ISOMETRIES OF THE HYPERBOLIC PLANE

KATHY SNYDER

Abstract. In this paper I will define the hyperbolic plane and describe and classify its isometries. I will conclude by showing how these isometries can be represented as 2×2 matrices.

Contents

1. Isometries of $\mathbb{H}^2$

The hyperbolic plane is defined as the upper half plane, without boundary, endowed with a certain metric. We denote the set of points by

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

Definition 1.1. A *path* from a to b for $a, b \in \mathbb{H}^2$ is a map $\gamma\colon [0; 1] \to \mathbb{H}^2$ such that γ is differentiable, $\gamma(0) = a$, and $\gamma(1) = b$.

We may also find it useful to write $\gamma(t) = (x_\gamma(t), y_\gamma(t))$.

Definition 1.2. We define the *arclength* of a path as

$$L(\gamma) = \int_0^1 \frac{|\gamma'(t)|}{y_\gamma(t)}\, dt.$$

This allows us to define a metric on $\mathbb{H}^2$ as follows:

Definition 1.3. The distance between two points, $a, b \in \mathbb{H}^2$, is defined as $d(a, b)$ $= \inf\{L(\gamma) : \gamma$ a path from a to $b\}$.

Let us verify that this is a metric:

Proof. (1) $d(a, b) \geq 0$ because $|\gamma'(t)|$ and $y_\gamma(t)$ are ≥ 0, so therefore the integral is nonnegative as well. If $d(a, b) = 0$, then let $a = (a_1, a_2)$ and $b = (b_1, b_2)$, and suppose $a \neq b$. Then WLOG assume $a_1 \neq b_1$. Then for any path γ from a to b, $x_\gamma(t)$ is not constant. Then $L(\gamma) \geq 0$. So it must be that $a = b$. Now, suppose $a = b$. Then the constant path $\gamma(t) = a$ is a path from a to b. However, it has length $L(\gamma) = 0$ because $|\gamma'(t)| = 0$. So $d(a, b) \leq 0$; but $d(a, b)$ is at least 0. So $d(a, b) = 0$.

Date: August 2009.

(2) Every path from a to b is also a path from b to a if one takes the same path backwards. That is, for each h in the image of a path, with preimage t, we define a new path by taking the preimage of h to be $1-t$. So the preimage of a becomes 1 instead of 0, and so on. Therefore $d(a,b) = d(b,a)$.

(3) Finally we must show that $d(a,c) \leq d(a,b) + d(b,c)\ \forall\ a,b \in \mathbb{H}^2$. Suppose that $d(a,c) > d(a,b) + d(b,c)$. Then $\exists\ \varepsilon > 0$ such that $d(a,c) = d(a,b) + d(b,c) + \varepsilon$. Then there exist paths γ_1 from a to b and γ_2 from b to c such that $L(\gamma_1) < d(a,b) + \frac{\varepsilon}{2}$ and $L(\gamma_2) < d(b,c) + \frac{\varepsilon}{2}$. Then we can construct a path γ' from a to c by concatenating paths γ_1 and γ_2 and reparametrizing them. Note that γ_1 and γ_2 might meet at a cusp, and therefore their concatenation might not be differentiable. However, since we can approximate their concatenation with a path that is differentiable and has arclength arbitrarily close to the arclength of the concatenation, this is not a problem. Then we have $L(\gamma') = L(\gamma_1) + L(\gamma_2) < d(a,b) + d(b,c) + \varepsilon < d(a,c)$, which is a contradiction, since $d(a,c)$ is the infimum of all paths from a to c. Therefore, the triangle inequality must hold. □

We are particuluary interested in when the distance between two points is realized by an actual path.

Definition 1.4. A *geodesic* is a path γ from a to b such that $L(\gamma) = d(a,b)$.

Later in the paper, we will show that there exists a unique geodesic between every two points. However, at the moment we are able to show this much:

Lemma 1.5. *For any two points a and b on a vertical line, the line segment between them is the unique geodesic from a to b.*

Proof. Without loss of generality, let our two points be of the form $(1,n)$, $(1,m)$, where $m > n$. Then let $\gamma : t \mapsto (1, n + (m-n)t)$. Then $L(\gamma) = \int_0^1 \frac{|m-n|}{n+(m-n)t}\,dt$. Let $u = n + (m-n)t$. Then $du = (m-n)dt$ So we have $L(\gamma) = \int_n^m \frac{1}{u}\,du = \ln m - \ln n = \ln \frac{m}{n}$.

Now I claim that this path is the unique geodesic from $(1,n)$ to $(1,m)$. Suppose that there exists another path $\mu \neq \gamma$ up to parameterization, such that μ has shorter arclength than γ. Suppose x_μ is non-constant. Then $x_\mu'(t) > 0$ for some nontrivial subinterval of $[0;1]$, since x_μ is continuous. Then it is true that $\sqrt{x_\mu'(t)^2 + y_\mu'(t)^2} > \sqrt{0 + y_\mu'(t)^2}$. That is, $|\mu'(t)|$ is larger when x_μ is nonconstant than it is when x_μ is constant, for any given y_μ. Therefore, the integral $L(\mu)$ is also smaller when x_μ is constant. So for any path from $(1,n)$ to $(1,m)$, we can always find a shorter path by making the x_μ coordinate constant. Therefore we can assume that $x_\mu(t) = 1$, since we are assuming μ is a path with shortest arclength.

