the possible number of fixed points: one or none. Isometries with one point—rotations—look completely similar, but the set of isometries with no fixed points—which in the Euclidean case is just translations—is more complicated in the hyperbolic case, including both parabolic and hyperbolic isometries.

An important difference in the structure of the isometry groups comes from the following observation. Recall that a subgroup H of a group G is *normal* if for any $h \in H$ and $g \in G$ the conjugate $g^{-1}hg$ remains in H. It is not hard to show that in the group of isometries of the Euclidean plane, translations form a normal subgroup; the situation in the hyperbolic case is rather different.

Exercise 4.19. Prove that the group of isometries of the hyperbolic plane has no non-trivial normal subgroups, i.e. the only normal subgroups are the whole group and the trivial subgroup containing only the identity.

Another example of a difference between the two cases comes when we consider the decomposition of orientation preserving isometries into reflections—this is possible in both the Euclidean and the hyperbolic planes, and any orientation preserving isometry can be had as a product of two reflections. In the Euclidean plane, there are two possibilities—either the lines of reflection intersect, and the product is a rotation, or the lines are parallel, and the product is a translation. In the hyperbolic plane, there are three possibilities for the relationship of the lines (geodesics) of reflection: once again, they may intersect or be parallel (i.e. have a common point at infinity), but now a new option arises; they may also be ultraparallel (see Figure 4.17). We will discuss this in more detail shortly.

Exercise 4.20. Prove that the product of reflections in two geodesics in the hyperbolic plane is elliptic, parabolic or hyperbolic, respectively, depending on whether the two axes of reflection intersect, are parallel, or are ultraparallel.

a.6. *Orientation reversing isometries.* Using representation (4.9) and following the same strategy, we try to look for fixed points of orientation reversing isometries. The fixed point equation takes the form

$$c|z|^2 + dz - a\bar{z} - b = 0.$$

Separating real and imaginary parts, we get two cases:

(1) $d + a = 0$. In this case, there is a whole geodesic of fixed points, and the transformation is a reflection in this geodesic, which geometrically is represented as inversion (if the geodesic is a half-circle) or the usual sort of reflection (if the geodesic is a vertical ray).

(2) $d+a \neq 0$. In this case, there are two fixed points on the (extended) real line, and the geodesic connecting these points is preserved, so the transformation is a glide reflection, and can be written as the composition of reflection in this geodesic and a hyperbolic isometry with this geodesic as its axis.

Thus the picture for orientation reversing isometries is somewhat more similar to the Euclidean case.

b. Geometric interpretation of isometries. From the synthetic point of view, the fundamental difference between Euclidean and hyperbolic geometry is the failure of the parallel postulate in the latter case. To be more precise, suppose we have a geodesic (line) γ and a point p not lying on γ, and consider the set of all geodesics (lines)

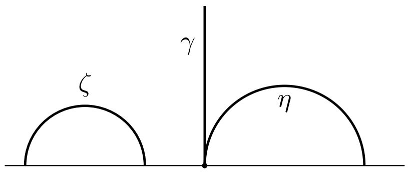

Figure 4.17. Parallels and ultraparallels.

through p which do not intersect γ. In the Euclidean case, there is exactly one such geodesic, and we say that it is parallel to γ. In the hyperbolic case, not only are there many such geodesics, but they come in two different classes, as shown in Figure 4.17.

The curves γ, η, and ζ in Figure 4.17 are all geodesics, and neither η nor ζ intersects γ in H^2. However, η and γ both approach the same point on the ideal boundary, while ζ and γ do not exhibit any such asymptotic behaviour. We say that η and γ are *parallel*, while ζ and γ are *ultraparallel*.

Each point x on the ideal boundary corresponds to a family of parallel geodesics which are asymptotic to x, as shown in Figure 4.14. The parallel geodesics asymptotic to ∞ are simply the vertical lines, while the parallel geodesics asymptotic to some point $x \in \mathbb{R}$ form a sort of bouquet of curves.

