11

More Dynamics on Quotients of the Hyperbolic Plane

In addition to the geodesic flow, whose study we started in Chapter 9, we introduced the natural action of $\mathrm{PSL}_2(\mathbb{R})$ on $X = \Gamma \backslash \mathrm{PSL}_2(\mathbb{R})$. In this chapter we will study this natural action in more detail. We will show ergodicity[93] of the horocycle flow, mixing of $\mathrm{PSL}_2(\mathbb{R})$, and go on to deduce from the mixing property of the geodesic flow an "almost unique ergodicity" property for the horocycle flow, which we will refer to as an instance of *rigidity of invariant measures*. Finally, we shall use this together with the ergodic decomposition to establish equidistribution for individual orbits of the horocycle flow.

In many ways the horocycle flow is complementary to the geodesic flow considered in Chapter 9. It has already featured in the proof of ergodicity for the geodesic flow, and this link between the two flows will become stronger in this chapter, where we will use them alternately to prove stronger and stronger statements about both flows. We will see that despite this close linkage, the two flows have fundamentally different dynamical behaviors – indeed they are in many senses opposite extremes. For instance, as discussed in Section 9.7, the geodesic flow has an abundance of invariant measures, while (as already mentioned) we will see that the horocycle flow exhibits rigidity of invariant measures.

In this chapter we will switch back and forth between a geometric and an algebraic point of view. For the latter we will consider the slightly more general setting of quotients of $\mathrm{SL}_2(\mathbb{R})$ instead of quotients of $\mathrm{PSL}_2(\mathbb{R})$.

11.1 Dirichlet Regions

In Chapter 9 we constructed a standard fundamental domain for the discrete group $\mathrm{PSL}_2(\mathbb{Z})$ in the group $\mathrm{PSL}_2(\mathbb{R})$, and used its geometry to understand the relationship between the geodesic flow and the Gauss map. Here we present a generalization[94] of Proposition 9.18 which will give a description of the geometry of fundamental domains for other Fuchsian groups. This description will be needed later in the discussion of dynamics on more general quotients.

Definition 11.1. *Let Z be a locally compact metric space carrying an action of a countable group Γ by homeomorphisms. The action is said to be* properly discontinuous *if for any compact set $P \subseteq Z$ the set $\{\gamma \in \Gamma \mid \gamma P \cap P \neq \varnothing\}$ is finite. A measurable set $F \subseteq Z$ is a* fundamental domain *if $|\Gamma z \cap F| = 1$ for all $z \in Z$. An open set $F \subseteq Z$ is called an* open fundamental domain *for the action if*

(1) *if $g_1 \neq g_2$ then $g_1 F \cap g_2 F = \varnothing$, and*
(2) $\bigcup_{g \in G} g\overline{F} = Z$.

Thus, for example, the interior of the set F in Proposition 9.18 is an open fundamental domain* for the natural action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{H}$.

Recall from Definition 9.17 that a Fuchsian group is a discrete subgroup $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$. Write d for the hyperbolic metric as in Section 9.1.

Lemma 11.2. *An infinite subgroup $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$ is a Fuchsian group if and only if its action on $\mathbb{H}$ is properly discontinuous.*

Proof. If Γ is not discrete then we may choose a sequence of elements (g_n) with $g_n \neq e$ for all $n \geqslant 1$ and $g_n \to e$ as $n \to \infty$. If P is a compact set containing an open set, then $g_n P \cap P \neq \varnothing$ for all large n, showing that the action of Γ is not properly discontinuous.

Conversely, assume that Γ is discrete. Then $\{g \in \Gamma \mid gP \cap P \neq \varnothing\}$ will be finite for any compact P if the set $B = \{g \in \mathrm{SL}_2(\mathbb{R}) \mid gP \cap P \neq \varnothing\}$ is compact. (This follows easily from Exercise 9.2.2, but we give a concrete argument here which effectively solves it.) Since the set P is compact, B is certainly closed, so it is enough to show that B is a bounded set in $\mathrm{SL}_2(\mathbb{R})$ when viewed as a subset of $\mathbb{R}^4$.

By compactness, there are constants $R, \varepsilon > 0$ such that every $w \in P$ has $|w| \leqslant R$ and $\Im(w) \geqslant \varepsilon$. It follows that if $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in B$ (and so $gz \in P$ for some $z \in P$), then $\left|\frac{az+b}{cz+d}\right| \leqslant R$ and $\Im\left(\frac{az+b}{cz+d}\right) = \frac{\Im(z)}{|cz+d|^2} \geqslant \varepsilon$ by equation (9.15). Thus

$$|cz + d|^2 \leqslant \frac{1}{\varepsilon}\Im(z) \leqslant \frac{R}{\varepsilon}$$

and

$$|az + b|^2 \leqslant R^2 |cz + d|^2 \leqslant \frac{R^3}{\varepsilon}.$$

Since z belongs to some fixed compact subset of $\mathbb{H}$, this readily implies that the coefficients in the matrices of B lie in a bounded subset of $\mathbb{R}^4$. □

* Just as in the discussion after Proposition 9.18, we will be interested in cases where F does not differ from a fundamental domain F' too much. In the case we will consider we will only need to take the union of F with some subset of the lower-dimensional boundary ∂F to obtain F'.

Definition 11.3. *Let Γ be an infinite Fuchsian group, and let $p \in \mathbb{H}$ be a point not fixed by any element of Γ other than the identity. Then the set*

$$D = D(p) = \{z \in \mathbb{H} \mid \mathsf{d}(z,p) < \mathsf{d}(z,\gamma p) \text{ for all } \gamma \in \Gamma \smallsetminus \{e\}\}$$

is called a Dirichlet region *for Γ.*

Notice that Dirichlet regions always exist; since a Fuchsian group is countable and any non-trivial element can only fix at most two points in $\mathbb{C}$ (indeed, at most one in $\mathbb{H}$), there must be points in $\mathbb{H}$ not fixed by any non-identity element. Moreover, a Dirichlet region is the intersection of the hyperbolic half planes

$$\{z \in \mathbb{H} \mid \mathsf{d}(z,p) < \mathsf{d}(z,\gamma p)\}$$

for each $\gamma \in \Gamma \smallsetminus \{e\}$ (see Lemma 11.4 for a justification of the terminology).

Lemma 11.4. *For any $\gamma \in \mathrm{PSL}_2(\mathbb{R})$ the open set*

$$D_\gamma = \{z \in \mathbb{H} \mid \mathsf{d}(z,p) < \mathsf{d}(z,\gamma p)\}$$

is the connected component of $\mathbb{H} \smallsetminus L_\gamma$ containing p, where L_γ is the geodesic in $\mathbb{H}$ defined by the equation

$$\mathsf{d}(z,p) = \mathsf{d}(z,\gamma p).$$

It follows that a Dirichlet region is connected and convex in the sense that the hyperbolic geodesic path joining any two points in D lies entirely inside D.

Proof. To see that L_γ is a geodesic and the description of D_γ is valid, notice that both depend only on the points p and γp. We may apply an isometry $g \in \mathrm{PSL}_2(\mathbb{R})$ to map those two points to $-r+\mathrm{i}$ and $r+\mathrm{i}$ respectively; choosing r suitably we can ensure that $\mathsf{d}(-r+\mathrm{i}, r+\mathrm{i}) = \mathsf{d}(p,\gamma p)$, and then the existence of g follows from Proposition 9.4. However, the set of points in $\mathbb{H}$ equidistant from the points $-r+\mathrm{i}$ and $r+\mathrm{i}$ is precisely the upper half of the y-axis. Clearly, D_γ is convex (which is again easily seen in the case $p = -r+\mathrm{i}$ and $\gamma p = r+\mathrm{i}$) and the intersection of convex sets is again convex. Therefore D is convex as claimed. □

Lemma 11.5. *Any Dirichlet region for an infinite Fuchsian group Γ is an open fundamental domain for the action of Γ on $\mathbb{H}$. The boundary of a Dirichlet region is made up of geodesic segments contained in geodesics defined by*

$$L_\gamma = \{z \in \mathbb{H} \mid \mathsf{d}(z,p) = \mathsf{d}(z,\gamma p)\}$$

for $\gamma \in \Gamma \smallsetminus \{e\}$.

PROOF. Let $D = D(p)$ be a Dirichlet region. Since the action of Γ is properly discontinuous by Lemma 11.2, we have that for any $z \in \mathbb{H}$ there are only finitely many $\gamma \in \Gamma$ with

$$\mathsf{d}(\gamma z, p) \leqslant \mathsf{d}(z, p) + 1,$$

say $\gamma_1, \ldots, \gamma_n$. Note that for $z' \in B_{1/2}(z)$, this list of elements will include those $\gamma \in \Gamma$ for which $\mathsf{d}(\gamma z', p) \leqslant \mathsf{d}(z', p)$. In particular, there is some $w \in \Gamma z$ with

$$\mathsf{d}(w, p) \leqslant \mathsf{d}(\gamma z, p) = \mathsf{d}(z, \gamma^{-1} p)$$

for all $\gamma \in \Gamma$. Without loss of generality, assume that $z = w$. If $z \in D$ then a point z' close to z belongs to $D_{\gamma_1} \cap \cdots \cap D_{\gamma_n}$ by Lemma 11.4, so that $z' \in D$ and hence D is open. If $z \notin D$ then z belongs to some of the boundaries of the sets $D_{\gamma_1}, \ldots, D_{\gamma_n}$ – assume that

$$z \in L_{\gamma_1} \cap \cdots \cap L_{\gamma_m} \cap D_{\gamma_{m+1}} \cap \cdots \cap D_{\gamma_n}.$$

Then by Lemma 11.4 the geodesic path ϕ joining z to p will have $\phi(0) = z \notin D$ but $\phi(t) \in D$ for $t \in (0, 1]$, so $z \in \overline{D}$.

To see property (1) of Definition 11.1, assume that points z and w in D have $z = \gamma w$ for some $\gamma \in \Gamma \smallsetminus \{e\}$. Then

$$\mathsf{d}(w, p) < \mathsf{d}(w, \gamma^{-1} p) = \mathsf{d}(z, p)$$

and

$$\mathsf{d}(z, p) < \mathsf{d}(z, \gamma p) = \mathsf{d}(w, p),$$

which is a contradiction. □

As before, we write

$$\overline{\mathbb{H}} = \mathbb{H} \cup \partial\mathbb{H}$$

for the union of $\mathbb{H}$ with its boundary $\partial\mathbb{H} = \mathbb{R} \cup \{\infty\}$.

Let D be a subset of $\mathbb{H}$. We will also sometimes write $\partial^{\overline{\mathbb{H}}} D$ and $\overline{D}^{\overline{\mathbb{H}}}$ for boundaries and closures taken in the set $\overline{\mathbb{H}}$. Given a finite set of points in $\overline{\mathbb{H}}$, we can define a convex polygon by successively taking points on geodesic paths connecting two vertices or points obtained earlier. The smallest set of points that can be used to define a given polygon is the set of vertices. Alternatively, given a unit-speed parametrization of a subset of ∂D, we will refer to the points where the parametrization is not locally along a geodesic as *vertices* of D. Moreover, any point of $\overline{D}^{\overline{\mathbb{H}}} \cap \partial\mathbb{H}$ will be called a vertex also.

Lemma 11.6. *The boundary of a Dirichlet region ∂D is the union of at most countably many connected components. Each connected component of ∂D is the image of a piecewise geodesic path $\phi : \mathbb{R} \to \mathbb{H}$. Either D is a convex polygon and this path periodically traverses ∂D, or any such path connects two (not necessarily distinct) points of (vertices of D in) $\partial\mathbb{H}$.*

PROOF. Fix some $R > 1$ and write B_R for the hyperbolic ball of radius R around p, the chosen point defining D. Then we may find finitely many elements $\gamma_0 = I, \gamma_1, \ldots, \gamma_n \in \Gamma$ such that if $z \in B_R$ and $\gamma(z) \in B_R$ for some $\gamma \in \Gamma$ then $\gamma = \gamma_i$ for some $i, 0 \leqslant i \leqslant n$ (since the action of Γ is properly discontinuous). This implies that

$$D \cap B_R = \bigcap_{i=1}^{n} D_{\gamma_i} \cap B_R$$

and

$$(\partial D) \cap B_R = \left(\partial \bigcap_{i=1}^{n} D_{\gamma_i} \right) \cap B_R.$$

It follows by induction on n that $\partial \bigcap_{i=1}^{n} D_{\gamma_i}$ can be described as in the lemma. If for some $R > 0$ we have $D \subseteq B_R$ then $D = \bigcap_{i=1}^{n} D_{\gamma_i}$ is a convex polygon.

If not, then letting $R \to \infty$ (increasing n as needed) we see that for every R the boundary $(\partial D) \cap B_R$ is as claimed. Notice that as R increases, the number of connected components of $(\partial D) \cap B_R$ can increase (when a new D_{γ_i} is needed to describe $D \cap B_R$ and its boundary does not connect to the previous boundary pieces) or decrease (when the boundary pieces contained in some L_{γ_i} and some L_{γ_j} connect).

Fix some $z \in \partial D$ and let ϕ_R be a piecewise geodesic parametrization of the boundary component of $\bigcap_{i=1}^{n} D_{\gamma_i}$ normalized to have $\phi_R(0) = z$. As $R \to \infty$ the paths ϕ_R converge to some path ϕ – indeed for fixed t, $\phi_R(t) = \phi(t)$ when R is sufficiently large. After applying some isometry we may assume without loss of generality that $z = \mathrm{i}$, part of the boundary of D is a segment in the imaginary axis, and $D \subseteq \{z \mid \Re(z) > 0\}$ (see Figure 11.1). Then it follows that $\Re(\phi(t))$ is eventually monotone, both for $t \to \infty$ and for $t \to -\infty$. Therefore, the limits $\lim_{t\to\infty} \phi(t)$ and $\lim_{t\to-\infty} \phi(t)$ taken in $\overline{\mathbb{H}}$ exist and belong to $\partial\mathbb{H}$. □

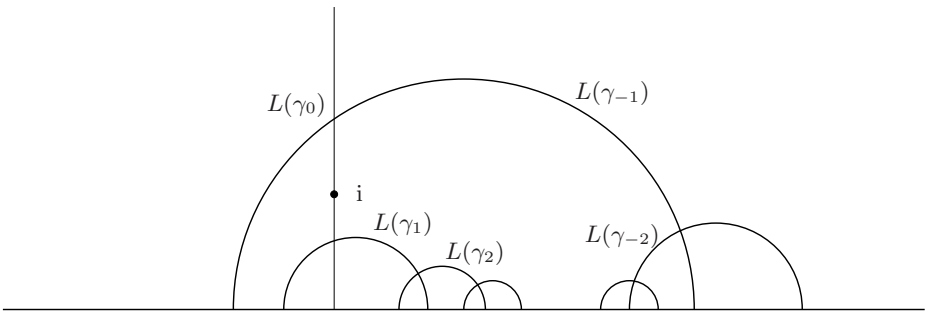

Fig. 11.1. A possible (and representative) scenario in the proof of Lemma 11.6.

As mentioned above, the sides of a Dirichlet region are segments of geodesics. A much less obvious result is that a Dirichlet region for a lattice in $\mathrm{PSL}_2(\mathbb{R})$ is in fact a hyperbolic polygon.

Theorem 11.7. *A Dirichlet region for a lattice in* $\mathrm{PSL}_2(\mathbb{R})$ *has finitely many sides. That is, it is a convex hyperbolic polygon.*

Before proving this result, we show how the hyperbolic area form from Lemma 9.16 gives a simple classical formula for the hyperbolic area of a polygon (a region bounded by finitely many geodesics).

Proposition 11.8. [Gauss–Bonnet formula] *Let P be a hyperbolic n-sided convex polygon in $\mathbb{H}$ with $n \geqslant 3$ vertices in $\overline{\mathbb{H}}$, with angles $\alpha_1, \ldots, \alpha_n$ at the n vertices. Then the hyperbolic area of P is*

$$(n-2)\pi - (\alpha_1 + \cdots + \alpha_n).$$

Here we measure angles between geodesics intersecting in $\mathbb{H}$ using the inner product at the intersection point; equivalently this is the angle in $\mathbb{C}$ between the circles (one of which might be a line) at the intersection point. For a vertex in $\partial\mathbb{H}$ we set the angle to be zero since the circles are tangential there.

Proof of Proposition 11.8. Assume for the purposes of an induction that the formula holds for all polygons with no more than $(n-1)$ sides, and let P be a polygon with n sides as in the statement of the lemma. By cutting off one triangle T (see Figure 11.2) to leave an $(n-1)$-gon Q we see that

$$\begin{aligned}\mathrm{Area}(P) &= \mathrm{Area}(Q) + \mathrm{Area}(T)\\ &= (n-3)\pi - (\alpha_1 + \cdots + \alpha_{n-3} + (\alpha_{n-2} - \beta_2) + (\alpha_{n-1} - \beta_1))\\ &\qquad\qquad +\pi - (\beta_1 + \beta_2 + \alpha_n)\\ &= (n-2)\pi - (\alpha_1 + \cdots + \alpha_n),\end{aligned}$$

showing that the result reduces to the case of a triangle.

