AN INTRODUCTION TO HYPERBOLIC GEOMETRY MAT4510/3510

BJØRN JAHREN

(Version: August 17, 2010)

Introduction.

In Euclid's axiom system the parallel axiom has always caused the most trouble. Already from the beginning it was recognized as less obvious than the other axioms, and during more than two thousand years of fascinating mathematical history geometers were trying to either prove it from the other axioms or replace it by something more obvious but with the same consequences. Today we know that the reason they did not succeed is that there exist geometries where the axiom is not satisfied. One may wonder why this was not realized earlier, but we must remember that geometry throughout all this time was concerned with a description of the world "as it is", and in the real world a statement like the parallel axiom must either be true or not true. Euclid's axioms do not *define* geometry; they describe more precisely what kind of arguments we are allowed to use when proving new results about the geometry of the world around us.

But in the 19th century the development of mathematics and mathematical thinking finally brought freedom from this purely descriptive approach to geometry, allowing mathematicians like Lobachevski, Bolyai and Gauss to realize that one might construct perfectly valid geometries where all the other axioms of Euclid hold, but where the parallel axiom fails.

The first concrete such models were constructed by *Beltrami* in 1868, and all the models we shall present here are due to him, even if some of them have names after other mathematicians.

Instead of Euclid's axiom system we shall start with Hilbert's axioms, and we define a *hyperbolic geometry* to be an incidence geometry with betweenness and congruence where Hilbert's axioms hold, except that the parallel axiom is replaced by

H: Given a line l and a point $P \notin l$, there are at least two lines through P which do not intersect l.

We start with a heuristic discussion that may, hopefully, serve as a motivation for our models for hyperbolic geometry. Discussing Hilbert's axiom system we observed that an open, convex subset K of the Euclidean plane is a candidate for such a geometry if we define 'lines' to be intersections between K and Euclidean lines (i. e. open *chords*), and betweenness is defined as in $\mathbb{R}^2$. The *Beltrami–Klein model* $\mathbb{K}$ of the hyperbolic plane utilizes a particularly simple such convex set: the interior of the unit disk. Then the incidence- and betweenness axioms will remain satisfied, as will Dedekind's

axiom. However, there will clearly exist *infinitely many* lines parallel to a line l through a point P outside the line, hence axiom H will hold instead of axiom P. (Recall that we call two lines parallel if they do not intersect.)

The only missing ingredient of a hyperbolic geometry is therefore a notion of *congruence* satisfying Hilbert's axioms C1–C6. Clearly the usual, Euclidean definition of congruence does not work, since the fundamental axiom C1 breaks down. (Although C2–C6 are still satisfied!) But, inspired by the Euclidean definition of congruence as equivalence under the action of the Euclidean group of transformations, it is natural to see if there is an analogous group of homeomorphisms of the unit disk that might work.

An absolutely essential property these homeomorphisms should have is that they should map all chords to chords. This property is rather difficult to study directly, but there is a geometric trick that will enable us to find sufficiently many such maps, using some elementary results of complex function theory! The trick is to map $\mathbb{K}$ to open subsets of $\mathbb{C}$ by certain homeomorphisms mapping chords to circular arcs. Then the problem is reduced to finding homeomorphism mapping circles to circles, and this is much simpler, leading to the theory of *Möbius transformations*. The Beltrami–Klein model $\mathbb{K}$ is obtained by transporting back the resulting congruence notion.

However, since the theory is computationally (as well as in other respects) much simpler in the homeomorphic models in $\mathbb{C}$ — *the Poincaré disk* $\mathbb{D}$ and *Poincaré's upper half–plane* $\mathbb{H}$ — these are the models mostly studied. They will be models for hyperbolic geometry where the "lines" are circular arcs (and certain straight lines) in $\mathbb{C}$. The Beltrami–Klein model will only be used for geometric motivation, except for a short discussion in an Appendix.

Here is an overview of the contents of these notes. The preparatory Section 1 discussed the transformation of $\mathbb{K}$ into the other models and relations between them. The Möbius transformations — especially those preserving the upper half–plane — are introduced and studied in some depth in Sections 2 and 3. These transformations can be used to define a congruence relation, giving $\mathbb{H}$ the structure of a hyperbolic plane. This is verified in Section 4. In Sections 5 and 6 we define distance and angle measures in $\mathbb{H}$, and in Section 7 we translate everything done so far to the disk model $\mathbb{D}$. Each of the models has its own advantages, and this is exploited in the remaining sections, where we study arc length and area (Section 8) and trigonometry (Section 9).

Some notation: In the different models we are going to introduce ($\mathbb{K}$, $\mathbb{B}$, $\mathbb{D}$, $\mathbb{H}$), the 'lines' of the geometry will be different types of curves. We shall call these curves $\mathbb{K}$–lines, $\mathbb{B}$–lines etc., or simply *hyperbolic* lines if the model is understood or if it doesn't matter which model we use. For example, the $\mathbb{K}$–lines are the open chords in the interior of the unit circle in the Euclidean plane. Similarly, many of our constructions will take place in standard Euclidean $\mathbb{R}^2$ and $\mathbb{R}^3$, and then 'lines', 'circles' etc. will refer to the usual Euclidean notions.

1. Stereographic projection

As a set, $\mathbb{K}$ is just the interior of the unit disk in $\mathbb{R}^2$:

$$\mathbb{K} = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 1\}.$$

Consider $\mathbb{R}^2$ as the subspace of $\mathbb{R}^3$ where the last coordinate is 0, and let $\mathbb{B}$ be the lower open hemisphere

$$\mathbb{B} = \{(x, y, z) \in R^3 \,|\, x^2 + y^2 + z^2 = 1, z < 0\}.$$

Vertical projection then defines a homeomorphism $\mathbb{K} \approx \mathbb{B}$, mapping the chords in $\mathbb{K}$ onto (open) semi–circles in $\mathbb{B}$ meeting the boundary curve $\{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 = 1\}$ orthogonally. (Perhaps the easiest way to see this is to consider the image of a chord as the intersection between $\mathbb{B}$ and the plane which contains the chord and is parallel to the z-axis. Then, by symmetry, the image is half of the intersection of this plane with the sphere.) Defining such half–circles as $\mathbb{B}$–lines, we obtain another model $\mathbb{B}$ with the same properties as $\mathbb{K}$.

We now use *stereographic projection* to map $\mathbb{B}$ back to $\mathbb{R}^2$. The version of stereographic projection that we shall use here is the homeomorphism $\mathbb{S}^2 - (0, 0, 1) \approx \mathbb{R}^2$ defined as follows: If P is a point in $\mathbb{S}^2 - (0, 0, 1)$, there is a uniquely determined straight line in $\mathbb{R}^3$ through P and (0,0,1), and this line meets $\mathbb{R}^2$ in a unique point. This defines a map $\Phi : \mathbb{S}^2 - (0, 0, 1) \to \mathbb{R}^2$ which clearly is both injective and surjective. A simple argument using similar triangles (see fig.1) shows that Φ is given by the formula

$$\Phi(x, y, z) = \left(\frac{x}{1 - z}, \frac{y}{1 - z} \right), \tag{1.1}$$

and the inverse map is given by

$$\Phi^{-1}(u, v) = \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1} \right). \tag{1.2}$$

These maps are both continuous, hence inverse homeomorphisms.

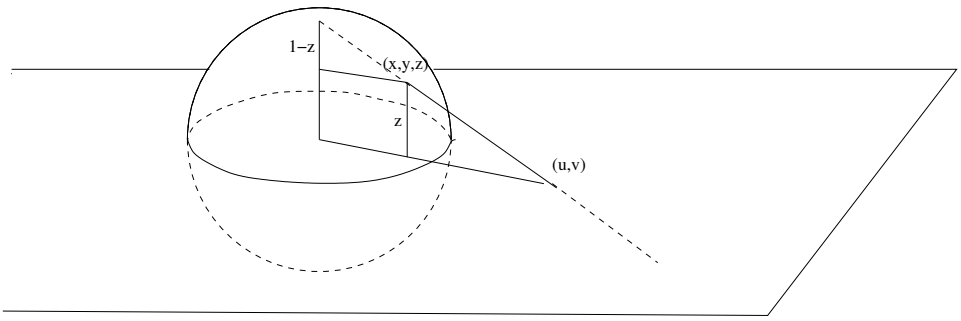

Fig. 1. Stereographic projection

In the following two Lemmas we state some important properties of stereographic projection.

Lemma 1.1. *Let $\mathcal{C}$ be a circle on $\mathbb{S}^2$.*

(i) *If $(0, 0, 1) \notin \mathcal{C}$, then $\Phi(\mathcal{C})$ is a circle in $\mathbb{R}^2$.*

(ii) *If* $(0,0,1) \in \mathcal{C}$*, then* $\Phi(\mathcal{C} - (0,0,1))$ *is a straight line in* $\mathbb{R}^2$

Proof. The circle $\mathcal{C}$ is the intersection between $\mathbb{S}^2$ and a plane defined by an equation $ax + by + cz = d$, say, and $(0,0,1) \in \mathcal{C}$ if and only if $c = d$. Now substitute x, y and z from formula (1.1 for Φ and get

$$\frac{2au}{u^2+v^2+1} + \frac{2bv}{u^2+v^2+1} + \frac{c(u^2+v^2-1)}{u^2+v^2+1} = d\,.$$

Clearing denominators and collecting terms then yields

$$(c-d)(u^2+v^2) + 2au + 2bv = c + d\,.$$

This is the equation of a line if $c = d$ and a circle if $c \neq d$. □

To formulate the next Lemma, recall that to give a curve an *orientation* is to choose a sense of direction along the curve. In all cases of interest to us, this can be achieved by choosing a nonzero tangent vector at every point, varying continuously along the curve. The *angle* between two oriented curves intersecting in a point P is then the angle between the tangent vectors at P.

Lemma 1.2. Φ *preserves angles — i. e. if* $\mathcal{C}$ *and* $\mathcal{C}'$ *are oriented circles on* S^2 *intersecting in a point* P *at an angle* θ*, then their images under* Φ *intersect in* $\Phi(P)$ *at the same angle.*

Remark. Here we are only interested in unoriented angles, i. e. we do not distinguish between the angle between $\mathcal{C}$ and $\mathcal{C}'$ and the angle between $\mathcal{C}'$ and $\mathcal{C}$. Then we can restrict to angles between 0 and π, and this determines θ uniquely. Note that in this range θ is also determined by $\cos(\theta)$. (With appropriate choices of orientations of $\mathbb{S}^2$ and $\mathbb{R}^2$ the result is also true for oriented angles, but we shall not need this.)

Proof. By rotational symmetry around the z-axis we may assume that the point P lies in a fixed meridian, so we assume that $P = (0, y, z)$ with $y \geq 0$. Furthermore, it clearly suffices to compare each of the circles with this meridian, i. e. we may assume $\mathcal{C}'$ is the circle $x = 0$, with oriented tangent direction $(0, -z, y)$ at the point $(0, y, z)$. The image of this circle under Φ is the y-axis with tangent direction $(0, 1)$.

Observe that Φ can be extended to $\{(x, y, z) \in \mathbb{R}^3 | z < 1\}$ by the same formula (1.1), and that tangential curves will map to tangential curves (by the chain rule). Hence we can replace $\mathcal{C}$ by any curve in $\mathbb{R}^3$ with the same oriented tangent as $\mathcal{C}$ in P — e. g. a straight line. This line can be parametrized by

$$\theta(t) = (0, y, z) + t(\alpha, \beta, \gamma) = (t\alpha, y + t\beta, z + t\gamma),$$

where $\alpha^2 + \beta^2 + \gamma^2 = 1$ and $(\alpha, \beta, \gamma) \cdot (0, y, z) = \beta y + \gamma z = 0$. The angle u between this line and the meridian $\mathcal{C}'$ is determined by

$$\cos u = (\alpha, \beta, \gamma) \cdot (0, -z, y) = -\beta z + \gamma y\,.$$

Now consider the image of this line under Φ. This is parametrized by

$$\omega(t) = \Phi(\theta(t)) = \left(\frac{t\alpha}{1 - z - t\gamma}, \frac{y + t\beta}{1 - z - t\gamma} \right),$$

(Restrict t such that $1 - z - t\gamma > 0$.) It is geometrically obvious that this is again a straight line, and to see this from the formula for $\omega(t)$, note that

$$\begin{aligned}\omega(t) - \omega(0) = \omega(t) - \Phi(P) &= \left(\frac{t\alpha}{1-z-t\gamma}, \frac{y+t\beta}{1-z-t\gamma} - \frac{y}{1-z}\right) \\ &= \frac{t}{1-z-t\gamma}\left(\alpha, \frac{\beta - \beta z + \gamma y}{1-z}\right).\end{aligned}$$

(Straightforward calculation.) But this has constant direction given by the vector $V = (\alpha, \dfrac{\beta - \beta z + \gamma y}{1-z})$.

It remains to check that the angle v between V and the positive y-axis is equal to u, or, equivalently, that $\cos u = \cos v = \dfrac{V \cdot (0,1)}{||V||}$.

Recall that $\cos u = -\beta z + \gamma y$ and $\beta y + \gamma z = 0$. Then

$$z \cos u = -\beta z^2 + \gamma y z = -\beta z^2 - \beta y^2 = -\beta,$$

since $y^2 + z^2 = 1$. Similarly, $y \cos u = \gamma$. It follows that

$$\frac{\beta - \beta z + \gamma y}{1-z} = \frac{-z\cos u + \cos u}{1-z} = \cos u\,,$$

hence $V = (\alpha, \cos u)$. Moreover, $\beta^2 + \gamma^2 = z^2 \cos^2 u + y^2 \cos^2 u = \cos^2 u$, so $||V|| = \alpha^2 + \cos^2 u = \alpha^2 + \beta^2 + \gamma^2 = 1$. But then $\cos v = \cos u$. □

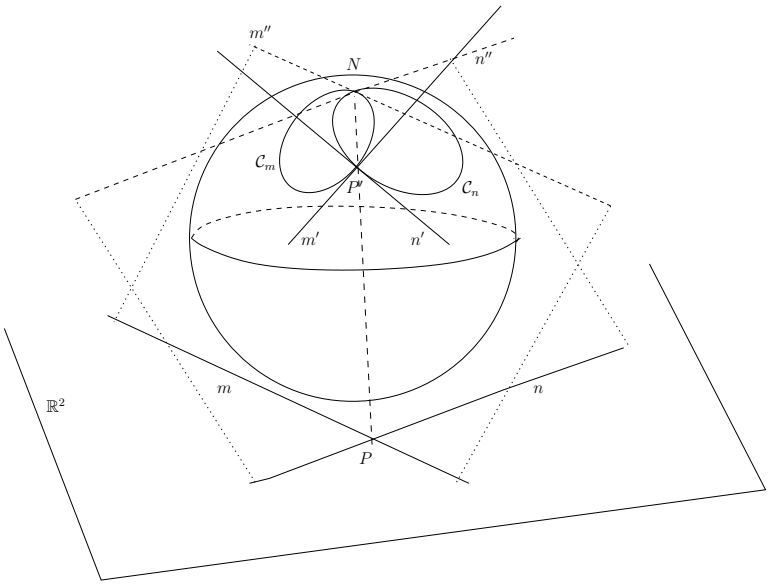

FIG. 2. Stereographic projection preserves angles

Figure 2 illustrates a geometric proof of Lemma 1.2. N is the "north pole", P' is a point on $\mathbb{S}^3$ and $P = \Phi(P)$. The lines m and n in $\mathbb{R}^2$ intersect in P. The crucial observation is that the image under Φ^{-1} of a line is a circle — the intersection between S^3 and the plane through the line and

N. Moreover, the tangent line at a point Q of this circle is the intersection between the plane and the tangent plane of S^3 at Q.

Accordingly, let $\mathcal{C}_m$ be the circle $\Phi^{-1}(m)$, and let m' and m'' be the tangents to $\mathcal{C}_m$ at P' and N. $\mathcal{C}_n$, n', n'' are defined similarly from n. Then by symmetry the angle between m' and n' is the same as the angle between m'' and n''. But since the tangent plane of S^3 at N is parallel to $\mathbb{R}^2$, the angle m'' is parallel to m and n'' is parallel to n. Hence the angle between m'' and n'' is the same as the angle between m and n.

There are, of course, many ways to identify $\mathbb{B}$ with an open hemisphere of $\mathbb{S}^2$, and we obtain a homeomorphic image in the plane by stereographic projection as long as we avoid the point (0,0,1). The Lemmas above imply that every such model will be bounded by a circle or a straight line in $\mathbb{R}^2$, and the 'lines' will correspond to circular arcs or straight lines meeting the bounding curve orthogonally. We will use of two models obtained this way, both being named after the great French mathematician Henri Poincaré:

1. *Poincaré's disk model* $\mathbb{D}$ is obtained by choosing $\mathbb{B}$ to be the lower hemisphere, as before. Then the image is again the interior of the unit circle, but the $\mathbb{D}$-lines are now either diameters or circular arcs perpendicular to the boundary circle.

2. The *Poincaré upper half–plane* $\mathbb{H} = \{(x, y) \,|\, y > 0\}$ is the image of the open hemisphere $\{(x, y, z) \in \mathbb{S}^2 \,|\, y > 0\}$. The $\mathbb{H}$–lines are either half–circles with center on the x-axis or straight lines parallel to the y-axis.

When we analyze these models we shall henceforth identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$ and make use of the extra structure and tools we have available there (complex multiplication, complex function theory etc.). Thus, as sets we make the identifications $\mathbb{D} = \{z \in \mathbb{C} \,|\, |z| < 1\}$ and $\mathbb{H} = \{z \in \mathbb{C} \,|\, \operatorname{Im} z > 0\}$.

$\mathbb{D}$ is the most symmetric model and therefore often the one best suited for geometric arguments. But we will see that $\mathbb{H}$ is better for analyzing and describing the notion of congruence. Therefore this where we begin our analysis.

Notation: In both models the hyperbolic lines have natural extensions to the boundary curve. (The unit circle for $\mathbb{D}$ and the real line for $\mathbb{H}$.) These extension points we refer to as *endpoints* of the hyperbolic lines, although they are not themselves points on the lines. Analogously, we also say that ∞ is an endpoint of a vertical $\mathbb{H}$–line.

Exercises.

1.1. Derive the formulae for Φ and its inverse.

1.2. We can also define stereographic projection from $(0, 0, -1)$ instead of $(0, 0, 1)$. Let Φ_- be the resulting map.
Determine $\Phi_- \circ \Phi^{-1}$. (We identify $\mathbb{R}^2$ with $\mathbb{C}$.)

1.3. If F is an identification between the two hemispheres we use in the definitions of $\mathbb{H}$ and $\mathbb{D}$, the map $\Phi \circ F \circ \Phi^{-1}$ will be a homeomorphism between

the two models. Find a formula for such a homeomorphism. (Choose F as simple as possible.)

1.4. (a) Show that $z \mapsto z^{-1} : \mathbb{C} - \{0\} \to \mathbb{C} - \{0\}$ corresponds to a rotation of $\mathbb{S}^2$ via stereographic projection.

(b) Which self–map of $\mathbb{C} - \{0\}$ does the antipodal map $x \mapsto -x$ on $\mathbb{S}^2 - \{0, 0, \pm 1\}$ correspond to?

2. Congruence in $\mathbb{H}$. Möbius transformations.

As noted before, we define congruence In Euclidean geometry as equivalence under the *Euclidean group* $E(2)$ of "rigid movements", generated by orthogonal linear transformations $x \mapsto Ax$ and translations $x \mapsto x + b$. This means that the congruence relation $\cong$ is defined by

Segments: $AB \cong A'B' \iff$ there is a $g \in E(2)$ such that $g(AB) = A'B'$.
Angles: $\angle BAC \cong \angle B'A'C' \iff$ there is a $g \in E(2)$ such that $g(\overrightarrow{AB}) = \overrightarrow{A'B'}$ and $g(\overrightarrow{AC}) = \overrightarrow{A'C'}$.

Observe that

> Every element of $E(2)$ maps straight lines to straight lines, and if A and A' are points on the lines l and l', there is a $g \in E(2)$ such that $g(l) = l'$ and $g(A) = A'$.

We now wish to do something similar in the case of $\mathbb{H}$. Motivated by the Euclidean example, we will look for a group G of bijections of $\mathbb{H}$ to itself such that

> Every element of G maps $\mathbb{H}$–lines to $\mathbb{H}$–lines, and if A and A' are points on the $\mathbb{H}$-lines l and l', there is a $g \in G$ such that $g(l) = l'$ and $g(A) = A'$.

We will show that there exists such a group, consisting of so–called *Möbius transformations* preserving the upper half–plane.

