CONTINUED FRACTIONS, HYPERBOLIC GEOMETRY AND QUADRATIC FORMS
COURSE NOTES for MATH 497A REU PROGRAM, SUMMER 2001 REVISED NOVEMBER 2001

Svetlana Katok
Department of Mathematics
The Pennsylvania State University
University Park, PA 16802, U.S.A.

Lecture 1
Continued fractions

The theory of continued fractions is closely related to the Gauss reduction theory for indefinite integral quadratic forms translated into the matrix language. More precisely, quadratic forms equivalent in the *narrow sense* correspond to $SL(2,\mathbb{Z})$–equivalent matrices, while quadratic forms equivalent in the *wide sense* correspond to $GL(2,\mathbb{Z})$–equivalent matrices. The second theory can be expressed using the regular "plus" continued fractions, and the first – via the theory of "minus" continued fractions. This theory is much less known, and we will present it in detail.

We will see in subsequent lectures how convenient the minus continued fractions are in geometric applications, in particular for coding closed geodesics (see §4.2).

1.1. Theory of "minus" continued fractions

Let α be arbitrary real number. We define a sequence of integers $\{a_i\}$, $i = 0, 1, 2, \ldots$ and a sequence of real numbers $\{\alpha_i\}$, $i = 1, 2 \ldots$ by

$$a_0 = [\alpha] + 1,\ \ \alpha_1 = \frac{1}{a_0 - \alpha},$$

and inductively,

$$a_n = [\alpha_n] + 1,\ \ \alpha_{n+1} = \frac{1}{a_n - \alpha_n}. \tag{1.1.1}$$

Next we define a sequence of rational numbers

$$r_n = (a_0, a_1, \ldots, a_{n-1}, a_n) = a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{a_n}}}}, \quad n \geq 0.$$

THEOREM 1.1. *We have* $\lim_{n\to\infty} r_n = \alpha$, *i.e., any real number* α *is represented as an infinite "minus continued fraction"*

$$\alpha = (a_0, a_1, \dots, a_{n-1}, \dots) = a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{\dots}}}}.$$

PROOF. (1) We claim that $a_i \geq 2$ for all $i \geq 1$.

For $i \geq 1$ we have $\alpha_i = 1/a_{i-1} - \alpha_{i-1}$. If α_{i-1} is not an integer, we have $0 < a_{i-1} - \alpha_{i-1} < 1$ and $\alpha_i > 1$, which implies $a_i > 2$. If α_{i-1} is an integer, then $a_{i-1} - \alpha_{i-1} = 1$ and hence $a_i = 2$, which proves our claim.

(2) We now define two sequences of integers $\{p_n\}$ and $\{q_n\}$, $n \geq -2$, inductively:

$$\begin{aligned} p_{-2} = 0,\ p_{-1} = 1;\ p_i = a_i p_{i-1} - p_{i-2} \text{ for } i \geq 0, \\ q_{-2} = -1,\ q_{-1} = 0;\ q_i = a_i q_{i-1} - q_{i-2} \text{ for } i \geq 0. \end{aligned} \tag{1.1.2}$$

We will prove that $r_n = p_n/q_n$. The three following statements ((2.1)-(2.3)) are proved by induction.

(2.1) We have $1 = q_0 < q_1 < q_2 < \dots < q_n < \dots$, and so $\lim_{n\to\infty} q_n = \infty$.

We have $q_1 = a_1 q_0 - q_{-1} = a_1 \geq 2$, the basis of induction. Now assume that $1 = q_0 < q_1 \dots < q_{n-1}$.Then

$$q_n = a_n q_{n-1} - q_{n-2} \geq 2q_{n-1} = q_{n-1} + (q_{n-1} - q_{n-1}) > q_{n-1},$$

as claimed.

(2.2) Suppose that

$$(a_0, a_1, \dots, a_{n-1}, x) = a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{x}}}}$$

Then for any $x \geq 1$,

$$(a_0, a_1, \dots, a_{n-1}, x) = \frac{xp_{n-1} - p_{n-2}}{xq_{n-1} - q_{n-2}}.$$

This follows by induction from the definition of $\{p_n\}$ and $\{q_n\}$.

(2.3) We have $p_{i-1}q_i - p_i q_{i-1} = 1$ for all $i \geq 1$.

Using (1.1.2) repeatedly, we see that

$$p_{i-1}q_i - p_i q_{i-1} = p_{i-2}q_{i-1} - p_{i-1}q_{i-2} = \dots = p_{-2}q_{-1} - p_{-1}q_{-2} = 1.$$

Then using (2.2) with $x = a_n$, we obtain $r_n = p_n/q_n$.

(3) The sequence $\{r_n\}$ is monotone decreasing since, by (2.3),

$$\frac{p_n}{q_n} - \frac{p_{n+1}}{q_{n+1}} = \frac{1}{q_n q_{n+1}} > 0$$

and bounded from below by $a_0 - 1$, hence has a limit, which is a real number.

In order to prove that this limit is α, we write $\alpha = (a_0, a_1, \dots, a_{n-1}, \alpha_n)$, where α_n is defined in (1.1.1). Then, using (2.3) again, we obtain

$$\begin{aligned}\frac{p_{n-1}}{q_{n-1}} - \alpha &= \frac{p_{n-1}}{q_{n-1}} - \frac{\alpha_n p_{n-1} - p_{n-2}}{\alpha_n q_{n-1} - q_{n-2}} \\ &= \frac{1}{q_{n-1}(\alpha_n q_{n-1} - q_{n-2})} \leq \frac{1}{q_{n-1}} \to 0.\end{aligned}$$

Thus, $\lim_{n\to\infty} r_n = \alpha$. □

Conversely, let $a_0, a_1, \dots, a_n, \dots$ be an infinite sequence of integers with $a_1, a_2, \dots \geq 2$. We define two sequences of integers $\{p_n\}$ and $\{q_n\}$, $n \geq -2$ by (1.1.2) and prove as in Theorem 1.1 that

$$r_n = \frac{p_n}{q_n} = (a_0, a_1, \dots, a_n),$$

and that the sequence $\{r_n\}$ tends to a limit, namely to the real number α with $a_0 = [\alpha] + 1$. If we apply the same procedure to the sequence $a_1, a_2, \dots$, we obtain, as the limit of the corresponding sequence, the real number

$$\alpha_1 = \lim_{n\to\infty} (a_1, a_2, \dots, a_n).$$

By the properties of limits, we conclude that

$$\alpha_1 = \frac{1}{a_0 - \alpha},$$

and by induction obtain the recursive relations (1.1.1).

Thus there is a one–to–one correspondence between the set of real numbers α and the set of infinite sequences $a_0, a_1, \dots, a_n, \dots$ with integers $a_0 \in \mathbb{Z}$ and $a_i \geq 2$ for $i \geq 1$. Using this correspondence, we will prove the following statement.

Theorem 1.2. *A real number α is rational if and only if from some point on all the a_i in its infinite minus continued fraction are equal to* 2 *(i.e., if there exists a positive integer n such that $\alpha_k = 2$ for all $k \geq n$).*

Proof. First we notice that

$$(2, 2, \dots) = 1,$$

since it is a limit of the sequence $2, 3/2, 4/3, \dots n+1/n, \dots$, and this sequence tends to 1.

If $\alpha = n \in \mathbb{Z}$, we have $n = (n+1, 2, 2, \dots)$.

Now assume $\alpha \in \mathbb{Q}$, is not an integer, i.e. $\alpha = c_0/d_0$, then the "tails" are also rational numbers written in the least terms,

$$\alpha_1 = (a_1, a_2, \dots) = \frac{c_1}{d_1},$$

$$\alpha_2 = (a_2, a_3, \dots) = \frac{c_2}{d_2},$$

$$\dots\dots$$

$$\alpha_n = (a_n, a_{n+1}, \dots) = \frac{c_n}{d_n},$$

$$\dots\dots\dots$$

We want to prove that if α_N is an integer for some N, then it has a tail of 2's. Assume not, i.e., suppose all the $\alpha_n = c_n/d_n$ are not integers, i.e., all the d_n are greater than 1. Since $a_n \geq 2$ for $n \geq 1$, we have $c_n/d_n > 1$. We will show that $d_0 > d_1 > \dots$, and therefore for some N we must have $d_N = 1$. But

$$\frac{c_0}{d_0} = a_0 - \frac{1}{\frac{c_1}{d_1}} = a_0 - \frac{d_1}{c_1}, \text{ i.e., } \frac{c_0}{d_0} + \frac{d_1}{c_1} = a_0.$$

Similarly

$$\frac{c_n}{d_n} + \frac{d_{n+1}}{c_{n+1}} = a_n, \text{ or } c_n c_{n+1} + d_n d_{n+1} = a_n d_n c_{n+1}.$$

Since c_{n+1} divides the other two terms of the above formula, we conclude that $c_{n+1} | d_n d_{n+1}$. But $(c_{n+1}, d_{n+1}) = 1$, therefore $c_{n+1} | d_n$. Hence we conclude that $c_{n+1} < d_n$. Using the fact that $c_{n+1} > d_{n+1}$ again, we obtain $d_{n+1} < d_n$.

Conversely, if α has a tail of 2's, we have $\alpha = (a_0, a_1, \dots, 1)$, hence α is a rational number. □

DEFINITION. A real number is called a *quadratic irrationality* if it is a real root of the quadratic equation $ax^2 + bx + c$ with coefficients $a, b, c \in \mathbb{Z}$, $c \neq 0$, while the discriminant $D = b^2 - 4ac$ is positive and is not a perfect square.

THEOREM 1.3. *A real number α is a quadratic irrationality if and only if its minus continued fraction expansion is eventually periodic with the periodic part being anything but a repeated* 2.

PROOF. Let α be a quadratic irrationality. Then it can be written in the form

$$\alpha = \alpha_0 = \frac{m_0 + \sqrt{D}}{\ell_0},$$

where $m_0, \ell_0, D \in \mathbb{Z}$, $\ell_0 \neq 0$, and $D > 0$ is not a square. Let

$$\alpha = (a_0, a_1, \dots, a_n, \dots) = a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{a_n - \cfrac{1}{\ddots}}}}}.$$

Then α_n is the “tail” of the minus continued fraction for α,

$$\alpha_n = (a_n, a_{n+1}, \dots) = a_n - \cfrac{1}{a_{n+1} - \cfrac{1}{\ddots}},$$

and $\alpha = (a_0, a_1, \dots, a_{n-1}, \alpha_n)$. It is proved by induction that all the α_n are quadratic irrationalities of the form $\alpha_n = (m_n + \sqrt{D})/\ell_n$ with integral m_n and ℓ_n, $\ell_n \neq 0$, which satisfy the following recurrent relations:

$$m_{n+1} = a_n\ell_n - m_n, \;\; \ell_{n+1}\ell_n = m_{n+1}^2 - D. \tag{1.1.3}$$

Let $\alpha_n' = (m_n - \sqrt{D})/\ell_n$. Using (2.2) of Theorem 1.1, we can write

$$\alpha_0 = (a_0, a_1, \dots, a_{n-1}, \alpha_n) = \frac{\alpha_n p_{n-1} - p_{n-2}}{\alpha_n q_{n-1} - q_{n-2}}.$$

Taking the conjugates of both sides, we obtain,

$$\alpha_0' = (a_0, a_1, \dots, a_{n-1}, \alpha_n') = \frac{\alpha_n' p_{n-1} - p_{n-2}}{\alpha_n' q_{n-1} - q_{n-2}}.$$

Solving for α_n' we obtain

$$\alpha_n' = \frac{\alpha_0' q_{n-2} - p_{n-2}}{\alpha_0' q_{n-1} - p_{n-1}} = \frac{q_{n-2}}{q_{n-1}} \cdot \frac{\alpha_0' - \frac{p_{n-2}}{q_{n-2}}}{\alpha_0' - \frac{p_{n-1}}{q_{n-1}}}.$$

Since both $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_{n-2}}{q_{n-2}}$ tend to $\alpha_0 \neq \alpha_0'$, the second fraction tends to 1. Since the sequence $\{r_n\} = \{p_n/q_n\}$ is monotone decreasing and the sequence $\{q_n\}$ is increasing, the second fraction tends to 1 from below, and we see that $\frac{q_{n-2}}{q_{n-1}} < 1$. Thus we conclude that there exists an $N > 1$ such that for all $n > N$ we have $0 < \alpha_n' < 1$. Since $a_n \geq 2$, we have $\alpha_n > 1$. It follows that

$$0 < \frac{m_n - \sqrt{D}}{\ell_n} < 1, \;\; \frac{m_n + \sqrt{D}}{\ell_n} > 1. \tag{1.1.4}$$

Since $\alpha_n - \alpha_n' = 2\sqrt{D}/\ell_n > 0$, we conclude that $\ell_n > 0$, and therefore (1.1.4) implies that $|m_n - \ell_n| < \sqrt{D}$, and hence can take only finitely many values for a given D. We have $D - (m_n - \ell_n)^2 > 0$, and this expression also can take only finitely many values. Using (1.1.3), we can write

$$D - (m_n - \ell_n)^2 = D - m_n^2 - \ell_n^2 + 2m_n\ell_n = -\ell_n\ell_{n-1} - \ell_n^2 + 2m_n\ell_n$$
$$= \ell_n(-\ell_{n-1} - \ell_n + 2m_n).$$

Thus $\ell_n | (D - (m_n - \ell_n)^2)$, hence ℓ_n takes only finitely many values, and so does m_n. Therefore, for some $j \neq k$, $\alpha_j = \alpha_k$, i.e., the "tails" of α coincide. More precisely, this implies $a_j = [\alpha_j] + 1 = [\alpha_k] + 1 = a_k$, and so on, i.e., the minus continued fraction expansion of α is eventually periodic. Then the periodic part cannot be a repeated 2 since in this case α would be rational by Theorem 1.2.

Conversely, it is easy to see that if α has an eventually periodic continued fraction expansion, it is a root of a quadratic equation with integer coefficients. Since the periodic part of its minus continued fraction expansion is anything but a repeated 2, α is irrational by Theorem 1.2, so the second root is irrational as well, and α is a quadratic irrationality. $\square$

THEOREM 1.4. *Let α be a quadratic irrationality. Then α has a purely periodic minus continued fraction expansion if and only if $\alpha > 1$, $0 < \alpha' < 1$, where α' is the second root of the same quadratic equation $ax^2 + bx + c = 0$.*

PROOF. If α is purely periodic, $\alpha = (\overline{a_1, \ldots, a_r})$, then $\alpha > 1$ since $a_1 \geq 2$. We continue this periodic sequence by setting $a_{i+r} = a_i$ for all $i \in \mathbb{N}$. Then we can define

$$x_i = \frac{1}{(\overline{a_{i-1}, \ldots, a_{i-r}})}.$$

Then $1/(x_{i+1}) = a_i - x_i$, or $x_i = a_i - 1/(x_{i+1})$. It follows that

$$x_1 = \frac{1}{(\overline{a_r, \ldots, a_1})}$$

satisfies the equation

$$x_1 = (a_1, a_2, \ldots, a_{r-1}, x_1) = a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{a_{r-1} - \cfrac{1}{x_1}}}},$$

which is the same quadratic equation that α satisfies, but $0 < x_1 < 1$, and so x_1 is different from α. Therefore $x_1 = \alpha'$, so that $0 < \alpha' < 1$.

