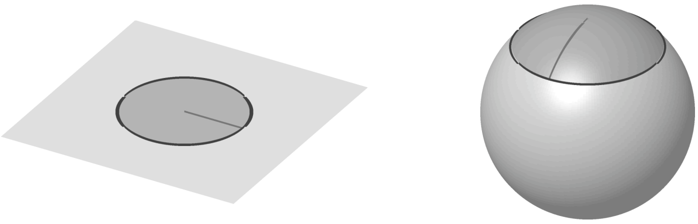

Figure 4.5. Relating curvature to the circumference of a circle.

the plane with radius r (Figure 4.5). We will see that

$$\text{circumference} = 2\pi r - cr^3 + o(r^3)$$

where c is a constant related to the curvature. Upon integration, we will obtain an expression for the area of the disc as

$$\text{area} = \pi r^2 - \frac{c}{4}r^4 + o(r^4).$$

b. The hyperbolic plane: two conformal models.

b.1. *The upper half-plane model.* In order to exhibit a surface with constant *negative* curvature, we pull a proverbial rabbit from our sleeve, or hat, or some other piece of proverbial clothing, and give without motivation the definition of the upper half-plane model of hyperbolic geometry due to Henri Poincaré, arguably the greatest mathematician since Gauss and Riemann. Our surface will be H^2, defined as

$$H^2 = \{\, (x, y) \in \mathbb{R}^2 \mid y > 0 \,\} = \{\, z \in \mathbb{C} \mid \operatorname{Im} z > 0 \,\},$$

where it is useful to keep in mind the formulation in terms of complex numbers in order to describe the isometry group of H^2.

The metric on H^2 is given by a conformal change of the standard metric:

$$ds^2 = \frac{dx^2 + dy^2}{y^2}. \tag{4.3}$$

The fact that the denominator vanishes when $y = 0$ gives some justification for the fact that we consider only the upper half-plane, and not the entire plane. From (4.3) it is apparent that Euclidean lengths are increased when y is small, and decreased when y is large; Figure 4.6

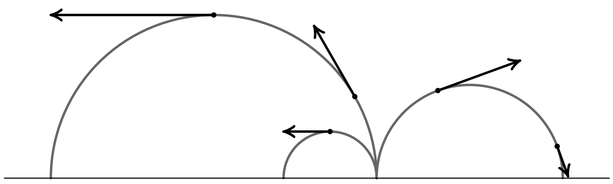

Figure 4.6. Unit tangent vectors in the hyperbolic plane.

shows some unit tangent vectors. All of these have unit length in the hyperbolic metric, and so their Euclidean lengths vary as y varies.

In order to show that H^2 has constant curvature, we will show that isometries act transitively. To see this, it will suffice to exhibit two particular classes of isometries.

(1) *Translations.* Given a real number t, the translation by t which takes z to $z + t$ (or in real coordinates, (x, y) to $(x + t, y)$) is an isometry since the metric does not depend on the horizontal coordinate x.

(2) *Homotheties.* For any $\lambda > 0$, the map $z \mapsto \lambda z$ turns out to be an isometry; this is most easily seen by writing the metric as

$$ds = \frac{(dx^2 + dy^2)^{\frac{1}{2}}}{y}$$

from which it is clear that multiplying both x and y by λ does not change ds.

Since any composition of these two types of isometries is itself an isometry, the isometry group acts transitively on H^2; given $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, we can first scale z_1 by y_2/y_1 so that the imaginary parts are the same, and then translate by the difference in the real parts. It follows that H^2 has constant curvature.

Acting transitively on the surface itself is not the whole story, however; in the case of the sphere and the Euclidean plane, the isometry group acts transitively not only on the surface, but also on the unit tangent bundle.

By way of explaining this last statement, recall the general fact that given any smooth map $f\colon S \to S$, the Jacobian Df_p at a point p

defines a linear transformation between the tangent spaces T_pS and $T_{f(p)}S$, so that the pair (f, Df) acts on the tangent bundle as

$$\begin{aligned}(f, Df)\colon \quad & TS \to TS, \\ & (p, v) \mapsto (f(p), Df_p v).\end{aligned}$$

Now f is an isometry iff Df acts isometrically on each tangent space; in particular, it must preserve the norm. Thus we restrict our attention to tangent vectors of norm one, which form the *unit tangent bundle*; for each isometry f acting on S, the pair (f, Df) acts isometrically on the unit tangent bundle of S.

For both S^2 and $\mathbb{R}^2$, this action is transitive; given any two points $p, q \in S$ and unit tangent vectors $v \in T_pS$, $w \in T_qS$, there exists an isometry $f\colon S \to S$ such that

$$\begin{aligned} f(p) &= q, \\ Df_p(v) &= w. \end{aligned}$$

To see that a similar property holds for H^2, we must consider all the isometries and not just those generated by the two classes mentioned so far. For example, we have not yet considered the orientation reversing isometry $(x, y) \mapsto (-x, y)$.

We will prove later (Proposition 4.14) that every orientation preserving isometry of H^2 has the form

$$f\colon z \mapsto \frac{az + b}{cz + d},$$

where $a, b, c, d \in \mathbb{R}$. This condition guarantees that f fixes the real line, which must hold for any isometry of H^2. We also require that $ad - bc \neq 0$, since otherwise the image of f is a single point; in fact, we must have $ad - bc > 0$; otherwise f swaps the upper and lower half-planes.

As given, f appears to depend on four real parameters, while considerations similar to those in the analysis of the isometry groups of S^2 and $\mathbb{R}^2$ suggest that three parameters ought to be sufficient. Indeed, scaling all four coefficients by a factor $\lambda > 0$ leaves the transformation f unchanged, but scales the quantity $ad - bc$ by λ^2; hence we may require in addition that $ad - bc = 1$, and now we see that f belongs to a three-parameter group.

