CRASH COURSE IN BASIC HYPERBOLIC GEOMETRY
GEOMETRIC LITERACY

These notes are intended to be elementary enough to be undergraduate-readable, requiring as little background as possible. Comments welcome!

I am updating this continually in Sept 2013. This version last updated September 5, 2013.

1. $\mathbb{H}$, its topology, its boundary, its action by $SL_2(\mathbb{R})$

Definition 1. The **Riemann sphere** $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is the topological sphere that you get from stereographically projecting the unit sphere $S^2 \subset \mathbb{R}^3$ to the xy-plane and adding the north pole as a point at infinity. An explicit map for stereographic projection is $(x, y, z) \mapsto \left(\frac{x}{1-z}, \frac{y}{1-z}\right)$. The **upper half-plane** is

$$\mathbb{H} = \{z = x + iy \in \mathbb{C} \mid y > 0\} \subset \hat{\mathbb{C}}.$$

Definition 2. For any field F, the **general linear group** $GL_2(F)$ is defined to be two-by-two matrices with entries in F that have non-zero determinant. The **special linear group** $SL_2(F)$ is the subgroup of $GL_2(F)$ where the matrices are required to have determinant one.

Definition 3. The **Möbius transformations** are the maps $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ of the form $f(z) = \frac{az+b}{cz+d}$ for $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$. They are extended to ∞ by continuity (so $f(\infty) = a/c$).

Definition 4. In $\mathbb{C}$, an **open ball** is the set of points within a certain distance of a fixed point:

$$B_r(p) := \{q \in \mathbb{C} \mid d(p, q) < r\}.$$

In $\hat{\mathbb{C}}$, an open ball centered at infinity is declared to be a set

$$B_r(\infty) := \{q \in \mathbb{C} \mid d(q, 0) > R\} \cup \{\infty\},$$

where R is related to r by stereographic projection: $R = \sqrt{\frac{1+\sqrt{1-r^2}}{1-\sqrt{1-r^2}}}$.

A set S in the Riemann sphere $\hat{\mathbb{C}}$ is **open** if at every point, a sufficiently small open ball around the point is contained in S:

$$\forall p \in S,\ \exists \epsilon > 0 \text{ such that } B_\epsilon(p) \subset S.$$

For a map $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, we say it is **continuous** if for all open balls $U \subset \hat{\mathbb{C}}$, the set $f^{-1}(U)$ is open as well.

A **homeomorphism** of $\hat{\mathbb{C}}$ is a map $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ that is bijective, continuous, and with continuous inverse. To check that f is a homeomorphism, it suffices to find its inverse function $f^{-1} : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ and check that both f and f^{-1} map open sets to open sets. For this, it suffices to check that both f and f^{-1} map open balls to open sets.

A **homothety** is multiplication by a real scalar: $z \mapsto kz$ for $k \in \mathbb{R}$.

An **affine map** is a map of the form $f(z) = az + b$ for $a, b \in \mathbb{C}$.

Notice that the Möbius transformations are a group action of $GL_2(\mathbb{C})$ on $\hat{\mathbb{C}}$, via $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right).z = \frac{az+b}{cz+d}$.

Exercise 1. Check that composition of Möbius transformations is encoded by matrix multiplication in the sense that for two matrices $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ and the corresponding Möbius transformations f and g, we have $f(g(z)) = A.(B.z) = (AB).z$.

Also check that two matrices $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$ correspond to the same Möbius transformation if and only if they are multiples of each other. In particular, the only matrices that induce the identity transformation are $\left(\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda \end{smallmatrix}\right) = \lambda I$.

Recall that a **group representation** is a homomorphism $\rho : G \to GL(V)$ for some vector space V. The previous exercise shows that we have a representation of the Möbius group into $GL_2(\mathbb{C}) = GL(\mathbb{C}^2)$, but that it is not faithful (injective). Let us define the projective linear groups as the quotients

$$PGL_2(\mathbb{C}) = {}^{GL_2(\mathbb{C})}\big/_{A \sim \lambda A} = {}^{GL_2(\mathbb{C})}\big/_{\mathbb{C}^\times}$$

and similarly

$$PSL_2(\mathbb{R}) = {}^{SL_2(\mathbb{R})}\big/_{A \sim kA} = {}^{SL_2(\mathbb{R})}\big/_{\pm}.$$

Note that because we've quotiented by the kernel of the representation, we can identify the Möbius group with $PGL_2(\mathbb{C})$ exactly.