A recurrent theme in our description of isometries has been the construction of orthogonal families of curves. Given the family of parallel geodesics asymptotic to x, one may consider the family of curves which are orthogonal to these geodesics at every point; such curves are called *horocycles*. As shown in Figure 4.14, the horocycles for the family of geodesics asymptotic to ∞ are horizontal lines, while the horocycles for the family of geodesics asymptotic to $x \in \mathbb{R}$ are Euclidean circles tangent to $\mathbb{R}$ at x.

The reason horocycles are sometimes called limit circles is illustrated by the following construction: fix a point $p \in H^2$ and a geodesic ray γ which starts at p. For each $r > 0$ consider the circle of radius r with centre on γ which passes through p; as $r \to \infty$, these circles converge to the horocycle orthogonal to γ.

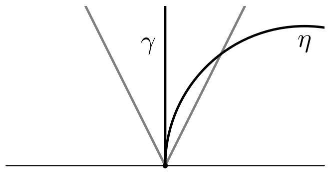

Figure 4.18. Distance between parallel geodesics.

What do we mean by this last statement? In what sense do the circles 'converge' to the horocycle? For any fixed value of r, the circle in the construction lies arbitrarily far from some points on the horocycle (those which are 'near' the ideal boundary), and so we certainly cannot expect any sort of uniform convergence in the hyperbolic metric. Rather, convergence in the hyperbolic plane must be understood as convergence of pieces of fixed, albeit arbitrary large, length—that is, given $R > 0$, the arcs of length R lying on the circles in the above construction with p at their midpoint do in fact converge uniformly to a piece of the horocycle, and R may be taken as large as we wish.

The situation is slightly different in the model, where we do have genuine uniform convergence, as the complete (Euclidean) circles representing (hyperbolic) circles converge to the (Euclidean) circle representing the horocycle.

This distinction between the intrinsic and extrinsic viewpoints raises other questions; for example, the above distinction between parallel and ultraparallel geodesics relies on this particular model of H^2 and the fact that points at infinity are represented by real numbers, and so seems rooted in the extrinsic description of H^2. Can we distinguish between the two sorts of asymptotic behaviour intrinsically, without reference to the ideal boundary?

It turns out that we can; given two ultraparallel geodesics γ and η, the distance from γ to η grows without bound; that is, given any $C \in \mathbb{R}$, there exists a point $z \in \gamma$ such that no point of η is within a distance C of z. On the other hand, given two parallel geodesics, this distance remains bounded, and in fact goes to zero.

To see this, let γ be the imaginary axis; then the equidistant curves are Euclidean lines through the origin, as shown in Figure 4.18, and η is a Euclidean circle which is tangent to γ at the origin. The distance from γ to the equidistant curves is a function of the slope of the lines; steeper slope corresponds to smaller distance, and the points in between the curves are just the points which lie within that distance of γ. But now for any slope of the lines, η will eventually lie between the two equidistant curves, since its slope becomes vertical as it approaches the ideal boundary, and hence the distance between γ and η goes to zero.

One can see the same result by considering a geodesic η which is parallel to γ not at 0, but at ∞; then η is simply a vertical Euclidean line, which obviously lies between the equidistant curves for large enough values of y.

To get an idea of how quickly the distance goes to 0 in Figure 4.18, recall that the hyperbolic distance between two nearby points is roughly the Euclidean distance divided by the height y, and that the Euclidean distance between a point on the circle η in Figure 4.18 and the imaginary axis is roughly y^2 for points near the origin; hence

$$\text{hyperbolic distance} \sim \frac{\text{Euclidean distance}}{y} \sim \frac{y^2}{y} = y \to 0$$

With this understanding of circles, parallels, ultraparallels, and horocycles, we can now return to the task of giving geometric meaning to the various categories of isometries. In each case, we found two families of curves which intersect each other orthogonally; one of these will comprise geodesics which are carried to each other by the isometry, and the other family will comprise curves which are invariant under the isometry.

In the elliptic case **(E)**, the isometry f is to be thought of as rotation around the unique fixed point p by some angle α; the two families of curves are shown in Figure 4.13. Given $v \in T_pH^2$, denote by γ_v the unique geodesic passing through p with $\gamma'(p) = v$. Then we have

$$f\colon \{\gamma_v\}_{v\in T_pH^2} \to \{\gamma_v\}_{v\in T_pH^2}$$
$$\gamma_v \mapsto \gamma_w$$

where $w \in T_pH^2$ is the image of v under rotation by α in the tangent space. Taking the family of curves orthogonal to the curves γ_v at each point of H^2, we have the one-parameter family of circles

$$\{\eta_r\}_{r\in(0,\infty)}$$

each of which is left invariant by f.