The condition $ad-bc=1$ is obviously reminiscent of the condition $\det A = 1$ for a 2×2 matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In fact, if given such a matrix A we denote the transformation given above by f_A, then a little algebra verifies that

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

and so the isometry group of H^2 is isomorphic to $SL(2,\mathbb{R})$, the group of 2×2 real matrices with unit determinant, modulo the provision that $f_I = f_{-I} = \mathrm{Id}$, and so we must take the quotient of $SL(2,\mathbb{R})$ by its centre $\{\pm I\}$. This quotient is denoted $PSL(2,\mathbb{R})$, and hence we will have

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

once we show that f_A is an isometry for every $A \in SL(2,\mathbb{R})$, and that every isometry is of this form. One way to prove the first statement (the second will be Proposition 4.14) is to show that every such f_A can be decomposed as a product of isometries which have one of the following three forms:

$$\begin{aligned} z &\mapsto z+t, \\ z &\mapsto \lambda z, \\ z &\mapsto -\frac{1}{z}, \end{aligned}$$

where $t \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$ define one-parameter families of isometries. This is equivalent to showing that $SL(2,\mathbb{R})$ is generated by the matrices

$$\left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \;\middle|\; t \in \mathbb{R} \right\} \bigcup \left\{ \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \;\middle|\; \lambda \in \mathbb{R}^+ \right\} \bigcup \left\{ \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right\}.$$

We have seen already that the first two transformations preserve the metric (4.3). To see that $z \mapsto \tilde{z} = -1/z$ is an isometry, one must suffer through a small amount of algebra and use the fact that for $z = x+iy$ we have

$$\tilde{z} = -\frac{1}{z} = \frac{-1}{x+iy} = -\frac{x-iy}{x^2+y^2} = \frac{-x+iy}{x^2+y^2},$$

which allows us to compute

$$d\tilde{x} = \frac{(x^2-y^2)\,dx - 2xy\,dy}{(x^2+y^2)^2}$$

along with a similar formula for $d\tilde{y}$; together, these let us deduce that $d\tilde{s} = ds$, showing that the map is an isometry.

Later we will give other proofs that any fractional linear transformation with real coefficients and non-vanishing determinant is a hyperbolic isometry.

b.2. *The disc model.* Remember that at least one motivation for considering the hyperbolic plane was to provide an ideal model of a surface of negative curvature.[1] In attempting to define curvature via excess or defect in the length of a small circle or area of a small disc, and to calculate it explicitly for the hyperbolic plane, we will find that our life is made easier by the introduction of a different model, which is also due to Poincaré. This is given by an open unit disc, for which the boundary of the disc plays the same role as was played by the real line with respect to H^2 (the so-called *ideal boundary*). The metric is given by

$$ds^2 = \frac{4(dx^2 + dy^2)}{(1 - x^2 - y^2)^2}, \tag{4.4}$$

and we may see that this model is the image of H^2 under a conformal transformation, for example

$$z \mapsto \frac{iz + 1}{z + i}. \tag{4.5}$$

An advantage of this model is that rotation around the origin is an isometry, and so hyperbolic circles around the origin are simply Euclidean circles in the plane with the same centre—of course, the hyperbolic radius is different from the Euclidean radius. This rotation is exactly the one type of isometry which does not have a convenient 'natural' representation in the upper half-plane model; thus it is useful to switch back and forth between the two models depending on the type of symmetry for which a particular problem calls.

b.3. *Embedded surfaces.* It is natural to ask whether one can realise the hyperbolic plane as a surface in $\mathbb{R}^3$. This turns out not to be possible for the whole plane (although the proof is not simple); however, there are surfaces in $\mathbb{R}^3$ whose intrinsic geometry is *locally* isometric

[1] There are of course plenty of other reasons—it is sufficient to recall that the geometry of the hyperbolic plane is the original non-Euclidean geometry where all the standard axioms except for the fifth postulate hold.

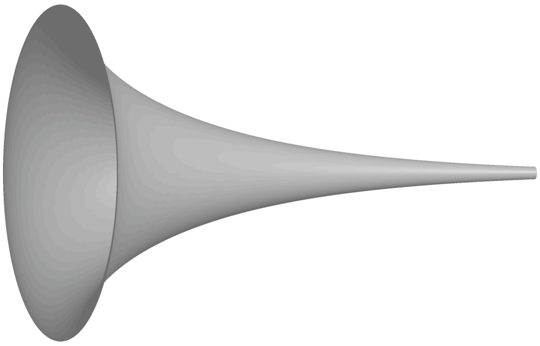

Figure 4.7. A pseudosphere.

to that of hyperbolic plane, in the same manner as the cylinder, for example, is locally Euclidean, despite not being globally isometric to $\mathbb{R}^2$.

The classic example of such a surface is the *pseudosphere* (Figure 4.7), the surface of revolution around the x-axis of the curve in the xz-plane called a *tractrix*, which is given parametrically by

$$(x,z)(t) = \left(t - \frac{\sinh t}{\cosh t}, \frac{1}{\cosh t} \right)$$

where $t \geq 0$. In order to see that the pseudosphere is locally isometric to the hyperbolic plane, one introduces coordinates on the pseudosphere in which the Riemannian metric induced from $\mathbb{R}^3$ has the same form as in the upper half-plane model of the hyperbolic plane.

c. Geodesics and distances on H^2. On an arbitrary surface with a Riemannian metric, the process of defining an explicit distance function and describing the geodesics can be quite tortuous. For the two spaces of constant curvature that we have already encountered, the solution turns out to be quite simple; on the Euclidean plane, geodesics are straight lines and the distance between two points is given by Pythagoras' formula, while on the sphere, geodesics are great circles and the distance between two points is proportional to the central angle they subtend.

One might expect, then, that the situation on H^2 exhibits a similar simplicity, and this will in fact turn out to be the case. Let us first consider two points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ with equal real parts $x_1 = x_2 = x$ and $y_2 > y_1$. Then it is fairly straightforward to see that the shortest path between z_1 and z_2 is a vertical line. For this curve we have

$$(4.6) \quad \ell(\gamma) = \int_{y_1}^{y_2} \left\| \frac{d}{dt}(x + it) \right\|_{x+it} dt = \int_{y_1}^{y_2} \frac{1}{t}\, dt = \log y_2 - \log y_1$$

and the length of any other curve will be greater than this value due to the contribution of the horizontal components of the tangent vectors—we will present this argument in more detail in the next lecture. It follows that vertical lines are geodesics in H^2.

