from w to the center of C_i and Q maps that to the segment from Qw to the center of D_f. Thus, this segment is mapped into itself and so, as above, the attracting fixed point must lie in this segment. $\square$

The argument behind the proof of (b), which says fixed points must lie in $\bar{D}_f \cup \bar{D}_i$, also shows that if $|f(z) - z|$ is small, z must be close to $\bar{D}_f \cup \bar{D}_i$:

Theorem 9.2.33. *Let $f \in \mathfrak{F}$ with $f(\infty) \neq \infty$. Let D_i and D_f be the initial and final disks. Then either $z \in D_i$ or*

$$\operatorname{dist}(z, D_f) \leq |z - f(z)| \tag{9.2.70}$$

Remarks. 1. Since we will talk about another metric in the next section, we emphasize that $\operatorname{dist}(\,\cdot\,, D_f)$ is here in the Euclidean metric.

2. This implies

$$\operatorname{dist}(z, D_f \cup D_i) \leq |z - f(z)| \tag{9.2.71}$$

Proof. If $z \notin D_i$, then $f(z) \in \bar{D}_f$, so (9.2.70) holds. $\square$

By (9.2.60), we have

$$D_i = \{z \mid |f'(z)| > 1\} \tag{9.2.72}$$

and

$$\mathbb{C} \setminus \bar{D}_i = \{z \mid |f'(z)| < 1\} \tag{9.2.73}$$

Remarks and Historical Notes. Given how fundamental FLTs are to so many parts of mathematics, it is unfortunate how little they are discussed in basic texts (which, e.g., do not discuss the hyperbolic, parabolic, elliptic splitting), and that this discussion does not talk about projective space. The textbook description of the Riemann sphere is via stereographic projection—admittedly useful—but not as basic as the $\mathbb{P}$ point of view.

Most of the material in this section is classical (from the nineteenth century), although our discussion has some more modern elements. Key figures in these classical developments are Möbius, Schwarz, Klein, and especially Poincaré.

The use of isometric circles and the representation $f = QR$ for nonloxodromic transformations was emphasized especially by Ford; see, for example, [138].

If $\tilde{R} = QRQ$, the reflection in the isometric circle for f^{-1}, then $f^2 = QRQR = \tilde{R}R$, something that can easily be proven directly. It is simple in various ways to use geometric structures defined by f to get f^2 as a product of reflections. The neat thing about Ford's idea of using a perpendicular bisector is that it "takes the square root."

9.3 MÖBIUS TRANSFORMATIONS

In this section, we will discuss FLTs that take $\mathbb{D}$ onto $\mathbb{D}$ (equivalently, take $\mathbb{D}$ into $\mathbb{D}$ and $\partial\mathbb{D}$ to $\partial\mathbb{D}$). Of course, by Theorem 9.2.13, the FLTs, which are bijections of any disk or half-plane, are conjugate to bijections of the disk, so this section could also describe analytic bijections of, say, $\mathbb{C}_+$. That said, there are often good reasons to study $\mathbb{C}_+$ (as we will explain in the Notes). But we will need $\mathbb{D}$ later, so we study these maps in this guise.

An FLT, which is a bijection of $\mathbb{D}$, we will call a *Möbius transformation*. We use $\mathcal{M}$ for the family of Möbius transformations. This is nonstandard terminology since "Möbius transformation" is typically used as a synonym for FLT, but it is useful to have a standard term.

It will be very useful to have Möbius transformations that map any point in $\mathbb{D}$ to any other point. As usual, if we do it for a fixed endpoint, we can do it for any other, for if f_{z_0} takes z_0 to 0, then $f_{w_0}^{-1} f_{z_0}$ maps z_0 to w_0.

Proposition 9.3.1. *Let $z_0 \in \mathbb{D}$. Then*

$$f_{z_0}(z) = \frac{z - z_0}{1 - \bar{z}_0 z} \tag{9.3.1}$$

maps $\mathbb{D}$ onto $\mathbb{D}$ and has $f_{z_0}(z_0) = 0$.

