Clay Mathematics Proceedings
Volume 8, 2008

Fuchsian groups, geodesic flows on surfaces of constant negative curvature and symbolic coding of geodesics

Svetlana Katok

Dedicated to the memory of my father Boris Abramovich Rosenfeld (1917-2008)

Contents

Introduction

These are the notes of an introductory four-lectures course given in the Summer School in Pisa. Lectures I and II cover hyperbolic geometry and the theory of Fuchsian groups; the material of these lectures is mostly an adaptation from my book "Fuchsian groups" [**13**]. Lecture III describes the geodesic flow on the surfaces of constant negative curvature and establishes its dynamical properties. Lecture IV is devoted to continued fractions and their relation to coding of geodesics. The material of this lecture is based on the survey article [**16**] in view of an ongoing project of the author with Ilie Ugarcovici and Don Zagier, but does not include the new results that will be published elsewhere.

Lecture I. Hyperbolic geometry

1. Models of hyperbolic geometry

Our first model of hyperbolic geometry is obtained similarly to the model of elliptic geometry on the unit sphere S^2 in $\mathbb{R}^3$,

$$S^2 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}.$$

The metric (arc length) on S^2 is induced from the Euclidean metric on $\mathbb{R}^3$,

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 \tag{1.1}$$

which corresponds to the standard inner product on $\mathbb{R}^3$,

$$(x, y) = x_1y_1 + x_2y_2 + x_3y_3.$$

The geodesics for this metric (i.e. the length minimizing curves) lie on planes through $0 \in \mathbb{R}^3$ and are arcs of great circles. The group of orientation-

preserving isometries of S^2 is the group $SO(3)$ that preserves the standard inner product $(\ ,\ )$ on $\mathbb{R}^3$.

If instead of the metric (1.1) we consider a *pseudo-metric* in $\mathbb{R}^3$:

$$ds_h^2 = dx_1^2 + dx_2^2 - dx_3^2, \tag{1.2}$$

corresponding to the bilinear symmetric form of signature $(2,1)$:

$$(x,y)_{2,1} = x_1y_1 + x_2y_2 - x_3y_3,$$

then the upper fold of the hyperboloid

$$H^2 = \{(x_1,x_2,x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 - x_3^2 = -1, \quad x_3 > 0\}$$

represents a model of hyperbolic geometry in two dimensions. Notice that outside of H^2 $(x,x)_{2.1} > -1$ and inside H^2 $(x,x)_{2.1} < -1$.

First, we check that the pseudo-metric (1.2) induces Riemannian metric on H^2. Let $x = (x_1,x_2,x_3) \in H^2$. Define $x^\perp = \{y \in \mathbb{R}^3 \mid (y,x)_{2,1} = 0\}$. It is a plane passing though 0, and $x + x^\perp$ is a plane passing through x.

PROPOSITION 1.1. *The tangent plane T_xH^2 to the hyperboloid H^2 at the point x is given by $T_xH^2 = x + x^\perp$, and $(\ ,\ )_{2,1}$ restricted to $x^\perp$ is positive-definite, hence gives a scalar product on T_xH^2, i.e. a Riemannian metric on H^2.*

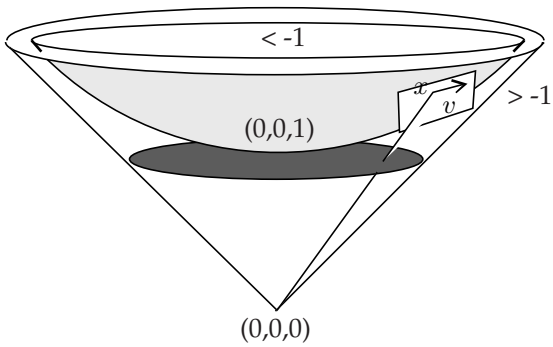


FIGURE 1.1. The hyperboloid model

PROOF. The upper fold of the hyperboloid is given by the equation $x_3 = \sqrt{x_1^2 + x_2^2 + 1}$. Let $x + v \in T_xH^2$. A tangent vector v at x is a linear combination of two basic tangent vectors,

$$v = a(1,0,\frac{\partial x_3}{\partial x_1}) + b(0,1,\frac{\partial x_3}{\partial x_2}) = (a,b,\frac{ax_1 + bx_2}{x_3}).$$

We see that $(v,x)_{2,1} = 0$, i.e. $v \in x^\perp$, hence $T_xH^2 \subset x + x^\perp$, and since T_xH^2 is a plane, $T_xH^2 = x + x^\perp$.

Let $v \in x^\perp$. By convexity of the hyperboloid, $x + v$ is outside of H^2, hence

$$-1 < (x+v, x+v)_{2,1} = (x,x)_{2,1} + 2(x,v)_{2,1} + (v,v)_{2,1} = -1 + (v,v)_{2,1}.$$

Therefore for all $v \in x^\perp$, $(v,v)_{2,1} > 0$. □

The geodesics for this metric also lie on planes through $0 \in \mathbb{R}^3$. The group of orientation-preserving isometries of H^2 is $SO(2,1)$, the group preserving the bilinear symmetric form $(x,y)_{2,1}$,

$$SO(2,1) = \{A \in SL(3,\mathbb{R}) \mid {}^TASA = S, \text{ where } S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}\}.$$

Other models of the hyperbolic plane are obtained from the **hyperboloid model** described above.

The Beltrami-Klein model. The group $G = SO(2,1)$ acts transitively on the upper fold of the hyperboloid H^2 by linear transformations. The cone given by the equation $x_1^2 + x_2^2 - x_3^2 = 0$ lies outside of H^2 and is asymptotic to it, as illustrated on Figure **??**. The intersection of this cone with the plane $x_3 = 1$ tangent to H^2 is the circumference $\partial\mathcal{U} = \{x_1^2 + x_2^2 = 1\}$. Consider the central projection σ of H^2 to the plane $x_3 = 1$ from the origin $0 \in \mathbb{R}^3$. We have

$$\sigma(x_1, x_2, x_3) = (\eta_1, \eta_2),$$

where $\eta_1 = \frac{x_1}{x_3}, \eta_2 = \frac{x_2}{x_3}$. We have $\eta_1^2 + \eta_2^2 = 1 - \frac{1}{x_3^2}$, hence $\sigma(H^2) = \mathcal{U} = \{\eta_1^2 + \eta_2^2 < 1\}$. Equivalently, $\mathcal{U}$ may be viewed as the space of *negative* vectors $x \in \mathbb{R}^3$, i.e. such that $(x,x)_{2,1} < 0$, which will be useful later. The metric on $\mathcal{U}$ is induced by the hyperbolic metric d_h on H^2:

$$d_h^*(\eta_1, \eta_2) = d_h(\sigma^{-1}\eta_1, \sigma^{-1}\eta_2).$$

Geodesics in H^2 are mapped to chords of the unit disc $\mathcal{U}$, which thus become geodesics with respect to the hyperbolic metric d_h^* on $\mathcal{U}$ (in what follows we will omit $*$ in most cases). We define the action of G on $\mathcal{U}$ so that it commutes with σ. It follows that G acts on $\mathcal{U}$ by fractional linear transformations: for (η_1, η_2) and $g = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$

$$g(\eta_1, \eta_2) = (\frac{a_{11}\eta_1 + a_{12}\eta_2 + a_{13}}{a_{31}\eta_1 + a_{32}\eta_2 + a_{33}}, \frac{a_{21}\eta_1 + a_{22}\eta_2 + a_{23}}{a_{31}\eta_1 + a_{32}\eta_2 + a_{33}}).$$

Notice that this model is not angle-true. Two geodesics which meet at the boundary are in fact asymptotically tangent.

The hemispherical model. Let $\eta_1 = \frac{x_1}{x_3}, \eta_2 = \frac{x_2}{x_3}, \eta_3 = \frac{1}{x_3}$. Then from the equation of H^2 we obtain the equation of the unit hemisphere

$$\eta_1^2 + \eta_2^2 + \eta_3^2 = 1, \ \eta_3 > 0$$

This model is obtained from the Beltrami-Klein model by the orthogonal projection of the unit disc $\mathcal{U}$ to the hemisphere. The geodesics in this model are arcs of circles on the hemisphere orthogonal to the disc—the boundary of the hemispere.

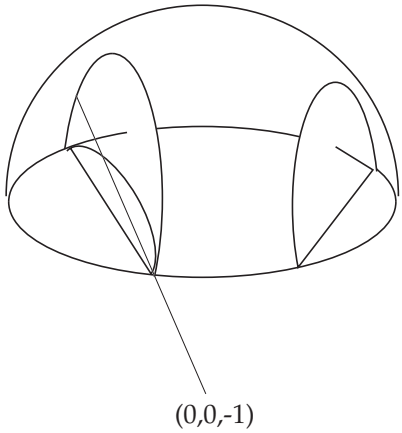


FIGURE 1.2. The hemispherical, Beltrami-Klein, and Poincaré disc models

The Poincaré disc model. The stereographic projection of the hemisphere from the point $(0, 0, -1)$ onto the plane $\eta_3 = 0$ (i.e. to the same unit disc $\mathcal{U}$) maps geodesics in the hemisphere to arcs of circles orthogonal to the boundary $\partial\mathcal{U}$. This gives us a new model in the unit disc, the Poincaré disc model. If we go from the Beltrami-Klein model to the Poincaré model (through the hemisphere) we notice that the end points of the geodesics are preserved and each point with polar coordinates (r, φ) is mapped to the point on the same radius (r', φ), where $r' = \frac{r}{\sqrt{1-r^2}+1}$.

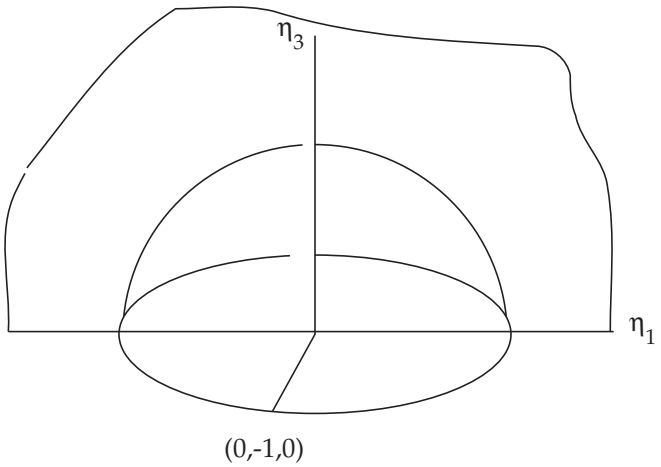


FIGURE 1.3. The Poincaré upper half-plane model

The Poincaré upper half-plane model. The stereographic projection of the upper hemisphere from the point $(0,-1,0)$ onto the plane $\eta_2 = 0$ give the model in the half-plane $\mathcal{H} = \{(\eta_1, \eta_3),\ \eta_3 > 0\}$.

Since the stereographic projection is *conformal* (i.e. preserves angles) the hemispherical model and its derivatives, the Poincaré disc model and the Poincaré upper half-plane model are angle-true.

Models as homogeneous spaces. All three models are obtained algebraically as homogeneous spaces G/K due to the accidental isomorphisms $SL(2,\mathbb{R}) \approx SU(1,1) \approx SO(2,1)$.

In the Beltrami-Klein model

$$G = SO(2,1), \quad K = \{\begin{pmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}\}.$$

In the Poincaré disc model, $G = SU(1,1)$, the group that preserves the Hermitian form on $\mathbb{C}^2$, $< z, w >= z_1\bar{w}_1 - z_2\bar{w}_2$ (for $z = (z_1, z_2)$ and $w = (w_1, w_2)$), is

$$SU(1,1) = \{g \in SL(2,\mathbb{C}) \mid g = \begin{pmatrix} a & c \\ \bar{c} & \bar{a} \end{pmatrix}\},$$

and

$$K = \{\begin{pmatrix} e^{i\varphi} & 0 \\ 0 & e^{-i\varphi} \end{pmatrix}\}.$$

The homogeneous space G/K can be identified with the "projectivized" space of the *negative* vectors in $\mathbb{C}^2$ ($< z, z >< 0$), analogous to that discussed above for $\mathbb{R}^3$, or, in homogeneous coordinates, with the unit disc in $\mathbb{C}$,

$$\mathcal{U} = \{z \in \mathbb{C} \mid |z| < 1\}.$$

In the Poincaré upper half-plane model, $G = SL(2,\mathbb{R})$, and $K = SO(2)$, the stabilizer of the point $i \in \mathcal{H}$. Here the homogeneous space G/K is identified with the upper half-plane

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

by the following construction. Each matrix in $SL(2,\mathbb{R})$ can be written as a product of upper-triangular and orthogonal (the Iwasawa decomposition):

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \sqrt{y} & \frac{x}{\sqrt{y}} \\ 0 & \frac{1}{\sqrt{y}} \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},$$

where $x, y \in \mathbb{R},\ y > 0$. Then $\pi : G/K \to \mathcal{H}$ given by

$$\pi(g) = g(i) = \frac{ai + b}{ci + d} = x + iy = z$$

does the identification.

In the last two conformal models, the corresponding group G acts by fractional-linear transformations: for $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $g(z) = \frac{az+b}{cz+d}$.

2. The hyperbolic plane

Let $\mathcal{H} = \{z \in \mathbb{C} \mid \mathrm{Im(z)} > 0\}$ be the upper-half plane. Equipped with the metric

$$ds = \frac{\sqrt{dx^2 + dy^2}}{y}, \tag{2.1}$$

it becomes a model of the *hyperbolic* or *Lobachevski plane* (see Exercise 2). 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

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

Using this fact and elementary geometric 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 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) 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

$$\langle \zeta_1, \zeta_2 \rangle = \frac{(\zeta_1, \zeta_2)}{\mathrm{Im(z)}^2}, \tag{2.2}$$

which is a scalar multiple of the Euclidean inner product $(\zeta_1, \zeta_2) = \xi_1\xi_2 + \eta_1\eta_2$. 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). Therefore the geometry in $\mathcal{H}$ is *non-Euclidean*. The metric in (2.1) 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 by

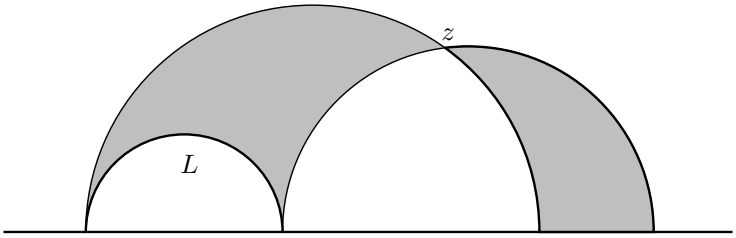


FIGURE 2.1. Geodesics in the upper half-plane

$$h(\gamma) = \int_0^1 \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{y(t)} dt. \tag{2.3}$$

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 2.1. *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) *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 3. □

Consider the group $SL(2, \mathbb{R})$ of real 2×2 matrices with determinant one. It acts on $\mathcal{H}$ by *Möbius transformations* if we assign to each $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$ the transformation

$$T_g(z) = \frac{az + b}{cz + d}. \tag{2.4}$$

PROPOSITION 2.2. *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

$$\mathrm{Im(w)} = \frac{\mathrm{w} - \overline{\mathrm{w}}}{2\mathrm{i}} = \frac{(\mathrm{ad} - \mathrm{bc})(\mathrm{z} - \overline{\mathrm{z}})}{2\mathrm{i}|\mathrm{cz} + \mathrm{d}|^2} = \frac{\mathrm{Im(z)}}{|\mathrm{cz} + \mathrm{d}|^2}. \tag{2.5}$$

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

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 (2.4) 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 (2.4). 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 2.3. *The group $PSL(2,\mathbb{R})$ acts on $\mathcal{H}$ by homeomorphisms.*

DEFINITION 2.4. 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 $\mathrm{Isom}(\mathcal{H})$.

THEOREM 2.5. *Möbius transformations are isometries, i.e., we have the inclusion $PSL(2,\mathbb{R}) \subset \mathrm{Isom}(\mathcal{H})$.*

PROOF. Let $T \in PSL(2,\mathbb{R})$. By Theorem 2.3 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

$$\frac{dw}{dz} = \frac{a(cz+d) - c(az+b)}{(cz+d)^2} = \frac{1}{(cz+d)^2}. \tag{2.6}$$

By (2.5) we have

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

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 follows from this immediately. □

3. Geodesics

THEOREM 3.1. *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 4, 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, and by Theorem 2.5 we conclude that the geodesic between z_1 and z_2 is the segment of L joining them. □

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 (2.3) imply the following statement.

COROLLARY 3.2. *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 3.3. *Any isometry of $\mathcal{H}$, and, in particular, any transformation from $PSL(2, \mathbb{R})$, maps geodesics into geodesics.*

PROOF. The same argument as in the Euclidean case using Corollary 3.2 works here as well. □

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 3.4. *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 4, let us choose a transformation $T \in PSL(2,\mathbb{R})$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-ratio 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 3.5. *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) $\cosh \rho(z,w) = 1 + \frac{|z-w|^2}{2\mathrm{Im(z)Im(w)}}$;

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

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

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

PROOF. We will prove that (e) holds. By Theorem 2.5, the left-hand side is invariant under any transformation $T \in PSL(2,\mathbb{R})$. By Exercise 5, 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. □

4. Isometries

We have seen that transformations in $PSL(2,\mathbb{R})$ are isometries of the hyperbolic plane $\mathcal{H}$ (Theorem 2.5). The next theorem identifies all isometries of $\mathcal{H}$ in terms of Möbius transformations and symmetry in the imaginary axis.

THEOREM 4.1. *The group* $\mathrm{Isom}(\mathcal{H})$ *is generated by the Möbius transformations from $PSL(2,\mathbb{R})$ 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 3.3, 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 4, 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

$$g \circ \varphi(z) = z \text{ or } -\overline{z}. \tag{4.1}$$

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 Exercise 6), only one of the equations (4.1) holds for all $z \in \mathcal{H}$. If $g \circ \varphi(z) = z$, then $\varphi(z)$ is a Möbius transformation of the form (2.4). If $g \circ \varphi(z) = -\overline{z}$, we have

$$\varphi(z) = \frac{a\overline{z} + b}{c\overline{z} + d} \text{ with } ad - bc = -1, \tag{4.2}$$

which proves the theorem. □

Thus we have characterized all the isometries of $\mathcal{H}$. The sign of the determinant of the corresponding matrix in (2.4) or (4.2) 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 (4.2) 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 $SL(2, \mathbb{R})$ into 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 $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$ with trace $t = a + d$ 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 action of the group $PSL(2, \mathbb{R})$ extends from $\mathcal{H}$ to its *Euclidean boundary* $\mathbb{R} \cup \{\infty\}$, hence $PSL(2, \mathbb{R})$ acts on the *Euclidean closure* of $\mathcal{H}$, denoted by $\widetilde{\mathcal{H}}$. 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

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

Notice that $\lambda_i = cw_i + d$ $(i = 1, 2)$ are the eigenvalues of the matrix A. A 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, namely, $w_i = x_i/y_i$. The derivative at the fixed point w_i can be written in terms of the eigenvalue λ_i as

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

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 $b/(1 - a)$.

Definition 4.2. A fixed point w of a transformation $f : \widetilde{\mathcal{H}} \to \widetilde{\mathcal{H}}$ 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. The geodesic in $\mathcal{H}$ connecting them is called the *axis* of T and is denoted by $C(T)$. By Theorem 3.3, T maps $C(T)$ onto itself, and $C(T)$ is the only geodesic with this property. Let λ be the eigenvalue of a matrix corresponding to T with $|\lambda| > 1$. Then the matrix of T is conjugate to the diagonal matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \frac{1}{\lambda} \end{pmatrix}$ that corresponds to the Möbius transformation

$$\Lambda(z) = \lambda^2 z, \tag{4.3}$$

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 4 and 9).

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 (4.3), the Euclidean rays in the upper half-plane issuing 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 7). Such rays have an important property: they are equidistant from the axis $C(\Lambda) = I$ (see Exercise 8), 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 4.1).

A useful notion in understanding how hyperbolic transformations act is that of an 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 4.1 and Exercise 10).

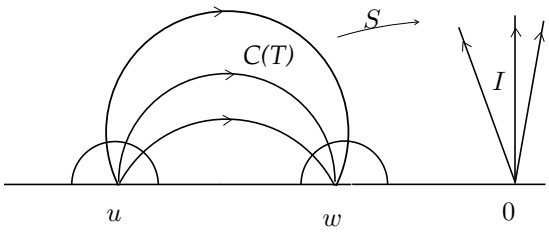


FIGURE 4.1. Hyperbolic transformations

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 $s \in \mathbb{R} \cup \{\infty\}$, i.e. "at infinity". 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 s (see Figure 4.2 and Exercise 12); we denote a horocycle through a point s at infinity by $\omega(s)$. Figure 4.2 illustrates a family of horocycles through a given point $s \in \mathbb{R}$ and $s = \infty$.

If $c \neq 0$, then the isometric circles for T and T^{-1} are tangent to each other (see Exercise 11). If $c = 0$, then there is no unique circle with the

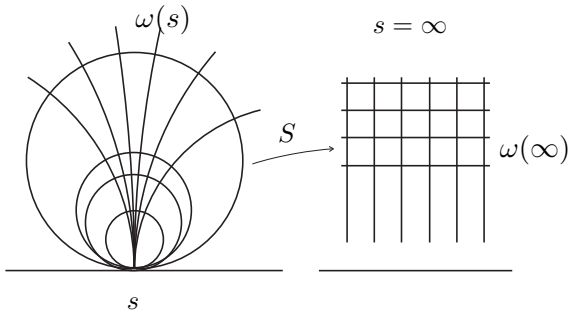


FIGURE 4.2. Parabolic transformations

isometric property: 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

$$f(z) = \frac{zi+1}{z+i} \tag{4.4}$$

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

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

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

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

Geodesics in the unit disc model are circular arcs and diameters orthogonal to the *principle circle* $\Sigma = \{z \in \mathbb{C} \mid |z| = 1\}$, the *Euclidean boundary* of $\mathcal{U}$. 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 13 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 vertical diameter.

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 14), 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 A. 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 $R(z) = -\overline{z}$, a 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 that intersects I at the 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

$$TRT^{-1} \tag{4.5}$$

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* (see Exercise 16).

DEFINITION 4.3. 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 of the circle Q. Inversion may be described by a geometric construction (see Exercise 15). 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 the definition 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

$$(z_1 - k)(\overline{z} - \overline{k}) = r^2. \tag{4.6}$$

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

$$z_1 = \frac{k\overline{z} + r^2 - |k|^2}{\overline{z} - \overline{k}}. \tag{4.7}$$

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

THEOREM 4.4. *Every isometry of $\mathcal{H}$ is a product of not more than three reflections in geodesics in $\mathcal{H}$.*

PROOF. By Theorem 4.1 it suffices to show that each transformation from the group $PSL(2,\mathbb{R})$ 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 11). 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 (4.6) 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. □

5. 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 5.1. *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 22. Since

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

we can write

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

□

COROLLARY 5.2. *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

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

this identity 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

$$\mu(A) = \int_A \frac{dxdy}{y^2}, \tag{5.1}$$

provided this integral exists.

THEOREM 5.3. *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. It follows immediately from the preservation of Riemannian metric (Theorem 5.1). Here is a direct calculation as well. When we performed the change of variables $w = T(z)$ in the line integral of Theorem 2.5, 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

$$\frac{\partial(u,v)}{\partial(x,y)} := \det \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \Big(\frac{\partial u}{\partial x}\Big)^2 + \Big(\frac{\partial v}{\partial x}\Big)^2 = |T'(z)|^2 = \frac{1}{|cz+d|^4}. \tag{5.2}$$

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 four types of hyperbolic triangles, depending on whether $0, 1, 2$, or 3 vertices belong to $\mathbb{R} \cup \{\infty\}$ (see Figure 5.1).

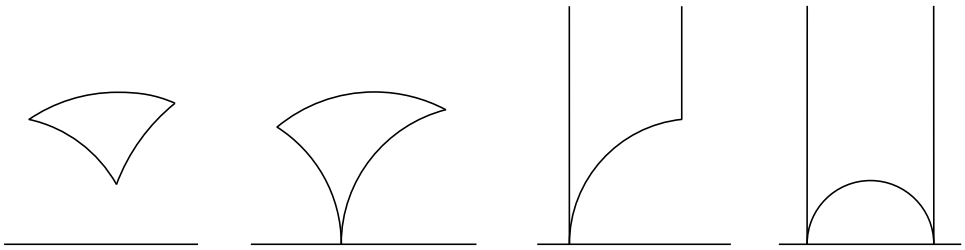


FIGURE 5.1. Hyperbolic triangles

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

THEOREM 5.4 (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 5.2), and prove the formula in this case.

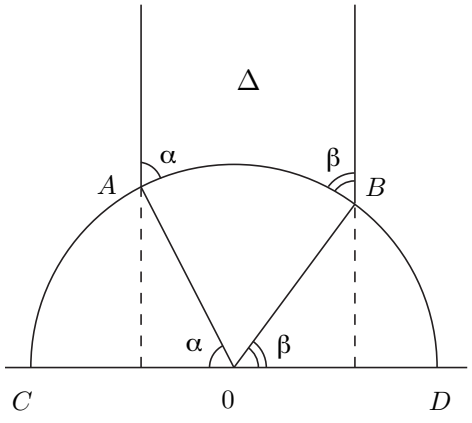


FIGURE 5.2. Proof of the Gauss-Bonnet formula

The angle at infinity is equal to 0, and let us assume that the other two angles are equal to α and β. Since the angle measure in the hyperbolic plane coincides with the Euclidean angle measure, the angles A0C and B0D are

equal to α and β, respectively, as angles with mutually perpendicular sides. Assume the vertical geodesics are the lines $x = a$ and $x = b$. Then

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

The substitution $x = \cos\theta$ $(0 \le \theta \le \pi)$ 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 5.3.