Now, assuming x_μ is constant, what about $y_\mu(t)$? Suppose $y_\mu(t) \neq a + (b-a)t$, up to parameterization. That is, the image $y_\mu(t)$ does not simply monotonically increase from a to b, but also decreases in some places. Then we define an alternate map y_μ^* as follows: suppose after some $s \in [0;1]$ such that $y_\mu(s) \leq m$, y_μ begins to decrease. Since y_μ is continuous, and since $y_\mu(1) = m$, there exists some $r > s$, $r \in [0;1]$, such that $y_\mu(r) = y_\mu(s)$, by the Intemediate Value Theorem. Then for all $x \in [s;r]$ we define $y_\mu^*(t) = y_\mu(s)$. In other words, where y_μ has a valley or U shape in its image, y_μ^* remains constant. Let us consider the arclength of y_μ^* compared to y_μ. In particular, y_μ^* only differs from y_μ on the valleys, so let us consider the arclength function on one of those intervals. Now, since $x_\mu(t)$ is constant, $|\mu'(t)| = |y'_\mu(t)|$. Now $\frac{d}{dt}y^*_\mu = 0$ on the interval, but $y'_\mu(t) < 0$ for the subinterval on which y_μ is decreasing. So $|y'_\mu(t)| > 0$ for some subinterval of positive measure. Therefore, in our arclength function, the numerator of $\int_0^1 \frac{|y^{*\prime}_\mu(t)|}{y^*_\mu(t)}\,dt$ is larger than the numerator of $\int_0^1 \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt$. Furthermore, since $y_\mu(t)$ decreases in the valley, the denomenator of $\frac{|y'_\mu(t)|}{y_\mu(t)}$ is smaller than the denominator of $\frac{|y^{*\prime}_\mu(t)|}{y^*_\mu(t)}$, since $y^*_\mu(t) \geq y_\mu(t)$ for all t in the interval we are considering. Then overall, the fraction $\frac{|y^{*\prime}_\mu(t)|}{y^*_\mu(t)} \leq \frac{|y'_\mu(t)|}{y_\mu(t)}$. So $L(\mu^*) \leq L(\mu)$.

However, there is one more case to consider. What if y_μ begins to decrease at some $s \in [0;1]$ such that $y_\mu(s) > m$? In this case, y_μ goes above m and then comes back down, creating a hill over some interval $[c;d] \subset [0;1]$. Let us break up the arclength integral as follows (remember that since x_μ is constant, $|\mu'(t)| = |y'_\mu(t)|$):

$$L(\mu) = \int_0^c \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt + \int_c^d \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt + \int_d^1 \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt.$$

Now I claim that

$$\int_0^c \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt + \int_d^1 \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt \geq d((1,n),(1,m)).$$

This is true because we could define a new path μ^* such that $y_\mu(t) = m$ on the interval $[c;d]$. (Actually, there might be cusps in this modified path, but we could approximate this new path by a differentiable one, so this is not a problem.) Then $|\frac{d}{dt}y^*_\mu(t)| = 0$ on $[c;d]$, so $\int_c^d \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt$ would also be 0. So $L(\mu^*) \leq L(\mu)$. In particular, this gives us that $L(\mu) \geq L(\mu^*) \geq d((1,n),(1,m))$. But we can do even better than this, because we know that the derivative $y'_\mu(t)$ is nonzero for some subinterval of $[c;d]$, because we picked this interval such that y_μ was decreasing on some part of it. Therefore $|y'_\mu(t)| > 0$ on some subinterval of $[c;d]$, and so the integral

$$\int_c^d \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt > 0$$

is positive as well. Therefore we have

$$L(\mu) \geq d((1,n),(1,m)) + \int_c^d \frac{|y'_\mu(t)|}{y_\mu(t)}\,dt > d((1,n),(1,m)).$$

So μ is not the minimal path from $(1,n)$ to $(1,m)$.

Therefore we can conclude that γ as defined above is, up to parameterization, the unique geodesic from $(1,n)$ to $(1,m)$. $\square$

Now we are ready to define some isometries of $\mathbb{H}^2$.

Definition 1.6. A *translation* $\mathrm{T}_s : \mathbb{H}^2 \to \mathbb{H}^2$ is defined by $\mathrm{T}_s(x,y) = (x+s,y)$.

Definition 1.7. For $\lambda > 0$, a *dilation* $\mathrm{D}_\lambda : \mathbb{H}^2 \to \mathbb{H}^2$ is defined by $\mathrm{D}_\lambda(x,y) = (\lambda x, \lambda y)$.

First of all, it is clear that T_s and D_λ are continuous and differentiable because this is true in each coordinate. Furthermore, they map paths from a to b to paths from the image of a to the image of b, because the composition of differentiable function is a differentiable function.

So, if T_s and D_λ preserve the arclength of each path, they will also preserve distance, since distance is the infimum of the arclength of all paths.

Theorem 1.8. T_s *and* D_λ *are isometries of* $\mathbb{H}^2$.

Proof. Let $a, b \in \mathbb{H}^2$, and let γ be a path from a to b.