In the parabolic case **(P)**, the map f is just horizontal translation $z \mapsto z+1$. Note that by conjugating this map with a homothety, and a reflection if necessary, we obtain horizontal translation by any distance, so any horizontal translation is conjugate to the canonical example. Given $t \in \mathbb{R}$, let γ_t be the vertical line $\operatorname{Re} z = t$, then the geodesics γ_t are all asymptotic to the fixed point ∞ of f, and we have

$$f\colon \{\gamma_t\}_{t\in\mathbb{R}} \to \{\gamma_t\}_{t\in\mathbb{R}}$$
$$\gamma_t \mapsto \gamma_{t+1}$$

The invariant curves for f are the horocycles, which in this case are horizontal lines η_t, $t \in \mathbb{R}$. For a general parabolic map, the fixed point x may lie on $\mathbb{R}$ rather than at ∞; in this case, the geodesics and horocycles asymptotic to x are as shown in the second image in Figure 4.14. The invariant family of geodesics consists of geodesics parallel to each other.

Finally, in the hyperbolic case **(H)**, the standard form is $f_A(z) = \lambda^2 z$ for $\lambda = e^t$, and the map is simply a homothety from the origin. There is exactly one invariant geodesic, the imaginary axis, and the other invariant curves are the equidistant curves, which in this case are Euclidean lines through the origin. The curves orthogonal to these at each point are the geodesics γ_r ultraparallel to each other, shown in Figure 4.16, where γ_r is the unique geodesic passing through the point ir and intersecting the imaginary axis orthogonally. The map f_A acts on this family by taking γ_r to $\gamma_{\lambda^2 r}$.

In the general hyperbolic case, the two fixed points will lie on the real axis, and the situation is as shown in Figure 4.15. The invariant geodesic η_0 is the half-circle connecting the fixed points, and the equidistant curves are the other circles passing through those two points. The family of orthogonal curves are the geodesics intersecting η_0 orthogonally, as shown in the picture.

Lecture 30.

a. Area of triangles in different geometries. In our earlier investigations of spherical and elliptic geometry (by the latter we mean the geometry of the projective plane with metric inherited from the sphere), we found that the area of a triangle was proportional to its *angular excess*, the amount by which the sum of its angles exceeds π. For a sphere of radius R, the constant of proportionality was $R^2 = 1/\kappa$, where κ is the curvature of the surface.

In Euclidean geometry, the existence of any such formula was precluded by the presence of similarity transformations, diffeomorphisms of $\mathbb{R}^2$ which expand or shrink the metric by a uniform constant.

In the hyperbolic plane, we find ourselves in a situation reminiscent of the spherical case. We will find that the area of a hyperbolic triangle is proportional to the angular *defect*, the amount by which the sum of its angles falls short of π, and that the constant of proportionality is again given by the reciprocal of the curvature.

We begin with a simple observation, which is that every hyperbolic triangle does in fact have angles whose sum is less than π (otherwise the above claim would imply that some triangles have area ≤ 0).

For that we use the open disc model of the hyperbolic plane, and note that given any triangle, we can use an isometry to position one of its vertices at the origin; thus two of the sides of the triangle will be (Euclidean) lines through the origin, as shown in Figure 4.19. Then because the third side, which is part of a Euclidean circle, is convex in the Euclidean sense, the sum of the angles is less than π.

This implies the remarkable 'fourth criterion of equality of triangles' above and beyond the three criteria which are common to both the Euclidean and hyperbolic planes.

Proposition 4.17. *Two geodesic triangles with pairwise equal angles are isometric.*

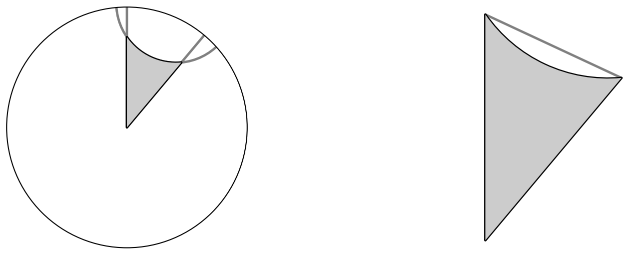

Figure 4.19. A hyperbolic triangle has angles whose sum is less than π.