From complex function theory we know that meromorphic functions can be thought of as functions $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$, where $\overline{\mathbb{C}}$ is the *extended complex plane* or the *Riemann sphere* $\mathbb{C} \cup \{\infty\}$, with a topology such that stereographic projection extends to a homeomorphism $\overline{\Phi} : \mathbb{S}^2 \approx \overline{\mathbb{C}}$. Lemma 1.1 says that $\overline{\Phi}$ maps circles to curves in $\overline{\mathbb{C}}$ that are either circles in $\mathbb{C}$ or of the form $l \cup \{\infty\}$, where l is a (real) line in $\mathbb{C}$. It is convenient to call all of these curves $\overline{\mathbb{C}}$*–circles.*

By Lemma 1.2 the angle between oriented such circles at an intersection point in $\mathbb{C}$ is the same as between the corresponding circles on $\mathbb{S}^2$. Moreover, if they intersect in two points, the two angles will be the same. Hence we can also define the angle at ∞ between two circles intersecting there — i. e. two lines in $\mathbb{C}$. If they intersect in a point $P \in \mathbb{C}$, the angle of intersection at ∞ is the same as the angle at P. If they are parallel, the angle of intersection at ∞ is 0. With this definition, Φ is also angle preserving at ∞.

For a meromorphic function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ to be a homeomorphism, it must have exactly one pole and one zero — hence it must have the form

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

where $a, b, c, d \in \mathbb{C}$. Solving the equation $w = f(z)$ with respect to z we get

$$z = g(w) = \frac{dw-b}{-cw+a}\,,$$

and (formally) substituting back again:

$$f(g(w)) = \frac{(ad-bc)w}{ad-bc} \quad \text{and} \quad g(f(z)) = \frac{(ad-bc)z}{ad-bc}\,.$$

Therefore f is invertible with g as inverse if $ad-bc \neq 0$. If $ad-bc = 0$ these expressions have no meaning, but it is easy to see that in that case $f(z)$ is constant. Hence we have:

The function $f(z) = \dfrac{az+b}{cz+d}$ *defines a homeomorphism of* $\overline{\mathbb{C}}$ *if and only if* $ad-bc \neq 0$.

Such a function is called a *fractional linear transformations* — FLT for short. Here are some important properties of FLT's:

Lemma 2.1. (i) *An FLT maps* $\overline{\mathbb{C}}$*–circles to* $\overline{\mathbb{C}}$*–circles.*
(ii) *An FLT preserves angles between* $\overline{\mathbb{C}}$*–circles.*

Proof. (i) Note that we may write the equations of both circles and straight lines in $\mathbb{C}$ as $\lambda(x^2+y^2)+\alpha x+\beta y+\gamma = 0$, where $\lambda, \alpha, \beta, \gamma$ are real numbers and $z = x+iy$. $\lambda = 0$ gives a straight line and $\lambda \neq 0$ a circle. Using that $x^2+y^2 = z\bar{z}$, $x = (z+\bar{z})/2$ and $y = (z-\bar{z})/2i = (\bar{z}-z)i/2$, we can write the equation as

$$\lambda z\bar{z} + \mu z + \bar{\mu}\bar{z} + \gamma = 0\,,$$

where $\mu = (\alpha - i\beta)/2$. Hence we need to show that if z satisfies such an equation and $w = f(z)$ for an FLT f, then w satisfies a similar equation. This can be checked by writing $z = f^{-1}(w) = \dfrac{aw+b}{cw+d}$ and substituting:

$$\lambda\frac{aw+b}{cw+d}\cdot\frac{\overline{aw+b}}{\overline{cw+d}} + \mu\frac{aw+b}{cw+d} + \overline{\mu}\frac{\overline{aw+b}}{\overline{cw+d}} + \gamma = 0\,.$$

If we multiply this equation by $(cw+d)(\overline{cw+d})$ and simplify, we end up with an expression just like the one we want.

(ii) This is a consequence of a general fact in complex function theory. We say that a differentiable map is angle–preserving, or *conformal*, if it maps two intersecting curves to curves meeting at the same angle. It then follows from the geometric interpretation of the derivative that a complex function is conformal in a neighborhood of any point where it is analytic with nonzero derivative.

For a more direct argument in our case, see Exercise 2.3. □

Remark 2.2. As in Lemma 1.2 it is not difficult to show that the same result is true for oriented angles (with a suitable notion of orientation that also applies to $\infty \in \overline{\mathbb{C}}$), but we do not need that. An angle–preserving map (in

the orientable sense) is called *conformal*, and it follows from the geometric interpretation of the derivative that a complex function is conformal in a neighborhood of any point where it is analytic with nonzero derivative.

The oriented version of Lemma 1.2 says that stereographic projection also is conformal.

The word 'linear' in FLT is related to the following remarkable observation:

Let $f(z) = \dfrac{az+b}{cz+d}$ and $g(z) = \dfrac{a'z+b'}{c'z+d'}$. A little calculation gives

$$(f \circ g)(z) = f(g(z)) = \frac{a\,g(z)+b}{c\,g(z)+d} = \frac{(aa'+bc')z+(ab'+bd')}{(ca'+dc')z+(cb'+dd')}.$$

This formula tells us two things. Firstly, it means that the composition of two FLT's is a new FLT. We showed earlier that the inverse of an FLT is an FLT, and the identity map is trivially also an FLT. ($z = \frac{1z+0}{0z+1}$.) Hence the set of fractional linear transformations forms a *group* under composition. This group will be denoted $M\ddot{o}b^+(\mathbb{C})$. (The "even complex Möbius transformations".)

Secondly, it is possible to calculate with FLT's as with *matrices:* Evidently the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ determines f completely, and the condition $ad - bc \neq 0$ means precisely that this matrix is invertible. In the same way g is determined by $\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$, and the calculation above shows that $f \circ g$ is determined by the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$.

The set of invertible 2×2–matrices over $\mathbb{C}$ forms a group — *the general linear group* $GL_2(\mathbb{C})$ — and we have shown that there is a surjective group homomorphism from $GL_2(\mathbb{C})$ onto $M\ddot{o}b^+(\mathbb{C})$. This homomorphism is not injective, since if $k \neq 0$, then $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $\begin{bmatrix} ka & kb \\ kc & kd \end{bmatrix} = k \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ will determine the same map. However, this is the only ambiguity (Exercise 2.3), and we get an *isomorphism* between the group $M\ddot{o}b^+(\mathbb{C})$ of fractional linear transformations and the quotient group $PGL_2(\mathbb{C}) = GL_2(\mathbb{C})/D$ (*projective* linear group), where $D = \{uI \,|\, u \in \mathbb{C} - \{0\}\}$ and I is the identity matrix.

(A more conceptual explanation of the connection between FLT's and linear algebra belongs to *projective geometry* — see Exercise 2.11).

The next Lemma and its Corollary tell us exactly what freedom we have in prescribing values of fractional linear transformations. In fact, it provides us with a method of constructing FLT's with prescribed values.

Lemma 2.3. *Given three distinct points z_1, z_2 og z_3 i $\overline{\mathbb{C}}$. Then there exists a uniquely determined FLT f such that $f(z_1) = 1$, $f(z_2) = 0$ and $f(z_3) = \infty$.*

Proof. Existence: Suppose first that none of the z_i's is at ∞. Then we define

$$f(z) = \frac{z - z_2}{z - z_3} \cdot \frac{z_1 - z_3}{z_1 - z_2}.$$

In the three other cases:

$$\text{if } z_1 = \infty: \qquad f(z) = \frac{z - z_2}{z - z_3},$$

$$\text{if } z_2 = \infty: \qquad f(z) = \frac{z_1 - z_3}{z - z_3},$$

$$\text{if } z_3 = \infty: \qquad f(z) = \frac{z - z_2}{z_1 - z_2}.$$

Uniqueness: Suppose $g(z)$ has the same properties and consider the composition $h = g \circ f^{-1}$. This is a new FLT with $h(1) = 1$, $h(0) = 0$ and $h(\infty) = \infty$. The last condition implies that h must have the form $h(z) = az + b$, and the first two conditions then determine $a = 1$ og $b = 0$. Thus $(g \circ f^{-1})(z) = z$ for all z, hence $f = g$. □

Definition 2.4. The element $f(z) \in \overline{\mathbb{C}}$ depends on the four variables (z, z_1, z_2, z_3) and is denoted $[z, z_1, z_2, z_3]$. It is defined as an element of $\overline{\mathbb{C}}$ whenever z_1, z_2 and z_3 are distinct points of $\overline{\mathbb{C}}$ and has the following geometric interpretation:

If z_2 and z_3 span a Euclidean segment S, every point in S will divide it in two and we can compute the ratio between the lengths of the pieces. If we do this for two points z and z_1 in S, then $|[z, z_1, z_2, z_3]|$ is the quotient of the two ratios we obtain. Because of this, $[z, z_1, z_2, z_3]$ is traditionally called the *cross–ratio* of the four points, and it plays a very important role in geometry. We shall meet it again later, and some of its properties are given below, in Proposition 2.10.

Corollary 2.5. *Given two triples (z_1, z_2, z_3) og (w_1, w_2, w_3) of distinct points in $\overline{\mathbb{C}}$. Then there exists a unique FLT f such that $f(z_i) = w_i,\ i = 1, 2, 3$. If all six points lie in $\overline{\mathbb{R}}$, then f may be expressed with real coefficients — i. e. $f(z) = \dfrac{az + b}{cz + d}$ with a, b, c, d all real.*

Remark 2.6. The *existence* of such f means that the group $M\ddot{o}b^+(\mathbb{C})$ acts *transitively* on the set of such triples. The *uniqueness* says that if two fractional linear transformations have the same values at three points, then they are equal. In particular, an FLT fixing three points is the identity map.

Note that $f(z)$ is characterized by the equation

$$[f(z), w_1, w_2, w_3] = [z, z_1, z_2, z_3]\,.$$

Proof. By Lemma 2.3 we can find unique FLT's h and g such that $h(z_1) = 1,\ h(z_2) = 0,\ h(z_3) = \infty$, and $g(w_1) = 1,\ g(w_2) = 0\ g(w_3) = \infty$. Let $f = g^{-1}h$. Then $f(z_i) = w_i,\ i = 1, 2, 3$.

Suppose also f' maps z_i to w_i. Then gf and gf' are both FLT's as in Lemma 2.3, and because of the uniqueness $gf = gf'$. Consequently $f = f'$.

The final assertion of the Corollary follows from the formulae in the proof of 2.3. They show that h and g have real coefficients, hence so does $f = g^{-1}h$. □

Our next observation is that $M\ddot{o}b^+(\mathbb{C})$ also acts transitively on the set of $\overline{\mathbb{C}}$*–circles.* The reason for this is that three distinct points in $\overline{\mathbb{C}}$ determine a unique $\overline{\mathbb{C}}$–circle containing all of them.

Corollary 2.7. *Given two circles $\mathcal{C}_1$ and $\mathcal{C}_2$ in $\overline{\mathbb{C}}$. Then there exists a fractional linear transformation f such that $f(\mathcal{C}_1) = \mathcal{C}_2$.*

Proof. Choose three distinct points (z_1, z_2, z_3) on $\mathcal{C}_1$ and (w_1, w_2, w_3) on $\mathcal{C}_2$, and let f be as in the Corollary above. Then $f(\mathcal{C}_1)$ is a $\overline{\mathbb{C}}$–circle which contains w_1, w_2 and w_3 — i. e. $\mathcal{C}_2$. □

We now want to determine the fractional linear transformations f which restrict to homeomorphisms of the upper half–plane. Such an f is characterized by $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, and $\operatorname{Im} f(z) > 0$ if $\operatorname{Im} z > 0$. Here $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\} \subset \overline{\mathbb{C}}$.

Proposition 2.8. *A fractional linear transformation restricts to a homeomorphism of $\mathbb{H}$ if and only if it can be written on the form $f(z) = \dfrac{az+b}{cz+d}$, where a, b, c, d are real and $ad - bc = 1$.*

Such FLT's preserve $\mathbb{H}$–lines.

Proof. Corollary 2.5 says that if $f(z) = \dfrac{az+b}{cz+d}$ restricts to a homeomorphism of $\mathbb{H}$, then a, b, c, d can be chosen to be real, since $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$. Conversely, $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$ if a, b, c, d are real.

A short calculation gives

$$f(z) = \frac{(az+b)(c\bar{z}+d)}{(cz+d)(c\bar{z}+d)} = \frac{ac|z|^2 + (ad+bc)\operatorname{Re} z + bd}{|cz+d|^2} + \frac{(ad-bc)\operatorname{Im} z}{|cz+d|^2}\, i\,. \tag{2.1}$$

It follows that f preserves the upper half–plane if and only if $ad - bc > 0$. Hence, if we multiply a, b, c and d by $1/\sqrt{ad-bc}$, f is as asserted.

The last claim follows immediately from the fact that fractional linear transformations preserve $\overline{\mathbb{C}}$–circles and angles between them. Every $\mathbb{H}$–line determines a $\overline{\mathbb{C}}$–circle which meets $\mathbb{R}$ orthogonally, and since f preserves angles and $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, the images of these circles will also meet $\mathbb{R}$ orthogonally. □

The fractional linear transformations restricting to homeomorphisms of $\mathbb{H}$ form a subgroup of of $M\ddot{o}b^+(\mathbb{C})$ denoted $M\ddot{o}b^+(\mathbb{H})$. It can also be described using matrices, as follows:

Let $SL_2(\mathbb{R})$ be *the special linear group* — the group of real 2×2–matrices with determinant 1. The only multiples of the identity matrix in $SL_2(\mathbb{R})$ are $\pm I$, hence, arguing as before, we get an isomorphism between $M\ddot{o}b^+(\mathbb{H})$ and the quotient group $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/(\pm I)$.

$M\ddot{o}b^+(\mathbb{C})$ does not contain all circle–preserving homeomorphisms of $\overline{\mathbb{C}}$. *Complex conjugation* is also circle–preserving but not even complex analytic. (Define $\overline{\infty} = \infty$.) We define the group of *complex Möbius transformations,* $M\ddot{o}b(\mathbb{C})$, to be the group of homeomorphisms of $\overline{\mathbb{C}}$ generated by the fractional linear transformations and complex conjugation.

Proposition 2.9. (1) *Every complex Möbius transformation can be written on exactly one of the forms*

$$f(z) = \frac{az+b}{cz+d} \quad or \quad f(z) = \frac{a\bar{z}+b}{c\bar{z}+d}, \quad where \;\; a,\, b,\, c,\, d \in \mathbb{C} \text{ and } ad-bc = 1\,.$$

(2) *The complex Möbius transformations preserving* $\mathbb{H}$ *can be written either as*

(i) $f(z)=\dfrac{az+b}{cz+d}$, *where* $a,\, b,\, c,\, d \in \mathbb{R}$ *and* $ad-bc=1$, *or as*

(ii) $f(z)=\dfrac{a\bar{z}+b}{c\bar{z}+d}$, *where* $a,\, b,\, c,\, d \in \mathbb{R}$ *og* $ad-bc=-1$.

Proof. (1) Let S be the subset of the set of homeomorphisms of $\mathbb{C}$ given by such expressions. Clearly $S \subseteq \mathit{Möb}(\mathbb{C})$, and S contains $\mathit{Möb}^+(\mathbb{C})$ and complex conjugation. Therefore it suffices to check that S is closed under composition and taking inverses, and this is an easy calculation. Moreover, the second expression is not even complex differentiable, hence no function can be written both ways.

To obtain determinant 1 we divide numerator and denominator by a square root of $ad-bc$.

(2) $f(z)$ can be written in one of the two types in (1). In the first case, the result is given in Proposition 2.8. If $f(z) = \dfrac{a\bar{z}+b}{c\bar{z}+d}$, then $g(z) = -\overline{f(z)}$ can be written as in (i). But then $f(z) = -\overline{g(z)}$ automatically has the form given in (ii). □

The representations in Proposition 2.9 is not unique, but it follows from the result in Exercise 2.3 that it is unique up to multiplication of $(a,\, b,\, c,\, d)$ by ± 1.

Let $\mathit{Möb}(\mathbb{H})$ be the group of Möbius transformations restricting to homeomorphisms of $\mathbb{H}$. We have shown that every element in $\mathit{Möb}(\mathbb{H})$ can be written as one of the two types in (2) of Proposition 2.9, and therefore we call these elements the *real Möbius transformations.* Note that complex conjugation, i. e. reflection in the real axis, is not in $\mathit{Möb}(\mathbb{H})$, but $f(z) = -\bar{z}$, reflection in the imaginary axis, is. $\mathit{Möb}(\mathbb{H})$ is generated by this reflection and $\mathit{Möb}^+(\mathbb{H})$.

We use the notation $\mathit{Möb}^-(\mathbb{H})$ for the elements in $\mathit{Möb}(\mathbb{H})$ of type (ii). These do not form a subgroup, but $\mathit{Möb}(\mathbb{H})$ is the disjoint union of $\mathit{Möb}^+(\mathbb{H})$ and $\mathit{Möb}^-(\mathbb{H})$. In fact, $\mathit{Möb}^+(\mathbb{H}) \subset \mathit{Möb}(\mathbb{H})$ is a normal subgroup of index two, and $\mathit{Möb}^-(\mathbb{H})$ is the coset containing $-\bar{z}$.

$\mathit{Möb}(\mathbb{H})$ is the group we shall use to define congruence in $\mathbb{H}$, but before we show that Hilbert's axioms hold, we will analyze the elements in $\mathit{Möb}(\mathbb{H})$ closer and show that they can be classified into a few very simple standard types.

We end this section with some properties satisfied by the cross ratio.

Proposition 2.10. *Assume that* z, z_1, z_2 *and* z_3 *are four distinct points in* $\overline{\mathbb{C}}$, *and let* $\rho = [z, z_1, z_2, z_3]$. *Then*

(i) $[z_1, z, z_2, z_3] = [z, z_1, z_3, z_2] = \dfrac{1}{\rho}$, and $[z, z_2, z_1, z_3] = 1 - \rho$.

(ii) *z, z_1, z_2 and z_3 all lie on the same $\overline{\mathbb{C}}$–circle if and only if the cross–ratio $[z, z_1, z_2, z_3]$ is real.*

(iii) $[g(z), g(z_1), g(z_2), g(z_3)] = [z, z_1, z_2, z_3]$ *if g is a fractional linear transformation.*

Proof. (i) Recall that the mapping $w \mapsto f(w) = [w, z_1, z_2, z_3]$ is the fractional linear transformation which is uniquely determined by its values $1, 0$ and ∞ at the points z_1, z_2 and z_3, respectively. Then the identities $[w, z_1, z_3, z_2] = \dfrac{1}{f(w)}$ and $[w, z_2, z_1, z_3] = 1 - f(w)$ follow easily by inspection. Setting $w = z$ proves two of the identities.

Note that since $f(z_2) = 0$ and $f(z_3) = \infty$, ρ is not 0 or ∞. Therefore $g(w) = \dfrac{1}{\rho} f(w)$ defines a new fractional linear transformation. But then $g(z) = 1$, $g(z_2) = 0$ and $g(z_3) = \infty$ — hence $g(w) = [w, z, z_2, z_3]$. Consequently,

$$[z_1, z, z_2, z_3] = g(z_1) = \frac{1}{\rho}[z_1, z_1, z_2, z_3] = \frac{1}{\rho}\,.$$

(ii) Let $\mathcal{C}$ be the unique $\overline{\mathbb{C}}$–circle containing the three points z_1, z_2 and z_3. Then $f(\mathcal{C})$ must be the unique $\overline{\mathbb{C}}$–circle containing $1, 0$ and ∞ — i. e. $\overline{\mathbb{R}}$. Likewise, $f^{-1}(\overline{\mathbb{R}}) = \mathcal{C}$. Thus $z \in \mathcal{C}$ if and only if $f(z) \in \overline{\mathbb{R}}$. But since $z \neq z_3$, $f(z) \in \overline{\mathbb{R}}$ means $f(z) \in \mathbb{R}$.

(iii) Let $h(w) = [g(w), g(z_1), g(z_2), g(z_3)]$. This is a composition of two FLT's — hence h is also an FLT. By inspection, $h(z_j) = [z_j, z_1, z_2, z_3]$ for $j = 1, 2, 3$. Therefore $h(w) = [w, z_1, z_2, z_3]$ for all w, by uniqueness. □

Remark 2.11. (1) The three transpositions (1,2), (2,3) and (3,4) generate the whole group S_4 — the group of permutations of four letters. Hence (i) can be used to determine the cross ratio of any permutation of the points z, z_1, z_2 and z_3. For examples, see Exercise 2.8.

It follows that $[z, z_1, z_2, z_3]$ can be defined as long as *three* of the points z, z_1, z_2 and z_3 are distinct, and it can be considered as a fractional linear transformation in each of the variables separately. This observation will be used repeatedly later without any further comment.

(2) The identity in (iii) is not valid for *all* Möbius transformations g. For example, if $g(z) = \bar{z}$, then $[g(z), g(z_1), g(z_2), g(z_3)] = \overline{[z, z_1, z_2, z_3]}$. (Exercise 2.9.)

Exercises.

2.1. Discuss what conditions λ, μ, γ must satisfy for $\lambda z\bar{z} + \mu z + \bar{\mu}\bar{z} + \gamma = 0$ to define an $\mathbb{H}$–line.

2.2. Let the circle $\mathcal{C}$ be given by the equation $|z - z_0| = r$, and let f be the function $f(z) = 1/z$. When is $f(\mathcal{C})$ a straight line $\cup\{\infty\}$?