(1) $\mathrm{T}_s \circ \gamma = (x_\gamma(t) + s, y_\gamma(t))$. Then $\frac{d}{dt}\,\mathrm{T}_s \circ \gamma = (x'_\gamma(t), y'_\gamma(t)) = \gamma'(t)$, so $L(\mathrm{T}_s \circ \gamma) = \int_0^1 \frac{|\gamma'(t)|}{y_\gamma(t)}\,dt = L(\gamma)$.
(2) $\mathrm{D}_\lambda \circ \gamma = (\lambda x(t), \lambda y(t))$. Therefore, $\frac{d}{dt}\,\mathrm{D}_\lambda \circ \gamma = \lambda\gamma'(t)$. So $L(\mathrm{D}_\lambda \circ \gamma) = \int_0^1 \frac{|\lambda\gamma'(t)|}{\lambda y_\gamma(t)}\,dt = \int_0^1 \frac{|\gamma'(t)|}{y_\gamma(t)}\,dt$, since $\lambda > 0$; but that is equal to $L(\gamma)$.

Then, as we observed above, T_s and D_λ are isometries because they preserve arclength, and therefore distance. □

It is of interest to us to consider the conjugates of these two types of isometries:

$$\mathrm{T}_s \circ \mathrm{D}_\lambda \circ \mathrm{T}_{-s}\colon (x, y) \mapsto (x - s, y) \mapsto ((\lambda x - \lambda s, \lambda y) \mapsto (\lambda(x - s) + s, \lambda y).$$

This corresponds to a dilation by λ about the boundary point $(s, 0)$ (i.e. the vertical line (s, y) is invariant under this transformation).

We also have

$$\mathrm{D}_\lambda \circ \mathrm{T}_s \circ \mathrm{D}_{\frac{1}{\lambda}}\colon (x, y) \mapsto (\frac{1}{\lambda}x, \frac{1}{\lambda}y) \mapsto (\frac{1}{\lambda}x + s, \frac{1}{\lambda}y) \mapsto (x + \lambda s, y),$$

which amounts to the translation $\mathrm{T}_{s\lambda}$.

At this point we are equipped to show $\mathbb{H}^2$ is homogeneous, meaning that no point in $\mathbb{H}^2$ is special.

Lemma 1.9. *For any* $a, b \in \mathbb{H}^2$*, there exists an isometry that maps* a *to* b*.*

Proof. Let $(a_1, a_2) \neq (b_1, b_2)$ be two points in $\mathbb{H}^2$. Then

$$\mathrm{T}_{b_1 - \frac{b_2 a_1}{a_2}} \circ \mathrm{D}_{\frac{b_2}{a_2}}\colon (a_1, a_2) \mapsto (\frac{b_2 a_1}{a_2}, b_2) \mapsto (b_1, b_2).$$

Note that since a_2, $b_2 > 0$, the dilation is well-defined. □

Definition 1.10. We define a map R: $\mathbb{H}^2 \to \mathbb{H}^2$ as follows: $\mathrm{R}(x, y) = (\frac{x}{x^2+y^2}, \frac{y}{x^2+y^2})$.

First let us understand what this map does. Consider any point on the unit circle. That is, any point (x, y) such that $x^2 + y^2 = 1$. In this case R takes (x, y) to $(\frac{x}{1}, \frac{y}{1})$, so R fixes all points on the unit circle. Next, consider any line of the form (x, ax). R: $(x, ax) \mapsto (\frac{x}{(a^2+1)x^2}, \frac{ax}{(a^2+1)x^2})$. Thus, these lines are invariant under R. If we look at a single line (x, ax), the two sides of the line, one inside the circle, and the other outside, are exchanged by R. For this reason, in the future we will refer to R as a "reflection".

Theorem 1.11. R *is an isometry.*

Proof. Let $a, b \in \mathbb{H}^2$ and let $\gamma(t)$ be a path from a to b. We will find it easier to consider $\gamma(t)$ as a function in polar coordinates, as follows:

$$\gamma(t) = (r(t)\cos(\theta(t)), r(t)\sin(\theta(t))).$$

Then

$$\mathrm{R} \circ \gamma(t) = \Big(\frac{1}{r(t)}\cos(\theta(t)), \frac{1}{r(t)}\sin(\theta(t))\Big).$$

So we have

$$L(\mathrm{R}(\gamma)) = \int_0^1 \frac{|(\frac{-1}{r(t)^2}r(t)\cos\theta(t) - \frac{1}{r(t)}\sin\theta(t), \frac{-1}{r(t)^2}r'(t)\sin\theta(t) + \frac{1}{r(t)}\cos\theta(t))|}{\frac{1}{r(t)}\sin\theta(t)}\, dt$$

which when we simplify, yields $\int_0^1 \frac{\sqrt{r'(t)^2+r(t)^2}}{\sin\theta(t)}\, dt$.
On the other hand, we have

$$\begin{aligned} L(\gamma) &= \int_0^1 \frac{|(r'(t\cos\theta(t) - r(t)\sin\theta(t), r'(t)\sin\theta(t) + r(t)\cos\theta(t))}{r(t)\sin\theta(t)}\, dt \\ &= \int_0^1 \frac{\sqrt{r'(t)^2 + r(t)^2}}{r(t)\sin\theta(t)}\, dt. \end{aligned}$$

So we can conclude that $L(\mathrm{R}(\gamma)) = L(\gamma)$. Therefore, R preserves arclength, and therefore, R preserves distance. □