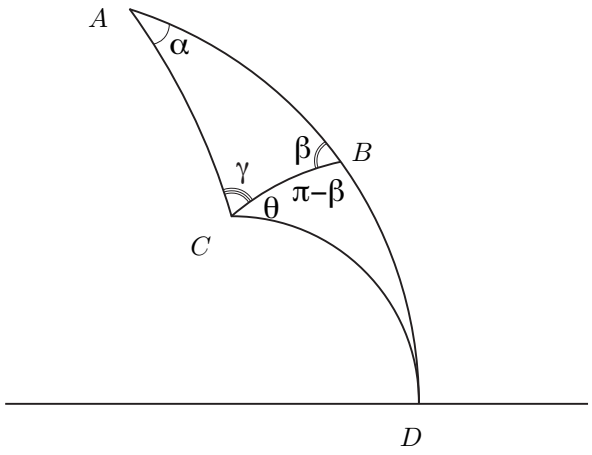


FIGURE 5.3. A general case in the Gauss-Bonnet formula

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 5.4 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 5.5. *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 14), 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 5.4. Therefore C must belong to the side $A'C'$, and hence the sides BC and $B'C'$ intersect at a point X (see Figure 5.4); 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). Then the sum of the angles of the triangle $B'XB$ is at least π, in contradiction with Theorem 5.4.

□

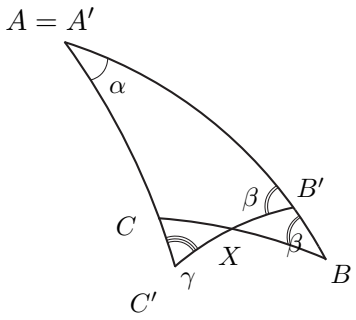


FIGURE 5.4. There are no similar triangles in hyperbolic geometry

6. Hyperbolic trigonometry

Let us consider a general hyperbolic triangle with sides of hyperbolic length a, b, c and opposite angles α, β, γ. We assume that α, β, and γ are positive (so a, b, and c are finite) and prove the following results.

THEOREM 6.1. *(i) The Sine Rule:* $\frac{\sinh\, a}{\sin \alpha} = \frac{\sinh\, b}{\sin \beta} = \frac{\sinh\, c}{\sin \gamma}$.

(ii) The Cosine Rule I: $\cosh\, c = \cosh\, a \cosh\, b - \sinh\, a \sinh\, b \cos \gamma$.

(iii) The Cosine Rule II: $\cosh\ c = \frac{\cos\alpha\cos\beta+\cos\gamma}{\sin\alpha\sin\beta}$.

REMARK. Note that Cosine Rule II implies that if two triangles have the same angles, then their sides are also equal, and therefore it has no analogue in Euclidean geometry. It also gives an alternative proof of Theorem 5.5.

PROOF OF (ii). Let us denote the vertices opposite the sides a, b, c by v_a, v_b, v_c respectively. We shall use the model $\mathcal{U}$ and may assume that $v_c = 0$ and $\operatorname{Im} v_a = 0$, $\operatorname{Re}\ v_a > 0$ (see Figure 6.1). By Exercise 20(iv) we have

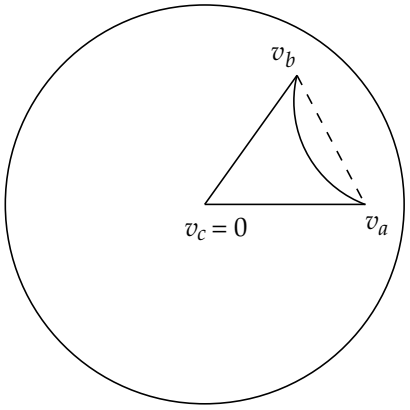


FIGURE 6.1. The Cosine Rule I

$$v_a = \tanh\frac{1}{2}\rho(0, v_a) = \tanh(\frac{1}{2}b), \tag{6.1}$$

and similarly,

$$v_b = e^{i\gamma}\tanh(\frac{1}{2}a), \tag{6.2}$$

We have $c = \rho(v_a, v_b)$, and from Exercise 20(iii)

$$\cosh\ c = \sinh^2[\frac{1}{2}\rho(v_a, v_b)] + 1 = \frac{2|v_a - v_b|^2}{(1-|v_a|^2)(1-|v_b|^2)} + 1. \tag{6.3}$$

The right-hand side of expression (6.3) is equal to $\cosh a\cosh b - \sinh a\sinh b\cos\gamma$ by Exercise 23, and hence (ii) follows. □

PROOF OF (i). Using (ii) we obtain

$$\left(\frac{\sinh c}{\sin\gamma}\right)^2 = \frac{\sinh^2 c}{1-\left(\frac{\cosh a\cosh b-\cosh c}{\sinh a\sinh b}\right)^2}. \tag{6.4}$$

The Sine Rule will be valid if we prove that the expression on the right-hand side of (6.4) is symmetric in a, b, and c. This follows from the symmetry of

$$(\sinh a\sinh b)^2 - (\cosh a\cosh b - \cosh c)^2$$

which is obtained by a direct calculation. □

PROOF OF (iii). Let us write A for $\cosh a$, B for $\cosh b$, and C for $\cosh c$. The Cosine Rule I yields

$$\cos\gamma = \frac{(AB - C)}{(A^2-1)^{\frac{1}{2}}(B^2-1)^{\frac{1}{2}}}$$

and so

$$\sin^2\gamma = \frac{D}{(A^2-1)(B^2-1)}$$

where $D = 1 + 2ABC - (A^2 + B^2 + C^2)$ is symmetric in A, B, and C. The expression for $\sin^2\gamma$ shows that $D \geq 0$. Using analogous expressions for $\cos\alpha$, $\sin\alpha$, $\cos\beta$, and $\sin\beta$ we observe that if we multiply both the numerator and denominator of

$$\frac{\cos\alpha\ \cos\beta + \cos\gamma}{\sin\alpha\ \sin\beta}$$

by the positive value of

$$(A^2-1)^{\frac{1}{2}}(B^2-1)^{\frac{1}{2}}(C^2-1)^{\frac{1}{2}}$$

we obtain

$$\frac{\cos\alpha\cos\beta + \cos\gamma}{\sin\alpha\sin\beta} = \frac{[(BC-A)(CA-B) + (AB-C)(C^2-1)]}{D} = C$$

□

THEOREM 6.2. (Pythagorian Theorem) If $\gamma = \frac{\pi}{2}$ we have $\cosh c = \cosh a \cosh b$.

PROOF. Immediate from the Cosine Rule I. □

Exercises

1. Prove that the metric in the Poincaré disc model is given by

$$ds^2 = \frac{4(d\eta_1^2 + d\eta_2^2)}{(1-(\eta_1^2+\eta_2^2))^2}.$$

2. Prove that the metric in the upper half-plane model is given by

$$ds^2 = \frac{d\eta_1^2 + d\eta_3^2}{\eta_3^2}.$$

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

4. 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.

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

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

6. Prove that isometies are continuous maps.

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

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

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

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

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

10. 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 $I(T^{-1})$.

11. 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.

12. 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.

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

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

14. 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 $T(z_1) = z_2$.

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

16. Prove that the transformation (4.5) is an inversion in the circle corresponding to the geodesic L.

17. Prove that any orientation-preserving isometry T of the unit disc $\mathcal{U}$ is an inversion in $I(T)$ followed by a reflection in the straight line L, the Euclidean bisector between the centers of the isometric circles $I(T)$ and $I(T^{-1})$.

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

19. 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 $AB \neq BA$.

20. Use the map f (4.4) to derive the formulae for the hyperbolic distance in the unit disc model similar to those in Theorem 3.5, for $z, w \in \mathcal{U}$:

(i) $\rho(z,w) \in \ln \frac{|1-z\bar{w}|+|z-w|}{|1-z\bar{w}|-|z-w|}$,

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

(iii) $\sinh^2[\frac{1}{2}\rho(z,w)] = \frac{|z-w|^2}{(1-|z|^2)(1-|w|^2)}$,

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

21. Justify the calculations in (5.2) 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 classical *Cauchy-Riemann equations*) and

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

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

22. 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

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} \mapsto \xi + i\eta = \zeta,$$

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.

23. Show that the right-hand side of expression (6.3) is equal to $\cosh a \cosh b - \sinh a \sinh b \cos\gamma$.

Lecture II. Fuchsian Groups and Their Fundamental Regions

7. The group $PSL(2,\mathbb{R})$

Let $S\mathcal{H}$ be the unit tangent bundle of the upper half-plane $\mathcal{H}$. It is homeomorphic to $\mathcal{H} \times S^1$. Let us parametrize it by local coordinates (z,ζ), where $z \in \mathcal{H}$, $\zeta \in \mathbb{C}$ with $|\zeta| = \mathrm{Im}(z)$. (Notice that with this parametrization, $\|\zeta\| = 1$ (see (2.2)), so that ζ is a unit tangent vector.) The group $PSL(2,\mathbb{R})$

acts on $S\mathcal{H}$ by the differentials: for $T : z \to \frac{az+b}{cz+d}$, $T(z,\zeta) = (T(z),\, DT(\zeta))$, where

$$DT(\zeta) = \frac{1}{(cz+d)^2}\,\zeta. \tag{7.1}$$

As any group, $PSL(2,\mathbb{R})$ acts on itself by left multiplication. The next result connects these two actions.

THEOREM 7.1. *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 multiplication corresponds to the action of $PSL(2,\mathbb{R})$ on $S\mathcal{H}$ induced by its action on $\mathcal{H}$ by fractional linear transformations.*

PROOF. Let (i,ι) be a fixed element of $S\mathcal{H}$, where ι 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 $S\mathcal{H}$. There exists a unique $T \in PSL(2,\mathbb{R})$ sending the imaginary axis to the geodesic passing through z and tangent to ζ (Exercise 4) so that $T(i) = z$. By (7.1) we have $DT(\iota) = \zeta$, and hence

$$T(i,\iota) = (z,\zeta). \tag{7.2}$$

It is easy to see that the map $(z,\zeta) \to T$ is a homeomorphism between $S\mathcal{H}$ and $PSL(2,\mathbb{R})$.

For $S \in PSL(2,\mathbb{R})$, suppose that $S(z,\zeta) = (z',\zeta')$. By (7.2) $S(z,\zeta) = ST(i,\iota)$, and hence $S(z,\zeta) \to ST$, and the last assertion follows. □

Let $d\ell = \sqrt{ds^2 + d\theta^2}$ be a Riemannian metric on $S\mathcal{H}$, where ds is the hyperbolic metric on $\mathcal{H}$ (2.1), and $\theta = \frac{1}{2\pi}\arg(\zeta)$; and let $dv = d\mu d\theta$ be a volume on $S\mathcal{H}$, where $d\mu$ is the hyperbolic area on $\mathcal{H}$ (5.1).

PROPOSITION 7.2. *The metric $d\ell$ and the volume dv on $S\mathcal{H}$ are $PSL(2,\mathbb{R})$-invariant.*

PROOF. This can be seen by a direct calculation. Let $f(z) = \frac{az+b}{cz+d} \in PSL(2,\mathbb{R})$. In local coordinates $(z,\zeta) \mapsto (f(z),(Df)(\zeta)) = (z',\zeta')$. The metric $d\ell$ on $S\mathcal{H}$ is a norm in the tangent space to $S\mathcal{H}$:

$$\|(dz,d\zeta)\|^2 = \frac{|dz|^2}{y^2} + (d\phi)^2.$$

Since each summand is invariant:

$$\frac{|dz'|^2}{(\mathrm{Imf(z)})^2} = \frac{|f'(z)|^2|dz|^2}{(\mathrm{Imf(z)})^2} = \frac{|dz|^2}{y^2} \text{ and } (d\phi')^2 = (d\phi)^2,$$

the invariance of the Riemannian metric $d\ell$ follows. The invariance of the volume dv follows from the invariance of the metric. □

Thus, besides being a group, $PSL(2,\mathbb{R})$ is also a topological space. Convergence in $PSL(2,\mathbb{R})$ can be expressed in the matrix language.
If $g_n \to g$ in $PSL(2,\mathbb{R})$, this means that there exist matrices $A_n \in SL(2,\mathbb{R})$ representing g_n such that $\lim_{n\to\infty} \|A_n - A\| = 0$, where $\|\cdot\|$ is a norm on $SL(2,\mathbb{R})$ induced from $\mathbb{R}^4$.

DEFINITION 7.3. A subgroup Γ of Isom$(\mathcal{H})$ is called *discrete* if the induced topology on Γ is a discrete topology, i.e. if Γ is a discrete set in the topological space Isom$(\mathcal{H})$.

It follows that Γ is discrete if and only if $T_n \to Id$, $T_n \in \Gamma$ implies $T_n = Id$ for sufficiently large n.

8. Discrete and properly discontinuous groups

DEFINITION 8.1. A discrete subgroup of Isom $(\mathcal{H})$ is called a *Fuchsian group* if it consists of orientation-preserving transformations, in other words, a Fuchsian group is a discrete subgroup of $PSL(2,\mathbb{R})$.

For any discrete group Γ of Isom $(\mathcal{H})$, its subgroup Γ^+ of index ≤ 2 consisting of orientation-preserving transformations is a Fuchsian group. Thus the main ingredient in the study of discrete subgroups of isometries of $\mathcal{H}$ is the study of Fuchsian groups. The action of $PSL(2,\mathbb{R})$ on $\mathcal{H}$ lifts to the action on its unit tangent bundle $S\mathcal{H}$ by isometries (Proposition 7.2), thus sometimes it is useful to consider Fuchsian groups as discrete groups of isometries of $S\mathcal{H}$. Certain discrete subgroups of Lie groups are called *lattices* by analogy with lattices in $\mathbb{R}^n$ that are discrete groups of isometries of $\mathbb{R}^n$. The latter have the following important property: their action on $\mathbb{R}^n$ is *discontinuous* in the sense that every point of $\mathbb{R}^n$ has a neighborhood which is carried outside itself by all elements of the lattice except for the identity. In general, discrete groups of isometries do not have such discontinuous behavior, for if some elements have fixed points these points cannot have such a neighborhood. However, they satisfy a slightly weaker discontinuity condition. First we need several definitions.

Let X be a locally compact metric space, and let G be a group of isometries of X.

DEFINITION 8.2. A family $\{M_\alpha \mid \alpha \in A\}$ of subsets of X indexed by elements of a set A is called *locally finite* if for any compact subset $K \subset X$, $M_\alpha \bigcap K \neq \emptyset$ for only finitely many $\alpha \in A$.

REMARK. Some of the subsets M_α may coincide but they are still considered different elements of the family.

DEFINITION 8.3. For $x \in X$, a family $Gx = \{g(x) \mid g \in G\}$ is called the G-orbit of the point x. Each point of Gx is contained with a multiplicity equal to the order of G_x, the *stabilizer of x in G*.

DEFINITION 8.4. We say that a group G acts properly discontinuously on X if the G-orbit of any point $x \in X$ is locally finite.

Since X is locally compact, a group G acts properly discontinuously on X if and only if each orbit has no accumulation point in X, and the order of the stabilizer of each point is finite. The first condition, however, is equivalent to the fact that each orbit of G is discrete. For, if $g_n(x) \to s \in X$, then for any $\varepsilon > 0$, $\rho(g_n(x), g_{n+1}(x)) < \varepsilon$ for sufficiently large n, but since g_n is an isometry, we have $\rho(g_n^{-1}g_{n+1}(x), x) < \varepsilon$, which implies that x is an accumulation point for its orbit Gx, i.e. Gx is not discrete. In fact, the discreteness of all orbits already implies the discreteness of the group (see Corollary 8.7 for subgroups of $PSL(2,\mathbb{R})$).

EXAMPLE B. Let us consider a group consisting of all transformations

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

It is called the *modular group* and denoted by $PSL(2,\mathbb{Z}) \approx SL(2,\mathbb{Z})/\{\pm 1_2\}$.

It is clearly a discrete subgroup of $PSL(2,\mathbb{R})$ and hence a Fuchsian group.

Our next task is to show that a $\Gamma \subset PSL(2,\mathbb{R})$ is a Fuchsian group if and only if it acts properly discontinuously on $\mathcal{H}$.

LEMMA 8.5. *Let $z_0 \in \mathcal{H}$ be given and let K be a compact subset of $\mathcal{H}$. Then the set*

$$E = \{T \in PSL(2,\mathbb{R}) \mid T(z_0) \in K\}$$

is compact.

PROOF. $PSL(2,\mathbb{R})$ is topologized as a quotient space of $SL(2,\mathbb{R})$. Thus we have a continuous map $\psi : SL(2,\mathbb{R}) \to PSL(2,\mathbb{R})$ defined by

$$\psi\begin{bmatrix} a & b \\ c & d \end{bmatrix} = T, \text{ where } T(z) = \frac{az+b}{cz+d}.$$

If we show that

$$E_1 = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL(2,\mathbb{R}) \mid \frac{az_0+b}{cz_0+d} \in K \right\}$$

is compact then it follows that $E = \psi(E_1)$ is compact. We prove that E_1 is compact by showing it is closed and bounded when regarded as a subset of $\mathbb{R}^4$ (identifying $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with (a,b,c,d)). We have a continuous map $\beta : SL(2,\mathbb{R}) \to \mathcal{H}$ defined by $\beta(A) = \psi(A)(z_0)$. $E_1 = \beta^{-1}(K)$, thus it follows that E_1 is closed as the inverse image of the closed set K.

We now show that E_1 is bounded. As K is bounded there exists $M_1 > 0$ such that

$$\left|\frac{az_0+b}{cz_0+d}\right| < M_1,$$

for all $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in E_1$.

Also, as K is compact in $\mathcal{H}$, there exists $M_2 > 0$ such that

$$\mathrm{Im}\left(\frac{az_0+b}{cz_0+d}\right) \geq M_2.$$

(2.5) implies that the left-hand side of this inequality is $\mathrm{Im}(z_0)/|cz_0+d|^2$ so that

$$|cz_0+d| \leq \sqrt{\left(\frac{\mathrm{Im}(z_0)}{M_2}\right)},$$

and thus

$$|az_0+b| \leq M_1\sqrt{\left(\frac{\mathrm{Im}(z_0)}{M_2}\right)},$$

and we deduce that a, b, c, d are bounded. □

THEOREM 8.6. *Let Γ be a subgroup of $PSL(2,\mathbb{R})$. Then Γ is a Fuchsian group if and only if Γ acts properly discontinuously on $\mathcal{H}$.*

PROOF. We first show that a Fuchsian group acts properly discontinuously on $\mathcal{H}$. Let $z \in \mathcal{H}$ and K be a compact subset of $\mathcal{H}$. We use Lemma 8.5 to see that $\{T \in \Gamma \mid T(z) \in K\} = \{T \in PSL(2,\mathbb{R}) \mid T(z) \in K\} \bigcap \Gamma$ is a finite set (it is the intersection of a compact and a discrete set), and hence Γ acts properly discontinuously. Conversely, suppose Γ acts properly discontinuously, but it is not a discrete subgroup of $PSL(2,\mathbb{R})$. Then there exists a sequence $\{T_k\}$ of distinct elements of Γ such that $T_k \to \mathrm{Id}$ as $k \to \infty$. Let $s \in \mathcal{H}$ be a point not fixed by any of T_k. Then $\{T_k(s)\}$ is a sequence of points distinct from s and $T_k(s) \to s$ as $k \to \infty$. Hence every closed hyperbolic disc centered at s contains infinitely many points of the Γ-orbit of s, i.e. Γ does not act properly discontinuously, a contradiction. □

COROLLARY 8.7. *Let Γ be a subgroup of $PSL(2,\mathbb{R})$. Then Γ acts properly discontinuously on $\mathcal{H}$ if and only if for all $z \in \mathcal{H}$, Γz, the Γ-orbit of z, is a discrete subset of $\mathcal{H}$.*

PROOF. Suppose Γ acts properly discontinuously on $\mathcal{H}$, hence each Γ-orbit is a locally finite family of points, hence a discrete set of $\mathcal{H}$. Conversely, suppose Γ does not act properly discontinuously on $\mathcal{H}$ and hence by Theorem 8.6 is not discrete. Repeating the argument in the proof of Theorem 8.6, we construct a sequence $\{T_k(s)\}$ of points distinct from s such that $T_k(s) \to s$, hence the Γ-orbit of the point s is not discrete. □

Corollary 8.7 implies the following: if $z \in \mathcal{H}$ and $\{T_n\}$ is a sequence of distinct elements in Γ such that $\{T_n(z)\}$ has a limit point $\alpha \in \mathbb{C} \cup \{\infty\}$, then $\alpha \in \mathbb{R} \cup \{\infty\}$.

9. Definition of a fundamental region

We are going to be concerned with fundamental regions of mainly Fuchsian groups, however it is convenient to give a definition in a slightly more general situation. As in §8, let X be a locally compact metric space, and Γ be a group of isometries acting properly discontinuously on X.

DEFINITION 9.1. A closed region $F \subset X$ (i.e. a closure of a non-empty open set $\overset{\circ}{F}$, called the interior of F) is defined to be a *fundamental region* for Γ if

(i) $\bigcup_{T\in\Gamma} T(F) = X$,

(ii) $\overset{\circ}{F} \cap T(\overset{\circ}{F}) = \emptyset$ for all $T \in \Gamma \setminus \{\mathrm{Id}\}$.

The set $\partial F = F \setminus \overset{\circ}{F}$ is called the *boundary* of F. The family $\{T(F) \mid T \in \Gamma\}$ is called the *tessellation* of X.

We shall prove in §10 that any Fuchsian group possesses a nice (connected and convex) fundamental region. Now we give an example in the simplest situation.

EXAMPLE C. Let Γ be the cyclic group generated by the transformation $z \to 2z$. Then the semi-annulus shown in Figure 9.1(a) is easily seen to be a fundamental region for Γ. It is already clear from this example that a fundamental region is not uniquely determined by the group: an arbitrary small perturbation of the lower semicircle determines a perturbation of the upper semicircle, and gives yet another fundamental region shown in Figure 9.1(b).

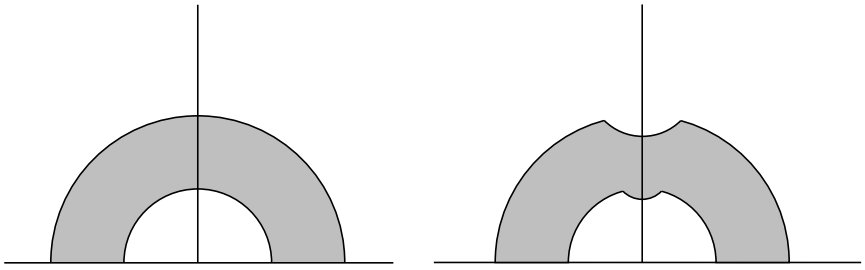


FIGURE 9.1. Fundamental domains for Example C

THEOREM 9.2. *Let F_1 and F_2 be two fundamental regions for a Fuchsian group Γ, and $\mu(F_1) < \infty$. Suppose that the boundaries of F_1 and F_2 have zero hyperbolic area. Then $\mu(F_2) = \mu(F_1)$.*

PROOF. We have $\mu(\overset{\circ}{F}_i) = \mu(F_i), i = 1, 2$. Now

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

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,

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

Since F_1 is a fundamental region

$$\bigcup_{T\in\Gamma} T(F_1) = \mathcal{H},$$

and therefore

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

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 $\mu(F_2) \geq \mu(F_1)$. Hence $\mu(F_2) = \mu(F_1)$. □

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. An example of a Fuchsian group with a fundamental region of infinite area is the group generated by $z \to z + 1$ (see also Example C above). Obviously, a compact fundamental region has finite area. Non-compact regions also may have finite area For example, for $\Gamma = PSL(2,\mathbb{Z})$ the fundamental region, which will be described in §10 (Example B), is a hyperbolic triangle with angles $\frac{\pi}{3}, \frac{\pi}{3}, 0$. By the Gauss-Bonnet formula (Theorem 5.4) its area is finite and is equal to $\pi - \frac{2\pi}{3} = \frac{\pi}{3}$.

THEOREM 9.3. *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 exists $w \in F$ and $T \in \Gamma$ such that $z = T(w)$. We have $T = ST_i$ for some $S \in \Lambda$ and some $i, 1 \le i \le n$. Therefore

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

Since $T_i(w) \in F_1, z$ is in the Λ-orbit of some point of F_1. Hence the union of the Λ-images of F_1 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 need to prove that $S = \mathrm{Id}$. Let $\varepsilon > 0$ be so small that $B_\varepsilon(z)$ (the open hyperbolic disc of radius ε centered at z) is contained in $\overset{\circ}{F}_1$. Then $B_\varepsilon(z)$ has a non-empty intersection with exactly k of the images of $\overset{\circ}{F}$ under $T_1, \cdots, T_n$, where $1 \le k \le n$. Suppose these images are $T_{i_1}(\overset{\circ}{F}), \cdots, T_{i_k}(\overset{\circ}{F})$. Let $B_\varepsilon(S(z)) = S(B_\varepsilon(z))$ have a non-empty intersection with $T_j(\overset{\circ}{F})$ say, $1 \le j \le n$. It follows that $B_\varepsilon(z)$ has a non-empty intersection with $S^{-1}T_j(\overset{\circ}{F})$ so that $S^{-1}T_j = T_{i_\ell}$ where $1 \le \ell \le k$. Hence

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

so that $T_j = T_{i_\ell}$ and $S = \mathrm{Id}$. Hence $\overset{\circ}{F}_1$ contains precisely one point of each Λ-orbit. □

PROOF OF (ii). This follows immediately, as $\mu(T(F)) = \mu(F)$ for all $T \in PSL(2, \mathbb{R})$, and $\mu(T_i(F) \cap T_j(F)) = 0$ for $i \ne j$. □

10. The Dirichlet region

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

$$D_p(\Gamma) = \{z \in \mathcal{H} \mid \rho(z, p) \le \rho(z, T(p)) \text{ for all } T \in \Gamma\}. \tag{10.1}$$

By the invariance of the hyperbolic metric under $PSL(2, \mathbb{R})$ this region can also be defined as

$$D_p(\Gamma) = \{z \in \mathcal{H} \mid \rho(z, p) \le \rho(T(z), p) \text{ for all } T \in \Gamma\}. \tag{10.2}$$

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

$$\{z \in \mathcal{H} \mid \rho(z, p) \le \rho(z, T_1(p))\} \tag{10.3}$$

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 as the Γ-orbit of p is discrete (Corollary 8.7), $D_p(\Gamma)$ contains a neighborhood of p. In order to describe the set (10.3) we join the points p and $T_1(p)$ by a geodesic segment and construct a line given by the equation

$$\rho(z, p) = \rho(z, T_1(p)).$$

DEFINITION 10.1. A *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 $[z_1, z_2]$.