Proof. We will use the disc model. Without loss of generality, we may assume that both triangles have one vertex at the centre O and that two of their sides lie on the same radii. Thus we have triangles AOB and $A'OB'$ where the vertices A and A' lie on one radius, and B and B' on another. The angles OAB and $OA'B'$ are equal and so are angles OBA and $OB'A'$.

Now there are two possibilities; either the arcs AB and $A'B'$ intersect, or they do not. Assume first that they intersect at some point C. Then the triangle ACA' has two angles which add to π, which is impossible. Hence without loss of generality, we may assume that arc AA' lies inside the triangle OBB'. Then the sum of the angles of the geodesic quadrangle $AA'BB'$ is equal to 2π, which is again an impossibility since it can be split into two geodesic triangles, at least one of which must therefore have angles whose sum is $\geq \pi$. This contradiction implies $A = A'$ and $B = B'$. □

b. Area and angular defect in hyperbolic geometry. Our proof of the area formula is due to Gauss, and follows the exposition in Coxeter's book "Introduction to Geometry" (sections 16.4 and 16.5). It is essentially a synthetic proof, and as such does not give us a value for the constant of proportionality; to obtain that value, we must turn to analytic methods. We will also deviate slightly from the true synthetic approach by using drawings in the two models of H^2.

As with so many things, non-Euclidean geometry was first discovered and investigated by Gauss, who kept his results secret because he had no proof that his geometry was consistent. Eventually, the

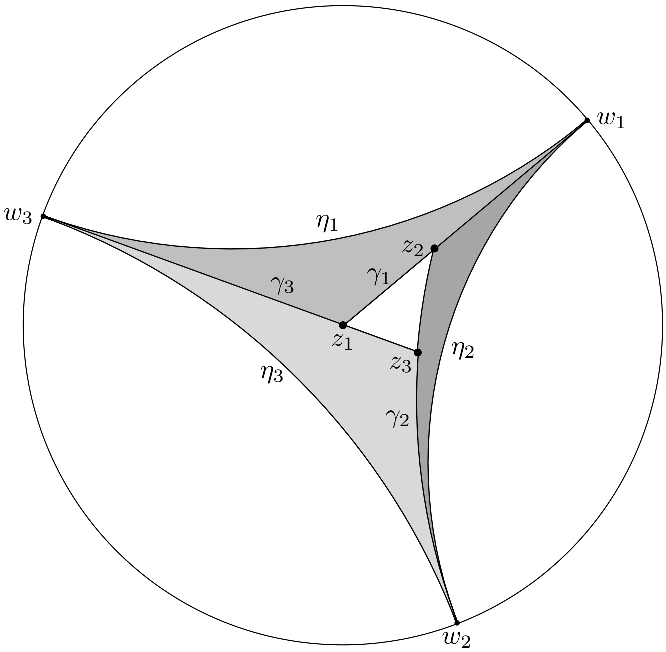

Figure 4.20. Computing the area of a hyperbolic triangle.

introduction of several models (of which the Poincaré half-plane and open disc models were not the earliest) showed that hyperbolic geometry is consistent, contingent upon the consistency of Euclidean geometry; a contradiction in the former would necessarily lead to a contradiction in the latter.

Theorem 4.10. *Given a hyperbolic triangle Δ with angles α, β, and γ, the area A of Δ is given by*

$$A = \frac{1}{-\kappa}(\pi - \alpha - \beta - \gamma) \tag{4.12}$$

where κ is the curvature, whose value is -1 for the standard upper half-plane and open disc models.