Conversely, let $\alpha > 1$ and $0 < \alpha' < 1$. We first notice that in the proof of Theorem 1.3 we already showed that if α is reduced, then it is eventually periodic, i.e., that there exist a j, $j \neq k$, such that $a_j = a_k$, $a_{j+1} = a_{k+1}$, and so on. It remains to show that it is actually purely periodic, i.e., that $a_{j-1} = a_{k-1}$ as well. We have

$$\alpha_{i+1} = \frac{1}{a_i - \alpha_i}.$$

We have seen in the proof of Theorem 1.3 that all "tails" α_i, $\alpha_1 = \alpha$, are also quadratic irrationalities with the same D as α, therefore

$$\alpha'_{i+1} = \frac{1}{a_i - \alpha'_i}. \tag{1.1.5}$$

We claim that for all $i \geq 1$, $0 < \alpha'_i < 1$. This is seen by induction: for $i = 1$, $0 < \alpha'_1 = \alpha' < 1$. Suppose $0 < \alpha'_i < 1$. Then $a_i - \alpha'_i > 1$, and by (1.1.5), we have $0 < \alpha'_{i+1} < 1$. Since

$$a_{i-1} = \frac{1}{\alpha'_i} + \alpha'_{i-1},$$

using the claim above we see that

$$a_{i-1} = \left[\frac{1}{\alpha'_i}\right] + 1,$$

where $[\cdot]$ denotes the integer part of a number. This allows us to conclude that $\alpha_j = \alpha_k$ implies $a_{j-1} = a_{k-1}$. $\square$

1.2. Theory of "plus" continued fractions

This theory is very similar, and is presented as a series of problems.

Let $a_0, a_1, \ldots, a_n, \ldots$ be an infinite sequence of positive integers. We define two sequences of integers $\{p_n\}$ and $\{q_n\}$, $n \geq -2$, inductively:

$$\begin{aligned} p_{-2} = 0,\ p_{-1} = 1;\ p_i = a_i p_{i-1} + p_{i-2} \text{ for } i \geq 0, \\ q_{-2} = 1,\ q_{-1} = 0;\ q_i = a_i q_{i-1} + q_{i-2} \text{ for } i \geq 0 \end{aligned} \tag{1.2.1}$$

1. Prove that $1 = q_0 \leq q_1 < q_2 < \cdots < q_n < \ldots$.

Now we define

$$\langle a_0, a_1, \ldots, a_{n-1}, x\rangle = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{x}}}}$$

2. Prove that for any $x \geq 1$ we have

$$\langle a_0, a_1, \ldots, a_{n-1}, x\rangle = \frac{x p_{n-1} + p_{n-2}}{x q_{n-1} + q_{n-2}}.$$

3. Prove that $p_{i-1} q_i - p_i q_{i-1} = (-1)^i$ for $i \geq 1$.

Let $r_n = \langle a_0, \ldots, a_n\rangle$. By (1.2.1) and Problem 2, we obtain

$$r_n = \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}} = \frac{p_n}{q_n}. \tag{1.2.2}$$

4. Show that $\{r_n\}$ is a sequence such that

$$r_0 < r_2 < r_4 < \cdots < r_5 < r_3 < r_1,$$

that the limit $\lim_{n\to\infty} r_n$ exists, and this limit is a real number.

Conversely, let α be any real number. We define a sequence of integers $\{a_i\}$, $i = 0, 1, 2, \ldots$ and a sequence of real numbers $\{\alpha_i\}$, $i = 1, 2 \ldots$, inductively:

$$a_0 = [\alpha],\ \alpha_1 = \frac{1}{\alpha - a_0}\ \ a_n = [\alpha_n],\ \alpha_{n+1} = \frac{1}{\alpha_n - a_n}.$$

This process will terminate if α_n is an integer for some n.

5. Prove that $a_i \geq 1$ for those $i \geq 1$ for which a_i is defined.

Now we can define a sequence $r_n = p_n/q_n$ as in (1.2.2), which can be infinite or finite, as explained above.

6. Prove that if the sequence $\{r_n\}$ is infinite, then $\lim_{n\to\infty} r_n = \alpha$; if the sequence is finite, $\{r_0, r_1, \ldots, r_n\}$, then $\alpha = r_n$.

Moreover, we deduce that the *convergents* $r_n = p_n/q_n$ are the best approximations of α:

7. If α is irrational, then

$$\left|\alpha - \frac{p_n}{q_n}\right| \leq \frac{1}{q_n^2}.$$

Thus there is a one–to–one correspondence between the set of real numbers α and the set of sequences (finite and infinite) $a_0, a_1, \ldots, a_n, \ldots$ with integers $a_0 \in \mathbb{Z}$ and $a_i \geq 1$ for $i \geq 1$. Using this correspondence, prove the following statements.

8. The real number α is rational if and only if the algorithm described above terminates, i.e., α_n is an integer for some n.

9. The real number α is a quadratic irrationality, i.e., a real root of the equation $ax^2 + bx + c$ with coefficients $a, b, c \in \mathbb{Z}$, $c \neq 0$, that has two distinct real roots, if and only if its continued fraction expansion is eventually periodic.

Problem 8 can be solved using the Euclidean algorithm. Problem 9 can be solved by an argument similar to the one used to prove Theorem 1.3. It is pretty standard and appears in textbooks on Number Theory treating the "plus" continued fractions.

There is a neat formula transforming one type of continued fractions to the other.

10. Let $\langle m_0, m_1, m_2, \dots \rangle$ ($m_i \in \mathbb{Z}$, $m_1, m_2, \dots \geq 1$) be a "plus" continued fraction expansion of a real number. Show that its minus continued fraction expansion is given by the formula on the right:

$$\langle m_0, m_1, m_2, \dots \rangle = (m_0 + 1, \underbrace{2, 2 \dots, 2}_{m_1 - 1}, m_2 + 2, \underbrace{2, 2 \dots, 2}_{m_3 - 1}, m_4 + 2, \dots).$$

Lecture 2
Hyperbolic geometry

2.1. The hyperbolic plane

Our main object of study in this section will be the upper half–plane $\mathcal{H} = \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$. Equipped with the metric

$$ds = \frac{\sqrt{dx^2 + dy^2}}{y}, \tag{2.1.1}$$

it becomes a model of the *hyperbolic* or *Lobachevski plane*. We will see that the *geodesics* (i.e., the shortest curves with respect to this metric) will be straight lines and semicircles orthogonal to the real line

$$\mathbb{R} = \{z \in \mathbb{C} \mid \mathrm{Im}(z) = 0\}.$$

By elementary geometry considerations, one easily shows that any two points in $\mathcal{H}$ can be joined by a unique geodesic, and that from any point in $\mathcal{H}$ in any direction one can draw a geodesic. We will measure the distance between two points in $\mathcal{H}$ along the geodesic connecting them. It is clear that that any geodesic can be continued indefinitely, and that one can draw a circle centered at a given point with any given radius.

The tangent space to $\mathcal{H}$ at a point z is defined as the space of tangent vectors at z. It has the structure of a 2–dimensional real vector space or of a 1–dimensional complex vector space: $T_z\mathcal{H} \approx \mathbb{R}^2 \approx \mathbb{C}$. The Riemannian metric (2.1.1) is induced by the following inner product on $T_z\mathcal{H}$: for $\zeta_1 = \xi_1 + i\eta_1$ and $\zeta_2 = \xi_2 + i\eta_2$ in $T_z\mathcal{H}$, we put

$$\langle \zeta_1, \zeta_2 \rangle = \frac{\xi_1\xi_2 + \eta_1\eta_2}{\mathrm{Im}(z)^2},$$

which is a scalar multiple of the Euclidean inner product.

We define the *angle* between two geodesics in $\mathcal{H}$ at their intersection point z as the angle between their tangent vectors in $T_z\mathcal{H}$. Using the formula

$$\cos\varphi = \frac{\langle \zeta_1, \zeta_2 \rangle}{\|\zeta_1\|\|\zeta_2\|} = \frac{(\zeta_1, \zeta_2)}{|\zeta_1||\zeta_2|},$$

where $\|\,\|$ denotes the norm in $T_z\mathcal{H}$ corresponding to the inner product $\langle\, , \rangle$, and $|\, , |$ denotes the norm corresponding to the inner product $(\, ,)$, we see that this notion of angle measure coincides with the Euclidean angle measure.

The first four axioms of Euclid hold for this geometry. However, the fifth postulate of Euclid's *Elements*, the axiom of parallels, does not hold: there is more than one geodesic passing through the point z not lying in the geodesic L that does not intersect L (see Fig. 2.1.1).

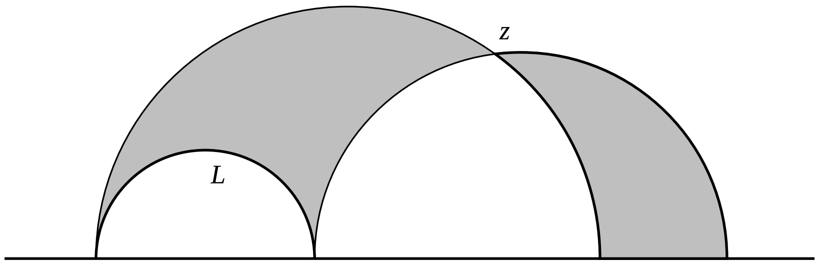

FIGURE 2.1.1. Geodesics in the upper half-plane

Therefore the geometry in $\mathcal{H}$ is *non–Euclidean*. The metric in (2.1.1) is said to be the *hyperbolic metric*. It can be used to calculate the length of curves in $\mathcal{H}$ the same way the Euclidean metric $\sqrt{dx^2 + dy^2}$ is used to calculate the length of curves on the Euclidean plane. Let $I = [0, 1]$ be the unit interval, and $\gamma : I \to \mathcal{H}$ be a piecewise differentiable curve in $\mathcal{H}$,

$$\gamma(t) = \{v(t) = x(t) + iy(t) \mid t \in I\}.$$

The length of the curve γ is defined as

$$h(\gamma) = \int_0^1 \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{y(t)} dt. \tag{2.1.2}$$

We define the *hyperbolic distance* between two points $z, w \in \mathcal{H}$ by setting

$$\rho(z, w) = \inf h(\gamma),$$

where the infimum is taken over all piecewise differentiable curves connecting z and w.

PROPOSITION 2.5. *The function $\rho : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined above is a distance function, i.e., it is*

(a) *nonnegative: $\rho(z, z) = 0$; $\rho(z, w) > 0$ if $z \neq w$;*
(b) *symmetric: $\rho(u, v) = \rho(v, u)$;*
(c) *and satisfies the triangle inequality: $\rho(z, w) + \rho(w, u) \geq \rho(z, u)$.*

PROOF. It is easily seen from the definition that (b), (c), and the first part of property (a) hold. The second part follows from Exercise 11. □

The group $SL(2, \mathbb{R})$ acts on $\mathcal{H}$ by *Möbius transformations* as follows. To each $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$, we assign a transformation

$$T_g(z) = \frac{az + b}{cz + d}. \tag{2.1.3}$$

PROPOSITION 2.6. *Any Möbius transformation T_g maps $\mathcal{H}$ into itself.*

PROOF. We can write

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

Therefore

$$\operatorname{Im}(w) = \frac{w - \overline{w}}{2i} = \frac{(ad-bc)(z-\overline{z})}{2i|cz+d|^2} = \frac{\operatorname{Im}(z)}{|cz+d|^2}. \tag{2.1.4}$$

Thus $\operatorname{Im}(z) > 0$ implies $\operatorname{Im}(w) > 0$. □

One can check directly that if $g, h \in SL(2,\mathbb{R})$, then $T_g \circ T_h = T_{gh}$ and $T_g^{-1} = T_{g^{-1}}$. It follows that each T_g, $g \in SL(2,\mathbb{R})$ is a bijection, and thus we obtain a *representation* of the group $SL(2,\mathbb{R})$ by Möbius transformations of the upper–half plane $\mathcal{H}$. In fact, the two matrices g and $-g$ give the same Möbius transformation, so formula (2.1.3) actually gives a representation of the quotient group $SL(2,\mathbb{R})/\{\pm 1_2\}$ (where 1_2 is the 2×2 identity matrix) denoted by $PSL(2,\mathbb{R})$, which we will identify with the group of Möbius transformations of the form (2.1.3). Notice that $PSL(2,\mathbb{R})$ contains all transformations of the form

$$z \to \frac{az+b}{cz+d} \quad \text{with} \quad ad - bc = \Delta > 0,$$

since by dividing the numerator and the denominator by $\sqrt{\Delta}$ we obtain a matrix for it with determinant equal to 1. In particular, $PSL(2,\mathbb{R})$ contains all transformations of the form $z \to az + b$ ($a, b \in \mathbb{R}$, $a > 0$). Since transformations in $PSL(2,\mathbb{R})$ are continuous, we have the following result.

THEOREM 2.7. *The group $PSL(2,\mathbb{R})$ acts on $\mathcal{H}$ by homeomorphisms.*

DEFINITION. A transformation of $\mathcal{H}$ onto itself is called an *isometry* if it preserves the hyperbolic distance in $\mathcal{H}$.

Isometries clearly form a group; we will denote it by $\operatorname{Isom}(\mathcal{H})$.

THEOREM 2.8. *Möbius transformations are isometries, i.e., we have the inclusion $PSL(2,\mathbb{R}) \subset \operatorname{Isom}(\mathcal{H})$.*

PROOF. Let $T \in PSL(2,\mathbb{R})$. By Theorem 2.7 T maps $\mathcal{H}$ onto itself. Let $\gamma : I \to \mathcal{H}$ be the piecewise differentiable curve given by $z(t) = x(t) + iy(t)$. Let

$$w = T(z) = \frac{az+b}{cz+d};$$

then we have $w(t) = T(z(t)) = u(t) + iv(t)$ along the curve γ. Differentiating, we obtain

$$\frac{dw}{dz} = \frac{a(cz+d) - c(az+b)}{(cz+d)^2} = \frac{1}{(cz+d)^2}. \tag{2.1.5}$$

By (2.1.4) we have

$$v = y/|cz+d|^2, \text{ therefore } |\frac{dw}{dz}| = \frac{v}{y}.$$

Thus

$$h(T(\gamma)) = \int_0^1 \frac{|\frac{dw}{dt}|dt}{v(t)} = \int_0^1 \frac{|\frac{dw}{dz}||\frac{dz}{dt}|dt}{v(t)} = \int_0^1 \frac{|\frac{dz}{dt}|dt}{y(t)} = h(\gamma).$$

The invariance of the hyperbolic distance follow from this immediately. $\square$

2.2. Geodesics

THEOREM 2.9. *The geodesics in $\mathcal{H}$ are semicircles and the rays orthogonal to the real axis $\mathbb{R}$.*

PROOF. Let $z_1, z_2 \in \mathcal{H}$. First consider the case in which $z_1 = ia, z_2 = ib$ with $b > a$. For any piecewise differentiable curve $\gamma(t) = x(t) + iy(t)$ connecting ia and ib, we have

$$h(\gamma) = \int_0^1 \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{y(t)} dt \geq \int_0^1 \frac{|\frac{dy}{dt}|dt}{y(t)} \geq \int_0^1 \frac{\frac{dy}{dt}dt}{y(t)} = \int_a^b \frac{dy}{y} = \ln \frac{b}{a},$$

but this is exactly the hyperbolic length of the segment of the imaginary axis connecting ia and ib. Therefore the geodesic connecting ia and ib is the segment of the imaginary axis connecting them.