Isometries preserve geodesics, and hence the image of a vertical line under any of the isometries discussed above is also a geodesic. Horizontal translation and scaling by a constant will map a vertical line to another vertical line, but the map $z \mapsto -1/z$ behaves differently. This map is the composition of reflection about the imaginary axis with the map $z \mapsto -1/\bar{z}$, and the latter is simply inversion in the unit circle. We encountered this map in Exercise 1.7 as the map

$$(x, y) \mapsto \left(\frac{x}{x^2 + y^2}, \frac{y}{x^2 + y^2} \right)$$

which arises as the transition map between stereographic projections from the north and south poles. It may be checked that this map takes lines to circles and circles to lines (with the exception of lines through the origin, which are mapped into themselves, and circles centred at the origin, which are taken into other circles centred at the origin); in particular, vertical lines are mapped to circles whose centres lie on the x-axis, and hence half-circles in H^2 with centres on the real axis are also geodesics.[2]

Because the three classes of isometries just mentioned generate the isometry group of H^2, which acts transitively on the tangent bundle, these are all the geodesics.

[2] In the next lecture we will prove that any fractional linear transformation $z \mapsto \frac{az+b}{cz+d}$, where a, b, c, d are arbitrary *complex* numbers such that $ad - bc \neq 0$, maps lines and circles into lines and circles.

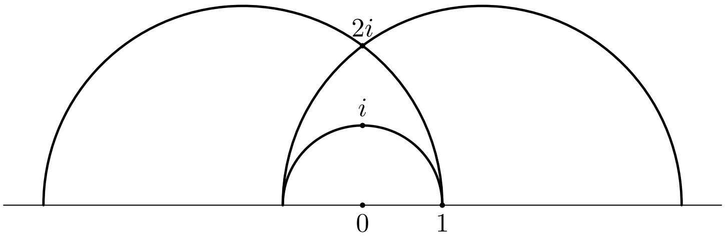

Figure 4.8. Failure of the parallel postulate in H^2.

With this characterisation of geodesics in hand, we can immediately see that Euclid's parallel postulate fails in the hyperbolic plane; given the upper half of the unit circle, which is a geodesic, and the point $2i$, which is a point not on that geodesic, there are many geodesics passing through $2i$ which do not intersect the upper half of the unit circle, as shown in Figure 4.8.

We now come to the question of giving a formula for the distance between two points $z_1, z_2 \in H^2$. Distance must be an isometric invariant, and must also be additive along geodesics. We may construct a geodesic connecting z_1 and z_2 by drawing the perpendicular bisector of the line segment between them and taking the intersection of this bisector with the real line. The circle centred at this point of intersection which passes through z_1 and z_2 will be the geodesic we seek.

As shown in Figure 4.9, let w_1 and w_2 be the points at which this circle intersects the real line; we will prove later (Lemma 4.7) that the *cross-ratio*

$$(z_1, z_2; w_1, w_2) = \frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2} \tag{4.7}$$

is preserved by all isometries of H^2. It turns out to be multiplicative along geodesics, not additive; if we place a third point z_3 between z_1 and z_2 along the circle as in Figure 4.9, we will have

$$\left|\frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2}\right| = \left|\frac{z_1 - w_1}{z_3 - w_1} \div \frac{z_1 - w_2}{z_3 - w_2}\right| \times \left|\frac{z_3 - w_1}{z_2 - w_1} \div \frac{z_3 - w_2}{z_2 - w_2}\right|.$$

Hence to obtain a true distance function which is additive along geodesics, we must take the logarithm of the cross-ratio. Notice from

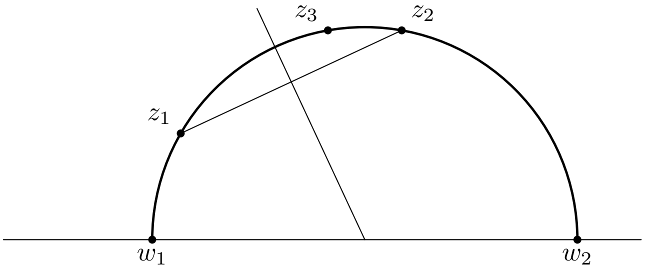

Figure 4.9. Using cross-ratio to define distance.

equation (4.6) that

$$d(iy_1, iy_2) = \log |(iy_1, iy_2; 0, \infty)|.$$

Since every pair of points can be mapped by an isometry to a pair of points on the imaginary axis, invariance of the cross-ratio implies that

$$d(z_1, z_2) = \log \left| \frac{z_1 - w_1}{z_2 - w_1} \right| - \log \left| \frac{z_1 - w_2}{z_2 - w_2} \right|. \tag{4.8}$$

Exercise 4.12. Prove the following formula for the hyperbolic distance between two points z_1 and z_2 in the upper half-plane:

$$d(z_1, z_2) = \log \frac{|z_1 - \bar{z}_2| + |z_1 - z_2|}{|z_1 - \bar{z}_2| - |z_1 - z_2|}.$$

Lecture 27

a. Detailed discussion of geodesics and isometries in the upper half-plane model. One of our key examples throughout this course has been the flat torus, a surface whose name indicates that it is a surface of constant zero curvature, and which has Euler characteristic zero. We have also seen that the sphere, which has positive Euler characteristic, has constant positive curvature.

From our considerations of the hyperbolic plane, which we will continue in this lecture, we will eventually see that a sphere with m handles, $m \geq 2$, which is a surface of negative Euler characteristic, can be endowed with a metric under which it has constant negative curvature.

These examples suggest that there might be some connection between curvature and Euler characteristic; this is the content of the *Gauss-Bonnet Theorem*, which we will come to later on.

For the time being, we postpone further discussion of curvature until we have examined the hyperbolic plane in greater detail. Recall the Poincaré upper half-plane model:

$$H^2 = \{\, (x,y) \in \mathbb{R}^2 \mid y > 0\} = \{\, z \in \mathbb{C} \mid \operatorname{Im} z > 0 \,\}.$$

The hyperbolic metric on the upper half-plane is given by a conformal change of the Euclidean metric:

$$ds^2 = \frac{dx^2 + dy^2}{y^2}.$$

Visually, this means that to obtain hyperbolic distances from Euclidean ones, we stretch the plane near the real axis, where $y = \operatorname{Im} z$ is small, and shrink it far away from the real axis, where y is large. Thus if we take a vertical strip which has constant Euclidean width, such as

$$X = \{\, (x,y) \in H^2 \mid 0 \le x \le 1 \,\},$$

and glue the left and right edges together, we will obtain a sort of funnel, or trumpet, in the hyperbolic metric, which is very narrow at large values of y, and flares out hyperbolically as y goes to 0. Part of this construction (at the narrow end of the funnel) is realised on the surface of the pseudosphere mentioned in Lecture 26(b.3).