Proof. The proof of the analogous formula for the sphere involved partitioning it into segments and using an inclusion-exclusion formula. This relied on the fact that the area of the sphere is finite; in our

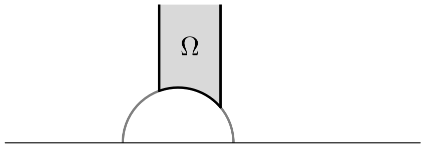

Figure 4.21. A singly asymptotic triangle.

present case, we must be more careful, as the hyperbolic plane has infinite area. However, we can recover a setting in which a similar proof works by considering *asymptotic triangles*, which turn out to have finite area.

The idea is as follows: let z_1, z_2, z_3 denote the vertices of the triangle, and without loss of generality, take z_1 to be the origin in the open disc model. As shown in Figure 4.20, draw the half-geodesic γ_1 which begins at z_1 and passes through z_2; similarly, draw the half-geodesics γ_2 and γ_3 beginning at z_2 and z_3, and passing through z_3 and z_1, respectively. Let w_j denote the point at infinity approached by γ_j as it nears the boundary of the disc.

Now draw three more geodesics, as shown in the picture; η_1 is to be asymptotic to w_3 and w_1, η_2 is to be asymptotic to w_1 and w_2, and η_3 is to be asymptotic to w_2 and w_3. Then the region T_0 bounded by η_1, η_2, and η_3 is a *triply asymptotic triangle.* If we write T_j for the *doubly asymptotic triangle* whose vertices are z_j, w_j, and w_{j-1}, we can decompose T_0 as the disjoint union

$$T_0 = T_1 \cup T_2 \cup T_3 \cup \Delta$$

and so the area $A(\Delta)$ may be found by computing the areas of the regions T_j, provided they are finite.

Since these regions are not bounded, it is not at first obvious why they should have finite area. We begin by making two observations concerning triply asymptotic triangles.

First, all triply asymptotic triangles are isometric. That is, given $w_1, w_2, w_3 \in \partial D^2$ and $\tilde{w}_1, \tilde{w}_2, \tilde{w}_3 \in \partial D^2$ with the same orientation, Lemma 4.9 guarantees the existence of a unique fractional linear transformation f taking w_j to $\tilde{w}_j$, which must then preserve ∂D^2 and map the interior to the interior, and hence is an isometry of H^2.

Secondly, a triply asymptotic triangle will have finite area iff each of its 'arms' does, where by an 'arm' we mean the section of the triangle which approaches infinity. How do we compute the area of such an arm? A prototypical example is the singly asymptotic triangle shown in Figure 4.21, where we use the half-plane model and choose ∞ as the point on the ideal boundary, so two of the geodesics are vertical lines. The infinitesimal area element at each point is given by $\frac{1}{y^2}\,dx\,dy$ where dx and dy are Euclidean displacements, and so the area of the shaded region Ω is

$$A(\Omega) = \int_\Omega \frac{1}{y^2}\,dx\,dy$$

which converges as $y \to \infty$, and hence Ω has finite area. It follows that the area of a triply asymptotic triangle is finite, and independent of our choice of triangle; denote this area by μ. Note that any hyperbolic triangle is contained in a triply asymptotic triangle, and so every hyperbolic triangle must have area less than μ.

In order to complete our calculations for A, we must find a formula for the areas of the doubly asymptotic triangles T_1, T_2, and T_3 (the shaded triangles in Figure 4.20). Note first that by using an isometry to place the non-infinite vertex of a doubly asymptotic triangle at the origin, we see that the area depends only on the angle at the vertex. Given an angle θ, let $f(\theta)$ denote the area of the doubly asymptotic triangle with angle $\pi - \theta$, so that if θ_j is the angle in the triangle at the vertex z_j, then $A(T_j) = f(\theta_j)$.

We may obtain a triply asymptotic triangle as the disjoint union of two doubly asymptotic triangles with angles $\pi - \alpha$, $\pi - \beta$ where $\alpha + \beta = \pi$, and hence

$$f(\alpha) + f(\beta) = \mu$$

Similarly, we may obtain a triply asymptotic triangle as the disjoint union of three doubly asymptotic triangles with angles $\pi - \alpha$, $\pi - \beta$, and $\pi - \gamma$ where $(\pi-\alpha)+(\pi-\beta)+(\pi-\gamma) = 2\pi$ and hence $\alpha+\beta+\gamma = \pi$, so we have