Now consider the case of arbitrary points z_1 and z_2. Let L be the unique Euclidean semicircle or a straight line connecting them. Then by Exercise 12, there exists a transformation in $PSL(2,\mathbb{R})$ which maps L into the positive imaginary axis. This reduces the problem to the particular case studied above, so that by Theorem 2.8 we conclude that the geodesic between z_1 and z_2 is the segment of L joining them. $\square$

Thus we have proved that any two points z and w in $\mathcal{H}$ can be joined by a unique geodesic, and the hyperbolic distance between them is equal to the hyperbolic length of the geodesic segment joining them; we denote the latter by $[z,w]$. This and the additivity of the integral (2.1.2) imply the following statement.

COROLLARY 2.10. *If z and w are two distinct points in $\mathcal{H}$, then*

$$\rho(z,w) = \rho(z,\xi) + \rho(\xi,w)$$

if and only if $\xi \in [z,w]$.

THEOREM 2.11. *Any isometry of $\mathcal{H}$, in particular, any transformation in $PSL(2,\mathbb{R})$, maps geodesics into geodesics.*

PROOF. The same argument as in the Euclidean case works here. $\square$

The *cross–ratio* of distinct points $z_1, z_2, z_3, z_4 \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is defined by the following formula:

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

THEOREM 2.12. *Suppose $z, w \in \mathcal{H}$ are two distinct points, the geodesic joining z and w has endpoints $z*, w* \in \mathbb{R} \cup \{\infty\}$, and $z \in [z*, w]$. Then*

$$\rho(z, w) = \ln(w, z*; z, w*).$$

PROOF. Using Exercise 12, in $PSL(2, \mathbb{R})$ let us choose a transformation T which maps the geodesic joining z and w to the imaginary axis. By applying the transformations $z \mapsto kz$ $(k > 0)$ and $z \mapsto -1/z$ if necessary, we may assume that $T(z*) = 0$, $T(w*) = \infty$ and $T(z) = i$. Then $T(w) = ri$ for some $r > 1$, and

$$\rho(T(z), T(w)) = \int_1^r \frac{dy}{y} = \ln r.$$

On the other hand, $(ri, 0; i, \infty) = r$, and the theorem follows from the invariance of the cross–ration under Möbius transformations, a standard fact from complex analysis (which can be checked by a direct calculation). □

We will derive several explicit formulas for the hyperbolic distance involving the hyperbolic functions

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}, \quad \tanh z = \frac{\sinh x}{\cosh x}.$$

THEOREM 2.13. *For $z, w \in \mathcal{H}$, we have*

(a) $\rho(z, w) = \ln \frac{|z - \overline{w}| + |z - w|}{|z - \overline{w}| - |z - w|}$;

(b) $cosh \rho(z, w) = 1 + \frac{|z-w|^2}{2\mathrm{Im}(z)\mathrm{Im}(w)}$;

(c) $sinh[\frac{1}{2}\rho(z, w)] = \frac{|z-w|}{2(\mathrm{Im}(z)\mathrm{Im}(w))^{1/2}}$;

(d) $cosh[\frac{1}{2}\rho(z, w)] = \frac{|z-\overline{w}|}{2(\mathrm{Im}(z)\mathrm{Im}(w))^{1/2}}$;

(e) $tanh[\frac{1}{2}\rho(z, w)] = |\frac{z-w}{z-\overline{w}}|$.

PROOF. We will prove that (e) holds. By Theorem 2.8, the left–hand side is invariant under any transformation $T \in PSL(2, \mathbb{R})$. By Exercise 13, the right–hand side is also invariant under any $T \in PSL(2, \mathbb{R})$. Therefore if is sufficient to check the formula for the case when $z = i$, $w = ir$, $(r > 1)$. The right–hand side is equal to $(r - 1)/(r + 1)$. The left–hand side is equal to $\tanh[\frac{1}{2} \ln r]$. A simple calculation shows that these two expressions are equal. The other formulas are proved similarly. □

Exercises

11. Prove that if $z \neq w$, then $\rho(z, w) > 0$.

12. Let L be a semicircle or a straight line orthogonal to the real axis which meets the real axis at a point α. Prove that the transformation

$T(z) = -(z-\alpha)^{-1} + \beta \in PSL(2,\mathbb{R})$, for an appropriate value of β, maps L to the positive imaginary axis.

13. Prove that for $z, w \in \mathcal{H}$ and $T \in PSL(2,\mathbb{R})$, we have

$$|T(z) - T(w)| = |z-w||T'(z)T'(w)|^{1/2}.$$

2.3. Isometries

We have seen that transformations in $PSL(2,\mathbb{R})$ are isometries of the hyperbolic plane $\mathcal{H}$ (Theorem 2.8). The next theorem identifies all isometries of $\mathcal{H}$ in terms of Möbius transformations and symmetry in the imaginary axis.

THEOREM 2.14. *The group* $\mathrm{Isom}(\mathcal{H})$ *is generated by Möbius transformations in* $PSL(2,\mathbb{R})$ *together with the transformation* $z \mapsto -\overline{z}$. *The group* $PSL(2,\mathbb{R})$ *is a subgroup of* $\mathrm{Isom}(\mathcal{H})$ *of index two.*

PROOF. Let φ be any isometry of $\mathcal{H}$. By Theorem 2.11, φ maps geodesics into geodesics. Let I denote the positive imaginary axis. Then $\varphi(I)$ is a geodesic in $\mathcal{H}$, and according to Exercise 12, there exists an isometry $T \in PSL(2,\mathbb{R})$ that maps $\varphi(I)$ back to I. By applying the transformations $z \mapsto kz$ $(k>0)$ and $z \mapsto -1/z$, we may assume that $g \circ \varphi$ fixes i and maps the rays (i,∞) and $(i,0)$ onto themselves. Hence, being an isometry, $g \circ \varphi$ fixes each point of I. The same (synthetic) argument as in the Euclidean case shows that

$$g \circ \varphi(z) = z \text{ or } -\overline{z}. \tag{2.3.1}$$

Let z_1 and z_2 be two fixed points on I. For any point z not on I, draw two hyperbolic circles centered at z_1 and z_2 and passing through z. These circles intersect in two points, z and $z' = -\overline{z}$, since the picture is symmetric with respect to the imaginary axis (note that a hyperbolic circle is a Euclidean circle in $\mathcal{H}$, but with a different center). Since these circles are mapped into themselves under the isometry $g \circ \varphi$, we conclude that $g \circ \varphi(z) = z$ or $g \circ \varphi(z) = -\overline{z}$. Since isometries are continuous (see Excercise 14), only one of the equations (2.3.1) holds for all $z \in \mathcal{H}$. If $g \circ \varphi(z) = z$, then $\varphi(z)$ is a Möbius transformation of the form (2.1.3). If $g \circ \varphi(z) = -\overline{z}$, we have

$$\varphi(z) = \frac{a\overline{z}+b}{c\overline{z}+d} \text{ with } ad-bc=-1, \tag{2.3.2}$$

which proves the theorem. □

Thus we have identified all the isometries of $\mathcal{H}$. The sign of the determinant of the corresponding matrix in (2.1.3) or (2.3.2) determines the *orientation* of an isometry. We will refer to transformations in $PSL(2,\mathbb{R})$ as *orientation–preserving* isometries and to transformations of the form (2.3.2) as *orientation–reversing* isometries.

Now we will study and classify these two types of isometries of the hyperbolic plane $\mathcal{H}$.

Orientation–preserving isometries. The classification of matrices in $SL(2,\mathbb{R})$ in hyperbolic, elliptic, and parabolic depended on the absolute value of their trace, and hence makes sense in $PSL(2,\mathbb{R})$ as well. A matrix $A \in SL(2,\mathbb{R})$ with trace t is called *hyperbolic* if $|t| > 2$, *elliptic* if $|t| < 2$, and *parabolic* if $|t| = 2$. Let

$$T(z) = \frac{az+b}{cz+d} \in PSL(2,\mathbb{R}).$$

The fixed points of T are found by solving the equation

$$z = \frac{az+b}{cz+d}, \quad \text{i.e.,} \quad cz^2 + (d-a)z - b = 0.$$

We obtain

$$w_1 = \frac{a-d+\sqrt{(a+d)^2-4}}{2c}, \quad w_2 = \frac{a-d-\sqrt{(a+d)^2-4}}{2c}.$$

We see that if T is hyperbolic, then it has two fixed points in $\mathbb{R}\cup\{\infty\}$, if T is parabolic, it has one fixed point in $\mathbb{R}\cup\{\infty\}$, and if T is elliptic, it has two complex conjugate fixed points, hence one fixed point in $\mathcal{H}$. A Möbius transformation T fixes ∞ if and only if $c = 0$, and hence it is in the form $z \mapsto az + b$ $(a, b \in \mathbb{R},\ a > 0)$. If $a = 1$, it is parabolic; if $a \neq 0$, it is hyperbolic and its second fixed point is $b/(1-a)$. The fixed point w_i of T can be expressed in terms of the eigenvector $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ with eigenvalue λ_i: $w_i = x_i/y_i$. In terms of the eigenvalue λ_i the derivative at the fixed point w_i can be written as itself:

$$T'(w_i) = \frac{1}{(cw_i+d)^2} = \frac{1}{\lambda_i^2}.$$

Definition. A fixed point w of a transformation $f : \mathcal{H} \to \mathcal{H}$ is called *attracting* if $|f'(w)| < 1$, and it is called *repelling* if $|f'(w)| > 1$.

Now we are ready to summarize what we know from linear algebra about different kinds of transformations in $PSL(2,\mathbb{R})$ and describe the action of Möbius transformations in $\mathcal{H}$ geometrically.

1. Hyperbolic case. A hyperbolic transformation $T \in PSL(2,\mathbb{R})$ has two fixed points in $\mathbb{R}\cup\{\infty\}$, one attracting, denoted by u, the other repelling, denoted by w. A geodesic in $\mathcal{H}$ connecting them is called the *axis* of T and is denoted by $C(T)$. By Theorem 2.11 T maps $C(T)$ onto itself, and $C(T)$ is the only geodesic with this property. Let λ be the eigenvalue of T with $|\lambda| > 1$. Then the matrix of T is conjugate to the diagonal matrix $\begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$, which corresponds to the Möbius transformation

$$\Lambda(z) = \lambda^2 z, \tag{2.3.3}$$

i.e., there exists a transformation $S \in PSL(2, \mathbb{R})$ such that $STS^{-1} = \Lambda$. The conjugating transformation S maps the axis of T, oriented from u to w, to the positive imaginary axis I, oriented from 0 to ∞, which is the axis of Λ (cf Exercises 12 and 17).

In order to see how a hyperbolic transformation T acts on $\mathcal{H}$, it is useful to look at the all its iterates T^n, $n \in \mathbb{Z}$. If $z \in C(T)$, then $T^n(z) \in C(T)$ and $T^n(z) \to w$ as $n \to \infty$, while $T^n(z) \to u$ as $n \to -\infty$. The curve $C(T)$ is the only geodesic which is mapped onto itself by T, but there are other T–invariant curves, also "connecting" u and w. For the standard hyperbolic transformation (2.3.3), the Euclidean rays in the upper half–plane emanating from the origin are obviously T–invariant. If we define the distance from a point z to a given geodesic L as $\inf_{v \in L} \rho(z, v)$, we see that the distance is measured over a geodesic passing through z and orthogonal to L (Exercise 15). Such rays have an important property: they are equidistant from the axis $C(\Lambda) = I$ (see Exercise 16), and hence are called *equidistants*. Under S^{-1} they are mapped onto equidistants for the transformation T, which are Euclidean circles passing through the points u and w (see Figure 2.3.1).

A useful notion in understanding how hyperbolic transformations act is that of *isometric circle*. Since $T'(z) = (cz + d)^{-2}$, the Euclidean lengths are multiplied by $|T'(z)| = |cz + d|^{-2}$. They are unaltered in magnitude if and only if $|cz + d| = 1$. If $c \neq 0$, then the locus of such points z is the circle

$$\left| z + \frac{d}{c} \right| = \frac{1}{|c|}$$

with center at $-d/c$ and radius $1/|c|$. The circle

$$I(T) = \{z \in \mathcal{H} \mid |cz + d| = 1\}$$

is called the *isometric circle* of the transformation T. Since its center $-d/c$ lies in $\mathbb{R}$, we immediately see that isometric circles are geodesics in $\mathcal{H}$. Further, $T(I(T))$ is a circle of the same radius, $T(I(T)) = I(T^{-1})$, and the transformation maps the outside of $I(T)$ onto the inside of $I(T^{-1})$ and vice versa (see Figure 2.3.1 and Exercise 18).

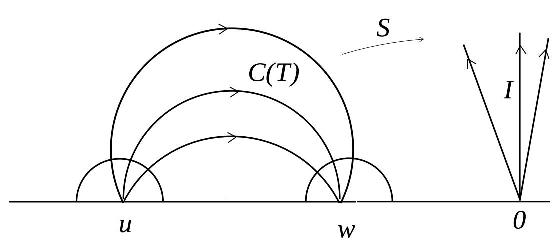

FIGURE 2.3.1. Hyperbolic transformations

If $c = 0$, then there is no circle with the isometric property: all Euclidean lengths are altered.

2. Parabolic case. A parabolic transformation $T \in PSL(2, \mathbb{R})$ has one fixed point in $p \in \mathbb{R} \cup \{\infty\}$. The transformation T has one eigenvalue $\lambda = \pm 1$ and is conjugate to the transformation $P(z) = z + b$ for some $b \in \mathbb{R}$, i.e., there exists a transformation $S \in PSL(2, \mathbb{R})$ such that $P = STS^{-1}$. The transformation P is an Euclidean translation, and hence it leaves all horizontal lines invariant. Horizontal lines are called *horocycles* for the transformation P. Under the map S^{-1} they are sent to invariant curves (horocycles) for the transformation T. Horocycles for T are Euclidean circles tangent to the real line at the parabolic fixed point p (see Figure 2.3.2 and Exercise 20).