Also notice that from our definition of $\hat{\mathbb{C}}$, we can define circles in $\hat{\mathbb{C}}$ to be images under stereographic projection of circles on the sphere. That means they are either (a) circles in $\mathbb{C}$ or (b) $L \cup \{\infty\}$, for L a line in $\mathbb{C}$.

Here are some facts about Möbius transformations and related geometry. When we write $A.S$ for a set $S \subset \hat{\mathbb{C}}$, we mean apply A to the whole set: $A.S := \{A.z \mid z \in S\}$.

(1) $PSL_2(\mathbb{R})$ also acts on $\hat{\mathbb{C}}$ by fractional linear transformations, and this preserves $\mathbb{H}$.

    Exercise 2. Check that $\mathbb{H}$ is preserved by $PSL_2(\mathbb{R})$.

(2) The closure of $\mathbb{H}$ in $\hat{\mathbb{C}}$ includes the real line and the point at infinity, which together form a circle in $\hat{\mathbb{C}}$. This is the "boundary at infinity" of $\mathbb{H}$.

(3) The group of all Möbius transformations is generated by the affine maps plus $J(z) = -1/z$. $J(z)$ is called **inversion in the unit circle**; it maps the unit circle to itself, and swaps the inside and the outside.

(4) $PGL_2(\mathbb{C})$ acts triply-transitively on $\hat{\mathbb{C}}$: for all $x, y, z \in \hat{\mathbb{C}}$ that are distinct from each other and $x', y', z' \in \hat{\mathbb{C}}$ that are distinct from each other, there is some $A \in PGL_2(\mathbb{C})$ such that $A.x = x'$, $A.y = y'$, $A.z = z'$. In fact, there is exactly one Möbius transformation that carries (x, y, z) to (x', y', z').

(5) Three distinct points in $\hat{\mathbb{C}}$ uniquely determine a circle.

(6) Möbius transformations are homeomorphisms of $\hat{\mathbb{C}}$.

(7) Möbius transformations map circles to circles: if C is a circle and $A \in PGL_2(\mathbb{C})$, then $A.C$ is also a circle.

(8) Möbius transformations preserve angle: if two curves C and C' have tangent lines L and L' meeting at angle θ, then $A.C$ and $A.C'$ also have tangent lines meeting at angle θ.

Proof of (3). Calculation! If $c = 0$, then the transformation is affine and we are done. If not, say $m(z) = \frac{az+b}{cz+d}$. But then

$$m(z) = \frac{(az+b)c}{(cz+d)c} = \frac{acz+bc}{c^2z+cd} = \frac{acz+ad-(ad-bc)}{c^2z+cd} = \frac{a}{c} - \frac{ad-bc}{c^2z+cd}.$$

But then if $f(z) = c^2z + cd$ and $J(z) = -1/z$ and $h(z) = (ad-bc)z + \frac{a}{c}$, we finally get $m(z) = h \circ J \circ f(z)$. □

Proof of (4). First we'll show that we can find a Möbius transformation sending $(0, 1, \infty)$ to any triple (z, w, q) of distinct points in $\hat{\mathbb{C}}$. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then $A.0 = b/d$, $A.1 = \frac{a+b}{c+d}$, and $A.\infty = a/c$. Now z and q can't both be ∞ since they are distinct. Assume without loss of generality that $q \neq \infty$. Then $a/c \neq \infty$, so I can let $c = 1$. Then it follows from the given equations that $a = q$, $d = \frac{q-w}{w-z}$, and $b = z \cdot \frac{q-w}{w-z}$. So I have found an A as desired. Why is this unique? One way to see it is that if I chose any other value for c in the process above, the whole matrix would have changed by a multiple, and as we saw above, this would induce the same Mobius transformation. Another way to see it is to suppose that two different matrices A and B both send $(0, 1, \infty)$ to (z, w, q). In that case, $A^{-1}B$ fixes $(0, 1, \infty)$. But then $a/c = \infty \implies c = 0$, and $b/d = 0 \implies b = 0$, and $\frac{a+b}{c+d} = 1 \implies a = c$, so $A^{-1}B = \left(\begin{smallmatrix} a & 0 \\ 0 & a \end{smallmatrix}\right)$ is a multiple of the identity matrix, which means A and B differ by a multiple.