2.3. Show that any FLT can be written as a composition of maps of the following three simple types:

(i) Translations $z \mapsto z + b$, $b \in \mathbb{C}$,

(ii) Linear maps $z \mapsto kz,\ k \in \mathbb{C} - \{0\}$,

(iii) Taking inverse $z \mapsto \dfrac{1}{z}$.

Use this to give another proof of Lemma 2.1. (Hint: You may find Exercise 1.4a useful.)

2.4. Assume $ad - bc \neq 0$. Show that $\dfrac{az+b}{cz+d} = \dfrac{a'z+b'}{c'z+d'}$ for every z if and only if there exists a $k \neq 0$ such that $\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} = k \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

2.5. Use the method of Corollary 2.5 to find explicit fractional linear transformations mapping $\mathbb{H}$ onto $\mathbb{D}$ and vice versa. (Compare with Exercise 1.3.)

2.6. Describe all the elements in $M\ddot{o}b(\mathbb{H})$ that map the imaginary axis to itself.

2.7. (a) Show that $M\ddot{o}b(\mathbb{H})$ acts transitively on the set of all triples of distinct points in $\overline{\mathbb{R}}$. Deduce that $M\ddot{o}b(\mathbb{H})$ acts transitively on the set of pairs of lines in $\mathbb{H}$ with one common endpoint.

(b) Show that $M\ddot{o}b^+(\mathbb{H})$ acts transitively on the set of all *pairs* of points in $\overline{\mathbb{R}}$.

(b) Show that $M\ddot{o}b(\mathbb{H})$ does not act transitively on the set of pairs of points in $\mathbb{H}$.

2.8. Using Remark 2.11, show that if $[z_1, z_2, z_3, z_4] = \rho$, then $[z_3, z_4, z_1, z_2] = \rho$ and $[z_3, z_2, z_1, z_4] = \dfrac{\rho}{1-\rho}$.

2.9. Show that $[g(z), g(z_1), g(z_2), g(z_3)] = \overline{[z, z_1, z_2, z_3]}$ for all $g \in M\ddot{o}b^-(\mathbb{H})$.

2.10. Show that $M\ddot{o}b(\mathbb{H})$ is isomorphic to the group $PGL_2(\mathbb{R}) = GL_2(\mathbb{R})/D$, where $D = \{uI | u \in \mathbb{R} - \{0\}\}$.

2.11. Let $CP^1 = (\mathbb{C}^2 - \{0\})/\sim$, where $\sim$ is the equivalence relation which identifies v and λv, for all $v \in \mathbb{C}^2 - \{0\}$ and $\lambda \in \mathbb{C} - \{0\}$. ($CP^1$ is called *the complex projective line.*)

Show that multiplication by a matrix in $GL_2(\mathbb{C})$ induces a bijection of CP^1 with itself.

Verify that $(z_1, z_2) \mapsto z_1/z_2$ defines a bijection $CP^1 \approx \overline{\mathbb{C}}$, and show that via this bijection, multiplication with the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ corresponds to the fractional linear transformation $\dfrac{az+b}{cz+d}$.

Why does it now follow immediately that this correspondence is a group homomorphism $GL_2(\mathbb{C}) \to M\ddot{o}b^+(\mathbb{C})$?

3. Classification of real Möbius transformations

Since the Möbius tranformations play such an important rôle in the theory, we would like to know as much as possible about them, both geometrically and algebraically. The results of this section can be interpreted in

both directions. Geometrically we show that up to coordinate shifts, Möbius transformations can be given one of three possible "normal forms", from which it is easy to see how they act on $\mathbb{H}$. Algebraically this translates into a classification into conjugacy classes of $\mathit{Möb}(\mathbb{H})$.

Since $\mathit{Möb}(\mathbb{H})$ is isomorphic to a matrix group, this classification could be done completely with tools from linear algebra. What we will do is equivalent to this, but interpreted in our geometric language.

The key to the classification of matrices is the study of eigenvectors, and it is not difficult, using Exercise 2.11, to see that in $\mathit{Möb}(\mathbb{C})$ this corresponds to analyzing the *fixpoints* of the transformations, i. e. the solutions in $\overline{\mathbb{C}}$ of the equation $z = f(z)$. By a "change of coordinates" we reduce to a situation where the fixpoint set is particularly nice. Then we can more easily read off the properties of f.

Let us first consider the subgroup $\mathit{Möb}^+(\mathbb{H}) \subset \mathit{Möb}(\mathbb{H})$. We have seen that an element here can be written $f(z) = \dfrac{az+b}{cz+d}$, where $ad - bc = 1$. We assume from now on that f has this form and is not the identity.

Observe that as a map $\overline{\mathbb{C}} \to \overline{\mathbb{C}}$, f has ∞ as fixpoint if and only if $c = 0$. If also $a = d$, this is the only fixpoint — otherwise we have one more, namely $z = -b/(a-d)$, which is a real number. In other words, if $c = 0$ we have either one or two fixpoints, and they lie in $\overline{\mathbb{R}}$.

If $c \neq 0$, the equation $z = \dfrac{az+b}{cz+d}$ is equivalent to the equation

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

with roots

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

Using that $ad - bc = 1$, we can simplify the square root and write

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

We see that we should distinguish between three cases:

- $(a+d)^2 = 4$: Exactly one real root
- $(a+d)^2 > 4$: Two real roots
- $(a+d)^2 < 4$: Two complex roots

The number $a + d$ is the *trace* of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and it is invariant under conjugation by elements of $GL_2(\mathbb{C})$. The trace of $-\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $-(a+d)$, so it follows that the number

$$\tau(f) = (a+d)^2$$

is invariant under conjugation of f by elements in $\mathit{Möb}^+(\mathbb{H})$. In fact, it is also invariant under conjugation by $z \mapsto -\bar{z}$, hence invariant under conjugation by *every* element of $\mathit{Möb}(\mathbb{H})$.

Note that if $c = 0$, then $ad = 1$ and $(a+d)^2 = (a-d)^2 + 4 \geq 4$, with equality if and only if $a = d$. Taking into account the discussion above of the

case $c = 0$, we can distinguish between the following three cases, regardless of the value of c:

- Exactly one fixpoint in $\overline{\mathbb{R}}$, when $\tau(f) = 4$,
- Two fixpoints in $\overline{\mathbb{R}}$, when $\tau(f) > 4$,
- Two fixpoints in $\mathbb{C}$, when $\tau(f) < 4$.

Let us consider more closely each of the three cases:

Case (1): *One fixpoint in* $\overline{\mathbb{R}}$.

If $c = 0$, we may then write $f(z) = z + \beta$, i. e. f is a translation parallel to the x–axis. If $c \neq 0$, the unique fixpoint is $q = (a - d)/2c \in \mathbb{R}$, and we can find an $h \in M\ddot{o}b^+(\mathbb{H})$ mapping q to ∞. (Choose e. g. $h(z) = -1/(z-q)$.) Then the composition $g = h \circ f \circ h^{-1}$ is also an element of $M\ddot{o}b^+(\mathbb{H})$, and g has ∞ as unique fixpoint. Hence, as above, g has the form

$$g(z) = z + \gamma$$

for a real number γ. In this case we say that f is of *parabolic* type.

In fact, we can do even better: Choose a point $p \in \mathbb{R}$, $p \neq q$ and such that also $f(p) \in \mathbb{R}$. Then $f(p) \neq p$ and $f(p) \neq q$, since q is the only fixpoint. Define h by $h(z) = [z, f(p), p, q]$ (cross ratio) if this is in $M\ddot{o}b^+(\mathbb{H})$; otherwise set $h(z) = -[z, f(p), p, q]$. Then $h(p) = 0$, $h(q) = \infty$ and $h(f(p)) = 1$ or -1. The composition $g = h \circ f \circ h^{-1}$ still has ∞ as its only fixpoint, hence it has to be a translation $g(z) = z + b$. But then $b = g(0) = h(f(p)) = \pm 1$.

Hence any parabolic transformation is to a translation of the form $z+1$ or $z-1$. These two translations are conjugate in $M\ddot{o}b(\mathbb{H})$, but not in $M\ddot{o}b^+(\mathbb{H})$. (See Exercise 3.5.)

We think of such a conjugation as a change of coordinates: writing $f = h^{-1} \circ g \circ h$, we see that $f(z)$ is obtained by first moving z to $h(z)$, then applying g and finally moving back again by h^{-1}.

g fixes the point ∞ and translates horizontally all straight lines orthogonal to the real axis, i. e. $\mathbb{H}$–lines ending in ∞. Since h and h^{-1} both map $\mathbb{H}$–lines to $\mathbb{H}$–lines, we see that f must map $\mathbb{H}$–lines ending in q to $\mathbb{H}$–lines of the same type.

Figure 3 illustrates this in more detail. If g translates the vertical lines horizontally, it must also preserve the horizontal lines (dashed lines in the left figure). Mapped back by h^{-1} these become circles, but these circles are now *tangent* to the x–aksis, as in the figure to the right. It follows that f also must preserve such circles.

Remark 3.1. Such (Euclidean) circles, tangent to $\overline{\mathbb{R}}$ at a point p, we call *horocircles* at p. They can also be characterized by the property that they are orthogonal to all hyperbolic lines ending at the point of tangency (See Exercise 3.3.). Another characterization is discussed in Exercise 7.5.

Case (2): If f has *two* fixpoints in $\overline{\mathbb{R}}$, we say that it is of *hyperbolic* type. This happens precisely when $\tau(f) = (a + d)^2 > 4$.

Let $h \in M\ddot{o}b^+(\mathbb{H})$ be a real FLT mapping the two fixpoints to 0 and ∞. Then $g = h \circ f \circ h^{-1} \in M\ddot{o}b^+(\mathbb{H})$ has 0 og ∞ as fixpoints. It also maps the imaginary axis to an $\mathbb{H}$–line l, But since the end points are fixed, the end

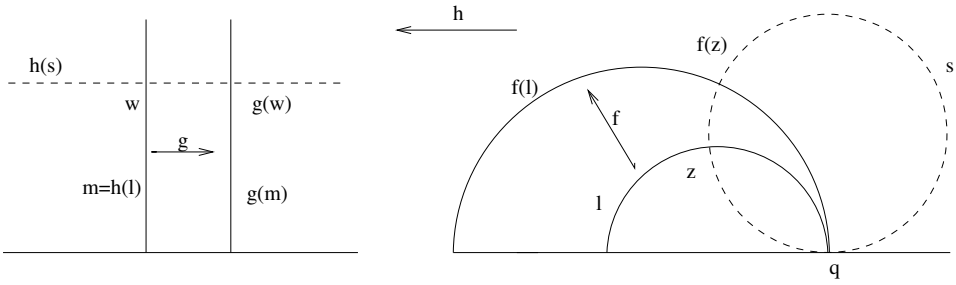

FIG. 3. Parabolic Möbius transformation

points of l are again 0 and ∞, so l must also be the imaginary axis. In particular there is a positive real number η such that $g(i) = \eta i$, and since we know that g is uniquely determined by the values at the three points 0, i og ∞, we see that

$$g(z) = \eta z$$

for all z. Thus, up to a change of coordinates, f is just multiplication by a real number.

Write $\eta = \lambda^2$, where we also may assume $\lambda > 0$. Since we must have $\tau(g) = \tau(f)$, λ satisfies the equation

$$\lambda + \frac{1}{\lambda} = |a + d| \,.$$

This equation has two roots λ and $1/\lambda$, and the corresponding functions $\lambda^2 z$ and z/λ^2 are conjugate by the transformation $z \mapsto -1/z$ — hence they are both realized by different choices of h. It follows that if we also choose $\eta > 1$, $g(z) = \eta z$ is uniquely determined. Moreover, this $g(z)$ is invariant under conjugation by $-\bar{z}$. Therefore there is a one–one correspondence between conjugacy classes of hyperbolic elements and real numbers > 1. (In both $M\ddot{o}b^+(\mathbb{H})$ and $M\ddot{o}b(\mathbb{H})$.)

Hyperbolic transformations behave as in figure 4. g preserves straight lines through 0, and these are mapped by h^{-1} to circular arcs or straight lines through the fixpoints of f, but the image of the imaginary axis is the only such curve meeting the x–axis orthogonally. Hence this is an $\mathbb{H}$–line between the two fixpoints of f, and it is mapped to itself by f. We call this $\mathbb{H}$–line the *axis* of f.

An element of $M\ddot{o}b^+(\mathbb{H})$ of hyperbolic type with axis l is often called a "translation along l".

Case (3): The final case is when we have *two complex fixpoints.* This happens when $\tau(f) < 4$, and we then say that f is of *elliptic* type. Since the two fixpoints are the roots of a real, quadratic equation, they are complex conjugate. In particular, there is exactly one in the upper half–plane and none in $\overline{\mathbb{R}}$. Let p be the fixpoint in $\mathbb{H}$ and set $h(z) = \dfrac{z - \operatorname{Re} p}{\operatorname{Im} p}$, such that $h \in M\ddot{o}b^+(\mathbb{H})$ and $h(p) = i$. This time $g = h \circ f \circ h^{-1}$ will be a real fractional linear transformation fixing i.

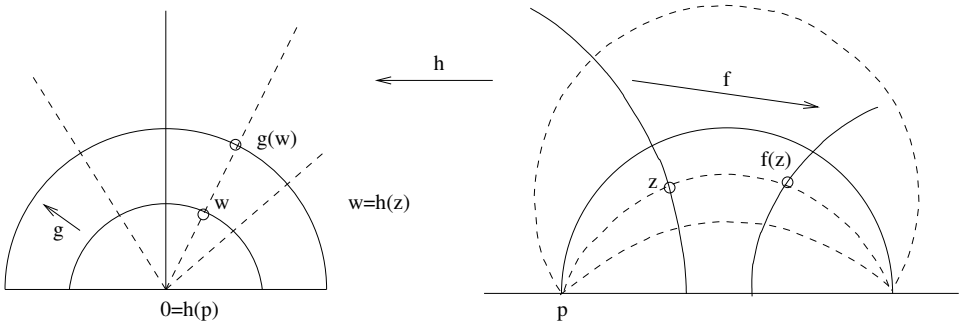

FIG. 4. Hyperbolic Möbius transformation

Let us analyze this g. Write g as $g(z) = \dfrac{\alpha z + \beta}{\gamma z + \delta}$, where α, β, γ, $\delta \in \mathbb{R}$. $g(i) = i$ means that $\alpha i + \beta = -\gamma + \delta i$ — i. e. $\alpha = \delta$ and $\beta = -\gamma$. If we substitute this into the equation $\alpha\delta - \beta\gamma = 1$, we see that $\alpha^2 + \beta^2 = 1$. Then we may write $\alpha = \cos(\theta)$, $\beta = \sin(\theta)$ for some real number θ, and we have

$$g(z) = g_\theta(z) = \frac{\cos(\theta)z + \sin(\theta)}{-\sin(\theta)z + \cos(\theta)}\,.$$

The matrix $\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$ describes a (clock-wise) rotation by the angle θ in $\mathbb{R}^2$. We can think of g also as a kind of rotation, since it keeps i fixed and maps the $\mathbb{H}$–lines through i to lines of the same type. Figure 5 shows the image of the imaginary axis. (Note that $g_\theta(0) = \tan\theta$.)

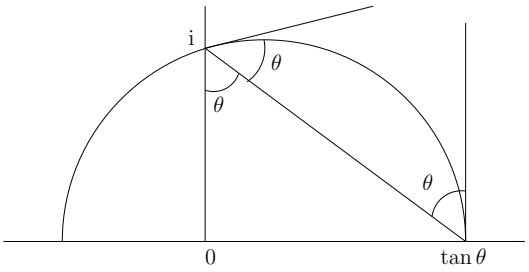

FIG. 5. Elliptic Möbius transformation

From the picture we see that the imaginary axis is rotated by an angle 2θ. But an easy calculation shows that $g_\theta g_{\theta'} = g_{\theta+\theta'}$ — hence g_θ will rotate *any* $\mathbb{H}$–line through i by an angle 2θ. In particular, g_π =id, and $g_{\theta+\pi} = g_\theta$.

Later, after we have introduced Poincaré's disk–model, the analogy with Euclidean rotations will become even clearer. See also Exercise 3.6.

If g_θ is conjugate to $g_{\theta'}$ in $M\ddot{o}b^+(\mathbb{H})$, then $g_\theta = g_{\theta'}$. The reason for this is that if $h^{-1}g_\theta h$ has i as fixpoint, then g_θ has $h(i)$ as fixpoint — hence $h(i) = i$. But then $h = g_\phi$ for some ϕ; thus g_θ and h commute. It follows that there is a one–one correspondence between conjugacy classes in $M\ddot{o}b^+(\mathbb{H})$ of elliptic elements and angles $\theta \in (0, \pi)$.

On the other hand, if $h(z) = -\bar{z}$, then $h^{-1}g_\theta h = g_{-\theta} = g_{\pi-\theta}$, so the conjugacy classes in $M\ddot{o}b(\mathbb{H})$ are in one–one correspondence with $\theta \in (0, \pi/2)$.

Let us sum up what we have done so far:

Proposition 3.2. *Suppose* $f(z) = \dfrac{az+b}{cz+d}$, *where* a, b, c, $d \in \mathbb{R}$ *and* $ad - bc = 1$, *and let* $\tau(f) = (a+d)^2$. *Then, as an element of* $M\ddot{o}b^+(\mathbb{H})$, f *is of*

- *Parabolic type, conjugate to* $z \mapsto z+1$ *or* $z-1$, *if* $\tau(f) = 4$,
- *Hyperbolic type, conjugate to exactly one* $z \mapsto \eta z$ *with* $\eta > 1$, *if* $\tau(f) > 4$,
- *Elliptic type, conjugate to a unique* $z \mapsto g_\theta = \dfrac{\cos(\theta)z + \sin(\theta)}{-\sin(\theta)z + \cos(\theta)}$ *with* $\theta \in (0, \pi)$, *if* $\tau(f) < 4$.

In $M\ddot{o}b(\mathbb{H})$ the only differences are that there is only one conjugacy class of elements of parabolic type (see Exercise 3.5), and g_θ is conjugate to $g_{\pi-\theta}$.

We say that an element of $M\ddot{o}b^+(\mathbb{H})$ is given on *normal form* if it is written as a conjugate of one of these standard representatives.

We now move on to $M\ddot{o}b^-(\mathbb{H})$ and consider a transformation of the form $f(z) = \dfrac{a\bar{z}+b}{c\bar{z}+d}$, where a, b, c, $d \in \mathbb{R}$ and $ad - bc = -1$.

Again we will look for fixpoints of $f(z)$. As before, $z = \infty$ is a fixpoint if and only if $c = 0$. If $z \neq \infty$, the equation $f(z) = z$ is now equivalent to $c|z|^2 + dz - a\bar{z} - b = 0$, or

$$c(x^2 + y^2) - (a-d)x - b = 0, \tag{3.1}$$

$$(a+d)y = 0. \tag{3.2}$$

We consider the two cases $a + d = 0$ and $a + d \neq 0$ separately.

First, let $a + d = 0$. In this case equation (3.2) is trivially satisfied, so we have only one equation (3.1). This describes an $\mathbb{H}$–line: the vertical line $x = \dfrac{b}{d-a}$ when $c = 0$ and the semi–circle with center $(a/c, 0)$ and radius $1/|c|$ if $c \neq 0$. (Note that $a \neq d$ if $c = 0$, since then $ad = -1$.) Hence f fixes an entire $\mathbb{H}$–line and interchanges the two components of its complement in $\mathbb{H}$. More precisely, by a suitable conjugation as above, we may assume that the fixed H–line is the imaginary axis. Then $c = b = 0$ and $a = -d = \pm 1$ — hence $f(z) = -\bar{z}$. Thus all such transformations are conjugate to the horizontal reflection in the imaginary axis.

If the fixpoint set is another vertical line l, we can choose h to be a horizontal translation. Hence f must be horizontal reflection in l.

If $c \neq 0$ we can also write

$$f(z) = \frac{a}{c} + \frac{1/c}{c\bar{z}+d} = \frac{a}{c} + \frac{1/c^2}{\bar{z} - a/c}. \tag{3.3}$$

This has the general form

$$g(z) = m + \frac{r^2}{\overline{z-m}} = m + r^2\frac{z-m}{|z-m|^2}\,.$$

If $\mathcal{C}$ is the circle with center m and radius r, $g(z)$ maps points outside $\mathcal{C}$ to points inside and vice versa, and it leaves the circle itself fixed. More precisely, we see that $g(z)$ lies on the (Euclidean) ray from m through z, and such that the product $|w-m||z-m|$ is equal to r^2. This is a very important geometric construction called "inversion in the circle $\mathcal{C}$".

By analogy we will also call the horizontal reflection in a vertical line l "inversion in l". Thus inversions are precisely the transformations in $M\ddot{o}b^-(\mathbb{H})$ such that $a+d=0$, and all inversions are conjugate. Note that there are two particularly simple representatives for this conjugacy class: the horizontal reflection $z \mapsto -\overline{z}$ and the map $z \mapsto 1/\overline{z}$, which is inversion in the circle $|z|=1$.