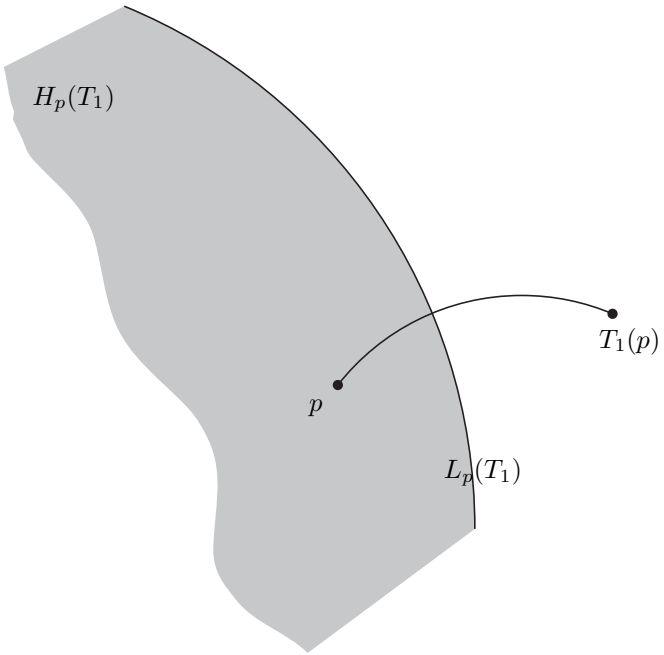


FIGURE 10.1. Construction of the Dirichlet region

LEMMA 10.2. *A line given by the equation*

$$\rho(z, z_1) = \rho(z, z_2) \tag{10.4}$$

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

PROOF. We may assume that $z_1 = i, z_2 = ir^2$ with $r > 0$: thus $w = ir$ and the perpendicular bisector is given by the equation $|z| = r$. On the other hand, by Theorem 3.5(b) (10.4) is equivalent to

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

which simplifies to $|z| = r$. □

We shall 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 (10.3) by $H_p(T_1)$ (see Figure 10.1). 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 10.3. *If p is not fixed by any element of $\Gamma \setminus \{\mathrm{Id}\}$, 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 $z_0 \in \Gamma z$ with the smallest $\rho(z_0, p)$. Then $\rho(z_0, p) \leq \rho(T(z_0), p)$ for all $T \in \Gamma$, and by (10.2) $z_0 \in D_p(\Gamma)$. 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 $D_p(\Gamma)$, they cannot lie in the same Γ-orbit. If $\rho(z, p) = \rho(T(z), p)$ for some $T \in \Gamma \setminus \{\mathrm{Id}\}$, 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 $D_p(\Gamma)$, $\rho(z, p) < \rho(T(z), p)$ for all $T \in \Gamma \setminus \{\mathrm{Id}\}$. If two points z_1, z_2 lie in the same Γ-orbit, this implies $\rho(z_1, p) < (z_2, p)$ and $\rho(z_2, p) < (z_1, p)$, a contradiction. Thus the interior of $D_p(\Gamma)$ contains at most one point of each Γ-orbit. Being an intersection of closed half-planes, $D_p(\Gamma)$ is closed and convex. Thus $D_p(\Gamma)$ is path-connected, hence connected. □

EXAMPLE B. $\Gamma = PSL(2, \mathbb{Z})$. It is easily verified that ki $(k > 1)$ is not fixed by any non-identity element of the modular group, so choose $p = ki$, where $k > 1$. We shall show that the region

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

illustrated in Figure 10.2 is the Dirichlet region for Γ centered at p.

First, the isometries $T(z) = z + 1$, $S(z) = -1/z$ are in Γ; and, as can be

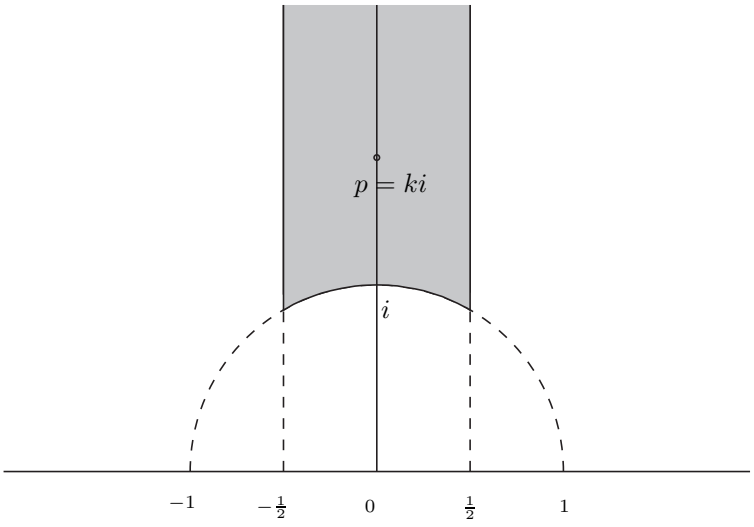


FIGURE 10.2. A Dirichlet region for $PSL(2, \mathbb{Z})$

easily verified, the three geodesic sides of F are $L_p(T), L_p(T^{-1})$ and $L_p(S)$.

This shows that $D_p(\Gamma) \subset F$. If $D_p(\Gamma) \neq F$, there exists $z \in \overset{\circ}{F}$ and $h \in \Gamma$ such that $h(z) \in \overset{\circ}{F}$. We shall now show that this cannot happen. Suppose that

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

Then

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

since $|z| > 1$ and $\mathrm{Re}(z) > -\frac{1}{2}$. This lower bound is an integer: it is non-negative and is not zero (this would be possible only if $c = d = 0$, which contradicts $ad - bc = 1$). Therefore it is at least 1 and $|cz+d| > 1$. Hence

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

Exactly the same argument holds with z, h replaced by $h(z)$, h^{-1}, and a contradiction is reached: thus $D_p(\Gamma) = F$.

In the rest of this section, Γ will be a discrete group of orientation-preserving isometries of the unit disc $\mathcal{U}$ (sometimes also referred to as a Fuchsian group). We assume that 0 is not an elliptic fixed point, i.e. that $c \neq 0$ for all $T(z) = \frac{az+\bar{c}}{cz+\bar{a}}$ in the group Γ. We define

$$R_0 = \overline{\bigcap_{T\in\Gamma} \widehat{I}(T) \bigcap \mathcal{U}},$$

the closure of the set of points in $\mathcal{U}$ which are exterior to the isometric circles of all transformations in the group Γ. We shall prove that R_0 is a fundamental region for Γ, called the *Ford fundamental region.*

THEOREM 10.4. *R_0 is a fundamental region for Γ.*

PROOF. We shall prove that R_0 is a Dirichlet region $D_0(\Gamma)$, and the theorem will follow from Theorem 10.3. The perpendicular bisector I of the geodesic segment $[T^{-1}(0), 0]$ is a geodesic in the unit disc model, hence the arc of an Euclidean circle orthogonal to the circle at infinity (see Figure 10.3). Since both geodesic segments $[T^{-1}(0), 0]$ and $[0, T(0)]$ are segments of the radii of the unit disc, $T(I)$ is the perpendicular bisector of $[0, T(0)]$ which is the arc of an Euclidean circle of the same radius. Thus the transformation T maps I to $T(I)$ without alteration of Euclidean lengths, and therefore the perpendicular bisector of $[0, T(0)]$ is the isometric circle $I(T^{-1})$. □

THEOREM 10.5. *Given any infinite sequence of distinct isometric circles $I_1, I_2, \cdots$ of transformations of a Fuchsian group Γ with radii $r_1, r_2, \cdots$ we have $\lim_{n\to\infty} r_n = 0$.*

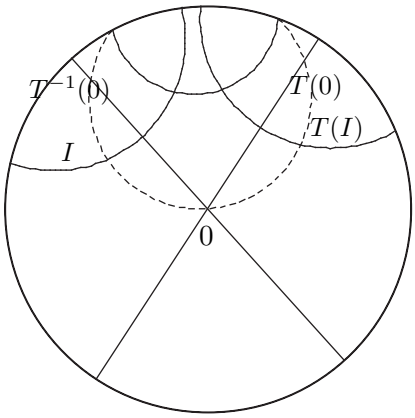


FIGURE 10.3. Ford region is Dirichlet region

PROOF. The transformations are of the form

$$T(z) = \frac{az + \bar{c}}{cz + \bar{a}} \ (a, c \in \mathbb{C}, |a|^2 - |c|^2 = 1). \tag{10.5}$$

Recall that the radius of $I(T)$ is equal to $\frac{1}{|c|}$. Let $\varepsilon > 0$ be given. There are only finitely many $T \in \Gamma$ with $|c| < 1/\varepsilon$. 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. □

11. Structure of a Dirichlet region

Dirichlet regions for Fuchsian groups can be quite complicated. They are bounded by geodesics in $\mathcal{H}$ and possibly by segments of the real axis. If two such geodesics intersect in $\mathcal{H}$, their point of intersection is called a *vertex* of the Dirichlet region. It will be shown that the vertices are isolated (see Proposition 11.3 below) so that a Dirichlet region is bounded by a union of (possibly, infinitely many) geodesics and possibly segments of the real axis (see Figure 11.1 for the unit disc model). We shall be interested in the

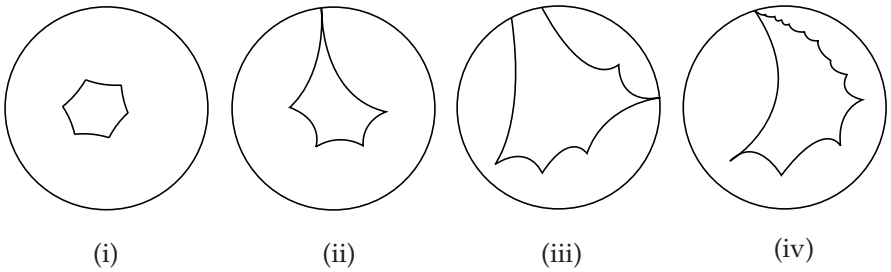


FIGURE 11.1. Dirichlet regions in the unit disc model

tessellation of $\mathcal{H}$ formed by a Dirichlet region F and all its images under

Γ (called *faces*): $\{T(F) \mid T \in \Gamma\}$. This tesselation will be referred to as a *Dirichlet tessellation*. (See Figure 11.2 for a Dirichlet tessellation for the modular group.) The next theorem shows that the Dirichlet tessellation has nice local properties.

DEFINITION 11.1. A fundamental region F for a Fuchsian group Γ is called *locally finite* if the tessellation $\{T(F) \mid T \in \Gamma\}$ is locally finite (see the definition of locally finite family of subsets in §8).

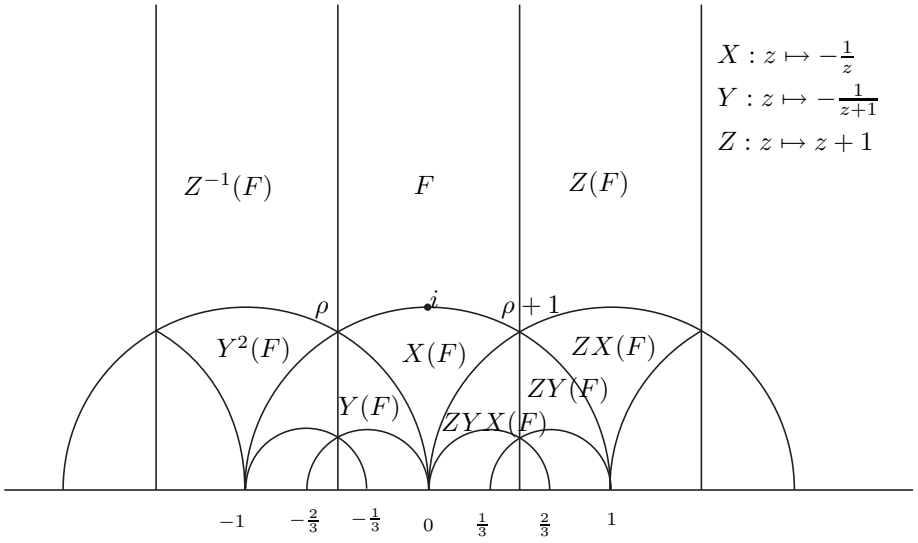


FIGURE 11.2. Dirichlet tessellation for the modular group

THEOREM 11.2. *A Dirichlet region is locally finite.*

PROOF. Let $F = D_p(\Gamma)$, where p is not fixed by any element of $\Gamma \setminus \{\mathrm{Id}\}$. Let $a \in F$, and let $K \subset \mathcal{H}$ be a compact neighborhood of a. Suppose that $K \bigcap T_i(F) \neq \emptyset$ for some infinite sequence $T_1, T_2, \cdots$ of distinct elements of Γ. Let $\sigma = \underset{z \in K}{sup} \rho(p, z)$. Since $\rho(p, z) \leq \rho(p, a) + \rho(a, z)$, for all $z \in K$, and K is bounded, σ is finite. Let $w_j \in K \bigcap T_j(F)$. Then $w_j = T_j(z_j)$ for $z_j \in F$, and by the triangle inequality,

$$\begin{aligned} \rho(p, T_j(p)) &\leq \rho(p, w_j) + \rho(w_j, T_j(p)) \\ &= \rho(p, w_j) + \rho(z_j, p) \\ &\leq \rho(p, w_j) + \rho(w_j, p) \quad (\text{as } z_j \in D_p(\Gamma)) \\ &\leq 2\sigma \end{aligned}$$

Thus the infinite set of points $T_1(p), T_2(p), \cdots$ belongs to a compact hyperbolic ball with center p and radius 2σ, but this contradicts the properly discontinuous action of Γ. □

PROPOSITION 11.3. *The vertices of a Dirichlet region are* isolated, *that is every vertex of F has a neighborhood containing no other vertices of F.*

PROOF. If x lies on the side of the Dirichlet region $F = D_p(\Gamma)$, then there exists $T_x \in \Gamma$ such that $\rho(p, x) = \rho(T_x p, x)$, hence $\rho(p, x) = \rho(p, T_x^{-1} x)$. Now assume that a vertex $v \in \mathcal{H}$ is not isolated, i.e. there is a sequence of vertices $v_i \in F$ such that $v_i \to v$. According to the above remark, choose T_i such that $\rho(p, v_i) = \rho(p, T_i v_i)$. We have

$$\begin{aligned}\rho(v, T_i v_i) &\le \rho(v, v_i) + \rho(v_i, T_i v_i) \le \rho(v, v_i) \\ &\quad + \rho(v_i, p) + \rho(p, T_i v_i) = \rho(v, v_i) + 2\rho(v_i, p) \\ &\le \rho(v, v_i) + 2\rho(v_i, v) + 2\rho(v, p).\end{aligned}$$

Hence for any $\varepsilon > 0$ $\rho(v, T_i v_i) < 2\rho(v, p) + \varepsilon$, for large i, which means that $T_i v_i \in K$ for all $i > N$ where K is a compact region in $\mathcal{H}$ which contradicts the local finiteness of F. □

COROLLARY 11.4. *A compact Dirichlet region has a finite number of vertices.*

We call two points $u, v \in \mathcal{H}$ *congruent* if they belong to the same Γ-orbit. First, notice that two points in a fundamental region F may be congruent only if they belong to the boundary of F. Suppose now that F is a Dirichlet region for Γ, and let us consider congruent vertices of F. The congruence is an equivalence relation on the vertices of F and the equivalence classes are called *cycles*. If u is fixed by an elliptic element S, then $v = Tu$ is fixed by the elliptic element TST^{-1}. Thus if one vertex of the cycle is fixed by an elliptic element, then all the vertices of that cycle are fixed by conjugate elliptic elements. Such a cycle is called an *elliptic cycle* and the vertices are called *elliptic vertices*. The number of elliptic cycles is equal to the number of non-congruent elliptic points in F.

Since the Dirichlet region F is a fundamental region, it is clear that every point $w \in \mathcal{H}$ fixed by an elliptic element S' of Γ lies on the boundary of $T(F)$ for some $T \in \Gamma$. Hence $u = T^{-1}(w)$ lies on the boundary of F and is fixed by the elliptic element $S = T^{-1}S'T$. Since Γ is a Fuchsian group, S has a finite order k. Suppose first that $k \ge 3$: then as S is an isometry fixing u which maps geodesics to geodesics, u must be a vertex whose angle θ is at most $2\pi/k$. (See Figure 11.2 where the angle at the elliptic fixed point ρ of order 3 is $2\pi/6$.) The hyperbolically convex region F is bounded by a union of geodesics. The intersection of F with these geodesics is either a single point or a segment of a geodesic. These segments are called *sides* of F. If S has order 2, its fixed point might lie on the interior of a side of F. In this case, S interchanges the two segments of this side separated by the fixed point. We will include such elliptic fixed points as vertices of F, the angle at such vertex being π. Thus a *vertex* of F is a point of intersection in $\mathcal{H}$ of two bounding geodesics of F or a fixed point of an elliptic element of

order 2. (All the previous definitions such as conjugate, elliptic cycles, etc. apply to this extended set of vertices.)

If a point in $\mathcal{H}$ has a nontrivial stabilizer in Γ, this stabilizer is a finite cyclic subgroup of Γ by Exercise 27; it is a *maximal finite cyclic subgroup* of Γ by Exercise 28. Conversely, every maximal finite cyclic subgroup of Γ is a stabilizer of a single point in $\mathcal{H}$. We can summarize the above as:

THEOREM 11.5. *There is a one-to-one correspondence between the elliptic cycles of F and the conjugacy classes of non-trivial maximal finite cyclic subgroups of Γ.*

EXAMPLE B. Let Γ be the modular group. The Dirichlet region F in Figure 9 has vertices in $\mathcal{H}$ at $\rho = \frac{-1+\sqrt{3}}{2}$, $\rho + 1 = \frac{1+\sqrt{3}}{2}$ and i. These are stabilized by the cyclic subgroups generated by $z \mapsto \frac{-z-1}{z}$, $z \mapsto \frac{z-1}{z}$, and $z \mapsto -\frac{1}{z}$, respectively. The verteces ρ and $\rho + 1$ belong to the same cycle since they are congruent via $z \to z + 1$. Each of them is fixed by an elliptic element of order 3. It is easy to check that these two verteces form an elliptic cycle. The point i is fixed by an elliptic element of order 2, and i is the only such point. Thus $\{i\}$ is an elliptic cycle consisting of just one vertex. By Theorem 11.5, the modular group has two conjugacy classes of maximal finite cyclic subgroups, one consisting of groups of order 2, the other consisting of groups of order 3.

DEFINITION 11.6. The orders of non-conjugate maximal finite cyclic subgroups of Γ are called the *periods* of Γ.

Each period is repeated as many times as there are conjugacy classes of maximal finite cyclic subgroups of that order. Thus the modular group has periods 2, 3.

A parabolic element can be considered as an elliptic element of infinite order; it has a unique fixed point in $\mathbb{R}\bigcup\{\infty\}$. Hence if a point in $\mathbb{R}\bigcup\{\infty\}$ has a non-trivial stabilizer in Γ all elements of which have only this fixed point, then this stabilizer is a *maximal (cyclic) parabolic subgroup* of Γ, and every maximal parabolic subgroup of Γ is a stabilizer of a single point in $\mathbb{R}\bigcup\{\infty\}$. Let F be a Dirichlet region for Γ with parabolic elements. It will be shown in §14 that in this case F is not compact (Theorem 14.2), and if additionally $\mu(F) < \infty$, then F has at least one *vertex at infinity*, i.e. two bounding geodesics of F meet there (Theorem 14.3). Moreover, each vertex at infinity is a parabolic fixed point for a maximal parabolic subgroup of Γ (Theorem 14.6), and non-congruent verteces at infinity of F are in a one-to-one correspondence with conjugasy classes of maximal parabolic subgroups of Γ (Corollary 14.7). If we allow infinite periods, the period ∞ will occur the same number of times as there are conjugacy classes of maximal parabolic subgroups. This number is called the *parabolic class number of* Γ. It is easily calculated that in the modular group every parabolic element is conjugate to $z \to z + n$ for some $n \in \mathbb{Z}$, so that the modular group has

periods 2, 3, ∞. The angle at a vertex at infinity is 0. With this convention, the Dirichlet region for the modular group described in §10 has a vertex at ∞ whose angle is $\frac{\pi}{\infty} = 0$.

The following result relates the sum of angles at all elliptic verticies belonging to an elliptic cycle with the order of that cycle.

THEOREM 11.7. *Let F be a Dirichlet region for Γ. Let $\theta_1, \theta_2, \cdots, \theta_t$ be the internal angles at all congruent vertices of F. Let m be the order of the stabilizer in Γ of one of these vertices. Then $\theta_1 + \cdots + \theta_t = 2\pi/m$.*

REMARKS. 1. As F is locally finite, there are only finitely many vertices in a congruent cycle.

2. As the stabilizers of two points in a congruent set are conjugate subgroups of Γ, they have the same order.

3. If a vertex is not a fixed point, we have $m = 1$ and $\theta_1 + \cdots + \theta_t = 2\pi$.

PROOF. Let $v_1, \ldots, v_t$ be the vertices of the congruent set, the internal angles being $\theta_1, \ldots, \theta_t$. Let

$$H = \left\{\mathrm{Id}, S, S^2, \ldots, S^{m-1}\right\}$$

be the stabilizer of v_1 in Γ. Then each $S^r(F)$ $(0 \le r \le m-1)$ has a vertex at v_1 whose angle is θ_1. Suppose $T_k(v_k) = v_1$ for some $T_k \in \Gamma$. Then the set of all elements which map v_k to v_1 is HT_k, a coset which has m elements, so the $S^r T_k(F)$ have v_1 as a vertex with an angle of θ_k. On the other hand, if a region $A(F)$ $(A \in \Gamma)$ has v_1 as a vertex, then $A^{-1}(v_1) \in F$, hence $A^{-1}(v_1) = v_i$ for some i, $1 \le i \le t$. Thus $A \in HT_i$, and $A(F)$ has been included in the above description. So we have mt regions surrounding v_1. These regions are distinct, for if $S^r T_k(F) = S^q T_l(F)$, then $S^r T_k = S^q T_l$, and hence $r = q$ and $k = l$. We conclude then that

$$m(\theta_1 + \cdots + \theta_t) = 2\pi.$$

□

We now consider the congruence of sides. Let s be a side of F, a Dirichlet region for a Fuchsian group Γ. If $T \in \Gamma \setminus \{\mathrm{Id}\}$ and $T(s)$ is a side of F, then s and $T(s)$ are called *congruent sides*. But $T(s)$ is also a side of $T(F)$ so that $T(s) \subseteq F \bigcap T(F)$. If a side of F has a fixed point of an elliptic element S of order 2 on it then S interchanges the two segments of this side. It is convenient to regard these two segments as distinct sides separated by a vertex. With this convention, one observes that for each side of F there exists another side of F congruent to it. There cannot be more than two sides in a congruent set. For, suppose that for some $T_1 \in \Gamma \setminus \{\mathrm{Id}\}$, $T_1(s)$ is also a side of F; then $T_1(s) = F \bigcap T_1(F)$. Thus $s = T_1^{-1}(F) \bigcap F = T^{-1}(F) \bigcap F$ so that $T_1^{-1}(F) \equiv T^{-1}(F)$ which implies $T_1 = T$. Thus the sides of F fall into congruent pairs. Hence if the number of sides of a Dirichlet region is finite, it is always even.

EXAMPLE B. The two vertical sides of the fundamental region for the modular group found in §10 (Figure 10.2) are congruent via the transformation $z \to z+1$. The arc of the unit circle between ρ and $\rho+1$ is the union of two sides: $[\rho, i]$ and $[i, \rho+1]$, congruent via the elliptic transformation of order 2, $z \to -1/z$.

THEOREM 11.8. *Let $\{T_i\}$ be the subset of Γ consisting of those elements which pair the sides of some fixed Dirichlet region F. Then $\{T_i\}$ is a set of generators for Γ.*

PROOF. Let Λ be the subgroup generated by the set $\{T_i\}$. We have to show that $\Lambda = \Gamma$. Suppose that $S_1 \in \Lambda$, and that $S_2(F)$ is adjacent to $S_1(F)$, i.e. they share a side. Then $S_1^{-1}S_2(F)$ is adjacent to F. Hence $S_1^{-1}S_2 = T_k$ for some $T_k \in \{T_i\}$; and since $S_2 = S_1T_k$ we conclude that $S_2 \in \Lambda$. Suppose now $S_3(F)$ intersects $S_1(F)$ in a vertex v. Then $S_1^{-1}S_3(F)$ intersects F in a vertex $u = S_1^{-1}v$. By Theorem 11.2, there can only be finitely many faces with vertex u, and F can be "connected" with $S_1^{-1}S_3(F)$ by a finite chain of faces in such a way that each two consecutive ones share a side. Hence we can apply the above argument repeatedly to show that $S_3 \in \Lambda$. Let $X = \bigcup_{S\in\Lambda} S(F)$, $Y = \bigcup_{S\in\Gamma\setminus\Lambda} S(F)$. Then $X \bigcap Y = \emptyset$. Clearly $X \cup Y = \mathcal{H}$, so if we show that X and Y are closed subsets of $\mathcal{H}$, then as $\mathcal{H}$ is connected and $X \neq \emptyset$, we must have $X = \mathcal{H}$ and $Y = \emptyset$. This would show that $\Lambda = \Gamma$ and the result will follow.

We now show that any union $\bigcup V_j(F)$ of faces of the tessellation is closed. Suppose $\{z_i\}$ is an infinite sequence of points of $\bigcup V_j(F)$ which tends to some limit $z_0 \in \mathcal{H}$. Then $z_0 \in T(F)$ for some $T \in \Gamma$, and by Theorem 11.2, any neighborhood N of z_0 intersects only finitely many of the $V_j(F)$. Therefore, one face of this finite family, say $V_m(F)$, must contain a subsequence of $\{z_i\}$ tending to z_0. Since $V_m(F)$ is closed, $z_0 \in V_m(F) \subseteq \bigcup V_j(F)$. Thus $\bigcup V_j(F)$ is closed, and, in particular, X and Y are closed. □

EXAMPLE B. Theorem 11.8 implies that the modular group is generated by $z \to z+1$ and $z \to -1/z$.