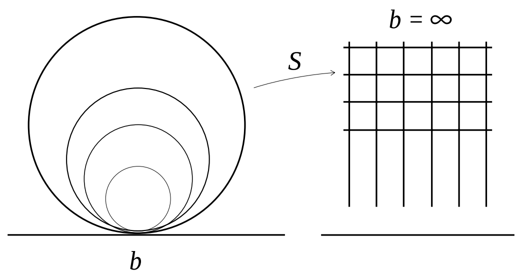

FIGURE 2.3.2. Parabolic transformations

If $c \neq 0$, then the isometric circles for T and T^{-1} are tangent to each other (see Exercise 19). If $c = 0$, then there is no unique circle with the isometric property: since in this case T is an Euclidean translation, all Euclidean lengths are unaltered.

3. Elliptic case. An elliptic transformation $T \in PSL(2, \mathbb{R})$ has a unique fixed point $e \in \mathcal{H}$. It has the eigenvalues $\lambda = \cos\varphi + i \sin\varphi$ and $\overline{\lambda} = \cos\varphi - i \sin\varphi$, and it is easier to describe its simplest form in the *unit disc model* of hyperbolic geometry: $\mathcal{U} = \{z \in \mathbb{C} \mid |z| < 1\}$. The map

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

is a homeomorphism of $\mathcal{H}$ onto $\mathcal{U}$. The distance in $\mathcal{U}$ is induced by means of the hyperbolic distance in $\mathcal{H}$:

$$\rho(z, w) = \rho(f^{-1}z, f^{-1}w) \ (z, w \in \mathcal{U}).$$

The readily verified formula

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

implies that this distance in $\mathcal{U}$ is derived from the metric

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

Isometries of $\mathcal{U}$ are the conjugates of isometries of $\mathcal{H}$, i.e., we can write

$$S = f \circ T \circ f^{-1} \ (T \in PSL(2, \mathbb{R})).$$

Exercise 21 shows that orientation–preserving isometries of $\mathcal{U}$ are of the form

$$z \mapsto \frac{az + \overline{c}}{cz + \overline{a}} \ (a, c \in \mathbb{C}, a\overline{a} - c\overline{c} = 1),$$

and the transformation corresponding to the standard reflection $R(z) = -\overline{z}$ is also the reflection of $\mathcal{U}$ in the imaginary axis.

Let us return to our elliptic transformation $T \in PSL(2, \mathbb{R})$ that fixes $e \in \mathcal{H}$. Conjugating T by f, we obtain an elliptic transformation of the unit disc $\mathcal{U}$. Using an additional conjugation by an orientation–preserving isometry of $\mathcal{U}$ if necessary (see Exercise 22), we bring the fixed point to 0, and hence bring T to the form $z \mapsto e^{2i\varphi} z$. In other words, an elliptic transformation with eigenvalues $e^{i\varphi}$ and $e^{-i\varphi}$ is conjugate to a rotation by 2φ.

Example 1 . Let $z \mapsto -1/z$ be the elliptic transformation given by the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Its fixed point in $\mathcal{H}$ is i. It is a transformation of order 2 since the identity in $PSL(2, \mathbb{R})$ is $\{-1_2, 1_2\}$, and hence is a half–turn. In the unit disc model its matrix is conjugate to the matrix $\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$.

Orientation–reversing isometries. The simplest orientation–reversing isometry of $\mathcal{H}$ is the transformation $R(z) = -\overline{z}$, which is the reflection in the imaginary axis I, and hence it fixes I pointwise. It is also a hyperbolic reflection in I, i.e., if for each point z we draw a geodesic through z, orthogonally to I and intersecting I at a point z_0, then $R(z) = z'$ is on the same geodesic and $\rho(z', z_0) = \rho(z, z_0)$. Let L be any geodesic in $\mathcal{H}$ and $T \in PSL(2, \mathbb{R})$ be any Möbius transformation. Then the transformation

$$TRT^{-1} \tag{2.3.4}$$

fixes the geodesic $L = T(I)$ pointwise and therefore may be regarded as a "reflection in the geodesic L". In fact, it is the well–known geometrical transformation called *inversion in a circle*.

Definition. Let Q be a circle in $\mathbb{R}^2$ with center K and radius r. Given any point $P \neq K$ in $\mathbb{R}^2$, a point P_1 is called *inverse* to P if

(a) P_1 lies on the ray from K to P,
(b) $|KP_1| \cdot |KP| = r^2$.

The relationship is reciprocal: if P_1 is inverse to P, then P is inverse to P_1. We say that P and P_1 are *inverse with respect to Q*. Obviously, inversion fixes all points in the circle Q. Inversion may be described by a geometric construction (see Exercise 23). We will derive a formula for it. Let P, P_1 and K be the points z, z_1, and k in $\mathbb{C}$. Then Definition 2.3 can be rewritten as

$$|(z_1 - k)(z - k)| = r^2, \quad \arg(z_1 - k) = \arg(z - k).$$

Since $\arg(z - k) = -\arg(\overline{z} - \overline{k})$, both equations are satisfied if and only if

$$(z_1 - k)(\overline{z} - \overline{k}) = r^2. \tag{2.3.5}$$

This gives us the following formula for the inversion in a circle:

$$z_1 = \frac{k\overline{z} + r^2 - |k|^2}{\overline{z} - \overline{k}}. \tag{2.3.6}$$

Now we are able to prove a theorem for isometries of the hyperbolic plane similar to a result in Euclidean geometry.

THEOREM 2.15. *Every isometry of $\mathcal{H}$ is a product of not more than three reflections in geodesics in $\mathcal{H}$.*

PROOF. By Theorem 2.14 it suffices to show that each transformation in $PSL(2,\mathbb{R})$ is a product of two reflections. Let

$$T(z) = \frac{az + b}{cz + d}.$$

First consider the case for which $c \neq 0$. Then both T and T^{-1} have well–defined isometric circles (see Exercise 19). They have the same radius $1/|c|$ and their centers are on the real axis at $-d/c$ and a/c, respectively. We will show that $T = R \circ R_{I(T)}$, where $R_{I(T)}$ is the reflection in the isometric circle $I(T)$, or inversion, and R is the reflection in the vertical geodesic passing through the midpoint of the interval $[-d/c, a/c]$. To do this, we use formula (2.3.5) for inversion:

$$R_{I(T)}(z) = \frac{-\frac{d}{c}\overline{z} + \frac{1}{c^2} - \frac{d^2}{c^2}}{\overline{z} + \frac{d}{c}} = \frac{-d(\overline{z} + \frac{d}{c}) + \frac{1}{c}}{c\overline{z} + d}.$$

The reflection in the line $x = (a - d)/2c$ is given by the formula

$$R(z) = -\overline{z} + 2\frac{a - d}{2c}.$$

Combining the two, we obtain

$$R \circ R_{I(T)} = \frac{az + b}{cz + d}.$$

Now if $c = 0$, the transformation T may be either parabolic $z \mapsto z + b$ or hyperbolic $z \mapsto \lambda^2 z + b$, each fixing ∞. In the first case the theorem follows from the Euclidean result for translations. For $T(z) = \lambda^2 z + b$, it is easy to see that the reflections should be in circles of radii 1 and λ centered at the second fixed point. $\square$

Exercises

14. Prove that isometies are continuous maps.

15. (a) Prove that there is a unique geodesic through a point z orthogonal to a given geodesic L.

(b)* Give a geometrical construction of this geodesic.

(c) Prove that for $z \notin L$, $\inf_{v \in L} \rho(z, v)$ is achieved on the geodesic described in (a).

16. Prove that the rays in $\mathcal{H}$ emanating from the origin are equidistant form the positive imaginary axis I.

17. Let $A \in PSL(2, \mathbb{R})$ be a hyperbolic transformation, and $B = SAS^{-1}$ ($B \in PSL(2, \mathbb{R})$) be its conjugate. Prove that B is also hyperbolic and find the relation between their axes $C(A)$ and $C(B)$.

18. Prove that isometric circles $I(T)$ and $I(T^{-1})$ have the same radius, and that the image of $I(T)$ under the transformation T is $I(T^{-1})$.

19. Prove that

(a) T is hyperbolic if and only if $I(T)$ and $I(T^{-1})$ do not intersect;
(b) T is elliptic if and only if $I(T)$ and $I(T^{-1})$ intersect;
(c) T is parabolic if and only if $I(T)$ and $I(T^{-1})$ are tangential.

20. Prove that the horocycles for a parabolic transformation with a fixed point $p \in \mathbb{R}$ are Euclidean circles tangent to the real line at p.

21. Show that orientation–preserving isometries of $\mathcal{U}$ are of the form

$$z \mapsto \frac{az + \overline{c}}{cz + \overline{a}} \qquad (a, c \in \mathbb{C}, a\overline{a} - c\overline{c} = 1).$$

22. Prove that for any two distinct points $z_1, z_2 \in \mathcal{H}$ there exists a transformation $T \in PSL(2, \mathbb{R})$ such that $T(z_1) = z_2$.

23. Give a geometric construction of inversion in a given circle Q in the Euclidean plane $\mathbb{R}^2$.

24. Prove that the transformation (2.3.4) is an inversion in the circle corresponding to the geodesic L.

25. Prove that two hyperbolic transformations in $PSL(2, \mathbb{R})$ commute if and only if their axes coincide.

26. Let $A \in PSL(2, \mathbb{R})$ be hyperbolic and $B \in PSL(2, \mathbb{R})$ be an elliptic transformation different from the identity. Prove that $AB \neq BA$.

27. Using the hyperbolic trigonometric functions, find formulas for the hyperbolic distance in the disk model, similar to those for the half-plane model.

2.4. Hyperbolic area and the Gauss–Bonnet formula

Let T be a Möbius transformation. The *differential* of T, denoted by DT, at a point z is the linear map that takes the tangent space $T_z\mathcal{H}$ onto $T_{T(z)}\mathcal{H}$ and is defined by the 2×2 matrix

$$DT = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}.$$

THEOREM 2.16. *Let $T \in PSL(2,\mathbb{R})$. Then DT preserves the norm in the tangent space at each point.*

PROOF. For $\zeta \in T_z\mathcal{H}$, we have $DT(\zeta) = T'(z)\zeta$ by Exercise 29. Since

$$|T'(z)| = \frac{\operatorname{Im}(T(z))}{\operatorname{Im}(z)} = \frac{1}{|cz+d|^2},$$

we can write

$$\|DT(\zeta)\| = \frac{|DT(\zeta)|}{\operatorname{Im}(T(z))} = \frac{|T'(z)||\zeta|}{\operatorname{Im}(T(z))} = \frac{|\zeta|}{\operatorname{Im}(z)} = \|\zeta\|.$$

□

COROLLARY 2.17. *Any transformation in $PSL(2,\mathbb{R})$ is conformal, i.e., it preserves angles.*

PROOF. It is easy to prove the *polarization identity*, which asserts that for any $\zeta_1, \zeta_2 \in T_z\mathcal{H}$ we have

$$\langle \zeta_1, \zeta_2 \rangle = \frac{1}{2}(\|\zeta_1\|^2 + \|\zeta_2\|^2 - \|\zeta_1 - \zeta_2\|^2);$$

it implies that the inner product and hence the absolute value of the angle between tangent vectors is also preserved. Since Möbius transformations preserve orientation, the corollary follows. □

Let $A \subset \mathcal{H}$. We define the *hyperbolic area* of A by the formula

$$\mu(A) = \int_A \frac{dxdy}{y^2}$$

provided this integral exists.

THEOREM 2.18. *Hyperbolic area is invariant under all Möbius transformations $T \in PSL(2,\mathbb{R})$, i.e., if $\mu(A)$ exists, then $\mu(A) = \mu(T(A))$.*

PROOF. When we performed the change of variables $w = T(z)$ in the line integral of Theorem 2.8, the coefficient $|T'(z)|$ appeared (it is the coefficient responsible for the change of Euclidean lengths). If we carry out the same change of variables in the plane integral, the Jacobian of this map will appear, since it is responsible for the change of the Euclidean areas. Let $z = x + iy$, and $w = T(z) = u + iv$.

The Jacobian is the determinant of the differential map DT and is customarily denoted by $\partial(u,v)/\partial(x,y)$. Thus

$$\frac{\partial(u,v)}{\partial(x,y)} := \det\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 = |T'(z)|^2 = \frac{1}{|cz+d|^4}. \quad (2.4.1)$$

We use this expression to compute the integral

$$\begin{aligned} \mu(T(A)) &= \int_{T(A)} \frac{dudv}{v^2} = \int_A \frac{\partial(u,v)}{\partial(x,y)} \frac{dxdy}{v^2} \\ &= \int_A \frac{1}{|cz+d|^4} \frac{|cz+d|^4}{y^2} dxdy = \mu(A), \end{aligned}$$

as claimed. □

A *hyperbolic triangle* is a figure bounded by three segments of geodesics. The intersection points of these geodesics are called the *vertices* of the triangle. We allow vertices to belong to $\mathbb{R} \cup \{\infty\}$. There are 4 types of hyperbolic triangles, depending on whether $0, 1, 2$, or 3 vertices belong to $\mathbb{R} \cup \{\infty\}$ (see Figure 4.4).

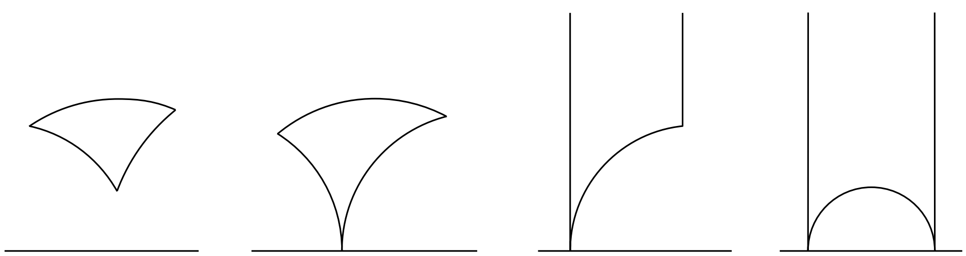

FIGURE 2.4.1. Hyperbolic triangles

The Gauss–Bonnet formula shows that the hyperbolic area of a hyperbolic triangle depends only on its angles.

THEOREM 2.19 (Gauss-Bonnet). *Let Δ be a hyperbolic triangle with angles α, β, and γ. Then $\mu(\Delta) = \pi - \alpha - \beta - \gamma$.*

PROOF. First we consider the case in which one of the vertices of the triangle belongs to $\mathbb{R} \cup \{\infty\}$. Since transformations from $PSL(2, \mathbb{R})$ do not alter the area and the angles of a triangle, we may apply the transformation from $PSL(2, \mathbb{R})$ which maps this vertex to ∞ and the base to a segment of the unit circle (as in Figure 2.4.2), and prove the formula in this case.