So now we know that there's a unique Möbius transformation mapping $(0, 1, \infty)$ to (z, w, q), and therefore also mapping any (z, w, q) to $(0, 1, \infty)$. Composing two such maps lets us map any triple to any other triple. □

Proof of (5). If one of the three points is ∞ or if all three points are collinear, then the circle is the line through the points in the plane and we are done. So assume without loss of generality that z, w, q are non-collinear points in $\mathbb{C}$.

First consider z and w. Let L be the line of all points equidistant from z and w (that is, the perpendicular bisector of the segment between z and w). For a point p on L, let $f(p) = d(p, z) - d(p, q)$. It is clear that $f(p)$ is a real number that changes continuously as p is moved along L; it is also clear that $f(p)$ is sometimes positive and sometimes negative. By the intermediate value theorem, there must be some point p_0 on L such that $f(p_0) = 0$. But then $d(p_0, z) = d(p_0, q)$, and since it is on L we also have $d(p_0, z) = d(p_0, w)$. So p_0 is equidistant from all three points, and therefore a Euclidean circle centered at p_0 goes through all three points.

Three distinct points $z, w, q \in \hat{\mathbb{C}}$ can not lie on more than one circle, because distinct circles intersect in at most two points. □

Proof of (6). Möbius transformations are clearly bijections, since they have inverses that are also maps from $\hat{\mathbb{C}} \to \hat{\mathbb{C}}$. We need to see that Möbius transformations map open sets to open sets. It suffices to show that open balls are mapped to open sets by affine maps and by inversion in the circle. This takes four cases.

Case 1: Affine maps (compositions of rotation, homothety, and translation) clearly send open balls in $\mathbb{C}$ to other open balls in $\mathbb{C}$.

Case 2: Affine maps applied to $B_r(\infty)$. The ball is everything in $\mathbb{C}$ outside a Euclidean circle of radius R, plus the point at infinity. Affine maps preserve infinity: $f(\infty) = \infty$. They take Euclidean circles to Euclidean circles, so the image $f(B_r(\infty))$ is everything outside a new Euclidean circle plus the point at infinity. To see that this is an open set, the only tricky point is to check that there is some $B_\epsilon(\infty)$ contained in it. But $B_\epsilon(\infty)$ is everything outside of a certain circle, whose radius gets big as ϵ gets small. Simply choose ϵ so small that that circle completely contains the boundary of $f(B_r(\infty))$.

Case 3: J applied to open balls $B_r(p) \subset \mathbb{C}$.

If the open ball does not contain 0, then the image does not contain ∞ and J is clearly a bicontinuous map on $\mathbb{C}$ (by calculus).

If the open ball does contain 0, then there is some $B_s(0)$ inside $B_r(p)$ and it suffices to check that $J(B_s(0))$ is open. Now $J(0) = \infty$ and J clearly maps the circle of radius M about the origin to the circle of radius $1/M$ about the origin. Thus J maps $B_s(0)$ to the outside of some circle around the origin, plus the point at infinity. That is, $J(B_s(0)) = B_t(\infty)$ for an appropriate t.

Case 4: J applied to $B_r(\infty)$. By exactly similar reasoning, J maps $B_r(\infty)$ to $B_t(0)$ for an appropriate t.

This completes the proof that affine maps and $J(z)$ both take all open balls to open sets. □

Proof of (7). Affine maps $f(z) = az + b$ are the composition of a rotation, a homothety, and a translation. To be precise, if $a = re^{i\theta}$, then $f(z) = t_b \circ \rho_\theta \circ h(z)$, where $h(z) = rz$. This is the composition of three maps that evidently preserve circles, so $f(z)$ does as well.