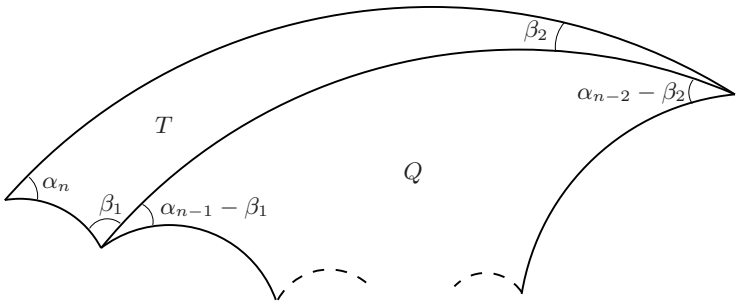

Fig. 11.2. Decomposing a hyperbolic n-gon into a triangle and an $(n-1)$-gon.

It is therefore enough to show that the formula holds when $n = 3$, so let T be a triangle with angles $\alpha_1, \alpha_2, \alpha_3$. If T has a vertex z_1 at infinity (this means the corresponding angle is zero), then we may apply a suitable transformation from $\mathrm{PSL}_2(\mathbb{R})$ to place the other two vertices on the unit circle (Lemma 9.16 shows that such a transformation preserves area) obtaining the triangle shown in Figure 11.3. Then

$$\mathrm{Area}(T) = \int_{\cos(\pi-\alpha_2)}^{\cos(\alpha_3)} \left(\int_{\sqrt{1-x^2}}^{\infty} \frac{\mathrm{d}y}{y^2} \right) \mathrm{d}x = \pi - (\alpha_2 + \alpha_3).$$

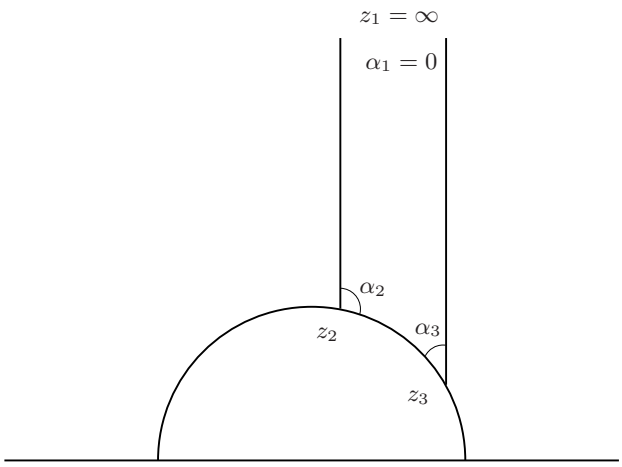

Fig. 11.3. A hyperbolic triangle with one vertex at infinity.

Now assume that all three vertices lie in $\mathbb{H}$ and assume without loss of generality that z_1 and z_2 do not have the same real part. Continue the geodesic through z_1 and z_2 to meet the real axis at w (a point at infinity), as shown in Figure 11.4. Then the triangle $z_1 z_3 w$ has interior angles $\pi - \alpha_1, \beta, 0$ while the triangle $z_2 z_3 w$ has interior angles $\alpha_2, \alpha_3 + \beta, 0$.

Thus

$$\begin{aligned} \mathrm{Area}(z_1 z_2 z_3) &= \mathrm{Area}(z_2 z_3 w) - \mathrm{Area}(z_1 z_3 w) \\ &= \pi - (\alpha_2 + \alpha_3 + \beta) - (\pi - (\pi - \alpha_1 + \beta)) \\ &= \pi - (\alpha_1 + \alpha_2 + \alpha_3), \end{aligned}$$

showing the formula holds for a triangle. □

Proof of Theorem 11.7. Recall that the Haar measure on $\mathrm{PSL}_2(\mathbb{R})$ is (under the isomorphism to $\mathrm{T}^1\mathbb{H}$) given by

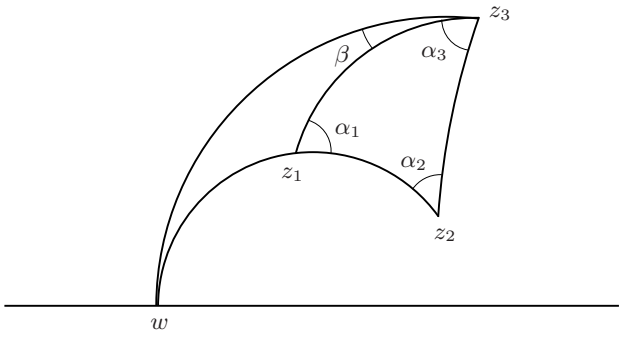

Fig. 11.4. A hyperbolic triangle with all vertices in $\mathbb{H}$.

$$\mathrm{d}m = \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}\theta,$$

where $\theta \in [0, 2\pi)$ corresponds to the angle of the unit vector, and

$$\mathrm{d}A = \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y$$

is the hyperbolic area form (see Lemma 9.16 and Proposition 9.19). This shows that a Dirichlet region $D = D(p)$ (which is a fundamental domain by Lemma 11.5) for a lattice Γ must have finite hyperbolic area.

Recall that the interior angle of a vertex at infinity is zero. We claim that the Gauss–Bonnet formula (Proposition 11.8) shows that the number of vertices at infinity cannot exceed $\frac{1}{\pi}\operatorname{Area}(D) + 2$, and in particular is finite. For if we had more points on the closure of D (taken in $\overline{\mathbb{H}}$) that lie on $\partial\mathbb{H}$ we could take such points $z_1, \ldots, z_n$ with $n > \frac{1}{\pi}\operatorname{Area}(D) + 2$ and consider the convex polygon P generated by them. By the Gauss–Bonnet formula we have $\operatorname{Area}(P) = (n-2)\pi > \operatorname{Area}(D)$. To obtain a contradiction, approximate each z_i by some $w_i \in D$, so that the resulting polygon Q generated by the w_i will satisfy $Q \subseteq D$ by convexity of D and have $\operatorname{Area}(Q) > \operatorname{Area}(D)$, which is impossible. Below we will ignore the boundaries for arguments like this one, and simply say that P is essentially contained in D.

We proceed by again using the Gauss–Bonnet formula to describe the boundary ∂D of D that is contained in $\mathbb{H}$. For example, it follows that the boundary of D lying in $\mathbb{H}$ can only have finitely many connected components. Let C be one of the finitely many connected components of ∂D. Pick an arbitrary point in C and a direction. Let $x_1, x_2, \ldots$ be the vertices in C along that chosen direction, and let $\omega_1, \omega_2, \ldots$ be the corresponding internal angles. Of course we would like to show that there can be only finitely many vertices. As a first step towards this, we claim that all but finitely many of the angles must be close to π. That is, ∂D is almost straight at most of its vertices.

Write T_k for the triangle with vertices x_k, x_{k+1}, p and write $\alpha_k, \beta_k, \varepsilon_k$ for the respective internal angles as shown in Figure 11.5. Thus $\omega_k = \beta_{k-1} + \alpha_k$

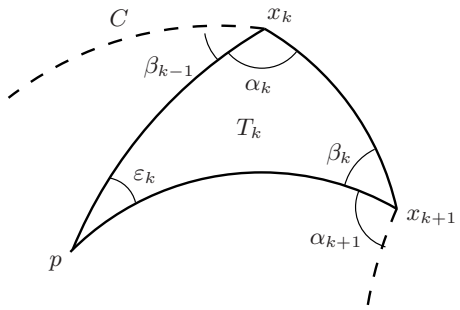

Fig. 11.5. A connected component C of $\partial(D)$.

for $1 \leqslant k$. Since D is convex, for any n the triangles $T_1, \ldots, T_{n-1}$ all lie essentially in D and are essentially disjoint, so

$$\begin{aligned}\sum_{k=1}^{n-1} \mathrm{Area}(T_k) &= \sum_{k=1}^{n-1} \pi - (\alpha_k + \beta_k + \varepsilon_k) \\ &= \pi - \alpha_1 - \beta_{n-1} + \sum_{k=2}^{n-1} (\pi - \omega_k) - \sum_{k=1}^{n-1} \varepsilon_k \\ &\leqslant \mathrm{Area}(D) < \infty.\end{aligned}$$

Now the angles ε_k are those for disjoint arcs at p, and so $\sum_{k=1}^{n-1} \varepsilon_k \leqslant 2\pi$. Therefore

$$\sum_{k=1}^{n-1} (\pi - \omega_k) \leqslant \alpha_1 + \beta_{n-1} + \pi + \mathrm{Area}(D)$$

is uniformly bounded. This establishes our claim. In particular, it follows that we can only have

$$(\pi - \omega_k) > \tfrac{\pi}{4}$$

for finitely many k, so

$$\left|\{k \mid \omega_k \leqslant \tfrac{3\pi}{4}\}\right| < \infty. \tag{11.1}$$

Since we have already established that there are only finitely many connected components of ∂D this holds across all vertices of ∂D. We will now use this claim twice to establish that there are only finitely many vertices.

We show next that any vertex of D has only finitely many other vertices of D in its Γ-orbit. Choose a vertex x_k of D in $\mathbb{H}$, let $\gamma_1 x_k = x_k, \gamma_2 x_k, \ldots$ (with $\gamma_1 = I$, $\gamma_i \in \Gamma$, and $\gamma_i x_k \neq \gamma_j x_k$ for $i \neq j$) be the vertices of D that lie on the orbit of x_k under the action of Γ, and let the internal angles of D at these vertices be $\delta_1, \delta_2, \ldots$.

For any $j \geqslant 1$, x_k is a vertex of $\gamma_j^{-1} D$ with internal angle δ_j. The sets

$$D = \gamma_1^{-1}D, \gamma_2^{-1}D, \dots$$

are pairwise disjoint since the Dirichlet region D is an open fundamental domain by Lemma 11.5. Hence

$$\delta_1 + \delta_2 + \cdots \leqslant 2\pi. \tag{11.2}$$

It follows from equation (11.1) that the number of vertices

$$x_k = \gamma_1 x_k, \gamma_2 x_k, \dots, \gamma_n x_k$$

on the Γ-orbit of x_k must be finite.

We finish the proof by showing that every class of these finite subsets of orbits described above has to contain at least one vertex whose internal angle is smaller than $\frac{2\pi}{3} < \frac{3\pi}{4}$. Together with the bound (11.1) this then implies that there are only finitely many vertices of D.

To do this, choose again a vertex x_k, let $\gamma_1, \dots, \gamma_n$ be as above, and let

$$\Gamma_k = \{\gamma \in \Gamma \mid \gamma x_k = x_k\}$$

be the stabilizer of x_k in Γ. The group Γ_k is conjugate to a discrete subgroup of the compact group PSO(2) and so must be finite. Using Γ_k we now refine equation (11.2) as follows. The images of D under the action of Γ which have x_k as a vertex are precisely those of the form $\gamma\gamma_i^{-1}D$ for all $\gamma \in \Gamma_k$ and $1 \leqslant i \leqslant n$. Since D is a fundamental domain, these are all disjoint and together they essentially cover a neighborhood of x_k, so

$$|\Gamma_k|(\delta_1 + \cdots + \delta_n) = 2\pi.$$

Since $0 < \delta_j < \pi$ for $1 \leqslant j \leqslant n$, we must have $n \geqslant 3$, and hence $\delta_j \leqslant \frac{2\pi}{3}$ for some j. As explained above, this implies the desired statement. □

The geometry of a Dirichlet region for a lattice detects whether or not the lattice is uniform.

Lemma 11.9. *A lattice $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$ is uniform (that is, $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$ is compact) if and only if every vertex of any Dirichlet region for Γ lies in $\mathbb{H}$ (that is, has compact closure).*

Proof. Let D be a Dirichlet region for Γ. If the boundary of D lies in $\mathbb{H}$, then the closure of D is a compact subset of $\mathbb{H}$. The compact subset

$$F = \{g \in \mathrm{PSL}_2(\mathbb{R}) \mid g(\mathrm{i}) \in \overline{D}\}$$

then maps continuously onto $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$, so Γ is uniform.

Conversely, assume that Γ is uniform. By Proposition 9.14 there is, for every $x \in X$ a neighborhood $B_r^X(x)$ which is the isometric image of the neighborhood $B_r^G(g)$ where $x = \Gamma g$. By compactness of $X = \Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$

there is a finite subcover of the open cover defined by the set $B^X_{r=r(x)}(x)$, and so there exists some bounded subset $B \subseteq \mathrm{PSL}_2(\mathbb{R})$ which maps onto X. Let $p \in \mathbb{H}$ be not fixed by any non-trivial element of Γ, and let $D = D(p)$ be the Dirichlet region defined by p. Then for any $z \in \mathbb{H}$ there is some $g \in \mathrm{PSL}_2(\mathbb{R})$ with $g(p) = z$, some $\gamma \in \Gamma$ with $\gamma g \in B$, and hence $\gamma(z) \in B(p)$. By the argument in the proof of Lemma 11.2 (that is, the properness of the $\mathrm{PSL}_2(\mathbb{R})$-action on $\mathbb{H}$), the set $B(p) \subseteq \mathbb{H}$ has compact closure. It follows that the vertices of D lie in $\mathbb{H}$. □

Let D be a Dirichlet domain for a lattice Γ in $\mathrm{PSL}_2(\mathbb{R})$. Points of the closure of D taken in $\overline{\mathbb{H}}$ that belong to $\partial\mathbb{H}$ (that is, those that are elements of $\overline{D}^{\overline{\mathbb{H}}} \cap \partial\mathbb{H}$) are called *cusps*. Thus, for example, if $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ then $\infty \in \partial\mathbb{H}$ is a cusp. More precisely, if there are several points in $\overline{D}^{\overline{\mathbb{H}}} \cap \partial\mathbb{H}$ then we identify points on the same Γ-orbit (for the natural action of $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$ on $\partial\mathbb{H} = \overline{\mathbb{R}}$) – cusps are then the equivalence classes. We will see an example where this distinction is important in the next section.

Exercises for Section 11.1

Exercise 11.1.1. Show that a lattice in $\mathrm{PSL}_2(\mathbb{R})$ is finitely generated. Hint: Choose a Dirichlet region D for the lattice, use Theorem 11.7 to show that there are only finitely many elements γ with $\overline{D} \cap \gamma\overline{D} \neq \varnothing$, and finally show that these elements must generate the lattice.

Exercise 11.1.2. Show that the fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$ in $\mathrm{PSL}_2(\mathbb{R})$ constructed in Section 9.4 is a Dirichlet domain.

11.2 Examples of Lattices

In order to avoid the impression that $\mathrm{PSL}_2(\mathbb{Z})$ is the only interesting lattice in $\mathrm{PSL}_2(\mathbb{R})$, in this section we will discuss some other lattices.

Notice first that the canonical map $\mathrm{SL}_2(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm I_2\}$ is two-to-one and the push-forward of the Haar measure on $\mathrm{SL}_2(\mathbb{R})$ under this map gives the Haar measure on $\mathrm{PSL}_2(\mathbb{R})$. As before, we may identify the Haar measure on $\mathrm{PSL}_2(\mathbb{R})$ with the measure m described in Section 9.4.1 (which we use to normalize the measure on $\mathrm{SL}_2(\mathbb{R})$). It follows that we can determine whether a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ is a lattice by analyzing a fundamental region for its action on $\mathbb{H}$. In particular, it follows that any lattice in $\mathrm{PSL}_2(\mathbb{R})$ gives a lattice in $\mathrm{SL}_2(\mathbb{R})$ by taking the pre-image under the map $\mathrm{SL}_2(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R})$.

11.2.1 Arithmetic and Congruence Lattices in $\mathrm{SL}_2(\mathbb{R})$

Notice that any discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ which contains a lattice is itself a lattice. Moreover, a finite-index subgroup Λ of a lattice Γ is also a lattice, since in this case the union of finitely many copies of a fundamental domain for Γ will form a fundamental domain for Λ. A *principal congruence lattice of* $\mathrm{SL}_2(\mathbb{R})$ is a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ of the form

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid a \equiv d \equiv 1, c \equiv b \equiv 0 \pmod{N} \right\}$$

for some $N \geqslant 1$. A *congruence lattice* is a lattice that contains a principal congruence lattice. A discrete subgroup Λ with the property that $\Lambda \cap \mathrm{SL}_2(\mathbb{Z})$ has finite index in both Λ and in $\mathrm{SL}_2(\mathbb{Z})$ is called an *arithmetic lattice of* $\mathrm{SL}_2(\mathbb{R})$ (equivalently, a lattice is called arithmetic if it has the property that $\Lambda \cap \mathrm{SL}_2(\mathbb{Z})$ is also a lattice).

We note that there are other arithmetic and congruence lattices that are not constructed from $\mathrm{SL}_2(\mathbb{Z})$ but from other types of integer lattices (see also Exercise 11.6.3).

Example 11.10. The subgroup $\Gamma(2)$ described above has index 6 in $\mathrm{SL}_2(\mathbb{Z})$ since $\mathrm{SL}_2(\mathbb{Z})/\Gamma(2) \cong \mathrm{SL}_2(\mathbb{F}_2)$ has order 6. The subgroup

$$\Gamma_0(2) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \pmod{2} \right\}$$

is a congruence lattice of index 3 in $\mathrm{SL}_2(\mathbb{Z})$. The index may be seen by applying the orbit-stabilizer theorem to the natural linear action of $\mathrm{SL}_2(\mathbb{Z})$ on the vector space $\mathbb{F}_2^2$; $\Gamma_0(2)$ is the stabilizer of $\begin{pmatrix} 1 \\ 0 \end{pmatrix} \in \mathbb{F}_2^2$, which has an orbit of size 3.