Let us find the conjugates of R by dilations and translations.
(1)

$$\begin{aligned} \mathrm{T}_s \circ \mathrm{R} \circ \mathrm{T}_{-s} \colon (x, y) &\mapsto (x - s, y) \\ &\mapsto \left(\frac{x - s}{(x-s)^2 + y^2}, \frac{y}{(x-s)^2 + y^2}\right) \\ &\mapsto \left(\frac{(x - s)}{(x-s)^2 + y^2} + s, \frac{y}{(x-s)^2 + y^2}\right), \end{aligned}$$

which fixes points satisfying $(x-s)^2 + y^2 = 1$, that is, the circle of radius 1 centered about the point $(s, 0)$.
(2)

$$\begin{aligned} \mathrm{D}_\lambda \circ \mathrm{R} \circ \mathrm{D}_{\frac{1}{\lambda}} \colon (x, y) &\mapsto \left(\frac{x}{\lambda}, \frac{y}{\lambda}\right) \\ &\mapsto \left(\frac{x/\lambda}{(x^2 + y^2)/\lambda^2}, \frac{y/\lambda}{(x^2 + y^2)/\lambda^2}\right) = \left(\frac{\lambda x}{x^2 + y^2}, \frac{\lambda y}{x^2 + y^2}\right) \\ &\mapsto \left(\frac{\lambda^2 x}{x^2 + y^2}, \frac{\lambda^2 y}{x^2 + y^2}\right). \end{aligned}$$

.

As we can see, by conjugating R by translations or dilations, we can get a reflection R about a semi-circle of any radius centered about any point on the x-axis.

Note, furthermore, that one special reflection is the one about a vertical line, for example, the y-axis, which maps (x, y) to $(x, -y)$. This can be thought of loosely as a reflection about the circle centered at infinity.

We define one more type of isometry, namely, a rotation. On the Euclidean plane, we can obtain any rotation by reflecting across two intersecting lines. As it turns out, we will define rotations in the hyperbolic plane as the composition of two reflections as well. Since these so-called rotations look considerably more odd than those in the Euclidean plane, we start by simply looking at a small neighborhood of the point fixed by the rotation.

Consider any point $p \in \mathbb{H}^2$. Let $T_p\mathbb{H}^2$ denote the tangent space of p; that is, the space of infinitesimal vectors centered at the point p. We can think of the tangent space as isometric to $\mathbb{R}^2$ with a similarly scaled metric. In $\mathbb{R}^2$, a rotation of angle θ can be obtained by composing two reflections about lines that intersect at an angle of $\theta/2$. In $\mathbb{H}^2$, we take two reflections about semi-circles intersecting at our point p, and in the tangent space $T_p\mathbb{H}^2$, the circles look like straight lines, so their composition gives us a rotation of the tangent space.

So let us show that for any θ, we can find two circles with centers on the x-axis that intersect at our point p, and whose tangent vectors at p intersect at angle θ.

Consider the two semi-circles of radius 1 centered at points on the x-axis. By taking two of these circles centered at the same point, the tangent vectors at the apex of the circles are the same vector and therefore form an angle of $0°$. If we move one of the circles continuously to the right, the angle formed by the tangent vectors changes continuously as it approaches $180°$, which it would achieve once the circles intersected on the x-axis, except that that point is not in $\mathbb{H}^2$. Then by the Intermediate Value Theorem, every angle $0° \leq \theta < 180°$ can be achieved at the intersection of the two circles. Then by dilations and translations, we can get these semi-circles to intersect at any point, at any angle.

We can conclude, therefore, that by composing two reflections about circles, we can obtain a rotation by any degree θ in the tangent space of the intersection point p. Since this map is the composition of two isometries, the so-called rotation acting on the whole space is an isometry as well. However, we would like to find some way to make sense of the notion of a point far away from p being rotated around p by some angle θ. In other words, we want to understand what a rotation does to points in $\mathbb{H}^2$, not just to the tangent space. Before we can attack this problem, we need to better understand geodesics in $\mathbb{H}^2$.

Lemma 1.12. *There exists a unique geodesic between any two points $a, b \in \mathbb{H}^2$.*

Proof. Let a and b be two points in $\mathbb{H}^2$. If $a = b$, then the constant path $\gamma(t) = a$ is a geodesic. If $a \neq b$ but a and b are on the same vertical line, then by Lemma 1.5 there exists a unique geodesic between the points. So suppose a and b are not on the same vertical line. WLOG we may assume $a = (1, 0)$. Then we can rotate b to the y-axis in the following manner: First, perform the R isometry. Next, perform the reflection about the vertical line $(0, y)$. This time, the image of b is flipped to the opposite side of the line $(0, y)$. Call the composition of these two reflections as ψ. Now, ψ is a rotation about the point $(1, 0)$. So there is a continuous family of rotations about the point $(1, 0)$ from this isometry back to the rotation by $0°$, as we argued above. So for some rotation ϕ, $\phi(b)$ lies on the same vertical line as $a = \phi(a)$. Then we know there exists a unique geodesic between $\phi(b)$ and $\phi(a)$ by Lemma 1.5. Call this geodesic as γ. Now, rotations are arclength preserving isometries, as we showed in the proof of Thm 1.10 and the subsequent comments. So ϕ^{-1} preserves arclength, and so $L(\gamma) = L(\phi^{-1}(\gamma))$. But ϕ^{-1} is also an isometry, and so $d(\phi(a), \phi(b)) = d(a, b)$. So $\phi^{-1}(\gamma)$ realizes the distance between a and b, that is, $\phi^{-1}(\gamma)$ is a geodesic. In fact, it is a unique geodesic, because γ is unique, and isometries are injective. So there exists a unique geodesic from a to b. $\square$

Note that just because Isometries preserve distance, does not immediately imply that they map the geodesic between two points to the geodesic between the images of those points. We need the following lemma to show that geodesics do, in fact, map to other geodesics under isometries.