$$f(\alpha) + f(\beta) + f(\gamma) = \mu$$

for such α, β, γ. We may rewrite the above two equations as

$$f(\alpha + \beta) + f(\pi - \alpha - \beta) = \mu$$
$$f(\alpha) + f(\beta) + f(\pi - \alpha - \beta) = \mu$$

and comparing the two gives

$$f(\alpha + \beta) = f(\alpha) + f(\beta)$$

so that f is in fact a linear function. Further, the limit $\alpha \to \pi$ corresponds to a doubly asymptotic triangle whose nonzero angle shrinks and goes to zero, and so the triangle becomes triply asymptotic; hence $f(\pi) = \mu$, and we have

$$f(\theta) = \frac{\mu}{\pi}\theta.$$

It follows that

$$\begin{aligned} A(\Delta) = T_0 - T_1 - T_2 - T_3 &= \mu - \frac{\mu}{\pi}(\theta_1 + \theta_2 + \theta_3) \\ &= \frac{\mu}{\pi}(\pi - \theta_1 - \theta_2 - \theta_3) \end{aligned}$$

and hence our formula is proved, with constant of proportionality $\frac{1}{-\kappa} = \frac{\mu}{\pi}$.

In order to calculate the coefficient of proportionality for the standard half-plane model consider the triply asymptotic triangle T in the upper half-plane bounded by the unit circle $|z| = 1$ and the vertical lines $\operatorname{Re} z = 1$ and $\operatorname{Re} z = -1$. The area of T is given by

$$\begin{aligned} \mu = \int_T \frac{1}{y^2}\,dx\,dy &= \int_{-1}^{1} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\,dy\,dx \\ &= \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\,dx = \int_{-\pi/2}^{\pi/2} d\theta = \pi \end{aligned}$$

using the substitution $x = \sin\theta$. This confirms the choice $\kappa = -1$ for the usual model. □

Note that the formula is valid not only for finite triangles, but also for asymptotic triangles, since taking a vertex to infinity is equivalent to taking the corresponding angle to zero.

The above proof that the area μ of a triply asymptotic triangle is finite relied on analytic methods, rather than purely synthetic ones. We sketch the purely synthetic proof given in Coxeter's book, which

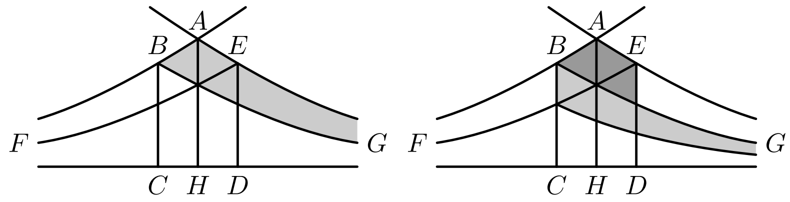

Figure 4.22. Decomposing an asymptotic triangle.

relies only on the fact that area is additive and that reflections are isometries. As before, it suffices to prove that the area of a singly asymptotic triangle is finite.

Consider such a triangle, given by the shaded region in Figure 4.22. Here we begin with the asymptotic triangle ABG and extend the geodesic AB to the point F at infinity. Then we draw the geodesic asymptotic to F and G and add the perpendicular AH, which bisects the angle at A. Note that all the curves in this picture represent geodesics—as this is a purely synthetic picture, it does not refer to either of the models, and in particular, does not include the ideal boundary. Reflecting BG in the line AH gives the geodesic EF; the geodesics BC, ED bisect the appropriate angles and meet the geodesic FG orthogonally.

The bulk of the proof is in the assertion that by repeated reflections first in ED and then in AH, the rest of the shaded region can be brought into the pentagon $ABCDE$. The first step is shown in Figure 4.22, and the details of the proof are left to the reader. Once it is established that ABG can be decomposed into triangles whose isometric images fill $ABCDE$ disjointly, it follows immediately that the area of ABG is finite, and the proof is complete.

Exercise 4.21. Find all the isometries which preserve a triply asymptotic triangle.

Exercise 4.22. Consider a line in the hyperbolic plane and a doubly asymptotic triangle for which this line is one of the sides. Assume the angle at the finite vertex is fixed, and find the locus of all finite vertices.