11.2.2 A concrete principal congruence lattice of $\mathrm{SL}_2(\mathbb{R})$

Let D be the convex 4-gon spanned by the points $-1, 0, 1, \infty \in \partial\mathbb{H}$ as in Figure 11.6.

Lemma 11.11. *The image of the lattice $\Gamma(2)$ (cf. Example* 11.10*) in* $\mathrm{PSL}_2(\mathbb{R})$ *is freely generated by* $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ *and* $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$*, and D (as in Figure* 11.6*) is its Dirichlet domain for the point $p = \mathrm{i}$.*

The proof that $\Gamma(2)/\{\pm I\}$ is a free group is based on a simple version of the so-called *ping-pong lemma.* Notice that the fact that $\Gamma(2)$ is a lattice with D as its fundamental region in $\mathbb{H}$ can also be deduced from Proposition 9.18 by analyzing $\Gamma(2)\backslash \mathrm{SL}_2(\mathbb{R})$. We will give an independent proof here.

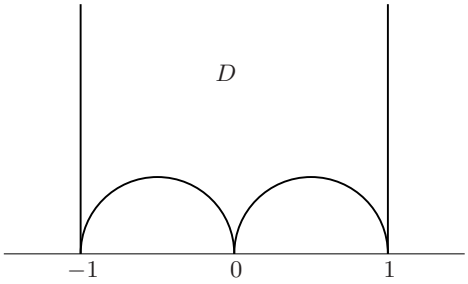

Fig. 11.6. The convex polygon D.

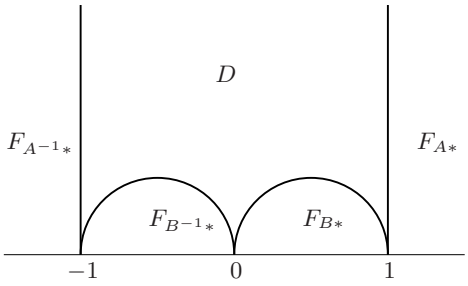

Fig. 11.7. Connected components of $\mathbb{H} \smallsetminus D$.

Proof of Lemma 11.11. We begin by analyzing the action of $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ on $\mathbb{H}$ with respect to the partition of $\mathbb{H}$ into D and the four connected components of $\mathbb{H} \smallsetminus D$, labeled as in Figure 11.7.

A calculation shows that

$$A\,(\mathbb{H} \smallsetminus F_{A^{-1}*}) \subseteq F_{A*}, \quad A^{-1}\,(\mathbb{H} \smallsetminus F_{A*}) \subseteq F_{A^{-1}*} \tag{11.3}$$

since the action of A on $\mathbb{H}$ is given by $A(z) = z + 2$. Similarly,

$$B\,(\mathbb{H} \smallsetminus F_{B^{-1}*}) \subseteq F_{B*}, \quad B^{-1}\,(\mathbb{H} \smallsetminus F_{B*}) \subseteq F_{B^{-1}*}. \tag{11.4}$$

This can be checked either by calculating the images of the geodesics (which may be done by finding the images of the end points of the geodesics in $\partial\mathbb{H}$) under B and B^{-1}, or by conjugating with $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, thereby reducing equation (11.4) to the case of equation (11.3).

Now let $w = A^{i_0} B^{j_1} A^{i_1} B^{j_2} \cdots A^{i_m} B^{\ell}$ be any reduced word in the generators A and B of $\Gamma_{r,s}$. We claim that the image $w(D)$ of D uniquely determines the exponents $i_0, j_1, i_1, j_2, \ldots, i_m, \ell \in \mathbb{Z}$, which implies that the group $\langle A, B \rangle$ is

freely generated. To prove the claim it is sufficient to show that $w(D) \subseteq F_{A^{\pm 1}*}$ if and only if the reduced word begins on the left with a positive power of $A^{\pm 1}$ respectively, and similarly for B. For words of length one (that is, for the generators A, B and their inverses), this follows from equation (11.3) and equation (11.4). Using these equations again in an induction argument then proves the claim.

By the Gauss–Bonnet formula (Proposition 11.8) the four-gon D has finite volume. We claim that D is a Dirichlet domain in $\mathbb{H}$ for the action of $\langle A, B\rangle$. Note that the vertical line $x = -1$ (or $x = 1$) coincides with the line L_γ in Lemma 11.5 where $\gamma = A^{-1}$ (or $\gamma = A$ respectively) for the point $p = \mathrm{i}$. Similarly, one may check that the other two geodesic boundaries of D are of the form L_γ for suitable γ. It follows that D is contained in the Dirichlet domain for $\langle A, B\rangle$ and $p = \mathrm{i}$. Moreover, the argument above shows that no two interior points of D are images of one another under the action of $\langle A, B\rangle$, so D is the Dirichlet region as required.

It remains to show that A, B and $-I$ together generate $\Gamma(2)$. This may be seen by the following version of the Euclidean algorithm. Let

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(2).$$

Notice that a is odd, c is even,

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a + 2c & b + 2d \\ c & d \end{pmatrix}, \tag{11.5}$$

and

$$\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ c + 2a & d + 2b \end{pmatrix}. \tag{11.6}$$

If $|a| > |c|$ then equation (11.5) can be used repeatedly to find $A^n\gamma = \begin{pmatrix} a' & b' \\ c & d \end{pmatrix}$ with $|a'| < |c|$. If $|c| > |a|$ then equation (11.6) can be used repeatedly to find $B^n\gamma = \begin{pmatrix} a & b \\ c' & d' \end{pmatrix}$ with $|c'| < |a|$. Iterating these two cases shows that there is an element $C \in \langle A, B\rangle$ such that $C\gamma = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $a = \pm 1$ and b even. Clearly $d = a$, and so $C\gamma = A^{b/2}(aI)$ as required. □

11.2.3 Uniform Lattices

In this section we outline two constructions of uniform lattices in $\mathrm{PSL}_2(\mathbb{R})$.

Lemma 11.12. *There is a uniform lattice in* $\mathrm{PSL}_2(\mathbb{R})$.

In the proof a convex n-gon will be called *regular* if all of its interior angles are equal and all its sides have the same length.

Sketch proof of Lemma 11.12. First notice that there is a regular four-gon D for which all of the internal angles are equal to $\frac{\pi}{3}$. To see this, draw two geodesics intersecting at i in a normal angle, and consider the four points on this pair of geodesics at distance t from i. For small values of t, the area of the regular four-gon spanned by these points is small, and the internal angles are close to $\frac{\pi}{2}$ (these are equivalent statements by the Gauss–Bonnet formula (Proposition 11.8)). As $t \to \infty$ these angles converge to zero, so for some t they must be equal to $\frac{\pi}{3}$ (it is clear that the angles vary continuously in t). In Figure 11.8 this is visualized in the *disc model*[(95)] of $\mathbb{H}$, which is obtained from $\mathbb{H}$ by applying an inversion with respect to a circle outside $\mathbb{H}$ but tangent to $\partial\mathbb{H}$ (for example, the map $\mathbb{H} \ni z \mapsto \frac{1}{z+\mathrm{i}} \in \mathbb{C}$).

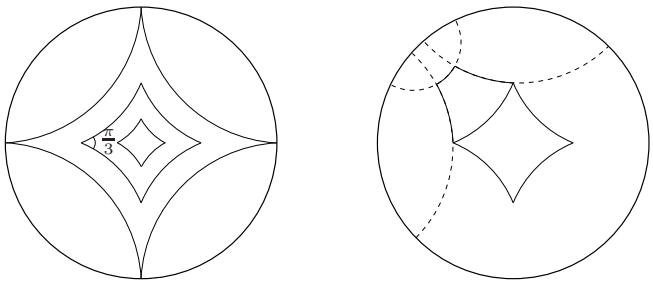

Fig. 11.8. A regular polyhedron and a tiling in the ball model for $\mathbb{H}$.

Now notice that 6 copies of D that are isometric under the action of $\mathrm{PSL}_2(\mathbb{R})$ can be put together edge-to-edge to cover a neighborhood of one of their vertices. By iterating this, we obtain a tiling of $\mathbb{H}$ by tiles isometric to D. That is, we end up with countably many isometric images of D with the property that any two of the images intersect either in one of the sides, or in one corner, or not at all. Notice that the same tiling can be constructed from any of the copies of D in the tiling by placing new copies of D edge-to-edge to already existing copies.

Now consider the group Γ of all matrices in $\mathrm{PSL}_2(\mathbb{R})$ that map the tiling onto itself. It is clear that Γ is discrete since the set of vertices of the copies of D in the tiling forms a discrete subset of $\mathbb{H}$ and must be mapped onto itself by Γ. Finally, any copy of D in the tiling can be mapped to any other by some element of Γ. It follows that D contains a fundamental region and so Γ is a uniform lattice. □

Another construction of co-compact lattices $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ comes from the uniformization theorem[(96)] which states that any connected Riemannian surface is a quotient of $\mathbb{H}$ or of $\mathbb{C}$, or it has a bijective holomorphic map to the

Riemann sphere $\overline{\mathbb{C}}$. Moreover, if the surface has genus two or more, it must be a quotient $\Gamma\backslash\mathbb{H}$. Thus it is enough to construct some Riemann surface of genus two, and one way to do this is illustrated in Figure 11.9.

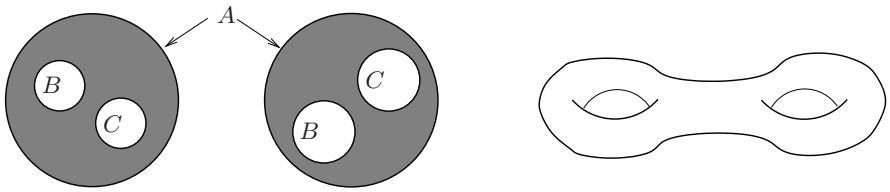

Fig. 11.9. A Riemann surface of genus two.

The Riemann surface is constructed by gluing the big circles labeled A together via the holomorphic inversion on the circle followed by the correct translation, and similarly for the smaller circles B and C. Strictly speaking, the manifold is defined by two charts which are small neighborhoods of the regions between the big and small circles and the chart maps. This produces a surface that topologically can be viewed as the surface of a three-dimensional body in the shape of a figure eight.

Another arithmetic construction of uniform lattices in $\mathrm{SL}_2(\mathbb{R})$ will be discussed in Exercise 11.6.3.

Exercises for Section 11.2

Exercise 11.2.1. Show rigorously that the tiling used in the sketch proof of Lemma 11.12 exists.

Exercise 11.2.2. Show that the *Hecke triangle group* G_n for $n \geqslant 3$, generated by

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and

$$T_n = \begin{pmatrix} 1 & 2\cos\frac{\pi}{n} \\ 0 & 1 \end{pmatrix}$$

is a non-uniform lattice in $\mathrm{SL}_2(\mathbb{R})$. Find the associated Dirichlet region for the point $p = 2\mathrm{i}$.

Exercise 11.2.3. This exercise shows how to contruct uncountably many non-conjugate lattices. Fix a parameter $x \in (-1, 1)$, and let

$$V_x = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t_x & 1 \end{pmatrix} \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix}.$$

The matrix V_x acts on $\mathbb{C}$ via Möbius transformations as in equation (9.1); verify that $V_x(x) = x$ for any t_x. Now add the requirement that $V_x(-1) = 1$, and check that this uniquely determines t_x in terms of x.

(a) Prove that Γ_x, the group generated by $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and V_x, is a lattice in $\mathrm{SL}_2(\mathbb{R})$, freely generated by $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ and V_x, and show that the domain illustrated in Figure 11.10 is a fundamental domain for Γ_x.

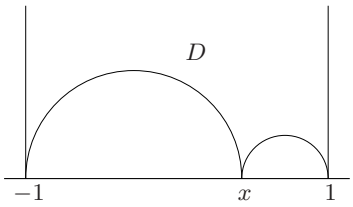

Fig. 11.10. A fundamental domain for Γ_x.

(b) Calculate the trace of the element

$$\gamma = V_x \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \in \Gamma_x,$$

and deduce that $\Gamma_x \backslash \mathrm{SL}_2(\mathbb{R})$ contains a periodic orbit, say of $\Gamma_x g$, for the geodesic flow associated to γ in the sense that $\gamma g = ga$ for some diagonal element a. Express the length of the period as a function of x.

(c) Show that for any lattice $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ there are only countably many periodic orbits for the geodesic flow on $\Gamma \backslash \mathrm{SL}_2(\mathbb{R})$.

(d) Show that if $\Gamma^g = g\Gamma g^{-1}$ is the conjugate of a lattice $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ by some element $g \in \mathrm{SL}_2(\mathbb{R})$, then the sets of lengths of the periodic orbits on $\Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ and on $\Gamma^g \backslash \mathrm{SL}_2(\mathbb{R})$ agree.

(e) Deduce that $\mathrm{SL}_2(\mathbb{R})$ has uncountably many non-conjugate lattices.

11.3 Unitary Representations, Mautner Phenomenon, and Ergodicity

The "Mautner phenomenon" has its origins in his study of geodesic flows [257]. The phenomenon refers to situations where invariance of an observable (a function) along orbits of one flow implies invariance along the orbits of certain transverse flows(97). An instance of this has already been seen in the proof of Theorem 9.21.

11.3.1 Three Types of Actions

For any discrete subgroup $\Gamma \leqslant \mathrm{SL}_2(\mathbb{R})$, the geodesic flow is defined on the space $X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$ by

$$R_{a_t}(x) = xa_t^{-1} = x \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

for $x \in X$ and $t \in \mathbb{R}$. The stable (and analogously unstable) horocycle flow is defined by

$$R_{u^-(s)}(x) = xu^-(-s) = x \begin{pmatrix} 1 & -s \\ 0 & 1 \end{pmatrix}$$

for $x \in X$ and $s \in \mathbb{R}$. This represents a unit-speed parametrization of the stable manifold of the geodesic flow g_t.

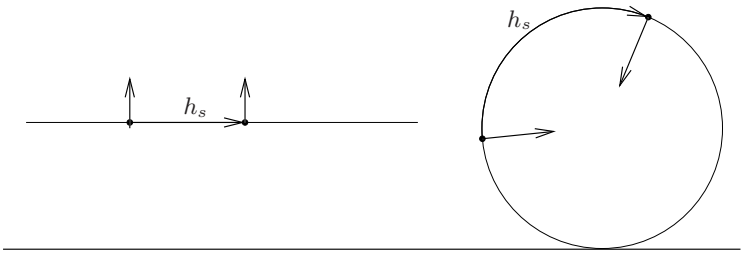

Fig. 11.11. The action of the horocycle flow.

Taken together, the subgroups

$$A = \left\{ \begin{pmatrix} e^{-t/2} & \\ & e^{t/2} \end{pmatrix} \mid t \in \mathbb{R} \right\}$$

and

$$U^- = \left\{ \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} \mid s \in \mathbb{R} \right\}$$

contain representatives of all but one type of element of $\mathrm{SL}_2(\mathbb{R})$ in the following sense. Recall that group elements $g_1, g_2 \in G$ are said to be *conjugate* if there is some $h \in G$ with $g_1 = hg_2h^{-1}$, and recall also the group $\mathrm{SO}_2(\mathbb{R})$ defined in Lemma 9.1.

Lemma 11.13. *Every $g \in \mathrm{SL}_2(\mathbb{R})$ is conjugate to an element of $\pm A$, $\pm U^-$, or $\mathrm{SO}_2(\mathbb{R})$.*

Proof. If $g \in \mathrm{SL}_2(\mathbb{R})$ is diagonalizable over $\mathbb{R}$, then it is conjugate to an element of A or $-A$; if it is diagonalizable over $\mathbb{C}$ but not over $\mathbb{R}$ then it is

conjugate to an element of $\mathrm{SO}_2(\mathbb{R})$. In fact the two eigenvalues must be λ, λ^{-1} for some $\lambda = e^{i\theta}$, and so in the right basis the matrix is a rotation by θ. If g is not diagonalizable, then it can only have one eigenvalue λ, which must satisfy $\det(g) = \lambda^2 = 1$. It follows that the Jordan normal form of either g or $-g$ belongs to U^-. □

Lemma 11.13 is useful because of the following result.

Lemma 11.14. *Let $\Gamma \leqslant G$ be a discrete subgroup of a closed linear group. Suppose that $g_2 = hg_1h^{-1}$ for $g_1, g_2, h \in G$. Then the maps R_{g_1} and R_{g_2} on $X = \Gamma\backslash G$ are conjugate via R_h. In particular, if Γ is a lattice, then the measure-preserving systems $(X, \mathscr{B}_X, m_X, R_{g_1})$ and $(X, \mathscr{B}_X, m_X, R_{g_2})$ are conjugate (measurably isomorphic).*

Proof. This is clear since $R_{g_2} = R_h R_{g_1} R_h^{-1}$, and if Γ is a lattice then R_h preserves the finite measure m_X. □

Thus the study of the dynamics of elements of $\mathrm{SL}_2(\mathbb{R})$ on X reduces to three cases (ignoring the possibility of a negative sign), namely:

- diagonalizable elements (that is, those elements with a conjugate in A), which are called *hyperbolic*;
- elements conjugate to an element of U^-, called *parabolic*; and
- elements that are conjugate to an element of $\mathrm{SO}(2)$, called *elliptic*.