Lemma 1.13. *Isometries preserve geodesics.*

Proof. Let a, b be two different points in $\mathbb{H}^2$ and let ϕ be an arbitrary isometry.

Now, we already know that translations, dilations, reflections, and rotations preserve geodesics. Furthermore, as we showed in L1.12, there exists a rotation f that rotates a and b to be on the same vertical line. Let $f(a) = c$ and $f(b) = d$. Now consider $\phi(a)$, $\phi(b)$. By the composition of a rotation, translation, and dilation, we can map $\phi(a)$ to c, and $\phi(b)$ to the same vertical line as c (by L1.9 and L1.12). Now, there are two points along the vertical line at distance $d(c, d)$ from c. One is d itself, the other point we will call e. Now, if $\phi(b)$ maps to e, let $c = (c_1, c_2)$ and reflect across the circle of radius c_2 centered at the point $(c_1, 0)$. The vertical line (c_1, y) is invariant under this reflection, and so e maps to d. Let g be this composition of a rotation, translation, dilation, and also the reflection if needed, such that $g(\phi(a)) = c$ and $f(\phi(b) = d$. Then we have that $\psi = g \circ \phi \circ f^{-1}$ satisfies $\psi(c) = c$ and $\psi(d) = d$. Now, we know that g and f^{-1} are geodesic-preserving isometries. So if we can show that ψ preserves the geodesic between c and d, then it follows that ϕ must preserve the geodesic between a and b. So we must show that an isometry that fixes two points also fixes the geodesic between them.

Let γ be the geodesic between c and d. Recall that γ is the vertical line segment between those two points. Let ξ be the image of γ under ψ. Now, every point on ξ is the image of some point on γ, and since ψ is an isometry, the distance between any point on ξ and the points c and d are preserved.

If p is an arbitrary point on ξ, then p has some preimage $q \in \gamma$, and q is some distance r from c and distance $d(c, d) - r$ from d. Then p also is distance r from c and distance $d(c, d) - r$ from d. Assume for a moment that $p \notin \gamma$. Let μ_1, μ_2 be the geodesics from c to p and from p to d, respectively. Now, the concatenation of μ_1 and μ_2 has arclength $d(c, d)$, but it may not be differentiable at p. However, we can find a differentiable path that approximates the concatenation of μ_1 and μ_2 and has arclength arbitrarily close to $d(c, d)$. Furthermore, we can approximate this path with another path that is differentiable and has arclength exactly $d(c, d)$, but still has points not in γ. Since γ is supposed to be the unique geodesic between c and d, this is a contradiction. It must be, therefore, that $p = q$. Since p was an arbitrary point, we can conclude that $\xi = \gamma$ up to parameterization. $\square$

For the next lemma, we introduce the idea of an **infinite geodesic**. An infinite geodesic is defined to be a curve in the plane such that for every two points on the curve, the segment of the curve between them is the shortest path, or geodesic, between them.

Lemma 1.14. *A point and a direction uniquely determine an infinite geodesic.*

Proof. Let $p \in \mathbb{H}^2$ and v be a tangent vector based at p. We claim there is a unique infinite geodesic passing through p to which v is tangent. Let us make an educated guess. Let γ be the circle centered on the x-axis such that p lies on the circle and

v is tangent to the circle at p. (If v is vertical, then we define γ to be the vertical line through p, which we may loosely think of a circle centered at infinity along the x-axis.) Then γ is indeed an infinite geodesic: Take two points, a and b on γ. Then, there exists a reflection about γ that fixes a and b. But we showed in Lemma 1.13 that if an isometry fixes two points, it also fixes the geodesic between them. But the only path from a to b that is fixed by our reflection is the segment of γ between a and b. So this must be the geodesic between them. Note that this implies that all infinite geodesics are circles with centers on the x-axis or vertical lines. $\square$

Now we are ready to return to the question of how to think of a rotation outside the tangent space. Let ϕ be a rotation by θ degrees about the point p. Let q be an arbitrary point that is not p. Then there exists a unique geodesic, γ from p to q. Consider the tangent vector τ to the geodesic γ at point p. Now, ϕ rotates the tangent space to p by θ degrees, and since τ, if scaled down, is in the tangent space, τ is also rotated θ degrees. Now, there is only one geodesic going through p in the direction of $\phi(\tau)$ (L1.14). Since ϕ preserves the distance between p and q, $\phi(q)$ is the point at distance $d(p, q)$ along that geodesic, in the direction of $\phi(\tau)$. In this sense, we can say that q is "rotated" about p.

2. Classification of the Isometries of $\mathbb{H}^2$

We are now ready to classify the isometries of the hyperbolic plane. As it turns out, we have already found all of them, as the following theorem shows.

Theorem 2.1. *If $\phi : \mathbb{H}^2 \to \mathbb{H}^2$ is an isometry, then ϕ is a translation, dilation, reflection, or rotation, or some composition of these.*

Proof. Case I) Suppose ϕ has at least two fixed points, p, q. Then the geodesic γ between p and q is fixed. That follows from the proof of L1.13. Now, rotate these two points til they are on a vertical line. Then any point on the vertical line is fixed: For any point b not between p and q, consider the geodesic from b and the midpoint $\frac{p+q}{2}$. Our isometry fixes the part of the geodesic that is the segment between $\frac{p+q}{2}$ and one of the endpoints, so it must fix the rest of the geodesic as well, because as we noted before (proof of L1.14), geodesics are always segments of circles or vertical lines, and not some concatenation of the two. So the vertical line is fixed. Let's call this line L.