Proof. f is analytic in $\{z \mid |z| < |z_0|^{-1}\}$ and so in a neighborhood of $\overline{\mathbb{D}}$. Moreover, $|f_{z_0}(e^{i\theta})| = |e^{i\theta} - z_0|/|e^{-i\theta} - \bar{z}_0| = 1$, so by the maximum principle, f maps $\mathbb{D}$ into $\mathbb{D}$. But by calculating, $f_{-z_0} \cdot f_{z_0} = 1$ since $\left(\begin{smallmatrix} 1 & -z_0 \\ -\bar{z}_0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & z_0 \\ \bar{z}_0 & 1 \end{smallmatrix}\right) = 1 - |z_0|^2 \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, so f is an analytic bijection of $\mathbb{D}$. Clearly, $f_{z_0}(z_0) = 0$. $\square$

The second main result that we will need to analyze all Möbius transformations is a general one about analytic bijections, which we do not know a priori are FLTs restricted to $\mathbb{D}$:

Theorem 9.3.2. *If $f\colon \mathbb{D} \to \mathbb{D}$ is an analytic bijection and $f(0) = 0$, then for some $\theta \in [0, 2\pi)$,*

$$f(z) = e^{i\theta} z \tag{9.3.2}$$

Proof. We begin with the Schwarz lemma (Proposition 2.3.4), which implies that $|f(z)| \le |z|$. But since f^{-1} also maps $\mathbb{D}$ to $\mathbb{D}$ and $f^{-1}(0) = 0$, we have that $|f^{-1}(z)| \le |z|$. Setting $w = f^{-1}(z)$, we see $|w| \le |f(w)|$, so $|f(z)/z| = 1$ on $\mathbb{D}$. By the maximum principle, $f(z)/z$ is constant. $\square$

Theorem 9.3.3. *If $f\colon \mathbb{D} \to \mathbb{D}$ is an analytic bijection, then f is a Möbius transformation. In fact, if $f(z_0) = 0$, then for some $\theta \in [0, 2\pi)$,*

$$f(z) = e^{i\theta} f_{z_0}(z) \tag{9.3.3}$$

where f_{z_0} is given by (9.3.1).

Proof. $f f_{z_0}^{-1}$ maps $\mathbb{D}$ onto $\mathbb{D}$ and takes 0 to 0, so this follows from Proposition 9.3.1 and Theorem 9.3.2. $\square$

The remarkable fact about this is that analytic bijections of $\mathbb{D}$ automatically have meromorphic continuations to all of $\mathbb{P}$. This is not quite as surprising as it might seem at first. If $|z_n| \to 1$, $f(z_n)$ cannot converge to a point, w_0, in $\mathbb{D}$ because $f(z)$ near w_0 means z must be near $f^{-1}(w_0)$, and so must have $|z|$ near $|f^{-1}(w_0)|$. Thus,

$|f(z)| \to 1$ as $|z| \to 1$. If we knew f had a continuous extension of $\overline{\mathbb{D}}$ to $\overline{\mathbb{D}}$, then we could extend f to $\mathbb{C} \cup \{\infty\}$ by

$$f(z) = \overline{f(1/\bar{z})}^{-1} \tag{9.3.4}$$

which is trivially meromorphic in $\mathbb{D} \cup \mathbb{C} \setminus \overline{\mathbb{D}}$ and analytic across $\partial\mathbb{D}$ by the reflection principle and the fact that $|f(e^{i\theta})| = 1$. There is a version of the Schwarz reflection principle that only requires that $\operatorname{Im} g$ vanishes. That can be applied to $i \log|f|$. In any event, we have (9.3.4) for any Möbius transformation.

In the last section, we saw that FLTs could be labeled by three complex variables, $f(0)$, $f(1)$, $f(\infty)$, so $\mathfrak{F}$ has real dimension 6. Here we saw that Möbius transformations are parametrized by one complex variable $z_0 = f^{-1}(0)$ and one real variable, so $\mathcal{M}$ is three-dimensional. Moreover, we see $\mathcal{M}$ topologically is $\mathbb{D} \times \partial\mathbb{D}$.

By Theorem 9.2.16, any $f \in \mathcal{M}$ is nonloxodromic. $f(z) = e^{i\theta} z$ is elliptic (it is f_T for $T = \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}$ has $\det(T) = 1$ and $\operatorname{Tr}(T) \in [-2, 2]$). f_{z_0} is hyperbolic since it is f_T for $T = (1 - |z_0|^2)^{-1/2} \begin{pmatrix} 1 & -z_0 \\ -\bar{z}_0 & 1 \end{pmatrix}$ has $\det(T) = 1$ and $\operatorname{Tr}(T) = 2/(1 - |z|^2)^{1/2} > 2$. The parabolic example is

$$f(z) = \frac{(1+i)z - i}{iz + 1 - i}$$

($T = \begin{pmatrix} 1+i & -i \\ i & 1-i \end{pmatrix}$ has determinant 1 and trace 2, and a little calculation shows $|f(e^{i\theta})| = 1$.) Thus, all nonloxodromic possibilities occur. Here is what one can say about fixed points:

Theorem 9.3.4. *Let $f \in \mathcal{M}$ not be the identity. Then*

(a) *If f is elliptic, it has one fixed point at z_0 in $\mathbb{D}$ and one fixed point in $\mathbb{C} \setminus \overline{\mathbb{D}}$ at $1/\bar{z}_0$.*

(b) *If f is hyperbolic or parabolic, all the fixed points of f lie in $\partial\mathbb{D}$.*

Proof. By (9.3.4), if $f \in \mathcal{M}$ has a fixed point z_0, then $1/\bar{z}_0$ is also a fixed point, so if there is a fixed point not in $\partial\mathbb{D}$, there is one, call it z_0 in $\mathbb{D}$.