Since the latter group is compact it exhibits little* interesting dynamics. For example, the ergodic measures for the $\mathrm{SO}(2)$-action are precisely images of the Haar measure $m_{\mathrm{SO}(2)}$. More precisely, consider a point $x \in X$ and the map

$$\phi_x(g) = R_g(x) = xg^{-1}$$

for $g \in \mathrm{SO}(2)$. Then $(\phi_x)_* (m_{\mathrm{SO}(2)})$ is an ergodic measure for the $\mathrm{SO}(2)$-action for which the resulting measure-preserving system is a factor of the action of $\mathrm{SO}(2)$ on itself. These are all of the ergodic measures (see Exercise 11.3.1). This shows that we have to exclude elements of $\mathrm{SO}(2)$ in the discussion of ergodicity.

11.3.2 Ergodicity

Theorem 11.15. *Let $\Gamma \leqslant \mathrm{SL}_2(\mathbb{R})$ be a lattice, and write $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. Let $g \in \mathrm{SL}_2(\mathbb{R})$ be an element that is not conjugate to an element of $\mathrm{SO}(2)$. Then R_g acts ergodically on $(X, \mathscr{B}_X, m_X)$.*

* There are nonetheless interesting dynamical properties of compact orbits (for example, one can ask about the behavior of compact orbits under translation by other group elements), but they are not relevant to the discussion in this chapter.

As discussed above, we have to consider two cases, namely (after replacing g by g^2 if necessary) elements of A and elements of U^-. Even though we already dealt with the former, we will give here a different proof covering both cases using the language of unitary representations.

Definition 11.16. *Let G be a metrizable group and let $\mathscr{H}$ be a Hilbert space. An action $G \times \mathscr{H} \to \mathscr{H}$ of G on $\mathscr{H}$ is a* unitary representation *if every $g \in G$ acts unitarily on $\mathscr{H}$ and for every $v \in \mathscr{H}$ the map $g \longmapsto g(v)$ is continuous with respect to the metric on G and the norm on $\mathscr{H}$.*

Lemma 11.17. *Let X be a locally compact metric space, and let μ be a probability measure on X. Assume that G is a metrizable group that acts continuously on X (see p. 229) and preserves the measure μ. Then the action of G on $L^2_\mu(X)$ defined by $g : f \mapsto f \circ g^{-1}$ is a unitary representation.*

In particular, this lemma may be applied to the right action on $X = \Gamma\backslash G$ where Γ is a lattice in a closed linear group G and $\mu = m_X$ is the Haar measure.

The induced action on $L^2_\mu(X)$ may also be written $(g(f))(x) = f(xg)$; we avoid the usual notation for the unitary operator associated to a measure-preserving system to avoid confusion with the distinguished subgroups $U^\pm$.

Proof of Lemma 11.17. Since every $g \in G$ preserves μ, we already know that $f \mapsto g(f)$ is unitary on $L^2_\mu(X)$. All that remains is to check the continuity requirement, and this follows from the more general result in Lemma 8.7. □

Since ergodicity of an action is characterized by the absence of invariant functions in L^2_0 (the space of square-integrable functions with integral zero) (see Chapter 2 and Exercise 2.3.5 in particular), the following proposition (Proposition 11.18) will become useful in our non-commutative setting. In a metric group G with a left-invariant metric d_G, write

$$B^G_\delta = B^G_\delta(I) = \{g \in G \mid \mathsf{d}_G(g, I) < \delta\}$$

for the metric open ball of radius $\delta > 0$ around the identity, and $B^G_\delta(h) = hB^G_\delta$ for the metric open ball of radius δ around $h \in G$. The following simple but powerful argument is due to Margulis [249].

Proposition 11.18. *Let $\mathscr{H}$ be a Hilbert space carrying a unitary representation of a metric group G. Suppose that $v_0 \in \mathscr{H}$ is fixed by some subgroup $L \subseteq G$. Then v_0 is also fixed by any other element $h \in G$ with the property that*

$$B^G_\delta(h) \cap LB^G_\delta L \neq \varnothing$$

for every $\delta > 0$.

In particular, this proposition applies in general for an element $g \in G$ (with $L = g^{\mathbb{Z}}$) and elements h contained in its *stable horospherical subgroup*

$$U_g^- = \{h \in G \mid g^n h g^{-n} \to I \text{ as } n \to \infty\}$$

or its *unstable horospherical subgroup*

$$U_g^+ = \{h \in G \mid g^n h g^{-n} \to I \text{ as } n \to -\infty\}.$$

As mentioned just after Theorem 11.15, this gives a simple (and less geometric) proof of Theorem 9.21. If $f \in L^2_{m_X}(X)$ is R_g-invariant, then by Proposition 11.18 it is also $R_{U_g^-}$ and $R_{U_g^+}$-invariant. In the case

$$g = a_t \in G = \mathrm{SL}_2(\mathbb{R}),$$

the subgroups $U_g^- = U^-$ and $U_g^+ = U^+$ generate all of $\mathrm{SL}_2(\mathbb{R})$, so the function f is $\mathrm{SL}_2(\mathbb{R})$-invariant and therefore equal to a constant almost everywhere. Note that here the invariance is always understood in $L^2(X)$, while in the proof of Theorem 9.21 on p. 309 we worked with concrete points.

Summarizing the discussion above gives the following more general result.

Corollary 11.19. *Let $\Gamma \leqslant G$ be a lattice in a closed linear group and let X be the homogeneous space $\Gamma\backslash G$. If $g \in G$ has the property that G is generated by U_g^- and U_g^+ then R_g is ergodic with respect to the Haar measure m_X.*

Proof of Proposition 11.18. Without loss of generality we may assume that $\|v_0\| = 1$. We define the auxiliary function (a so-called *matrix coefficient*)

$$p(h) = \langle h(v_0), v_0 \rangle$$

for $h \in G$. Notice that by the continuity requirement in Definition 11.16, $p(h)$ depends continuously on h. Moreover, for $g_1, g_2 \in L$ and $h \in G$,

$$p(g_1 h g_2) = \langle g_1 h(g_2(v_0)), v_0 \rangle = \left\langle h(v_0), g_1^{-1}(v_0) \right\rangle = p(h) \tag{11.7}$$

(since v_0 is fixed by $g_1, g_2 \in L$). Now $h \in G$ acts unitarily, so $\|h(v_0)\| = 1$. We claim that $p(h) = 1$ implies $h(v_0) = v_0$. This may be seen as a consequence of the fact that equality in the Cauchy–Schwartz inequality $|\langle v, w \rangle| \leqslant \|v\|\|w\|$ only occurs if v and w are linearly dependent.

Now let $h \in G$ be as in the statement of the proposition, and choose sequences $g_n \to e$ (the identity in G), (ℓ_n) and (ℓ'_n) in L with

$$\ell_n g_n \ell'_n \to h$$

as $n \to \infty$. Then, on the one hand, by equation (11.7) we have

$$p(\ell_n g_n \ell'_n) = p(g_n) \to p(e) = \|v_0\|^2 = 1$$

as $n \to \infty$, while on the other

$$p(\ell_n g_n \ell'_n) \to p(h)$$

as $n \to \infty$ by continuity. It follows that $p(h) = 1$, and so $h(v_0) = v_0$ by the claim above.

□

The proof of Theorem 11.15 would be easier if we were only interested in proving ergodicity of the horocycle flow, rather than of a single element. This distinction becomes important in equation (11.8), where we would in the former case be free to make the top right-hand entry in the final matrix vanish. In fact the case of the full horocycle flow is the only case that we will need later.

Proof of Theorem 11.15. By Lemmas 11.13 and 11.14 it is enough to consider the cases $g = a_t$ for some $t \neq 0$ and $g = u_s^-$ for some $s \neq 0$. The case $g = a_t$ is covered by Corollary 11.19. So let $g = u_s^-$ (notice that in this case Corollary 11.19 tells us nothing since $U^-_{u_s^-} = U^+_{u_s^-} = \{I_2\}$). Suppose that $f \in L^2_{m_X}(X)$ is invariant under g. We are going to apply Proposition 11.18. For $m, n \in \mathbb{Z}$,

$$\begin{aligned} h_\varepsilon = g^n u_\varepsilon^+ g^m &= \begin{pmatrix} 1 & ns \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \varepsilon & 1 \end{pmatrix} \begin{pmatrix} 1 & ms \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 + ns\varepsilon & (1 + ns\varepsilon)ms + ns \\ \varepsilon & 1 + ms\varepsilon \end{pmatrix}. \end{aligned} \tag{11.8}$$

For small ε we may choose n so that $1 + ns\varepsilon$ is close to 2, specifically so that

$$|(1 + ns\varepsilon) - 2| < s\varepsilon.$$

Now choose m with $\left| ms + \frac{ns}{1+ns\varepsilon} \right| < 1$, so that

$$|(1 + ns\varepsilon)ms + ns| < 3.$$

Let $\varepsilon \to 0$ and choose a subsequence for which the matrix h_ε converges to

$$h = \begin{pmatrix} 2 & r \\ 0 & \frac{1}{2} \end{pmatrix} \tag{11.9}$$

for some r, $|r| \leqslant 3$. For any $\delta > 0$ we can choose $\varepsilon > 0$ so that

$$h_\varepsilon \in B_\delta^G(h) \cap g^{\mathbb{Z}} B_\delta^G g^{\mathbb{Z}},$$

so by Proposition 11.18 we see that f is invariant under R_h. Since h is conjugate to an element of A, the theorem follows from the previous case. □

11.3.3 Mautner Phenomenon for $\mathbf{SL_2}(\mathbb{R})$

Examining the abstract arguments above suggests a general principle. This principle, first discovered and exploited in a geometric context by Mautner [257], was generalized by Moore [260] to the unitary setting.

Proposition 11.20. *Let $\mathscr{H}$ be a Hilbert space carrying a unitary representation of $\mathrm{SL}_2(\mathbb{R})$. Let $g \in \mathrm{SL}_2(\mathbb{R})$ be an element that is not conjugate to an element of $\mathrm{SO}(2)$. Then any vector $v_0 \in \mathscr{H}$ that is fixed by g is also fixed by all of $\mathrm{SL}_2(\mathbb{R})$.*

In ergodic theory this theorem gives a surprising corollary regarding the hereditary behavior of ergodicity. Notice that whenever a group G acts by measure-preserving transformations on a probability space, the same holds for any subgroup $H \subseteq G$. However, if the G-action is in addition ergodic, there is no reason to expect the H-action to be ergodic, since we expect more functions to be invariant under the action of the smaller group (see Exercise 8.1.2 for an abelian example of this). Nonetheless, by combining Proposition 11.20 with Lemma 11.17 we get the following corollary for $\mathrm{SL}_2(\mathbb{R})$.

Corollary 11.21. *Let X be a locally compact metric space with a Borel probability measure μ, and suppose that μ is ergodic for a measure-preserving action of $\mathrm{SL}_2(\mathbb{R})$. Then any element of $\mathrm{SL}_2(\mathbb{R})$ that is not conjugate to an element of $\mathrm{SO}(2)$ acts ergodically.*

Proof of Proposition 11.20. The proof is virtually the same as the proof of Theorem 11.15. An element $g \in \mathrm{SL}_2(\mathbb{R})$ that is not conjugate to an element of $\mathrm{SO}(2)$ is either conjugate to an element of A or to an element of U^-. In the former case we can apply Proposition 11.18 for g, any $h \in U^-$ and any $h \in U^+$ to conclude that v_0 is fixed under $\langle U^-, U^+ \rangle = \mathrm{SL}_2(\mathbb{R})$. In the latter case we proceed as before, using the matrix calculation in equation (11.8) and Proposition 11.18 to see that v_0 is fixed by some element h as in equation (11.9), placing us back in the first case. □

Exercises for Section 11.3

Exercise 11.3.1. For any point $x \in X = \Gamma \backslash \mathrm{SL}_2(\mathbb{R})$, define a map

$$\phi_x : \mathrm{SO}(2) \to X$$

by $\phi_x(g) = R_g(x) = xg^{-1}$ for $g \in \mathrm{SO}(2)$. Show that $(\phi_x)_* (m_{\mathrm{SO}(2)})$ is an ergodic measure for the $\mathrm{SO}(2)$-action on X, for which the resulting measure-preserving system is a factor of the action of $\mathrm{SO}(2)$ on itself. Show that any ergodic measure for the action of $\mathrm{SO}(2)$ is of this form.

Exercise 11.3.2. State and prove a generalization of Proposition 11.20 to unitary representations of $\mathrm{SL}_3(\mathbb{R})$ (and, more generally, to $\mathrm{SL}_d(\mathbb{R})$ for $d \geqslant 2$).

11.4 Mixing and the Howe–Moore Theorem

Using ergodicity of the horocycle flow we can now improve our understanding of the dynamical properties of the action of $\mathrm{SL}_2(\mathbb{R})$.

We say that a square matrix $M \in \mathrm{GL}_k(\mathbb{R})$ is *unipotent* if $(M-I)^k = 0$, that is if $M - I$ is *nilpotent*. In $\mathrm{SL}_2(\mathbb{R})$ a unipotent matrix is one which is conjugate to an element of U^-.

Theorem 11.22. *Let $\Gamma \leqslant \mathrm{SL}_2(\mathbb{R})$ be a lattice. Then the action of $\mathrm{SL}_2(\mathbb{R})$ on $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is mixing.*

11.4.1 First Proof of Theorem 11.22

In this section we will prove a stronger and more general result in the language of unitary representations, which will be related to Proposition 11.18 and which has Theorem 11.22 as a consequence. For this, the following notation will be useful. Let G be a locally compact group, and let $\alpha = (a_n)$ be a sequence of elements of G. Define

$$S(\alpha) = \left\{ g \in G \mid e \in \overline{\{a_n^{-1} g a_n \mid n \in \mathbb{N}\}} \right\}$$

where e is the identity element in G. The reader may think of the case $a_n = a_{t_n} \in A$ for a sequence $t_n \to \infty$, in which case it is easy to describe the set $S(\alpha)$. In fact, this is the only case that we will need later (in Section 11.5).

Proposition 11.23. *Let G be a locally compact group and let $\mathscr{H}$ be a Hilbert space carrying a unitary representation of G. Let $\alpha = (a_n) \in G^{\mathbb{N}}$ be a sequence in G, and suppose for some $v \in \mathscr{H}$ the sequence $a_n(v)$ converges in the weak*-topology to $v_0 \in \mathscr{H}$. Then $gv_0 = v_0$ for all g in the closure of the subgroup generated by the set $S(\alpha)$.*

Proof. Clearly it is sufficient to show that v_0 is fixed by each $g \in S(\alpha)$. So suppose that $g \in S(\alpha)$ and let a_{n_k} be a subsequence such that

$$\lim_{k\to\infty} a_{n_k}^{-1} g a_{n_k} = e. \tag{11.10}$$

Then for any $w \in \mathscr{H}$ we have (by definition of weak*-convergence)

$$\langle gv_0, w\rangle = \left\langle v_0, g^{-1}w\right\rangle = \lim_{k\to\infty} \left\langle a_{n_k} v, g^{-1} w\right\rangle = \lim_{k\to\infty} \langle g a_{n_k} v, w\rangle$$

and similarly

$$\langle v_0, w\rangle = \lim_{k\to\infty} \langle a_{n_k} v, w\rangle .$$

By the unitary property and the continuity of the representation, it follows that

$$|\langle gv_0, w\rangle - \langle v_0, w\rangle| = \lim_{k\to\infty} \left|\left\langle (a_{n_k}^{-1} g a_{n_k})v, a_{n_k}^{-1} w\right\rangle - \left\langle v, a_{n_k}^{-1} w\right\rangle\right|$$
$$\leqslant \lim_{k\to\infty} \left\|(a_{n_k}^{-1} g a_{n_k})v - v\right\| \|w\| = 0.$$

Since this holds for all $w \in \mathscr{H}$ we get $gv_0 = v_0$ as claimed. □

To be able to apply this we need to be able to find non-trivial elements of $S(\alpha)$. We do this now for $G = \mathrm{SL}_2(\mathbb{R})$ in preparation for the proof of Theorem 11.22.

Lemma 11.24. *Let $\alpha = (g_n)$ be a sequence in $\mathrm{SL}_2(\mathbb{R})$ converging to ∞ (that is, such that for any compact subset $K \subseteq \mathrm{SL}_2(\mathbb{R})$ there are only finitely many n with $g_n \in K$). Then the set $S(\alpha)$ contains a non-trivial unipotent element.*

Proof. Recall from the discussion of $\mathrm{SL}_2(\mathbb{R})$ as a closed linear group on p. 284 that the homomorphism $\phi : \mathrm{SL}_2(\mathbb{R}) \to \mathrm{GL}(\mathrm{Mat}_{22}(\mathbb{R}))$ defined by $(\phi(g))(v) = gvg^{-1}$ is a proper map. That is, the norm of $\phi(g_n)$ goes to infinity when (g_n) is a sequence that leaves compact subsets. Notice that $\mathrm{Mat}_{22}(\mathbb{R}) = \mathbb{R}I_2 \oplus \mathfrak{sl}_2(\mathbb{R})$ splits into a sum of two invariant subspaces on the first of which the action is trivial. Thus the norm of Ad_{g_n} goes to infinity; that is we can choose a sequence of vectors (v_n) in $\mathfrak{sl}_2(\mathbb{R})$ for which $\|v_n\| \to 0$ while $\|\mathrm{Ad}_{g_n}(v_n)\| = c > 0$ (where c is some fixed small constant chosen so that the exponential map is injective on the ball of radius $2c$). By exponentiating this sequence, and choosing an appropriate subsequence we get $h_n = \exp(v_n) \to I_2$ and $g_n h_n g_n^{-1} \to u \neq I_2$. Since for large n the element h_n is close to the identity, its eigenvalues are close to 1. The conjugated element $g_n h_n g_n^{-1}$ has the same eigenvalues, so the limit element has 1 as its only eigenvalue – thus u is unipotent and non-trivial. □

Proof of Theorem 11.22. Let (a_n) be a sequence in $\mathrm{SL}_2(\mathbb{R})$ converging to ∞. Let $f \in L^2(X)$ (with X as in the statement of the theorem); we wish to show that $a_n(f)$ converges in the weak*-topology to $\int f \,\mathrm{d}m_X$. Since

$$\|a_n(f)\|_2 = \|f\|_2,$$

we know that any subsequence of $(a_n(f))$ has a weak*-convergent subsequence (since the closed ball of radius $\|f\|_2$ is weak*-compact by the Alaoglu–Tychonoff theorem (Theorem B.6)).