Now we will present a detailed derivation of the distance formula (4.8), beginning with the special case (4.6). So we take two points $z_1 = x + iy_1$ and $z_2 = x + iy_2$ which lie on the same vertical half-line, where $y_1 < y_2$. The curve $\gamma\colon [y_1, y_2] \to H^2$ given by

$$\gamma(t) = x + it$$

has length given by

$$\ell(\gamma) = \int_{y_1}^{y_2} \|\gamma'(t)\|\, dt = \int_{y_1}^{y_2} \frac{1}{t}\, dt = \log y_2 - \log y_1.$$

To see that this is in fact minimal, let $\eta\colon [a,b] \to H^2$ be any smooth curve with $\eta(a) = z_1$, $\eta(b) = z_2$, and write $\eta(t) = x(t) + iy(t)$. Then we have

$$\begin{aligned}\ell(\eta) &= \int_a^b \|\eta'(t)\|\,dt = \int_a^b \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)}\,dt \\ &\geq \int_a^b \frac{|y'(t)|}{y(t)}\,dt \geq \int_a^b \frac{d}{dt}\log y(t)\,dt = \log y_2 - \log y_1\end{aligned}$$

with equality iff $x'(t) \equiv 0$ and $y'(t) > 0$. Hence vertical lines are geodesics in H^2.

To determine what the rest of the geodesics in H^2 look like, we will examine the images of vertical lines under isometries. First we give another proof (independently of any decomposition of the transformation into a product of simple ones) that *fractional linear transformations*

$$f \colon z \mapsto \frac{az+b}{cz+d},$$

where $a, b, c, d \in \mathbb{R}$ are such that $ad - bc = 1$, are indeed isometries of the hyperbolic plane. If we attempt to write f in terms of the real and imaginary parts of z, we quickly discover why the use of complex numbers to represent H^2 is so convenient:

$$\begin{aligned}f(x,y) &= f(x+iy) \\ &= \frac{ax+iay+b}{cx+icy+d} \\ &= \frac{ax+b+iay}{cx+d+icy} \cdot \frac{ax+b-iay}{cx+d-icy} \\ &= \frac{(ax+b)(cx+d)+acy^2+i(acxy+ady-acxy-bcy)}{(cx+d)^2+(cy)^2} \\ &= F(x,y) + \frac{iy}{(cx+d)^2+c^2y^2}.\end{aligned}$$

The exact form of the real part $F(x,y)$ is unimportant for our purposes here, since ds is independent of the value of x. It is important, however, to note that the denominator of the imaginary part is given by

$$(cx+d)^2 + c^2y^2 = |cx+d+icy|^2 = |cz+d|^2,$$

and hence if we write $f(x,y) = (\tilde{x}, \tilde{y})$, we have

$$\tilde{y} = \frac{y}{|cz+d|^2}.$$

How are we to show that this is an isometry? One conceivable plan of attack would be to compute the distance formula on H^2 and then show directly that the distance between $f(z_1)$ and $f(z_2)$ is the same as the distance between z_1 and z_2 for any two points $z_1, z_2 \in H^2$. This, however, requires computation of an explicit distance formula, which is in fact our ultimate goal. To avoid a vicious circle, we take the infinitesimal point of view and examine the action of f on tangent vectors. That is, we recall that given a map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, the Jacobian derivative Df is a linear map from $\mathbb{R}^2$ to $\mathbb{R}^2$ which takes tangent vectors at (x, y) to tangent vectors at $f(x, y)$. If f is in addition a holomorphic map from $\mathbb{C}$ to (shining) $\mathbb{C}$, then this map $Df_{(x,y)}$ will act on $\mathbb{R}^2$ ($\mathbb{C}$) as multiplication by a complex number $f'(z)$. Geometrically, this means that Df is the composition of a homothety (by the modulus of $f'(z)$) and a rotation (by the argument of $f'(z)$).

In the case of a fractional linear transformation given by the formula above, we have

$$\begin{aligned} f'(z) &= \frac{d}{dz}\frac{az+b}{cz+d} = \frac{a(cz+d)-c(az+b)}{(cz+d)^2} \\ &= \frac{ad-bc}{(cz+d)^2} = \frac{1}{(cz+d)^2} \end{aligned}$$

and hence, writing $f(x, y) = (\tilde{x}, \tilde{y})$ and recalling the form of $\tilde{y}$, we have

$$|f'(z)| = \frac{\tilde{y}}{y}.$$

Now f takes the point $z = x+iy \in H^2$ to the point $\tilde{z} = \tilde{x}+i\tilde{y}$, and Df_z takes the tangent vector $v \in T_zH^2$ to the vector $Df_zv \in T_{\tilde{z}}H^2$. Because Df_z is homothety composed with rotation, we have, *in the Euclidean norm on* $\mathbb{R}^2$,

$$\|Df(v)\|_{\mathrm{Euc}} = |f'(z)| \cdot \|v\|_{\mathrm{Euc}}.$$

The hyperbolic norm is just the Euclidean norm divided by the y-coordinate, and so we have

$$\|Df(v)\|_{\tilde{z}} = \frac{\|Df(v)\|_{\mathrm{Euc}}}{\tilde{y}} = \frac{|f'(z)|}{\tilde{y}}\|v\|_{\mathrm{Euc}} = \frac{1}{y}\|v\|_{\mathrm{Euc}} = \|v\|_z.$$

This is the infinitesimal condition for f to be an isometry; with this fact in hand, it quickly follows that f preserves the length of any curve γ, and hence preserves geodesics and the distances between points.

b. The cross-ratio. The knowledge that fractional linear transformations are isometries allows us to find the rest of the geodesics in H^2; these are simply the images under isometries of the vertical half-lines discussed earlier. This in turn will give us the tools we need to compute the explicit formula (4.8) for the distance between two points $z_1, z_2 \in H^2$. To this end, we make the following definition (the following discussion is valid in $\mathbb{C}$ generally, not just H^2):

Definition 4.6. Given $z_1, z_2, z_3, z_4 \in \mathbb{C}$, the *cross-ratio* is the complex number

$$(z_1, z_2; z_3, z_4) = \frac{z_1 - z_3}{z_2 - z_3} \div \frac{z_1 - z_4}{z_2 - z_4}.$$

This generalises (4.7), where the last two points were taken on the real line. It turns out that *any* fractional linear transformation, whether or not the coefficients lie in $\mathbb{R}$, preserves the cross-ratio.