The angle at infinity is equal to 0, and let us assume that the other two angles are equal to α and β. Then the angles A0C and B0D are equal to α and β, respectively, as angles with mutually perpendicular sides (this theorem from Euclidean geometry does not use the Fifth Postulate and is therefore true in hyperbolic geometry as well). Assume the vertical geodesics are the lines $x = a$ and $x = b$. Then

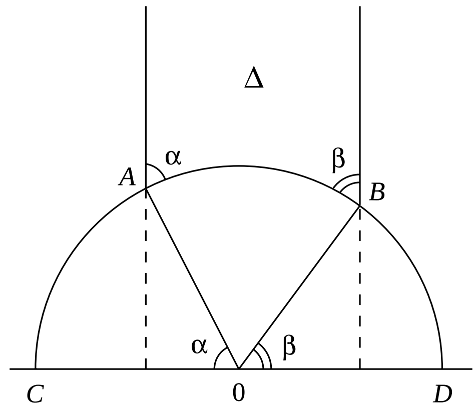

FIGURE 2.4.2

$$\mu(\Delta) = \int_\Delta \frac{dxdy}{y^2} = \int_a^b dx \int_{\sqrt{1-x^2}}^\infty \frac{dy}{y^2} = \int_a^b \frac{dx}{\sqrt{1-x^2}}.$$

The substitution $x = \cos\theta$ $(0 \leq \theta \leq \pi)$ gives

$$\mu(\Delta) = \int_{\pi-\alpha}^{\beta} \frac{-\sin\theta d\theta}{\sin\theta} = \pi - \alpha - \beta.$$

For the case in which Δ has no vertices at infinity, we continue the geodesic connecting the vertices A and B, and suppose that it intersects the real axis at the point D (if one side of Δ is a vertical geodesic, then we label its vertices A and B), and draw a geodesic from C to D. Then we obtain the situation depicted in Figure 2.4.3.

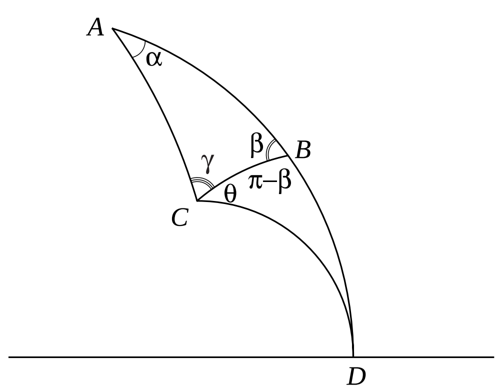

FIGURE 2.4.3

We denote the triangle ADC by Δ_1 and the triangle CBD by Δ_2. Our formula has already been proved for triangles such as Δ_1 and Δ_2 since the vertex D is at infinity. Now we can write

$$\begin{aligned}\mu(\Delta) = \mu(\Delta_1) - \mu(\Delta_2) &= (\pi - \alpha - \gamma - \theta) - (\pi - \theta - \pi + \beta) \\ &= \pi - \alpha - \beta - \gamma,\end{aligned}$$

as claimed. □

Theorem 2.19 asserts that the area of a triangle depends only on its angles, and is equal to the quantity $\pi - \alpha - \beta - \gamma$, which is called the *angular defect*. Since the area of a nondegenerate triangle is positive, the angular defect is positive, and therefore, in hyperbolic geometry the sum of angles of any triangle is less than π. We will also see that there are no similar triangles in hyperbolic geometry (except isometric ones).

THEOREM 2.20. *If two triangles have the same angles, then there is an isometry mapping one triangle into the other.*

PROOF. If necessary, we perform the reflection $z \mapsto -\overline{z}$, so that the respective angles of the triangles ABC and $A'B'C'$ (in the clockwise direction) are equal. Then we apply a hyperbolic transformation mapping A to A' (Exercise 22), and an elliptic transformation mapping the side AB onto the side $A'B'$. Since the angles CAB and $C'A'B'$ are equal, the side AC will be mapped onto the side $A'C'$. We must prove that B is then mapped to B' and C to C'. Assume B' is mapped inside the geodesic segment AB. If we had $C' \in [A, C]$, the areas of triangles ABC and $A'B'C'$ would not be equal, which contradicts Theorem 2.19. Therefore C must belong to the side $A'C'$, and hence the sides BC and $B'C'$ intersect at a point X (see Fig. 2.4.4); thus we obtain the triangle $B'XB$. Its angles are β and $\pi - \beta$ since the angles at the vertices B and B' of our original triangles are equal (to β). We see that, in contradiction with Theorem 2.19, the sum of the angles of the triangle $B'XB$ is at least π. □

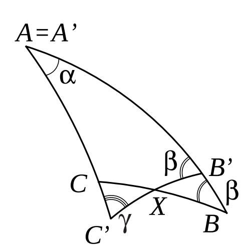

FIGURE 2.4.4

Exercises

28. Justify the calculations in (2.4.1) by checking that for the Möbius transformation

$$w = T(z) = \frac{az+b}{cz+d} \quad \text{with} \quad z = x+iy,\ w = u+iv$$

we have

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

(these are the classsical *Cauchy–Riemann equations*) and

$$T'(z) = \frac{dw}{dz} = \frac{1}{2}(\frac{\partial w}{\partial x} - i\frac{\partial w}{\partial y}) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x};$$

(*Hint*: express x and y in terms of z and $\overline{z}$ and use the Cauchy–Riemann equations.)

29. If we identify the tangent space $T_z\mathcal{H} \approx \mathbb{R}^2$ with the complex plane $\mathbb{C}$ by means of the map

$$\binom{\xi}{\eta} \mapsto \xi + i\eta = \zeta,$$

then $DT(\zeta) = T'(z)\zeta$, where in the left–hand side we have a linear transformation of $T_z\mathcal{H} \approx \mathbb{R}^2$, and in the right–hand side, the multiplication of two complex numbers.

Lecture 3
The modular group and its fundamental regions

3.1. Definition of fundamental regions and Fuchsian groups

Let Γ be a subgroup of the group of isometries $\mathrm{Isom}(\mathcal{H})$ of the hyperbolic plane $\mathcal{H}$. We call a subset $F \subset \mathcal{H}$ a *fundamental region* for Γ if it satifies the following conditions:

(a) F is a closed region in $\mathcal{H}$ bounded by a finite number of geodesics;
(b) the images $T(F)$ $(T \in \Gamma)$ cover the entire hyperbolic plane $\mathcal{H}$;
(c) for $T_1 \neq T_2$, the images $T_1(F)$ and $T_2(F)$ have no interior points in common.

We will denote the interior of the fundamental region F by $\mathring{F}$. Notice that not every subgroup of $\mathrm{Isom}(\mathcal{H})$ has fundamental regions. For instance, the entire group $\mathrm{Isom}(\mathcal{H})$ does not have any.

First let us give a simple example.

Example 2 Let Γ be the cyclic group generated by the transformation $z \to 2z$. Then the semi-annulus between the circles of radii 1 and 2 shaded in Figure 3.1.1(a) is easily seen to be a fundamental region for Γ. It is also clear from this example that a fundamental region is not uniquely determined by the group: for example, any semi-annulus between circles of radii r and $2r$ gives a fundamental region for this group (an example for $r = 3/2$ is shown in Figure 3.1.1(b)), but there are many others.

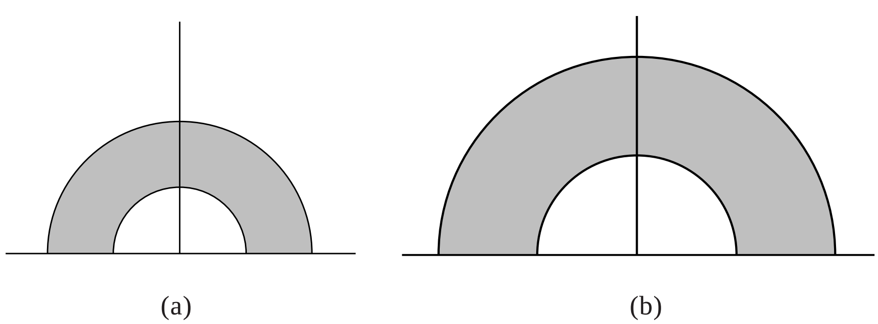

FIGURE 3.1.1. Fundamental regions for the group generated by $z \to 2z$

Besides being a group, $PSL(2,\mathbb{R})$ is also a topological space. More precisely, $SL(2,\mathbb{R})$ can be identified with the following subset of $\mathbb{R}^4$:

$$X = \{(a,b,c,d) \in \mathbb{R}^4 \mid ad - bc = 1\}.$$

The norm on $SL(2,\mathbb{R})$ is induced from $\mathbb{R}^4$: for $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $ad-bc = 1$, we define

$$\|A\| = \sqrt{a^2 + b^2 + c^2 + d^2}, \tag{3.1.1}$$

and $SL(2,\mathbb{R})$ is given the topology induced by the metric

$$d(A,B) = \|A - B\|. \tag{3.1.2}$$

Since $A \sim -A$ is an equivalence relation on $SL(2,\mathbb{R})$, the quotient space $SL(2,\mathbb{R})/_\sim = PSL(2,\mathbb{R})$, is topologized with the quotient topology. Exercise 30 shows that in fact $PSL(2,\mathbb{R})$ is a topological group. Since (2.3.2) implies that orientation–reversing isometries are given by matrices in $GL(2,\mathbb{R})$ with determinant -1, it follows that the whole group of isometries $\mathrm{Isom}(\mathcal{H})$ can be topologized using the same distance. Notice that since $\|A\| = \|-A\|$, the norm (3.1.1) is a well–defined function on $PSL(2,\mathbb{R})$, while the metric (3.1.2) is not. One natural way to introduce a metric on $PSL(2,\mathbb{R})$ is to represent it as a matrix group $S0_o(2,1)$, the other is via the so–called *chord metric* on the unit disc, obtained from the Euclidean metric on the unit sphere by means of the stereographic projection, but both are beyond the scope of these notes.

Convergence in $PSL(2,\mathbb{R})$ can be expressed in matrix language as follows. If $g_n \to g$ is in $PSL(2,\mathbb{R})$, then there exist matrices $A_n \in SL(2,\mathbb{R})$ representing g_n such that $\lim_{n\to\infty} \|A_n - A\| = 0$.

Definition. A subgroup Γ of $\mathrm{Isom}(\mathcal{H})$ is called *discrete* if $T_n \to \mathrm{Id}$ ($T_n \in \Gamma$) implies $T_n = \mathrm{Id}$ for sufficiently large n.

Remark. It is clear that if $\Gamma \subset \mathrm{Isom}(\mathcal{H})$ is discrete, then so are all subgroups of $\mathrm{Isom}(h)$ conjugate to it, i.e., for every $T \in \mathrm{Isom}(\mathcal{H})$, the subgroup $T^{-1}\Gamma T$ is discrete.

Definition. Any discrete subgroup of $PSL(2,\mathbb{R})$ is said to be a *Fuchsian group.*

Example 3 For an elliptic $\gamma_0 \in PSL(2,\mathbb{R})$, the group $\Gamma = \langle \gamma_0 \rangle$ is a Fuchsian group if and only if it is finite. In other words, Γ is Fuchsian if the eigenvalue λ of γ_0 is equal to $e^{\pi n/m}$ for some integers n and m, or if γ_0 is conjugate to a rotation by a rational multiple of 2π.

Example 4 If $\gamma_0 \in PSL(2,\mathbb{R})$ is parabolic, then $\Gamma = \langle \gamma_0 \rangle$ is a Fuchsian group. It is sufficient to look at the case in which $\gamma_0(z) = z + 1$. Assume that there is a sequence $\gamma_n \in \Gamma$ such that $\gamma_n \to \mathrm{Id}$ in $PSL(2,\mathbb{R})$. Then there are matrices $A_n \in SL(2,\mathbb{R})$ representing g_n for which $A_n \to 1_2$. Therefore there exists an $N > 0$ such that for all $n > N$ we have $A_n = \begin{pmatrix} 1 & a_n \\ 0 & 1 \end{pmatrix}$. But then $\| A_n - 1_2 \| \to 0$ as $n \to \infty$, so that $a_n \to 0$, and since a_n is an integer, this implies that for n large enough we have $a_n = 0$, i.e., $A_n = 1_2$.

Example 5 If $\gamma_0 \in PSL(2,\mathbb{R})$ is hyperbolic, then $\Gamma = \langle \gamma_0 \rangle$ is a Fuchsian group. Again, it is sufficient to look at the case in which γ_0 is in the standard form, i.e., $\gamma_0(z) = \lambda^2 z$, where $\lambda \neq 1$. If $\gamma_n \in \Gamma$ satisfies $\gamma_n \to \mathrm{Id}$ in $PSL(2,\mathbb{R})$, we choose matrices

$$A_n = \begin{pmatrix} \lambda^{a_n} & 0 \\ 0 & \lambda^{-a_n} \end{pmatrix} \in SL(2,\mathbb{R})$$

representing g_n so that $A_n \to 1_2$. But then $\| A_n - 1_2 \| \to 0$ as $n \to \infty$, hence $\lambda^{a_n} \to 1$, $\lambda^{-a_n} \to 1$. Since $\lambda > 1$ and a_n is an integer, this implies that $a_n = 0$ for n large enough, i.e., $A_n = 1_2$.

Discrete subgroups of $\mathrm{Isom}(\mathcal{H})$ have the following important property, which we state without proof.

THEOREM 3.21. *A subgroup $\Gamma \subset \mathrm{Isom}(\mathcal{H})$ is discrete if and only if it acts on $\mathcal{H}$* properly discontinuously*, i.e., so that the orbit of each point z, $\Gamma z = \{x \in \mathcal{H} \mid x = \gamma(z)$ for some $\gamma \in \Gamma\}$, has no accumulation point in $\mathcal{H}$.*

Example 6 Let us consider the group consisting of all transformations

$$z \mapsto \frac{az+b}{cz+d} \ (a,b,c,d \in \mathbb{Z},\ ad-bc=1).$$

It is called the *modular group* and is denoted by $PSL(2,\mathbb{Z})$. Let

$$\begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix} \to 1_2 \, .$$

Then $a_n \to 1$, $b_n \to 0$, $c_n \to 0$, and $d_n \to 1$. Since a_n, b_n, c_n, d_n are integers, this implies that there exists an $N > 0$ such that for all $n > N$ we have $a_n = 1$, $b_n = 0$, $c_n = 0$, and $d_n = 1$, i.e.,

$$\begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix} = 1_2 \, .$$

Therefore $PSL(2,\mathbb{Z})$ is a discrete subgroup of $PSL(2,\mathbb{R})$, i.e., a Fuchsian group.