So let $(a_{n_k}(f))$ be a sequence converging weak* to f_0. By Lemma 11.24, the set $S(\alpha)$ contains a non-trivial unipotent element of $\mathrm{SL}_2(\mathbb{R})$ where $\alpha = (a_{n_k})$. It follows that f_0 is invariant under a unipotent element by Proposition 11.23 and is therefore constant by Theorem 11.15. This implies that $f_0 = \int f \,\mathrm{d}m_X$ as claimed, proving the theorem. □

11.4.2 Vanishing of Matrix Coefficients for $\mathrm{PSL}_2(\mathbb{R})$

The proof in Section 11.4.1 also gives the following theorem due to Howe and Moore [160].

Theorem 11.25. *Let $\mathscr{H}$ be a Hilbert space carrying a unitary representation of $\mathrm{SL}_2(\mathbb{R})$ without any invariant vectors. Then for any $v, w \in \mathscr{H}$ the* matrix coefficients *$\langle gv, w\rangle$ for $g \in \mathrm{SL}_2(\mathbb{R})$ vanish at ∞:*

$$\langle g_n v, w\rangle \longrightarrow 0$$

as $g_n \to \infty$.

11.4.3 Second Proof of Theorem 11.22; Mixing of All Orders

In the section we present a different proof for mixing, which also leads to a proof of mixing of all orders[98]. The result and approach presented here is due to Mozes [262], [263] and it holds more generally than the case considered here.

Theorem 11.26. *Let X be a σ-compact metric space equipped with a continuous $\mathrm{SL}_2(\mathbb{R})$-action. Let μ be an $\mathrm{SL}_2(\mathbb{R})$-invariant ergodic probability measure on X. Then the $\mathrm{SL}_2(\mathbb{R})$-action is mixing of all orders with respect to μ. That is, for any $r \geqslant 1$, functions $f_0, f_1, \dots, f_{r-1} \in L^\infty(X)$ and $g^{(1)}, \dots, g^{(r-1)}$ in $\mathrm{SL}_2(\mathbb{R})$, we have*

$$\int f_0(x) f_1(g^{(1)}\bullet x) f_2(g^{(2)}\bullet x) \cdots f_{r-1}(g^{(r-1)}\bullet x)\,\mathrm{d}\mu(x) \to \int f_0\,\mathrm{d}\mu \cdots \int f_{r-1}\,\mathrm{d}\mu$$

as $g^{(i)} \to \infty$ and $g^{(i)}(g^{(j)})^{-1} \to \infty$ for $i \neq j$, $1 \leqslant i, j \leqslant r$.

We start with the special case $r = 2$, which is essentially the statement of Theorem 11.22.

Second Proof of Theorem 11.22. Suppose that $g_n = g_n^{(1)} \in \mathrm{SL}_2(\mathbb{R})$ eventually leaves any compact subset of $\mathrm{SL}_2(\mathbb{R})$. We wish to show that for $f_1, f_2 \in C_c(X)$ we have

$$\int f_1(x) f_2(g_n\bullet x)\,\mathrm{d}\mu \longrightarrow \int f_1\,\mathrm{d}\mu \int f_2\,\mathrm{d}\mu. \tag{11.11}$$

This then extends by approximation to functions $f_1, f_2 \in L^2_\mu(X)$, showing the mixing property (we will say more about this approximation argument in the proof of Theorem 11.26).

Consider the diagonal measure Δ on $X \times X$ defined by the relation

$$\int_{X\times X} F(x_1, x_2)\,\mathrm{d}\Delta(x_1, x_2) = \int_X F(x, x)\,\mathrm{d}m_X(x)$$

for all $F \in C_c(X \times X)$, and the push-forward

$$\mu_n = (I, g_n)_*\Delta,$$

where the action is simply the product action

$$(h_1, h_2)\cdot(x_1, x_2) = (h_1\cdot x_1, h_2\cdot x_2)$$

for $h_1, h_2 \in \mathrm{SL}_2(\mathbb{R})$. Then

$$\begin{aligned}\int_{X\times X} f_1(x_1)f_2(x_2)\,\mathrm{d}\mu_n(x_1,x_2) &= \int_{X\times X} f_1(x_1)f_2(g_n\cdot x_2)\,\mathrm{d}\Delta(x_1,x_2)\\ &= \int_X f_1(x)f_2(g_n\cdot x)\,\mathrm{d}\mu(x)\end{aligned}$$

gives precisely the left-hand side of equation (11.11). Therefore we wish to show that μ_n converges weakly to $\mu\times\mu$.

Note that μ_n projects to μ in both coordinates: that is, $(\pi_j)_*(\mu_n) = \mu$ for $j = 1, 2$, where π_j is as usual the projection $(x_1, x_2) \mapsto x_j$ onto the jth coordinate. However, μ_n is not a joining of the two systems since it is only invariant under the action of the subgroup

$$\{(h, g_n h g_n^{-1}) \mid h \in \mathrm{SL}_2(\mathbb{R})\} \tag{11.12}$$

obtained from conjugation of the diagonal subgroup by (I_2, g_n).

If μ_n does not converge to $\mu\times\mu$, then we may choose a subsequence which converges to a different limit ν say.

We claim first that ν is still a probability measure with $(\pi_j)_*(\nu) = \mu$ for $j = 1, 2$. If X is compact this is clear. If X is not compact, then for any $\varepsilon > 0$ we may choose a compact set $K \subseteq X$ of measure $\mu(K) > 1-\varepsilon$, and a function f_K in $C_c(X)$ with $0 \leqslant f_K \leqslant 1$ and $f_K(x) = 1$ for all $x \in K$. Then for any non-negative function $f \in C_c(X)$, the function

$$(x_1, x_2) \mapsto f(x_1)f_K(x_2)$$

is in $C_c(X\times X)$, and satisfies

$$0 \leqslant f(x_1) - f(x_1)f_K(x_2) \leqslant \|f\|_\infty\,(1 - f_K(x_2))$$

and so for any n we have

$$\begin{aligned}\int_{X\times X} f(x_1)f_K(x_2)\,\mathrm{d}\mu_n(x_1,x_2) &\leqslant \int_{X\times X} f(x_1)\,\mathrm{d}\mu_n(x_1,x_2)\\ &= \int_X f\,\mathrm{d}\mu \leqslant \int_{X\times X} f(x_1)f_K(x_2)\,\mathrm{d}\mu_n + \varepsilon\|f\|_\infty.\end{aligned}$$

In the limit as $n\to\infty$, the same property holds for ν. Decreasing ε and increasing both K and f_K monotonically gives

$$\int_{X\times X} f(x_1)\,\mathrm{d}\nu(x_1,x_2) = \int_X f\,\mathrm{d}\mu$$

for all $f \in C_c(X)$. Therefore, $(\pi_1)_*\nu = \mu$ and similarly $(\pi_2)_*\nu = \mu$.

Next we claim that ν has some invariance properties (inherited from equation (11.12)). More precisely, let $\alpha = (g_n)$ and let $u \in S(\alpha)$ be a non-trivial unipotent element (which exists by Lemma 11.24). Then we claim that ν is invariant under (I, u). So suppose without loss of generality that a sequence (h_n) in $\mathrm{SL}_2(\mathbb{R})$ satisfies $h_n \to I$ as $n \to \infty$, and has $g_n h_n g_n^{-1} \to u$. Let $f \in C_c(X \times X)$. Then, for any n,

$$\int f(x_1, x_2)\, \mathrm{d}\mu_n(x_1, x_2) = \int f(h_n{\cdot}x_1, (g_n h_n g_n^{-1}){\cdot}x_2)\, \mathrm{d}\mu_n(x_1, x_2)$$

by invariance under the subgroup in equation (11.12). For $n \to \infty$ we have

$$\int f(x_1, x_2)\, \mathrm{d}\mu_n(x_1, x_2) \to \int f(x_1, x_2)\, \mathrm{d}\nu(x_1, x_2)$$

and

$$\int f(h_n{\cdot}x_1, (g_n h_n g_n^{-1}){\cdot}x_2)\, \mathrm{d}\mu_n(x_1, x_2) \to \int f(x_1, u{\cdot}x_2)\, \mathrm{d}\nu(x_1, x_2),$$

where the latter follows by uniform continuity of f as $h_n \to e$, $g_n h_n g_n^{-1} \to u$, and so $h_n{\cdot}x_1 \to x_1$, $g_n h_n g_n^{-1}{\cdot}x_2 \to u{\cdot}x_2$ as $n \to \infty$ uniformly for (x_1, x_2) in a neighborhood of the support of f.

To summarize: $(\pi_j)_* \nu = \mu$ for $j = 1, 2$ and ν is invariant under (I_2, u) for some non-trivial unipotent $u \in \mathrm{SL}_2(\mathbb{R})$. We claim that this implies that $\nu = \mu \times \mu$. Clearly the σ-algebra

$$\mathscr{A} = \mathscr{B}_X \times \{\varnothing, X\}$$

consists of (I_2, u)-invariant sets, so that for almost every (x_1, x_2) the conditional measures $\mu^{\mathscr{A}}_{(x_1,x_2)}$ are supported on $[(x_1, x_2)]_{\mathscr{A}} = \{x_1\} \times X$, and are invariant under (I_2, u) (cf. Corollary 5.24). Writing $\nu_{(x_1,x_2)} = (\pi_2)_* \mu^{\mathscr{A}}_{(x_1,x_2)}$, we see that

$$\mu = (\pi_2)_* \nu = \int \nu_{(x_1,x_2)}\, \mathrm{d}\mu(x_1, x_2)$$

represents μ as an integral over u-invariant measures on X. Since μ is ergodic with respect to u by Corollary 11.21 we conclude that

$$\mu^{\mathscr{A}}_{(x_1,x_2)} = \delta_{x_1} \times \nu_{(x_1,x_2)} = \delta_{x_1} \times \mu$$

for ν-almost every (x_1, x_2). Together with the fact that $(\pi_1)_* \nu = \mu$, this implies that $\nu = \mu \times \mu$ (since $\mathscr{A} = \mathscr{B}_X \times \{\varnothing, X\}$). □

The proof of the general case of Theorem 11.26 proceeds along similar lines, using induction on the parameter r.

Proof of Theorem 11.26. First notice that it is enough* to consider continuous functions of compact support in $C_c(X)$. To see this, let $F_j \in L^\infty(X)$

* These reductions are used frequently in Chapter 7, where higher-order expressions play an important role.

and choose functions $f_j \in C_c(X)$ with $\|f_j - F_j\|_2 < \varepsilon$ and $\|f_j\|_\infty \leqslant \|F_j\|_\infty$ for $j = 0, 1, \ldots, r-1$. Then

$$\left| \int F_0(x) F_1(g^{(1)} \cdot x) \cdots F_{r-1}(g^{(r-1)} \cdot x) \, \mathrm{d}\mu(x) \right.$$
$$\left. - \int f_0(x) F_1(g^{(1)} \cdot x) \cdots F_{r-1}(g^{(r-1)} \cdot x) \, \mathrm{d}\mu(x) \right| \leqslant \varepsilon \|F_1\|_\infty \cdots \|F_{r-1}\|_\infty,$$

so replacing F_0 by f_0 comes at the cost of a fixed multiple of ε only. Repeating as necessary shows that it is enough to prove Theorem 11.26 for functions in $C_c(X)$.

As a result of formulating the result using continuous functions, we can again phrase the required conclusion in terms of weak*-convergence of a sequence of measures. Define the diagonal measure $\Delta = \Delta_r$ on X^r by

$$\int_{X^r} F(x_0, \ldots, x_{r-1}) \, \mathrm{d}\Delta(x_0, \ldots, x_{r-1}) = \int_X F(x, \ldots, x) \, \mathrm{d}\mu(x)$$

for all $F \in C_c(X^r)$. Then

$$\int_{X^r} f_0(x_0) f_1(x_1) \cdots f_{r-1}(x_{r-1}) \, \mathrm{d}(I, g^{(1)}, \ldots, g^{(r-1)})_* \Delta$$
$$= \int_X f_0(x) f_1(g^{(1)} \cdot x) \cdots f_{r-1}(g^{(r-1)} \cdot x) \, \mathrm{d}\mu(x)$$

for $f_j \in C_c(X)$ and $(I, g^{(1)}, \ldots, g^{(r-1)}) \in (\mathrm{SL}_2(\mathbb{R}))^r$ acting on X^r. Choose a sequence $(g_n^{(1)}, \ldots, g_n^{(r-1)})$ of $(r-1)$-tuples of elements of $\mathrm{SL}_2(\mathbb{R})$ with $g_n^{(i)} \to \infty$ and $g_n^{(i)}(g_n^{(j)})^{-1} \to \infty$ as $n \to \infty$ for $i \neq j$. We wish to show that

$$\mu_n = (I, g_n^{(1)}, \ldots, g_n^{(r-1)})_* \Delta$$

converges to $\mu \times \mu \times \cdots \times \mu$ in the weak*-topology as $n \to \infty$. Choosing a subsequence if necessary we can assume that μ_n converges in the weak*-topology to some limit ν.

Just as in the proof for $r = 2$ above, we have that ν is a probability measure with $(\pi_j)_* \nu = \mu$ for any of the coordinate projections

$$\pi_j(x_0, x_1, \ldots, x_{r-1}) = x_j.$$

By induction on r we may assume that the theorem holds for $r-1$ functions. This in fact translates to the refined statement that $(\pi_J)_* \nu = \mu^{|J|}$ on $X^{|J|}$ for any proper subset $J \subseteq \{0, 1, \ldots, r-1\}$. Here $\pi_J = \prod_{j \in J} \pi_j$ is the projection onto the coordinates corresponding to the subset J. To see this, fix some proper subset $J \subseteq \{0, 1, \ldots, r-1\}$ and functions $f_j \in C_c(X)$ for $j \in J$. Then (by definition, and in the case that X is not compact by the approximation argument used in the case $r = 2$)

$$\int_X \prod_{j\in J} f_j(x_j)\,\mathrm{d}\nu = \lim_{n\to\infty} \int \prod_{j\in J} f_j(g_n^{(j)}\cdot x_j)\,\mathrm{d}\mu$$
$$= \prod_{j\in J} \int f_j\,\mathrm{d}\mu$$

by the assumption on the sequences $\left(g_n^{(j)}\right)$ and the inductive assumption.

It is also clear that μ_n is invariant under the group

$$\left\{ \left(h, g_n^{(1)} h (g_n^{(1)})^{-1}, \ldots, g_n^{(r-1)} h (g_n^{(r-1)})^{-1}\right) \mid h \in \mathrm{SL}_2(\mathbb{R}) \right\},$$

from which we again want to derive a non-trivial invariance property of ν. This requires the following generalization of Lemma 11.24. There exists a subsequence of the $g_n^{(j)}$ (passing to this subsequence is suppressed in the notation for simplicity), a sequence (h_n) in $\mathrm{SL}_2(\mathbb{R})$, and a subset $J \subseteq \{1, \ldots, r-1\}$ such that $h_n \to I$ and $g_n^{(j)} h_n (g_n^{(j)})^{-1} \to I$ for $j \notin J$, while $g_n^{(j)} h_n (g_n^{(j)})^{-1} \to u_j$ for $j \in J$, where $u_j \in \mathrm{SL}_2(\mathbb{R})$ is non-trivial and unipotent for each $j \in J$ (see Exercise 11.4.1).

For notational simplicity we assume that $J = \{s, \ldots, r-1\}$, and we write $u = (u_j \mid j \in J)$ for the $|J|$-tuple of unipotent elements constructed as limits. By the same argument as that used in the case $r = 2$, this implies that ν is invariant under the action of u on the last $(r-s)$ coordinates of $X^r = Z$, namely

$$u\cdot(x_0, x_1, \ldots, x_{r-1}) = (x_0, \ldots, x_{s-1}, u_s\cdot x_s, \ldots, u_{r-1}\cdot x_{r-1})$$

for $(x_0, \ldots, x_{r-1}) \in X^r$. We define the σ-algebra

$$\mathscr{A} = \mathscr{B}_X \otimes \cdots \otimes \mathscr{B}_X \times \{\varnothing, X\}^{r-s}$$

which consists of u-invariant sets. Then the conditional measures $\nu_z^{\mathscr{A}}$ are, for ν-almost every $z \in X^r$, invariant under u by Corollary 5.24. This implies that

$$\mu^{r-s} = (\pi_J)_*\nu = \int_Z (\pi_J)_*\nu_z^{\mathscr{A}}\,\mathrm{d}\nu(z) \tag{11.13}$$

expresses μ^{r-s} as an integral of u-invariant probability measures $(\pi_J)_*\nu_z^{\mathscr{A}}$. However, as the action of u_j for $j = s, \ldots, r-1$ is mixing (and so, in particular weak-mixing) with respect to μ by Theorem 11.22, it follows from Theorem 2.36 that the action of u on X^{r-s} is ergodic with respect to μ^{r-s}. Therefore, equation (11.13) implies that

$$\nu_z^{\mathscr{A}} = \delta_{(z_0,\ldots,z_{s-1})} \times \mu^{r-s}$$

for ν-almost every $z \in X^r$ by Theorem 4.4. Furthermore, $\pi_{\{0,\ldots,s-1\}}\nu = \mu^s$ on X^s, so that we get $\nu = \mu^r$ on X^r as desired. This concludes the proof of the inductive step, and hence the theorem. □

Exercises for Section 11.4

Exercise 11.4.1. Prove the generalization of Lemma 11.24 used in the proof of Theorem 11.26 by analyzing the sequence of linear maps $\mathrm{Ad}_{g_n^{(j)}}$ on $\mathfrak{sl}_2(\mathbb{R})$.