Lemma 4.7. *Given any $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ and any $z_1, z_2, z_3, z_4 \in \mathbb{C}$, define w_1, w_2, w_3, w_4 by*

$$w_j = \frac{az_j + b}{cz_j + d}$$

for $1 \leq j \leq 4$. Then

$$(w_1, w_2; w_3, w_4) = (z_1, z_2; z_3, z_4).$$

Proof. Straightforward computation; substitute the expressions for w_i into the cross-ratio formula, clear denominators, and notice that constant and quadratic terms (in z_i) cancel out additively, while linear coefficients cancel multiplicatively, leaving the cross-ratio of the z_i as the result. □

As a simpler example of this general idea, one can notice that if we consider triples (z_1, z_2, z_3) of complex numbers, then the *simple ratio*

$$\frac{z_1 - z_3}{z_2 - z_3}$$

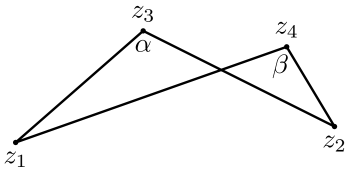

Figure 4.10. Interpreting the cross-ratio of four numbers.

is preserved by the linear map $z \mapsto az+b$ for any $a, b \in \mathbb{C}$. Indeed, the complex number $z_1 - z_3$ is represented by the vector pointing from z_3 to z_1, and similarly $z_2 - z_3$ is the vector from z_3 to z_2. Recall that the argument of the ratio of two complex numbers is given by the difference in their arguments; hence the argument of the above ratio is the angle made by the points z_1, z_3, z_2 taken in that order.

Furthermore, linear transformations are characterised by the fact that they preserve the simple ratio; this can easily be seen by fixing two points z_1 and z_2, and then expressing $f(z)$ in terms of z from the equality

$$\frac{z_1 - z}{z_2 - z} = \frac{f(z_1) - f(z)}{f(z_2) - f(z)}.$$

Later we will use the same argument to show that fractional linear transformations are characterised by the property of preserving the cross-ratio (Lemma 4.9).

As with the simple ratio, the cross-ratio can be interpreted geometrically. Let α be the angle made by z_1, z_3, z_2 in that order, and β the angle made by z_1, z_4, z_2, as in Figure 4.10. Then the argument of the cross-ratio is just $\alpha - \beta$. In particular, if $\alpha = \beta$, then the cross-ratio is a positive real number; this happens iff the points z_1, z_2, z_3, z_4 all lie on a circle with z_1 adjacent to z_2 and z_3 adjacent to z_4 as in the picture, or if they are collinear.

If $\alpha - \beta = \pi$, the four points still lie on a circle (or possibly a line), but now the order is changed; z_4 will have moved to a position between z_1 and z_2 on the circumference. The upshot of all of this is that the cross-ratio is a real number iff the four points lie on a circle or a line. Because fractional linear transformations preserve cross-ratios, we have proved the following theorem.

Theorem 4.8. *If γ is a line or a circle in $\mathbb{C}$ and $f\colon \mathbb{C} \to \mathbb{C}$ is a fractional linear transformation, then $f(\gamma)$ is also a line or a circle.*

There are other ways of proving this theorem, but they involve either a fair amount of algebra using the characterisations of lines and circles in terms of z and $\bar{z}$, or a synthetic argument which requires the decomposition of fractional linear transformations into maps of particular types.

It is worth noting that if we think of all this as happening on the Riemann sphere rather than on the complex plane, we can dispense with this business of 'lines and circles'. Recall that the Riemann sphere is the complex plane $\mathbb{C}$ together with a point at infinity; circles in the plane are circles on the sphere which do not pass through the point at infinity, and lines in the plane are circles on the sphere which do pass through the point at infinity. Fractional linear transformations also assume a nicer form, once we make the definitions

$$f(\infty) = \frac{a}{c}, \quad f\left(-\frac{d}{c}\right) = \infty.$$

Returning to the hyperbolic plane, we now make use of the fact that fractional linear transformations preserve angles (because they are conformal) and cross-ratios (as we saw above). In particular, the image of a vertical line under such a transformation f is either a vertical line, which we already know to be a geodesic, or a circle; because angles are preserved and because f preserves the real line (by virtue of having coefficients in $\mathbb{R}$), this circle must intersect $\mathbb{R}$ perpendicularly, and hence must have its centre on the real line.

This allows us to conclude our detailed derivation of the distance formula (4.8) by establishing that semicircles whose centre lies in $\mathbb{R}$ are also geodesics. Let f be a fractional linear transformation which maps the vertical half-line $\{\, z \in H^2 \mid \operatorname{Re} z = 0 \,\}$ to the semicircle $\{\, z \in H^2 \mid |z - a_0| = r \,\}$. Given two points z_1, z_2 lying on the semicircle, we have $z_1 = f(iy_1)$ and $z_2 = f(iy_2)$; hence $d(z_1, z_2) = d(iy_1, iy_2)$ since f is an isometry.

Furthermore, supposing without loss of generality that $y_1 > y_2$, we see that $f(0)$ and $f(\infty)$ are the two points where the circle intersects $\mathbb{R}$. Denote these by w_1 and w_2, respectively; then w_1 lies closer to z_1, and w_2 lies closer to z_2. Since f preserves cross-ratios, we have

$$\begin{aligned}(z_1, z_2; w_1, w_2) &= (iy_1, iy_2; 0, \infty)\\ &= \frac{iy_1 - 0}{iy_2 - 0} \div \frac{iy_1 - \infty}{iy_2 - \infty} = \frac{y_1}{y_2}\end{aligned}$$

and recalling that $d(iy_1, iy_2) = \log y_1 - \log y_2 = \log(y_1/y_2)$, the fact that f is an isometry implies

$$d(z_1, z_2) = \log(z_1, z_2; w_1, w_2).$$

If we remove the assumption that $y_1 > y_2$, we must take the absolute value of this quantity.