If $f(z_0) = z_0$, then $h \equiv g_{-z_0}^{-1} f g_{-z_0}$ maps zero to zero, and so is $h(z) = e^{i\theta} z$, which is elliptic, and thus f is elliptic. This proves (b).

All that remains is the proof that elliptic elements of $\mathcal{M}$ cannot have their fixed points on $\partial\mathbb{D}$. As we have seen, if f has a fixed point off $\partial\mathbb{D}$, it has a second at the reflected point. Thus, if f has a fixed point on $\partial\mathbb{D}$, it must have two. Let g be a map in $\mathfrak{F}$ that takes these two fixed points to zero and infinity and some other point, z_2, on $\partial\mathbb{D}$ to ± 1. g thus maps $\partial\mathbb{D}$ to $\mathbb{R}$ and so if we pick the ± 1 for $g(z_2)$ properly, $\mathbb{D}$ maps to $\mathbb{C}_+$. Since $h \equiv g f g^{-1}$ fixes zero and infinity and is elliptic, it has the form $h(z) = e^{i\theta} z$. No such map takes $\mathbb{C}_+$ to $\mathbb{C}_+$, which proves (a). $\square$

Remark. We will see later (see the discussion after Proposition 9.3.8) a geometric way to understand why parabolic and hyperbolic maps have their fixed points on $\partial\mathbb{D}$.

Obviously, if $f, g \in \mathcal{M}$ are conjugate in $\mathcal{M}$, they are conjugate in $\mathfrak{F}$ but, in principle (and in practice!), they could be conjugate in $\mathfrak{F}$ but not in $\mathcal{M}$. Put differently, if $C \subset \mathcal{F}$ is a class in $\mathcal{F}$ and $C \cap \mathcal{M} \neq \emptyset$, $C \cap \mathcal{M}$ is one or more classes in $\mathcal{M}$. Here is the breakdown:

Theorem 9.3.5. (a) *Each hyperbolic conjugacy class in $\mathcal{F}$ intersects $\mathcal{M}$. Two hyperbolic elements in $\mathcal{M}$ are conjugate in $\mathcal{M}$ if and only if they are conjugate in $\mathcal{F}$. Hyperbolic conjugacy classes in $\mathcal{M}$ are labeled by $a \in (0, 1)$ with*

$$f_a(z) = \frac{z-a}{1-az} \tag{9.3.5}$$

The associated T in $\mathbb{SU}(1,1)$ has $\mathrm{Tr}(T) = 2/(1-|a|^2)^{1/2}$.

(b) *Each elliptic conjugacy class in $\mathcal{F}$ intersects $\mathcal{M}$, and for $\theta \in (0, \pi/2)$, its intersection is two classes in $\mathcal{M}$ labeled by $\pm\theta$. The $\mathcal{F}$-class with $\theta = \pi/2$ $(\mathrm{Tr}(T) = 0)$ intersects $\mathcal{M}$ in a single class of $\mathcal{M}$. All elliptic classes are labeled by $\theta \in \pm(0, \pi/2)$. An element in the class is*

$$f_\theta(z) = e^{2i\theta} z \tag{9.3.6}$$

The associated trace is $2\cos\theta$.

(c) *The single parabolic class in $\mathcal{F}$ intersects $\mathcal{M}$ and the intersection is two classes of $\mathcal{M}$ of which representative elements are*

$$f_\pm(z) = \frac{(1 \pm i)z \mp i}{iz + 1 \mp i} \tag{9.3.7}$$

These have $\mathrm{Tr}(T) = 2$.

Remark. The $f_\pm$ in (9.3.7) has

$$f_\pm^{(n)}(0) = \frac{n^2}{1+n^2} \pm \frac{in}{1+n^2}$$

and iterates approach 1 asymptotically tangent to $\partial\mathbb{D}$ but from the top (resp. bottom) for f_+ (resp. f_-). In $\mathcal{F}$, they are conjugate via $g(z) = z^{-1}$, but that maps $\mathbb{D}$ to $\mathbb{C} \setminus \overline{\mathbb{D}}$ and is not in $\mathcal{M}$.