11.5 Rigidity of Invariant Measures for the Horocycle Flow

In this section we will discuss the set of probability measures on the homogeneous space $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ that are invariant under the horocycle flow. Recall that

$$U^- = \left\{ u^-(s) = \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} \mid s \in \mathbb{R} \right\}$$

and that the horocycle flow is defined by

$$h(s)\cdot x = R_{u^-(s)}(x) = x \begin{pmatrix} 1 & -s \\ & 1 \end{pmatrix}$$

for any $x \in \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. If Γ is a lattice, which we will assume here, then m_X is an invariant and ergodic probability measure for the horocycle flow by Theorem 11.15. If Γ is cocompact (that is, if X is compact) then the horocycle flow is uniquely ergodic[(99)] by a theorem of Furstenberg [101]. For the general case we do not have unique ergodicity. For example, if $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ then our reference vector $x_0 = (\mathrm{i}, \mathrm{i})$ (the vector pointing upwards based at $\mathrm{i} \in \mathbb{H}$) has a periodic orbit under the horocycle flow – and a periodic orbit supports an invariant measure, showing that the flow is not uniquely ergodic. In algebraic terms x_0 corresponds to $I \in \mathrm{SL}_2(\mathbb{R})$, and the identity

$$h(1)\cdot \mathrm{SL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z}) \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \mathrm{SL}_2(\mathbb{Z})$$

shows that the orbit consists of a periodic cycle. Figure 11.12 shows the periodic cycle for the reference vector x_0, which consists of all vectors pointing upwards based at all points in the fundamental domain of the form $\mathrm{i} + t$ with $-\frac{1}{2} \leqslant t \leqslant \frac{1}{2}$ since the horocycle flow moves vectors that point upwards horizontally without changing their direction.

Even though in general the horocycle is not uniquely ergodic, it is possible to describe all probability measures that are invariant and ergodic under the horocycle flow, as shown by Dani [61]. The proof we present is different from the original proofs of Furstenberg and Dani, and is due to Margulis.

Theorem 11.27. *Let $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ be a lattice and let $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. Let μ be a probability measure that is invariant and ergodic under the horocycle flow $h(s)$ for $s \in \mathbb{R}$. Then either $\mu = m_X$ or μ is the unique invariant measure*

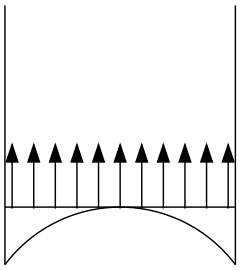

Fig. 11.12. The periodic cycle for the reference vector x_0.

*supported on a periodic orbit** *for* U^-*. If* X *is compact the only possibility is* $\mu = m_X$ *(that is, there is unique ergodicity) since there are no periodic orbits for* U^-*.*

Our method of proof requires the assumption that Γ is a lattice, but this is not necessary for the results.

11.5.1 Existence of Periodic Orbits; Geometric Characterization

We begin by explaining the relationship between compactness of X and the existence of periodic orbits for U^-.

Lemma 11.28. *Let* $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ *be a discrete subgroup. Assume that the point* $x_0 \in X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ *is periodic for* U^-*. Then* $R_{a_t}(x_0)$ *diverges to infinity in* X *(that is, leaves any compact subset of* X*). In particular, if* X *is compact then there are no periodic orbits for* U^-*.*

Proof. As discussed in Section 9.3.3, the space X is locally isomorphic to $\mathrm{SL}_2(\mathbb{R})$. That is, for any $x \in X$ there is an injectivity radius $r_x > 0$ such that the map

$$\begin{aligned} B_{r_x}^{\mathrm{SL}_2(\mathbb{R})} &\longrightarrow X \\ x &\longmapsto xg \end{aligned}$$

is an isometry. Moreover, if $K \subseteq X$ is compact then there exists some uniform $r > 0$ that serves as an injectivity radius across all points of K. This puts an immediate constraint on the length ℓ_0 of a possible periodic orbit of $x_0 \in K$ for the horocycle flow as follows. There exists some $s > 0$ (depending only on r) such that for $x \in K$ and $\ell \in \mathbb{R}\smallsetminus\{0\}$ with $|\ell| < s$ we have $h(\ell)\cdot x \neq x$, so that $|\ell_0| \geqslant s$.

* As a measure-preserving dynamical system, in this case the system (X, μ, R_{u^-}) for $u^- \in U^-$ is conjugate to the rotation flow $(\mathbb{T} = \mathbb{R}/\mathbb{Z}, m_{\mathbb{T}}, R_t)$ for $t \in \mathbb{R}$.

Suppose that $x_0 \in X$ is a periodic cycle of length ℓ_0, so

$$h(\ell_0)\cdot x_0 = x_0 u^-(-\ell_0) = x_0.$$

Then $R_{a_t}(x_0)$ satisfies

$$h(e^{-t}\ell_0)\cdot R_{a_t}(x_0) = x_0 a_t^{-1} u^-(-e^{-t}\ell_0) = x_0 u^-(-\ell_0) a_t^{-1} = R_{a_t}(x_0),$$

so $R_{a_t}(x_0)$ is periodic with period length $e^{-t}\ell_0$ and so $R_{a_t}(x_0) \notin K$ for large enough t. This shows that $R_{a_t}(x_0) \to \infty$ for $t \to \infty$, and hence that there cannot be any periodic points if X is compact. □

The converse for lattices relies on the geometry of the Dirichlet region discussed in Section 11.1.

Lemma 11.29. *If $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is not compact but is of finite volume, then there are periodic orbits for the horocycle flow in X. In fact, to every cusp of X there corresponds precisely a one-parameter family of periodic U^--orbits in X parameterized by the action of the diagonal subgroup A. More precisely, for one point x with periodic U^--orbit we get precisely one element, namely $R_{a_t}(x)$, of all other periodic U^--orbits that are associated to the same cusp, by letting $t \in \mathbb{R}$ vary. Moreover, $x \in X$ is periodic for U^- if and only if $R_{a_t}(x) \to \infty$ for $t \to \infty$.*

As the proof of this lemma will show, the result is easily verified for the case $\Gamma = \mathrm{SL}_2(\mathbb{Z})$.

Proof of Lemma 11.29. Recall from Section 11.1 that the Dirichlet region D defined by $p \in \mathbb{H}$ consists of all points $y \in \mathbb{H}$ with

$$\mathsf{d}(p, y) = \min_{\gamma \in \Gamma} \mathsf{d}(p, \gamma(y)),$$

that it is a hyperbolic convex n-gon for some n, that it represents a fundamental domain for the action of Γ, and that the n-gon D has at least one point on the boundary $\overline{\mathbb{R}} = \partial\mathbb{H}$ since X is not compact by Lemma 11.9. We will refer to the points of $\overline{D}^{\mathbb{H}} \cap \partial\mathbb{H}$ as boundary vertices, while cusps are as before equivalence classes of boundary vertices.

We claim that for every boundary vertex r there is a non-trivial unipotent $\gamma \in \Gamma$ fixing r. To do this, we first show that it is sufficient that there be a non-central element $\gamma \in \Gamma$ fixing r. By conjugating Γ with some element of $\mathrm{SL}_2(\mathbb{R})$ we may assume that $r = \infty$. Now any matrix $\gamma \in \Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ with $\gamma(\infty) = \infty$ has the form $\gamma = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $ad = 1$ and $b \in \mathbb{R}$. For such an element, $\gamma(p) = \frac{a}{d}p + \frac{b}{d}$ has imaginary part $\Im(\gamma(p)) = \frac{a}{d}\Im(p)$ with $\frac{a}{d} > 0$. Replacing γ by γ^{-1} if necessary, we may assume that $\frac{a}{d} \geqslant 1$. If $\frac{a}{d} > 1$, then $\Im(\gamma(p)) > \Im(p)$. In this case, the geodesic line

$$L_\gamma = \{z \in \mathbb{H} \mid \mathsf{d}(z, p) = \mathsf{d}(z, \gamma(p))\}$$

as in Lemma 11.5 is a half-circle in $\mathbb{C}$ rather than a vertical line. In particular, in this case any $z \in \mathbb{H}$ with big enough imaginary part is closer to $\gamma(p)$ than it is to p, which contradicts the assumption that ∞ is a boundary vertex of the Dirichlet region defined by p. We must therefore have $\frac{a}{d} = 1$, so $a = d = \pm 1$. Thus if a non-central element $\gamma \in \Gamma$ fixes the boundary vertex r, then γ^2 is a non-trivial unipotent element fixing r.

To prove the claim that for any boundary vertex r there is a non-trivial element $\gamma \in \Gamma$ fixing r, we start by considering an edge E which makes up one of the pieces of the boundary of D near r. Then $E \subseteq L_{\gamma_1}$ for some $\gamma_1 \in \Gamma$, with L_{γ_1} as in Lemma 11.5. We claim that $E = \overline{D} \cap L_{\gamma_1}$ may also be written

$$E = \overline{\gamma_1(D)} \cap L_{\gamma_1};$$

that is, the two hyperbolic convex n-gons D and $\gamma_1(D)$ meet full edge to full edge in E. To see this, consider a point $y \in \overline{\gamma_1(D)} \cap L_{\gamma_1}$. Then

$$\mathsf{d}(y, p) = \mathsf{d}(y, \gamma_1(p)) = \min_{\gamma \in \Gamma} \mathsf{d}(y, \gamma(p)).$$

We claim that every point on the geodesic from y to p belongs to D. If not, then there is some z in that geodesic segment with $\mathsf{d}(y, p) = \mathsf{d}(y, z) + \mathsf{d}(z, p)$ (since z lies on the geodesic joining y to p) and with $\mathsf{d}(z, \gamma(p)) \leqslant \mathsf{d}(z, p)$. This would imply that

$$\mathsf{d}(y, \gamma(p)) \leqslant \mathsf{d}(y, z) + \mathsf{d}(z, \gamma(p)) \leqslant \mathsf{d}(y, z) + \mathsf{d}(z, p) = \mathsf{d}(y, p) \leqslant \mathsf{d}(y, \gamma(p)),$$

and, moreover, that the path from y to z and on to $\gamma(p)$ (via the pieces of geodesics) is in fact a length minimizing path. By Proposition 9.4 this can only happen if y, z and $\gamma(p)$ all belong to the same geodesic, which implies that $\gamma(p) = p$ and therefore $\gamma = \pm I$. This implies that y lies in $\overline{D} \cap L_{\gamma_1}$, and so $\overline{D} \cap L_{\gamma_1} \subseteq \overline{\gamma_1(D)} \cap L_{\gamma_1}$. The same argument implies the reversed inclusion.

If there are no other boundary vertices (as, for example, in the case $r = \infty$ and $\Gamma = \mathrm{SL}_2(\mathbb{Z})$) or if none of the other boundary vertices are equivalent to r (as, for example, in the case $r = \infty$ in the example of Section 11.2.2), then we claim that γ_1 fixes r. From the previous paragraph, we know that D meets $\gamma_1(D)$ full edge to full edge in E, so there is an edge E' of D that is mapped under γ_1 to E. As E' has the boundary vertex $r_1 = {\gamma_1}^{-1}(r)$, we must have $r_1 = r$, and so $\gamma_1(r) = r$ as claimed.

In general, the boundary vertex r_1 satisfying $\gamma_1(r_1) = r$ might not coincide with r (as, for example, in the case $r = 1$ and $r_1 = -1$ arising in Section 11.2.2). In the following argument we will make use of pairs (E', r'), where E' is an edge of D and r' is both a boundary vertex of D and an endpoint of E'. We may think of these pairs as *directed edges*. Starting with the directed edge (E, r), there is an edge E_1' of D that is mapped under γ_1 to E (that is, the directed edge (E_1', r_1) is mapped to (E, r)). Let (E_1, r_1) be that directed edge with the property that E_1 is the other edge of D meeting E_1' at the vertex r_1. If $r_1 \neq r$, then there is some $\gamma_2 \in \Gamma$

with $E_1 \subseteq L_{\gamma_2}$. In particular, D meets $\gamma_2(D)$ in the edge E_1. Hence there is a directed edge (E_2', r_2) that is mapped under γ_2 to (E_1, r_2). Let (E_2, r_2) be the other directed edge meeting E_2' in r_2, and so on. Geometrically speaking we move from D to the next copy $\gamma_1(D)$ of D near r and then move through $\gamma_1(D)$ to the next copy $\gamma_1\gamma_2(D)$ and so on. Formally, we obtain a sequence of directed edges $(E, r), (E_1', r_1), (E_1, r_1), (E_2', r_2), (E_2, r_2), \ldots$, and elements $\gamma_1, \gamma_2, \cdots \in \Gamma$ with the properties that

- The directed edge (E_{j+1}', r_{j+1}) is mapped under γ_{j+1} to (E_j, r_j); that is, $\gamma_{j+1}(r_{j+1}) = r_j$ and $\gamma_{j+1}(E_{j+1}') = E_j$ for $j \geqslant 1$. Similarly, γ_1 maps (E_1', r_1) to (E, r).
- The directed edges (E_j', r_j) and (E_j, r_j) are precisely the two directed edges meeting in the boundary vertex r_j for $j \geqslant 1$.

Note that $E_{j+1}' \subseteq L_{\gamma_{j+1}^{-1}}$ and that E_j is the edge of D that is mapped under γ_{j+1}^{-1} to E_{j+1}. In particular, the directed edge (E_{j+1}, r_{j+1}) determines (E_j, r_j) in the same way that (E_j, r_j) determines (E_{j+1}, r_{j+1}). Therefore, there exists some $n \geqslant 1$ for which $E_n = E$ and $r_n = r$. In symbols, we have

$$r = \gamma_1\gamma_2 \cdots \gamma_n(r_n) = \gamma(r_n) = \gamma(r)$$

where $\gamma = \gamma_1\gamma_2 \cdots \gamma_n$. It remains to show that γ is non-trivial. Again assume that $r = \infty$, so that E is a vertical line, and that $\gamma_1(D)$ is to the right of E. Then E_1 is the right edge of D rising vertically to ∞. By induction, we deduce that $\gamma_1\gamma_2 \cdots \gamma_n(D)$ is to the right of D, so γ is non-central. This completes the proof of the claim that every boundary vertex of D is fixed by a non-trivial unipotent γ.

Given a non-trivial unipotent $\gamma \in \Gamma$ there is an element $g \in \mathrm{SL}_2(\mathbb{R})$ with $g^{-1}\gamma g = u^-(s) \in U^-$. Therefore $h(s)\cdot\Gamma g = \Gamma g u^-(-s) = \Gamma g$ is periodic for the horocycle flow.

It remains to demonstrate the correspondence between one-parameter families of periodic orbits for the horocycle flow and cusps. For this, notice that if we remove from D its intersection with a big compact ball we are left with as many connected components as there are boundary vertices (indeed, what is left will be a union of disjoint open neighborhoods of the boundary vertices). However, as discussed, boundary vertices are identified under the action of Γ and correspondingly some of the edges meeting those vertices are identified under the natural map $D \to \Gamma\backslash\mathbb{H}$. In other words, the cusps correspond to connected components of $\Gamma\backslash\mathbb{H}$ after removing from the latter a big compact subset Ω. By Lemma 11.28, any point x which is periodic for U^- diverges under R_{a_t} as $t \to \infty$, so $R_{a_t}(x)$ must approach one and only one of the cusps as $t \to \infty$.

Now consider the point $z \in D$ and the vector $v \in \mathrm{T}_z\mathbb{H}$ corresponding to $R_{a_{t_0}}(x)$ for some large enough t_0 for which $R_{a_t}(x)$ stays in the given neighborhood of the appropriate cusp for all $t \geqslant t_0$. Without loss of generality $z = x + \mathrm{i}y$ belongs to the neighborhood of the boundary vertex $r = \infty$

and we may assume that $y > 1$. We claim that v points straight up, which then shows that the corresponding element of $\mathrm{SL}_2(\mathbb{R})$ has the form $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$. Since we are interested in parameterizing the periodic orbits, it then follows that any two such periodic orbits are on the same orbit of the subgroup

$$A = \left\{ \begin{pmatrix} a & b \\ & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}) \right\}.$$

To prove the claim, suppose that v does not point straight up. Then the associated geodesic line is a semi-circle, and some future point $R_{a_t}(x)$ will have imaginary part equal to 1. Define K to be the compact segment of the line $y = 1$ in $\mathbb{C}$ starting at i and ending at $\gamma(\mathrm{i}) \in \mathbb{R} + \mathrm{i}$, where $\gamma \in \Gamma$ is a unipotent element fixing ∞. Since K is compact, we may assume that Ω contains the image of K in $\Gamma\backslash\mathbb{H}$. This is a contradiction since $R_{a_t}(x)$ by hypothesis does not return to Ω for $t > t_0$.