Next, assume $a+d \neq 0$. Then $y=0$ by (3.2), so there are no fixpoints in $\mathbb{H}$. In equation (3.1) we distinguish between the cases $c=0$ and $c \neq 0$.

If $c=0$, we get $x = b/(d-a)$, but then we also have the fixpoint (in $\overline{\mathbb{C}}$) $z=\infty$, so f must map the vertical line $x=b/(d-a)$ to itself. (But without fixpoints.)

If $c \neq 0$, (3.1) has two finite solutions $x = \dfrac{a-d \pm \sqrt{(a+d)^2+4}}{2c}$. Hence $f(z)$ has two fixpoints on the real axis, and f must map the $\mathbb{H}$–line with these two points as endpoints to itself.

Thus, in both cases f preserves an $\mathbb{H}$–line l, and conjugating with a transformation mapping l to the imaginary axis, we obtain a function of the form $g(z) = -k^2\overline{z}$ — a composition of the reflection (inversion) in the y–axis and a hyperbolic transformation with the same axis.

Note that as $k^2(-\overline{z}) = -\overline{(k^2 z)}$, these two transformations *commute.* Conjugating back, we see that we have written f as a composition of two commuting transformations — a hyperbolic transformation h and an inversion in the axis of h.

Proposition 3.3. *Let $f \in M\ddot{o}b^-(\mathbb{H})$ have the form $f(z) = \dfrac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, $d \in \mathbb{R}$ and $ad-bc=-1$.*

- *If $a+b=0$, then f an inversion, conjugate to reflection in the imaginary axis.*
- *If $a+b \neq 0$, f can be written $f=gh$, where g is an inversion and g and h commute. Moreover, this decomposition is unigue and h is of hyperbolic type.*

Proof. It only remains to prove the uniqueness statement. So, suppose $f = gh$ where g is inversion in a line ℓ. If $z \in \ell$ we have

$$g(h(z)) = h(g(z) = h(z)\,,$$

i. e. $h(z)$ is a fixpoint for g. Hence $h(z) \in \ell$. It follows that

$$f(\ell) = h(g(\ell)) = h(\ell) = \ell\,.$$

Since f has no fixpoint in $\mathbb{H}$, it must fix the two endpoints of ℓ. This means that f determines ℓ, hence also g. It is now clear that g and h are uniquely determined as the two transformations constructed above. □

It is worth pointing out that if we do not require that the components commute, there are many ways of decomposing an element of $M\ddot{o}b^-(\mathbb{H})$ into a product of an inversion and an element of $M\ddot{o}b^+(\mathbb{H})$. Trivial such decompositions are given by the formulae

$$\frac{a\overline{z}+b}{c\overline{z}+d} = \frac{(-a)(-\overline{z})+b}{(-c)(-\overline{z})+d} = -\overline{\left(\frac{(-a)z+(-b)}{cz+d}\right)}.$$

More interesting, perhaps; if $c \neq 0$, we can generalize (3.3) and write (using $ad-bc=-1$):

$$\frac{a\overline{z}+b}{c\overline{z}+d} = \frac{a}{c} + \frac{1/c}{c\overline{z}+d} = \frac{a+d}{c} + \left(-\frac{d}{c} + \frac{1}{c^2}\frac{z-(-d/c)}{|z-(-d/c)|^2}\right). \tag{3.4}$$

This is a composition of an inversion and a *parabolic* transformation. For more on decompositions of Möbius transformations, see excercises 3.9 and 3.10.

Note the following, which is implicit in what we have done:

- An element in $M\ddot{o}b^-(\mathbb{H})$ is an inversion if and only if its trace $a+d$ is 0. (This condition is independent of whether we have normalized the coefficients or not.)
- An inversion in an $\mathbb{H}$–line l is characterized, as an element of $M\ddot{o}b(\mathbb{H})$, by having all of l as fixpoint set.

Exercises.

3.1. Classify the following maps and write them explicitly as conjugates of mappings on normal form.

$$\frac{4z-3}{2z-1}, \qquad -\frac{1}{z-1}, \qquad \frac{z}{z+1}.$$

3.2. Discuss the classification of Möbius transformations in terms of matrix representations, without assuming determinant 1.

3.3. Show geometrically that the horocircles at a point $p \in \overline{\mathbb{R}}$ are orthogonal to all hyperbolic lines with p as one endpoint.

3.4. Explain what a hyperbolic transformation f does to the horocircles at the endpoints of the axis of f, and also to the other $\mathbb{H}$–linjes sharing the same endpoint.

3.5. Show that all parabolic transformations are conjugate in $M\ddot{o}b(\mathbb{H})$. Show that the translations $z \mapsto z+1$ and $z \mapsto z-1$ are *not* conjugate in $M\ddot{o}b^+(\mathbb{H})$.

3.6. Fix a z in $\mathbb{H}$, $z \neq 0$. Show that as θ varies, the points $g_\theta(z)$ all lie on the same circle in $\mathbb{C}$.

(Hint: if $\cos\theta \neq 0$,write $g_\theta(z) = \dfrac{\tan\theta + z}{-\tan\theta\, z + 1}$, and think of this as a function of $\tan\theta$.)

3.7. Show that $f \in \mathit{Möb}^-(\mathbb{H})$ is an inversion if and only if it has a fixpoint in $\mathbb{H}$.

3.8. Show that an inversion in a circle $\mathcal{C} \subset \mathbb{C}$, considered as a map on $\mathbb{C}$ minus the center of $\mathcal{C}$, has the following properties:

(a) It maps straight lines outside $\mathcal{C}$ to circles inside $\mathcal{C}$ and through its center.
(b) Circles intersecting $\mathcal{C}$ ortogonally are mapped to themselves.

(These are important results about inversions that are usually proved by geometric arguments. Her they should follow quite easily from what we now know about Möbius transformations.)

3.9. Show that every element in $\mathit{Möb}^+(\mathbb{H})$ may be written as the composition of two inversions.

3.10. Show that $\mathit{Möb}(\mathbb{H})$ is generated by inversions, and show that $\mathit{Möb}^+(\mathbb{H})$ ($\mathit{Möb}^-(\mathbb{H})$) consists of those elements that can be written as a composition of an even (odd) number of inversions.

4. Hilbert's axioms and congruence in $\mathbb{H}$

We are now ready to prove that the upper half–plane provides a model for that hyperbolic plane, with congruence based on the action of $\mathit{Möb}(\mathbb{H})$.

Recall that, using a combination of orthogonal and stereographic projections, we have identified the open unit disk $\mathbb{K} \subset \mathbb{R}^2$ with the upper half–plane $\mathbb{H} \subset \mathbb{C}$, such that chords in $\mathbb{K}$ correspond to what we have called $\mathbb{H}$–lines — vertical lines or semicircles with center on the real axis in $\mathbb{C}$. $\mathbb{K}$ inherits incidence and betweenness relations from $\mathbb{R}^2$, hence we obtain corresponding relations in $\mathbb{H}$. Automatically all of Hilbert's axioms I1–3 and B1–4 for these relations hold, as does Dedekind's axiom. In this section we introduce a congruence relation and show that it satisfies Hilbert's axioms C1–6. Since, as we have observed, the parallel axiom is replaced by the hyperbolic axiom, we will then have completed the construction of a hyperbolic geometry.

Remark 4.1. Betweenness for points on a line in the Euclidean plane can be formulated via homeomorphisms between the line and $\mathbb{R}$ or intervals in $\mathbb{R}$, hence the same is true for $\mathbb{H}$–lines, if we use the subspace topology from $\mathbb{C}$. On $\mathbb{R}$ the easiest definition is:

$$a * b * c \iff a < b < \quad \text{or} \quad a > b > c\,.$$

(Equivalently: $(a-b)(b-c) > 0$.) The simplest such homeomorphisms are projections to the imaginary axis from the vertical lines and to the real axis from the half–circles. It follows that betweenness for points on an $\mathbb{H}$–line ℓ can be characterized by

- $x * y * z \iff \operatorname{Im} x * \operatorname{Im} y * \operatorname{Im} z$ if ℓ is vertical,
- $x * y * z \iff \operatorname{Re} x * \operatorname{Re} y * \operatorname{Re} z$ otherwise.

Before we go on, we need a more precise notation for lines, rays etc. If z_1, z_2 are two points of $\mathbb{H}$, we write $\overleftrightarrow{z_1z_2}$ for the uniquely determined hyperbolic line containing them, $\overrightarrow{z_1z_2}$ for the ray from z_1 containing z_2 and $[z_1, z_2]$ for the segment between z_1 and z_2 — i. e. $[z_1, z_2] = \overrightarrow{z_1z_2} \cap \overrightarrow{z_2z_1}$. An $\mathbb{H}$–line l is uniquely determined by its endpoints p and q in $\overline{\mathbb{R}}$, and therefore we will also write $l = (p, q)$. With this notation, the identity $\overleftrightarrow{z_1z_2} = (p, q)$ will mean that the uniquely determined $\mathbb{H}$–line containing z_1 and z_2 has endpoints p and q. Similarly, we may also write $[z_1, q) = \overrightarrow{z_1q} = \overrightarrow{z_1z_2}$, expressing that q is the endpoint of the ray $\overrightarrow{z_1z_2}$.

We say that z_1 is the *vertex* and q the *endpoint* of the ray $[z, q)$. An *angle* is then an unordered pair of rays with the same vertex, where the two rays do not lie on the same line. We use the notation $\angle uzv$ for the unordered pair $\{\overrightarrow{zu}, \overrightarrow{zv}\}$, where $z \in \mathbb{H}$ and u, v are either in $\mathbb{H}$ or in $\overline{\mathbb{R}}$.

The congruence relation in $\mathbb{H}$ is now defined as follows:

Congruence of segments: $[z_1, z_2] \cong [w_1, w_2] \iff g([z_1, z_2]) = [w_1, w_2]$ for some $g \in \mathit{Möb}(\mathbb{H})$.

Congruence of angles: $\angle uzv \cong \angle u'z'v' \iff g(\overrightarrow{zu}) = \overrightarrow{z'u'}$ and $g(\overrightarrow{zv}) = \overrightarrow{z'v'}$ for some $g \in \mathit{Möb}(\mathbb{H})$. (Notation: $g(\angle uzv) = \angle u'z'v'$.)

The existence parts of the congruence statements say that there are enough Möbius transformations to move angles and segments freely around in $\mathbb{H}$, whereas the uniqueness means that there are not too many such transformations. The technical results we need are contained in the following Lemmas:

Lemma 4.2. *Suppose z_j lies on an $\mathbb{H}$–line l_j, with endpoints p_j and q_j, for $j = 1, 2$. Then there is a uniquely determined $f \in \mathit{Möb}^+(\mathbb{H})$ such that $f(p_1) = p_2$, $f(q_1) = q_2$ and $f(z_1) = z_2$ — hence also $f(l_1) = l_2$.*

In particular we have, for example, $f([z_1, q_1)) = [z_2, q_2)$. But since a ray determines the line containing it, we get

Corollary 4.3. *$\mathit{Möb}^+(\mathbb{H})$ acts transitively on the set of all rays: In fact, given two rays σ_1 and σ_2 with vertices z_1 and z_2, there is a unique $f \in \mathit{Möb}^+(\mathbb{H})$ such that $f(z_1) = z_2$ and $f(\sigma_1) = \sigma_2$.*

Lemma 4.4. *(i) An element in $\mathit{Möb}^+(\mathbb{H})$ is completely determined by its values at two points in $\mathbb{H}$.*

(ii) Suppose the segments $[z_1, z_2]$ and $[w_1, w_2]$ are congruent. Then there is a uniquely determined $f \in \mathit{Möb}^+(\mathbb{H})$ such that $f(z_1) = w_1$ and $f(z_2) = w_2$.

Lemma 4.5. *Given two rays σ_1 and σ_2 with a common vertex z_0. Then there is a unique inversion g such that $g(z_0) = z_0$, $g(\sigma_1) = \sigma_2$ and $g(\sigma_2) = \sigma_1$.*

Proof of Lemma 4.2. By Exercise 2.7b there exists a $g \in \mathit{Möb}^+(\mathbb{H})$ such that $g(p_1) = p_2$ and $g(q_1) = q_2$. Then automatically $g(l_1) = l_2$. Now let h be a hyperbolic transformation with axis l_2 such that $h(g(z_1)) = z_2$. Then $f = hg$ satisfies all the required properties.

Uniqueness follows from Corollary 2.5. □

Proof of Lemma 4.4. (i) Two points in $\mathbb{H}$ determine a unique line l, and the endpoints of l must map to the endpoints of $f(l)$, in such a way that betweenness relations are preserved. Hence the values of f at *four* points are determined, and the uniqueness follows from uniqueness in Corollary 2.5.

(ii) Assume that $g([z_1, z_2]) = [w_1, w_2]$ for some $g \in \mathit{Möb}(\mathbb{H})$. If $g \in \mathit{Möb}^-(\mathbb{H})$, we replace g by $k \circ g$, where k is the inversion in the $\mathbb{H}$–line $\overleftrightarrow{w_1 w_2}$. Therefore we may assume that $g \in \mathit{Möb}^+(\mathbb{H})$.

The problem is that we might have $g(z_1) = w_2$ and $g(z_2) = w_1$. If so, choose an $h \in \mathit{Möb}^+(\mathbb{H})$ such that $h(\overleftrightarrow{w_1 w_2})$ is the imaginary axis, and write $h(w_1) = \omega_1 i$, $h(w_2) = \omega_2 i$. If we define $k(z) = -\omega_1\omega_2/z$, we see that k interchanges $\omega_1 i$ and $\omega_2 i$. Then $h^{-1}kh$ will interchange w_1 og w_2, and we let $f = h^{-1}khg$. □

Proof of Lemma 4.5. It follows easily from Lemma 4.2 that we can find an $h \in \mathit{Möb}^+(\mathbb{H})$ mapping σ_1 to σ_2. (h has z_0 as fixpoint and must necessarily be elliptic.) Let g' be inversion in the line containing σ_1 and define $g = hg'$. Then $g \in \mathit{Möb}^-(\mathbb{H})$ and has a fixpoint $z_0 \in \mathbb{H}$, hence is an inversion. Clearly $g(\sigma_1) = h(\sigma_1) = \sigma_2$.

Uniqueness: Suppose g' is another inversion with the same properties. Then $g^{-1}g'$ is an element of $\mathit{Möb}^+(\mathbb{H})$ mapping both σ_1 and σ_2 to themselves. Therefore it has three fixpoints z_0, q_1 and q_2 (in $\overline{\mathbb{C}}$), hence it must be the identity map. Thus $g = g'$. □

We are now ready to prove that Hilbert's congruence axioms C1–6 are satisfied. The axioms are:

The axioms for congruence of segments:

C1: Given a segment $[z_1, z_2]$ and a ray σ with vertex w_1, there is a uniquely determined point w_2 on σ such that $[w_1, w_2] \cong [z_1, z_2]$.

C2: $\cong$ is an equivalence relation on the set of segments.

C3: If $z_1 * z_2 * z_3$ and $w_1 * w_2 * w_3$ and both $[z_1, z_2] \cong [w_1, w_2]$ and $[z_2, z_3] \cong [w_2, w_3]$, then also $[z_1, z_3] \cong [w_1, w_3]$.

The axioms for congruence of angles:

C4: Given a ray $[w, q)$ and an angle $\angle uzv$, there are unique angles $\angle p_1wq$ and $\angle p_2wq$ on opposite sides of $[w, q)$ such that $\angle p_1wq \cong \angle p_2wq \cong \angle uzv$.

C5: $\cong$ is an equivalence relation on the set of angles.

C6: (SAS) Given triangles $z_1z_2z_3$ og $w_1w_2w_3$. If $[z_1, z_2] \cong [w_1, w_2]$, $[z_1, z_3] \cong [w_1, w_3]$ and $\angle z_2z_1z_3 \cong \angle w_2w_1w_3$, then the two triangles are congruent — i. e. we also have $[z_2, z_3] \cong [w_2, w_3]$, $\angle z_1z_2z_3 \cong \angle w_1w_2w_3$ and $\angle z_2z_3z_1 \cong \angle w_2w_3w_1$.

C2 and **C5** follow immediately since we have defined congruence by a group action.

C1. The segment $[z_1, z_2]$ defines a ray $\overrightarrow{z_1 z_2}$, and by Corollary 4.3 there exists an $f \in \mathit{Möb}^+(\mathbb{H})$ such that $f(z_1) = w_1$ and $f(\overrightarrow{z_1 z_2}) = \sigma$. If we put $w_2 = f(z_2)$, we clearly get $[w_1, w_2] \cong [z_1, z_2]$.

Suppose that w_2' is another point on σ such that $[w_1, w_2'] \cong [z_1, z_2]$. By Lemma 4.4 we can find $h \in \mathit{Möb}^+(\mathbb{H})$ such that $h(z_1) = w_1$ and $h(z_2) = w_2'$. But then also $h(\overrightarrow{z_1 z_2}) = \overrightarrow{w_1 w_2'} = \sigma$, and by the uniqueness in Corollary 4.3 we must have $h = f$. Therefore $w_2 = f(z_2) = h(z_2) = w_2'$.

C3. Here we use Lemma 4.4(ii), saying that there exists a $g \in \mathit{Möb}^+(\mathbb{H})$ such that $g(z_1) = w_1$ and $g(z_2) = w_2$. Then $w_3' = g(z_3)$ is on the ray $\overrightarrow{w_2 w_3}$, and g defines congruences $[z_2, z_3] \cong [w_2, w_3']$ and $[z_1, z_3] \cong [w_1, w_3']$. But by the uniqueness in C1 we must then have $w_3' = w_3$.

C4. We may assume that u and v are the endpoints of the rays in $\angle uzv$. To construct the angles is easy: let $g \in \mathit{Möb}^+(\mathbb{H})$ be such that $g([z, u)) = [w, q)$, and define $\angle p_1 wq$ as $g(\angle uzv)$. Let $j \in \mathit{Möb}^-(\mathbb{H})$ be inversion in the $\mathbb{H}$–line l containing $[w, q)$, and let $\angle p_2 wq = jg(\angle uzv) = j(\angle p_1 wq)$. Since j interchanges the two sides of l, then $\angle p_1 wq$ and $\angle p_2 wq$ must lie on opposite sides of $[w, q)$.

It remains to prove uniqueness. Suppose $\angle pwq = h(\angle uzv)$, where $h \in \mathit{Möb}(\mathbb{H})$. This means that $h([z, u))$ is either $[w, q)$ or $[w, p)$, and using Lemma 4.5, we may, if necessary, compose h with an inversion interchanging $[w, q)$ and $[w, p)$ — hence we may assume $h([z, u)) = [w, q)$. If $h \in \mathit{Möb}^+(\mathbb{H})$, then $h = g$ because of the uniqueness in Corollary 4.3 once again, and $[w, p) = [w, p_1)$. If $h \in \mathit{Möb}^-(\mathbb{H})$, then $jh \in \mathit{Möb}^+(\mathbb{H})$, and since $jh([z, u)) = j([w, q)) = [w, q)$, $jh = g$. Therefore $h = jg$, and $[w, p) = [w, p_2)$.

C6. By assumption we have $\angle w_2 w_1 w_3 = g(\angle z_2 z_1 z_3)$, for some $g \in \mathit{Möb}(\mathbb{H})$, and after applying Lemma 4.5, if necessary, we may assume that $g(\overrightarrow{z_1 z_2}) = \overrightarrow{w_1 w_2}$ and $g(\overrightarrow{z_1 z_3}) = \overrightarrow{w_1 w_3}$. But uniqueness in C1 then means that $g(z_2) = w_2$ and $g(z_3) = w_3$. Hence it follows that also $g([z_2, z_3]) = [w_2, w_3]$, $g(\angle z_1 z_2 z_3) = \angle w_1 w_2 w_3$ and $g(\angle z_2 z_3 z_1) = \angle w_2 w_3 w_1$.

□

Remark 4.6. Except for **C6**, we could have defined congruence using the smaller group $\mathit{Möb}^+(\mathbb{H})$. The remaining axioms would still hold.

5. Distance in $\mathbb{H}$

Now that we have established the existence of a model for hyperbolic geometry based on the upper half–plane $\mathbb{H}$, it is time to start investigating the geometric structure itself. Classical geometry has a rich theory of triangles and circles, we can measure distances and angles, and there is a trigonometric theory relating them. Of great practical importance are formulae for arc lengths and area. To what extent can we do the same in hyperbolic geometry? Many classical geometric arguments do not use the parallel axiom, and these will automatically be valid in hyperbolic geometry, as well. So, which results carry over and which do not? And, what can we say when they do not?

Naturally, we can only scratch the surface here, but in the next sections we shall develop some of the basic theory. Enough, hopefully, to give a feeling for what the hyperbolic world looks like.

We start with the fundamental concept of *distance*, and we will show that there is a distance function in $\mathbb{H}$ which characterizes congruence, just as in Euclidean geometry. Thus, we want to define a *metric* on $\mathbb{H}$ — i. e. a function $d : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ such that

(d1) $d(z, w) \geq 0$, *and* $d(z, w) = 0$ *if and only if* $z = w$,
(d2) $d(z, w) = d(w, z)$ *for all* $z, w \in \mathbb{H}$,
(d3) $d(z, w) \leq d(z, u) + d(u, w)$ *for all* $z, u, w \in \mathbb{H}$.