12. Connection with Riemann surfaces and homogeneous spaces

Let Γ be a Fuchsian group acting on the upper half-plane $\mathcal{H}$, and F be a fundamental region for this action. The group Γ induces a natural projection (continuous and open) $\pi : \mathcal{H} \to \Gamma\setminus\mathcal{H}$, and the points of $\Gamma\setminus\mathcal{H}$ are the Γ-orbits. The restriction of π to F identifies the congruent points of F that necessarily belong to its boundary ∂F, and makes $\Gamma\setminus F$ into an oriented surface with possibly some *marked points* (which correspond to the elliptic cycles of F) and *cusps* (which correspond to non-congruent vertices at infinity of F), also known as an *orbifold*. Its topological type is determined

by the number of cusps and by its *genus*—the number of handles if we view the surface as a sphere with handles. If F is locally finite, the quotient space $\Gamma\backslash\mathcal{H}$ is homeomorphic to $\Gamma\backslash F$ ([**5**], Theorem. 9.2.4), hence by choosing F to be a Dirichlet region which is locally finite by Theorem 11.2, we can find the topological type of $\Gamma\backslash\mathcal{H}$. We have seen in §9 (Theorem 9.2) that the area of a fundamental region (with nice boundary) is, if finite, a numerical invariant of the group Γ. Since the area on the quotient space $\Gamma\backslash\mathcal{H}$ is induced by the hyperbolic area on $\mathcal{H}$, 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 by Proposition 11.3, F has finitely many sides, and the quotient space $\Gamma\backslash\mathcal{H}$ is compact. We shall see in §14 (Corollary 14.4) that 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 [**26**].

Since Γ acts on $PSL(2,\mathbb{R})$ by left multiplication one can form the homogeneous space $\Gamma\backslash PSL(2,\mathbb{R})$. We have seen (Theorem 7.1) that $PSL(2,\mathbb{R})$ can be interpreted as the unit tangent bundle of the upper half-plane. It is easy to see (Exercise 24) that if F is a fundamental region for Γ in $\mathcal{H}$, SF is a fundamental region for Γ in $PSL(2,\mathbb{R})$. It also can be shown (see Exercise 25) that if Γ contains no elliptic elements, the homeomorphism described in Theorem 7.1 induces an homeomorphism of the corresponding quotient spaces. If Γ contains elliptic elements, an analogous result holds; however, the structure of the fibered bundle is violated in a finite number of marked points.

Since the fiber in $S(\Gamma\backslash\mathcal{H})$ over each point of $\Gamma\backslash\mathcal{H}$ is compact, $\Gamma\backslash\mathcal{H}$ is compact if and only if $S(\Gamma\backslash\mathcal{H})$ is compact.

13. Fuchsian groups of cofinite volume

THEOREM 13.1. (Siegel's Theorem) *If Γ is such that $\mu(\Gamma\backslash\mathcal{H}) < \infty$, then any Dirichlet region $F = D_p(\Gamma)$ has finitely many sides.*

PROOF. [**7**] Since the vertices of $D_p(\Gamma)$ are isolated (Proposition 11.3), any compact subset $K \subset \mathcal{H}$ contains only finitely many vertices. This takes care of the case in which F is compact. Now suppose that F is not compact.

The main ingredient of the proof is an estimation of the angles ω at vertices of the region F. More precisely, we are going to prove that

$$\sum_{\omega}(\pi - \omega) \leq \mu(F) + 2\pi, \tag{13.1}$$

where the sum is taken over all vertices of F lying in $\mathcal{H}$. We first notice that F is a star-like generalized polygon, and that the boundary of F, ∂F, is not necessarily connected. Let us connect all vertices of F with the point p by geodesics and consider the triangles thus obtained. Let

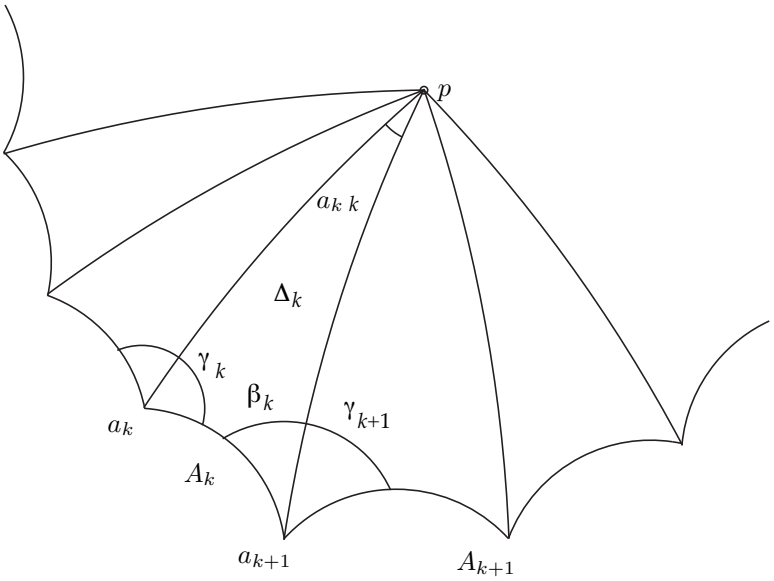


FIGURE 13.1. Proof of Siegel's Theorem

$\dots A_m, A_{m+1}, \dots, A_n, \dots$ be a connected set of geodesic segments in ∂F with vertices $\dots a_m, a_{m+1}, \dots, a_{n+1}, \dots$ (Figure 13.1).

We assume that this set is unbounded in both directions. We denote the triangle with the side A_k by Δ_k, its angles by $\alpha_k, \beta_k, \gamma_k$, and the angle between A_k and A_{k+1} by ω_k; thus we have

$$\omega_k = \beta_k + \gamma_{k+1}.$$

By the Gauss-Bonnet formula (Theorem 5.4) we have

$$\mu(\Delta_k) = \pi - \alpha_k - \beta_k - \gamma_k.$$

Thus

$$\sum_{k=m}^{n} \alpha_k + \sum_{k=m}^{n} \mu(\Delta_k) = \pi - \gamma_m - \beta_n + \sum_{k=m}^{n-1} (\pi - \omega_k). \tag{13.2}$$

The left-hand side of this equality is bounded since $\sum \alpha_k \le 2\pi$ and $\sum \mu(\Delta_k) \le \mu(F)$, hence the right-hand side is also bounded. It follows that $\sum(\pi - \omega_k)$ converges, and the following limits exist:

$$\lim_{m \to -\infty} \gamma_m = \gamma_\infty, \ \lim_{n \to \infty} \beta_n = \beta_\infty.$$

Let us show that

$$\pi - \gamma_\infty - \beta_\infty \ge 0. \tag{13.3}$$

Since only finitely many segments $\{A_k\}$ may be a bounded distance from the point p, we have $a_k \to \infty$ as $k \to \infty$. Thus $\rho(p, a_{k+1}) > \rho(p, a_k)$ for infinitely many values of k, and for these values, as follows, for instance, from the Sine Rule (Theorem 6.1(i)), we have $\gamma_k > \beta_k$. On the other hand,

$\beta_k + \gamma_k \le \pi$ and thus $\beta_k \le \pi/2$. Therefore $\beta_\infty \le \pi/2$. Similarly, $\gamma_\infty \le \pi/2$, and (13.3) follows.

Let $m \to -\infty, n \to \infty$. Taking into account (13.3) we obtain from 13.2 a limit inequality

$$\sum_{k=-\infty}^{\infty} \alpha_k + \sum_{k=-\infty}^{\infty} \mu(\Delta_k) \ge \sum_{k=-\infty}^{\infty} (\pi - \omega_k). \tag{13.4}$$

The inequality is obtained under the assumption that the connected set of segments $\{A_k\}$ is unbounded in both directions. Similar arguments apply in other cases when the connected set of segments is bounded at least in one direction. Adding up all these inequalities, we obtain a desired estimate

$$2\pi + \mu(F) \ge \sum_{\omega} (\pi - \omega), \tag{13.5}$$

where the sum is taken over all vertices of F which lie a finite hyperbolic distance from the point p, i.e. in $\mathcal{H}$.

Now we are going to prove, using this estimate, that the number of vertices which lie a finite distance from the point p is finite. Let a be a vertex and $a^{(1)} = a$, $a^{(2)}, \dots, a^{(n)}$ all vertices congruent to a. If we denote the angle at vertex $a^{(i)}$ by $\omega^{(i)}$, we have by Theorem 11.7

$$\omega^{(1)} + \omega^{(2)} + \cdots + \omega^{(n)} = 2\pi, \tag{13.6}$$

if a is not a fixed point for any $T \in \Gamma - \{\mathrm{Id}\}$; and

$$\omega^{(1)} + \omega^{(2)} + \cdots + \omega^{(n)} = 2\pi/m, \tag{13.7}$$

if a is a fixed point of order m. Since F is convex, $\omega^{(i)} < \pi$, and for each cycle of the type (13.6), we have $n \ge 3$, and hence

$$\sum_{i=1}^{n} (\pi - \omega^{(i)}) = (n-2)\pi > \pi. \tag{13.8}$$

Comparing (13.8) with (13.5) we conclude that the number of cycles where a is not a fixed point for any $T \in \Gamma \setminus \{\mathrm{Id}\}$ is finite. For each cycle of the type (13.7) we have

$$\sum_{i=1}^{n} (\pi - \omega^{(i)}) = (n - \frac{2}{m})\pi > \frac{\pi}{3}. \tag{13.9}$$

Comparing (13.9) with (13.5) we conclude that the number of elliptic cycles of order ≥ 3 is finite. Finally, any elliptic fixed point of order 2 belongs to a segment of ∂F between two vertices which are not elliptic points of order 2, hence we see that the number of elliptic cycles of order 2 is also finite. Thus we have proved that there are only finitely many vertices at a finite distance from the point p.

It remains to show that the number of vertices at infinity is also finite. Let us take any N vertices at infinity: $B_1, \dots, B_N$. It is obvious that there exists a hyperbolic polygon F_1 bounded by a finite number of geodesics and

contained inside F such that its vertices at infinity are $B_1, \ldots, B_N$. An argument similar to that in the proof of (13.2) shows that the hyperbolic area of F_1 satisfies the following equation:

$$\sum_{\omega}(\pi - \omega) = 2\pi + \mu(F_1),$$

where ω are the angles at the vertices of F_1, and the sum is taken over all vertices of F_1. Since $\omega = 0$ for all vertices at infinity, we have

$$\pi N \leq 2\pi + \mu(F_1) \leq 2\pi + \mu(F).$$

Thus N is bounded from above, and the theorem follows. □

14. Cocompact Fuchsian groups

DEFINITION 14.1. A Fuchsian group is called *cocompact* if the quotient-space $\Gamma\backslash\mathcal{H}$ is compact.

The following results reveal the relationship between cocompactness of Γ and the absence of parabolic elements in Γ.

THEOREM 14.2. *If a Fuchsian group Γ has a compact Dirichlet region, then Γ contains no parabolic elements.*

PROOF. Let F be a compact Dirichlet region for Γ and

$$\eta(z) = \inf\{\rho(z, T(z)) \mid T \in \Gamma \setminus \{\mathrm{Id}\}, T \text{ not elliptic}\}.$$

Since the Γ-orbit of each $z \in \mathcal{H}$ is a discrete set (Corollary 8.7) and $T(z)$ is continuous, $\eta(z)$ is a continuous function of z and $\eta(z) > 0$. Therefore, as F is compact, $\eta = \inf\{\eta(z) \mid z \in F\}$ is attained and $\eta > 0$. If $z \in \mathcal{H}$, there exists $S \in \Gamma$ such that $w = S(z) \in F$. Hence, if $T_0 \in \Gamma \setminus \{\mathrm{Id}\}$ is not elliptic,

$$\rho(z, T_0(z)) = \rho(S(z), S(T_0(z)) = \rho(w, ST_0S^{-1}(w)) \geq \eta,$$

and therefore

$$\inf\{\rho(z, T_0(z)) \mid z \in \mathcal{H},\ T_0 \text{ not elliptic}\} = \eta > 0.$$

Now suppose that Γ contains a parabolic element T_1. If for some $R \in PSL(2, \mathbb{R})$, $\Gamma_1 = R\Gamma R^{-1}$ then $R(F)$ will be a compact fundamental region for Γ_1. Thus by conjugating Γ in $PSL(2, \mathbb{R})$ we may assume that $T_1(z)$ or $T_1^{-1}(z)$ is the transformation $z \to z + 1$. However, by Theorem 3.5(c), $\rho(z, z + 1) \to 0$ as $\mathrm{Im}(z) \to \infty$, a contradiction. □

THEOREM 14.3.

(i) *If Γ has a non-compact Dirichlet region, then the quotient space $\Gamma\backslash\mathcal{H}$ is not compact.*
(ii) *If a Dirichlet region $F = D_p(\Gamma)$ for a Fuchsian group Γ has finite hyperbolic area but is not compact, then it has at least one vertex at infinity.*

PROOF. Let $F = D_p(\Gamma)$ be a non-compact Dirichlet region for Γ. We consider all oriented geodesic rays from the point p; each geodesic ray is uniquely determined by its direction l at the point p. Since F is a hyperbolically convex region, a geodesic ray in the direction l either intersects ∂F in a unique point or the whole geodesic ray lies inside F. Hence we can define a function $\tau(l)$ to be the length of a geodesic segment in the direction l inside F, $\tau(l)$ being equal to ∞ in the latter case. Obviously, $\tau(l)$ is a continuous function of l at the points where $\tau(l) < \infty$. Therefore if $\tau(l) < \infty$ for all l, the function $\tau(l)$ is bounded; hence the region F is compact. Thus if F is not compact, there are some directions l for which $\tau(l) = \infty$. After the identification of the congruent points of ∂F, we obtain a non-compact orbifold $\Gamma\backslash\mathcal{H}$ and (i) follows. To prove (ii), let us consider one such direction l_0. The intersection of the geodesic ray from p in the direction l_0 with the set of points at infinity belongs to $\partial_0 F$, the Euclidean boundary of F. By Theorem 13.1, F has finitely many sides, hence $\partial_0 F$ consists of finitely many free sides and vertices at infinity. Since $\mu(F) < \infty$, it is easy to see that $\partial_0 F$ cannot contain any free sides. Therefore this intersection is a vertex at infinity, and (ii) follows. □

COROLLARY 14.4. *The quotient space of a Fuchsian group* Γ, $\Gamma\backslash\mathcal{H}$, *is compact if and only if any Dirichlet region for* Γ *is compact.*

Let $p \in \mathcal{H}$ and $z(t)$, $0 \le t < \infty$, be a geodesic ray from the point p. Let $B_t(p)$ be a hyperbolic circle centered at $z(t)$ and passing through the point p. Exercise 26 asserts that the limit of $B_t(p)$, as $t \to \infty$, exists. It is a Euclidean circle passing through p and through the end of the geodesic $z(t)$, s, corresponding to $t = \infty$. It is orthogonal to the geodesic $z(t)$ at s, hence is tangent to the real axis, and therefore is a horocycle (see §4, Exercise 12, and Figure 4.2). Since the geodesic ray through p is determined by its direction l, the horocycle depends on p and l and is denoted by $\omega(p, l)$. Notice

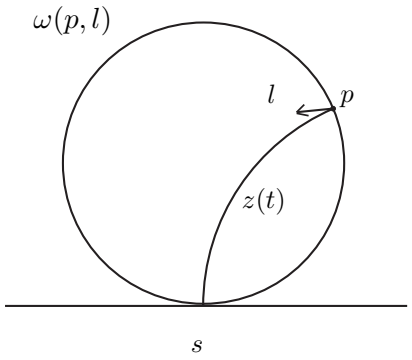


FIGURE 14.1. A horocycle

that horocycles are not hyperbolic circles, however they may be considered

as circles of infinite radius. A horocycle through a point s at infinity is denoted by $\omega(s)$.

THEOREM 14.5. *Let S be a transformation in $PSL(2,\mathbb{R})$ fixing a point $s \in \mathbb{R}$. Then S is parabolic if and only if for each horocycle through $s, \omega(s)$, we have $S(\omega(s)) = \omega(s)$.*

PROOF. Suppose first that S is parabolic, and $R \in PSL(2,\mathbb{R})$ is such that $R(s) = \infty$. Then $S_0 = R \circ S \circ R^{-1}$ is a parabolic transformation fixing ∞, and therefore $S_0(z) = z + h$, $h \in \mathbb{R}$. Since S_0 is a Euclidean translation, it maps each horizontal line to itself. Since a linear fractional transformation maps circles and straight lines into circles and straight lines and preserves angles, we conclude that horocycles are mapped into horocycles. Thus $S(\omega(s)) = \omega(s)$.

Conversely, suppose S maps each horocycle $\omega(s)$ onto itself. Making the same conjugation as above, we move the fixed point s to ∞. Then $S(z) = az + b$. The condition that each horizontal line is mapped into itself implies that $a = 1$. Hence S is a parabolic element. □

THEOREM 14.6. *Suppose Γ has a non-compact Dirichlet region $F = D_p(\Gamma)$ with $\mu(F) < \infty$. Then*

(i) each vertex of F at infinity is a parabolic fixed point for some $T \in \Gamma$.

(ii) If ξ is a fixed point of some parabolic element in Γ, then there exists $T \in \Gamma$ s.t. $T(\xi) \in \partial_0(F)$.

PROOF OF (i). Let b be a vertex of F at infinity. Let us consider all images $S(F)$, $S \in \Gamma$, which have the point b as a vertex. Obviously, there are infinitely many of them. Let $b^{(1)} = b$, $b^{(2)}, \ldots, b^{(n)}$ be all vertices of F congruent to b:

$$b^{(k)} = T_k(b) \quad (k = 1, \ldots, n).$$

We know from Theorem 13.1 that the number of such vertices is finite. Any image of F which has the point b as a vertex has a form

$$TT_i^{-1}(F) \ (i = 1, \ldots, n),$$

where T is any element of Γ which fixes the point b. Since there are infinitely many such images, and since T_i is only taken from a finite set of elements, we conclude that there are infinitely many elements $T \in \Gamma$ fixing b.

We shall show now that any such element T is a parabolic element. Suppose T is not parabolic. Let us consider a geodesic $z(t), 0 \leq t \leq \infty$, parametrized by its length, connecting the points p and $b (z(0) = p, z(\infty) = b)$. (See Figure 14.2.) Since F is a Dirichlet region the whole geodesic lies inside F and

$$\rho(p, z(t)) < \rho(T(p), z(t)), \quad 0 \leq t < \infty. \tag{14.1}$$

Consider a horocycle $\omega(b)$ containing the point p. Since by our assumption T is not a parabolic transformation, $T(p)$ does not belong to $\omega(b)$. Then

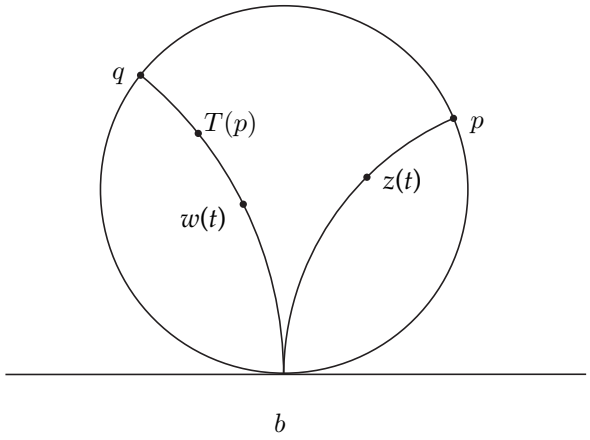


FIGURE 14.2. A vertex at infinity

by Exercise 29 either $T(p)$ or $T^{-1}(p)$ lies inside $\omega(b)$. We may assume then that $T(p)$ lies inside $\omega(b)$. Let $w(t)$ be a geodesic passing through $T(p)$ and b. Let q be a second point of intersection of $\omega(b)$ and $w(t)$; we choose the parametrization of $w(t)$ by its length such that $w(0) = q$. We notice first that $\rho(z(t), w(t)) \to 0$ as $t \to \infty$. In order to see this, we conjugate Γ so that its action on $\mathcal{H}$ gives: $b = \infty, z(t) = a + it, w(t) = c + it$ $(t \geq t_0 > 0)$. Then using Theorem 3.5(c), we obtain

$$\sinh\left[\frac{1}{2}\rho(z(t), w(t))\right] = \frac{|a-c|}{2t} \to 0 \text{ as } t \to 0,$$

and the claim follows. We have

$$\begin{aligned} t = \rho(p, z(t)) = \rho(q, w(t)) &= \rho(q, T(p)) + \rho(T(p), w(t)) \\ &\geq \rho(q, T(p)) + \rho(T(p), z(t)) - \rho(z(t), w(t)), \end{aligned}$$

and hence for sufficiently large t, we have

$$\rho(p, z(t)) > \rho(T(p), z(t)),$$

contradiction with (14.1). □

PROOF OF (ii). See Exercise 30. □

We leave the proof of the following Corollary (Exercise 31).

COROLLARY 14.7. *There is a one-to-one correspondence between non-congruent vertices at infinity of a Dirichlet fundamental region for a non-cocompact Fuchsian group Γ with $\mu(\Gamma\backslash\mathcal{H}) < \infty$ and conjugacy classes of maximal parabolic subgroups of Γ.*

The following result is a direct consequence of Theorems 13.1, 14.3, and 14.6.

COROLLARY 14.8. *A Fuchsian group* Γ *is cocompact if and only if* $\mu(\Gamma\backslash\mathcal{H}) < \infty$ *and* Γ *contains no parabolic elements.*

15. The signature of a Fuchsian group

We now assume that Γ has a compact fundamental region F. By Corollary 11.4 F has finitely many sides, and hence finitely many vertices, finitely many elliptic cycles, and by Theorem 11.5, a finite number of periods, say $m_1, \ldots, m_r$. As we have seen in §3.6 the quotient space $\Gamma\backslash\mathcal{H}$ is an orbifold, i.e. a compact, oriented surface of genus g with exactly r marked points. In this case we say that Γ has *signature* $(g; m_1, m_2, \ldots, m_r)$.

THEOREM 15.1. *Let* Γ *have signature* $(g; m_1, \ldots, m_r)$. *Then*

$$\mu(\Gamma\backslash\mathcal{H}) = 2\pi[(2g-2) + \sum_{i=1}^{r}\left(1 - \frac{1}{m_i}\right)].$$

PROOF. The area of the quotient space was defined in the beginning of §3.6: $\mu(\Gamma\backslash\mathcal{H}) = \mu(F)$ where F is a Dirichlet region. By Theorem 11.5 F has r elliptic cycles of vertices. (As described in §12, we include the interior point of a side fixed by an elliptic element of order 2 as a vertex whose angle is π, and then regard this side as being composed of two sides separated by this vertex.) By Theorem 11.7, the sum of angles at all elliptic vertices is $\sum_{i=1}^{4}\frac{2\pi}{m_i}$. Suppose there exist s other cycles of vertices. Since the order of the stabilizers of these vertices is equal to 1, the sum of angles at all these vertices is equal to $2\pi s$. Thus the sum of all angles of F is equal to

$$2\pi[\left(\sum_{i=1}^{r}\frac{1}{m_i}\right) + s].$$

The sides of F are matched up by elements of Γ. If we identify those matched sides, we obtain an orbifold of genus g. If F has n such sets of identified sides, we obtain a decomposition of $\Gamma\backslash\mathcal{H}$ into $(r+s)$ vertices, n edges, and 1 simply connected face. By the Euler formula,

$$2 - 2g = (r+s) - n + 1.$$

Exercise 32 gives a formula for the hyperbolic area of a hyperbolic polygon. Using it, we obtain

$$\mu(F) = (2n-2)\pi - 2\pi[\left(\sum_{i=1}^{r}\frac{1}{m_i}\right) + s] = 2\pi[(2g-2) + \sum_{i=1}^{r}\left(1 - \frac{1}{m_i}\right)].$$

□

It is quite surprising that the converse to Theorem 15.1 is also true, i.e. that there exists a Fuchsian group with a given signature. This first appeared in Poincaré's paper on Fuchsian groups [**23**], but the rigorous proof was given much later and is based on a more general result of Maskit [**21**].

THEOREM 15.2. (Poincaré's Theorem) *If $g \geq 0$, $r \geq 0$, $m_i \geq 2$ $(1 \leq i \leq r)$ are integers and if*

$$(2g-2) + \sum_{i=1}^{r} \left(1 - \frac{1}{m_i}\right) > 0,$$

then there exists a Fuchsian group with signature $(g; m_1, \ldots, m_r)$.

The sketch of the proof of Theorem 15.2 given in [**13**]; it is illustrated on the following example.