The argument above also shows that if $R_{a_t}(x) \to \infty$ for any $x \in X$ then this orbit must eventually get close to a single cusp, and this happens (assuming that $\infty \in \partial\mathbb{H}$ represents this cusp) again only if v points upwards. In this case, however, the horocycle flow moves horizontally and the unipotent element γ fixing ∞ shows that x is a periodic point for U^-. □

11.5.2 Proof of Measure Rigidity for the Horocycle Flow

Just as in the proof of Lemma 11.28 we will use the geodesic flow to study stretches of U^--orbits in the proof of Theorem 11.27. For this our main tool will be the mixing property of the geodesic flow R_{a_t} on X.

Proof of Theorem 11.27. Let μ be an invariant and ergodic probability measure for the horocycle flow $x \mapsto h(s)\cdot x = R_{u^-(s)}(x)$ for $s \in \mathbb{R}$ and $x \in X$. Recall that $x_0 \in X$ is generic for μ if for all $f \in C_c(X)$ the time averages converge in the sense that

$$\frac{1}{S}\int_0^S f\left(h(s)\cdot x_0\right)\,\mathrm{d}s \longrightarrow \int f\,\mathrm{d}\mu \tag{11.14}$$

as $S \to \infty$, and that μ-almost every $x_0 \in X$ is generic for μ. Let x_0 be such a generic point. If x_0 is periodic for the horocycle flow, then μ must be the image of the one-dimensional Lebesgue measure on this periodic orbit. So assume now that x_0 is not periodic. We will study the time averages as in equation (11.14) and show for some subsequence of times S that the averages converge to $\int f\,\mathrm{d}m_X$. Since for a point x_0 that is generic for μ the time averages converge to $\int f\,\mathrm{d}\mu$ we deduce that $\int f\,\mathrm{d}\mu = \int f\,\mathrm{d}m_X$ for all f in $C_c(X)$, and so $\mu = m_X$.

Since we may assume that $x_0 \in X$ is not periodic for U^-, we know from Lemma 11.29 that there exists a sequence of times $t_n \to \infty$ for the geodesic

flow and a fixed compact subset $K \subseteq X$ such that $R_{a_{t_n}}(x_0) \in K$ for $n \geqslant 1$. Proposition 11.30 therefore finishes the proof. □

Proposition 11.30. *Let $K \subseteq X$ be a compact set. Then there exists some constant $\eta > 0$ with the following property for all $x_0 \in X$. Suppose that (t_n) is a sequence in $\mathbb{R}$ with $t_n \to \infty$, and with $R_{a_{t_n}}(x_0) \in K$ for all n. Then*

$$\frac{1}{\eta \mathrm{e}^{t_n}} \int_0^{\eta \mathrm{e}^{t_n}} f\left(h(s) \cdot x_0\right) \mathrm{d}s \longrightarrow \int f \, \mathrm{d}m_X \tag{11.15}$$

as $n \to \infty$ for all $f \in C_c(X)$.

The basic idea of the proof – ignoring for the moment the slightly mysterious constant η – is as follows. Since f is uniformly continuous, we may replace the left-hand side of equation (11.15) by an integral over a slightly thickened tubular neighborhood B_n of the piece

$$\left\{x_0 u^-(-s) \mid s \in [0, \mathrm{e}^{t_n}]\right\} = x_0 u^-\left(-[0, \mathrm{e}^{t_n}]\right) \tag{11.16}$$

of the U^--orbit of x_0. We wish to do this in such a way that the image of B_n under $R_{a_{t_n}}$ is easy to describe. Note that the piece of the U^--orbit in equation (11.16) is mapped under $R_{a_{t_n}}$ to the set

$$R_{a_{t_n}}(x_0) u^-\left(-[0,1]\right).$$

Below we will define a set $Q_\delta \subseteq \mathrm{PSL}_2(\mathbb{R})$ which contains $u^-\left(-[0,1]\right)$ and which may be described as a cube with one of the sides being $u^-\left(-[0,1]\right)$. Define

$$B_n = R_{a_{t_n}}^{-1}\left(R_{a_{t_n}}(x_0) Q_\delta\right) = x_0\left(a_{t_n}^{-1} Q_\delta a_{t_n}\right).$$

We will show that B_n is a slight thickening of the set defined in equation (11.16). Using this we will be able to show

$$\frac{1}{e^{t_n}} \int f\left(h(s) \cdot x_0\right) \mathrm{d}s \underset{\varepsilon}{\approx} \frac{1}{m_X(B_n)} \int_{B_n} f(x) \, \mathrm{d}x = \frac{1}{m_X(B_n)} \left\langle f, \chi_{B_n} \right\rangle.$$

Moreover, B_n was defined as a pre-image under $R_{a_{t_n}}$, and together with the mixing property we expect

$$\left\langle f, \chi_{B_n} \right\rangle \longrightarrow \int f \, \mathrm{d}m_X \cdot m_X\left(R_{a_{t_n}}(x_0) Q_\delta\right) = \int f \, \mathrm{d}m_X \cdot m_X(B_n). \tag{11.17}$$

However, B_n is not defined as the pre-image of a fixed set in X, so the mixing statement does not apply directly. Roughly speaking, the set B_n is defined as the pre-image of a set whose "shape" Q_δ is fixed but whose position $R_{a_{t_n}}(x_0)$ is allowed to vary. Here the assumption that $R_{a_{t_n}}(x_0) \in K$ is crucial – it will allow us to use mixing to prove equation (11.17).

To make the above outline formal, we need a basic decomposition lemma for the Haar measure[(100)].

Lemma 11.31. *Let G be a σ-compact unimodular group and let $S, T \subseteq G$ be closed subgroups with the property that $S \cap T = \{e\}$ and the product set ST contains a neighborhood of $e \in G$. Let $\phi : S \times T \to G$ be the product map $\phi(s,t) = st \in ST \subseteq G$. Then the Haar measure m_G restricted to ST is proportional to the push-forward $\phi_* \left(m_S^\ell \times m_T^r\right)$ where m_S^ℓ is the left Haar measure on S and m_T^r is the right Haar measure on T.*

Clearly if $G \smallsetminus ST$ has Haar measure zero, then the above lemma gives a complete description of m_G in the coordinate system defined by the subgroups S and T.

We will apply the lemma in the case where $G = \mathrm{PSL}_2(\mathbb{R})$, $S = U^-$, and

$$T = U^+A = \left\{ \begin{pmatrix} a & \\ t & a^{-1} \end{pmatrix} \middle| \, t \in \mathbb{R}, a \in \mathbb{R}^\times \right\}.$$

Notice that $S = K$ and $T = U^+A$ is another choice, that was already discussed implicitly in Lemma 9.16.

Proof of Lemma 11.31. Since $S \cap T = \{e\}$, an element $g \in G$ has at most one decomposition as $g = st$ with $s \in S$ and $t \in T$. Therefore for compact subsets $K_S \subseteq S$ and $K_T \subseteq T$ the map ϕ restricted to $K_S \times K_T$ is a homeomorphism, so $\phi^{-1} : ST \to S \times T$ is measurable. The same applies to the map $\psi : S \times T \to ST$ defined by $\psi(s,t) = \phi(s,t^{-1}) = st^{-1}$.

Let $\nu = \left(\psi^{-1}\right)_* m_G$. We consider $S \times T$ as a σ-compact group by using coordinatewise multiplication. Then, for $B \subseteq S \times T$ and a point $(s,t) \in S \times T$, we have

$$\nu\left((s,t)B\right) = m_G\left(\psi\left((s,t)B\right)\right) = m_G\left(s\psi(B)t^{-1}\right) = m_G\left(\psi(B)\right) = \nu(B),$$

since G is unimodular. It follows that ν is a left Haar measure on $S \times T$ and so must be proportional to $m_S^\ell \times m_T^\ell$. Now ϕ and ψ differ only by the inverse in the second component, and the inverse map sends m_T^ℓ to a measure proportional to m_T^r, so the lemma follows. □

As mentioned above, we are interested in the case $G = \mathrm{PSL}_2(\mathbb{R})$, $S = U^-$, and $T = U^+A$. Clearly, in this case $S \cap T = \{e\}$ and S, T are closed subgroups of G. All that needs to be checked is that ST contains a neighborhood of e. This may be seen* from the inverse function theorem, since both $S \times T$ and G are three-dimensional, and the derivative of the multiplication map has full rank at the identity (e,e). We choose the Haar measure $m_{U^-}^\ell$ to be the usual Lebesgue measure on $\mathbb{R}$ under the identification of $s \in \mathbb{R}$ with $u^-(s) \in U^-$,

* More formally, we can consider the Lie algebras $\mathfrak{u}^-$ of U^- and $\mathfrak{t}$ of $T = U^+A$ and the map $\mathfrak{u}^- \times \mathfrak{t} \to \mathfrak{g}$ defined on a neighborhood of 0 by

$$(v,w) \mapsto \log(\exp(v)\exp(w)).$$

It may be checked that the derivative of this map at 0 is the embedding $\mathfrak{u}^- \times \mathfrak{t} \to \mathfrak{g}$ defined by $(v,w) \mapsto v + w$.

and we choose the right Haar measure m_T^r on T so that m_G restricted to U^-T coincides with the product measure of $m_{U^-}^\ell$ and m_T^r.

Proof of Proposition 11.30. Let $\eta = \eta(K) > 0$ be chosen so that

$$u^-\left(-[0,\eta]\right) B_\eta^T \ni g \longmapsto yg$$

is injective for all $y \in K$.

Let f be a function in $C_c(X)$. It is sufficient to prove equation (11.15) for non-negative functions, so assume that $f(x) \geqslant 0$ for all $x \in X$. Fix $\varepsilon > 0$ and choose by uniform continuity of f some $\delta \in (0,\delta)$ such that

$$\mathsf{d}(x,y) < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Recall that $T = U^+A$ and define

$$Q_\delta = u^-\left(-[0,\eta]\right) B_\delta^T.$$

Then by choice of η, for any $y \in K$ the map $g \mapsto yg$ is injective on Q_δ. Let

$$B_n = R_{a_{t_n}}^{-1}\left(R_{a_{t_n}}(x_0)Q_\delta\right) = x_0\left(a_{t_n}^{-1}Q_\delta a_{t_n}\right) \subseteq x_0\left(u^-\left(-[0,\eta\mathrm{e}^{t_n}]\right) B_\delta^T\right),$$

where the last inclusion may be seen by noting that conjugation by $a_{t_n}^{-1}$ contracts U^+ and expands U^-. Now for any $s \in [0,\eta\mathrm{e}^{t_n}]$ and $h \in B_\delta^T$ we have

$$\mathsf{d}_X\left(x_0u^-(-s), x_0u^-(-s)h\right) \leqslant \mathsf{d}_G(e,h) < \delta$$

and so

$$|f(x_0u^-(-s)) - f(x_0u^-(-s)h)| < \varepsilon.$$

By Lemma 11.31 and the discussion after it, we deduce that

$$\frac{1}{m(B_n)}\int_{B_n} f(x)\,\mathrm{d}m_X = \frac{1}{m(B_n)}\int_{a_{t_n}^{-1}Q_\delta a_{t_n}} f(x_0g)\,\mathrm{d}m_G(g)$$

$$= \frac{1}{\eta\mathrm{e}^{t_n}}\int_0^{\eta\mathrm{e}^{t_n}} \frac{1}{m_T^r(a_{t_n}^{-1}B_\delta^T a_{t_n})}\int_{a_{t_n}^{-1}B_\delta^T a_{t_n}} f(x_0u^-(-s)h)\,\mathrm{d}m_T^r(h)\,\mathrm{d}s$$

is within ε of

$$\frac{1}{\eta\mathrm{e}^{t_n}}\int_0^{\eta\mathrm{e}^{t_n}} f(x_0u^-(-s))\,\mathrm{d}s.$$

Next we are going to construct finitely many subsets of X to which the mixing property can be applied. Note that

$$\overline{Q}_\delta = u^-(-[0,\eta])\overline{B_\delta^T}$$

is compact, and the set $Q_\delta^o = u^-(-(0,\eta))B_\delta^T$ has

$$m_G(Q_\delta) = m_G(\overline{Q}_\delta) = m_G(Q_\delta^o).$$

It follows by regularity that there is a compact subset $P_\delta \subseteq Q_\delta^o$ and an open set $R_\delta \supseteq \overline{Q}_\delta$ with

$$m_G\left(R_\delta \smallsetminus P_\delta\right) < \frac{\varepsilon}{m_G(Q_\delta)}. \tag{11.18}$$

This implies that $B_\kappa^G P_\delta \subseteq Q_\delta$ and $B_\kappa^G Q_\delta \subseteq R_\delta$ for a sufficiently small $\kappa > 0$. By compactness we may choose finitely many points $y_1, \ldots, y_\ell \in K$ with

$$K \subseteq y_1 B_\kappa^G \cup \cdots \cup y_\ell B_\kappa^G.$$

Since the geodesic flow is mixing, we have

$$\frac{m_G(P_\delta)}{m_G(Q_\delta)} \int f \,\mathrm{d}m_G - \varepsilon \leqslant \left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{y_i P_\delta} \circ R_{a_{t_n}} \right\rangle \tag{11.19}$$

and

$$\left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{y_i R_\delta} \circ R_{a_{t_n}} \right\rangle \leqslant \frac{m_G(R_\delta)}{m_G(Q_\delta)} \int f \,\mathrm{d}m_G + \varepsilon \tag{11.20}$$

for $i = 1, \ldots, \ell$ and all large enough n.

Therefore, if $x = R_{a_{t_n}}(x_0) \in y_i B_\eta^G$ then $y_i P_\delta \subseteq x Q_\delta \subseteq y_i R_\delta$, and since f is non-negative this shows that

$$\left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{y_i P_\delta} \circ R_{a_{t_n}} \right\rangle \leqslant \left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{B_n} \right\rangle \leqslant \left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{y_i R_\delta} \circ R_{a_{t_n}} \right\rangle$$

and so, by the inequalities (11.19)–(11.20) and (11.18),

$$(1-\varepsilon) \int f \,\mathrm{d}m_X - \varepsilon \leqslant \left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{B_n} \right\rangle \leqslant (1+\varepsilon) \int f \,\mathrm{d}m_X + \varepsilon.$$

By combining this with the earlier statement that

$$\left\langle f, \frac{1}{m_G(Q_\delta)} \chi_{B_n} \right\rangle$$

and

$$\frac{1}{\eta \mathrm{e}^{t_n}} \int_0^{\eta \mathrm{e}^{t_n}} f(x_0 u^-(-s)) \,\mathrm{d}s$$

are ε-close, the proposition and Theorem 11.27 follow. □

While the proof of Proposition 11.30 shows that either x is periodic for the horocycle flow or that certain long ergodic averages of a function $f \in C_c(X)$ are close to the integral for the Haar measure, it does not establish that in the latter case x is in fact generic for m_X (unless X is compact). This will be proved in Section 11.7.

Exercises for Section 11.5

Exercise 11.5.1. Suppose that $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is compact. Let $u \in \mathrm{SL}_2(\mathbb{R})$ be a nontrivial unipotent matrix. Prove that the map $R_u : X \to X$ is uniquely ergodic. Generalize the statement and proof to non-compact quotients by lattices. (Note that Theorem 11.27 deals with the case of the $\mathbb{R}$-flow only.)

Exercise 11.5.2. Suppose that $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is not compact. Show that

$$\frac{1}{t(y_n)} \int_0^{t(y_n)} f\left(h(t)\cdot y_n\right)\, \mathrm{d}t \longrightarrow \int f \,\mathrm{d}m_X$$

as $n \to \infty$ for any $f \in C_c(X)$, if each $y_n \in X$ is periodic with least period $t(y_n)$ for the horocycle flow, and $t(y_n) \to \infty$ as $n \to \infty$.

Exercise 11.5.3. Let $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ be the quotient by a lattice Γ. Prove (without invoking ergodic decomposition Theorem 6.2) that a probability measure μ on X, invariant under the horocycle flow $h(s)$ for all $s \in \mathbb{R}$, that gives zero measure to the set of all periodic orbits of the horocycle flow must be the Haar measure of X.

11.6 Non-escape of Mass for Horocycle Orbits

We have seen that the geodesic flow R_{a_t} and the horocycle flow $h(s) = R_{u^-(s)}$ have fundamentally different types of behavior. For example, we briefly discussed in Section 9.7.2 the fact that orbits for the geodesic flow can be quite erratic; on the other hand we will show in the next section that an orbit for the horocycle flow will either be periodic (and hence compact) or will be equidistributed in the ambient space. Recall that the latter property means

$$\frac{1}{T} \int_0^T f\left(h(t)\cdot x\right)\, \mathrm{d}t \longrightarrow \int_X f \,\mathrm{d}m_X$$

as $T \to \infty$. In order to prove this, we first wish to show that any limit of a sequence of measures of the form

$$\frac{1}{T_n} \int_0^{T_n} \delta_{x u^-(-t)}\, \mathrm{d}t \tag{11.21}$$

with $T_n \to \infty$ as $n \to \infty$, is indeed a probability measure. This property – that limit points of sequences of uniform measures on long orbits are probability measures – is often called *quantitative non-divergence* or *non-escape of mass*[101]. Clearly this does not hold for the geodesic flow: for example, the orbit of the point (i, i) in $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ is strictly divergent (see Figure 9.1 and the discussion in Section 9.7.2) and so any limit measure along this orbit of the geodesic flow must be zero.