Proof. (a), (c) It is easier to look at the conjugate of $\mathcal{M}$ that maps $\mathbb{C}_+$ to $\mathbb{C}_+$, that is, $\mathbb{SL}(2, \mathbb{R})$. In the hyperbolic case, we can find a conjugate in $\mathbb{SL}(2, \mathbb{R})$ that takes any hyperbolic map to one whose fixed points are 0 and ∞ and with 0 the attracting fixed point. The classes in $\mathbb{SL}(2, \mathbb{R})$ are thus $z \mapsto az$ with $a \in (0, 1)$, as they are in $\mathbb{SL}(2, \mathbb{C})$. In the parabolic case, we can take the fixed point to infinity. The map is then $T_b(z) = z + b$ with $b \in \mathbb{R} \setminus 0$. By a scaling map in $\mathbb{SL}(2, \mathbb{R})$, we can conjugate that to $T_{\pm 1}$ but T_{+1} and T_{-1} are not conjugate in $\mathbb{SL}(2, \mathbb{R})$. The conjugacy in $\mathbb{SL}(2, \mathbb{C})$ is by $z \to -z$, which maps $\mathbb{C}_+$ to $\mathbb{C}_-$.

(b) By conjugating with f_{z_0}, we can suppose the elliptic map has zero as a fixed point, so of the form (9.3.6). For distinct θ's, these are not conjugate in $\mathcal{M}$, although conjugation with $z \to 1/z$ takes f_θ to $f_{-\theta}$. $\square$

Next, we want to discuss the Ford representation when $f \in \mathcal{M}$. Note that $f \in \mathcal{M}$ has $f(\infty) = \infty$ if and only if $f(0) = 0$, so the condition that f not leave ∞ fixed is $f(z) \not\equiv e^{i\theta} z$.

Theorem 9.3.6. *Let $f \in \mathcal{M}$ not be a rotation about* 0. *Then the isometric circle of f has a center outside $\overline{\mathbb{D}}$ and is orthogonal to $\partial\mathbb{D}$. $z = 0$ lies outside both the initial and final disks for f and on the* (*Euclidean*) *perpendicular bisection of the line between the center of D_i and D_f. $f(0)$ lies in D_f.*

Proof. We know f maps $\mathbb{C}\setminus\overline{\mathbb{D}}$ to itself, so $f^{-1}(\infty) \in \mathbb{C}\setminus\overline{\mathbb{D}}$, which says the center of the circle lies outside $\overline{\mathbb{D}}$. We know $f = f_T$ for $T = \left(\begin{smallmatrix} a & c \\ \bar{c} & \bar{a} \end{smallmatrix}\right)$. Then $f^{-1}(\infty) = -\frac{\bar{a}}{\bar{c}}$ and $f(\infty) = \frac{a}{\bar{c}}$. Since $|f^{-1}(\infty)| = |f(\infty)|$, they are equidistant from 0, which means that 0 lies on the perpendicular bisector of the line between $f^{-1}(\infty)$ and $f(\infty)$. Thus, in the Ford factorization of $f = QR$, Q maps $\mathbb{D}$ to $\mathbb{D}$, so $R = QF$ maps $\mathbb{D}$ to $\mathbb{D}$. By Theorem 9.2.23, the isometric circle is orthogonal to $\partial\mathbb{D}$.

With $f = f_T$ and $T = \left(\begin{smallmatrix} a & c \\ \bar{c} & \bar{a} \end{smallmatrix}\right)$ and $|a|^2 - |c|^2 = 1$, we have that $|\bar{c}z + \bar{a}| = 1$ is the isometric circle. Since $|\bar{c}\cdot 0 + \bar{a}| = |a| > 1$, (if f is not a rotation), 0 is outside $\bar{D}_i$. Since D_f is the initial circle for f^{-1}, 0 is also outside $\bar{D}_f$. $f(0) \in D_f$ since $\mathbb{C}\setminus \bar{D}_i$ is mapped to D_f by f (see Theorem 9.2.32). $\square$

Remarks. 1. There is a quantitative way of seeing that $f(0)$ lies inside D_f, namely, since $f(0) = \frac{c}{\bar{a}}$, $f(\infty) = \frac{a}{\bar{c}}$, so $|f(0) - f(\infty)| = \frac{1}{|ca|}$ since $|a|^2 - |c|^2 = 1$. On the other hand, r_f is $|c|^{-1}$, so $|a| > 1$ implies $|f(0) - f(\infty)| < r_f$.

2. This theorem illustrates Theorem 9.2.32. If f is parabolic, C_i and C_f intersect on $\partial\mathbb{D}$ (since C_i and C_f are orthocircles). If T is elliptic, C_i and C_f intersect in points inside and outside. If T is hyperbolic, the line from center to center intersects $\partial\mathbb{D}$, giving the fixed point on that line segment.