The desired relation with geometry leads us to we require that it also should have the following properties:

(d4) *If* z, u, w *are distinct points in* $\mathbb{H}$, *then* $d(z, w) = d(z, u) + d(u, w)$ *if and only if* $u \in [z, w]$. *("Distance is measured along* $\mathbb{H}$*–lines".)*

(d5) $d(z, w) = d(z', w')$ *if and only if there exists a* $g \in \mathit{Möb}(\mathbb{H})$ *such that* $g(z) = z'$ *and* $g(w) = w'$.
("Two segments are congruent if and only if they have the same lengths".)

We first observe that (d4) and (d5) determine the metric almost completely, if it exists. Given z, w, there is a unique $g \in \mathit{Möb}^+(\mathbb{H})$ such that $g(z) = i$ and $g(w) = ti$, where $t \geq 1$. (This is Lemma 4.2 applied to the point i on the imaginary axis and the point z on the $\mathbb{H}$–line through z and w. If $g(w)$ has imaginary part less than 1, we can replace g by $k \circ g$, where $k(z) = -1/z$.) Then, because of (d5), we must have $d(z, w) = d(i, ti)$, hence d is completely determined by the function $f : [1, \infty) \to [0, \infty)$ defined by $f(t) = d(i, ti)$.

We can say more about the function f. Let s and t be two numbers greater than or equal to 1. Then $si \in [i, sti]$ — hence from (d4) (and (d1), if s or t is 1) we have:

$$f(st) = d(i, sti) = d(i, si) + d(si, sti) = d(i, si) + d(i, ti) = f(s) + f(t)\,.$$

The third inequality follows from (d5) applied to $h(z) = sz$ and $z = i$.) But then one can show that $f(t) = C \ln(t)$ for some real (positive) constant C.

(See Exercise 5.1 for the case when f is differentiable in one point. But it is easy to see that f must be *increasing,* and a famous Theorem due to Lebesgue says that an increasing function is differentiable almost everywhere.)

The choice of constant C just amounts to a scaling, and any positive number could be used in the following. We choose to set $C = 1$. If ti and si are two points on the imaginary axis with $s < t$, we then have

$$d(si, ti) = d(i, \frac{t}{s}i) = \ln(\frac{t}{s})\,.$$

But then also

$$d(ti, si) = d(si, ti) = \ln(\frac{t}{s}) = -\ln(\frac{s}{t})\,,$$

hence $d(si, ti)$ is equal to $|\ln(\frac{t}{s})|$ for any two points si and ti on the imaginary axis. In particular, $d(i, ti) = |\ln(t)|$ for *every* $t > 0$.

This leads to the following formula for the metric d if $z \neq w$:

$$d(z,w) = |\ln(|\frac{g(z)}{g(w)}|)| = |\ln(|\frac{i}{g(w)}|)| = |\ln(|g(w)|)|\,,$$

where $g \in M\ddot{o}b^+(\mathbb{H})$ maps z to i and w to another point on the imaginary axis.

From the proof of Lemma 4.2 we can give an explicit expression for the transformation g. If p and q are the endpoints of the unique hyperbolic line through z and w and such that z is between p and w, we can set

$$g(w) = [w,z,p,q]\,i,$$

such that

$$d(z,w) = |\ln|[w,z,p,q]||\,. \tag{5.1}$$

By Proposition 2.10(i) this is independent of the ordering of p and q, and can also we used when $w = z$ (in which case we get 0, independent of p and q). Thus d is a well–defined function on all of $\mathbb{H} \times \mathbb{H}$. We now have to show that the function d satisfies (d1-d5) and hence defines the metric we want.

(d1) is obvious, and (d2) follows again from Proposition 2.10(i).

Consider next (d5):

Let h be an element in $M\ddot{o}b(\mathbb{H})$. Then $h(p)$ and $h(q)$ are the endpoints of the line through $h(z)$ and $h(w)$. If $h \in M\ddot{o}b(\mathbb{H})$,

$$d(h(z),h(w)) = |\ln|[h(w),h(z),h(p),h(q)]|| = |\ln|[w,z,p,q]|| = d(z,w)$$

by Proposition 2.10 and Exercise 2.9.

Conversely, assume that $d(z,w) = d(z',w')$, where we assume $z \neq w$ and $z' \neq w'$ — the other case being trivial. Let p,q and p',q' be the endpoints of the lines $\overleftrightarrow{zw}$ and $\overleftrightarrow{z'w'}$. Then $[w',z',p',q'] = [w,z,p,q]$ or $1/[w,z,p,q]$, and by interchanging p' and q', if necessary, we may assume $[w',z',p',q'] = [w,z,p,q]$. Let $h(u) = [u,z,p,q]$ and $h'(u) = [u,z',p',q']$, and set $g = (h')^{-1}h$. Then $g(z) = (h')^{-1}(1) = z'$ and $g(w) = (h')^{-1}([w,z,p,q]) = (h')^{-1}([w',z',p',q']) = w'$.

It remains to prove the triangle inequality (d3) with the additional property (d4). We need the following lemma:

Lemma 5.1. *For every z,w we have $d(z,w) \geq |\ln(\operatorname{Im} w/\operatorname{Im} z)|$, with equality if and only if $\operatorname{Re} z = \operatorname{Re} w$.*

Proof. If $\operatorname{Re} z = \operatorname{Re} w$ we can choose $p = \operatorname{Re} z$ and $q = \infty$, and we have $[w,z,p,q] = (w-p)/(z-p) = \operatorname{Im} w/\operatorname{Im} z$.

Henceforth, assume $\operatorname{Re} z \neq \operatorname{Re} w$. The inequality is trivially satisfied if $\operatorname{Im} z = \operatorname{Im} w$, hence we also assume $\operatorname{Im} z \neq \operatorname{Im} w$, and since $d(z,w) = d(w,z)$, we may choose the labeling such that $\operatorname{Im} w > \operatorname{Im} z$. Then

$$d(z,w) = d(w,z) = |\ln(|\frac{w-q}{w-p}\,\frac{z-p}{z-q}|)| = |\ln(\frac{|w-q|}{|w-p|}/\frac{|z-q|}{|z-p|})|\,,$$

where p and q are the endpoints of $\overleftrightarrow{zw}$. This expression does not change if we interchange p and q, hence we may assume that $p * w * z * q$. We now have the situation illustrated in figure 6 (or its mirror image).

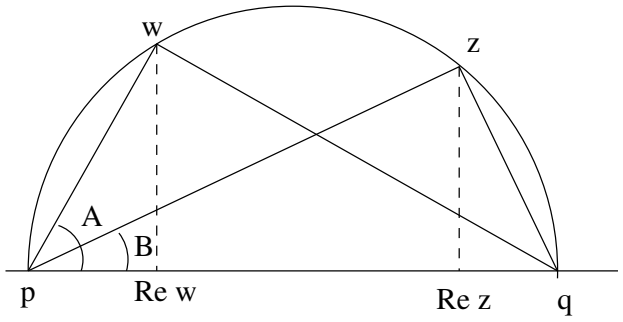

FIG. 6

From the figure we see that $\dfrac{|w-q|}{|w-p|} = \tan A$ and $\dfrac{|z-q|}{|z-p|} = \tan B$. But we also have $\tan A = \operatorname{Im} w/|\operatorname{Re} w - p|$ and $\tan B = \operatorname{Im} z/|\operatorname{Re} z - p|$, hence we get

$$\frac{|w-q|}{|w-p|}\Big/\frac{|z-q|}{|z-p|} = \frac{\tan A}{\tan B} = \frac{\operatorname{Im} w}{\operatorname{Im} z}\cdot\frac{|\operatorname{Re} z - p|}{|\operatorname{Re} w - p|}\,.$$

Because of the way we have chosen p, w, z and q, we have the inequalities $\dfrac{\operatorname{Im} w}{\operatorname{Im} z} > 1$ and $\dfrac{|\operatorname{Re} z - p|}{|\operatorname{Re} w - p|} > 1$. Therefore $\dfrac{|w-q|}{|w-p|}\Big/\dfrac{|z-q|}{|z-p|} > \dfrac{\operatorname{Im} w}{\operatorname{Im} z} > 1$, and

$$d(z,w) = \ln\Big(\frac{|w-q|}{|w-p|}\Big/\frac{|z-q|}{|z-p|}\Big) > \ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} z}\Big)\,.$$

Hence we have strict inequality when $\operatorname{Re} z \neq \operatorname{Re} w$ and equality when $\operatorname{Re} z = \operatorname{Re} w$. □

To prove the triangle inequality, it is enough to consider the case $z = i$, $w = ti$, where $t > 1$, since we can always move to this situation by (d5). If u is a third point, we get

$$\begin{aligned} d(z,u) + d(u,w) &\geq |\ln(\operatorname{Im} u/\operatorname{Im} z)| + |\ln(\operatorname{Im} w/\operatorname{Im} u)| \geq \\ &\quad |\ln(\operatorname{Im} u/\operatorname{Im} z) + \ln(\operatorname{Im} w/\operatorname{Im} u)| = |\ln(\operatorname{Im} w/\operatorname{Im} z)| = d(z,w)\,. \end{aligned}$$

This is (d3). We have equality if and only if $\operatorname{Re} u = 0$ and $\ln(\operatorname{Im} u/\operatorname{Im} z)$ and $\ln(\operatorname{Im} w/\operatorname{Im} u)$ have the same sign — i. e. if and only if u also lies on the imaginary axis and $1 = \operatorname{Im} z \leq \operatorname{Im} u \leq \operatorname{Im} w$. But this means precisely that $u \in [z,w]$, and (d4) follows. □

Exercises.

5.1. Show that a function $f : (1,\infty) \to \mathbb{R}$ which is differentiable in one point and satisfies the equation $f(st) = f(s) + f(t)$ for all s and t must be equal to $C\ln(t)$ for some positive constant C. (Hint: Show that F is differentiable everywhere and compute its derivative.)

5.2. The distance between two subsets U and V of a metric space is defined as $d(U,V) = \inf\{d(u,v) \mid u \in U,\ v \in V\}$.

Let l_1 and l_2 be two hyperbolic lines with a common endpoint. Show that $d(l_1,l_2) = 0$.

5.3. What does the set of points in $\mathbb{H}$ having the same, fixed distance to the y-axis look like?

5.4. Assume the function $f : \mathbb{H} \to \mathbb{H}$ preserves hyperbolic distance. (I. e. $d(f(x), f(y)) = d(x, y)$ for all $x, y \in \mathbb{H}$.) Prove that $f \in M\ddot{o}b(\mathbb{H})$.

5.5. Show that the metric d defines the usual topology (as a subspace of $\mathbb{R}^2$) in the upper half–plane. (This is not so easy to prove at this stage, but you may try it as a challenge. After section 7 this problem will be much simpler. See Exercise 7.6.)

6. Angle measure. $\mathbb{H}$ as a conformal model

Whereas the definition of distance required a fair amount of work, it turns out that angle measure is much simpler, due to the fact that all the congruence transformations are conformal as maps of $\mathbb{C}$ (Lemma 2.1ii).

A point z on a line divides the line into two rays with z as common vertex. A third ray from z determines two angles with these rays, and these angles are called *supplements* of each other. A given angle then has two supplementary angles, but they are easily seen to be congruent.

Definition. An angle is said to be a *right angle* if it is congruent to its supplements.

The usual (absolute) angular measure in $\mathbb{R}^2$ is a function which to any angle associates a number between 0 and $2R$, where R is the number we associate to a right angle (usually $\pi/2$ or 90 degrees), such that two angles are congruent (in the Euclidean sense) if and only if they are associated to the same number. The function is also *additive*, in the following precise sense:

Suppose A and D are on opposite sides of $\overleftrightarrow{BC}$, and suppose the angles $\angle ABC$ and $\angle CBD$ have angle measures U and V, respectively, If $U + V < 2R$, then the angle $\angle ABD$ has measure $U+V$. (The conditions mean that we are talking about a *sum*, rather than a *difference* of angles, and that we are only considering angles smaller than two right angles.) If we normalize the angle measure by requiring that right angles have the measure R and every number in the interval $(0, 2R)$ is realized by some angle, this determines the angle measure uniquely. Note that what determines the angle measure is then essentially the concept of *congruence.*

In $\mathbb{H}$ an angle measure should have exactly the same properties, except that two angles now should have the same measure if and only if they are congruent in the *hyperbolic* sense — i. e. there exists a Möbius transformation mapping one to the other. Now we define the *Euclidean angle* between two rays with common vertex to be the angle between their tangents at the vertex. Then Lemma 2.1 says that fractional linear transformations preserve the Euclidean angle measure, and complex conjugation trivially does the same. Therefore all Möbius transformations, and in particular those in

$M\ddot{o}b(\mathbb{H})$, also preserve Euclidean angle measure. In fact, the converse is also true, so we have:

Lemma 6.1. *Two angles A and B are congruent if and only if they have the same Euclidean angle measure.*

Proof. We only need to prove the *if* part. Let $A = \angle xyz$ and $B = \angle uvw$, and suppose they both have Euclidean angle measure θ. By Hilbert's axiom C4 we may reduce to the case $\overrightarrow{yx} = \overrightarrow{vu} = [i, 0)$. (Notation from section 4.) But there are exactly two rays from i making an angle θ with $[i, 0)$ — one on each side of the imaginary axis — and they are mapped to each other by the reflection $z \mapsto -\bar{z}$, which fixes $[i, 0)$. □

It follows that we can use the same measure of angles in $\mathbb{H}$ as in the Euclidean plane containing it.

We express this by saying that the Poincaré upper half–plane $\mathbb{H}$ is a *conformal* model for hyperbolic geometry. This is one of the properties that makes this model much more useful that the Beltrami–Klein model $\mathbb{K}$, which is not conformal.

Note that Lemma 6.1 then also says that two angles are congruent if and only if they have the same size.

Since stereographic projection preserves angles, it follows that the hemisphere model $\mathbb{B}$ also is conformal, and hence so is every other model obtained from it by stereographic projections. In particular, this is true for Poincaré's disk model $\mathbb{D}$, which we will investigate in the next section.

Exercises

6.1. Show that if l is a hyperbolic line and z is point not on l, then there is a unique line l' which contains z and which meets l orthogonally.

6.2. Show that if l_1 and l_2 are two lines which do not have a common endpoint and which do not intersect, then there exists a line m which meets both orthogonally.

7. Poincaré's disk model $\mathbb{D}$

Because of its rotational symmetry, the *Poincaré disk model* $\mathbb{D}$ will in certain respects have great advantages over $\mathbb{H}$. Therefore this section is devoted to a closer study of this model. Having two different models to our disposal enables us in each situation to choose the one best suited. We will see examples of this in the last two sections.

We begin by transferring everything we have done with the upper half–plane model $\mathbb{H}$ to $\mathbb{D}$. We will do this using the bijection $G = \Phi \circ F \circ \Phi^{-1}$, where Φ is stereographic projection and $F : B_1 \approx B_2$ is an identification between the hemispheres $B_1 = \{(x, y, z)|y > 0\}$ and $B_2 = \{(x, y, z)|z < 0\}$. (Cfr. Exercise 1.3.) If we choose $F(x, y, z) = (x, z, -y)$, the formulae (1.1) and (1.2) give:

$$G(u,v) = \left(\frac{2u}{u^2 + (v+1)^2} \, , \frac{u^2 + v^2 - 1}{u^2 + (v+1)^2} \right) ,$$

or, if we write G in terms of complex numbers $z = u + iv$:

$$G(z) = \frac{z + \bar{z} + i(z\bar{z} - 1)}{|z + i|^2} = \frac{(iz+1)(\bar{z} - i)}{(z+i)(\bar{z} - i)} = \frac{iz+1}{z+i} \, .$$

This is a fractional linear transformation which restricts to a bijection $G : \mathbb{H} \approx \mathbb{D}$. We see that $G(0) = -i$, $G(1) = 1$ and $G(-1) = -1$, and this determines G uniquely, by Corollary 2.5. (In fact, we could have used this to define G.) Observe also that $G(\infty) = i$ and $G(i) = 0$.

There are, of course, many other possible FLT's identifying $\mathbb{H}$ and $\mathbb{D}$, but G is particularly simple and will be our preferred choice.

G preserves angles and $\overline{\mathbb{C}}$–circles, hence it maps the circle through 0, 1 and -1, i. e. $\overline{\mathbb{R}}$, to the circle through $-i$, 1 and -1, i. e. $\mathbb{S}^1 = \{z \in \mathbb{C} \,|\, |z| = 1\}$. Therefore the $\mathbb{H}$–lines are mapped to either circular arcs meeting $\mathbb{S}^1$ orthogonally, or diameters. These curves are the hyperbolic lines in the disk model, or *$\mathbb{D}$–lines*.

Since G preserves angles, this model will also be *conformal* — i. e. the hyperbolic angle measure is the same as the Euclidean measure.

The group of real Möbius transformations $M\ddot{o}b(\mathbb{H})$ corresponds to a group $M\ddot{o}b(\mathbb{D})$ of transformations of $\mathbb{D}$ which preserve angles and hyperbolic lines, and which generally has the same properties with respect to $\mathbb{D}$ as $M\ddot{o}b(\mathbb{H})$ to $\mathbb{H}$. $M\ddot{o}b(\mathbb{D})$ is defined by

$$f \in M\ddot{o}b(\mathbb{D}) \iff G^{-1} f G \in M\ddot{o}b(\mathbb{H}),$$

In other words: every element $f \in M\ddot{o}b(\mathbb{D})$ can be written as GgG^{-1} for some $g \in M\ddot{o}b(\mathbb{H})$, and every transformation of this form is in $M\ddot{o}b(\mathbb{D})$. It follows that $M\ddot{o}b(\mathbb{D})$ and $M\ddot{o}b(\mathbb{H})$ are conjugate subgroups of $M\ddot{o}b(\mathbb{C})$, hence they are isomorphic as abstract groups. We will also use the notation $M\ddot{o}b^+(\mathbb{D})$ and $M\ddot{o}b^-(\mathbb{D})$ for the subgroup and coset corresponding to $M\ddot{o}b^+(\mathbb{H})$ and $M\ddot{o}b^-(\mathbb{H})$.

Congruence in $\mathbb{D}$ can now be defined as in $\mathbb{H}$, but using the group $M\ddot{o}b(\mathbb{D})$ instead of $M\ddot{o}b(\mathbb{H})$. The mappings G and G^{-1} will then preserve congruence.

To see what the transformations in $M\ddot{o}b(\mathbb{D})$ look like, we use the matrix representation of Möbius transformations. G corresponds to the matrix $\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$, and for G^{-1} we will use the formula $G^{-1}(z) = \dfrac{iz-1}{-z+i}$, corresponding to the matrix $\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix}$.

Consider first $M\ddot{o}b^+(\mathbb{D})$. If $g(z) = \dfrac{az+b}{cz+d}$, a,b,c,d real, GgG^{-1} will correspond to

$$\begin{aligned}\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix} &= \begin{bmatrix} ia+c & ib+d \\ a+ic & b+id \end{bmatrix}\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix} \\ &= \begin{bmatrix} -a+ci-ib-d & -ia-c-b+id \\ ia-c-b-id & -a-ic+ib-d \end{bmatrix} \\ &= \begin{bmatrix} -a-d+(c-b)i & -b-c-(a-d)i \\ -b-c+(a-d)i & -a-d-(c-b)i \end{bmatrix}.\end{aligned}$$

The last matrix has the form $\begin{bmatrix} \alpha & \beta \\ \bar\beta & \bar\alpha \end{bmatrix}$, with $\alpha = -a-d+(c-b)i$ and $\beta = -b-c-(a-d)i$. This means that elements in $M\ddot{o}b^+(\mathbb{H})$ give rise to complex fractional linear transformations of the form

$$g(z) = \frac{\alpha z+\beta}{\bar\beta z+\bar\alpha}. \tag{7.1}$$

On the other hand, it is easy to see that every pair of complex numbers (α,β) can be written uniquely on this form, with a,b,c and d real. A simple calculation (compare determinants) gives $\alpha\bar\alpha-\beta\bar\beta = 4(ad-bc)$, hence $\dfrac{\alpha z+\beta}{\bar\beta z+\bar\alpha}$ defines an element in $M\ddot{o}b^+(\mathbb{D})$ if and only if $\alpha\bar\alpha-\beta\bar\beta>0$, and we can normalize α and β such that $\alpha\bar\alpha-\beta\bar\beta=1$. Note that $(k\alpha,k\beta)$ defines the same function as (α,β) only if k is *real*. Hence we can only normalize by multiplication by real numbers. Any such normalization will preserve the sign of $\alpha\bar\alpha-\beta\bar\beta$.