THEOREM 3.22. *Suppose F_1 and F_2 are two fundamental regions for a Fuchsian group Γ, and the area of F_1, $\mu(F_1)$ is finite. Then $\mu(F_2) = \mu(F_1)$.*

PROOF. Since the boundary of each fundamental region consists of finitely many geodesics, we have $\mu(\overset{\circ}{F}_i) = \mu(F_i), i = 1, 2$. Now

$$F_1 \supseteq F_1 \cap \Big(\bigcup_{T \in \Gamma} T(\overset{\circ}{F}_2) \Big) = \bigcup_{T \in \Gamma} (F_1 \cap T(\overset{\circ}{F}_2)).$$

Since $\overset{\circ}{F}_2$ is the interior of a fundamental region, the sets $F_1 \cap T(\overset{\circ}{F}_2)$ are disjoint, and since μ is $PSL(2,\mathbb{R})$-invariant, we have

$$\mu(F_1) \geq \sum_{T\in\Gamma} \mu(F_1 \cap T(\overset{\circ}{F}_2)) = \sum_{T\in\Gamma} \mu(T^{-1}(F_1) \cap \overset{\circ}{F}_2) = \sum_{T\in\Gamma} \mu(T(F_1) \cap \overset{\circ}{F}_2).$$

Since F_1 is a fundamental region, we have

$$\bigcup_{T\in\Gamma} T(F_1) = \mathcal{H}, \text{ and } \bigcup_{T\in\Gamma} (T(F_1) \cap \overset{\circ}{F}_2) = \overset{\circ}{F}_2.$$

Hence

$$\sum_{T\in\Gamma} \mu(T(F_1) \cap \overset{\circ}{F}_2) \geq \mu(\bigcup_{T\in\Gamma} T(F_1) \cap \overset{\circ}{F}_2) = \mu(\overset{\circ}{F}_2) = \mu(F_2)$$

Interchanging F_1 and F_2, we obtain $\mu(F_2) \geq \mu(F_1)$; so hence $\mu(F_2) = \mu(F_1)$. $\square$

Thus we have proved a very important fact: the area of a fundamental region, if it is finite, is a numerical invariant of the group Γ. Examples 3, 4, and 5 give Fuchsian groups with fundamental regions of infinite area. Obviously, any compact fundamental region has a finite area, although groups with such fundamental regions are not easy to construct. Noncompact regions may also have finite area. A very important example of this kind, involving $\Gamma = PSL(2,\mathbb{Z})$, will be discussed in detail in §3.2.

THEOREM 3.23. *Let Γ be a discrete group of isometries of the upper half-plane $\mathcal{H}$, and Λ be a subgroup of Γ of index n. If*

$$\Gamma = \Lambda T_1 \cup \Lambda T_2 \cup \cdots \cup \Lambda T_n$$

is a decomposition of Γ into Λ-cosets and if F is a fundamental region for Γ, then

(i) *$F_1 = T_1(F) \cup T_2(F) \cup \cdots \cup T_n(F)$ is a fundamental region for Λ,*
(ii) *if $\mu(F)$ is finite and the hyperbolic area of the boundary of F is zero, then $\mu(F_1) = n\mu(F)$.*

PROOF OF (i). Let $z \in \mathcal{H}$. Since F is a fundamental region for Γ, there exist $w \in F$ and $T \in \Gamma$ such that $z = T(w)$. We have $T = ST_i$ for some $S \in \Lambda$ and some i with $1 \leq i \leq n$. Therefore

$$z = ST_i(w) = S(T_i(w)).$$

Since $T_i(w) \in F_1$, the point z is in the Λ-orbit of some point of F_1. Hence the union of the images of F_1 under the action of elements of Λ is $\mathcal{H}$.

Now suppose that $z \in \overset{\circ}{F}_1$ and that $S(z) \in \overset{\circ}{F}_1$, for $S \in \Lambda$. We must prove that $S = \mathrm{Id}$. Let $\epsilon > 0$ be so small that $B_\epsilon(z)$ (the open hyperbolic disc of radius ϵ centered at z) is contained in $\overset{\circ}{F}_1$. Then $B_\epsilon(z)$ has a nonempty intersection with exactly k of the images of $\overset{\circ}{F}$ under the maps $T_1, \cdots, T_n$, where $1 \leq k \leq n$. Suppose these images are $T_{i_1}(\overset{\circ}{F}), \cdots, T_{i_k}(\overset{\circ}{F})$. Suppose the set $B_\epsilon(S(z)) = S(B_\epsilon(z))$ has a nonempty intersection with $T_j(\overset{\circ}{F})$ say, $1 \leq j \leq n$. It follows that $B_\epsilon(z)$ has a nonempty intersection with $S^{-1}T_j(\overset{\circ}{F})$, so that $S^{-1}T_j = T_{i_1}$, where $1 \leq 1 \leq k$. Hence

$$\Lambda T_j = \Lambda S^{-1} T_j = \Lambda T_{i_1},$$

which implies that $T_j = T_{i_1}$ and $S = \mathrm{Id}$. Thus $\overset{\circ}{F}_1$ contains precisely one point of each Λ-orbit. □

Proof of (ii). This follows immediately since $\mu(T(F)) = \mu(F)$ for all $T \in PSL(2, \mathbb{R})$, and $\mu(T_i(F) \cap T_j(F)) = 0$ for $i \neq j$. □

Exercises

30. Prove that group multiplication and taking inverses are continuous with respect to the topology on $PSL(2, \mathbb{R})$.

3.2. Construction of fundamental regions

Suppose Γ is an arbitrary Fuchsian group and the point $p \in \mathcal{H}$ is not fixed by any element of $\Gamma - \{\mathrm{Id}\}$. We define the *Dirichlet region for* Γ *centered at* p to be the set

$$D_p(\Gamma) = \{z \in \mathcal{H} \mid \rho(z, p) \leq \rho(z, T(p)) \text{ for all } T \in \Gamma\}. \tag{3.2.1}$$

For each fixed $T_1 \in PSL(2, \mathbb{R})$,

$$\{z \in \mathcal{H}\} \mid \rho(z, p) \leq \rho(z, T_1(p))\} \tag{3.2.2}$$

is the set of points z which are closer in the hyperbolic metric to p than to $T_1(p)$. Clearly, $p \in D_p(\Gamma)$, and since the Γ-orbit of p is discrete, $D_p(\Gamma)$ contains a neighborhood of p. In order to describe the set (3.2.2), we connect the points p and $T_1(p)$ by a geodesic segment and construct the line given by the equation $\rho(z, p) = \rho(z, T_1(p))$.

Definition. The *perpendicular bisector* of the geodesic segment $[z_1, z_2]$ is the unique geodesic through w, the mid-point of $[z_1, z_2]$, orthogonal to $[z_1, z_2]$ (Figure 3.2.1).

Lemma 3.24. *The line given by the equation*

$$\rho(z, z_1) = \rho(z, z_2) \tag{3.2.3}$$

is the perpendicular bisector of the geodesic segment $[z_1, z_2]$.

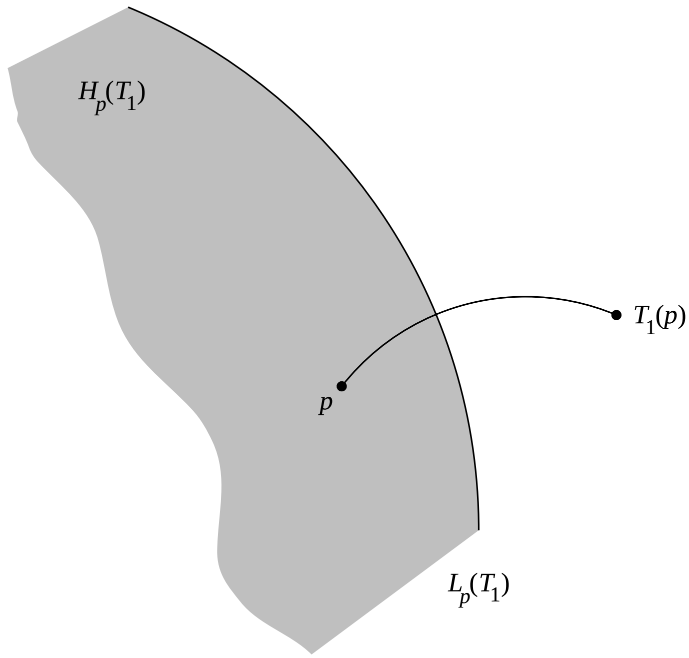

FIGURE 3.2.1

PROOF. We may assume that $z_1 = i, z_2 = ir^2$ with $r > 0$; then $w = ir$ and the perpendicular bisector is given by the equation $|z| = r$. On the other hand, by Theorem 2.13(b), relation (3.2.3) is equivalent to

$$\frac{|z - z_1|^2}{y} = \frac{|z - z_2|^2}{r^2 y}$$

which simplifies to $|z| = r$. □

We denote the perpendicular bisector of the geodesic segment $[p, T_1(p)]$ by $L_p(T_1)$, and the hyperbolic half-plane containing p described in (3.2.2) by $H_p(T_1)$ (see Figure 3.2.1). Thus $D_p(\Gamma)$ is the intersection of hyperbolic half-planes,

$$D_p(\Gamma) = \bigcap_{T \in \Gamma,\, T \neq \mathrm{Id}} H_p(T),$$

and thus is a *hyperbolically convex region.*

THEOREM 3.25. *If p is not fixed by any element of $\Gamma - \{\mathrm{Id}\}$, then $D_p(\Gamma)$ is a connected fundamental region for Γ.*

PROOF. Let $z \in \mathcal{H}$, and Γz be its Γ-orbit. Since Γz is a discrete set, there exists a $z_0 \in \Gamma z$ with the smallest value of $\rho(z_0, p)$. Then we have $\rho(z_0, p) \leq \rho(T(z_0), p)$ for all $T \in \Gamma$. By the invariance of the hyperbolic metric under $PSL(2, \mathbb{R})$, the region in (3.2.1) can also be defined as

$$D_p(\Gamma) = \{z \in \mathcal{H} \mid \rho(z,p) \leq \rho(T(z),p) \text{ for all } T \in \Gamma\}. \tag{3.2.4}$$

Therefore $z_0 \in D_p(\Gamma)$. Thus $D_p(\Gamma)$ contains at least one point from every Γ-orbit.

Next we show that if z_1, z_2 are in the interior of $D_p(\Gamma)$, then they cannot lie in the same Γ-orbit. If $\rho(z,p) = \rho(T(z),p)$ for some $T \in \Gamma - \{\mathrm{Id}\}$, then $\rho(z,p) = \rho(z, T^{-1}(p))$ and hence $z \in L_p(T^{-1})$. Then either $z \notin D_p(\Gamma)$ or z lies on the boundary of $D_p(\Gamma)$; hence if z is in the interior of $D_p(\Gamma)$, so that $\rho(z,p) < \rho(T(z),p)$ for all $T \in \Gamma - \{\mathrm{Id}\}$. If two points z_1, z_2 lie in the same Γ–orbit, we have $\rho(z_1,p) < (z_2,p)$ and $\rho(z_2,p) < (z_1,p)$, a contradiction. Thus the interior of $D_p(\Gamma)$ contains at most one point in each Γ-orbit. Being the intersection of closed half-planes, $D_p(\Gamma)$ is closed and convex. Thus $D_p(\Gamma)$ is path-connected, hence connected. □

Example 7 Let $\Gamma = PSL(2,\mathbb{Z})$. It is easy to check that the point $p = ki$ $(k > 1)$ is not fixed by any nonidentity element of Γ. We are going to show that the Dirichlet region $D_p(\Gamma)$ is the set F,

$$F = \left\{ z \in \mathcal{H} \mid |z| \geq 1, \mathrm{Re}(z) \leq \frac{1}{2} \right\}$$

is illustrated in Figure 3.2.2.

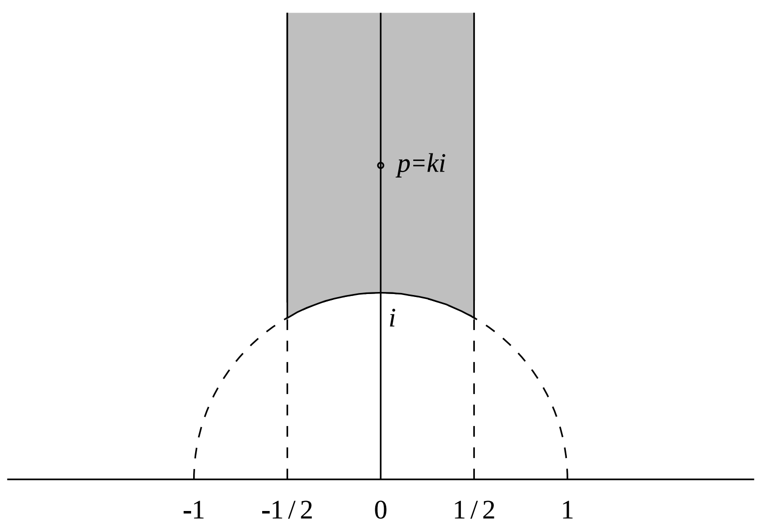

FIGURE 3.2.2. Fundamental region for $SL(2,\mathbb{Z})$

The isometries T and S, where $T(z) = z+1$, and $S(z) = -1/z$, are in G, and the geodesic sides of F are the perpendicular bisectors of the segments $[p, T(p)]$, $[p, T^{-1}(p)]$, and $[p, S(p)]$ respectively. This shows that $D_p \subset F$. Suppose $D_p \neq F$, then there exists a point $z \in \mathring{F}$ and $T \in \Gamma$ such that $T(z) \in \mathring{F}$. We write

$$T(z) = \frac{az+b}{cz+d} \ (a,b,c,d \in \mathbb{Z},\ ad-bc=1).$$

Then

$$|cz+d|^2 = c^2|z| + 2\mathrm{Re}(z)cd + d^2 > c^2 + d^2 - |cd| = (|c|-|d|)^2 + |cd|,$$

since $|z| > 1$ and $\mathrm{Re}(z) > -1/2$. The lower bound is nonnegative integer. It cannot be 0 since this would imply $c = d = 0$, which contradicts the equality $ad - bc = 1$. Therefore it is at least 1, so that $|cz + d| > 1$, and thus

$$\mathrm{Im}\, T(z) = \frac{\mathrm{Im}(z)}{|cz+d|^2} < \mathrm{Im}(z).$$

The same argument with z and T replaced by $T(z)$ and T^{-1} gives us the reverse inequality, $\mathrm{Im}(z) < \mathrm{Im}(T(z)$, a contradiction. Therefore $D_p(\Gamma) = F$. The fundamental region F is a hyperbolic triangle with angles $\pi/3, \pi/3, 0$. By the Gauss-Bonnet formula (Theorem 2.19), its area is finite and is equal to $\pi - (2\pi/3) = \pi/3$.