Now consider some c not on L. We want to determine what the image of c under ϕ might be. Now, there exists some unique point $m \in L$ such that m is the closest point on the line to c. This is true because on any closed interval $\varepsilon \leq y \leq n$, the function $d((x, y), c)$ has a minimum. Since as $y \to 0$ or $y \to \infty$, the distance from (x, y) to c approaches infinity, a global minimum must occur on some such interval. Furthermore, it can be shown, although we will not in this paper, that the minimum $m = (x, y)$ is unique.

Now, for a given $m = (m_1, m_2) \in L$ and a distance d, what points are such that they are distance d from m and m is the closest point on L to them? Consider the circle C defined by the equation

$$(x - m_1)^2 + y^2 = m_2^2.$$

This is the circle centered at $(m_1, 0)$ with radius m_2. Note that it is an infinite geodesic. Let w, v be the points distance d away from m lying on C, one on each side of the line L.

Now let f be the reflection across the circle C. Note that f fixes C, and in particular, the points m, v, and w and the geodesics between them. Now if the closest point to w on L was some point $n \neq m$, then $f(n) \neq n$ since f only fixes m out of all the points on L. However, since f is an isometry, $f(n)$ would also be the closest point to w. This is a contradiction since the closest point is unique. The same holds for the point v. So m is the closest point on L to w and to v.

If we let g be the reflection across the line L, then C is invariant under g. This is true because, considering the equation $(x - m_1)^2 + y^2 = m_2^2$, if $(x + m_1, y) \in C$, then $(m_1 - x, y) \in C$ as well. And since g is an isometry, $d(g(w), m) = d(w, m)$, so $g(w)$ is exactly the point at distance d along the geodesic C on the opposite side of L from, which is just v. So w and v are reflections of each other across the line L.

Now let us return to our consideration of ϕ. Let c be an arbitrary point not on the line L fixed by ϕ. What might $\phi(c)$ be? We know that there exists some closest point $m \in L$ to c. But ϕ is an isometry, and so the distances between c and every point on L are preserved. In particular, if m is the closest point to c, then it is also the closest point to $\phi(c)$. But as we showed above, there are only two points distance $d(c, m)$ away from m such that m is closest to both of them. So $\phi(c)$ is either c itself or the reflection of c across L, which we will denote as c'.

Now consider two arbitrary points c, $e \notin L$. I claim that if c and e are on the same side of L, then so are $\phi(c), \phi(e)$, and analogously, if c and e are on opposite sides of L, then so are $\phi(c)$ and $\phi(e)$. We will break this into cases.

Case 1) Suppose c and e are on the same side of L. Let γ be the geodesic between them. Then I claim that γ does not intersect L.

Suppose that it did. Recall that geodesics are vertical lines or circles centered on the x-axis. If γ is itself a vertical line, then it will not intersect L at all, as it will be parallel with L. If γ is a circle, then it intersects L exactly once, because the upper half of a circle is a well-defined function over its domain and therefore every element in the domain has exactly one element as its image (that is, it passes the vertical line test). So if γ intersected L, this would mean that c and e were on opposite sides of L, since the circle is continuous and only crosses L once. But c and e are on the same side of the line, so γ does not intersect L.

Therefore, $\phi(\gamma)$ won't intersect L either. Each point on L is fixed, so its preimage is itself, and no point of γ is in the preimage of L. However, if $\phi(\gamma)$ does not intersect L, then $\phi(c)$ and $\phi(e)$ must lie on the same side of L. This is because L separates $\mathbb{H}^2$ and $\phi(\gamma)$ is a continuous path.

Case 2) If c and e are on different sides of L, a similar argument shows that $\phi(c)$ and $\phi(e)$ must also lie on different sides of L.

Therefore, if for some $c \notin L$, $\phi(c) = c$, then for any other point e, $\phi(e) = e$, because this is the only possible image of e that remains on the same (or opposite) side of L as c, if e is on the same (or opposite) of L as c. Similarly, if for some $c \notin L$, $\phi(c)$ is the reflection of c across L, then for any other point e, the same holds. Therefore, ϕ is either the identity or the reflection across the line L.

Case II) Suppose that ϕ is an isometry with exactly one fixed point, p. Consider the tangent space of p. Considering the tangent space as $\mathbb{R}^2$, ϕ acts on it fixing the origin. Then if we think of the tangent space as isometric to $\mathbb{R}^2$, the type of isometry that only fixes one point is a rotation. So ϕ is a rotation of the tangent space, and therefore, as we showed before, a rotation of the whole plane.

Case III) Suppose that ϕ is an isometry with no fixed points. Then let $x \in \mathbb{H}^2$ be an arbitrary point. Then x has a unique preimage, $y = \phi^{-1}(x)$. By L1.9, there exists a translation and dilation mapping x to y. Let this translation composed with a dilation be ψ. Then $\psi \circ \phi$ is an isometry with at least one fixed point. Then this case reduces to either Case I or II. Then ϕ is a rotation, reflection, or the identity, composed with a translation and dilation. $\square$

3. The Hyperbolic Plane and $\mathrm{PSL}_2(\mathbb{R})$

$\mathrm{SL}_2(\mathbb{R})$, the group of 2×2 matrices over $\mathbb{R}$ with determinant $+1$, turn out to have a correspondence with the orientation-preserving isometries of $\mathbb{H}^2$.