Definition. An *orthocircle* is a circle or line in $\mathbb{C}$ that intersects $\partial\mathbb{D}$ in two points with orthogonal intersections.

The *extended Möbius transformations* are those *extended FLTs* that map $\mathbb{D}$ onto $\mathbb{D}$. The set of such maps we denote by $\widetilde{\mathcal{M}}$. Since c is such a map, one easily sees:

Proposition 9.3.7. *Every $f \in \widetilde{\mathcal{M}}$ is of the form g or gc for some $g \in \mathcal{M}$. A reflection is an extended Möbius transformation if and only if the line or circle in which one reflects is an orthocircle.*

Proof. The first statement is immediate and the second follows from Theorem 9.2.23. $\square$

One big difference between $\mathcal{M}$ and $\mathfrak{F}$ is that there is a Riemannian metric (on $\mathbb{D}$) that is left fixed by all elements of $\mathcal{M}$, while there cannot be such a metric on $\mathbb{P}$ left invariant by all elements of $\mathfrak{F}$ since:

Proposition 9.3.8. *If (X, ρ) is a metric space, $f\colon X \to X$ an isometry* (*i.e., $\rho(f(x), f(y)) = \rho(x, y)$ for all x, y*), *then there cannot be an x_0 and $x_\infty \neq x_0$ so that $f^{(n)}(x_0) \to x_\infty$.*

Proof. Since f is continuous, $f(f^{(n)}(x_0)) \to f(x_\infty)$ but $f(f^{(n)}(x_0)) = f^{(n+1)}(x_0)$, so x_∞ is a fixed point. But then $\rho(f^{(n+1)}(x_0), x_\infty) = \rho(f^{(n+1)}(x_0), f(x_\infty)) = \rho(f^{(n)}(x_0), x_\infty) = \cdots = \rho(x_0, x_\infty) \neq 0$. Thus, $f^{(n)}(x_0)$ does not converge to x_∞. This contradiction proves the result. $\square$

Thus, isometries cannot have attracting fixed points, so there is no metric (let alone Riemann metric) on $\mathbb{P}$ in which hyperbolic or parabolic maps are isometries. The reason we can define a metric on $\mathbb{D}$ in which hyperbolic or parabolic maps are isometries is that the attracting fixed points are not in $\mathbb{D}$ (but in $\partial\mathbb{D}$). This will not be a problem because the metric will diverge as we approach $\partial\mathbb{D}$.

The following calculation is the key to the invariant metric:

Theorem 9.3.9. *Let f be an extended Möbius transformation. Then*

$$|f'(z)| = \frac{1 - |f(z)|^2}{1 - |z|^2} \tag{9.3.8}$$

Proof. If g is an antilinear extended Möbius transformation, then $f = cg$ is in $\mathcal{M}$ and $|f'(z)| = |g'(z)|$ and $|f(z)| = |g(z)|$, so (9.3.8) for f implies it for g, that is, we can suppose $f \in \mathcal{M}$, that is, $f = f_T$ with

$$T = \begin{pmatrix} a & c \\ \bar{c} & \bar{a} \end{pmatrix} \tag{9.3.9}$$

where $\det(T) = |a|^2 - |c|^2$.

As we computed in (9.2.60),

$$|f'(z)| = \frac{1}{|\bar{c}z + \bar{a}|^2} \tag{9.3.10}$$

On the other hand, since (the cross-terms cancel)

$$|az + c|^2 - |\bar{c}z + \bar{a}|^2 = (|a|^2 - |c|^2)(|z|^2 - 1)$$

we see that

$$|f(z)|^2 - 1 = \frac{|z|^2 - 1}{|\bar{c}z + \bar{a}|^2} \tag{9.3.11}$$

(9.3.11) and (9.3.10) imply (9.3.8). $\square$

The standard Euclidean Riemannian structure will be called $(dz)^2$. The *Poincaré metric* on $\mathbb{D}$ is defined to be the one associated to the Riemann structure

$$(1 - |z|^2)^{-2}\,(dz)^2 \tag{9.3.12}$$

Put differently, the length of a smooth curve $\gamma\colon [0, 1] \to \mathbb{D}$ is

$$L(\gamma) = \int_0^1 |\gamma'(s)|(1 - |\gamma(s)|^2)^{-1}\,ds \tag{9.3.13}$$

and

$$\rho(x, y) = \inf\{L(\gamma) \mid \gamma(0) = x,\ \gamma(1) = y\} \tag{9.3.14}$$

Theorem 9.3.10. *Let $g \in \widetilde{\mathcal{M}}$. Then g preserves the Poincaré–Riemann structure* (9.3.12)*, the length* (9.3.13)*, and the metric* (9.3.14).