Exercise 3. Recall that if $z = x + iy$, then its complex conjugate is $\overline{z} = x - iy$. By converting z to x and y coordinates, show that the set of solutions to the equation

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$$

is a circle for every $\alpha, \gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$. (It may be a point or the empty set if $\alpha\gamma \geq \beta\overline{\beta}$. It is a straight line when $\alpha = 0$.)

Exercise 4. Using the previous exercise, show that $J(z) = -1/z$ preserves circles. To do this, assume that z is on a circle (it is a solution to an equation as in the previous problem.) Now let $w = -1/z$ and find a similar equation satisfied by w.

This completes the proof of (7). □

Exercise 5. Check that $f(x + iy) = x + 2iy$, $(f(\infty) = \infty)$ is a homeomorphism of $\hat{\mathbb{C}}$ that does not preserve circles.

Exercise 6. Find the fixed points of $\left(\begin{smallmatrix} 2 & 5 \\ 3 & -1 \end{smallmatrix}\right)$. Where does it map the unit circle? Where does it map the real line?

$SL_2(\mathbb{Z})$ is the set of two-by-two matrices with integer entries and determinant one. It is a subgroup of $SL_2(\mathbb{R})$, and we let $PSL_2(\mathbb{Z})$ be the corresponding subgroup of $PSL_2(\mathbb{R})$.

2. The metric

$\mathbb{H}$ is the hyperbolic plane; our basic model of it is the upper half-plane in $\hat{\mathbb{C}}$. The classical way of turning this into a metric space is with the metric $ds = \frac{dz}{y}$. (Here, $dz = \sqrt{dx^2 + dy^2}$ is the Euclidean distance element.) In other words, this distance element gives $\mathbb{H}$ the structure of a length space (in fact of a Riemannian metric): the distance between two points is the infimum over paths γ connecting the points of the integral against ds over γ. The boundary of $\mathbb{H}$ in $\hat{\mathbb{C}}$ is $\partial\mathbb{H} = \mathbb{R} \cup \{\infty\}$, which is a circle, as we've seen. The action of $SL_2(\mathbb{R})$ not only preserves $\mathbb{H}$, but it acts by homeomorphisms and the action preserves $\partial\mathbb{H}$ as well.

Orthocircles are semi-circles in $\mathbb{H}$ (that is, pieces of either lines or Euclidean circles) which are perpendicular to the real axis.

Disk model. There is a second model of the hyperbolic plane which is sometimes very useful for visualizing things more symmetrically. Consider the Möbius transformation f that sends $0 \mapsto -i$, $i \mapsto 0$, $\infty \mapsto i$. Using techniques from the previous handout, we can solve and find that $f(z) = \frac{iz+1}{z+i}$. Since this transformation also maps $1 \mapsto 1$, it takes three points on the boundary $\partial\mathbb{H}$ (namely $0, 1, \infty$) to three points on the unit circle (namely $-i, 1, i$) and it takes i to the origin. We know that f maps circles to circles and is a homeomorphism, so it must take the upper half-plane to the unit disk (the inside of the unit circle). This gives us a second model of the hyperbolic plane, called the **disk model**. We can define the metric on the disk by "pushing forward" the metric on $\mathbb{H}$. That means that if you want to know how far apart two points in the disk are, you just take f^{-1} of the points and measure their distance in $\mathbb{H}$. Just for your information, this works out to a formula for the metric in the disk model that looks a bit different:

$$ds' = \frac{2\ dz}{1 - |z|^2}.$$

We will save the notation $\mathbb{H}$ for the upper half-plane and use $\mathbb{D}$ when we are referring to the disk model.

Here's an important remark: we can tell by inspection that horizontal translations are isometries of $(\mathbb{H}, ds)$, because ds depends only on dx, dy, and y, but not x. Similarly, we can tell by inspection that rotations ρ_θ about the origin are isometries of $(\mathbb{D}, ds')$, because ds' depends only on dz (preserved by rotation since it is a Euclidean isometry) and $|z|$ (distance from the origin is clearly preserved by rotation).