If we consider the complex numbers as ordered pairs $z = (x, y)$, where $x = Re(z)$, $y = Im(z)$, then we can describe elements of the hyperbolic plane as elements $z \in \mathbb{C}$ such that $Im(z) > 0$. Then we define an action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{C}$ as follows:

For a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and an element $z \in \mathbb{C}$ with $Im(z) > 0$, we say that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d}.$$

Note that since $\det(A) = 1$, either $c \neq 0$ or $d \neq 0$, so the denominator is nonzero.

Note also, that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d} = \frac{-az-b}{-cz-d} = \begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix} \cdot z = (-1) \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z.$$

So for the action we defined, if $A \in \mathrm{SL}_2(\mathbb{R})$, then $A \cdot z = (-1)A \cdot z$. Since we do not want to distinguish between two matrices whose actions on $\mathbb{C}$ are the same, we define the following equivalence relation: $A \sim B$ iff $A = \pm B$. We denote $\mathrm{SL}_2(\mathbb{R})/\sim$ by $\mathrm{PSL}_2(\mathbb{R})$.

Let us consider three types of matrices:

(1) $\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \cdot z = \frac{z+s}{1} = z + s = (Re(z) + s, Im(z))$. This corresponds to a translation, T_s in $\mathbb{H}^2$.

(2) $\begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix} \cdot z = \frac{\lambda z}{1/\lambda} = \lambda^2 z$. This corresponds to a dilation D_{λ^2} in $\mathbb{H}^2$. Note that since $\lambda^2 > 0$, this dilation is well-defined.

(3) $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot z = \frac{1}{-z}$. This corresponds to the reflection R composed with the reflection about the y-axis, which is a rotation and therefore orientation preserving, as in fact are all three of the above matrices.

Theorem 3.1. $\mathrm{PSL}_2(\mathbb{R})$ *is generated by the matrices* $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $\begin{pmatrix} \lambda & 0 \\ 0 & \frac{1}{\lambda} \end{pmatrix}$, *and* $\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$.

Proof. We want to show that for an arbitrary matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{R})$, we can reduce it to the identity matrix only by multiplying by the above three matrix types. Consider

$$\begin{aligned}
&\begin{pmatrix} \frac{1}{c} & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & -\frac{c}{a} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -\frac{a}{c} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{c} & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & -\frac{c}{a} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & b - \frac{ad}{c} \\ c & d \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{c} & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} 1 & -\frac{c}{a} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} c & d \\ 0 & \frac{ad}{c} - b \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{c} & 0 \\ 0 & c \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & \frac{ad}{c} - b \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 \\ 0 & ad - bc \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
\end{aligned}$$

because $ad - bc = 1$, as the determinant of our matrix is 1. Note that each matrix type was used. □

In terms of the hyperbolic plane, this means that every matrix in $\mathrm{PSL}_2(\mathbb{R})$ can be decomposed into a composition of orientation-preserving isometries, and therefore that every element of $\mathrm{PSL}_2(\mathbb{R})$ is itself an orientation-preserving isometry of $\mathbb{H}^2$. And by our classification of isometries, we know that every orientation-preserving isometry of $\mathbb{H}^2$ is some composition of translations, dilations, and rotations. We already know how to represent translations and dilations as matricies, so let us show that there is a matrix representation for any rotation, as well, and then we will know that there is a bijective correspondence between $\mathbb{H}^2$ and $\mathrm{PSL}_2(\mathbb{R})$.

Under the action defined above, the reflection R corresponds to the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Now, this matrix is not orientation preserving, and it has determinant -1. However, if we recall, in section 1, we obtained arbitrary rotations by composing the reflection R with dilations and translations to move it around or make it larger in diameter, and then composed it with another similarly altered reflection R. So to get an arbitrary rotation, we take some product of translation and dilation matrices, along with two of these R matrices, arranged in some order. But the determinant of the product of matrices is the product of the determinants. And since the dilations and translations have determinant 1, and the two R matrices have determinant -1, the determinant of the product is 1, which means that the product is in $\mathrm{SL}_2(\mathbb{R})$. Therefore every rotation can be represented as a matrix in $\mathrm{PSL}_2(\mathbb{R})$.

It would be convenient if just by looking at the matrix we could tell what kind of isometry it was. As it turns out, this can be achieved simply by considering the absolute value of the trace of the matrix.

Before, we classified isometries by looking at fixed points. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$. Then z is a fixed point iff $\frac{az+b}{cz+d} = z$. This is equivalent to the equation

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

By the Fundamental Theorem of Algebra we know that every polynomial has a root over the complex number. Since roots represent fixed points, we are interested in where these roots are. Remember that the upper half plane is only those elements of $\mathbb{C}$ with positive imaginary part, so the fixed points may not be in the hyperbolic plane at all.