The corresponding computation for $M\ddot{o}b^-(\mathbb{D})$ is slightly different because of the complex conjugation involved. If $g(z) = \dfrac{a\bar z+b}{c\bar z+d}$ with a,b,c,d real, then

$$GgG^{-1}(z) = G\left(\frac{a\overline{G^{-1}(z)}+b}{c\overline{G^{-1}(z)}+d}\right) = G\left(\frac{a\overline{G^{-1}}(\bar z)+b}{c\overline{G^{-1}}(\bar z)+d}\right),$$

where $\overline{G^{-1}}$ now is the fractional linear transformation corresponding to the *conjugate* of the matrix for G^{-1}, i. e. $\begin{bmatrix} -i & -1 \\ -1 & -i \end{bmatrix} = -\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$. But this matrix defines the same transformation as $\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$, hence $\overline{G^{-1}} = G$. One possible matrix of coefficients for GgG^{-1} is therefore

$$\begin{aligned}&\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix} = \cdots \\ &\qquad \cdots = \begin{bmatrix} -a+d+(c+b)i & c-b-(a+d)i \\ -(c-b)+(a+d)i & a-d+(c+b)i \end{bmatrix} = \begin{bmatrix} \alpha' & \beta' \\ -\bar\beta' & -\bar\alpha' \end{bmatrix},\end{aligned}$$

with $\alpha' = -a + d + (c + b)i$ and $\beta' = c - b - (a + d)i$. But $\begin{bmatrix} \alpha' & \beta' \\ -\bar{\beta}' & -\bar{\alpha}' \end{bmatrix}$ determines the same function as

$$\begin{bmatrix} i\alpha' & i\beta' \\ -i\bar{\beta}' & -i\bar{\alpha}' \end{bmatrix} = \begin{bmatrix} i\alpha' & i\beta' \\ \overline{i\beta'} & \overline{i\alpha'} \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \bar{\beta} & \bar{\alpha} \end{bmatrix},$$

with $\alpha = i\alpha'$ and $\beta = i\beta'$. Here $\alpha\bar{\alpha} - \beta\bar{\beta} = \alpha'\bar{\alpha}' - \beta'\bar{\beta}' = -4(ad - bc)$, which is positive if $g \in \mathit{M\ddot{o}b}^-(\mathbb{H})$. Therefore we again may normalize such that $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$

Hence $\mathit{M\ddot{o}b}^-(\mathbb{H})$ corresponds to maps of the form

$$f(z) = \frac{\alpha\bar{z} + \beta}{\bar{\beta}\bar{z} + \bar{\alpha}}, \tag{7.2}$$

with $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$.

Examples. (1) Complex conjugation in $\mathbb{D}$. Then $\alpha = 1$, $\alpha' = -i$ and $\beta = \beta' = 0$. The equations above give $-a + d = 0$, $c + b = -1$, $a + d = 0$ and $c - b = 0$. Therefore $a = d = 0$ and $c = b = -1/2$. Hence complex conjugation in $\mathbb{D}$ corresponds to $f(z) = (-\frac{1}{2})/(-\frac{1}{2})\bar{z} = 1/\bar{z} = z/|z|^2$ in $\mathbb{H}$, i. e. the inversion in the circle ($\mathbb{H}$–line) $|z| = 1$.

However, we could also argue like this: Complex conjugation is an element in $\mathit{M\ddot{o}b}(\mathbb{D})$ which fixes the whole horizontal diameter ℓ. Hence the corresponding element in $\mathit{M\ddot{o}b}(\mathbb{H})$ must fix all of $G^{-1}(\ell)$, which must be the $\mathbb{H}$–line containing i and having endpoints 1 and -1. But, by the classification results of section 3, this is the inversion $z \mapsto 1/\bar{z}$.

(2) Similarly we see that inversion (reflection) in the imaginary axis $\mathbb{H}$ corresponds to reflection in the imaginary axis in $\mathbb{D}$.

(3) Let us now determine the elements in $\mathit{M\ddot{o}b}(\mathbb{D})$ which have 0 as a fixpoint. For $\mathit{M\ddot{o}b}^+(\mathbb{D})$ this means $\dfrac{\alpha \cdot 0 + \beta}{\bar{\beta} \cdot 0 + \bar{\alpha}} = 0$ — i. e. $\beta = 0$. The condition $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$ gives $|\alpha| = 1$, and we can write $\alpha = \mathrm{e}^{i\theta}$ for some $\theta \in \mathbb{R}$. Hence any element in $\mathit{M\ddot{o}b}^+(\mathbb{D})$ which has 0 as fixpoint can be written $f(z) = \alpha z/\bar{\alpha} = \alpha^2 z = e^{i2\theta} z$. But this is the formula for a rotation by the angle 2θ, written as complex multiplication. Conversely, any such rotation is an element in $\mathit{M\ddot{o}b}^+(\mathbb{D})$.

A short calculation shows that the transformation $G^{-1}fG$ of $\mathbb{H}$ that this rotation corresponds to is the elliptic transformation $g_\theta(z) = \dfrac{\cos\theta z + \sin\theta}{-\sin\theta z + \cos\theta}$ studied in section 3.

Similar considerations show that an element of $\mathit{M\ddot{o}b}^-(\mathbb{D})$ having 0 as fixpoint has the form $f(z) = e^{i\theta}\bar{z}$, which is a reflection in a line (diameter in D) forming an angle of $\theta/2$ with the x-axis. (For example, we can write $e^{i\theta}\bar{z} = e^{i\theta/2}\overline{e^{-i\theta/2}z}$, which means that we get $e^{i\theta}\bar{z}$ from z by first rotating by the angle $-\theta/2$, then reflecting in the x-axis, and finally rotating back by the angle $\theta/2$.) These mappings — rotations and reflections — form the group of orthogonal linear transformations in dimension 2, corresponding to the group $O(2)$ of orthogonal 2×2–matrices. Hence we have shown that the set of elements of $\mathit{M\ddot{o}b}(\mathbb{D})$ fixing 0 is precisely the group $O(2)$ acting on $\mathbb{R}^2$, restricted to the open unit disk $\mathbb{D}$.

Note that via the isomorphism with $M\ddot{o}b(\mathbb{H})$ (conjugation by G), the subgroup of rotations (isomorphic to $(SO(2))$ corresponds to the *elliptic* transformations of $\mathbb{H}$ having i as fixpoint. This is the identification of these transformations with rotations that we promised in section 3.

More generally, also in $M\ddot{o}b(\mathbb{D})$ we have a classification of elements according to the behaviour of fixpoints, and we can define parabolic, hyperbolic and elliptic elements as before, corresponding to the elements of the same types in $M\ddot{o}b(\mathbb{H})$. Similarly, in $M\ddot{o}b^-(\mathbb{D})$ the inversions are the transformations with a $\mathbb{D}$–line of fixpoints, corresponding to the inversions in $M\ddot{o}b^-(\mathbb{H})$. Cfr. Exercise 7.2-3.

We can also transfer the metric d from $\mathbb{H}$ to $\mathbb{D}$. Let us write $d_{\mathbb{H}}$ for d as defined in section 5. Then the formula

$$d_{\mathbb{D}}(z_1, z_2) = d_{\mathbb{H}}(G^{-1}(z_1), G^{-1}(z_2))$$

will define a metric on $\mathbb{D}$ such that G and G^{-1} become inverse *isometries* (distance-preserving maps). In particular, the analogues of conditions (d1)-(d5) in section 5 will automatically hold. We will now derive a more explicit expression for $d_{\mathbb{D}}$.

Recall that $d_{\mathbb{H}}(G^{-1}(z_1), G^{-1}(z_2)) = |\ln(|[G^{-1}(z_1), G^{-1}(z_2), P, Q]|)|$, where P, Q are the endpoints of the $\mathbb{H}$–line through $G^{-1}(z_1)$ and $G^{-1}(z_2)$. But G^{-1} maps $\mathbb{D}$–lines to $\mathbb{H}$–lines, so $\{P, Q\} = \{G^{-1}(p), G^{-1}(q)\}$, where p and q are the endpoints of the $\mathbb{D}$–line through z_1 and z_2. Hence

(7.3)

$$d_{\mathbb{D}}(z_1, z_2) = |\ln(|[G^{-1}(z_1), G^{-1}(z_2), G^{-1}(p), G^{-1}(q) = |\ln(|[z_1, z_2, p, q]||\,,$$

by Proposition 2.10(iii). This is completely analogous to the formula for $d_{\mathbb{H}}$, but now all four points are in $\mathbb{C}$, so we can always write

$$d_{\mathbb{D}}(z_1, z_2) = |\ln(|\frac{z_1 - p}{z_1 - q}\,\frac{z_2 - q}{z_2 - p}|)|\,.$$

The extra symmetry in $\mathbb{D}$ can be used to study the metric in more detail, and in particular it will enable us to derive formulae for $d_{\mathbb{D}}(z_1, z_2)$ not involving the endpoints p and q.

First observe that since rotations around the origin are isometries, the distance from 0 to a point z must be equal to the distance from 0 to r, where $r = |z| \in [0, 1) \in \mathbb{D}$. But the endpoints of the $\mathbb{D}$–line through 0 and r are 1 and -1, so the distance formula gives

$$d_{\mathbb{D}}(0, z) = d_{\mathbb{D}}(0, r) = |\ln(|\frac{0 - (-1)}{0 - 1}\,\frac{r - 1}{r - (-1)}|)| = \ln(\frac{1 + r}{1 - r})\,.$$

This equation can also be solved for r, yielding

$$r = \tanh(\frac{d_{\mathbb{D}}(0, z)}{2})\,. \tag{7.4}$$

To find the distance between two arbitrary points z_1 and z_2, we now first move z_1 to 0 by an isometry $f \in M\ddot{o}b^+(\mathbb{D})$, and then use the formula above for the distance betweeen 0 and $f(z_2)$.

If $f(z) = \dfrac{az+b}{\bar{b}z+\bar{a}}$ satisfies $f(z_1) = 0$, then $b = -az_1$, Therefore we can write

$$f(z) = \frac{a(z-z_1)}{-\overline{a}\overline{z_1}z+\bar{a}} = \frac{a}{\bar{a}}\left(\frac{z-z_1}{-\bar{z}_1 z+1}\right).$$

Introducing the notation $\rho = |f(z_2)| = \dfrac{|z_2-z_1|}{|1-\bar{z}_1 z_2|}$, we now have

$$d_{\mathbb{D}}(z_1,z_2) = \ln(\frac{1+\rho}{1-\rho}), \quad \text{or} \quad \rho = \tanh(\frac{d_{\mathbb{D}}(z_1,z_2)}{2}).$$

Another expression is obtained from the second of these by using the formula of Exercise 7.8b:

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{\tanh^2(d_{\mathbb{D}}/2)}{1-\tanh^2(d_{\mathbb{D}}/2)} = \frac{\rho^2}{1-\rho^2}.$$

Substituting for ρ, we have

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{\dfrac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2}}{1-\dfrac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2}} = \frac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2-|z_2-z_1|^2}.$$

The denominator here is

$$\begin{aligned}(1-\bar{z}_1 z_2)(1-z_1\bar{z}_2) - (z_1-z_2)(\bar{z}_1-\bar{z}_2) = \\ 1-\bar{z}_1 z_2 - z_1\bar{z}_2 + |z_1|^2|z_2|^2 - (|z_1|^2 - z_1\bar{z}_2 - \bar{z}_1 z_2 + |z_2|^2) = \\ 1-|z_1|^2-|z_2|^2+|z_1|^2|z_2|^2 = (1-|z_1|^2)(1-|z_2|^2).\end{aligned}$$

Hence we get

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{|z_2-z_1|^2}{(1-|z_1|^2)(1-|z_2|^2)}.$$

Finally we make use of the identity $\cosh(2w) = 2\sinh^2(w)+1$ with $w = d_{\mathbb{D}}/2$, and obtain:

$$\cosh(d_{\mathbb{D}}(z_1,z_2)) = 1 + \frac{2|z_2-z_1|^2}{(1-|z_1|^2)(1-|z_2|^2)}. \tag{7.5}$$

This is perhaps the most common formula for the metric in $\mathbb{D}$. But now we may also go back to $\mathbb{H}$ using the isometry G, to obtain a similar fomula for $d_{\mathbb{H}}$:

$$d_{\mathbb{H}}(w_1,w_2) = d_{\mathbb{D}}(G(w_1),G(w_2)) = d_{\mathbb{D}}(\frac{iw_1+1}{w_1+i}, \frac{iw_2+1}{w_2+i}).$$

A simple calculation gives

$$\frac{iw_2+1}{w_2+i} - \frac{iw_1+1}{w_1+i} = \cdots = \frac{2w_1-2w_2}{(w_2+i)(w_1+i)}, \quad \text{and}$$

$$1 - |\frac{iw+1}{w+i}|^2 = \frac{(w+i)(\bar{w}-i) - (iw+1)(-i\bar{w}+1)}{|w+i|^2} =$$
$$\cdots = \frac{2\,i\bar{w} - 2\,iw}{|w+i|^2} = \frac{4\,\mathrm{Im}\,w}{|w+i|^2}\,.$$

Substituting these expressions into the formula for $d_{\mathbb{D}}$ above, we obtain the formula

$$\cosh(d_{\mathbb{H}}(w_1, w_2)) = 1 + \frac{|w_2 - w_1|^2}{2\,(\mathrm{Im}\,w_1)(\mathrm{Im}\,w_2)}\,. \tag{7.6}$$

These equations (7.5 and 7.6) express the metrics in $\mathbb{D}$ and $\mathbb{H}$ as functions of the points only, without referring to the endpoints of lines. Note that the function $\cosh(t)$ is increasing for $t \geq 0$, hence the equations determine the metrics uniquely.

Exercises.

7.1. Show that the restriction $G^{-1}|\,\mathbb{S}^1 : \mathbb{S}^1 \to \overline{\mathbb{R}}$ of the fractional linear transformation G is the analogue of stereographic projection from $i \in \mathbb{S}^1$.

7.2. Discuss a classification of the elements of $M\ddot{o}b^+(\mathbb{D})$ analogous to the classification of elements of $M\ddot{o}b^+(\mathbb{H})$ in section 3.

7.3. Show that an element of $M\ddot{o}b(\mathbb{D})$ has a $\mathbb{D}$–line of fixpoints if and only if it is the restriction of an inversion in a $\overline{\mathbb{C}}$–circle. (Hence the term 'inversion' is well–defined.)

Show that $\dfrac{\alpha\bar{z} + \beta}{\bar{\beta}\bar{z} + \bar{\alpha}}$ determines an inversion if and only if $|\alpha|^2 - |\beta|^2 > 0$ and $\mathrm{Re}\,\beta = 0$.

7.4. (a) Show that *hyperbolic circles*, i. e. subsets of $\mathbb{D}$ of the form $\{z \in \mathbb{D} \,|\, d_{\mathbb{D}}(z, z_0) = r\}$, where z_0 is a fixed point and $r > 0$, also are Euclidean circles.

(b) The same problem with $\mathbb{H}$ instead of $\mathbb{D}$.

7.5. Fix a point z_0 on a hyperbolic line l, and consider (hyperbolic) circles through z_0 with center on l. Show that as the center approaches an endpoint of l, the circle approaches a horocircle.

7.6. Show that $d_{\mathbb{D}}$ defines the subspace topology on $\mathbb{D} \subset \mathbb{C}$.

(Hint: Show first (1) $d_{\mathbb{D}} : \mathbb{D} \times \mathbb{D} \to [0, \infty)$ is continuous in the Euclidean topology on $\mathbb{D}$, and (2) $d_{\mathbb{D}}(z_1, z_2) \geq |z_1 - z_2|$ for all z_1 and z_2 in D.)

Why have we now also solved Exercise 5.5?

7.7. Prove a converse to Exercise 5.2, i. e. show that if l_1 and l_2 are two lines which do not intersect and $d(l_1, l_2) = 0$, then they have a common endpoint. (You might need hint (2) in Exercise 7.6.)

Same question for $\mathbb{D}$.

7.8. In this and the next two sections we use several relations between the hyperbolic functions. Verify the following formulae:

(a) $\cosh^2 x - \sinh^2 x = 1$,

(b) $\sinh^2 x = \dfrac{\tanh^2 x}{1-\tanh^2 x}$,
(c) $\sinh 2x = 2\sinh x \cosh x$,
(d) $\cosh 2x = \cosh^2 x + \sinh^2 x$,
(e) $\dfrac{1+\tanh^2 x}{1-\tanh^2 x} = \cosh 2x$,
(f) $\dfrac{2\tanh x}{1-\tanh^2 x} = \sinh 2x$.

8. Arc–length and area in the hyperbolic plane

Suppose that $\mathcal{C}$ is a curve in a metric space, given by a parametrization $z(t),\ t \in [a,b]$, and assume for simplicity that z is injective. The *arc–length* of $\mathcal{C}$ is defined as

$$\sup_{a=t_0<t_1<\cdots<t_n=b} \sum_i d(z(t_i)z(t_{i+1}))\,, \tag{8.1}$$

provided this number is finite. (The supremum is taken over all partitions $a = t_0 < t_1 < \cdots < t_n = b$ of the interval $[a,b]$.) If so, we say that the curve is *rectifiable.* This is clearly the case if $z(t)$ satisfies a *Lipschitz* condition on $[a,b]$ — i. e. if there exists a constant K such that

$$d(z(t_1), z(t_2)) \le K|t_1 - t_2|$$

for all t_1, t_2 in $[a,b]$.

If $\mathcal{C}$ is rectifiable, then the arc–length of the restriction of z to $[a,t]$ exists for every $t \in [a,b]$ and defines a continuous, nondecreasing function $s(t)$ on $[a,b]$. This is not the place to discuss the general theory, but one can show that if the limit

$$\sigma(t) = \lim_{h\to 0} \frac{d(z(t+h), z(t))}{|h|},$$

exists and is continuous at every t, then $s(t)$ is given as the integral

$$s(t) = \int_a^t \sigma(\tau)\, d\tau\,.$$

Hence

$$s'(t) = \sigma(t) = \lim_{h\to 0} \frac{d(z(t+h), z(t))}{|h|}\,. \tag{8.2}$$

In particular, this condition will be satisfied in $\mathbb{R}^2$, $\mathbb{H}$ or $\mathbb{D}$ whenever $z(t)$ is $\mathcal{C}^1$ as a curve in $\mathbb{R}^2$ — i. e. whenever both component functions are continuously differentiable.

The following is an important observation: Suppose $g : X \to Y$ is an isometry between (possibly different) metric spaces X and Y, and let $z(t),\ t \in [a,b]$ is a curve in X with image curve $gz(t)$ in Y. Then, if z has arc–length s, gz will also have arc–length s — i. e. *arc length is preserved by isometries.* This follows immediately from the definition (8.1), since we then always will have $d(gz(t_{i+1}), gz(t_i)) = d(z(t_{i+1})z(t_i))$.

Example 8.1. If $z(t)$ parametrizes a segment of a hyperbolic line, condition (d4) for a metric in in chapter 5 implies that the arc-length is equal to the hyperbolic distance between the endpoints. Moreover, an obvious generalization of the triangle inequality (d3) shows that $d(z(a), z(b)) \leq \Sigma_i d(z(t_i), d(z_{i+1}))$ for every partition $a = t_0 < t_1 < \cdots < t_n = b$ of $[a, b]$. Hence the hyperbolic line is the shortest possible curve between the two points.

Example 8.2. In $\mathbb{R}^2$ we have the formula

$$s'(t) = \sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}\,, \tag{8.3}$$

where $z(t) = (x(t), y(t))$, provided $z(t)$ is continuously differentiable. We may also write this as $(\frac{ds}{dt})^2 = (\frac{dx}{dt})^2 + (\frac{dy}{dt})^2$. This formula is valid for any parametrization, and we express this by the relation

$$ds^2 = dx^2 + dy^2\,. \tag{8.4}$$

(This relation can be given a precise interpretation as an equation in an appropriate vector space, but here it suffices to think about it as an invariant notation for (8.3)).

We will now derive analogous expressions for the arc-length in the two models $\mathbb{H}$ and $\mathbb{D}$ for the hyperbolic plane. To distinguish between the two cases we write $d_{\mathbb{H}}$ and $d_{\mathbb{D}}$ for the metrics.