Proof. It suffices to prove preservation of the Riemann structure. Since g is conformal or anticonformal, it preserves angles, so we need only show infinitesimal lengths get mapped properly. The mapping is, of course, by $|f'(z)|$. (9.3.8) is precisely this statement, that is,

$$\frac{|df|}{1-|f|^2} = \frac{|dz|}{1-|z|^2} \tag{9.3.15}$$

□

The metric has a $\frac{1}{2}(1-|z|)^{-1}$ divergence as $|z| \to 1$ whose integral diverges logarithmically, so we expect $\rho(0,z)$ to look like $\frac{1}{2}\log(1-|z|)^{-1}$ as $|z| \uparrow 1$. That is part of the following:

The set $\mathbb{D}$ with the Poincaré metric is called the $\mathbb{D}$*-model of the hyperbolic plane*.

Theorem 9.3.11. (i) *The geodesic from* 0 *to* $z \in \mathbb{D}$ *is the straight line segment between them.*

(ii) *We have that* $\rho(z,0)$ *is given by*

$$\tanh(\rho(z,0)) = |z| \tag{9.3.16}$$

so that as $|z| \uparrow 1$,

$$\rho(z,0) = \tfrac{1}{2}\log((1-|z|)^{-1}) + \tfrac{1}{2}\log 2 + O(1-|z|) \tag{9.3.17}$$

(iii) *For any* $z, w \in \mathbb{D}$,

$$\tanh(\rho(z,w)) = \frac{|z-w|}{|1-\bar{z}w|} \tag{9.3.18}$$

(iv) *The geodesics in the* $\mathbb{D}$*-model of the hyperbolic plane are precisely segments of the orthocircles.*

Proof. (i) Because the Poincaré metric is conformal, for any curve from 0 to z, if $\hat{z} = z/|z|$, then

$$\begin{aligned} |\gamma'(s)|^2 &= [\mathrm{Re}(\gamma'(s)\hat{z})]^2 + [\mathrm{Im}(\gamma'(s)\hat{z})]^2 \\ &\geq \mathrm{Re}(\gamma'(s)\hat{z})^2 \end{aligned} \tag{9.3.19}$$

that is, the infinitesimal length is larger than its radial component. Since the metric is invariant under rotations,

$$|\gamma'(s)| \geq \frac{1}{1-|\gamma(s)|^2}\left|\frac{d|\gamma(s)|}{ds}\right| \tag{9.3.20}$$

with equality only if $\arg(\gamma(s))$ is constant. This shows the minimal length path has $\arg(\gamma(s))$ constant, and so it is the straight line.

(ii) By (i), $\gamma(s) = sz$, so

$$|\gamma'(s)| = \frac{|z|}{1-|\gamma(s)|^2}$$

and thus

$$\begin{aligned} \rho(0,z) &= \int_0^1 \frac{|z|\,ds}{1-|zs|^2} = \int_0^{|z|} \frac{dy}{1-y^2} \\ &= \mathrm{arctanh}(|z|) \end{aligned}$$

since $\frac{d}{dy}\,\mathrm{arctanh}(y) = (1-y^2)^{-1}$. This proves (9.3.16).

To get (9.3.17), we note (9.3.16) with $|z| = r$, we have

$$\frac{1 - e^{-2\rho}}{1 + e^{-2\rho}} = r \tag{9.3.21}$$

so

$$(1 - r)^{-1} = \left[\frac{2e^{-2\rho}}{1 + e^{-2\rho}}\right]^{-1} = \tfrac{1}{2}\, e^{2\rho} + \tfrac{1}{2} \tag{9.3.22}$$

which implies (9.3.17).

(iii) By the invariance of ρ under $f \in \mathcal{M}$,

$$\begin{aligned}\rho(z, w) &= \rho(f_z(z), f_z(w)) \\ &= \rho\left(0, \frac{w - z}{1 - \bar{z}w}\right)\end{aligned}$$

so (9.3.16) implies (9.3.18).

(iv) The geodesic from z to w is taken into the geodesic from 0 to $g_z(w)$ by g_z. Thus, this geodesic is the image under g_z^{-1} of a diameter, so a segment of an orthocircle. □

Remark. A convenient way of rewriting (9.3.21) is

$$e^{-2\rho(0,z)} = \frac{1 - |z|}{1 + |z|} \tag{9.3.23}$$

Notice that given an orthocircle and a point not on that circle, we can find multiple orthocircles that contain the point but do not intersect the original circle, for by a Möbius transformation, we can suppose the point is 0 and it is obvious that multiple diameters avoid a given orthocircle. That is, if parallel lines mean infinite geodesics, which are nonintersecting, Euclid's fifth postulate fails. This is a homogeneous geometry that is a realization of Lobachevsky's plane.