Subgroup structure. There are three special subgroups of $SL_2(\mathbb{R})$ that are worth understanding to know its structure:

$$K = \{\left(\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right)\}; \quad A = \{\left(\begin{smallmatrix} k & 0 \\ 0 & 1/k \end{smallmatrix}\right)\}; \quad N = \{\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)\}.$$

Recall that for rigid motions of the Euclidean plane, there is a structure theorem saying that every rigid motion of $\mathbb{R}^2$ is a product of a translation, a rotation, and possibly a reflection (Artin, page 159). There is a similar structure theorem here, called the **Iwasawa decomposition**. It says $SL_2(\mathbb{R}) = NAK = KAN$, so any fractional linear transformation can be written as a product of matrices of the three types above in an essentially unique way (the matrix $-I$ belongs to more than one of the subgroups, but this is a quibble).

In the upper half-plane model, you can think of K as being lopsided rotations around the point i (moving points along hyperbolic circles), A as being scaling maps $z \mapsto k^2 z$, and N as being horizontal translations $z \mapsto z + n$. In the disk model, the elements corresponding to K are true rotations around zero; the elements of A slide you away from a repelling point on the boundary and toward an attractor; and the elements of N move everything along **horocycle** orbits, or circles tangent to the boundary.

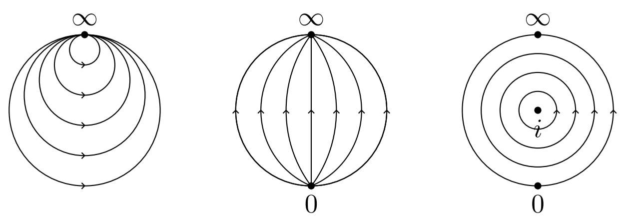

FIGURE 1. Actions on the unit disk by elements of N, A, and K respectively.

Metric properties. Now that we know all about Möbius transformations and have defined the hyperbolic plane as a metric space, we can study the metric properties of the $SL_2(\mathbb{R})$ action on $\mathbb{H}$.

(1) FLTs act by isometries. For instance, even the map $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$, which sends $z \mapsto 4z$, preserves distances in $\mathbb{H}$—surprising, but true!
(2) A circle in the hyperbolic sense is a circle in the Euclidean sense, but with a different center and radius. Note that this means that the topology induced on $\mathbb{H}$ by its metric matches with the usual Euclidean topology, since the set of open balls is the same.
(3) The geodesics in $\mathbb{H}$ are the orthocircles.
(4) Any two points in $\mathbb{H}$ are connected by a unique geodesic.
(5) If $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is translation and $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is inversion in the unit circle, then $PSL_2(\mathbb{Z}) = \langle T, J \rangle$.
(6) A fundamental domain for $PSL_2(\mathbb{Z})$ is the triangle bounded by the unit circle, $x = 1/2$, and $x = -1/2$. (It has one "ideal vertex," or vertex at infinity.)
(7) If a triangle (with geodesic segments as its sides) has angles α, β, γ, then its area is $\pi - \alpha - \beta - \gamma$.
(8) The circumference of a circle is an exponential function of its radius.

Proof of (1). Suppose $\gamma : [0, 1] \to \mathbb{H}$ is any rectifiable path in $\mathbb{H}$. We will show that the length of $T \circ \gamma$ is the same as the length of γ for any $T = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbb{R})$. Then since the distance between two points is the smallest length of a path between those points, and since fractional linear transformations preserve length, they must preserve distance as well.

Suppose that $\gamma(t) = (x(t), y(t))$ as usual, and let u, v be defined as the coordinates of the path $T \circ \gamma$. In other words, if $z = x + iy$, then let $w = u + vi = T.z = \frac{az+b}{cz+d}$.