In order to show that this analysis is complete, we must show that there are no other geodesics in H^2 other than those described here. This will follow once we know that any two points $z_1, z_2 \in H^2$ either lie on a vertical half-line or on a semicircle whose centre is in $\mathbb{R}$, and that any such half-line or semicircle can be obtained as the image of the imaginary axis under a fractional linear transformation.

The former assertion is straightforward, as described in the previous lecture (Figure 4.9). To see the latter, note that horizontal translation $z \mapsto z + t$ and homothety $z \mapsto \lambda z$ are both fractional linear transformations, and that using these, we can obtain any vertical half-line from any other, and any semicircle centred in $\mathbb{R}$ from any other. Thus we need only obtain a circle from a line, and this is accomplished by considering the image of the vertical line $\operatorname{Re} z = 1$ under the fractional linear transformation $z \mapsto -1/z$, which will be a circle of radius $1/2$ centred at $-1/2$.

Exercise 4.13. Prove that fractional linear transformations of the form

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

where $a, c \in \mathbb{C}$ satisfy $a\bar{a} - c\bar{c} = 1$, represent isometries of the hyperbolic plane in the disc model.

c. Circles in the hyperbolic plane. Theorem 4.8 raises a natural question: what is the intrinsic meaning of the curves in the hyperbolic plane which are represented in the models by lines, rays, intervals, circles, or arcs of circles?

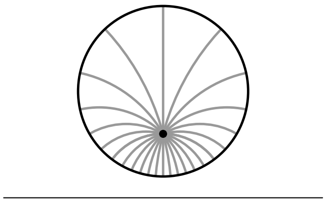

Figure 4.11. Hyperbolic centre and radii of a circle in H^2.

As we have seen, some of these are geodesics; in fact, a necessary and sufficient condition for that is that the curve (or its extension) cross the real line (in the half-plane model) or the unit disc (in the disc model) at a right angle. But what are the rest?

We have seen one example: in the disc model, the circles centred at the origin represent circles in the hyperbolic metric. Hence any image of such a circle under a fractional linear hyperbolic isometry, which must be a (Euclidean) circle by Theorem 4.8, also represents a hyperbolic circle. Now using the inverse of the transformation (4.5), these circles are mapped to circles in the upper half-plane, which thus also represent hyperbolic circles. In the upper half-plane, any circle can be mapped into any other circle by a linear transformation with real coefficients, and so we conclude that any circle inside the upper half-plane represents a hyperbolic circle. Applying (4.5), we reach the same conclusion for the disc model.

Finally, we need to show that *any* hyperbolic circle is represented this way. Let γ be a hyperbolic circle in the disc model and p be its centre in the hyperbolic metric. There is a hyperbolic isometry, represented by a fractional linear transformation, which maps p into the origin, and γ into a hyperbolic circle centred at the origin, which is represented by a Euclidean circle. Hence γ is a Euclidean circle as well. This carries over to the upper half-plane model, and so we have proved the following fact:

Proposition 4.13. *In both the upper half-plane and the disc models, circles in hyperbolic metric are represented by Euclidean circles; conversely, every Euclidean circle which lies inside the half-plane or the disc represents a hyperbolic circle.*

What about Euclidean circles which do not lie inside the upper half-plane or the disc, but which intersect the ideal boundary? They are not closed curves in H^2, and so cannot be circles; if they do not meet the ideal boundary at a right angle, they are not geodesics. So what are they? We will address this question in Lecture 29, where we make a more detailed study of the isometries of the hyperbolic plane.

Exercise 4.14. Calculate the hyperbolic radius and the hyperbolic centre of the circle in H^2 given by the equation

$$\|z - 2i - 1\|^2_{\mathrm{Euc}} = 9/4.$$

Lecture 28

a. Three approaches to hyperbolic geometry. As we continue to plan our assault on the mountain of hyperbolic geometry, there are three main approaches that we might take: the synthetic, the analytic, and the algebraic.

a.1. The first of these, the *synthetic* approach, proceeds along the same lines as the classical Euclidean geometry which is (or used to be, at any rate) taught as part of any high school education. One approaches the subject axiomatically, formulating several postulates and then deriving theorems from these basic assumptions. From this point of view, the only difference between the standard Euclidean geometry one learns in school and the hyperbolic non-Euclidean geometry we are investigating here is the failure of Euclid's fifth postulate, the parallel postulate, in our present case.

This postulate can be stated in many forms; the most common formulation is the statement that given a line and a point not on that line, there exists exactly one line through the point which never intersects the original line. One could also state that the measures of the angles of any triangle sum to π radians, or that there exist triangles with equal angles which are not isometric, and there are many other equivalent formulations.

In hyperbolic geometry, this postulate is no longer valid; however, any theorem of Euclidean geometry which does not rely on this postulate still holds. The common body of such results is known as *absolute* or *neutral geometry*, and the historical approach from the time of Euclid until the work of Lobachevsky and Bolyai in the nineteenth century was to attempt to prove that the parallel postulate in fact follows from the others. The synthetic approach, then, uses the result that if the parallel postulate can be added to the axioms of absolute geometry without fear of contradiction, then its negation can as well, and proceeds axiomatically assuming that negation.

a.2. The second approach is the *analytic* one, which we have made some use of thus far; one derives and then makes use of formulae for lengths, angles, and areas. This approach has the advantage of being the most general of the three, in that it can be applied to any surface, whereas both the synthetic and the algebraic approaches have limited applicability beyond the highly symmetric examples of the Euclidean and hyperbolic (and, to a certain extent, elliptic) planes. Hyperbolic trigonometry can be associated with this approach too.

a.3. For the time being, however, we will make use of the symmetry possessed by the hyperbolic plane, which allows us to take the third option, the *algebraic* approach. In this approach, we study the isometry group of H^2 and use properties of isometries to understand various aspects of the surface itself, a process in which linear algebra becomes an powerful and invaluable tool.

b. Characterisation of isometries. First, then, we must obtain a complete description of the isometries of H^2. We saw in the previous lecture that fractional linear transformations of the form

$$z \mapsto \frac{az+b}{cz+d}$$

are orientation preserving isometries of H^2 in the upper half-plane model for any $a, b, c, d \in \mathbb{R}$ with $ad - bc = 1$. But what about orientation reversing isometries? Since the composition of two orientation reversing isometries is an orientation preserving isometry, once we have understood the orientation preserving isometries it will suffice to exhibit a single orientation reversing isometry. Such an isometry is given by the map

$$z \mapsto -\bar{z}$$

which is reflection in the imaginary axis. By composing this with fractional linear transformations of the above form, we obtain a family of orientation reversing isometries of the form

$$z \mapsto \frac{-a\bar{z}+b}{-c\bar{z}+d}$$

where again, $a, b, c, d \in \mathbb{R}$ are such that $ad - bc = 1$. By changing the sign on a and c, we can write each of these isometries as

$$z \mapsto \frac{a\bar{z}+b}{c\bar{z}+d} \tag{4.9}$$

where $ad - bc = -1$.