Analogous to the fact that $\mathcal{M}$ is the set of holomorphic bijections of $\mathbb{D}$, we can describe all isometries.

Theorem 9.3.12. *Let $f\colon \mathbb{D} \to \mathbb{D}$ be any continuous function, which is an isometry in the Poincaré metric. Then $f \in \widetilde{\mathcal{M}}$.*

Remark. Since we have seen all $f \in \widetilde{\mathcal{M}}$ are isometries, we see $\widetilde{\mathcal{M}}$ is the set of all isometries.

Proof. Let $f(0) = z_0$, $f(\frac{1}{2}) = w_0$. Then $(g_{z_0} \circ f)(0) = 0$. Since $g_{z_0} \circ f$ is an isometry, $\rho((g_{z_0} \circ f)(\frac{1}{2}), 0) = \rho((g_{z_0} \circ f)(\frac{1}{2}), (g_{z_0} \circ f)(0)) = \rho(\frac{1}{2}, 0)$. Since $\rho(w, 0)$ is a monotone function of $|w|$, $|(g_{z_0} \circ f)(\frac{1}{2})| = \frac{1}{2}$. Thus, by following g_{z_0} by a rotation about zero, we find $h \in \mathcal{M}$, so $h \circ f$ takes 0 to 0 and $\frac{1}{2}$ to $\frac{1}{2}$.

It thus takes the geodesic from 0 to $\frac{1}{2}$ and its continuation setwise to itself, that is, $h \circ f$ maps $(-1, 1)$ to itself. Since $h \circ f$ is one-one and continuous, either $h \circ f[\mathbb{C}_+ \cap \mathbb{D}] \subset \mathbb{C}_+ \cap \mathbb{D}$ or in $\mathbb{C}_- \cap \mathbb{D}$. By replacing h by ch, we can be sure the image is in $\mathbb{C}_+ \cap \mathbb{D}$, that is, we can find $h \in \widetilde{\mathcal{M}}$ so that

$$(h \circ f)(0) = 0 \qquad (h \circ f)(\tfrac{1}{2}) = \tfrac{1}{2} \qquad (h \circ f)(\mathbb{C}_+ \cap \mathbb{D}) \subset \mathbb{C}_+ \cap \mathbb{D}$$

If we prove $h \circ f$ is the identity, then $f = h^{-1} \in \widetilde{\mathcal{M}}$.

Let w lie in $\mathbb{C}_+ \cap \mathbb{D}$. The two sets $S_0 = \{w_1 \mid \rho(w_1, 0) = \rho(w, 0)\}$ and $S_1 = \{w_1 \mid \rho(w_1, \frac{1}{2}) = \rho(w, \frac{1}{2})\}$ are circles (S_0 is a circle by (9.3.16) and S_1 is an image under a Möbius transformation of a circle about 0, and so a circle).

These circles are distinct (look at their real points) and contain w and $\bar{w}$. Since circles can intersect in at most two points, $S_1 \cap S_0 = \{w, \bar{w}\}$. But $(h \circ f)(w) \in S_1 \cap S_0$ and is in $\mathbb{C}_+$ so $(h \circ f)(w) = w$. Thus, $h \circ f = \mathbf{1}$ on $\mathbb{C}_+ \cap \mathbb{D}$ and similarly on $\mathbb{C}_- \cap \mathbb{D}$ and so, by continuity, on $\mathbb{D}$. $\square$

Next, we want to look at which points in $\mathbb{D}$ are closer to z than w. For Euclidean geometry, this is answered by the perpendicular bisector. The same is true here but the bisector is an orthocircle:

Theorem 9.3.13. *Fix $z_0 \neq z_1$ both in $\mathbb{D}$. Then*

$$\{w \mid \rho(w, z_0) = \rho(w, z_1)\}$$

is an orthocircle. Removing this orthocircle from $\mathbb{D}$ yields two open connected components with z_0 and z_1 in the two components. In the component with z_0, we have $\rho(w, z_0) < \rho(w, z_1)$, and vice versa within the other.