EXAMPLE D. Construction of a Fuchsian group with signature $(2; -)$. Since $r = 0$, a fundamental region is a regular hyperbolic octagon F_8 (see Figure 15.1) of hyperbolic area 4π. We call this group Γ_8. We suppose

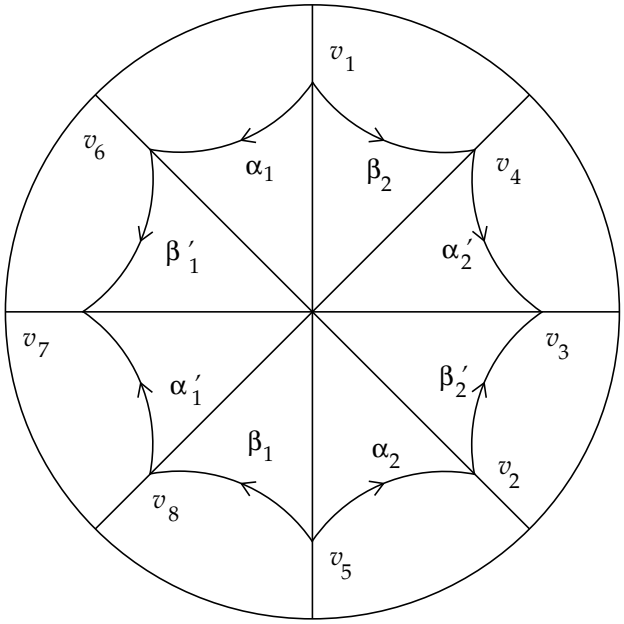


FIGURE 15.1. Fundamental region for the group Γ_8

that t_0 is chosen such that $\mu(N(t_0)) = 4\pi$. Then the area of each of the 8 isosceles hyperbolic triangles is equal to $\frac{\pi}{2}$; and since the angle at the origin is equal $\frac{\pi}{4}$, by the Gauss-Bonnet formula (Theorem 5.4) the two other angles are equal to $\frac{\pi}{8}$. The group Γ_8 is generated by 4 hyperbolic elements, A_1, A_2, B_1, and B_2, that identify the sides of F_8 as shown in Figure 15.1. Since all eight sides of F_8 are arcs of circles of the same Euclidean radius of

equal Euclidean length, the sides identified by a generator must be isometric circles of this generator and its inverse. This allows us to use elementary geometry to explicitely write down those generators. Let

$$A_2 = \begin{bmatrix} a & c \\ \bar{c} & \bar{a} \end{bmatrix}, \tag{15.1}$$

then the isometric circle $I(A_2)$ is given by the equation $|\bar{c}z + \bar{a}| = 1$. By Exercise 17, A_2 maps $I(A_2)$ onto $I(A_2^{-1})$ in such a way that the center of $I(A_2)$, $-\frac{\bar{a}}{\bar{c}}$, is mapped onto the center of $I(A_2^{-1})$, $\frac{a}{\bar{c}}$. But from Figure 15.1 we see that $\frac{ia}{\bar{c}} = -\frac{\bar{a}}{\bar{c}}$, which implies $a = \pm|a|(\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}})$. Let the radius of $I(A_2) = R$, and the distance of the center of $I(A_2)$ from the origin be d. By elementary geometric arguments, we have $d = R(1+\sqrt{2})$. On the other hand, $|c| = \frac{1}{R}$, and $d = \frac{|a|}{|c|} = R|a|$, hence $|a| = 1+\sqrt{2}$; and since $|a|^2 - |c|^2 = 1$, we have $|c| = \sqrt{2+2\sqrt{2}}$. Now let us choose the $+$ sign in the expression for a, i.e. $\mathrm{Arg}(a) = \frac{\pi}{4}$. Since $\mathrm{Arg}(-\frac{\bar{a}}{\bar{c}}) = \frac{\pi}{8}$, we obtain $\mathrm{Arg}(c) = -\frac{5\pi}{8}$. Using the formulas $\cos\frac{5\pi}{8} = -\frac{\sqrt{2-\sqrt{2}}}{2}$ and $\sin\frac{5\pi}{8} = \frac{\sqrt{2+\sqrt{2}}}{2}$, we obtain the expressions for the numbers a and c in (15.1):

$$a = \frac{2+\sqrt{2}}{2}(1+i), c = -\frac{\sqrt[4]{2}}{2}(\sqrt{2} + i(2+\sqrt{2})).$$

Other generators of the group Γ_8 can also be expressed in terms of parameters a and c as follows: $A_1 = \begin{bmatrix} a & -c \\ -\bar{c} & \bar{a} \end{bmatrix}$, $B_1 = \begin{bmatrix} \bar{a} & -\bar{c} \\ -c & a \end{bmatrix}, B_2 = \begin{bmatrix} \bar{a} & \bar{c} \\ c & a \end{bmatrix}$.

Let $R : \mathcal{H} \to \mathcal{U}$ be a map given by $R(z) = \frac{zi+1}{z+i}$, see (4.4). Then $\Gamma = R^{-1}\Gamma_8 R$ be a subgroup of $PSL(2,\mathbb{R})$ whose generators are:

$$A_2 = \begin{bmatrix} \mathrm{Re}(a) + \mathrm{Im}(c) & \mathrm{Im}(a) + \mathrm{Re}(c) \\ -(\mathrm{Im}(a) - \mathrm{Re}(c)) & \mathrm{Re}(a) - \mathrm{Im}(c) \end{bmatrix},$$

$$A_1 = \begin{bmatrix} \mathrm{Re}(a) - \mathrm{Im}(c) & \mathrm{Im}(a) - \mathrm{Re}(c) \\ -(\mathrm{Im}(a) + \mathrm{Re}(c)) & \mathrm{Re}(a) + \mathrm{Im}(c) \end{bmatrix},$$

$$B_1 = \begin{bmatrix} \mathrm{Re}(a) + \mathrm{Im}(c) & -\mathrm{Im}(a) - \mathrm{Re}(c) \\ \mathrm{Im}(a) - \mathrm{Re}(c) & -\mathrm{Re}(a) - \mathrm{Im}(c) \end{bmatrix},$$

$$B_2 = \begin{bmatrix} \mathrm{Re}(a) - \mathrm{Im}(c) & -\mathrm{Im}(a) + \mathrm{Re}(c) \\ \mathrm{Im}(a) + \mathrm{Re}(c) & \mathrm{Re}(a) + \mathrm{Im}(c) \end{bmatrix}.$$

As elements of $PSL(2,\mathbb{R})$, the generators are:

$$A_2 = \begin{bmatrix} \frac{(2+\sqrt{2})(1-\sqrt[4]{2})}{2} & \frac{(2+\sqrt{2})-\sqrt[4]{2}\sqrt{2}}{2} \\ -\frac{(2+\sqrt{2})+\sqrt[4]{2}\sqrt{2}}{2} & \frac{(2+\sqrt{2})(1+\sqrt[4]{2})}{2} \end{bmatrix},$$

$$A_1 = \begin{bmatrix} \frac{(2+\sqrt{2})(1+\sqrt[4]{2})}{2} & \frac{(2+\sqrt{2})+\sqrt[4]{2}\sqrt{2}}{2} \\ -\frac{(2+\sqrt{2})-\sqrt[4]{2}\sqrt{2}}{2} & \frac{(2+\sqrt{2})(1-\sqrt[4]{2})}{2} \end{bmatrix},$$

$$B_1 = \begin{bmatrix} \frac{(2+\sqrt{2})(1-\sqrt[4]{2})}{2} & \frac{-(2+\sqrt{2})+\sqrt[4]{2}\sqrt{2}}{2} \\ -\frac{-(2+\sqrt{2})-\sqrt[4]{2}\sqrt{2}}{2} & \frac{(2+\sqrt{2})(1+\sqrt[4]{2})}{2} \end{bmatrix},$$

$$B_2 = \begin{bmatrix} \frac{(2+\sqrt{2})(1+\sqrt[4]{2})}{2} & \frac{-(2+\sqrt{2})-\sqrt[4]{2}\sqrt{2}}{2} \\ -\frac{-(2+\sqrt{2})+\sqrt[4]{2}\sqrt{2}}{2} & \frac{(2+\sqrt{2})(1-\sqrt[4]{2})}{2} \end{bmatrix}.$$

It can be shown that Γ_8 is derived from the quaternion algebra over $\mathbb{Q}(\sqrt{2})$.

Exercises

24. Prove that if F is a fundamental region for a Fuchsian group Γ on $\mathcal{H}$, then SF is a fundamental region for Γ on $S\mathcal{H}$.

25. Prove that if Γ is a Fuchsian group without elliptic elements, then $S(\Gamma\backslash\mathcal{H})$ is homeomorphic to $\Gamma\backslash PSL(2,\mathbb{R})$.

26. Show that the limit of $B_t(p)$ as $t \to \infty$ is a Euclidean circle passing through p and the end of the geodesic $z(t)$ corresponding to $t = \infty$, and orthogonal to the geodesic $z(t)$.

27. Prove that an elliptic subgroup of $PSL(2,\mathbb{Z})$ is a Fuchsian group if and only of it is finite.

28. If $ST = TS$ then S maps the fixed-point set of T to itself.

29. Let T be a non-parabolic transformation fixing a point b at infinity, $\omega(b)$ be a horocycle, $p \in \omega(b)$. Prove that either $T(p)$ or $T^{-1}(p)$ lies inside $\omega(b)$.

30. Let Γ be a non-elementary Fuchsian group, F a locally finite fundamental region for Γ, and ξ a fixed point of some parabolic element in Γ, then there exists $T \in \Gamma$ s.t. $T(\xi) \in \partial_0(F)$.

31. Give a careful proof of Corollary 14.7.

32. Prove the Gauss-Bonnet formula for an n-sided star-like hyperbolic polygon Π with angles $\alpha_1, \ldots, \alpha_n$:

$$\mu(\Pi) = (n-2)\pi - \sum_{i=1}^{n} \alpha_i.$$

Lecture III. Geodesic flow

16. First properties

The geodesic flow $\{\tilde{\varphi}^t\}$ on $\mathcal{H}$ is defined as an $\mathbb{R}$-action on the unit tangent bundle $S\mathcal{H}$ which moves a tangent vector along the geodesic defined by this vector with unit speed. As was explained in §7, $S\mathcal{H}$ can be identified with $PSL(2,\mathbb{R})$, by sending v to the unique $g \in PSL(2,\mathbb{R})$ such that $z = g(i)$, $\zeta = g'(z)(\iota)$, where ι is the unit vector at the point i to the imaginary axis pointing upwards.

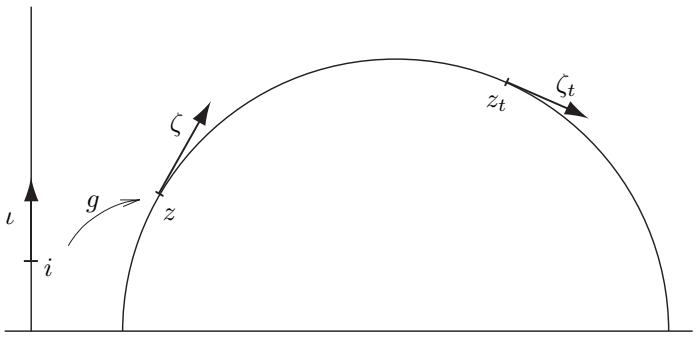


FIGURE 16.1. Geodesic flow on the upper half-plane $\mathcal{H}$

Under this identification the $PSL(2,\mathbb{R})$-action on $\mathcal{H}$ by Möbius transformations corresponds to left multiplications (Theorem 7.1, and the geodesic flow corresponds to the right multiplication by the one-parameter subgroup

$$a_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \text{ such that } \tilde{\varphi}^t(v) \leftrightarrow g a_t \,. \tag{16.1}$$

The orbit $\{ga_t\}$ projects to a geodesic through $g(i)$. The quotient space $\Gamma\backslash S\mathcal{H}$ can be identified with the unit tangent bundle of $M = \Gamma\backslash\mathcal{H}$, SM, although the structure of the fibered bundle is violated at elliptic fixed points and cusps (see §12 for details). The geodesic flow $\{\tilde{\varphi}^t\}$ on $\mathcal{H}$ descents to the *geodesic flow* $\{\varphi^t\}$ on the factor M via the projection $\pi : S\mathcal{H} \to SM$ of the unit tangent bundles (see e.g. [**15**, §5.3, 5.4] for more details).

In this chapter we will assume that $\mu(M) < \infty$.

THEOREM 16.1. *The geodesic flow $\{\varphi^t\} : SM \to SM$ preserves a smooth measure $dv = d\mu d\theta$ which is called the Liouville measure.*

PROOF. In [**15**, Theorem 5.3.6] a more general statement for geodesic flows on Riemannian manifolds is a corollary from the fact that geodesic flows are Hamiltonian flows. In this case the theorem follows from the fact that the volume $dv = d\mu d\theta$ is the left-invariant Haar measure on $PSL(2,\mathbb{R})$ that is also right-invariant since $PSL(2,\mathbb{R})$ is an unimodular group. □

The following theorem is crucial in establishing further important properties of the geodesic flow.

THEOREM 16.2. *The geodesic flow $\{\varphi^t\}$ is Anosov, i.e. there exists a $C^\infty\{\varphi^t\}$ - invariant decomposition of the tangent bundle to SM,$T(SM) = E^0 \oplus E^+ \oplus E^-$ such that*

(a) *The integral curves of E^0 are orbits of the geodesic flow.*
(b) *The integral curves of E^+ and E^- (we call them* stable *and* unstable *manifolds and denote them W^+ and W^-, respectively) are the unit normal vector fields to the horocycles orthogonal to the orbits of $\{\varphi^t\}$.*
(c) *There exist positive constants C and λ such that for any pair of points $x_1, x_2 \in SM$ lying on the same leaf of W^+ (or W^-),*

$$d^{W^\pm}(\varphi^t(x_1), \varphi^t(x_2)) \le Ce^{-\lambda|t|} d^{W^\pm}(x_1, x_2) \text{ for } t \ge 0, \ (t \le 0).$$

Here $d^{W^\pm}$ is the distance on the corresponding stable or unstable manifold.

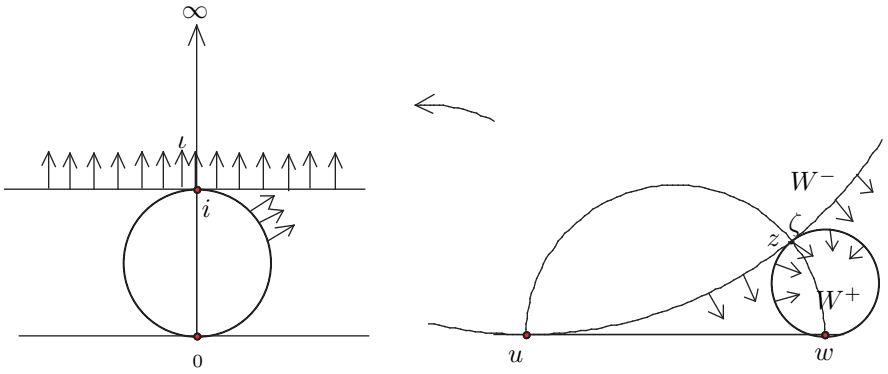


FIGURE 16.2. Stable and unstable manifolds

PROOF. For each $v = (z, \zeta) \in SM$, let $z(t)$ be the geodesic through v with the fixed points $w = z(\infty)$ and $u = z(-\infty)$. As we saw in §14, there are two horocycles on $\mathcal{H}$ passing through the point z: one tangent to the real axis at w, another - at u. Let $W^+(v)$ be the unit vector field containing v and normal to the horocycle passing through w, and $W^-(v)$ be the unit vector field containing v and normal to the horocycle passing through u. In order to prove the estimates in (c), we "move" the geodesic $z(t)$ to the positive imaginary axis by a transformation $\gamma \in PSL(2, \mathbb{R})$ so that $\gamma(z) = i$ (this can be done by Exercise 4), and make the calculations for this particular case. The stable manifold W^+ will be mapped to the upward unit vector field normal to the horocycle $H = \mathbb{R} + i$, and the unstable manifold W^- will be mapped to the outward unit vector field normal to the horocycle passing

through i and 0 (see Figure 16.2). Let $x_1 = (i, \iota)$, where ι is the unit vector tangent to the imaginary axis pointed upwards, and $x_2 = (i + x, \iota)$. Then

$$d^{W^+}(x_1, x_2) = x,$$

and after the time $t > 0$,

$$d^{W^+}(\varphi^t(x_1), \varphi^t(x_2)) = xe^{-t},$$

so the estimates hold with $C = \lambda = 1$. The estimates for the unstable manifold are obtained by the change of direction of the flow. □

DEFINITION 16.3. A point $v \in SM$ is call *nonwandering* with respect to the flow $\{\varphi^t\}$ if for every open set $U \ni v$ there is T such that $\varphi^T(U) \cap U \neq \emptyset$.

REMARK. It follows from Poincaré Recurrence Theorem that since the geodesic flow is volume-preserving, every point of SM is nonwandering [**15**, Theorem 4.1.18].

17. Dynamics of the geodesic flow

Let $x \in SM$ and W_x^+ be the stable manifold containing x. Denote by D_x^+ the set of all points w in W_x^+ with $d^{W^+}(x, w) < \delta_0$, where δ_0 will be chosen later (see Figure 17.1). For any point $w \in D_x^+$ we will denote by W_w^- the unstable manifold containing w, and by D_w^- the set of all $z \in W_w^-$ with $d^{W^-}(z, w) < \delta_0$.

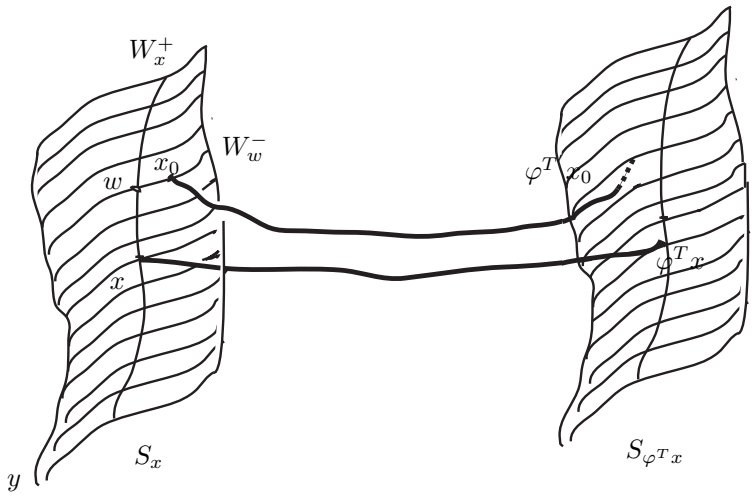


FIGURE 17.1. Dynamical parametrization of SM

Let $S_x = \{z \mid z \in D_w^- \text{ for some } w \in D_x^+\}$. For small enough δ_0, S_x is a submanifold of dimension 2 transversal to the orbit of $\{\varphi^t\}$. This construction gives us a convenient way to parameterize SM locally by the coordinates (t, u, v): we parameterize D_x^+ by the length u measured over the stable leaf,

and D^-_w by the length v measured over the unstable leaf. The coordinate t is the length on the orbit of the flow through x.

THEOREM 17.1 (Anosov Closing Lemma). *Suppose $x \in SM$ is such that $d(x, \varphi^T(x)) < \varepsilon$. Then*

(a) *there exists $x_0 \in SM$ whose orbit is closed, i.e. $\varphi^{T'}(x_0) = x_0$ and such that for some constant C, $d(x_0, x) < C\varepsilon$ and $|T' - T| < C\varepsilon$;*
(b) *for $0 \le t \le T$ and some other constant C'*

$$d(\varphi^t x_0, \varphi^t x) \le C'\varepsilon e^{-\min(t, T-t)}.$$

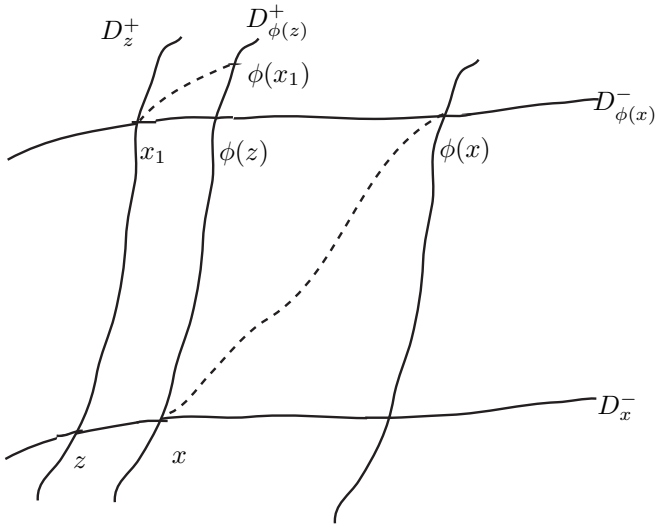


FIGURE 17.2. Proof of Anosov Closing Lemma

PROOF. Consider $S_x \ni x$, as in the beginning of this section, of some fixed size δ_0. Then there exists t_0 with $|t_0| < C_1\varepsilon$ for some constant C_1 and $\gamma \in \Gamma$ so that $\tilde\varphi^{T+t_0} S_x \ni d\gamma x$. Consider a map $\phi = (d\gamma)^{-1}\tilde\varphi^{T+t_0} : S_x \to S_x$. Since $d\gamma$ is an isometry, we have $d(x, \phi(x)) < C_1\varepsilon$. There exists $z \in \mathcal{D}^-_x$ such that $\phi(z) \in \mathcal{D}^+_x$ (see Figure 17.2). By property (c), $d^{W^-}(z, x) \le C_2\varepsilon e^{-T}$ for some constant C_2. Take $x_1 \in \mathcal{D}^-_{\phi(z)} \cap \mathcal{D}^+_z$, then $\phi(x_1) \in \mathcal{D}^+_{\phi(z)}$ and $d^{W^+}(\phi(x_1), \phi(z)) < C_3\varepsilon e^{-T}$ for some constant C_3. Therefore $d(x_1, \phi(x_1)) < C_4\varepsilon e^{-T}$. Continuing this process we get a fixed point $x_0 : \phi(x_0) = x_0$, i.e. $(d\gamma)x_0 = \varphi^{T'} x_0$ for some $T' : |T' - T| < C_1\varepsilon$.

We have $d(x, x_0) < C\varepsilon$. By construction $x_0 \in S_x$. Let $w \in D^+_x$ be a point such that $x_0 \in \mathcal{D}^-_w$. Then for some constant C_3, $d^{W^+}(x, w) < C_3\varepsilon$ and $d^{W^-}(w, x_0) < C_3\varepsilon$. For the same reason since $d(\varphi^T x, \varphi^{T'} x_0) < C\varepsilon$, we can conclude that for some t_1 with $|t_1| < C'_2\varepsilon$, $\varphi^{T'+t_1} x_0 \in S_{\varphi^T x}$ and that $d^{W^+}(\varphi^T x, \varphi^T w) < C_4\varepsilon$ and $d^{W^-}(\varphi^T w, \varphi^T x_0) < C_4\varepsilon$ for another constant

C_4. Property (c) implies that $d^{W^+}(\varphi^t x, \varphi^t w) < C_5\varepsilon e^{-t}$ and $d^{W^-}(\varphi^t w, \varphi^t x_0) \le C_5\varepsilon e^{-(T-t)}$ and therefore $d(\varphi^t x, \varphi^t x_0) \le C_6\varepsilon e^{-\min(t,T-t)}$. □

The following results are obtained by geometric considerations (see [**15**, §5.4] for cocompact case).

THEOREM 17.2. *Let $M = \mathcal{H}\backslash\Gamma$ and Γ be a Fuchsian group such that $\mu(M) < \infty$. Then the geodesic flow $\{\varphi^t\}$ has a dense orbit on SM, that is, it is* topologically transitive.

PROOF. We will prove that for any two nonempty open balls $U, V \in SM$ there is $t \in \mathbb{R}$ such that $\varphi^t(U) \cap V \neq \emptyset$, a property equivalent to topological transitivity [**15**, Lemma 1.4.2]. It is convenient to visualize this using the unit disc model $\mathcal{U}$ for the hyperbolic plane (see Figure 17.3).

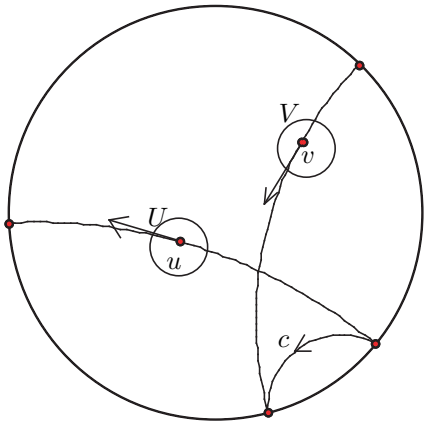


FIGURE 17.3. Topological transitivity of the geodesic flow

Using remark at the end of §16 and Theorem 17.1 we can find two periodic points, $u \in U$ and $v \in V$ (whose lifts to $S\mathcal{U}$ we also denote by u and v). Let c_u and c_v be geodesics in $\mathcal{U}$ such that $\dot{c_u} = u$ and $\dot{c_v} = v$. We may assume that $c_u(-\infty) \neq c_v(\infty)$, otherwise we replace c_u by $\gamma(c_u)$ for some $\gamma \in \Gamma$. Consider the geodesic c such that $c(\infty) = c_v(\infty)$ and $c(-\infty) = c_u(-\infty)$. By Theorem 16.2, for each $t \in \mathbb{R}$ we can find two numbers $g_u(t)$ and $g_v(t)$ such that

$$d(\dot{c_u}(g_u(t)), \dot{c}(t)) < e^{-t} \text{ as } t \to -\infty$$

and

$$d(\dot{c_v}(g_v(t)), \dot{c}(t)) < e^{-t} \text{ as } t \to \infty.$$

Since c_u and c_v project to closed geodesics on $\Gamma\backslash\mathcal{U}$, this shows that there exist t_1 and t_2 such that the projection of $\dot{c}(t_1)$ to SM is in U and the projection of $\dot{c}(t_2)$ to SM is in V. This yields the claim. □

The following important result follows immediately from Theorems 17.2 and 17.1:

COROLLARY 17.3. *Periodic orbits of the geodesic flow are dense in SM.*

THEOREM 17.4. *The Liouville measure $dv = d\mu d\phi$ on SM is ergodic under the geodesic flow.*

PROOF. The proof [**15**, Theorem 5.4.16] uses so-called "Hopf arguments", an important tool for hyperbolic dynamic. We will show that the ergodic average

$$f^+(x) = f_{\varphi^t}(x) := \lim_{T\to\infty} \frac{1}{T}\int_0^T f(\varphi^t(x))dt$$

is constant a.e. for every function $f \in L^1(SM, v)$, a property equivalent to ergodicity. It is sufficient to prove that for a continuous function f with compact support (hence uniformly continuous) since such functions are dense in $L^1(SM, v)$. Consider such an f. Then by Birkhoff Ergodic Theorem [**15**, Theorem 4.1.2]

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(\varphi^t(x))dt$$

exists a.e. First we show that $f^+(x)$ is constant on stable leaves. Suppose the limit exists for some $p \in SM$. We will prove that it also exists for all $q \in W^+(p)$ and is independent on q: Given $\varepsilon > 0$, there exists T_0 such that

$$f(\varphi^t(p)) - f(\varphi^t(q)) < \varepsilon$$

for all $t > T_0$ by uniform continuity. But this means that

$$|\frac{1}{T}\int_o^T (f(\varphi^t(p)) - f(\varphi^t(q)))dt| < \varepsilon$$

for sufficiently large T, as required. Since existence and the value of the limit is φ^t-invariant, $f^+(x) = f_{\varphi^t}(x)$ is, in fact, constant on weak stable manifolds, the integral manifold of $E^0 \oplus E^+$. Consider also the negative time average

$$f^-(x) = \lim_{T\to\infty} \frac{1}{T}\int_{-T}^0 f(\varphi^t(x))dt.$$

It exists and is constant a.e. on unstable manifolds. Furthermore, by a corollary to Birkhoff Ergodic Theorem [**15**, Proposition 4.1.3] $f^+(x) = f^-(x)$ a.e. In terms of the local C^1 coordinates on SM (t, u, v) introduced in the proof of Theorem 17.1, on a small open set U, by Fubini's Theorem, for t in a set C of full measure $f^+(t, u, v) = f^-(t, u, v)$ for almost all (u, v). But then for any such t_1 and t_2 the corresponding sets of (u, v) intersect since they both have full measure. Therefore, $f^+(t_1, u, v) = f^+(t_2, u, v)$ for a.e. (u, v), hence $f^+(x)$ is constant on C, as required. □

18. Livshitz's Theorem

A. Livshitz in [**20**] proved his theorem for Anosov systems on any compact manifold and he proved that the function he obtained was C^1. Guillemin and Kazhdan in [**9**] gave a proof of Livshitz's theorem for Anosov flows, and in [**10**] they gave a proof for geodesic flows on any compact surface and proved that the function is C^∞.