There is an alternative method of constructing fundamental regions. Since it is related to the isometric circle, it is more convenient to describe it for the unit disc model of the hyperbolic plane. Let Γ be a discrete group of orientation–preserving isometries of the unit circle $\mathcal{U}$. We assume that 0 is not an elliptic fixed point, i.e., that for all the transformations $T(z) = (az + \overline{c})/(cz + \overline{a})$ in the group Γ, $c \neq 0$. Let R_0 be the locus of points exterior to the isometric circles of all elements in Γ different from the identity,

$$R_0 = \overline{\bigcap_{T\in\Gamma-\{\mathrm{Id}} \}\hat{I}(T) \cap \mathcal{U}}.$$

THEOREM 3.26. *The set R_0 defined by the above formula is a fundamental region for Γ.*

PROOF. We will prove that R_0 is actually the Dirichlet region centered at the point 0, $D_0(\Gamma)$. This will follow immediately from the fact that for any $T \in \Gamma$, the perpendicular bisector of the segment $[0, T(0)]$ is the the isometric circle $I(T^{-1})$. We are using the formula for the unit disc model:

$$\cosh^2\left[\frac{1}{2}\rho(z,w)\right] = \frac{|1-z\overline{w}|^2}{(1-|z|^2)(1-|w|^2)},$$

which is the counterpart of formula (b) from Theorem 2.13. The perpendicular bisector is given by the equation $\rho(z, 0) = \rho(z, \overline{c}/\overline{a})$, which is equivalent to

$$\frac{1}{(1-|z|)^2} = \frac{|1-z\frac{c}{a}|}{(1-|z|)^2(1-\frac{|c|^2}{|a|^2})},$$

and ultimately to $|-cz+a| = 1$, which is the equation of the isometric circle $I(T^{-1})$. $\square$

THEOREM 3.27. *Suppose an infinite sequence of distinct isometric circles $I_1, I_2, \cdots$ of transformations of a Fuchsian group Γ with radii $r_1, r_2, \cdots$ is given. Then we have $\lim_{n\to\infty} r_n = 0$.*

PROOF. The transformations are of the form

$$T(z) = \frac{az+\bar{c}}{cz+\bar{a}} \; (a, c \in \mathbb{C}, |a|^2 - |c|^2 = 1). \tag{3.2.5}$$

Recall that the radius of $I(T)$ is equal to $1/|c|$. Let $\epsilon > 0$ be given. There are only finitely many $T \in \Gamma$ with $|c| < 1/\epsilon$. This follows from the discreteness of Γ and the relation $|a|^2 - |c|^2 = 1$. Hence there are only finitely many $T \in \Gamma$ with $I(T)$ of radius exceeding ϵ, and the theorem follows. $\square$

It is a general fact that elements of the group Γ which identify the sides of the Dirichlet fundamental region generate the group Γ. In particular, for the modular group, we have the following theorem.

THEOREM 3.28. *The group $PSL(2,\mathbb{Z})$ is generated by two elements, $T(z) = z+1$ and $S(z) = -1/z$.*

PROOF. (Suggested by A. Mezhirov.) Let $g(z) = \frac{az+b}{cz+d}$ be a transformation in $PSL(2,\mathbb{Z})$. We will show that it can be represented as the composition of a finite number of transformations T, T^{-1}, and S. Since $ad-bc = 1$, the integers a and c are relatively prime or one of them is equal to 0. If $a = 0$, either $b = -1$, $c = 1$, or vice versa. In the first case $T^{-d}S \circ g = 1_2$, and hence $g = S \circ T^d$, and in the second, $g = S \circ T^{-d}$. Similarly, if $c = 0$, $g(z) = z + b$, or $g(z) = z - b$, i.e., $g = T^b$, or $g = T^{-b}$.

Now assume $a, c \neq 0$. The algorithm of factorization of the matrix corresponding to g is essentially the Euclidean algorithm for finding the greatest common divisor of $|a|$ and $|c|$, which in this case is equal to 1. We may assume that $c > 0$. If $|a| \geq c$, we can write $|a| = qc + r$, where q, r are positive integers and $r < c$. If $a > 0$, then we apply T^{-q} to g to obtain the transformation

$$T^{-q} \circ g(z) = \frac{rz+b'}{cz+d},$$

and by applying S, we obtain

$$S \circ T^{-q} \circ g(z) = \frac{-cz-d}{rz+b'}.$$

If $a < 0$, we apply $S \circ T^q$ to g to obtain

$$S \circ T^{q} \circ g(z) = \frac{-cz-d}{rz-b''}.$$

In both cases, after the first step, we obtain the transformation given by $(a_1z+b_1)/(c_1z+d_1)$ with $|a_1| \geq |c_1|$, and $|a_1| < |a|$. In finitely many steps we arrive at the transformation $(a_nz+b_n)/(c_nz+d_n)$ with $a_n = \pm 1$ and $c_n = 0$, and this case has been already considered above. If $|a| < |c|$, we first apply the transformation S to reduce the problem to the case already considered. □

Exercises

31. Let $\Gamma(2)$ be the subgroup of $PSL(2,\mathbb{Z})$ which consists of transformations $z \to (az+b)/(cz+d)$ for which $a \equiv d \equiv 1 \pmod 2$ and $b \equiv d \equiv 0 \pmod 2$. Prove that $\Gamma(2)$ is a subgroup of $PSL(2,\mathbb{Z})$ of index 6. Prove that $\Gamma(2)$ is discrete and is generated by the transformations

$$A(z) = z+2 \text{ and } B(z) = \frac{z}{2z+1},$$

and find its Dirichlet fundamental region centered at $p = i$.

32. Show that the Dirichlet region can be described by using the Euclidean metric via the following formula:

$$D_p(\Gamma) = \left\{ z \in \mathcal{H} \mid \left| \frac{T(z)-p}{z-p} \right| \geq \frac{1}{|cz+d|} \text{ for all } T \in \Gamma \right\}.$$

33. Prove that

(i) T is hyperbolic if and only if $I(T)$ and $I(T^{-1})$ do not intersect;
(ii) T is elliptic if and only if $I(T)$ and $I(T^{-1})$ intersect;
(iii) T is parabolic if and only if $I(T)$ and $I(T^{-1})$ are tangential.

34. Let Γ be the group acting on $\mathcal{H}$ generated by the transformations

$$f(z) = 2z,\ g(z) = \frac{3z+4}{2z+3}.$$

Prove that Γ is discrete and find a Dirichlet fundamental region F for Γ.

Lecture 4
Coding closed geodesics on the modular surface

4.1. The modular surface and closed geodesics

The quotient $\Gamma\backslash\mathcal{H}$ of the hyperbolic plane $\mathcal{H}$ by the modular group $\Gamma = SL(2,\mathbb{Z})$ is said to be the *modular surface*. Let F be a fundamental region for Γ, and $\pi : \mathcal{H} \to \Gamma\backslash\mathcal{H}$ be the natural projection (continuous and open). The points of $\Gamma\backslash\mathcal{H}$ are the Γ–orbits. The restriction of π to F identifies the congruent points of F (which necessarily belong to its boundary ∂F) and takes $\Gamma\backslash F$ to an oriented surface with possibly some *marked points* (which correspond to the elliptic cycles of F) and *cusps* (which correspond to noncongruent vertices at infinity of F). Such a surface is known as an *orbifold*. Its topological type is determined by the number of cusps and by its *genus*, i.e., the number of handles in the surface viewed as a sphere with handles. If F is a Dirichlet fundamental region, the quotient space $\Gamma\backslash\mathcal{H}$ is homeomorphic to $\Gamma\backslash F$.

We haveseen in §3.1 (Theorem 3.22) that the area of a fundamental region (with nice boundary) is, if finite, a numerical invariant of the group Γ. Since the area in the quotient space $\Gamma\backslash\mathcal{H}$ is induced by the hyperbolic area in $\mathcal{H}$, the *hyperbolic area* of $\Gamma\backslash\mathcal{H}$, denoted by $\mu(\Gamma\backslash\mathcal{H})$, is well defined and equal to $\mu(F)$ for any fundamental region F. If Γ has a compact Dirichlet region F, then F has finitely many sides, and the quotient space $\Gamma\backslash\mathcal{H}$ is compact (in this case, Γ is called *co-compact*). If one Dirichlet region for Γ is compact, then all Dirichlet regions are compact. If, in addition, Γ acts on $\mathcal{H}$ without fixed points, $\Gamma\backslash\mathcal{H}$ is a compact *Riemann surface*–a 1-dimensional complex manifold–and its fundamental group is isomorphic to Γ. The above–mentioned material is discussed in detail in [**3**].

We shall view the standard fundamental region F for $SL(2,\mathbb{Z})$ as a quadrilateral, rather than a triangle, with the point i dividing the circular side of the boundary into two parts, left and right (see Fig. 3.2.2). Under the projection π, the left vertical side is identified with the right one by the transformation $T(z) = z + 1$, and the left circular side is identified with the right one by the transformation $S(z) = -1/z$. After the identifications, we obtain a topological sphere with two marked points corresponding to elliptic elements of order 2 and 3, and one cusp at infinity.

The *tangent bundle* to $\mathcal{H}$ is defined by

$$T\mathcal{H} = \{(z,\zeta) \mid z \in \mathcal{H}, \zeta \in T_z\mathcal{H}\},$$

and the *unit tangent bundle* is defined by

$$S\mathcal{H} = \{(z, \zeta) \mid z \in \mathcal{H}, \zeta \in T_z\mathcal{H}, \ \|\zeta\| = 1\},$$

where $\|\cdot\|$ is the norm in $T_z\mathcal{H}$ introduced in Section 2.1.

By Theorem 2.16, the group $PSL(2, \mathbb{R})$ acts on $T\mathcal{H}$ by differentials:

$$T(z, \zeta) = (T(z), DT_z(\zeta)), \tag{4.1.1}$$

this action being induced by the action on $\mathcal{H}$ by Möbius transformations.

THEOREM 4.29. *There is a homeomorphism between $PSL(2, \mathbb{R})$ and the unit tangent bundle $S\mathcal{H}$ of the upper-half plane $\mathcal{H}$ such that the action of $PSL(2, \mathbb{R})$ on itself by left multiplications corresponds to the action of $PSL(2, \mathbb{R})$ on $S\mathcal{H}$* (4.1.1).

PROOF. Let (i, ζ_0) be a fixed element of $S\mathcal{H}$, where $\mathbb{Z}_0$ is the unit vector at the point i tangent to the imaginary axis and pointed upwards, and let (z, ζ) be an arbitrary element of $S\mathcal{H}$. By Exercise 12, there exists a unique transformation $T \in PSL(2, \mathbb{R})$ sending the imaginary axis to the geodesic passing through z and tangent to ζ, and satisfying $T(i) = z$. Then $DT_i(\zeta_0) = \zeta$, and hence,

$$T(i, \zeta_0) = (z, \zeta). \tag{4.1.2}$$

The map $(z, \zeta) \mapsto T$ is a homeomorphism between $S\mathcal{H}$ and $PSL(2, \mathbb{R})$ (see Exercise 35).

For $S \in PSL(2, \mathbb{R})$, let $S(z, \zeta) = (z', \zeta')$. By (4.1.2), $S(z, \zeta) = ST(i, \zeta_0)$. Hence the element $S(z, \zeta)$ is identified with the transformation ST, and the last assertion follows. $\square$

PROPOSITION 4.30. *Closed geodesics on the modular surface $M = \Gamma\backslash\mathcal{H}$ are in one–to–one correspondence with conjugacy classes of hyperbolic elements in* Γ.

PROOF. Recall that a hyperbolic transformation $T \in PSL(2, \mathbb{R})$ has two fixed points in $\mathbb{R} \cup \{\infty\}$, one attracting, denoted by u, the other repelling, denoted by w. Let z be any point on the axis of T, the geodesic in $\mathcal{H}$ from u to w, which we denoted by $C(T)$, and let ζ be the unit vector tangent to $C(T)$. Then $T(z) \in C(T)$, and by Exercise 36, $DT(\zeta)$ is the unit vector tangent to $C(T)$ at the point $T(z)$. This means that the geodesic $C(T)$ will be closed on M.

Conversely, suppose C is a closed oriented geodesic on M. Let us lift it to $\mathcal{H}$, and assume that it intersects the given fundamental region F (otherwise we apply a transformation from $PSL(2, \mathbb{Z})$ to move it there). We follow the geodesic in its direction from u to w, and as soon at it reaches a side of ∂F, apply a transformation identifying this side with its image. Thus we obtain a geodesic on F, which closes up after finitely many steps. This means that there exists a finite string of generators of $PSL(2, \mathbb{Z})$, namely T, T^{-1}, S, such that after their successive application we return to our original geodesic, i.e., for some $\gamma_0 \in PSL(2,\mathbb{Z})$, we have $\gamma_0(C) = C$. If follows from the classification of elements in $PSL(2,\mathbb{Z})$ that γ_0 is hyperbolic, and that C is its axis, and that $z \in C$ implies that $\gamma_0^n(z)$ tends to u as $n \to \infty$, not to w. Therefore, if we want a hyperbolic element whose axis is the oriented geodesic C, we must take $\gamma = \gamma_0^{-1}$. Axes of transformations conjugate in $PSL(2,\mathbb{Z})$ produce the same closed geodesic on M. □

The element γ described above, which fixes a closed oriented geodesic C, is a "word" in the generators T, T^{-1}, S. It is easy to see that it must contain at least one S, an S cannot be followed by another S, and a T cannot be followed by a T^{-1} and vice versa. Thus it is a sequence of blocks, defined up to a cyclic permutation, consisting of T's and T^{-1}'s that are separated by S's. If we choose the initial point on the circular part $a_1 \cup a_2$ of the boundary of F, we see that the sequence always ends by an S. To each block of T's we assign a positive integer equal to its length, and to each block of T^{-1}'s, we assign a negative integer whose absolute value is equal to its length. Thus we obtain a finite sequence of integers $[n_1, n_2, \dots, n_m]$, defined up to a cyclic permutations, called the *geometric code* of C. It is clearly $PSL(2,\mathbb{Z})$–invariant and therefore we will refer to it also as the *geometric code* of the conjugacy class of γ and denote it by $[\gamma]$. Moreover, we have $\gamma = T^{n_1}S^{n_2}S \cdot T^{n_m}S$.

A convenient way to obtain the geometric code of C is to count the number of times C hits the vertical sides of the boundary of F, so that a positive integer is assigned to each block of hits on the right vertical side, and a negative one, to each block of hits on the left vertical side.

Fig. 4.1.1 shows the closed geodesic in F for the matrix

$$A = \begin{pmatrix} 15 & -8 \\ 2 & -1 \end{pmatrix}.$$

Following the closed geodesic in F, we obtain its geometric code $[A] = [6, -2]$.

The coding sequence of a geodesic passing through the vertex ρ of F in the clockwise direction obeys the convention that it must exit F through the right vertical side. The axis of

$$A_4 = \begin{pmatrix} 4 & -1 \\ 1 & 0 \end{pmatrix},$$

passing through the vertex ρ, and the corresponding closed geodesic in F, are shown in Fig. 4.1.2. According to our convention, its geometric code is $[4]$.