We start with $\mathbb{H}$. The distance formula (7.6) gives:

$$\cosh(d_{\mathbb{H}}(z(t+h), z(t))) = 1 + \frac{|z(t+h) - z(t)|^2}{2\,(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,. \tag{8.5}$$

To simplify notation we now write $d(h) = d_{\mathbb{H}}(z(t+h), z(t))$. By Taylor's formula for cosh we can write $\cosh w = 1 + \frac{w^2}{2} + \eta(w)w^2$, where $\lim_{w\to 0} \eta(w) = 0$. (8.5) then yields

$$1 + \frac{(d(h))^2}{2} + \eta(d(h))(d(h))^2 = 1 + \frac{|z(t+h) - z(t)|^2}{2\,(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,,$$

and hence

$$(\frac{d(h)}{|h|})^2(1 + 2\,\eta(d(h))) = \left|\frac{z(t+h) - z(t)}{h}\right|^2 \frac{1}{(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,. \tag{8.6}$$

It follows that if $z(t)$ is $\mathcal{C}^1$, then $\lim_{h\to 0}\left(\frac{d(z(t+h), z(t))}{|h|}\right)^2$ exist and is equal to $\left((\frac{dx}{dt})^2 + (\frac{dy}{dt})^2\right)/(\operatorname{Im} z(t))^2$. Since the expressions involved are positive, we get

$$s'(t) = \lim_{h\to 0} \frac{d(z(t+h), z(t))}{|h|} = \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{\operatorname{Im} z(t)}\,.$$

Analogous to the example $\mathbb{R}^2$ above we can also write (remember that $y = \operatorname{Im} z$)

$$ds^2 = \frac{dx^2 + dy^2}{y^2}\,. \tag{8.7}$$

For the Poincaré disk we can make a similar analysis. The only difference is that the factor $\dfrac{1}{(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}$ on the right hand side of formula (8.6) is replaced by $\dfrac{4}{(1-|z(t+h)|^2)(1-|z(t)|^2)}$. Thus, in this case we obtain

$$ds^2 = 4\frac{dx^2 + dy^2}{(1-x^2-y^2)^2}\,. \tag{8.8}$$

As in the Euclidean case (8.4) we should think of these formulae as a way to describe how to get ds/dt from a parametrization $z(t) = (x(t), y(t))$ of the curve. Thus, (8.8) means that in $\mathbb{D}$ we have

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

Example 8.3. Let us apply this to compute the arc–length (circumference) of the hyperbolic circle $\mathcal{C}$ with hyperbolic radius ρ. (Cfr. Exercise 7.5.) Since arc–length is preserved by isometries, we may assume that $\mathcal{C} \subset \mathbb{D}$ and with center in 0. $\mathcal{C}$ will then also be a *Euclidean* circle with center in 0, with Euclidean radius given by formula (7.4) — i. e. $r = \tanh(\rho/2)$. $\mathcal{C}$ may then be parametrized as $z(t) = re^{it} = (r\cos t, r\sin t),\ t \in [0, 2\pi]$, and we get $dx/dt = -r\sin t,\ dy/dt = r\cos t$ and $x^2 + y^2 = r^2$. Thus

$$(\frac{ds}{dt})^2 = 4\frac{(-r\sin t)^2 + (r\cos t)^2}{(1-r^2)^2} = \frac{4r^2}{(1-r^2)^2}\,.$$

Hence the arc–length of $\mathcal{C}$ is

$$s(\mathcal{C}) = \int_0^{2\pi} \frac{2r}{1-r^2}\,dt = \frac{4\pi r}{1-r^2} = \pi\frac{4\,\tanh(\frac{\rho}{2})}{1-\tanh^2(\frac{\rho}{2})} = 2\pi\sinh(\rho)\,.$$

Now recall that

$$2\sinh(\rho) = e^{\rho} - e^{-\rho} = 2\rho + \frac{\rho^3}{3} + \frac{\rho^5}{60} + \cdots\,.$$

It follows that the circumference of a circle is greater and increases faster as a function of the radius in the hyperbolic than in the Euclidean plane. More explicitly, we see that for small ρ the circumference is approximately equal to $2\pi\rho$ (i. e. the same formula as in the Euclidean Case), but when ρ is large, it increases approximately as πe^{ρ}.

Next we discuss *area* in the hyperbolic plane, and in particular we want to find the area of a triangle — i. e. the part of the plane bounded by the three segments between three points not on a line. Hence we do not need the most general concept of area possible, and we will limit our study to subsets of the plane bounded by a finite number of $\mathcal{C}^1$curves. ($\mathcal{C}^1$as curves in $\mathbb{R}^2$.) A

reasonable area function A should be *additive* in the sense that for two such subsets U and V we should have $A(U \cup V) + A(U \cap V) = A(U) + A(V)$, and the area of points and smooth curves should be 0. Furthermore, congruent sets should have the same area — in other words: Möbius–transformations should preserve area.

It is not hard to see that such a function will be determined up to a constant scaling–factor (as was also the case for the distance function), so we might just write down the formulae below and show that they satisfy these properties. But hopefully the following informal discussion will help to explain the geometric reason for the formulae and why they are normalized as they are.

Consider first $\mathbb{H}$. Le $\Omega \subset \mathbb{H} \subset \mathbb{R}^2$ be a set as described above, and think of the identity map as a parametrization (i. e. as a map from Ω considered as a subset of $\mathbb{R}^2$ to Ω as a subset of $\mathbb{H}$). As usual we now cover *the parameter set* Ω with Euclidean rectangles with vertices (x_i, y_j), where $\{x_i\}_i$ and $\{y_j\}_j$ are increasing sequences of real numbers. Let $R(i,j)$ be the rectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. Then $\sum_{R(i,j)\cap\Omega\neq\emptyset} A(R(i,j))$ will approximate $A(\Omega)$, and the approximation gets better as the rectangles get smaller. Therefore the area should be given by an integral of the form

$$A_{\mathbb{H}}(\Omega) = \iint_{\Omega} K(x,y)\, dx\, dy\,,$$

where

$$K(x,y) = \lim_{\Delta x, \Delta y \to 0} \frac{A_{\mathbb{H}}(R(\Delta x, \Delta y))}{\Delta x \Delta y}\,,$$

and $R(\Delta x, \Delta y)$ denotes the rectangle $[x, x+\Delta x] \times [y, y+\Delta y]$. The two edges $[x, x+\Delta x] \times \{y\}$ and $\{x\} \times [y, y+\Delta y]$ of $R(\Delta x, \Delta y)$ are curves meeting orthogonally in $\mathbb{H}$, hence $A_{\mathbb{D}}(R(\Delta x, \Delta y))$ ought to be approximated by the product of the hyperbolic lengths of these edges. It is therefore natural to normalize $A_{\mathbb{H}}$ by requiring

$$\lim_{\Delta x, \Delta y \to 0} \frac{A_{\mathbb{H}}(R(\Delta x, \Delta y))}{d_{\mathbb{H}}(x, x+\Delta x)\, d_{\mathbb{H}}(y, y+\Delta y)} = 1\,.$$

But from (8.7) we get

$$\lim_{\Delta x \to 0} \frac{d_{\mathbb{H}}(x, x+\Delta x)}{|\Delta x|} = \lim_{\Delta y \to 0} \frac{d_{\mathbb{H}}(y, y+\Delta y)}{|\Delta y|} = \frac{1}{y}\,.$$

Putting all this together we get $K(x,y) = \dfrac{1}{y^2}$, and hence

$$A_{\mathbb{H}}(\Omega) = \iint_{\Omega} \frac{dx\, dy}{y^2}\,. \tag{8.9}$$

This equation we now take to be our definition of the area function on $\mathbb{H}$, and the area is defined for every set for which the integral is defined.

This discussion can also be used to prove that the area is invariant under congruence — i. e. $A_{\mathbb{H}}(g\Omega) = A_{\mathbb{H}}(\Omega)$ if g is a Möbius transformation — but it may be instructive to see how this can also be verified from the formula.

Let $g(z) = \dfrac{az+b}{cz+d}$, with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$, and assume that $\Omega' = g\Omega$. In order to distinguish between Ω and Ω' we use the notation $z = x + iy$ for points in Ω and $w = u + iv$ for points in Ω'.

The formula for change of variables in a double integral gives

$$A_{\mathbb{H}}(\Omega') = \iint_{\Omega'} \frac{du\,dv}{v^2} = \iint_{\Omega} |J(g)(z)| \, \frac{dx\,dy}{(\operatorname{Im} g(z))^2} \,. \tag{8.10}$$

where $|J(g)|$ is the determinant of the Jacobian of g considered as a mapping between subsets of $\mathbb{R}^2$. But the Cauchy–Riemanns equations for the complex analytic mapping g imply that $|J(g)(z)| = |g'(z)|^2$. (Exercise 8.2.) In our case $g'(z) = (ad - bc)/(cz + d)^2$, hence $|J(g)(z)| = 1/|cz + d|^4$.

By formula (2.1) we have $\operatorname{Im}(g(z)) = \operatorname{Im} z/|cz + d|^2 = y/|cz + d|^2$. Substituting all this in (8.10), we get

$$A_{\mathbb{H}}(\Omega') = \iint_{\Omega} \frac{1}{|cz+d|^4} \, \frac{dx\,dy}{\left(\dfrac{y}{|cz+d|^2}\right)^2} = \iint_{\Omega} \frac{dx\,dy}{y^2} = A(\Omega) \,.$$

To show that *all* Möbius transformations preserve area it now suffices to observe that reflection in the imaginary axis, $\gamma(z) = -\bar{z}$, does, since $M\ddot{o}b(\mathbb{H})$ is generated by γ and $M\ddot{o}b^+(\mathbb{H})$. But γ preserves the y–coordinate and has Jacobian equal to -1, so (8.10) again gives $A_{\mathbb{H}}(\Omega') = A_{\mathbb{H}}(\Omega)$.

Before we apply this to compute the area of a hyperbolic triangle, we need to remark that in addition to the ordinary triangles determines by three vertices in $\mathbb{H}$, we can also consider *asymptotic* triangles, with one or more "vertices" in $\overline{\mathbb{R}}$ — so–called *ideal* vertices. The two edges meeting at an ideal vertex are then $\mathbb{H}$–lines or rays with this vertex as common endpoint. We talk about *simply*, *doubly* or *triply* asymptotic triangles if there are one, two or three ideal vertices.

Example 8.4. Area of a triangle. Every finite triangle in $\mathbb{H}$ is congruent to a triangle with one side along the imaginary axis and where the third vertex has positive real part. Figure 7 shows such a triangle, with vertices A, B and C. m and r are the center and the radius of the circular arc (hyperbolic line) $\overline{AB}$, spanned by A and B. The other lower–case letters denote the sizes of the obvious angles, thus for example, b is the angular measure of the hyperbolic angle $\angle ABC$, which is the same as the Euclidean angle between the tangents of the two circular arcs meeting at B.

First we compute the area of the region $ABQP$ bounded by the horizontal Euclidean segment $y = Y$ and the hyperbolic segments $[A, P]$, $[A, B]$ and $[B, Q]$. From the figure we see that we can parametrize x as $x = m - r\cos t$, $t \in [u, \pi - v]$. Then $dx = r \sin t\, dt$, and for every t, y ranges from

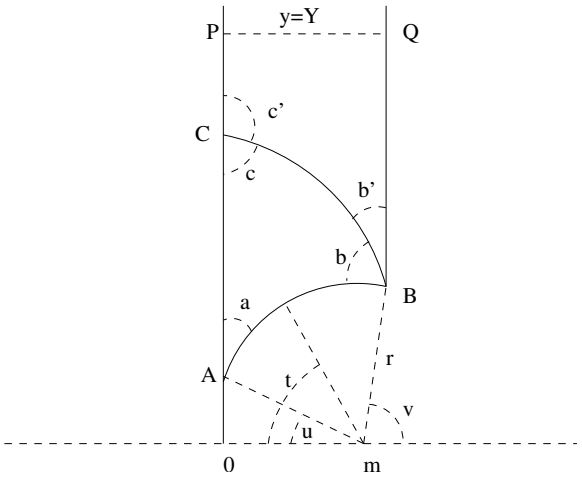

FIG. 7

$r \sin t$ to Y. We get

$$\begin{aligned}\iint_{ABQP} \frac{dx\,dy}{y^2} &= \int_u^{\pi-v} \left[\int_{r\sin t}^{Y} \frac{dy}{y^2}\right] r\sin t\,dt \\ &= \int_u^{\pi-v} \left[\frac{1}{r\sin t} - \frac{1}{Y}\right] r\sin t\,dt \\ &= \pi - u - v - \frac{m + r\cos v}{Y}\,.\end{aligned}$$

Note that as Y goes to ∞, this expression approaches $\pi - u - v$. This means that the asymptotic triangle with vertices A, B and ∞ has *finite* area equal to $\pi - (u + v)$. Now observe that the Euclidean lines mO and mA meet the $\mathbb{H}$–segments $[AP]$ and $[AB]$ orthogonally, hence $u = a$. Similarly $v = b + b'$, hence $u + v$ equals the sum of the angles of the triangle, since the third angle is 0. Letting one or both of the vertices A and B approach the real axis, we see that this formula remains valid even for doubly or triply asymptotic triangles.

The area of the finite triangle ABC is equal to the difference between the areas of two such asymptotic triangles — one with area $\pi-(a+b+b')$ and the other with area $\pi-(b'+c')$. Hence the area of ABC is $\pi-a-b-b'-\pi+b'+c' = \pi - a - b - c$, since $c + c' = \pi$. (See figure 7.) Thus we have proved

Proposition 8.5. *The area of a triangle with angles a, b and c is equal to*

$$\pi - (a + b + c)\,.$$

This formula is valid also for asymptotic triangles, i. e. if one or more of the angles are 0.

This is a striking result of fundamental importance. It says that the area only depends on the sum of the angles of the triangle, and since the area is always positive, the sum of the angles in a triangle is always less than π. We also see that the area of a triangle never exceeds π, and the maximal value π is only achieved by the triply asymptotic triangles, (which are all congruent, cfr. Exercise 2.7).

We may also consider the area function for the disk model $\mathbb{D}$. A similar analysis then leads to the formula

$$A_{\mathbb{D}}(\Omega) = \iint_{\Omega} \frac{4\,dx\,dy}{(1-x^2-y^2)^2}\,. \tag{8.11}$$

Using the change of variables–formula 8.10 with g our standard isometry G^{-1} between $\mathbb{D}$ and $\mathbb{H}$ (section 7), we see that G^{-1} — hence also G — is area preserving, in the sense that $A_{\mathbb{H}}(\Omega)$ is defined if and only if $A_{\mathbb{D}}(G(\Omega))$ is defined, and

$$A_{\mathbb{H}}(\Omega) = A_{\mathbb{D}}(G(\Omega))\,.$$

This equation could, of course, also have been used to *define* $A_{\mathbb{D}}$, given $A_{\mathbb{H}}$.

Because of the rotational symmetry in $\mathbb{D}$ it is often convenient to use polar coordinates $x = r\cos\theta,\ y = r\sin\theta$. Then the formula becomes

$$A_{\mathbb{D}}(\Omega) = \iint_{T} \frac{4r\,dr\,d\theta}{(1-r^2)^2}\,, \tag{8.12}$$

where T is the appropriate parameter set in the (r,θ)–plane.

Example 8.6. Let us use this to compute the area of a hyperbolic circular disk $\mathcal{D}$ with hyperbolic radius ρ. As in example 8.3 we may assume that the circle is a Euclidean circle with center 0. The Euclidean radius is then $R = \tanh(\rho/2)$ (7.4), and we can parametrize $\mathcal{D}$ by polar coordinates $x = r\cos\theta,\ y = r\sin\theta$, where $r \in [0,R]$ and $\theta \in [0, 2\,\pi]$. From formula (8.12) we get

$$\begin{aligned} A_{\mathbb{D}}(\mathcal{D}) = \int_0^R \left[\int_0^{2\pi} \frac{4r\,d\theta}{(1-r^2)^2}\right] dr \;&=2\pi \int_0^R \frac{4r\,dr}{(1-r^2)^2} \\ &=2\pi\left[\frac{2}{(1-r^2)}\right]_0^R = 4\pi\frac{R^2}{1-R^2}\,. \end{aligned}$$

But $R = \tanh(\rho/2)$, and the formula in Exercise 7.8 b gives

$$A_{\mathbb{D}}(\mathcal{D}) = 4\pi\frac{R^2}{1-R^2} = 4\pi\sinh^2(\frac{\rho}{2})\,.$$

Using more of Exercise 7.8 and Taylor expansion we get

$$A_{\mathbb{D}}(\mathcal{D}) = 2\pi(\cosh(\rho)-1) = \pi(\mathrm{e}^{\rho}+\mathrm{e}^{-\rho}-2) = \pi(\rho^2+\frac{\rho^4}{12}+\cdots)\,.$$

This means that the area of a circular disk is greater and increases faster with the radius in the hyperbolic plane than in the Euclidean plane. (Just as we observed for the circumference of a circle in example 8.3). In differential geometry this is expressed by saying that the hyperbolic plane has *negative curvature*.

It is also interesting to compare with geometry on a sphere of radius one. There the circumference of a circle of radius ρ is equal to $2\pi\sin\rho$, and the area is $2\pi\sin^2(\rho/2)$. Both are smaller and increase slower than in the Euclidean case. We say that the sphere has *positive curvature*, whereas the Euclidean plane has curvature 0.

Exercises.

8.1. Let $z_1 = a_1 + ib$ and $z_2 = a_2 + ib$ be two points in $\mathbb{H}$ with the same imaginary value. Let L be the hyperbolic arc–length of the *Euclidean* segment between z_1 and z_2. Compute L and show that $L > d_{\mathbb{H}}(z_1, z_2)$.

8.2. To show invariance of area under Möbius transformations we used that $|J(g)(z)| = |g'(z)|^2$ for a complex analytic function g. Verify this.

8.3. Show by calculation that the isometry $G : \mathbb{H} \to \mathbb{D}$ of section 7 is area preserving.

8.4. Find an expression for the area of a hyperbolic n–gon.

8.5. Let T_α be a doubly asymptotic triangle in $\mathbb{D}$ with one vertex in 0 and the angle there equal to α. Show that $\lim_{\alpha \to 0^+} A(T_\alpha) = \pi$, even though the triangles degenerate to a ray in the limit.

8.6. Let T be an asymptotic quadrilateral in $\mathbb{D}$ with one finite vertex with angle α and three ideal vertices.

a) Find a formula for the area of T and show that the area only depends on α.

b) Does α determine T up to congruence?

9. Trigonometry in the hyperbolic plane

In Euclidean geometry fundamental roles are played by the formulae known as the *Law of Sines* and the *Law of Cosines.* For instance, these formulae imply that certain combinations of three angles or sides in a triangle determine the triangle up to congruence ("congruence criteria") — in fact, they even provide simple ways of computing the remaining angles and sides. In this section we will derive similar formulae for *hyperbolic* triangles. For this it will be convenient to use Poincaré's disk model $\mathbb{D}$ for the hyperbolic plane, but the resulting formulae will be independent of which model we use. In particular, they will also hold in $\mathbb{H}$.

We first consider *finite* triangles, i. e. triples of points in the plane (the *vertices*) which do not lie on a common hyperbolic line. Each pair of vertices spans a segment, a line and two rays. The segments are the *sides* of the triangle, and the *angle* at a vertex is the pair of rays having this vertex in common. An arbitrary such triangle is congruent to one which has one vertex in 0 and another on the positive real axis, and where the third vertex has positive imaginary part: if u, z and w are the vertices, we can move u to 0, rotate such that z lands on the positive real axis, and, if necessary, use complex conjugation to obtain $\operatorname{Im} w > 0$. We then say that the triangle is in "standard position". (But note that for any given triangle there are six possible ways of doing this.) The angles and sides of this new triangle will be congruent to the corresponding angles and sides of the original triangle, so if we want to study relations between their sizes, we may assume that the triangle is in standard position.

Admitting a slight lack of precision, we will simplify our terminology and use the word 'side' interchangeably for a segment and its length, measured in the hyperbolic metric. Similarly, an 'angle' can mean both the actual

angle and its size, measured in radians. This is all in accordance with the usual (abuse of) language in Euclidean geometry.

Our general triangle will then have angles α, β and γ, and we denote the opposite sides by a, b and c, respectively. We may assume that α is the angle at 0 and β is the other angle on the real axis. The vertex on the real axis may be identified with a real number $r \in (0,1)$ and the third vertex can be written $w = se^{i\alpha}$, where $s \in (0,1)$ and $\alpha \in (0,\pi)$. Then the distance formula (7.5) yields:

$$\begin{aligned}\cosh a &= 1 + 2\frac{|r - se^{i\alpha}|^2}{(1-r^2)(1-s^2)} = 1 + 2\frac{r^2+s^2-2rs\cos\alpha}{(1-r^2)(1-s^2)} \\ &= \frac{1 - r^2 - s^2 + r^2s^2 + 2r^2 + 2s^2 - 4rs\cos\alpha}{(1-r^2)(1-s^2)} \\ &= \frac{1+r^2}{1-r^2}\,\frac{1+s^2}{1-s^2} - \frac{2r}{1-r^2}\,\frac{2s}{1-s^2}\,\cos\alpha.\end{aligned}$$

But $r = \tanh(c/2)$ (7.4), and therefore

$$\frac{1+r^2}{1-r^2} = \frac{1+\tanh^2(c/2)}{1-\tanh^2(c/2)} = \cosh c\,, \text{ and}$$
$$\frac{2r}{1-r^2} = \frac{2\tanh(c/2)}{1-\tanh^2(c/2)} = \sinh c\,.$$

Similar formulae hold for s and b, and, substituting in the above expression for $\cosh a$, we have proved

Proposition 9.1. *(The first Law of Cosines.)*

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos\alpha.$$

(Obviously there are two more such relations, obtained by permuting the vertices.)

Corollary 9.2. *(The hyperbolic Pythagorean Theorem.)*
If $\alpha = \pi/2$, *then* $\cosh a = \cosh b \cosh c$.