Here we will prove a version of Livshitz's Theorem for the geodesic flow on SM of finite volume and possibly with cusps. In this case, it is natural to formulate it for the class $\mathcal{BL}$ of bounded Lipschitz functions.

THEOREM 18.1. *Let Γ be a discrete subgroup of $PSL(2,\mathbb{R})$, $M = \Gamma\backslash\mathcal{H}$, $v(M) < \infty$, $X = S(M)$, and let $\{\varphi^t\}$ be the geodesic flow acting on X. Let $f \in \mathcal{BL}(X)$ be a function having zero integrals over all periodic orbits of $\{\varphi^t\}$. Then there exists a function F on X satisfying a Lipschitz condition with some constant α and differentiable in the direction of the geodesic flow and such that $\mathcal{D}F = f$, where $\mathcal{D} = \frac{d}{dt}$ is the operator of differentiation along the orbits of the flow $\{\varphi^t\}$.*

PROOF. Consider a point $x_0 \in X$ with a dense orbit $\mathcal{O}(x_0)$ and define a map $F : \mathcal{O}(x_0) \to \mathbb{R}$ by the formula $F(x) = \int_0^s f(\varphi^t(x_0))dt$ at $x = \varphi^s(x_0)$. We want to prove that F extends to a function on X satisfying a Lipschitz condition and $\mathcal{D}F = f$. We claim that $F(x)$ satisfies a Lipschitz condition on the orbit $\mathcal{O}(x_0)$. Let $y_1 = \varphi^{t_1}(x_0)$, $y_2 = \varphi^{t_1+T}(x_0) \in \mathcal{O}(x_0)$ and $d(y_1, y_2) < \varepsilon$. Then

$$F(y_2) - F(y_1) = \int_0^T f(\varphi^t(y_1))dt.$$

By Theorem 17.1 we can find a point $x_1 \in X$ such that $\varphi^{T'}x_1 = x_1$ for some $T' : |T' - T| < C_1\varepsilon$ and such that for $0 \le t \le T$, $d(\varphi^t x_1, \varphi^t y_1) < C_2\varepsilon$ for some constant C_2. Since the orbit $\mathcal{O}(x_1)$ is periodic, we have

$$\int_0^{T'} f(\varphi^t(x_1))dt = 0.$$

Therefore, using the estimates of Theorem 17.1(b) we have

$$|F(y_2) - F(y_1)| = |\int_0^T f(\varphi^t(y_1))dt - \int_0^{T'} f(\varphi^t(x_1))dt|$$
$$\le |\int_0^T f(\varphi^t(y_1)) - f(\varphi^t(x_1))dt| + |\int_T^{T'} f(\varphi^t(x_0)dt| < \alpha\varepsilon$$

for some constant α. The Lipschitz property of the function f is used to estimate the first term, and the boundeness to estimate the second. This proves the claim. Therefore F can be extended from the dense set to a function satisfying a Lipschitz condition on X. Since $\frac{d}{dt}F = f$ on the dense set it follows that F is differentiable in the direction of the geodesic flow and $\frac{d}{dt}F = f$ on X. □

Exercises

33. Prove that if the function f is C^1, then the function F is also C^1. *Hint:* show that F is differentiable in the directions of W_x^+ and W_x^- using the transversality of these curves in SM, the fact that they depend continuously on x, and the uniform convergence of the integral expression for derivative.

Lecture IV. Symbolic coding of geodesics

19. Representation of the geodesic flow as a special flow

A *cross-section* C for the geodesic flow is a subset of the unit tangent bundle SM visited by (almost) every geodesic infinitely often both in the future and in the past. In other words, every $v \in C$ defines an oriented geodesic $\gamma(v)$ on M which will return to C infinitely often. The function $f : C \to \mathbb{R}$ giving the *time of the first return* to C is defined as follows: if $v \in C$ and t is the time of the first return of $\gamma(v)$ to C, then $f(v) = t$. The map $R : C \to C$ defined by $R(v) = \varphi^{f(v)}(v)$ is called the *first return map*. Thus $\{\varphi^t\}$ can be represented as a *special flow* on the space

$$C^f = \{(v, s) \mid v \in C,\ 0 \le s \le f(v)\},$$

given by the formula $\varphi^t(v, s) = (v, s + t)$ with the identification $(v, f(v)) = (R(v), 0)$.

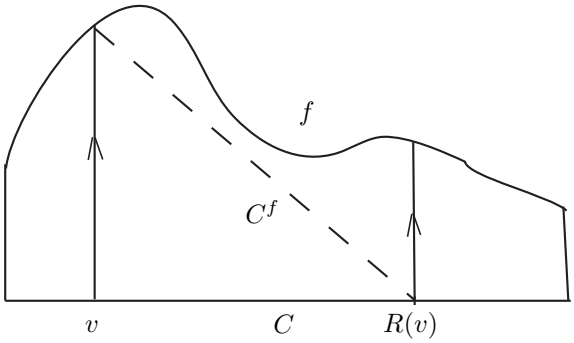


FIGURE 19.1. Geodesic flow is a special flow

Let $\mathcal{N}$ be a finite or countable alphabet, $\mathbb{N}^{\mathbb{Z}} = \{x = \{n_i\}_{i\in\mathbb{Z}} \mid n_i \in \mathbb{N}\}$ be the space of all bi-infinite sequences endowed with the Tikhonov (product) topology,

$$\sigma : \mathbb{N}^{\mathbb{Z}} \to \mathbb{N}^{\mathbb{Z}} \text{ defined by } \{\sigma x\}_i = n_{i+1}$$

be the left shift map, and $\Lambda \subset \mathbb{N}^{\mathbb{Z}}$ be a closed σ-invariant subset. Then (Λ, σ) is called a *symbolic dynamical system*. There are some important classes of such dynamical systems. The whole space $(\mathbb{N}^{\mathbb{Z}}, \sigma)$ is called the *Bernoulli shift*. If the space Λ is given by a set of simple transition rules which can be described with the help of a matrix consisting of zeros and ones, we say that

(Λ, σ) is a *one-step topological Markov chain* or simply a *topological Markov chain* (sometimes (Λ, σ) is also called a *subshift of finite type*). Similarly, if the space Λ is determined by specifying which $(k+1)$-tuples of symbols are allowed, we say that (Λ, σ) is a *k-step topological Markov chain* (a precise definition is given in Section 25).

In order to represent the geodesic flow as a special flow over a symbolic dynamical system, one needs to choose an appropriate cross-section C and code it, i.e. to find an appropriate symbolic dynamical system (Λ, σ) and a continuous surjective map $\mathfrak{C} : \Lambda \to C$ (in some cases the actual domain of $\mathfrak{C}$ is Λ except a finite or countable set of excluded sequences) defined such that the diagram

$$\begin{array}{ccc} \Lambda & \xrightarrow{\sigma} & \Lambda \\ {\scriptstyle \mathfrak{C}}\downarrow & & \downarrow{\scriptstyle \mathfrak{C}} \\ C & \xrightarrow{R} & C \end{array}$$

is commutative. We can then talk about *coding sequences* for geodesics defined up to a shift which corresponds to a return of the geodesic to the cross-section C. Notice that usually the coding map is not injective but only finite-to-one (see e.g. [**1**, §3.2 and §5]).

There are two essentially different methods of coding geodesics on surfaces of constant negative curvature. The geometric code, with respect to a given fundamental region, is obtained by a construction universal for all Fuchsian groups, while. The second method is specic for the modular group and is of arithmetic nature: it uses continued fraction expansions of the endpoints of the geodesic at infinity and a so-called reduction theory. They will be discussed in the subsequent sections.

20. Geometric coding

The Morse method. We first describe the general method of coding geodesics on a surface of constant negative curvature by recording the sides of a given fundamental region cut by the geodesic. This method first appeared in a paper by Morse [**22**] in 1921. However, in a 1927 paper, Koebe [**19**] mentioned an unpublished work from 1917, where the same ideas were apparently used. Starting with [**25**] Series called this method *Koebe-Morse*, but since this earlier work by Koebe has not been traced, we think it is more appropriate to call this coding method the *Morse method*. We will follow [**14**] in describing the Morse method for a finitely generated Fuchsian group Γ of cofinite hyperbolic area.

A Dirichlet fundamental region $\mathcal{D}$ of Γ always has an even number of sides identified by generators of Γ and their inverses (Theorems 13.1 and 11.8); we denote this set by $\{g_i\}$. We label the sides of $\mathcal{D}$ (on the inside) by elements of the set $\{g_i\}$ as follows: if a side s is identified in $\mathcal{D}$ with the side $g_i(s)$, we label the side s by g_i. By labeling all the images of s under Γ by the same generator g_i we obtain the labeling of the whole net

$\mathcal{S} = \Gamma(\partial\mathcal{D})$ of images of sides of $\mathcal{D}$, such that each side in $\mathcal{S}$ has two labels corresponding to the two images of $\mathcal{D}$ shared by this side. We assign to an oriented geodesic in $\mathcal{H}$ a bi-infinite sequence of elements of $\{g_i\}$ which label the successive sides of $\mathcal{S}$ this geodesic crosses.

We describe the *Morse coding sequence* of a geodesic in $\mathcal{H}$ under the assumption that it does not pass through any vertex of the net $\mathcal{S}$—we call such *general position geodesics*. (Morse called the coding sequences *admissible line elements*, and some authors [**25, 8**] referred to them as *cutting sequences*.) We assume that the geodesic intersects $\mathcal{D}$ and choose an initial point on it inside $\mathcal{D}$. After exiting $\mathcal{D}$, the geodesic enters a neighboring image of $\mathcal{D}$ through the side labeled, say, by g_1 (see Figure 20.1). Therefore this image is $g_1(\mathcal{D})$, and the first symbol in the code is g_1. If it enters

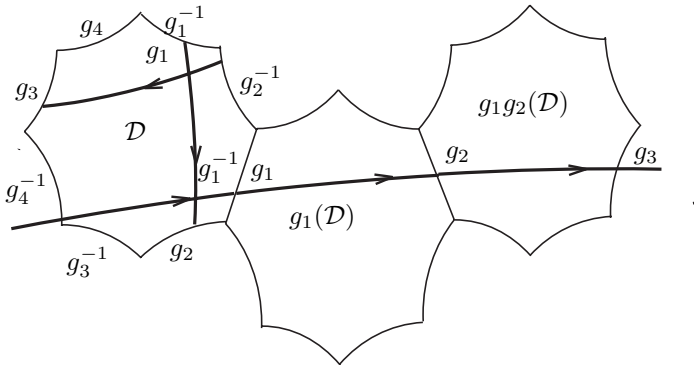


FIGURE 20.1. Morse coding

the second image of $\mathcal{D}$ through the side labeled by g_2, the second image is $(g_1 g_2 g_1^{-1})(g_1(\mathcal{D})) = g_1 g_2(\mathcal{D})$, and the second symbol in the code is g_2, and so on. Thus we obtain a sequence of all images of $\mathcal{D}$ crossed by our geodesic in the direction of its orientation: $\mathcal{D}$, $g_1(\mathcal{D})$, $g_1 g_2(\mathcal{D}), \dots$, and a sequence of all images of $\mathcal{D}$ crossed by our geodesic in the opposite direction: $g_0^{-1}(\mathcal{D})$, $(g_0 g_{-1})^{-1}(\mathcal{D}), \dots$. Thus, the Morse coding sequence is

$$[\dots g_{-1}, g_0, g_1, g_2, \dots].$$

By mapping the oriented geodesic segments between every two consecutive crossings of the net $\mathcal{S}$ back to $\mathcal{D}$ (as shown in Figure 20.1), we obtain a geodesic in $\mathcal{D}$. The coding sequence described above may be obtained by taking generators labeling the sides of $\mathcal{D}$ (on the outside) the geodesic hits consequently.

A geodesic on M is closed if and only if it is the projection of the axis of a hyperbolic element in Γ. For general position geodesics, a coding sequence is periodic if and only if the geodesic is closed. If a geodesic is the axis of a primitive hyperbolic element $g \in \Gamma$, i.e. a hyperbolic element which is not a

power of another element in Γ, we have

$$g = g_1 g_2 \cdots g_n$$

for some n. In this case the sequence is periodic with the least period $[g_1, g_2, \ldots, g_n]$.

An ambiguity in assigning a Morse code occurs whenever a geodesic passes though a vertex of $\mathcal{D}$: such geodesics have more than one code, and closed geodesics have non-periodic codes along with periodic ones (see [**8, 17**] for relevant discussions).

For free groups Γ with properly chosen fundamental regions, all reduced (here this simply means that a generator g_i does not follow or precede g_i^{-1}) bi-infinite sequences of elements from the generating set $\{g_i\}$ are realized as Morse coding sequences of geodesics on M (see [**25**]), but, in general, this is not the case. Even for the classical example of $\Gamma = PSL(2, \mathbb{Z})$ with the standard fundamental region F (Figure 10.2) no elegant description of admissible Morse coding sequences is known and probably does not exist. Important results in this direction were obtained in [**8**], where the admissible coding sequences were described in terms of forbidden blocks. The set of generating forbidden blocks found in [**8**] has an intricate structure attesting the complexity of the Morse code.

Geometric code for the modular surface. Let $\Gamma = PSL(2, \mathbb{Z})$, and $M = \Gamma \backslash \mathcal{H}$ be the modular surface. Recall that the generators of $PSL(2, \mathbb{Z})$ acting on $\mathcal{H}$ are $T(z) = z+1$ and $S(z) = -\frac{1}{z}$. The Morse code with respect to the standard fundamental region F can be assigned to any oriented geodesic γ in F (which does not go to the cusp of F in either direction), and can be described by a bi-infinite sequence of integers as follows. The boundary of F consists of four sides: left and right vertical, identified and labeled by T, and T^{-1}, respectively; left and right circular both identified and labeled by S (see Figure 20.2). It is clear from geometrical considerations that any oriented geodesic (not going to the cusp) returns to the circular boundary of F infinitely often. We first assume that the geodesic is in general position, i.e. does not pass through the corner $\rho = \frac{1}{2} + i\frac{\sqrt{3}}{2}$ of F (see Figure 20.2). We choose an initial point on the circular boundary of F and count the number of times it hits the vertical sides of the boundary of F moving in the direction of the geodesic. A positive integer is assigned to each block of hits of the right vertical side (or a block of T's in the Morse code), and a negative, to each block of hits of the left vertical side (or a block of T^{-1}'s). Moving the initial point in the opposite direction allows us to continue the sequence backwards. Thus we obtain a bi-infinite sequence of nonzero integers

$$[\gamma] = [\ldots, n_{-2}, n_{-1}, n_0, n_1, \ldots],$$

uniquely defined up to a shift, which is called the *geometric code* of γ. Moving the initial point in either direction until its return to one of the circular sides of F corresponds to a shift of the geometric coding sequence $[\gamma]$. Recall that a geodesic in general position is closed if and only if the coding

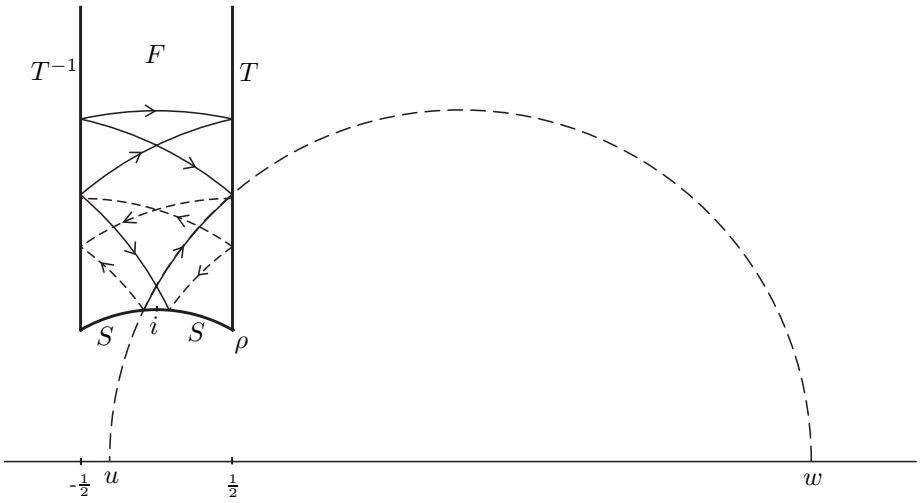


FIGURE 20.2. The fundamental region and a geodesic on M

sequence is periodic. We refer to the least period $[n_0, n_1, \dots, n_m]$ as its geometric code. For example, the geometric code of the closed geodesic on Figure 20.2 is $[4, -3]$.

A geodesic with geometric code $[\gamma]$ can be lifted to the upper half-plane $\mathcal{H}$ (by choosing the initial point appropriately) so that it intersects

$$T^{\pm 1}(F), \dots, T^{n_0}(F), T^{n_0}S(F), \dots, T^{n_0}ST^{n_1}S(F), \dots,$$

in the positive direction (the sign in the first group of terms is chosen in accordance with the sign of n_0, etc.) and

$$S(F), ST^{\mp 1}(F), \dots, ST^{-n_{-1}}(F), \dots, ST^{-n_{-1}}ST^{-n_{-2}}(F), \dots,$$

in the negative direction.

The case when a geodesic passes through the corner ρ of F was described to a great extent in [**8**, §7]. Such a geodesic has multiple codes obtained by approximating it by general position geodesics which pass near the corner ρ slightly higher or slightly lower. If a geodesic hits the corner only once it has exactly two codes. If a geodesic hits the corner at least twice, it hits it infinitely many times and is closed; if it hits the corner n times in its period, it has exactly $2n+2$ codes, i.e. shift-equivalence classes of coding sequences, some of which are not periodic. It is unknown, however, whether there is an upper bound on the number of shift-equivalence classes of coding sequences corresponding to closed geodesics [**8**, §9].

Canonical codes considered in [**14**] were obtained by the convention that a geodesic passing through ρ in the clockwise direction exits F through the right vertical side of F labeled by T (this corresponds to the approximation by geodesics which pass near the corner ρ slightly higher). According to this convention, the geometric codes of the axes of transformations

$A_4 = T^4S$, $A_{3,6} = T^3ST^6S$ and $A_{6,3} = T^6ST^3S$ are $[4], [3,6]$ and $[6,3]$, respectively. However, all these geodesics have other codes. For example, the axis of A_4 has a code $[2,-1]$ obtained by approximation by geodesics which pass near the corner ρ slightly lower, and two non-periodic codes for the same closed geodesic are

$$[\dots, 4, 4, 3, -1, 2, -1, 2, -1, 2, \dots] \text{ and } [\dots, 2, -1, 2, -1, 2, -1, 3, 4, 4, \dots].$$

For more details, see [**8, 17**].

21. Symbolic representation of geodesics via geometric code.

Let

$$\mathbb{N}^{\mathbb{Z}} = \{x = \{n_i\}_i \in \mathbb{Z} \mid n_i \in \mathbb{N}\}$$

be the set of all bi-infinite sequences on the alphabet $\mathbb{N} = \{n \in \mathbb{Z} \mid n \neq 0\}$, endowed with the Tykhonov product topology, and $\sigma : \mathbb{N}^{\mathbb{Z}} \to \mathbb{N}^{\mathbb{Z}}$ be the left shift map given by $\{\sigma x\}_i = n_{i+1}$. Let X_0 be the set of admissible geometric coding sequences for general position geodesics in M, and X be its closure in the Tykhonov product topology. It was proved in [**8**, Theorem 7.2] that every sequence in X is a geometric code of a unique oriented geodesic in M, and every geodesic in M has at least one and at most finitely many codes (see examples above). Thus X is a closed σ-invariant subspace of $\mathbb{N}^{\mathbb{Z}}$.

The cross-section for the geometric code and its partition. Since every oriented geodesic that does not go to the cusp of F in either direction returns to the circular boundary of F infinitely often, the set $B \subset SM$ consisting of all unit vectors in SM with base points on the circular boundary of F and pointing inside F (see Figure 21.1) is a cross-section which captures the geometric code.

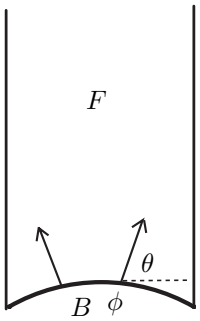


FIGURE 21.1. The cross-section B

We parameterize the cross-section B by the coordinates (ϕ, θ), where $\phi \in [-\pi/6, \pi/6]$ parameterizes the arc and $\theta \in [-\phi, \pi - \phi]$ is the angle the unit vector makes with the positive horizontal axis in the clockwise direction. The elements of the partition of B are labeled by the symbols of the alphabet $\mathbb{N}$, $B = \cup_{n\in\mathbb{N}} C_n$, and are defined by the following condition: $C_n = \{v \in B \mid n_0(v) = n\}$, i.e. C_n consists of all tangent vectors v in

B such that, for the coding sequence of the corresponding geodesic in $\mathcal{H}$, $n_0(x) = n$. Let $R : B \to B$ be the first return map. Since the first return to the cross-section exactly corresponds to the left shift of the coding sequence x associated to v, we have $n_0(R(v)) = n_1(v)$. The infinite geometric partition and its image under the return map R are sketched on Figure 21.2. Boundaries between the elements of the partition shown on

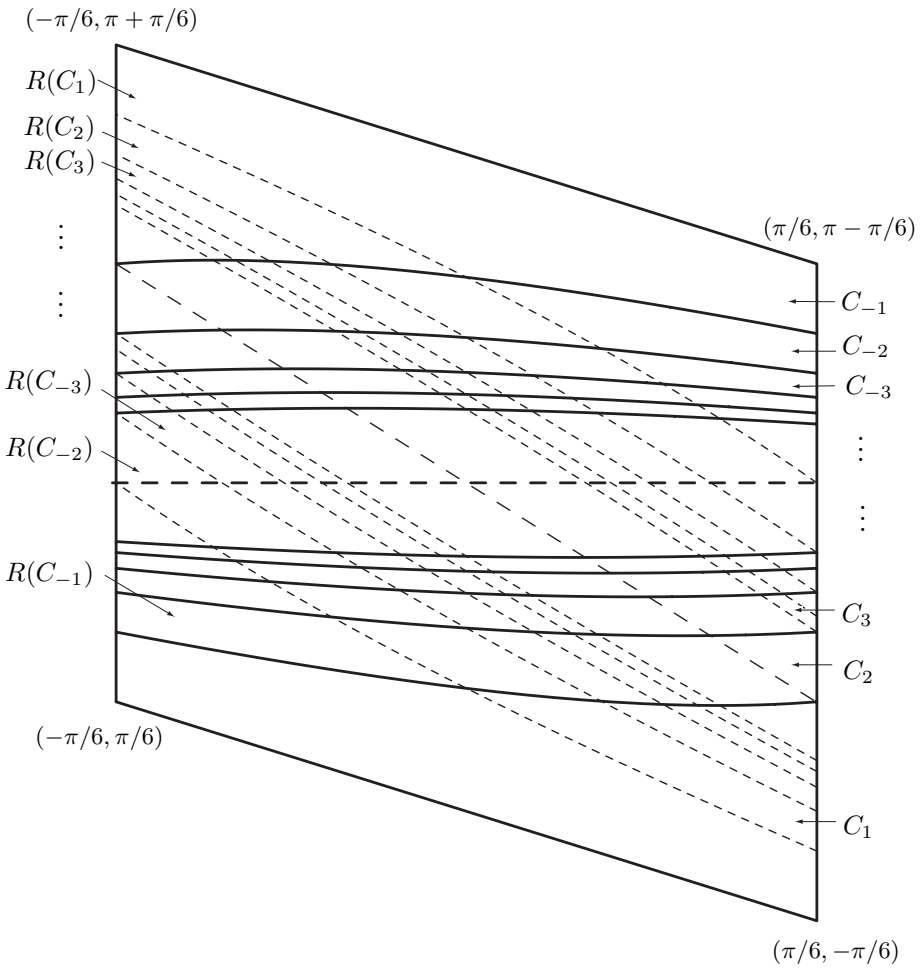


FIGURE 21.2. The infinite geometric partition and its image under the return map R

Figure 21.2 correspond to geodesics going into the corner; the two vertical boundaries of the cross-section B are identified and correspond to geodesics emanating from the corner. They have more than one code. For example, the codes $[4]$ and $[\dots, 2, -1, 2, -1, 2, -1, 3, 4, 4, 4, \dots]$ correspond to the point on the right boundary of B between C_4 and C_3, and the codes $[2, -1]$ and $[\dots, 4, 4, 4, 4, 3, -1, 2, -1, 2, -1, 2, \dots]$ correspond to the point on the

left boundary between C_2 and C_3 that are identified and are the four codes of the axis of A_4.