Let us solve this equation:

$$z = \frac{(a-d) \pm \sqrt{(d-a)^2 + 4bc}}{2c}.$$

Whether s is real or not depends on whether $(d-a)^2 + 4bc$ is nonnegative. So we have

$$\begin{gathered}(d-a)^2 + 4bc \geq 0 \implies d^2 + 2ad + a^2 + 4bc \geq 0 \implies \\ d^2 + a^2 + 2ad - 2ad - 2ad + 4bc \geq 0 \implies \\ (d+a)^2 - (4ad - 4bc) \geq 0 \implies (d+a)^2 - 4 \geq 0.\end{gathered}$$

This means that we have real roots if and only if

$$|d+a| = \operatorname{tr}(A) \geq 2.$$

However, note also that the characteristic polynomial of A is as follows:

$$\begin{aligned}f_A(t) = (t-a)(t-d) - bc &= t^2 + (a+d)t + (ad - bc) \\ &= t^2 + (a+d)t + 1\end{aligned}$$

which is zero when

$$t = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4}}{2}.$$

This equation also has real solutions iff

$$(a+d)^2 - 4 \geq 0 \implies |d+a| = \operatorname{tr}(A) \geq 2.$$

In other words, the matrix A corresponds to an isometry of $\mathbb{H}^2$ with fixed points with real solutions exactly when A has real eigenvalues.

Let us break this up into cases.

Case 1) $\operatorname{tr}(A) > 2$; that is, we have two real eigenvalues, or two real fixed points. Call these eigenvalues λ_1, λ_2. Since we are thinking about the hyperbolic plane, these real solutions lie on the real axis, which is not in the hyperbolic plane. Now, A is a real matrix. If we have two real eigenvalues, then there exist two distinct eigenvectors (in $\mathbb{R}^2$), one for each eigevalue. Then these vectors are linearly independent, and therefore can be a basis for $\mathbb{R}^2$. Then if we think of A as a linear transformation, it is similar to the same linear transformation under a basis change to the eigenbasis [2]. That is, for some matrix D, there exists an invertible matrix S such that $S^{-1}AS = D$. And D will be a diagonal matrix, since each basis vector gets sent to a scalar multiple of itself under this transformation. Now, similar matrices have the same trace and determinant [1]. So $D \in \mathrm{PSL}_2(\mathbb{R})$. However, not only is D diagonal, but since it has determinant 1, the two entries along the diagonal must be inverses of each other, since their product must be 1. So in fact, $D = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$. Therefore, if $\operatorname{tr}(A) > 2$, A is similar to a dilation matrix.

Case 2) $\operatorname{tr}(A) = 2$; i.e., there is exactly one fixed real point, or one distinct real eigenvalue. Call this eigenvalue as λ. In fact, if we look at the characteristic polynomial, the eigenvalue is $\frac{a+d}{2} = 1$ when $\operatorname{tr}(A) = 2$. So there exists some vector

v_1 such that $Av_1 = v_1$. Then if we extend v to a basis, v_1, v_2, then A is similar to some matrix of the form $\begin{pmatrix} 1 & ? \\ 0 & ?' \end{pmatrix}$. But $?'$ must be 1 as well because the matrix has determinant 1 (since it is similar to A). So in fact, $A \sim \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ for some $s \in \mathbb{R}$. That is, A is similar to a translation matrix.

Case 3) $\mathrm{tr}(A) < 2$; i.e., there are two complex eigenvalues, or two complex fixed points. However, if we look at the characteristic polynomial $t^2 + (a+d)t + 1$, it is clear that whatever the complex roots are, their sum must be $(a+d)$ which is a real number. Therefore, the two complex roots must be conjugate, that is, of the form $\alpha + i\beta$, $\alpha - i\beta$, where $\alpha, \beta \in \mathbb{R}$ and $\beta > 0$. Then $\alpha + i\beta$ is in the hyperbolic plane since its imaginary part is positive. So we have exactly one fixed point in the plane. Then by the classification of the isometries of the hyperbolic plane, A acting on the hyperbolic plane must be a rotation.

So now we are capable of looking at a matrix in $\mathrm{PSL}_2(\mathbb{R})$ and associating it with an orientation-preserving isometry of the hyperbolic plane simply by calculating its trace.

This shows that any matrix A in $\mathrm{PSL}_2(\mathbb{R})$ is similar to a translation, dilation, or the rotation matrix. However, we can conclude more than that. Let $A = S^{-1}XS$, where X is one of the three types of matrices in the above cases, and S is an invertible 2×2 matrix. Then since $S^{-1}S = \mathrm{I}$, $\det(S) = 1/\det(S^{-1})$. Now let $d = \sqrt{\det(S)}$, and let $S = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$. Then we can multiply S by a scalar as follows:

$$(1/d)S = \begin{pmatrix} p/d & q/d \\ r/d & s/d \end{pmatrix}.$$

The determinant of this matrix is $(1/\det(S))(ps - rq) = 1$. Therefore, we have

$$A = dS^{-1}XS(1/d).$$

So if we call our new matrix $Q = S(1/d)$, we have $A = Q^{-1}XQ$ and $Q \in \mathrm{PSL}_2(\mathbb{R})$. This means that not only is A similar to X, but also conjugate to it by another isometry in $\mathrm{PSL}_2(\mathbb{R})$. For example, if X is a dilation, we know that A, as an isometry, is conjugate to a dilation. In particular, since conjugation does not change the essential characteristics of an isometry – that is, it does not alter the angle of rotation nor the scale of a dilation or translation – we can conclude that every isometry in $\mathrm{PSL}_2(\mathbb{R})$ looks like a translation, a dilation, or a rotation.

This concludes our study of the hyperbolic plane and its isometries.

Acknowledgments. It is my pleasure to thank my mentors Tom Church and Katie Mann for helping me to learn this material.

References

[1] Laszlo Babai Apprentice Course, Lecture Notes, 2009, p 4.
[2] Laszlo Babai Apprentice Course, Lecture Notes, 2009, pp 9-10.