Proof. Suppose first $z_0 = ia$, $z_1 = -ia$ with $0 < a < 1$ and $\operatorname{Im} w > 0$, $w \in \mathbb{D}$. We claim $\rho(w, z_0) < \rho(w, z_1)$. By (9.3.18), this is equivalent to

$$|(w - ia)(1 + ia\bar{w})| < |(w + ia)(1 - ia\bar{w})| \tag{9.3.24}$$

$$\text{LHS} = A + B \qquad \text{RHS} = A - B$$

where

$$A = -ia + ia|w|^2 \qquad B = w + a^2\bar{w}$$

A is pure imaginary, so

$$|\operatorname{Re}(A + B)| = |\operatorname{Re} B| = |\operatorname{Re}(A - B)|$$

Since $|w| < 1$ and $|a| < 1$, $\operatorname{Im} A < 0$, and since $\operatorname{Im} w > 0$, $\operatorname{Im} B > 0$. Thus, $|\operatorname{Im}(A + B)| < |\operatorname{Im}(A - B)|$, proving (9.3.24).

This proves the result in the special case $z_0 = ia$, $z_1 = -ia$. In general, let w be the geodesic midpoint of the geodesic from z_0 to z_1. Let $g \in \mathcal{M}$ take w to 0. Since it preserves geodesics and hyperbolic lengths, it must map z_0 and z_1 to equidistant points from 0 on the same line through zero. By a further rotation, we see any pair is equivalent to the special case under a hyperbolic isometry. $\square$

Corollary 9.3.14. *Let r be a reflection in an orthocircle, C. Let w, z be on the same side of C (and not on C). Then*

$$\rho(w, z) < \rho(w, r(z)) \tag{9.3.25}$$

Proof. Since ρ is preserved by $\gamma \in \mathcal{M}$, we can suppose the orthocircle is $(-1, 1)$. Then C is the perpendicular bisector of points equidistant from z, $r(z) = \bar{z}$, and (9.3.25) is the final assertion of the theorem. $\square$

Theorem 9.3.15. *For any $f \in \mathcal{M}$, the hyperbolic perpendicular bisection of the hyperbolic line from* 0 *to* $f(0)$ *is the part of the boundary,* ∂D_f*, of the final circle,* D_f*, inside* $\mathbb{D}$*.*

Proof. f^{-1} is the reflection in ∂D_f followed by reflection in the line, L, which is the Euclidean bisector of the line between the centers of D_f and D_i. By Theorem 9.3.6, $0 \in L$, so for $w \in \mathbb{D} \cap \partial D_f$,

$$|f^{-1}(w)| = |w| \tag{9.3.26}$$

Since $\rho(0, z)$ is a function of $|z|$ only, we have

$$\rho(0, f^{-1}(w)) = \rho(0, w) \tag{9.3.27}$$

But since f is a ρ-isometry,

$$\rho(f(0), w) = \rho(0, f^{-1}(w)) \tag{9.3.28}$$

Thus, w lies on the hyperbolic perpendicular bisector. □

Remarks and Historical Notes. The fact that $\mathbb{D}$ has a metric in which all fractional linear automorphisms are isometries is a discovery of Poincaré. This metric has constant curvature -1. It is a remarkable fact that the other two simply connected Riemann surfaces (namely, $\mathbb{C}$ and $\mathbb{C} \cup \{\infty\}$) have natural constant curvature metrics—the flat metric on $\mathbb{C}$ and the spherical metric on $\mathbb{C} \cup \{\infty\}$. However, in these other cases, there are automorphisms that are not isometries.

For further discussion of the group $\mathbb{SU}(1, 1)$, see Sections 10.4 and 10.5 of [400].

The study of subgroups of $\mathbb{SL}(2, \mathbb{R}) \cong \mathbb{SU}(1, 1)$ has arithmetic significance because it contains matrices with integral coefficients. Indeed, $\mathbb{SL}(2, \mathbb{Z})$, the 2×2 matrices of determinant 1 with integral coefficients, is a subgroup. For this reason, the upper half-plane model is often more popular than the disk model.

Katok [216] proves Theorem 9.3.13 in the UHP model where the calculation is less messy.

9.4 FUCHSIAN GROUPS

In this section, we will say something about general Fuchsian groups as a preliminary to the study in the next two sections of the ones of interest for finite gap Jacobi matrices. This will hardly be a comprehensive look at the subject—our example, as we will explain in the next two sections, will be infinitely nicer than more typical cases, so we can avoid discussions of all sorts of subtleties. Our main theme here will be equivalences of various measures of discreteness and of a critical number called the Poincaré index.

Given $f \in \mathcal{M}$, there are various measures of how "large" f is, that is, how far it is from the identity. We can write $f = f_T$ with $\det(T) = 1$ and use $\|T\|$; we can look at $(1 - |f(0)|)^{-1}$, $e^{2\rho(f(0),0)}$, or $|f'(0)|^{-1}$, or replace $f(0)$ by $f(z)$ for some other $z \in \mathbb{D}$. Our initial goal will be to prove an equivalence in the quantitative sense of upper and lower bounds on ratios. We begin with what happens at a fixed