The coding map for the geometric code. It was proved in [**8**, Lemma 7.1] that if a sequence of general position geodesics is such that the sequence of their coding sequences converges in the product topology, then the sequence of these geodesics converges to a limiting geodesic uniformly. Since the tangent vectors in the cross-section B are determined by the intersection of the corresponding geodesics with the unit circle, we conclude that the sequence of images of the coding sequences under the map $\mathfrak{C} : X \to B$ converges to the image of the limiting coding sequence. This implies that the map $\mathfrak{C}$ is continuous.

Which geometric codes are realized? Not all bi-infinite sequences of nonzero integers are realized as geometric codes. For instance, the periodic sequence $\{\overline{8, 2}\}$ is not a geometric code since the geometric code of the axis of T^8ST^2S is $[6, -2]$, as can be seen on Figure 21.3 [**14**].

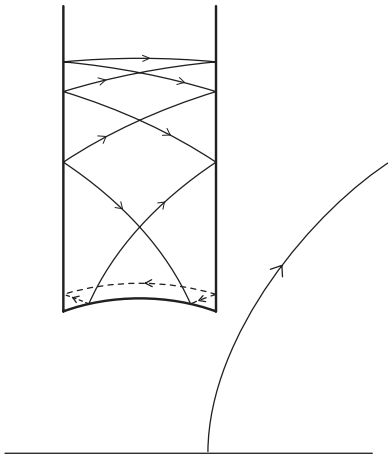


FIGURE 21.3. The geometric code of the axis of T^8ST^2S is $[6, -2]$

Figure 21.2 gives an insight into the complexity of the geometric code, where the elements C_n and their forward iterates $R(C_n)$ are shown. Each C_n is a curvilinear quadrilateral with two vertical and two "horizontal" sides, and each $R(C_n)$ is a curvilinear quadrilateral with two vertical and two "slanted" sides. The horizontal sides of C_n are mapped to vertical sides of $R(C_n)$, and the vertical sides of C_n are stretched across the parallelogram representing B and mapped to the "slanted" sides of $R(C_n)$.

If $n_0(v) = n$ and $n_1(v) = m$ for some vector $v \in B$, then $R(C_n) \cap C_m \neq \emptyset$. Therefore, as Figure 21.2 illustrates, the symbol 2 in a geometric code cannot be followed by 1, 2, 3, 4 and 5.

We say that C_m and $R(C_n)$ intersect "transversally" if their intersection is a curvilinear parallelogram with two "horizontal" sides belonging to the

horizontal boundary of C_m and two "slanted" sides belonging to the slanted boundary of $R(C_n)$. Notice that for each transverse intersection $R(C_n) \cap C_m$ its forward iterate under R stretches to a strip inside $R(C_m)$ between its two vertical sides. Hence, the symbol m can follow symbol n in a coding sequence.

We also observe that the elements C_m and $R(C_n)$ intersect transversally if and only if $|n| \geq 2$, $|m| \geq 2$, and

$$|1/n + 1/m| \leq 1/2.$$

This is a flow-invariant subset which constitutes the essential part of the set of geometrically Markov codes; see Theorems 25.2 and 25.4 in Section 25.

22. Arithmetic codings

In this section we describe a method of constructing arithmetic codes for geodesics on the modular surface M using expansions of the endpoints of their lifts to $\mathcal{H}$ in what we call *generalized minus continued fractions* [**18**]. Three classical continued fraction expansions described in that paper were defined using different integer-valued functions $(\cdot)$ (generalized "integer part" functions) that are included into a 2-parameter family of integer-valued functions suggested for consideration recently by Don Zagier,

$$(x)_{a,b} = \begin{cases} \lfloor x - a \rfloor & \text{if } x \leq a \\ 0 & \text{if } a < x < b \\ \lceil x - b \rceil & \text{if } x \geq b\,, \end{cases} \tag{22.1}$$

where $\lfloor x \rfloor$ denotes the integer part of x, $\lceil x \rceil = \lfloor x \rfloor + 1$, and

$$(a, b) \in \Delta = \{-1 \leq a \leq 0 \leq b \leq 1, b - a \geq 1\}.$$

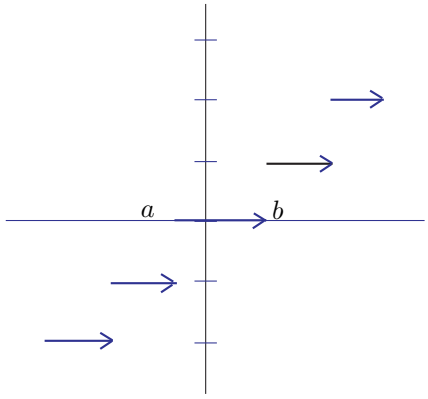


FIGURE 22.1. The function $(x)_{a,b}$

Any irrational number x can be expressed in a unique way as an infinite (a,b)-*continued fraction*

$$x = a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots}}}$$

that we will denote by $(a_0, a_1, \cdots)$ for short. The "digits" a_i, $i \geq 1$, are non-zero integers determined recursively by

$$n_0 = (x)_{a,b},\ x_1 = -\frac{1}{x - n_0},\ \text{and}\ n_i = (x_i)_{a,b},\ x_{i+1} = -\frac{1}{x_i - n_i}. \qquad (22.2)$$

The following theorem is a starting point of the study of what we call (a,b)-*continued fractions*, an ongoing project of the author with I. Ugarcovici and D. Zagier.

THEOREM 22.1. *Let $\{a_n\}$ be a sequence of integers defined by (22.2) and*

$$r_n = (a_0, a_1, \ldots, a_n) := a_0 - \cfrac{1}{a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots - \cfrac{1}{a_n}}}}.$$

Then the sequence r_n converges to x.

The three classical continued fraction expansions can be now described as follows.

***G*-expansion** $(a = -1, b = 0)$**.** The function

$$(x)_{-1,0} = \lceil x \rceil = \lfloor x \rfloor + 1$$

(that differs for integers from the classical ceiling function) gives the *minus continued fraction expansion* described in [**27**] and used in [**14**] for coding closed geodesics. Since the coding procedure for closed geodesic is the same as the Gauss reduction theory for indefinite integral quadratic forms, we refer to this expansion as the *Gauss-* or *G-expansion* and call the corresponding code *G-code*. G-codes for oriented geodesics, not necessarily closed, were introduced in [**11**]. The digits $n_0, n_1, \ldots$ of a G-expansion satisfy the condition $n_i \geq 2$, if $i \geq 1$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \ldots$ with $n_i \geq 2$ for $i \geq 1$ defines a real number whose G-expansion is $\lceil n_0, n_1, n_2, \ldots \rceil$.

***A*-expansion** $(a = -1, b = 1)$**.** The function

$$(x)_{-1,1} = \lceil x \rfloor = \begin{cases} \lfloor x \rfloor & \text{if } x > 0 \\ \lceil x \rceil & \text{if } x < 0 \end{cases}$$

gives an expansion which was used in [**18**] to reinterpret the classical Artin code *(A-code)*. This expansion has digits of alternating signs, and we call it the *A-expansion*. Conversely, any infinite sequence of nonzero integers with alternating signs $n_0, n_1, n_2, \dots$ defines a real number whose A-expansion is $\lceil n_0, n_1, n_2, \dots \rfloor$.

The G- and A-expansions satisfy the following properties:

(1) Two irrationals x, y are $PSL(2,\mathbb{Z})$-equivalent $\iff$ their expansions have the same tail, that is, if $x = (n_0, n_1, \dots)$ and $y = (m_0, m_1, \dots)$ then $n_{i+k} = m_{i+l}$ for some integers k, l and all $i \geq 0$;
(2) A real number x is a quadratic irrationality $\iff$ $(n_0, n_1, \dots)$ is eventually periodic;
(3) Let x and x' be conjugate quadratic irrationalities, i.e. the roots of a quadratic polynomial with integer coefficients. If $x = (\overline{n_0, n_1 \dots, n_k})$, then $\frac{1}{x'} = (\overline{n_k, \dots, n_1, n_0})$.

Let us remark that properties (2) and (3) are also valid for the regular continued fractions, while property (1) holds if one replaces $PSL(2,\mathbb{Z})$ by $PGL(2,\mathbb{Z})$.

***H*-expansion** $(a = -\frac{1}{2}, b = \frac{1}{2})$**.** This expansion is obtained using the function

$$(x)_{-\frac{1}{2},\frac{1}{2}} = \langle x \rangle$$

(the nearest integer to x). It was first used by Hurwitz [**12**] in order to establish a reduction theory for indefinite real quadratic forms, and we call it the *Hurwitz- or H-expansion*. The digits n_i $(i \geq 1)$ of an H-expansion satisfy $|n_i| \geq 2$, and if $|n_i| = 2$ then $n_i n_{i+1} < 0$. Conversely, any infinite sequence of integers $n_0, n_1, n_2, \dots$ with the above property defines an irrational number whose H-expansion is $\langle n_0, n_1, n_2, \dots \rangle$.

The H-expansion satisfies property (2), but not (1) and (3). There is a minor exception to property (1): it is possible for two irrationals not sharing the same tail to be $PSL(2,\mathbb{Z})$-equivalent, but this can happen if and only if one irrational has a tail of 3's in its H-expansion and the other one has a tail of -3's, i.e. the irrationals are equivalent to $r = (3 - \sqrt{5})/2$ ([**12, 6**]). Property (3) is more serious. In order to construct a meaningful code, we need to use a different expansion for $1/u$ (introduced also by Hurwitz) so that a property similar to (3) is satisfied. It uses yet another integer-valued function which is a part of the (a, b)-family:

$$(x)_{r-1,1-r} = \langle\langle x \rangle\rangle = \begin{cases} \langle x \rangle - \operatorname{sgn}(x) & \text{if} \quad \operatorname{sgn}(x)(\langle x \rangle - x) > r = (3 - \sqrt{5})/2, \\ \langle x \rangle & \text{otherwise,} \end{cases}$$

and is called the *H-dual expansion*. Now if $x = \langle \overline{n_0, n_1 \dots, n_k} \rangle$, then $\frac{1}{x'}$ has a purely periodic H-dual expansion $\frac{1}{x'} = \langle\langle \overline{n_k, \dots, n_1, n_0} \rangle\rangle$. The formula for $\langle\langle \cdot \rangle\rangle$ comes from the fact that if $x = \langle n_0, n_1, \dots \rangle$ then the entries n_i satisfy the asymmetric restriction: if $|n_i| = 2$, then $n_i n_{i+1} < 0$. For more details, see [**12, 6, 18**]; for a general definition of a dual expansion see §23.

Convergents. If $x = (n_0, n_1, \dots)$, then the *convergents* $r_k = (n_0, n_1, \dots, n_k)$ can be written as p_k/q_k where p_k and q_k are obtained inductively as:

$$p_{-2} = 0\,,\ p_{-1} = 1\,;\ p_k = n_k p_{k-1} - p_{k-2} \ \text{ for } k \geq 0$$
$$q_{-2} = -1\,,\ q_{-1} = 0\,;\ q_k = n_k q_{k-1} - q_{k-2} \ \text{ for } k \geq 0\,.$$

The following properties are shared by all three expansions:

- $1 = q_0 \leq |q_1| < |q_2| < \dots$;
- $p_{k-1}q_k - p_k q_{k-1} = 1$, for all $k \geq 0$.

The rates of convergence, however, are different. For the A- and H-expansions we have

$$\left| x - \frac{p_k}{q_k} \right| \leq \frac{1}{q_k^2}, \tag{22.3}$$

while for the G-expansion we only have

$$\left| x - \frac{p_k}{q_k} \right| \leq \frac{1}{q_k}. \tag{22.4}$$

A quadratic irrationality x has a purely periodic expansion if and only if x and x' satisfy certain *reduction inequalities*, which give us the notion of a *reduced geodesic* for each code.

Definition 22.2. An oriented geodesic in $\mathcal{H}$ going from u to w (with u, w irrationals) is called

- *G-reduced* if $0 < u < 1$ and $w > 1$;
- *A-reduced* if $|w| > 1$ and $-1 < \operatorname{sgn}(w)u < 0$;
- *H-reduced* if $|w| > 2$ and $\operatorname{sgn}(w)u \in [r-1, r]$.

Now we can describe a reduction algorithm which works for each arithmetic α-code, where $\alpha = G, A, H$. For the H-code we consider only geodesics whose end points are not equivalent to r.

Reduction algorithm. Let γ be an arbitrary geodesic on $\mathcal{H}$, with endpoints u and w, and $w = (n_0, n_1, n_2, \dots)$. We construct the sequence of real pairs $\{(u_k, w_k)\}$ $(k \geq 0)$ defined by $u_0 = u$, $w_0 = w$ and

$$w_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} w\,, \quad u_{k+1} = ST^{-n_k} \dots ST^{-n_1} ST^{-n_0} u\,.$$

Each geodesic with endpoints u_k and w_k is $PSL(2,\mathbb{Z})$-equivalent to γ by construction.

Theorem 22.3. *The above algorithm produces in finitely many steps an α-reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, i.e. there exists a positive integer ℓ such that the geodesic with end points u_ℓ and w_ℓ is α-reduced.*

To an α-reduced geodesic γ, we associate a bi-infinite sequence of integers

$$(\gamma) = (\dots, n_{-2}, n_{-1}, n_0, n_1, n_2, \dots)$$

called its *arithmetic code*, by juxtaposing the α-expansions of $w = (n_0, n_1, n_2, \dots)$ and $1/u = (n_{-1}, n_{-2}, \dots)$ (for the H-code we need to use the dual H-expansion of $1/u$).

REMARK. Any further application of the reduction algorithm to an α-reduced geodesic yields α-reduced geodesics whose codes are left shifts of the code of the initial α-reduced geodesic.

The proof of Theorem 22.3 follows the same general scheme for each code, but the notion of reduced geodesic is different in each case, and so are the properties of the corresponding expansions and estimates.

Now we associate to any oriented geodesic γ in $\mathcal{H}$ the α-code of a reduced geodesic $PSL(2,\mathbb{Z})$-equivalent to γ, which is obtained by the reduction algorithm described above.

THEOREM 22.4. *Each geodesic γ in $\mathcal{H}$ is $PSL(2,\mathbb{Z})$-equivalent to an α-reduced geodesic ($\alpha = G, A, H$). Two reduced geodesics γ and γ' in $\mathcal{H}$ having arithmetic codes $(\gamma) = (n_i)_{i=-\infty}^{\infty}$ and $(\gamma') = (n'_i)_{i=-\infty}^{\infty}$ are $PSL(2,\mathbb{Z})$-equivalent if and only if for some integer l and all integers i one has $n'_i = n_{i+l}$.*

23. Reduction theory conjecture

The notion of a "reduced" geodesic can be explained by studying a map on the boundary $\mathbb{R}\cup\{\infty\}$ associated with (a,b)-continued fraction expansion and its (natural) extension.

Let

$$\rho_{a,b,x} = \rho_x = \begin{cases} T & \text{if } x < a \\ S & \text{if } a \le x < b \\ T^{-1} & \text{if } x \ge b \end{cases}$$

($T(x) = x+1$ and $S(x) = -\frac{1}{x}$ are the generators of $SL(2,\mathbb{Z})$), and $f_{a,b} : \mathbb{R} \to \mathbb{R}$ be given by $f_{a,b}(x) = \rho_x(x)$.

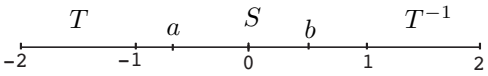


FIGURE 23.1. The action of $f_{a,b}$ on $\mathbb{R}$

Let $R_{a,b} : \mathbb{R}^2\setminus D \to \mathbb{R}^2\setminus D$ ($D = \{(u,w) \in \mathbb{R}^2 \mid u = w\}$ is the "diagonal") be a (natural) extension map of $f_{a,b}$ given by

$$R_{a,b}(u,w) = (\rho_w(u), \rho_w(w)).$$

It may be regarded as a map on oriented geodesics: R maps a geodesic from u to w to a geodesic from $\rho_w(u)$ to $\rho_w(w)$.

CONJECTURE 23.1. *The boundary map $f_{a,b}$* has a reduction theory*, i.e*

(a) *The extension map $R_{a,b}$ has an* attractor*, i.e. the smallest set $A_{a,b} \subsetneq \mathbb{R}^2$ such that*

(i) for (almost) every $(u,w) \in \mathbb{R}^2$ $\exists$ $N \ge 0$ such that $R^N(u,w) \in$

$A_{a,b}$;
(ii) if $(u,w) \in A_{a,b}$, *then for any* $n > 0$ $R^n_{a,b}(u,w) \in A$.

(b) *The map* $R_{a,b} : A_{a,b} \to A_{a,b}$ *is a bijection.*

(c) *The attractor* $A_{a,b}$ *consists of two (or one, in degenerate cases) connected components each having* finite rectangular structure, *i.e. bounded by step-functions with a finite number of steps.*

(d) *The orbit of any* $(u,w) \in A_{a,b}$ *returns to the subset* $\Lambda_{a,b} = R_{a,b}(A_{a,b} \cap \{a \le w \le b\})$.

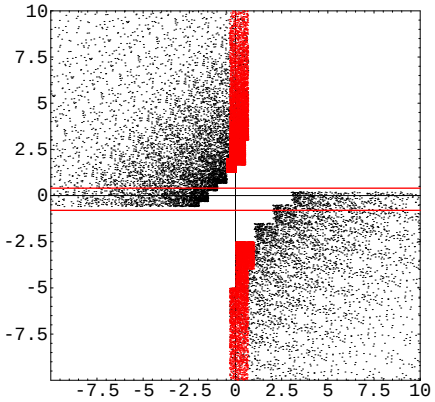


FIGURE 23.2. A typical attractor

Besides the classical cases, this conjecture has been proved for a large part of the parameter set Δ. A typical attractor $A_{a,b}$ $(a = -\frac{4}{5}, b = \frac{2}{5})$ is shown on Figure 23.2 with $\Lambda_{a,b}$ shown in red (dark grey).

DEFINITION 23.2. A geodesic in $\mathcal{H}$ from u to w is called (a,b)-*reduced* if $(u,w) \in \Lambda_{a,b}$.

The cross-section. If $(a,b) \in \Delta$, the (a,b)-reduced geodesic from u to w intersects the unit half-circle, and we define the *cross-section* as a subset of SM, $C_{a,b} = P \cup Q_1 \cup Q_2$, where P consists of all tangent vectors with base points in the circular boundary of F and pointing inward such that the corresponding geodesic is (a,b)-reduced; Q_1 consists of all tangent vectors with base points on the right vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $TS(\gamma)$ is (a,b)-reduced; Q_2 consists of all tangent vectors with base points on the left vertical side of F pointing inwards, such that if γ is the corresponding geodesic, then $T^{-1}S(\gamma)$ is (a,b)-reduced (see Figure 23.3). Thus $\Lambda_{a,b}$ is a (u,w)-parametrization of $C_{a,b}$. For the three classical expansions we obtain exactly the (u,w)-parametrization of the corresponding cross-sections given by the inequalities of Definition 22.2. It is easy to see that for the G-code Q_2 and the left half of P are absent.

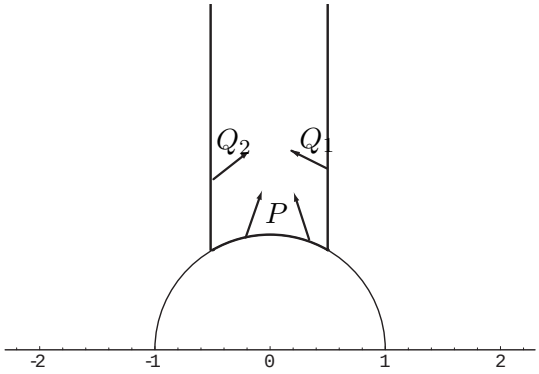
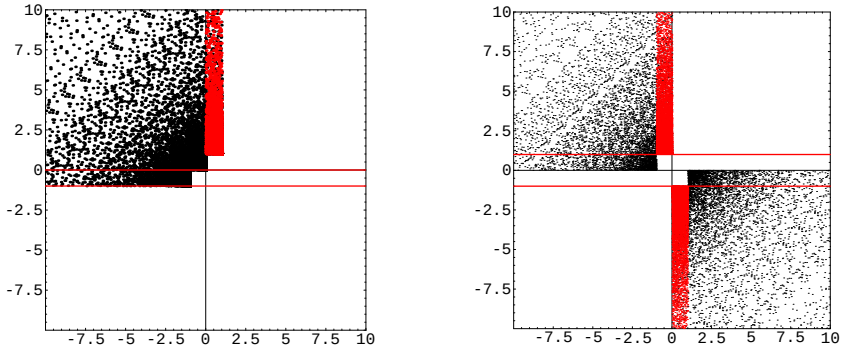


FIGURE 23.3. The cross-section $C_{a,b} = P \cup Q_1 \cup Q_2$

The dual expansions. We say that the (a, b)-expansion has a *dual* if



$(-1, 0)$ $(-1, 1)$

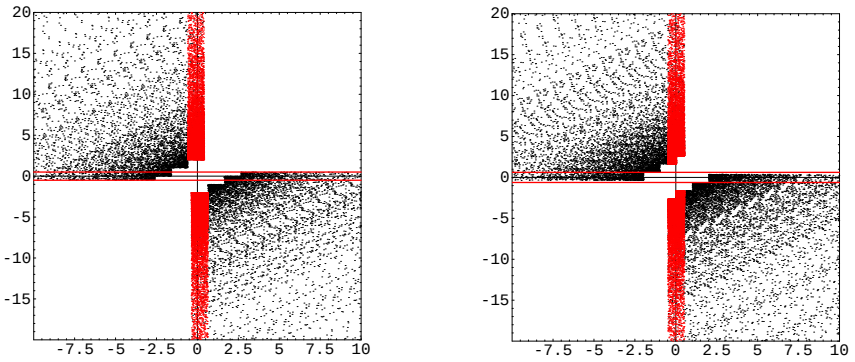
FIGURE 23.4. Attractors for self-dual Gauss and Artin expansions

the reflection of $A_{a,b}$ in the line $w = -u$ is an attractor for some (a', b')-expansion. If $(a', b') = (a, b)$, the (a, b)-expansion is called *self-dual*. Figure 23.4 shows the attractors for Gauss and Artin codes that are self-dual, and Figure 23.5 shows the attractors for the Hurwitz expansion and its dual.

Every oriented geodesic γ on M can be represented as a bi-infinite sequence of segments σ_i between successive returns to $C_{a,b}$. To each segment σ_i we associate the corresponding (a, b)-reduced geodesic γ_i on $\mathcal{H}$. Thus we obtain a sequence of reduced geodesics $\{\gamma_i\}_{i=-\infty}^{\infty}$ representing the geodesic γ. If one associates to γ_i its (a, b)-code, obtained by juxtaposing of expansion for $w = (n_0, n_1, \dots)_{a,b}$ and (dual) expansion for $1/u = (n_{-1}, n_{-2}, \dots)_{a',b'}$,

$$(\gamma) = (\dots n_{-2}, n_{-1}, n_0, n_1, n_2, \dots),$$

then $\gamma_{i+1} = ST^{-n_0}(\gamma_i)$ and the coding sequence is shifted one symbol to the left. Thus all (a, b)-reduced geodesics γ_i in the sequence produce the same, up to a shift, bi-infinite coding sequence, which we call the (a, b)-*code* of γ



$(-\frac{1}{2}, \frac{1}{2})$ $(r-1, 1-r)$

FIGURE 23.5. Attractors for the Hurwitz expansion and its dual

and denote by (γ). The left shift of the sequence corresponds to the return of the geodesic to the cross-section $C_{a,b}$.

EXAMPLE E. Let γ be a geodesic on $\mathcal{H}$ from $u = \sqrt{5}$ to $w = -\sqrt{3}$. The G-expansions are

$$w = \lceil -1, 2, \overline{2,3} \rceil\,, \quad 1/u = \lceil 1, \overline{2,6,2,2} \rceil\,.$$

First, we need to find an equivalent G-reduced geodesic. For this we use the reduction algorithm described above for G-expansions and construct the sequence (u_1, w_1), $(u_2, w_2), \dots$, until we obtain a G-reduced pair equivalent to (u, w). We have

$$w_1 = ST(w) = (1+\sqrt{3})/2, \quad u_1 = ST(u) = (1-\sqrt{5})/4\,,$$
$$w_2 = ST^{-2}(w_1) = 1 + 1/\sqrt{3}, \; u_2 = ST^{-2}(u_1) = (7-\sqrt{5})/11$$

and the pair (u_2, w_2) is already G-reduced. The G-expansions of $1/u_2$ and w_2 are

$$w_2 = \lceil \overline{2,3} \rceil\,, \quad 1/u_2 = \lceil 3, \overline{2,2,6,2} \rceil\,,$$

hence $\lceil \gamma \rceil = \lceil \overline{2,6,2,2}, 3, \overline{2,3} \rceil = \lceil \dots, 2,2,6,2,2,2,6,2,2,3,2,3,2,3,2,3 \dots \rceil$.

24. Symbolic representation of geodesics via arithmetic codes

Let $\mathcal{N}_G^{\mathbb{Z}}$ be the Bernoulli space on the infinite alphabet $\mathcal{N}_G = \{n \in \mathbb{Z} \mid n \geq 2\}$. We proved that each oriented geodesic which does not go to the cusp of M in either direction admits a unique G-code, $\lceil \gamma \rceil \in \mathcal{N}_G^{\mathbb{Z}}$ which does not contain a tail of 2's. Taking the closure of the set of such G-codes we obtain the entire space $\mathcal{N}_G^{\mathbb{Z}}$. Now, each bi-infinite sequence $x \in \mathcal{N}_G^{\mathbb{Z}}$ produces a geodesic on $\mathcal{H}$ from $u(x)$ to $w(x)$, where

$$w(x) = \lceil n_0, n_1, \dots \rceil\,, \; \frac{1}{u(x)} = \lceil n_{-1}, n_{-2}, \dots \rceil\,. \tag{24.1}$$

Notice that if a sequence has a tail of 2's then the oriented geodesic goes to the cusp. Thus the set of all oriented geodesics on M can be described symbolically as the Bernoulli space $X_G = \mathcal{N}_G^{\mathbb{Z}}$.