Now we claim that these are in fact all of the isometries of H^2. The following argument for the hyperbolic plane can in fact be made to work in much greater generality, and says that for any surface the isometry group is not 'too big'.

We will show that any isometry I is uniquely determined by the images of three points which do not lie on the same geodesic (recall Figure 1.20). Given that $I(A) = \tilde{A}$ and $I(B) = \tilde{B}$, let γ be the unique geodesic connecting A and B, and $\tilde{\gamma}$ the unique geodesic connecting $\tilde{A}$ and $\tilde{B}$. Then because $I(\gamma)$ is also a geodesic connecting $\tilde{A}$ and $\tilde{B}$, we must have $I(x) \in \tilde{\gamma}$ for every $x \in \gamma$. Furthermore, the distance along γ from x to A must be the same as the distance along $\tilde{\gamma}$ from $I(x)$ to $\tilde{A}$, and similarly for B. This requirement uniquely determines the point $I(x)$.

This demonstrates that the action of I on two points of a geodesic is sufficient to determine it uniquely on the entire geodesic. It follows that I is uniquely determined on the three geodesics connecting A, B, and C by its action on those three points; thus we know the action of I on a geodesic triangle. But now given any point $y \in S$, we may draw a geodesic through y which passes through two points of that triangle; it follows that the action of I on those two points, which we know, determines $I(y)$.

Thus we have established uniqueness, but not existence, of an isometry taking A, B, and C to $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$. Indeed, given two sets of three points, it is not in general true that some isometry carries one set to the other. As a minimal requirement, we see that the pairwise distances between the points must be the same; we must have $d(A, B) = d(\tilde{A}, \tilde{B})$ and so on. If our surface is symmetric enough, this condition will be sufficient, as is the case for the Euclidean plane and the round sphere; we will soon see that this is also the case for H^2. First, we prove a fundamental lemma concerning fractional linear transformations in general.

Lemma 4.9. *Let (z_1, z_2, z_3) and (w_1, w_2, w_3) be two triples of distinct points in the extended complex plane $\mathbb{C} \cup \{\infty\}$ (the Riemann sphere). Then there exist unique coefficients $a, b, c, d \in \mathbb{C}$ such that the fractional linear transformation*

$$f\colon z \mapsto \frac{az+b}{cz+d}$$

satisfies $f(z_j) = w_j$ for $j = 1, 2, 3$. Furthermore, any map from $\mathbb{C} \cup \{\infty\}$ to itself which preserves the cross-ratio is a fractional linear transformation.

Proof. Recall that fractional linear transformations preserve cross-ratios, and hence if for some $z \in \mathbb{C}$ the f we are looking for has $f(z) = w$, we must have

$$(z_1, z_2; z_3, z) = (w_1, w_2; w_3, w). \tag{4.10}$$

Using the expression for the cross-ratio, we have

$$\frac{(z_1 - z_3)(z_2 - z)}{(z_2 - z_3)(z_1 - z)} = \frac{(w_1 - w_3)(w_2 - w)}{(w_2 - w_3)(w_1 - w)},$$

and solving this equation for w in terms of z will give the desired fractional linear transformation:

$$w = \frac{w_1(z_1 - z_3)(w_2 - w_3)(z_2 - z) - w_2(z_2 - z_3)(w_1 - w_3)(z_1 - z)}{(z_1 - z_3)(w_2 - w_3)(z_2 - z) - (z_2 - z_3)(w_1 - w_3)(z_1 - z)}. \tag{4.11}$$

Since (4.10) implies (4.11) we also get the second statement. □

Proposition 4.14. *Given points $z_1, z_2, z_3, w_1, w_2, w_3 \in H^2$ satisfying $d(z_j, z_k) = d(w_j, w_k)$ for each pair of indices (j, k), there exists a unique isometry taking z_k to w_k. If the geodesic triangles z_1, z_2, z_3 and w_1, w_2, w_3 have the same orientation, this isometry is orientation preserving and is represented by a fractional linear transformation; otherwise it is orientation reversing and has the form* (4.9).

Remark. The first part of this proposition states that given two triangles in H^2 whose corresponding sides are of equal length, there

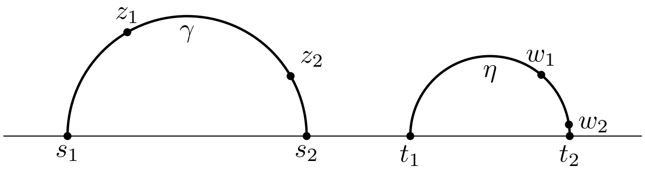

Figure 4.12. The images of two points determine a unique fractional linear transformation.

exists an isometry of H^2 taking one triangle to the other. This statement is true in Euclidean geometry as well, and in fact holds as a result in absolute geometry. As such, it could be proven in a purely synthetic manner; while such an approach does in fact succeed, we will take another path and use our knowledge of fractional linear transformations.

Notice that, while Lemma 4.9 gives us a fractional linear transformation which is a candidate to be an isometry, this candidate is the desired isometry only if the orientations of the triangles z_1, z_2, z_3 and w_1, w_2, w_3 coincide.

We first prove that the group of fractional linear transformations with real coefficients acts transitively on *pairs* of points (z_1, z_2), where the distance $d(z_1, z_2)$ is fixed. We then use the fact that a third point z_3 has only two possible images under an isometry, and that the choice of one of these as w_3 determines whether the isometry preserves or reverses orientation.