To see the relationship with the classical Pythagorean Theorem, we substitute the Taylor series for cosh:

$$1 + \frac{a^2}{2} + \frac{a^4}{4!} + \cdots = (1 + \frac{b^2}{2} + \frac{b^4}{4!} + \cdots)(1 + \frac{c^2}{2} + \frac{c^4}{4!} + \cdots)\,.$$

Multiplying out the parentheses and solving with respect to the a^2–term, we see that $a^2 = b^2 + c^2 +$ {terms of order at least 4}. Hence, for small triangles this is approximately the Euclidean Pythagorean Theorem.

Now we put $A = \cosh a$, $B = \cosh b$ and $C = \cosh c$. Then $\sinh a = \sqrt{A^2-1}$, etc., and the first law of cosines may be written

$$\sqrt{B^2-1}\sqrt{C^2-1}\cos\alpha = BC - A,$$

or, squaring on both sides:

$$(B^2-1)(C^2-1)(1-\sin^2\alpha) = (BC-A)^2\,.$$

We can solve this equation for $\sin^2\alpha$:

$$\sin^2\alpha = \frac{(B^2-1)(C^2-1)-(BC-A)^2}{(B^2-1)(C^2-1)} = \tag{9.1}$$
$$\cdots$$
$$= \frac{2\,ABC - A^2 - B^2 - C^2 + 1}{(B^2-1)(C^2-1)}\,.$$

Consequently, the quotient

$$\frac{\sin^2\alpha}{\sinh^2 a} = \frac{\sin^2\alpha}{A^2-1} = \frac{2\,ABC - A^2 - B^2 - C^2 + 1}{(A^2-1)(B^2-1)(C^2-1)}\,,$$

is completely symmetric in A, B, C. Hence we get exactly the same if we replace (α, a) by (β, b) or (γ, c), so we have shown:

$$\frac{\sin^2\alpha}{\sinh^2 a} = \frac{\sin^2\beta}{\sinh^2 b} = \frac{\sin^2\gamma}{\sinh^2 c}\,.$$

Since $\sin y$ is positive for $y \in (0,\pi)$ and $\sinh x$ is positive for all $x > 0$, we have proved

Proposition 9.3. *(The hyperbolic Law of Sines.)*

$$\frac{\sin\alpha}{\sinh a} = \frac{\sin\beta}{\sinh b} = \frac{\sin\gamma}{\sinh c}\,.$$

Since $\sinh x \approx x$ for small x, we see that for small triangles this is approximately the Euclidean sine relation.

For hyperbolic triangles there is also another cosine formula:

Proposition 9.4. *(The second Law of Cosines.)*

$$\cos\alpha = -\cos\beta\cos\gamma + \sin\beta\sin\gamma\cosh a.$$

(Again we obtain two additional formulae by permuting $\alpha,\ \beta$ and γ.)

Proof. From the first law of cosines we may write

$$\cos\alpha = \frac{BC - A}{\sqrt{(B^2-1)(C^2-1)}}\,,$$

and similarly for $\cos\beta$ and $\cos\gamma$. Substituting these three expressions, we get

$$\cos\alpha + \cos\beta\cos\gamma =$$
$$\frac{BC-A}{\sqrt{(B^2-1)(C^2-1)}} + \frac{(AC-B)(AB-C)}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}} =$$
$$\cdots = A\frac{1 - A^2 - B^2 - C^2 + 2\,ABC}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}}\,.$$

Using expressions analogous to 9.1 for $\sin\beta$ and $\sin\gamma$, we obtain

$$\sin\beta\sin\gamma = \frac{1 - A^2 - B^2 - C^2 + 2\,ABC}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}}\,.$$

Hence we have

$$\cos\alpha + \cos\beta\cos\gamma = A\sin\beta\sin\gamma = \sin\beta\sin\gamma\cosh a\,.$$

□

The first law of cosines is analogous to the classical, Euclidean version. One consequence is the SSS ('side–side–side') congruence criterion, stating that the lengths of all three sides determine all the angles; hence the whole triangle up to congruence. An important result that follows from this (as it does also in Euclidean geometry) is:

Proposition 9.5. *Congruence of angles can be characterized in terms of congruence of segments.*

Proof. let r, s be rays with vertex A and r', s' rays with vertex A'. Choose vertices $B \in r$ and $C \in s$, both different from A. Using aciom $C1$ we can now find points $B' \in r'$ and $C' \in s'$ such that $A'B' \cong AB$ and $A'C' \cong AC$. Then it follows from the SSS–criterion that $\angle(r,s) \cong \angle(r',s')$ if and only if $B'C' \cong BC$. □

The second cosine relation, however, does not really have a counterpart in Euclidean geometry. One striking consequence is that if we know all the angles of a triangle, the *sides* are also completely determined, and hence the whole triangle, up to congruence. Hence, in hyperbolic geometry similar triangles are congruent! This is the AAA congruence criterion, which also holds in *spherical geometry,* but definitely not in Euclidean geometry.

Note that this observation complements the area formula in Proposition 8.5, which says that the area is determined by the *sum* of the angles. In fact, if anything, the second law of cosines is another replacement for the result in Euclidean geometry saying that the sum of the angles in a triangle is π, in the sense that this is what it approximates for "small" triangles. To see this, note that the formula can be rewritten as

$$\cos(\beta+\gamma) - \cos(\pi-\alpha) = \sin\beta\sin\gamma(\cosh a - 1)\,.$$

If $a \to 0$, the right hand side will also go to 0. But cos is decreasing, hence injective, in the interval $(0,\pi)$ containing both $\pi-\alpha$ and $\beta+\gamma$, so this means that $\alpha+\beta+\gamma \to \pi$ as $a \to 0$. Hence $\alpha+\beta+\gamma \approx \pi$ for small triangles.

Because of the angle–sum formula in Euclidean geometry, two angles of a triangle determine the third. This is not true in hyperbolic geometry, but the second law of cosines says that two angles *and* the side between them determine the third angle. In both geometries the law of sines then determines the remaining two sides, and hence the whole triangle up to congruence. This is known as the ASA congruence criterion ('angle–side–angle').

It should be remarked that these congruence criteria, stating that certain combinations of three quantities determine the triangle up to congruence, can be proved geometrically from the axioms. (Note that axiom C6 is the congruence criterion SAS.) In fact, except for the AAA–criterion, which is only valid in hyperbolic geometry, this can be done without using any parallel axiom. Hence the same geometric proofs are valid in both Euclidean

and hyperbolic geometry. But the trigonometric formulae are needed in order to calculate the remaining quantities (angles and sides).

We conclude this section with some remarks on asymptotic triangles. These can be thought of as limiting positions of finite triangles as one or more vertices move to infinity. Recall that we call such vertices *ideal* vertices, and we call a triangle *simply, doubly* or *triply* asymptotic, depending on how many ideal vertices it has. The (size of the) *angle* at an ideal vertex is defined to be 0.

Let us consider more closely the three types of asymptotic triangles:

(i) Triply asymptotic. All three sides are then hyperbolic lines, and any two of these lines have a common endpoint in $\mathbb{R}$. By Exercise 2.7 (essentially Corollary 2.5), any triple of points can be mapped to any other triple by an element of $M\ddot{o}b(\mathbb{H})$. Hence any two triply asymptotic triangles are congruent.

(ii) Doubly asymptotic. Two of the sides are rays and the third is a hyperbolic line between their endpoints. Consider the triangle in the disk model $\mathbb{D}$. If we move the finite vertex to 0 by a Möbius transformation, we see that the two rays become radii in $\mathbb{D}$. Clearly the angle at 0 then determines the triangle up to congruence.

(iii) Simply asymptotic. Two of the sides are rays and the third is a finite segment of length c, say, between two finite vertices with angles α and β. The third angle is $\gamma = 0$.

Passing to the limit as $\gamma \to 0$, the second cosine relation will continue to hold, and we get

$$1 = \cos(0) = -\cos\alpha\cos\beta + \sin\alpha\sin\beta\cosh c\,.$$

This equation determines the third of the parameters α, β and c if the two others are given. For example, we have

$$\cosh c = \frac{1+\cos\alpha\cos\beta}{\sin\alpha\sin\beta}\,.$$

(See also Exercise 9.5.)

An important special case is when one of the angles, e. g. β, is $\pi/2$. (See Figure 8.) Then

$$\cosh c = \frac{1}{\sin\alpha}\,.$$

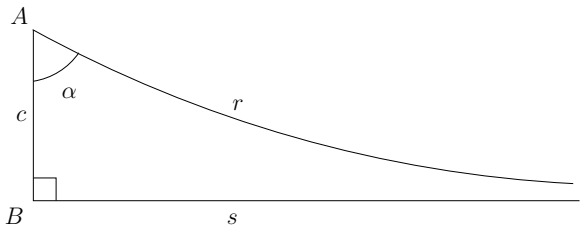

FIG. 8

This relation is usually given a different form, obtained by solving the equation with respect to e^{-c}:

$$\mathrm{e}^{-c} = \frac{1-\cos\alpha}{\sin\alpha} = \tan\frac{\alpha}{2}\,.$$

These relations are possibly the simplest manifestations of the close relationship between units of measurement of angles and lengths in the hyperbolic plane. In fact, using Lemma 6.1 it can be considered as another version of Proposition 9.5.

Some classical terminology: let ℓ be the line containing the ray s. The ray r is one of two *limiting parallel rays* to ℓ through the point A. The lines containing the limiting parallel rays are called *asymptotic parallel lines* to ℓ, and the angle α is the *asymptotic angle*. Parallel lines that are not asymptotic are called *ultraparallel*.

One final word: We have now developed enough of hyperbolic geometry to see that it differs dramatically from Euclidean geometry in the large. However, several results show that in the small, i. e. locally around a point, the geometries become approximations of each other. Examples are the calculations of area and circumference of a circle, and the remarks on the trigonometric formulae above. This is analogous to the fact that although the Earth is a sphere, locally it looks flat, and the errors we make by using Euclidean geometry on small regions are usually very small.

Exercises.

9.1. In a triangle two of the angles are α and β, and the length of the side opposite to the vertex with angle β is b. Explain how to find the remaining angles and sides.

Why does this establish the congruence criterion SAA?

9.2. Show that two of the angles of a finite triangle are equal if and only if the two sides opposite to these angles have the same length.

9.3. Use the hyperbolic sine relation to prove that in a hyperbolic triangle the greatest angle has the longest opposite side.

9.4. Show that the sides of a triangle all have the same length if and only if the three angles also are equal.

Show that if the sides have length a and the angles are α, then $2\cosh(a/2)\sin(\alpha/2) = 1$. (Hint: cut the triangle into two pieces.)

9.5. Show that if a simply asymptotic triangle has finite angles α and β and finite side c, then β is determined by α and c.

9.6. Suppose given a *quadrilateral* with one ideal vertex and three right angles. Then two of the sides have finite lengths a and b. Show that

$$\frac{1}{\cosh^2 a} + \frac{1}{\cosh^2 b} = 1\,.$$

9.7. Show that in a triangle with $\gamma = \pi/2$ the following formulae hold:

$$\begin{aligned}\sin\alpha &= \frac{\sinh a}{\sinh c}\\ \cos\alpha &= \frac{\tanh b}{\tanh c}\end{aligned}$$

9.8. Formulate and prove a 'converse' of Proposition 9.5.

9.9. Prove that a map $\mathbb{H} \to \mathbb{H}$ is a Möbius transformation if and only if it is distance preserving.

Appendix. Remarks on the Beltrami–Klein model

Since the Beltrami–Klein model played such an important part at the beginning of these notes, we should not end the discussion of hyperbolic geometry without some remarks on the missing bits of its geometry.

When we left it, we had all ingredients except congruence. Now we can define congruence as equivalence under transformations of the form $H^{-1}gH$, where H is our identification $\mathbb{K} \approx \mathbb{D}$ and $g \in \mathit{Möb}(\mathbb{D})$. This can be written out explicitly, but the formulae are ugly and not very enlightening. However, by Proposition 9.5 we now also know that congruence is completely determined by the distance measure, so we might instead ask what the distance formula looks like when transported back to $\mathbb{K}$. It turns out that this question does indeed have a nice and interesting answer.

Recall that the set of points of $\mathbb{K}$ is the interior of the unit disk in $\mathbb{R}^2$, and the 'lines' of the geometry are the chords in this disk.

Let Z_1, Z_2 be two points of $\mathbb{K}$. They span a unique chord which has endpoints P and Q on the boundary circle of $\mathbb{K}$ in $\mathbb{R}^2$. Let us denote the distance function on $\mathbb{K}$ by $d_{\mathbb{K}}$.

Proposition A.1. $\quad d_{\mathbb{K}}(Z_1, Z_2) = \dfrac{1}{2}|\ln|[Z_1, Z_2, P, Q]||.$

Proof. The identification $H : \mathbb{K} \approx \mathbb{D}$ is the composition of the vertical projection from $\mathbb{K}$ to the upper hemisphere $\mathbb{B}$ with equator disk $\mathbb{K}$ and stereographic projection from $\mathbb{B}$ to $\mathbb{D}$. Note that $\mathbb{K}$ and $\mathbb{D}$ coincide as sets.

Let $W_1 = H(Z_1)$ and $W_2 = H(Z_2)$. The endpoints P and Q are left fixed, so we have by definition

$$d_{\mathbb{K}}(Z_1, Z_2) = d_{\mathbb{D}}(W_1, W_2) = |\ln|[W_1.W_2, P, Q]||.$$

Hence the result follows from

Claim 1: $\quad [W_1.W_2, P, Q]^2 = [Z_1, Z_2, P, Q].$

Now we simplify notation and denote by AB both the Euclidean segment from A to B and its Euclidean length $|A - B|$. Then the equation in Claim 1 reads

$$\left(\frac{W_1P}{W_1Q}\frac{W_2Q}{W_2P}\right)^2 = \frac{Z_1P}{Z_1Q}\frac{Z_2Q}{Z_2P}.$$

Hence Claim 1 is a consequence of

Claim 2: Let Z be a point on the chord with endpoints P and Q, and let $W = H(Z)$. Then $\left(\dfrac{WP}{WQ}\right)^2 = \dfrac{ZP}{ZQ}$.

We will prove this by a geometric argument entirely within Euclidean geometry, referring to Figure 9. Here O is the center of the unit disk, and we have drawn both the $\mathbb{K}$–line and the $\mathbb{D}$–line with endpoints P and Q. Note that the $\mathbb{D}$–line has the radii OP and OQ as tangents. By the definition of π and Φ the point W lies on the line between Z and O, so W must be the point of intersection between the $\mathbb{D}$–line and Euclidean segment OZ.

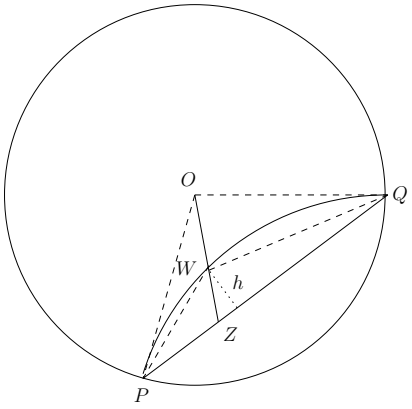

FIG. 9

Consider now the triangles ZPO and ZQO. With ZP and ZQ as base segments the two triangles have the same heights, and we have

$$\frac{ZP}{ZQ} = \frac{A(ZPO)}{A(ZQO)},$$

where the right hand side is the ratio between the areas of the triangles.

Likewise, the triangles ZPW and ZQW have the same heights, so we deduce

$$\frac{ZP}{ZQ} = \frac{A(ZPW)}{A(ZQW)} = \frac{A(ZPO) - A(ZPW)}{A(ZQO) - A(ZQW)} = \frac{A(PWO)}{A(QWO)}.$$

But the areas in the last fraction can also be calculated with the two radii OP and OQ as bases, yielding

$$\frac{ZP}{ZQ} = \frac{\frac{1}{2}OP \cdot PW \sin(\angle WPO)}{\frac{1}{2}OQ \cdot QW \sin(\angle WQO)} = \frac{PW \sin(\angle WPO)}{QW \sin(\angle WQO)}.$$

The $\mathbb{D}$–line through P, W, Q is part of a Euclidean circle, hence we can use the theorem of peripheral angles to obtain

$$\angle WPO = \angle WQZ \quad \text{and} \quad \angle WQO = \angle WPZ.$$

But from the figure we see that

$$PW \sin(\angle WPZ) = h = QW \sin(\angle WQZ).$$

Combining these observations, we get

$$\frac{PW}{QW} = \frac{\sin(\angle WQZ)}{\sin(\angle WPZ)} = \frac{\sin(\angle WPO)}{\sin(\angle WQO)}.$$

Substituting this in the expression for $\dfrac{ZP}{ZQ}$ above proves Claim 2. □

Remark. Comparing the formula in Proposition A.1 with the formula for $d_{\mathbb{D}}$ in (7.3), it might look as if $d_{\mathbb{K}}$ is just a rescaling of $d_{\mathbb{D}}$. However, this is not the case. The reason is that the points P, Q (p, q) are not the same in the two formulae — in Proposition A.1 they are the endpoints of the chord through Z_1, Z_2, but in (7.3) they are endpoints of the $\mathbb{D}$–line through the same points. The only case where they coincide is when Z_1 and Z_2 lie on a common diameter. Then we have $d_{\mathbb{K}}(Z_1, Z_2) = \frac{1}{2} d_{\mathbb{D}}(Z_1, Z_2)$.

The Beltrami–Klein model is not conformal, and angle measure is much more complicated. However, one fact is worth noting: it is easy to draw all $\mathbb{K}$–lines which are *orthogonal* to a given line! We distinguish between two cases:

(1) If the given line is a diameter, the $\mathbb{K}$–lines orthogonal to it are the chords that are orthogonal to it in the Euclidean sense.

(2) Assume the $\mathbb{K}$–line ℓ is not a diameter, and let its endpoints be P and Q. The tangents to the circle at ∞ at P and Q intersect in a point W outside the unit disk. The $\mathbb{K}$ lines orthogonal to ℓ are the chords contained in Euclidean lines through W. (See Figure 10a.)

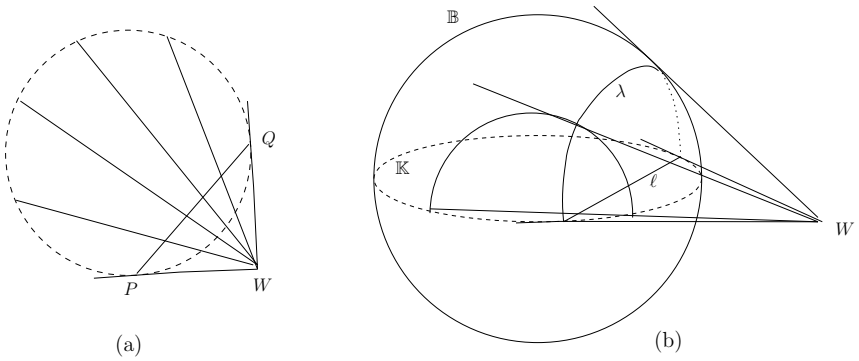

FIG. 10

Note that case (1) follows from case (2) by a limiting procedure. Therefore it suffices to consider (2). We will prove this by a geometric argument using the hemisphere model $\mathbb{B}$ and a little three–dimensional geometry. Cfr. Figure 10b. In this picture we think of $\mathbb{K}$ as lying in a standard $\mathbb{R}^2 \subset \mathbb{R}^3$, and $\mathbb{B} \subset S^2$ is the upper hemisphere.

Via vertical projection the $\mathbb{K}$–lines correspond to $\mathbb{B}$–lines, which are circular arcs (semi–circles) which lie in vertical planes and meet the equator circle at right (Euclidean) angles. In particular, ℓ correspond to one such semi–circle, which we call λ.

The reason we pass to $\mathbb{B}$ is that the hemisphere model is conformal, since it is related with $\mathbb{D}$ by stereographic projection. Therefore we now only need to determine all $\mathbb{B}$–lines which meet λ orthogonally in the Euclidean sense.

A $\mathbb{B}$–line γ meets λ orthogonally if and only if its tangent line at the intersection point does. But the union of all the lines meeting the circle containing λ ortogonally is easily seen to be a circular cone, and all the lines meet at its vertex. Since the tangents at P and Q in Figure 10a are two such lines, we see that the cone vertex is precisely the point W.

It now only remains to observe that since γ lies in a vertical plane, its tangents also lie in this plane. Hence the plane contains W, and its projection to R^2 will be a line also containing W. But the $\mathbb{K}$–line corresponding to γ is contained in this line.

Exercises

A.1. What kind of curves are 'horocircles' in $\mathbb{K}$?

A.2. Verify that inversion in a diameter in $\mathbb{K}$ is ordinary reflection in that diameter.

A.3. Show that two hyperbolic lines have a common perpendicular if and only they are ultraparallel, and show how it can be constructed in the Beltrami–Klein model.

A.4. Show that we can parametrize $\mathbb{K}$ by $(r, \theta) \mapsto (\tanh r \cos\theta, \tanh r \sin\theta)$, where r is the hyperbolic distance from the origin. ('Geodesic polar coordinates'.)