For the A-code, the set of all oriented geodesics (which do not go to the cusp) on M can be described symbolically as a countable one-step Markov chain $X_A \subset \mathcal{N}_A^{\mathbb{Z}}$ with the infinite alphabet $\mathcal{N}_A = \{n \in \mathbb{Z} \mid n \neq 0\}$ and transition matrix A,

$$A(n,m) = \begin{cases} 1 & \text{if } nm < 0, \\ 0 & \text{otherwise}. \end{cases} \tag{24.2}$$

For the H-code, recall first that the reduction algorithm and Theorem 22.4 are valid only for geodesics whose end points are not equivalent to r. Taking the closure of the set of all such H-codes, we obtain a set X_H containing also the bi-infinite sequences with a tail of 3's or -3's. These exceptional sequences are H-codes of some geodesics with one of the endpoints equivalent to r, but not of all such geodesics. More precisely, each exceptional geodesic with u equivalent to r has two H-codes (see Figure 24.1 for the only closed such geodesic), but not all exceptional geodesics with w equivalent to r can be H-reduced (see [**12**]).

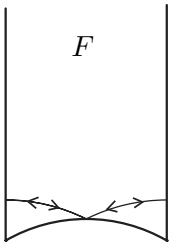


FIGURE 24.1. An exceptional geodesic with two H-codes, $\langle \overline{3} \rangle$ and $\langle \overline{-3} \rangle$

The set X_H is a countable one-step Markov chain $X_H \subset \mathcal{N}_H^{\mathbb{Z}}$ with infinite alphabet $\mathcal{N}_H = \{n \in \mathbb{Z} \mid |n| \geq 2\}$ and transition matrix H,

$$H(n,m) = \begin{cases} 0 & \text{if } |n| = 2 \text{ and } nm > 0, \\ 1 & \text{otherwise}. \end{cases} \tag{24.3}$$

Therefore, for $\alpha = G, A, H$, the space X_α is a closed shift-invariant subset of $\mathcal{N}_\alpha^{\mathbb{Z}}$.

Coding maps for arithmetic codes. As shown above, the coding map for each arithmetic α-code ($\alpha = G, A$), $\mathfrak{C}_\alpha : X_\alpha \to C_\alpha$ is a bijection between the cross-section C_α and the symbolic space $X_\alpha \subset \mathcal{N}_\alpha^{\mathbb{Z}}$. The map $\mathfrak{C}_H : X_H \to C_H$ is surjective, and essentially one-to-one: the only exception

is given by the H-codes corresponding to geodesics whose repelling endpoints are equivalent to r; for these exceptional H-codes the map is two-to-one.

The product topology on $\mathbb{N}_\alpha^{\mathbb{Z}}$ is induced by the distance function

$$d(x, x') = \frac{1}{m},$$

where $x = (n_i), x' = (n_i') \in \mathbb{N}_\alpha^{\mathbb{Z}}$, and $m = \max\{k \mid n_i = n_i' \text{ for } |i| \leq k\}$.

PROPOSITION 24.1. *The map $\mathfrak{C}_\alpha$ is continuous.*

PROOF. If $d(x, x') < \frac{1}{m}$, then the α-expansions of the attracting endpoints $w(x)$ and $w(x')$ of the corresponding geodesics given by (24.1) have the same first m digits. Hence the first m convergents of their α-expansions are the same, and by (22.4) and (22.3) $|w(x)-w(x')| < \frac{1}{m}$. Similarly, the first m digits of $\frac{1}{u(x)}$ and $\frac{1}{u(x')}$ are the same, and hence $|u(x)-u(x')| < \frac{u(x)u'(x)}{m} < \frac{1}{m}$. Therefore the geodesics are uniformly $\frac{1}{m}$-close. But the tangent vectors $v(x), v(x') \in C_\alpha$ are determined by the intersection of the corresponding geodesic with the unit circle. Hence, by making m large enough we can make $v(x')$ as close to $v(x)$ as we wish. □

Coding via (a, b)-continued fractions. If an (a, b)-expansion has a dual, one can code geodesics on the modular surface, as was explained in §23. The coding map $\mathfrak{C}_{\mathrm{a,b}} : X_{a,b} \to C_{a,b}$ is essentially a bijection and continuous. However, it is not known whether the set of admissible coding sequences $X_{a,b} \subset \mathbb{N}^{\mathbb{Z}}$ is always Markov.

The partition of the cross-section C_α. We parameterize the lift of the cross-section C_α to $S\mathcal{H}$, C_a by the coordinates (ϕ, θ), where $\phi \in [0, \pi]$ parameterizes the unit semicircle (counterclockwise) and $\theta \in [-\pi/2, (3\pi)/2]$ is the angle the unit vector makes with the positive horizontal axis (counterclockwise). The angle θ depends on ϕ and is determined by the condition that the corresponding geodesic is α-reduced.

The elements of the partition of C_a are labeled by the symbols of the corresponding alphabet $\mathcal{N}_\alpha$, $C_a = \bigcup_{n \in \mathbb{N}_\alpha} C_n$ and are defined by the following condition: C_n consists of all tangent vectors v in C_a such that for the coding sequence of the corresponding geodesic in $\mathcal{H}$, $n_0(x) = n$. The partitions of C_a (and therefore of C_α by projection) corresponding to the α-code ("the horizontal element") and their iteration under the first return map R to the cross-section C_a ("the vertical element") were obtained in [**18**], and are shown on Figures 24.2 and 24.3.

If we parameterize the cross-section C_a by using the coordinates u, w and the inequalities given in Definition 22.2, as was explained in §23, the pictures become even simpler, each element of the partition is a rectangle. We have chosen the coordinates (ϕ, θ) here to be consistent with the parametrization of the cross-section associated to the geometric code in §21.

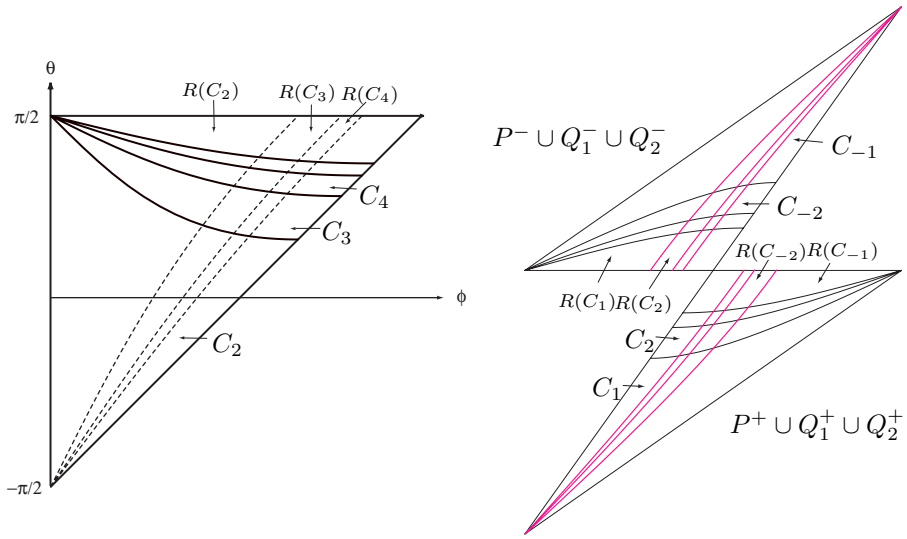


FIGURE 24.2. Infinite partition for the G-code (A-code, respectively) and its image under the return map R

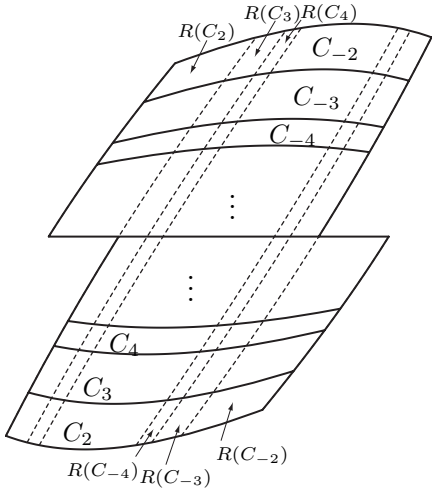


FIGURE 24.3. Infinite partition for the H-code and its image under the return map R

Some results of this section can be illustrated geometrically since the Markov property of the partition is equivalent to the Markov property of the shift space: the symbol m follows the symbol n in the coding sequence if and only if $R(C_n) \cap C_m \neq \emptyset$, and since all intersections are transversal, according to [**1**, Theorem 7.9], each partition is Markov.

25. Complexity of the geometric code

Deciding which bi-infinite sequences of nonzero integers are admissible geometric codes is a nontrivial task. We present some known classes of such admissible sequences, and show that the space X of all geometric codes is not a topological Markov chain.

The arithmetic codes we considered in §24 provide partial results: by identifying certain classes of geometric codes which coincide with arithmetic codes we obtain classes of admissible geometric codes. The first result of this kind was obtained in [**11**]:

THEOREM 25.1. *A bi-infinite sequence of positive integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *is an admissible geometric code if and only if*

$$\frac{1}{n_i} + \frac{1}{n_{i+1}} \le \frac{1}{2} \quad \text{for all } i \in \mathbb{Z}\,. \tag{25.1}$$

The corresponding geodesics are exactly those for which geometric codes coincide with G-codes.

The pairs forbidden by Theorem 25.1, $\{2,p\}$, $\{q,2\}$, $\{3,3\}$, $\{3,4\}$, $\{4,3\}$, $\{3,5\}$, and $\{5,3\}$—we call them *Platonic restrictions*—are of Markov type. More precisely, the set of all bi-infinite sequences satisfying relation (25.1) can be described as a one-step countable topological Markov chain $X_P \subset \mathbb{N}_G^{\mathbb{Z}}$, with the alphabet $\mathbb{N}_G$ and transition matrix P,

$$P(n,m) = \begin{cases} 1 & \text{if } 1/n + 1/m \le 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{25.2}$$

Clearly, X_P is a shift-invariant subset of X.

The geodesics identified in Theorem 25.1 have the property that all their segments in F are positively (clockwise) oriented. Following [**11**] we call them *positive geodesics*, and the corresponding class of sequences *positive coding sequences*.

A wider class of admissible coding sequences, which includes the positive ones, has been identified in [**17**]:

THEOREM 25.2. *Any bi-infinite sequence of integers* $\{\dots, n_{-1}, n_0, n_1, n_2, \dots\}$ *such that*

$$\left|\frac{1}{n_i} + \frac{1}{n_{i+1}}\right| \le \frac{1}{2} \quad \text{for } i \in \mathbb{Z} \tag{25.3}$$

is realized as a geometric code of a geodesic on M.

The set of all bi-infinite sequences satisfying relation (25.3) can be described as a one-step countable topological Markov chain, with the alphabet $\mathbb{N} = \{n \in \mathbb{Z} \mid n \ne 0\}$ and transition matrix M,

$$M(n,m) = \begin{cases} 1 & \text{if } |1/n + 1/m| \le 1/2\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{25.4}$$

We denote the associated one-step Markov chain by X_M. Clearly, X_M is a closed shift-invariant subset of X.

Following [**17**] we call the admissible geometric coding sequences identified in Theorem 25.2 and the corresponding geodesics, *geometrically Markov*. In [**18**] we show that the H-code comes closest to the geometric code:

THEOREM 25.3. *For any geometrically Markov geodesic whose geometric code does not contain* 1*'s and* -1*'s, the* H*-code coincides with the geometric code.*

The set X_M is a σ-invariant subset strictly included in X. For example, $[5, 3, -2]$ is an admissible geometric code, obtained as the code of the closed geodesic corresponding to the axis of $T^5ST^3ST^{-2}S$ (see Figure 25.1), but it is not geometrically Markov. Moreover, the latter is also an example of a non-geometrically Markov geodesic for which geometric and H-codes coincide. A natural question would be to characterize completely the class of geodesics for which the two codes coincide.

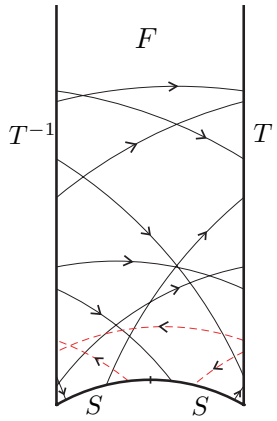


FIGURE 25.1. Geometric code $[5, 3, -2]$

The following theorems were proved in [**17**]:

THEOREM 25.4. *The set* X_M *is a maximal, transitive one-step countable topological Markov chain in the set of all geometric codes* X*.*

THEOREM 25.5. *The set* X_M *is the maximal symmetric (i.e. given by a symmetric transition matrix) one-step countable topological Markov chain in the set of all geometric codes* X*.*

The following result is an extension of a theorem proved in [**18**]:

THEOREM 25.6. *For any geometrically Markov geodesic whose geometric code consists of symbols with alternating signs, the* A*-code coincides with the geometric code.*

Unlike the spaces of admissible arithmetic codes X_G, X_A, and X_H which in §24 were proved to form topological Markov chains, the space of admissible geometric codes X is very complicated. In order to state the complexity

result proved in [**17**] we recall the notion of a k-step topological Markov chain defined on the alphabet $\mathbb{N}$ (see [**15**, §1.9] for the finite alphabet definition):

DEFINITION 25.7. Given an integer $k \geq 1$ and a map $\tau : \mathbb{N}^{k+1} \to \{0, 1\}$, the set

$$X_\tau = \{x \in \mathbb{N}^{\mathbb{Z}} \mid \tau(n_i, n_{i+1}, \ldots, n_{i+k}) = 1 \ \forall \ i \in \mathbb{Z}\}$$

with the restriction of the left-shift map σ to X_τ is called the *k-step topological Markov chain with alphabet $\mathbb{N}$ and transition map τ*.

Without loss of generality we always assume that the map τ is *essential*, i.e. $\tau(n_1, n_2, \ldots, n_{k+1}) = 1$ if and only if there exists a bi-infinite sequence in X_τ containing the $(k+1)$-block $\{n_1, n_2, \ldots, n_{k+1}\}$.

THEOREM 25.8. *The space X of geometric codes is not a k-step topological Markov chain, for any integer $k \geq 1$.*

The proof of this result is contained in [**16**].

26. Applications of arithmetic codes

Calculation of the return time for special flows. In §21 and §24 we have constructed four continuous surjective coding maps. The map $\mathfrak{C} : X \to B$ for the geometric code and the map $\mathfrak{C}_H : X_H \to C_H$ (for the H-code) are essentially one-to-one, (and finite-to-one everywhere) while the maps for the other two arithmetic codes, $\mathfrak{C}_\alpha : X_\alpha \to C_\alpha$ ($\alpha = G, A$) are bijections. In all cases the first return to the cross-section corresponds to the left-shift of the coding sequence. This provides four symbolic representations of the geodesic flow $\{\varphi^t\}$ on SM as a special flow over (Λ, σ), where $\Lambda = X_G, X_A, X_H, X$, with the ceiling function f being the time of the first return to the cross-section $C = C_G, C_A, C_H, B$, i.e. four symbolic representations of the geodesic flow on the space

$$\Lambda^f = \{(x, y) : x \in \Lambda, 0 \leq y \leq f(x)\} \tag{26.1}$$

as explained in §19.

For $\Lambda = X_G, X_A, X_H, X$ and $C = C_G, C_A, C_H, B$, respectively, the ceiling function $f(x)$ on Λ is the time of the first return of the geodesic $\gamma(x)$ to the cross-section C. The following theorem was proved in [**11**] for the G-code, and appeared for other arithmetic codes in [**18**], and for the geometric code in [**17**]. The same formula holds for all (a, b)-codes with $(a, b) \in \Delta$ as well. The proof for all codes is the same. A similar formula for Artin's original code has appeared earlier in [**24**].

THEOREM 26.1. *Let $x \in \Lambda$ and $w(x)$, $u(x)$ be the end points of the corresponding geodesic $\gamma(x)$. Then*

$$f(x) = 2 \log |w(x)| + \log g(x) - \log g(\sigma x),$$

where

$$g(x) = \frac{|w(x) - u(x)| \sqrt{w(x)^2 - 1}}{w(x)^2 \sqrt{1 - u(x)^2}}.$$

This formula was used to obtain topological entropy estimates in [**11**] and [**17**].

Factor-maps associated with arithmetic codes and their invariant measures. Let $(u, w) \in \Lambda_{a,b}$. Making a change of variables $x = -\frac{1}{w}$, $y = u$ we obtain a compact region $D_{a,b} \subset [a, b) \times [-1, 1]$. The reduction map in these coordinates

$$G_{a,b} : D_{a,b} \to D_{a,b}$$

is given by the formula

$$G_{a,b}(x, y) = \left(-\frac{1}{x} - \left(-\frac{1}{x} \right)_{a,b}, -\frac{1}{y - \left(-\frac{1}{x} \right)_{a,b}} \right).$$

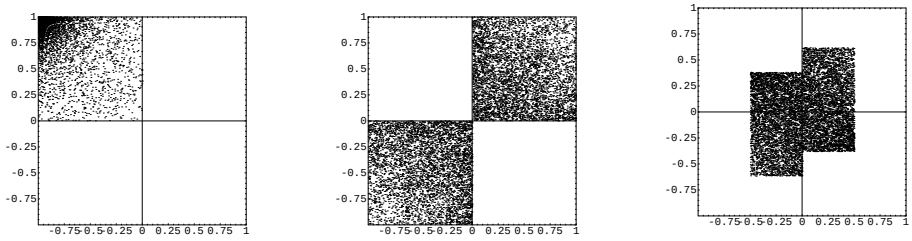
It may be considered a (natural) extension of the Gauss-type map $g_{a,b} : [a, b) \to [a, b)$,

$$g_{a,b}(x) = -\frac{1}{x} - \left(-\frac{1}{x} \right)_{a,b} ; \; g_{a,b}(0) = 0.$$

One sees immediately that the following diagram

$$\begin{array}{ccc} D_{a,b} & \xrightarrow{G_{a,b}} & D_{a,b} \\ \pi \downarrow & & \downarrow \pi \\ [a, b) & \xrightarrow{g_{a,b}} & [a, b) \end{array}$$

is commutative if $\pi(x, y) = x$.



$D_{-1,0}$ $D_{-1,1}$ $D_{-\frac{1}{2},\frac{1}{2}}$

FIGURE 26.1. The three classical attractors in compact form

In order to calculate the invariant measure for the map $g_{a,b}$ we use the parametrization of $S\mathcal{H}$ by (u, w, s), considered in [**2**]: a vector in $S\mathcal{H}$ is identified with (u, w, s), where u, w are the endpoints of the associated geodesic in $\mathcal{H}$, and s is the distance to a predetermined point on the geodesic (for example, the midpoint). In this parametrization the geodesic flow has a particularly simple form:

$$\varphi^t : (u, w, s) \mapsto (u, w, s + t). \tag{26.2}$$

The Liouville measure dv on $S\mathcal{H}$, introduced in §refs:first, in these coordinates is given by the formula

$$dv = \frac{du\,dw\,ds}{(w-u)^2}\,, \tag{26.3}$$

and its invariance under $\{\varphi^t\}$ follows immediately from (26.2) and (26.3). The measure on the cross-section $\Lambda_{a,b}$ invariant for the first return map is obtained by dropping ds: $dv_{\Lambda_{a,b}} = \frac{du\,dw}{(w-u)^2}$, and, by the above change of variables, the invariant measure on $D_{a,b}$ is given by $d\bar{v}_{D_{a,b}} = \frac{dx\,dy}{(1+xy)^2}$. The invariant measure on $[a,b)$ is obtained by integrating $d\bar{v}_{D_{a,b}}$ with respect to dy as explained in [**2**]. Thus, if we know the exact shape of the set $D_{a,b}$, we can calculate the invariant measure precisely.

G-code. In this case $g_G : [-1,0) \to [-1,0)$ is given by $g_G(x) = -\frac{1}{x} - \left\lceil -\frac{1}{x} \right\rceil$, $D_{-1,0} = [-1,0] \times [0,1]$, and the invariant measure for g_G on $[-1,0)$ is

$$d\mu_G = \frac{dx}{1+x}.$$

(See also [**3**] for a similar computation.)

A-code. In this case $g_A : [-1,1) \to [-1,1)$ is given by $g_A = -\frac{1}{x} - \left\lceil -\frac{1}{x} \right\rfloor$, $D_{-1,1} = \{[-1,0] \times [-1,0]\} \cup \{[0,1] \times [0,1]\}$, and the invariant measure is

$$d\mu_A = \left(\frac{\chi_{[-1,0]}}{1-x} + \frac{\chi_{[0,1]}}{1+x} \right) dx.$$

H-code. In this case $g_H : [-\frac{1}{2}, \frac{1}{2}) \to [-\frac{1}{2}, \frac{1}{2})$, $g_H = -\frac{1}{x} - \left\langle -\frac{1}{x} \right\rangle$, and the invariant measure is replace= by :

$$d\mu_H = \left(\frac{\chi_{[-\frac{1}{2},0]}}{(1+rx)(1+(r-1)x)} + \frac{\chi_{[0,\frac{1}{2}]}}{(1-rx)(1+(1-r)x)} \right) dx.$$

The formulae for $d\mu_A$ and $d\mu_H$ rectify the formulae given in [**16**].

Classical results proved using arithmetic codes. Artin [**4**] used regular continued fractions to prove the topological transitivity of the geodesic flow on the modular surface (i.e. the existence of a dense geodesic) and the density of closed geodesics. In fact, any Markov (a,b)-code, in particular, any arithmetic α-code ($\alpha = G, A, H$) can be used for this purpose since the Markov property allows us to list all admissible periodic coding sequences.

Exercises

34. Let $A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}, B = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}$, and $C = \begin{pmatrix} 0 & -1 \\ 1 & 4 \end{pmatrix}$. Use one of the classical continued fraction expansions to determine which of these matrices are conjugate in $PSL(2,\mathbb{Z})$.

35. Verify the formulae for the invariant measures $d\mu_G$, $d\mu_A$, $d\mu_H$.

References

1. R. Adler, *Symbolic dynamics and Markov partitions*, Bull. Amer. Math. Soc. **35** (1998), no. 1, 1–56.
2. R. Adler, L. Flatto, *Cross section maps for geodesic flows, I (The Modular surface)*, Birkhäuser, Progress in Mathematics (ed. A. Katok) (1982), 103–161.
3. R. Adler and L. Flatto, *The backward continued fraction map and geodesic flow*, Ergod. Th. & Dynam. Sys. **4** (1984), 487–492.
4. E. Artin, *Ein Mechanisches System mit quasiergodischen Bahnen*, Abh. Math. Sem. Univ. Hamburg **3** (1924), 170–175.
5. A. Beardon, *The Geometry of Discrete Groups*, Springer-Verlag, 1983.
6. D. Fried, *Reduction theory over quadratic imaginary fields*, J. Number Theory **110** (2005), no. 1, 44–74.
7. I.M. Gelfand, M.I. Graev, I.I. Pyatetskii-Shapiro, *Representation Theory and Automorphic Functions* (English translation). W.B. Saunders, Philadelphia, 1969.
8. D. J. Grabiner, J. C. Lagarias, *Cutting sequences for geodesic flow on the modular surface and continued fractions*, Monatsh. Math. **133** (2001), no. 4, 295–339.
9. V. Guillemin and D. Kazhdan, *O*n the cohomology of certain dynamical systems, Topology, **19** (1980), 291–299.
10. V. Guillemin and D. Kazhdan, *S*ome inverse spectral results for negatively curved 2-manifolds, Topology, **19** (1980), 301–312.
11. B. Gurevich, S. Katok, *Arithmetic coding and entropy for the positive geodesic flow on the modular surface*, Moscow Math. J. **1** (2001), no. 4, 569–582.
12. A. Hurwitz, *Über eine besondere Art der Kettenbruch-Entwicklung reeler Grossen*, Acta Math. **12** (1889) 367–405.
13. S. Katok, Fuchsian groups, University of Chicago Press, 1992
14. S. Katok, *Coding of closed geodesics after Gauss and Morse*, Geom. Dedicata **63** (1996), 123–145.
15. A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.
16. S. Katok and I. Ugarcovici, *Symbolic dynamics for the modular surface and beyond*, Bull. of the Amer. Math. Soc., **44**, no. 1 (2007), 87–132.
17. S. Katok, I. Ugarcovici, *Geometrically Markov geodesics on the modular surface*, Moscow Math. J. **5** (2005), 135–151.
18. S. Katok, I. Ugarcovici, *Arithmetic coding of geodesics on the modular surface via continued fractions*, 59–77, CWI Tract **135**, Math. Centrum, Centrum Wisk. Inform., Amsterdam, 2005.
19. P. Koebe, *Riemannsche Mannigfaltigkeiten und nicht euklidische Raumformen*, Sitzungsberichte der Preußischen Akademie der Wissenschaften, *I* (1927), 164–196; *II, III* (1928), 345–442; *IV* (1929), 414–557; *V, VI* (1930), 304–364, 504–541; *VII* (1931), 506–534.
20. A. Livshitz. Some homological properties of U -systems. Mat. Zametki **10** (1971), 555–564.
21. B. Maskit, On Poincaré's Theorem for fundamental polygons, Adv. Math. **7** (1971) 219–230.
22. M. Morse, *A one-to-one representation of geodesics on a surface of negative curvature*, Trans. Amer. Math. Soc. **22** (1921), 33–51.
23. H. Poincaré, Théorie des groupes Fuchsiens, Acta Math. 1 (1882) 1–62.
24. C. Series, *The modular surface and continued fractions*, J. London Math. Soc. (2) **31** (1985), 69–80.
25. C. Series, *Geometrical Markov coding of geodesics on surfaces of constant negative curvature*, Ergod. Th. & Dynam. Sys. **6** (1986), 601–625.

26. G. Springer, *Introduction to Riemann Surfaces*, 2nd ed., Chelsea Publ. Co., New York, 1981.
27. D. Zagier, *Zetafunkionen und quadratische Körper: eine Einführung in die höhere Zahlentheorie*, Springer-Verlag, 1982.

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802
E-mail address: katok_s@math.psu.edu