Proposition 4.15. *Given points $z_1, z_2, w_1, w_2 \in H^2$ with $d(z_1, z_2) = d(w_1, w_2)$, there exists a unique fractional linear transformation f satisfying $f(z_j) = w_j$ for $j = 1, 2$. This transformation f has real coefficients and hence is an isometry of H^2.*

Proof. Let γ be the geodesic connecting z_1 and z_2, and η the geodesic connecting w_1 and w_2. Let s_1 and s_2 be the two points where γ intersects $\mathbb{R}$, with s_1 nearer to z_1 and s_2 nearer to z_2, and define t_1 and t_2 similarly on η, as shown in Figure 4.12.

By Lemma 4.9, there exists a unique fractional linear transformation f with complex coefficients such that $f(s_1) = t_1$, $f(z_1) = w_1$, and $f(z_2) = w_2$. In order to complete the proof, we must show that f in fact preserves the real line, and hence has real coefficients.

Recalling our distance formula for H^2 in terms of the cross-ratio, the condition that $d(z_1, z_2) = d(w_1, w_2)$ can be rewritten as

$$(z_1, z_2; s_1, s_2) = (w_1, w_2; t_1, t_2).$$

From the proof of Lemma 4.9, this was exactly the formula that we solved for t_2 to find $f(s_2)$; it follows that $f(s_2) = t_2$. Since f is a conformal map which takes lines and circles to lines and circles, and since $\mathbb{R}$ intersects γ orthogonally at s_1 and s_2, the image of $\mathbb{R}$ is a line or circle which intersects η orthogonally at t_1 and t_2, and hence is in fact $\mathbb{R}$.

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

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 $z = f(z)$:

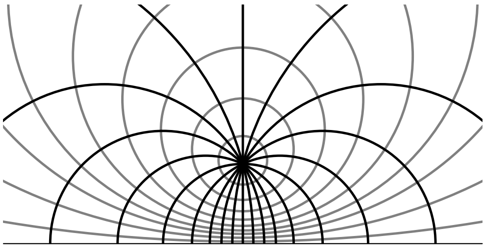

Figure 4.13. Geodesics passing through i and hyperbolic circles centred at i.

- **(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 $z_1, z_2 \in \mathbb{R}$.

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:

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

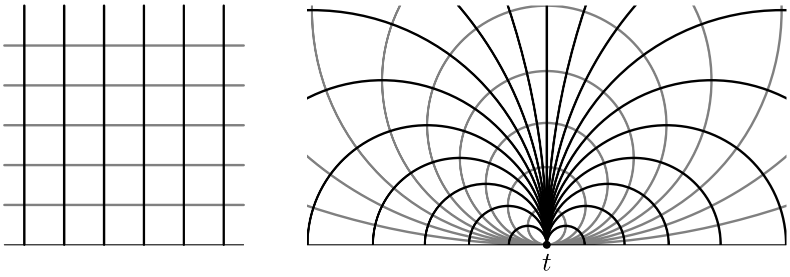

Figure 4.14. Parallel geodesics and horocycles for parabolic isometries.

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 semicircle 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:

(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.

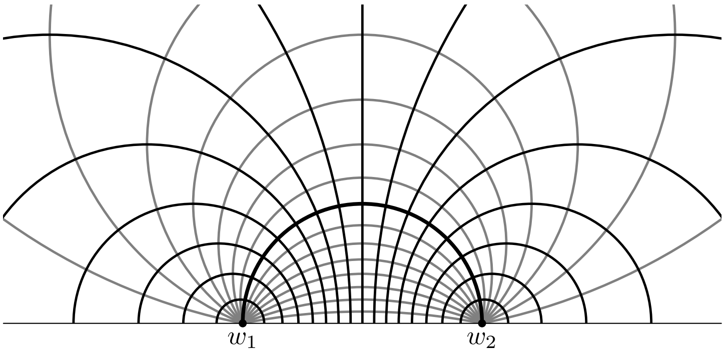

Figure 4.15. Orthogonal geodesics and equidistant curves for the geodesic connecting w_1 and w_2.

(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 ∞:

$$f: z \mapsto \lambda^2 z.$$

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.

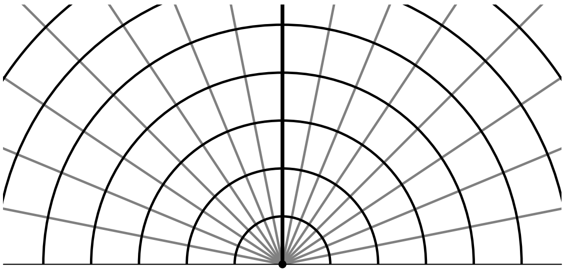

Figure 4.16. Orthogonal geodesics and equidistant curves for the imaginary axis.

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 preserve 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*

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, \qquad \alpha \in \mathbb{R}.$$

(P): *The* parabolic *matrix*

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

(H): *A* hyperbolic *matrix of the form*

$$\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}, \qquad t \in (0, \infty).$$

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 semicircle) 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.

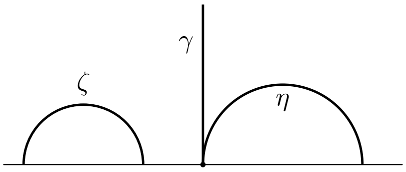

Figure 4.17. Parallels and ultraparallels.

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) 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.

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 arbitrarily 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

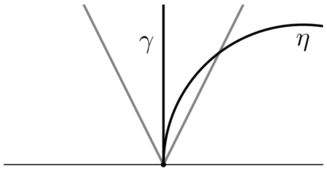

Figure 4.18. Distance between parallel geodesics.

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

$$\begin{aligned} f\colon \{\gamma_v\}_{v\in T_pH^2} &\to \{\gamma_v\}_{v\in T_pH^2}, \\ \gamma_v &\mapsto \gamma_w, \end{aligned}$$

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

$$\begin{aligned} f\colon \{\gamma_t\}_{t\in\mathbb{R}} &\to \{\gamma_t\}_{t\in\mathbb{R}}, \\ \gamma_t &\mapsto \gamma_{t+1}. \end{aligned}$$

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 semicircle connecting the fixed points, and the equidistant curves are the other circles passing through those two points. The family of orthogonal curves comprises 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.