Now some computations. By the quotient rule from calculus,

$$\frac{dw}{dz} = \frac{a(cz+d) - c(az+b)}{(cz+d)^2} = \frac{1}{(cz+d)^2}.$$

We have $w = \frac{az+b}{cz+d}$, and rationalizing the denominator gives

$$w = \frac{(az+b)(c\overline{z}+d)}{|cz+d|^2} = \frac{ac|z|^2 + adz + bc\overline{z} + bd}{|cz+d|^2}.$$

Just look at the imaginary part of both sides, and you get $v = \frac{(ad-bc)y}{|cz+d|^2} = \frac{y}{|cz+d|^2}$. Putting this together with the dw/dz calculation above, we get $\left|\frac{dw}{dz}\right| = \frac{v}{y}$. Finally,

$$\begin{aligned} \ell(T \circ \gamma) &= \int_0^1 \frac{\sqrt{(du/dt)^2 + (dv/dt)^2}}{v} dt = \int_0^1 \frac{|dw/dt|}{v} dt \\ &= \int_0^1 \frac{|dw/dz| \cdot |dz/dt|}{v} dt = \int_0^1 \frac{|dz/dt|}{y} dt = \ell(\gamma). \end{aligned}$$

□

Proof of (2). Here's a neat argument. We will consider C, the set of all points in $\mathbb{H}$ which are hyperbolic distance r from i. This is the hyperbolic circle of radius r centered at i. We will show that this is also a Euclidean circle. Since the metric on the disk model $\mathbb{D}$ is defined by pushing forward the metric on $\mathbb{H}$, we know that $f(C)$ is the set of points in $\mathbb{D}$ which are hyperbolic distance r away from the origin. Now consider rotation about the origin in $\mathbb{D}$. This is an isometry of $\mathbb{D}$, as remarked above, and it fixes the origin, so it must map a point at distance r from the origin to a point at distance r from the origin. Since $f(C)$ is the set of all such points, it is preserved by rotation, and it contains exactly one point on each radius of the unit disk, so it is a Euclidean circle centered at the origin. But f is a Möbius transformation, so we know it preserves circles. Thus C is also a Euclidean circle. It goes through the points $e^r i$ and $e^{-r} i$, so its Euclidean center must be $(e^r + e^{-r})i/2 = (\cosh r)i$ and its Euclidean radius must be $(e^r - e^{-r})/2 = \sinh r$.

Any other point can be mapped to i by an isometry, so all hyperbolic circles centered everywhere are also Euclidean circles. And finally, since we can obtain every circle of every radius in this way by changing r and moving the center, it also follows that every Euclidean circle is obtained in this way. □

Proof of (3). Because the metric ds on $\mathbb{H}$ is defined via lengths, that makes the hyperbolic plane automatically a length space. A geodesic is a path such that for any segment of it, the length equals the distance between the endpoints, and which is "paced right" (parametrized by arclength). We know that the imaginary axis is a geodesic from any ai to bi with parametrization $\gamma(t) = e^t i$: $\int_\gamma ds = \int_\gamma \frac{|dz|}{y} = \int_\gamma \frac{|dy|}{y} = \ln(b/a)$, and for any path we have the inequality $|dz| \geq |dy|$, so this is shortest-possible. All we need to see is that isometries (of any length space) map geodesics to geodesics, and we will know exactly what the rest of the geodesics in $\mathbb{H}$ are.

Suppose $\gamma : [a,b] \to X$ is an arbitrary geodesic segment in a length space X. Suppose $f : X \to X$ is an isometry of X. Then we want to see that $f \circ \gamma$ is a geodesic as well. Say the endpoints of γ are p and q. Then $\ell(\gamma) = d(p,q)$. But $f \circ \gamma$ is a path from $f(p)$ to $f(q)$. Because f preserves distance, we know that

$d(f(p), f(q)) = d(p,q)$. Similarly, we know that $\ell(f\circ\gamma) = \ell(\gamma)$ because every point along $f\circ\gamma$ is $f\circ\gamma(t)$ for some t, so

$$\ell(f\circ\gamma) = \sup \sum d(f\circ\gamma(t_i), f\circ\gamma(t_{i+1})) = \sup \sum d(\gamma(t_i), \gamma(t_{i+1})) = \ell(\gamma).$$

But this means we have found a path whose length equals the distance between its endpoints, so it is a geodesic, as desired. □