Exercises

35. Prove that the map $(z, \zeta) \mapsto T$ described in Theorem 4.29 is a homeomorphism.

36. Let L be a geodesic on $\mathcal{H}$ and ζ be the unit tangent vector to L at the point $z \in L$. Prove that under a Möbius transformation $T \in PSL(2,\mathbb{R})$,

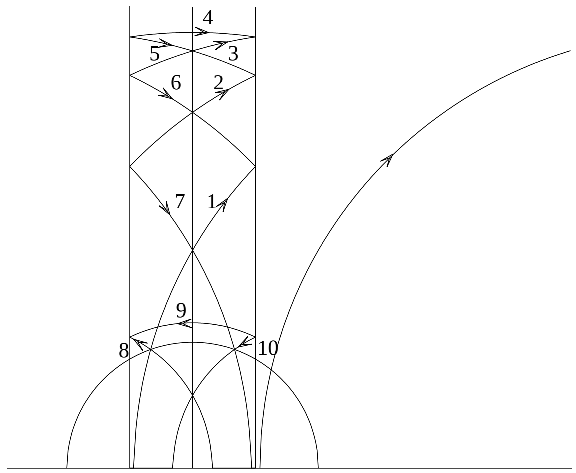

FIGURE 4.1.1. Closed geodesic in F

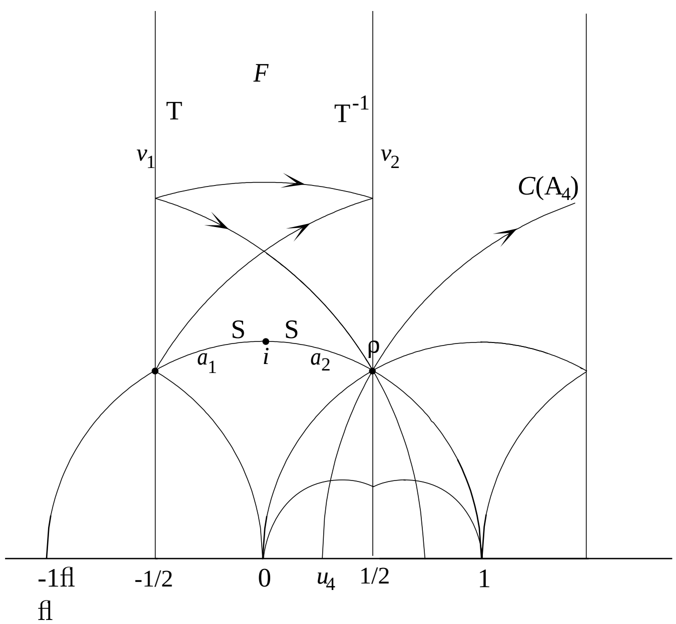

FIGURE 4.1.2. Closed geodesic in F corresponding to A_4

the vector $DT(\zeta)$ is the unit tangent vector to the geodesic $T(L)$ at the point $T(z)$.

4.2. Arithmetic coding of closed geodesics

We have seen in Lecture 1 that the minus continued fraction expansion of a quadratic irrationality is eventually periodic. Our goal is to prove that two matrices in $SL(2,\mathbb{Z})$ are conjugate in $SL(2,\mathbb{Z})$ if and only if the periods of their minus continued fraction expansions differ by a cyclic permutation.

PROPOSITION 4.31. *Two quadratic irrationalities are obtained from one another by an application of a transformation from $SL(2,\mathbb{Z})$ if and only if the periods in their minus continued fraction expansions are cyclic permutations of one another.*

PROOF. If two quadratic irrationalities have periods in their minus continued fraction expansions, which are cyclic permutations of one another, one can be obtained from the other by consecutive applications of transformations $T(z) = z+1$, $T^{-1}(z) = z-1$, and $S(z) = -1/z$. Since those transformations are in $SL(2,\mathbb{Z})$, the claim in this direction follows.

Since the transformations S and T generate $SL(2,\mathbb{Z})$ (see Theorem 3.28), it is sufficient to prove the converse only for these particular transformations. Let w be a quadratic irrationality:

$$w = (n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}}).$$

This representation is not unique: we can extend the part before the period by adding a period to it if we need to. Then obviously

$$T^{\pm 1}(w) = (n_0 \pm 1, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}}).$$

In order to deal with S, we first notice that if $n_0 \geq 2$, then

$$S(w) = (0, n_0, n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}})$$

which is a legitimate minus continued fraction expansion. We recall that the following relations between S and T hold (in fact, these relations define $SL(2,\mathbb{Z})$, but we do not use this here):

$$S^2 = \mathrm{Id}, \quad STSTST = \mathrm{Id},$$

where Id denotes the identity transformation. In the next argument we use the following consequences of the second relation:

$$STS = T^{-1}ST^{-1}, \quad ST^{-1}S = TST;$$

further, for $p \geq 2$ we have

$$ST^{-p}S = TS\underbrace{T^2S\dots T^2S}_{p-1 \text{ times}}T.$$

If $n_0 \leq -1$, we obtain

$$S(w) = (1, \underbrace{2, \dots, 2}_{-n_0-1 \text{ times}}, n_1+1, n_2, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}}).$$

If $n_0 = 0$, we can write

$$S(w) = (n_1, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}}).$$

Now let $n_0 = 1$. If $n_1 \geq 3$, we have

$$S(w) = (-1, n_1 - 1, n_2, \dots, n_k, \overline{n_{k+1}, \dots, n_{k+m}}).$$

Since w is irrational, there is an n_i in the period that is greater than 2, so we suppose that $n_s \geq 3$, and $n_i = 2$ for all $1 \leq i \leq s-1$. Then

$$S(w) = (-s, n_s - 1, \ldots, n_k, \overline{n_{k+1}, \ldots, n_{k+m}}),$$

concluding the proof of the proposition. □

The following lemma holds for all Fuchsian groups.

LEMMA 4.32. *Let Γ be a Fuchsian group, and let $\gamma_1, \gamma_2 \in \Gamma$ be hyperbolic elements having a common fixed point. Then their second fixed points also coincide, hence they have the same axis, and both are powers of a primitive matrix with the same axis.*

PROOF. After a standard conjugation, we may assume that both γ_1 and γ_2 fix ∞, so

$$\gamma_1(z) = \lambda z \ (\lambda > 1) \quad \text{and} \quad \gamma_2(z) = \mu z + k \ (\mu \neq 1,\ k \neq 0).$$

Then

$$\gamma_1^{-n}\gamma_2\gamma_1^n(z) = \lambda^{-n}(\mu(\lambda^n z) + k) = \mu z + \lambda^{-n}k.$$

Hence $\|\gamma_1^{-n}\gamma_2\gamma_1^n\| = \sqrt{\mu^2 + \lambda^{-2n}k^2 + 1}$ is bounded as n tends to ∞, and so the sequence $\{\gamma_1^{-n}\gamma_2\gamma_1^n\}$ contains a converging subsequence of distinct terms, a contradiction with the fact that Γ is discrete. Therefore $k = 0$, and so both γ_1 and γ_2 fix 0. □

THEOREM 4.33. *Two hyperbolic matrices A and B in $SL(2, \mathbb{Z})$ with the same trace are conjugate in $SL(2, \mathbb{Z})$ if and only if their attracting (repelling) fixed points have periods in their minus continued fraction expansions that are cyclic permutations of one another.*

PROOF. Let w_A and w_B be attracting fixed points of A and B, respectively, such that the periods in their minus continued fraction expansion differ by a cyclic permutation. Then by Proposition 4.31 there exists a $C \in SL(2, \mathbb{Z})$ such that $w_A = C w_B$. Then the matrices CBC^{-1} and A have the same fixed point w_A, and by Lemma 4.32, since they have the same trace, either $CBC^{-1} = A$ or $CBC^{-1} = A^{-1}$. But both w_A and w_B are attracting, w_A is attracting for both, A and CBC^{-1}, and therefore $CBC^{-1} = A$.

Conversely, suppose two matrices in $SL(2, \mathbb{Z})$ are conjugate. Then their attracting fixed points w_A and w_B are obtained from each other by an application of a matrix C from $SL(2, \mathbb{Z})$. Then by Proposition 4.31, the periods in the minus continued fraction expansions of w_A and w_B differ by a cyclic permutation. □

Thus, we have obtained a $PSL(2, \mathbb{Z})$–invariant of closed geodesics on M, namely, the period of the minus continued fraction expansions of their attracting fixed points (determined up to a cyclic permutation), i.e., a finite sequence of integers $(n_1, n_2, \ldots, n_m)$ defined up to a cyclic permutation, called its *arithmetic code.*

Exercises

37. Find the arithmetic code of the matrix

$$A = \begin{pmatrix} 15 & -8 \\ 2 & -1 \end{pmatrix}$$

and compare it with its geometric code.

4.3. Gauss Reduction Theory in matrix language

To any hyperbolic matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z})$$

we associate the integral binary quadratic form

$$Q_A(x, y) = cx^2 + (d - a)xy - by^2 \tag{4.3.1}$$

with discriminant $D = (a + d)^2 - 4 > 0$ (such forms are called *indefinite*). It is easy to see that D is not a perfect square (Exercise 38).

Conversely, to each integral indefinite binary quadratic form $Q(x, y) = Ax^2+Bxy+Cy^2$ whose discriminant $D = B^2-AC > 0$ is not a perfect square (we assume that the integers A, B, C have no common factor) we associate a geodesic on $\mathcal{H}$ connecting the roots of the corresponding quadratic equation

$$Az^2 + Bz + C = 0.$$

Its image in $M = PSL(2, \mathbb{Z})\backslash\mathcal{H}$ is closed since there exists a hyperbolic matrix $U \in SL(2, \mathbb{Z})$ with the same axis (this is not immediately obvious and rather delicate, see Exercise 41).

We introduce the following equivalence relation on the set of all integral binary quadratic forms with the same discriminant.

Definition. Two integral binary quadratic forms

$$Q_1(x, y) = A_1x^2 + B_1xy + C_1y^2 \text{ and } Q_2(x, y) = A_2x^2 + B_2xy + C_2y^2$$

are said to be *equivalent in the narrow sense* if there is a matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z})$$

such that $Q_2(ax + by, cx + dy) = Q_1(x, y)$.

The Gauss Reduction Theory for integral indefinite binary quadratic forms allows to determine when two quadratic forms with the same discriminant are equivalent in the narrow sense. It is easy to check that two hyperbolic matrices with the same trace are conjugate in $SL(2, \mathbb{Z})$ if and only if the corresponding quadratic forms (with the same discriminant) are equivalent in the narrow sense (Exercise 42). Thus the two theories are equivalent. Gauss's notion of "reduced" binary quadratic form translates into the following notion of "reduced" matrix, which is not connected with any particular fundamental region.

DEFINITION. A hyperbolic matrix in $SL(2,\mathbb{Z})$ is called *reduced* if its attracting and repelling fixed points, denoted by w and u respectively, satisfy

$$w > 1, \quad 0 < u < 1.$$

THEOREM 4.34. [Reduction Algorithm.] *There is a finite number of reduced matrices in $SL(2,\mathbb{Z})$ with a given trace t, $|t| > 2$. Any hyperbolic matrix in $SL(2,\mathbb{Z})$ with trace t can be reduced by a finite number of standard conjugations. Applied to a reduced matrix A, this conjugation gives another reduced one. Any reduced matrix conjugate to A is obtained from A by a finite number of standard conjugations. Thereby the set of reduced matrices is decomposed into disjoint cycles of conjugate matrices.*

PROOF. The proof of the first assertion is an adaptation for matrices of the proof in [**5**]. Suppose the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is reduced. Let $k = a - d - 2c$. By Exercise 39, we have $|k| < \sqrt{D}$, hence k can take only finitely many values for a given $D = t^2 - 4$. We have $D - k^2 = 4c(a + b - c - d) > 0$, therefore $c\,|(D - k^2)/4$ can also take only finitely many values. We can express a, b, and d in terms of c and k as follows:

$$a = \frac{t + k + 2c}{2}, \; b = \frac{D - k^2}{4c} - (k + c), \; d = \frac{t - k - 2c}{2},$$

thus the number of reduced matrices with given trace t is finite, and the first assertion of the theorem is proved.

Consider the attracting fixed point of A; it has a minus continued fraction expansion of the form

$$(n_0, n_1, \ldots, n_k, \overline{n_{k+1}, \ldots, n_{k+m}}),$$

where the notations are as in Lecture 1. Conjugating A by $S^{-1}T^{-n_0}$, we obtain a matrix $A_0 = S^{-1}T^{-n_0}AT^{n_0}S$, and inductively,

$$A_i = S^{-1}T^{-n_i}A_{i-1}T^{n_i}S$$

for $i = 1, 2, \ldots$. The attracting fixed point of the matrix

$$A_k = (S^{-1}T^{-n_k}S^{-1}\ldots T^{-n_1}S^{-1}T^{-n_0})A(ST^{-n_k}S\ldots T^{-n_1}ST^{-n_0})^{-1},$$

w, has a purely periodic minus continued fraction expansion

$$w = (\overline{n_{k+1}, \ldots, n_{k+m}}),$$

and according to Theorem 1.3, we have $w > 1$, $0 < u < 1$, i.e., A_k is reduced. Applying the same procedure to A_k, we obtain m reduced matrices in a sequence corresponding to the period of w.

Conversely, if two reduced matrices are conjugate, their attracting fixed points have pure periodic minus continued fraction expansions whose periods, by Proposition 4.31, differ by a cyclic permutation. Hence they belong to the same cycle and are obtained from one another by a finite number of standard conjugations. □

REMARK. Theorem 4.34 asserts the finiteness of the number of conjugacy classes of matrices in $SL(2,\mathbb{Z})$ with given trace t, which corresponds to the class number of the real quadratic field $\mathbb{Q}(\sqrt{t^2-4})$ (in the narrow sense), is a standard and very important fact in number theory. This fact, however, is much more general, it is valid for all Fuchsian groups of the first kind. For the co–compact Fuchsian groups it follows from the expansiveness of the geodesic flow and can be found e.g. in [**4**], pp. 212, 549, 569–70.

Exercises

38. Prove that if $a, d \in \mathbb{Z}$ and $|a = d| > 2$, $D = (a+d)^2 - 4$ is not a perfect square.

39. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be reduced, and $k = a - d - 2c$. Prove that $|k| < \sqrt{D}$.

40.* Prove that the length of a closed geodesic with the arithmetic code $(n_1, \ldots, n_m)$ is equal to

$$2 \log \prod_{i=1}^{m} w_i,$$

where $w_1, \ldots, w_m$ are the attracting fixed points of all reduced matrices corresponding to this closed geodesic.

41.* Let L be a geodesic on $\mathcal{H}$ connecting the roots of the quadratic equation

$$Az^2 + Bz + C = 0$$

with $D = B^2 - 4AC > 0$, not a perfect square. Prove that there exists a hyperbolic matrix $U \in SL(2,\mathbb{Z})$ with the same axis. (*Hint*: This problem is for those who know some algebraic number theory. The set of integral matrices having this axis is the real quadratic field $\mathbb{Q}(\sqrt{D})$, where U corresponds to a nontrivial unit of norm 1.)

42. Prove that two hyperbolic matrices with the same trace are conjugate in $SL(2,\mathbb{Z})$ if and only if the corresponding quadratic forms (with the same discriminant) are equivalent in the narrow sense.

Contents

Bibliography

1. A. Beardon, *The Geometry of Discrete Groups*, Springer-Verlag, New York, 1983.
2. S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145.
3. S. Katok, *Fuchsian groups*, University of Chicago Press, Chicago, London, 1992.
4. A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, Encyclopedia of Mathematics and its Applications, **54**, Cambridge University Press, New York, 1995.
5. D. Zagier, *Zetafunktionen und quadratische Körper: eine Einführung in die höhere Zahlentheorie*, Hochschultext, Springer-Verlag, Berlin, Heidelberg, New York, 1982.