Proof of (4). Take two points $z, w \in \mathbb{H}$. If they have the same real part, then they are both on a vertical line and on no other orthocircle, so that case is easy. Suppose they have different real parts, say $z = x + yi$ and $w = u + vi$, $x < u$. Then consider the function g which measures the difference in Euclidean distances $g(q) = |q - z| - |q - w|$. This function varies continuously along the x-axis, gives real-number output, $g(q) < 0$ for $q \to -\infty$ while $g(q) > 0$ for $q \to \infty$. To see that $g(q) < 0$ has solutions, note that $|q-z|^2 = (q-x)^2+y^2$ while $|q-w|^2 = (q-u)^2+v^2$.

$$\begin{aligned} |q-z| < |q-w| \iff |q-z|^2 < |q-w|^2 \iff &\quad (q-x)^2 - (q-u)^2 < v^2 - y^2 \\ \iff &\quad (2q - x - u)(u - x) < v^2 - y^2, \end{aligned}$$

and this can clearly be accomplished by choosing q to be a sufficiently large negative number.

My favorite theorem, the Intermediate Value Theorem, now ensures that there is some point q on the x-axis such that $g(q) = 0$, which means that this q is equidistant from z and w, which means that both w and z lie on a circle centered at q.

It's not hard to see that g is strictly monotone increasing, so the solution is unique. □

Proof of (5). We want to show that $PSL_2(\mathbb{Z}) = \langle T, J\rangle$. Clearly, since T and J are in $G = PSL_2(\mathbb{Z})$, the second is a subgroup of the first, so we just need to show that any $M \in G$ can be built as a word in the letters T and J.

First, we consider which elements of G fix the point ∞. The matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ fixes ∞ if and only if $c = 0$. Since $ad - bc = 1$ in G, this means that $ad = 1$, so that $\left(\begin{smallmatrix} 1 & r \\ 0 & 1 \end{smallmatrix}\right)$ $(r \in \mathbb{Z})$ are the only possibilities, up to a minus sign. But $\left(\begin{smallmatrix} 1 & r \\ 0 & 1 \end{smallmatrix}\right) = T^r$, so all elements of G that fix ∞ are in $\langle T, J\rangle$.

Next, let's consider the G-orbit of the point ∞. For any $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G$, it maps ∞ to a/c, so the orbit is $G.\infty = \mathbb{Q} \cup \{\infty\}$, the rationals plus infinity itself.

We would like to show that by putting together a product of T's and J's, we can also do the same thing, mapping ∞ to any rational number. For example,

$$\begin{aligned} T^a J T^{-b} J T^c J.\infty = T^a J T^{-b} J T^c.0 = T^a J T^{-b} J.(c) = T^a J T^{-b}.(-1/c) = \\ = T^a J.(-b - 1/c) = T^a.\left(\frac{1}{b + 1/c}\right) = a + \frac{1}{b + \frac{1}{c}}. \end{aligned}$$

Continuing in this way, it is clear that we can build up any continued fraction expansion of finite length, so the orbit of ∞ under $\langle T, J\rangle$ is all of $\mathbb{Q} \cup \{\infty\}$.

Now consider an arbitrary $M \in G$. We need to show that $M \in \langle T, J\rangle$. Suppose $M.\infty = \infty$. Then we are done, because $M = T^r$. If not, then $M.\infty = p/q$ for some rational number p/q. Let $[a_0, a_1, \ldots, a_n]$ be the continued fraction expansion of p/q and let $A = T^{\pm a_0} J T^{\pm a_1} J \ldots T^{a_n}$ be the word constructed as above such that $A.\infty = p/q$. Then $A^{-1}M$ fixes ∞, so it is equal to T^r for some r, so $M = AT^r$ and we are done. □

Proof of (6). There are lots of nice ways to prove this. It's particularly useful to invoke the idea of a **Dirichlet region**. Let Γ be a discrete subgroup of $PSL_2(\mathbb{R})$ (so it acts by isometries on $\mathbb{H}$) and let p be any point in $\mathbb{H}$ that is not fixed by anything but the identity in Γ. The Dirichlet region is the points that are closer to p than to anything else in the orbit $\Gamma.p$. Note this makes sense when the orbits are discrete.

$$D_p(\Gamma) = \{z \in \mathbb{H} : d(z,p) \le d(z,gp) \forall g \in \Gamma\}.$$

I claim that this always gives a fundamental domain for Γ, in a particularly nice form: it is a connected region bounded by a geodesic polygon, it tiles the space ($\Gamma.D_p(\Gamma) = \mathbb{H}$) and its interior contains no "redundancy," meaning at most one point from each orbit ($z \in D_p(\Gamma)^\circ \implies gz \notin D_p(\Gamma)^\circ, g \neq e$). It's not hard to establish that these are true. To see that it is bounded by a polygon, note that for each element gp in the orbit of p, the locus of points equidistant from p and gp is a geodesic (it's the perpendicular bisector of the segment connecting p and gp). Thus the Dirichlet region is the intersection of (possibly infinitely many) half-spaces bounded by geodesics. That the region tiles the entire space is fairly clear: every point $w \in \mathbb{H}$ has *some* point in the orbit which is closest to it; say w is closest to gp. Then $g^{-1}w$ is closer to p than to anything else in the orbit, so $g^{-1}w \in D_p(\Gamma)$, so $w \in g.D_p(\Gamma)$. This shows that every point in $\mathbb{H}$ is in some image of the Dirichlet tile. Finally, why no redundancy on the interior? For contradiction, suppose that there were some z and $w = gz$ which were both strictly closer to p than to anything else in the orbit of p. This says that $d(gz,p) < d(gz,gp)$ and $d(z,p) < d(z,g^{-1}p)$. Now $d(gz,gp) = d(z,p)$ because g is an isometry, and likewise $d(z,g^{-1}p) = d(gz,p)$. Putting these together, we get $d(gz,p) < d(gz,gp) = d(z,p)$ and at the same time $d(z,p) < d(z,g^{-1}p) = d(gz,p)$, which is a contradiction. That finishes the claim.

So now we want to find a Dirichlet region for $\Gamma = SL_2(\mathbb{Z})$. We know it is generated by T and J. Let's let $p = 2i$ since that is not fixed by anything in Γ (check this). Now $Tp = 2i+1$ and $T^{-1}p = 2i-1$ and $Jp = i/2$, so the set of points which are closer to p than to these three other points in the orbit is the triangle bounded by $x = 1/2$, $x = -1/2$, and the unit circle. This tells us that the Dirichlet region must be contained in that triangle, but it might be smaller—we would have to know the whole orbit of p to know for sure. It remains to check, therefore, that there is no redundancy in the triangle: we will assume that $z = x+iy$ and $w = gz = \frac{az+b}{cz+d} = u+iv$ are both on the interior of the triangle, and find a contradiction. Here's a burst of cleverness to finish it off.

$$|cz+d|^2 = c^2|z|^2 + 2cdx + d^2 > c^2 + d^2 - |cd| = (|c| - |d|)^2 + |cd|,$$

where the inequality is valid because $|x| < 1/2$ and $|z| > 1$, which is what it means to be inside the triangle. But the quantity on the right-hand side is an integer since $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL_2(\mathbb{Z})$ and it must be positive because $c = d = 0$ can't happen. This means that $|cz+d|$ is strictly greater than 1, but recalling that $v = y/|cz+d|^2$, we finally conclude that $v < y$. But the exact same argument holds in reverse (exchanging the roles of z and w), so we also get $y < v$, a contradiction. □

Proof of (7). (Coming soon.)

Proof of (8). Calculation! It suffices to prove this for a circle of radius r centered (hyperbolically) at i. As we saw, this has Euclidean center $(\cosh r)i = ki$ and Euclidean radius $\sinh r = R$.

$$\gamma(t) = ((\sinh r)\cos t,\ (\sinh r)\sin t + \cosh r) = (R\cos t, R\sin t + k).$$

Then the circumference is $\int_0^{2\pi} \frac{R}{R\sin t + k} dt$ and careful computations will tell you in the end that this equals $2\pi R = 2\pi \sinh r = \pi(e^r - e^{-r})$. $\square$