On a positive note, this kind of strict divergence seen in the geodesic orbit of (i,i) is clearly impossible for the horocycle flow. More precisely, for any $x \in \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ there exists a compact set $L \subseteq \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ and a sequence $t_n \nearrow \infty$ such that $h(t_n)\cdot x \in L$ for all $n \geqslant 1$. This property is often called *non-divergence* (and is due to Margulis in much more general situations). To see why this property holds for the horocycle flow, recall that horocycle orbits can be drawn in the upper half-plane $\mathbb{H}$ as circles touching the real axis, or as horizontal lines. This is easy to see since the latter is the orbit of (i,i), and Möbius transformations map horizontal lines either to horizontal lines or to circles tangent to the real axis (see Figure 9.3 and Section 9.2). We conclude that given any $x \in \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$, either x is periodic for the horocycle flow, in which case the orbit is compact and the non-divergence statement is trivial, or x is represented in the fundamental domain F from Figure 9.5 by a vector not pointing straight upwards. In the latter case we choose L to be the closure of the set of all vectors with base point $z \in F$ satisfying $\Im(z) = 1$. As the circle comes back (in both directions) to $\Im(z) = 1$, we can find some $t_1 > 0$ with $h(t_1)\cdot x \in L$. Applying the same argument with $h(t_1 + 1)\cdot x$ in place of x leads to some $t_2 > t_1$ for which $h(t_2)\cdot x \in L$, and so on.

We next state the quantitative non-divergence theorem, which (in a more general setting) is Dani's refinement of Margulis' non-divergence theorem.

Theorem 11.32. *For every lattice $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$, every compact subset K in $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$, and every $\varepsilon > 0$, there exists a compact subset $L = L(K,\varepsilon)$ of X such that*

$$m_{\mathbb{R}}\Big(\{t \mid 0 \leqslant t \leqslant T, h(t)\cdot x \notin L\}\Big) \leqslant \varepsilon T \tag{11.22}$$

for all $T > 0$ and all $x \in K$. Moreover, there is a compact set $L = L(\varepsilon) \subseteq X$ (independent of K and of x) such that for any $x \in X$ either x is periodic or there exists some $T_x > 0$ such that equation (11.22) *holds for all $T \geqslant T_x$.*

Notice that the first claim gives (in a more uniform way than needed here) the earlier claim, that for any $x \in X$ and any convergent sequence of measures of the form (11.21), the limit μ is a probability measure. Indeed, to show that $\mu(X) > 1 - \varepsilon$ one only needs to choose a continuous function $f \in C_c(X)$ with $\chi_L \leqslant f \leqslant 1$ and apply the definition of weak*-convergence, where L is chosen as in Theorem 11.32 for $K = \{x\}$.

Before starting the proof we record the following fundamental difference between the behavior of polynomials (which model some aspects of the horocycle flow) and the exponential function (which models some aspects of the geodesic flow).

Let $p \in \mathbb{R}[t]$ be a polynomial with small coefficients and with $p(T) = 1$. Then a fixed positive proportion of $t \in [0,T]$ has $p(t) > \frac{1}{2}$. Moreover, the Lebesgue measure of the set

$$\{t \in [0,T] \mid |p(t)| < \varepsilon\}$$

is (for small ε) small compared to T. Here it is important to note that the quality of these statements is independent of the size of T, which may be seen by rescaling the polynomial $p(t)$ on the interval $[0,T]$ to give the polynomial $q(t) = p(tT)$ on $[0,1]$. Neither of these properties hold for an exponential function $g(t) = ae^{\pm t}$.

11.6.1 The Space of Lattices and the Proof of Theorem 11.32 for $X_2 = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$

In proving this theorem it will be helpful to think of $X_2 = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ and, more generally, $X_d = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$, in a slightly different way. A lattice $\Lambda \subseteq \mathbb{R}^d$ is called *unimodular* if the quotient $\mathbb{R}^d/\Lambda$ has a fundamental domain of Lebesgue volume 1; equivalently if it has *covolume* 1. Any unimodular lattice has the form $\Lambda_g = g^{-1}\mathbb{Z}^d \subseteq \mathbb{R}^d$ for some $g \in \mathrm{SL}_d(\mathbb{R})$. When we wish to emphasize the meaning of a point $x \in X_d$ in the sense of a lattice in $\mathbb{R}^d$, then we will simply use the symbol $\Lambda \in X_d$.

Moreover, matrices $g_1, g_2 \in \mathrm{SL}_d(\mathbb{R})$ define the same lattice if and only if $g_2 g_2^{-1} \in \mathrm{SL}_d(\mathbb{Z})$. Thus X_d can be identified with the space of unimodular lattices. This gives a geometrical interpretation of the space X_d; in particular Mahler's compactness criterion [241], which says that a sequence of elements of X_d diverge to infinity if and only if the distance from the origin in $\mathbb{R}^d$ to the set of non-trivial lattice elements converges to zero.

Theorem 11.33. [Mahler compactness criterion] *A set $K \subseteq X_d$ has compact closure if and only if there is an $s > 0$ with the property that*

$$\Lambda \cap B_s(0) = \{0\}$$

for all $\Lambda \in K$.

One way of formulating this result is as follows. Lattices $\Lambda \subseteq \mathbb{R}^d$ are clearly discrete, and so compact subsets of lattices in $\mathbb{R}^d$ should be uniformly discrete.

Proof of Theorem 11.33. We start by showing that a compact subset K of X_d must have the uniform discreteness property. Suppose therefore that K has compact closure but for every $n \geqslant 1$ there is some $\Lambda_n \in K$ with

$$\Lambda_n \cap B_{1/n}(0) \neq \{0\}.$$

By compactness there exists some $\Lambda \in X_d$ with $\Lambda_n \to \Lambda$. By definition, this means that $\Lambda_n = g_n^{-1}\mathbb{Z}^d$, $\Lambda = g^{-1}\mathbb{Z}^d$, and we can choose g_n (which is only unique up to left multiplication by $\mathrm{SL}_d(\mathbb{Z})$) such that $\mathsf{d}(g_n, g) \to 0$ as $n \to \infty$. Equivalently, $g_n^{-1} \to g^{-1}$ as $n \to \infty$, which in terms of the lattices means that one can choose a basis $\{\mathbf{b}_j^{(n)} = g_n^{-1}\mathbf{e}_j \mid 1 \leqslant j \leqslant d\}$ of Λ_n which converges to a basis $\{\mathbf{b}_j = g^{-1}\mathbf{e}_j \mid 1 \leqslant j \leqslant d\}$ of Λ. However, as (g_n) and (g_n^{-1}) both

converge, it follows that the lattices $\Lambda_n = g_n^{-1}\mathbb{Z}^d$ cannot contain arbitrarily small elements of $\mathbb{R}^d$, which contradicts our choice and proves the easier half of the theorem.

We now claim that for every lattice Λ with $\Lambda \cap B_s(0) = \{0\}$ one can find a basis of vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \Lambda$ that belong to a given ball of radius depending on s. Choose $\mathbf{b}_1 \in \Lambda$ with the property that

$$\|\mathbf{b}_1\| = \min\{\|\mathbf{b}\| \mid \mathbf{b} \in \Lambda \smallsetminus \{0\}\}, \tag{11.23}$$

so $\|\mathbf{b}_1\| \geqslant s$, and $\mathbf{b}_1$ generates $\Lambda \cap \mathbb{R}\mathbf{b}_1$. Moreover, there is a constant C_d depending only on the dimension d with $\|\mathbf{b}_1\| \leqslant C_d$, since if all the vectors in $\Lambda \smallsetminus \{0\}$ are very long, then it cannot be unimodular by Minkowski's convex body theorem (which simply relies on the argument that if $B_{2r}(0)$ does not contain a lattice element of Λ, then $B_r(0)$ is mapped injectively onto $\mathbb{R}^d/\Lambda$, which makes the volume of the fundamental domain of Λ at least as large as the volume of $B_r(0)$). Define

$$W = (\mathbb{R}\mathbf{b}_1)^{\perp}$$

and

$$\Lambda_W = \pi_W(\Lambda)$$

where $\pi_W : \mathbb{R}^d \to W$ is the projection along the line $\mathbb{R}\mathbf{b}_1$. The $(d-1)$-dimensional lattice Λ_W need not be unimodular, but the covolume is $\frac{1}{\|\mathbf{b}_1\|}$, which is uniformly bounded away from 0 and infinity. So after rescaling by a bounded scalar, we may assume that the lattice Λ_W is unimodular. We claim that all the non-zero vectors in Λ_W have length bounded away from 0 by a uniform amount. For, if not, a very short vector $\pi_W(\mathbf{x}) \in \Lambda_W$ with $\mathbf{x} \in \Lambda$ would have the property that $\mathbf{x} + n\mathbf{b}_1$ is closer to 0 than $\mathbf{b}_1$ for some $n \in \mathbb{Z}$, contradicting equation (11.23) (see Figure 11.13).

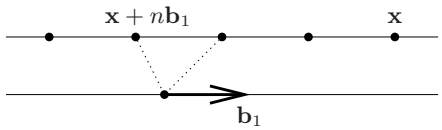

Fig. 11.13. Choosing the vector $\mathbf{x} + n\mathbf{b}_1$.

By induction, this shows that Λ_W has a basis consisting of vectors whose length is bounded uniformly away from 0 and infinity that can be lifted to complete the basis $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$. The length of $\mathbf{b}_j$ for $j \geqslant 2$ can be chosen to be less than $\sqrt{\|\pi_W(\mathbf{b}_j)\|^2 + \|\mathbf{b}_1\|^2}$, which is bounded by a number depending on d. It follows that for any sequence $\Lambda_n \in K$ we can write $\Lambda_n = g_n^{-1}\mathbb{Z}^d$ where the columns of g_n^{-1} are uniformly bounded (only depending on s). However, as $\det(g_n) = 1$ we know that g_n belongs to a fixed compact set of $\mathrm{SL}_d(\mathbb{R})$, so $K \subseteq X_d$ is contained in the compact image of that set, giving the result. □

We now have the background needed to proceed with the proof of quantitative non-divergence in the case $X_2 = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$. As the proof will show, the theorem simply relies on the polynomial (in fact, linear) behavior of orbits of the horocycle flow, together with the simple observation that a unimodular lattice in $\mathbb{R}^2$ cannot contain two linearly independent vectors of norm less than one.

Proof of Theorem 11.32. By Theorem 11.33 we have

$$K \subseteq \Omega_\delta = \{\Lambda \in X_2 \mid \Lambda \cap B_\delta(0) = \{0\}\}$$

for some $\delta \in (0,1)$. We need to choose some $\eta \in (0, \frac{\delta}{4})$ such that the set $L = \Omega_\eta$ satisfies equation (11.22) for all $\Lambda \in K$ and $T > 0$.

Note that an element $h(t)\cdot x = xu^-(-t)$ of the horocycle flow maps the lattice $\Lambda = g^{-1}\mathbb{Z}^2$ corresponding to $x = \Gamma g \in X_2$ to $u^-(t)\Lambda$. In other words, the horocycle flow corresponds to application of the matrix $\begin{pmatrix} 1 & t \\ & 1 \end{pmatrix}$, which fixes the x-axis and shears the y-axis towards the direction of the x-axis. Specifically, the image of a vector $\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ under $\begin{pmatrix} 1 & t \\ & 1 \end{pmatrix}$ is $\begin{pmatrix} \alpha_1 + t\alpha_2 \\ \alpha_2 \end{pmatrix}$ for any $t \in \mathbb{R}$, so the horocycle flow moves this point at linear speed determined by α_2 through the plane.

When applying the horocycle flow to Λ there could be different lattice elements $v \in \Lambda$ that at some time t give rise to a short vector $u^-(t)v$ in $u^-(t)\Lambda$, which prevents $u^-(t)\Lambda$ from belonging to $L = \Omega_\eta$. For this reason we let $\{v_1, v_2, \dots\} \subseteq \Lambda$ be a maximal set of mutually non-proportional primitive* elements of Λ. Notice that if a vector $u^-(t)v \in u^-(t)\Lambda \smallsetminus \{0\}$ has norm less than η, then v is a multiple of a primitive vector $v = nv_j$, and so $u^-(t)v_j$ will have norm less than η. Thus it is enough to consider only the orbits of the primitive vectors $v_1, v_2, \dots$.

Recall that $\Lambda \in \Omega_\delta$ by assumption, and so $\|v_i\| \geqslant \delta$ for $i = 1, 2, \dots$. Fix some $T > 0$. Then for each $i \geqslant 1$, we define

$$B_i = \{t \in [0,T] \mid u^-(t)v_i \in B_\eta^{\mathbb{R}^2}(0)\}$$

(the set of *bad times* in $[0,T]$) and

$$P_i = \{t \in [0,T] \mid u^-(t)v_i \in B_\delta^{\mathbb{R}^2}(0)\}$$

(the set of *protecting times* in $[0,T]$). By assumption $\eta < \delta$, so $B_i \subseteq P_i$. If $i \neq j$ then $P_i \cap P_j = \varnothing$, since if $t \in P_i \cap P_j$ then $u^-(t)v_i$ and $u^-(t)v_j$ both lie in the unimodular lattice $u^-(t)\Lambda$ and have norm less than $\delta < 1$, so they are linearly dependent†, and hence $i = j$ by construction.

* A vector $v \in \Lambda$ is called primitive if the equation $v = nw$ with $n \in \mathbb{Z}$ and $w \in \Lambda$ implies that $n = \pm 1$.

† The determinant of a 2×2 matrix (w_1, w_2) formed by the column vectors w_1, w_2 is bounded by the product of their Euclidean norms, since it is equal to $\|w_1\|\|w_2\| \cos\phi$, where ϕ is the angle between the vectors. Moreover, if w_1, w_2 lie in the lattice $\Lambda = \mathbf{b}_1\mathbb{Z} + \mathbf{b}_2\mathbb{Z}$, then $\mathrm{covolume}(\Lambda) = |\det(\mathbf{b}_1, \mathbf{b}_2)| \leqslant |\det(w_1, w_2)|$.

Finally, we claim that

$$m_{\mathbb{R}}(B_i) \leqslant \frac{8\eta}{\delta} m_{\mathbb{R}}(P_i). \tag{11.24}$$

This implies the first claim in the theorem. Indeed, summing over i and using disjointness of the sets P_i the inequality (11.24) gives

$$m_{\mathbb{R}}\left(\bigsqcup_{i\geqslant 1} B_i\right) \leqslant \frac{8\eta}{\delta} m_{\mathbb{R}}\left(\bigsqcup_{i\geqslant 1} P_i\right) = \frac{8\eta}{\delta} \sum_{i\geqslant 1} m_{\mathbb{R}}(P_i) \leqslant \frac{8\eta}{\delta} T.$$

However, as discussed above,

$$\begin{aligned} m_{\mathbb{R}}\left(\{t \mid u^-(t)\Lambda \notin \Omega_\eta\}\right) &= m_{\mathbb{R}}\left(\{t \mid \|u^-(t)v_i\| < \eta \text{ for some } i\}\right) \\ &= m_{\mathbb{R}}\left(\bigcup_{i\geqslant 1} B_i\right). \end{aligned}$$

Choosing $\eta \leqslant \frac{\varepsilon\delta}{8}$ gives the estimate needed.

To prove the inequality (11.24), let $v_i = \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$ so that

$$u^-(t)v_i = \begin{pmatrix}\alpha_1 + t\alpha_2\\ \alpha_2\end{pmatrix}.$$

Then $u^-(t)v_i$ moves at linear speed along a horizontal line, as in Figure 11.14.

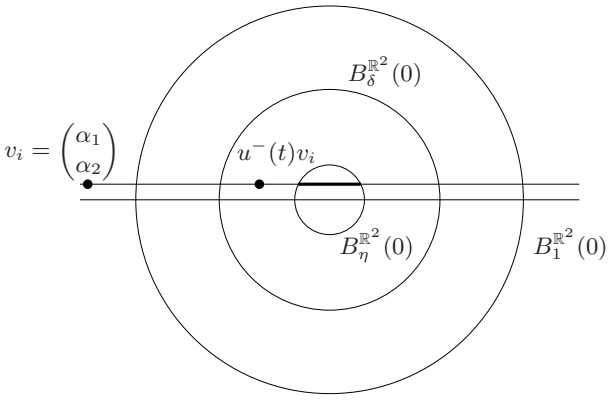

Fig. 11.14. The linear speed in the horocycle flow makes it easy to describe the lengths of the vectors in the orbit.

Clearly B_i and P_i are both intervals. To prove the inequality (11.24), we have to relate the lengths of these intervals to the speed α_2 of the vector $u^-(t)v_i$ and to the radii η and δ. If $B_i \subseteq [0, T]$ is empty, then there is

nothing to prove. This holds, in particular, if $\alpha_2 = 0$ or if $|\alpha_2| \geqslant \frac{\delta}{4} \geqslant \eta$. Otherwise, $P_i = [T_1, T_2]$ must contain some T' with $|\alpha_1 + T'\alpha_2| < \eta \leqslant \frac{\delta}{4}$ while $\|u^-(T_1)v\| = \delta$ gives $|\alpha_1 + T_1\alpha_2| \geqslant \frac{\delta}{2}$ since $|\alpha_2| \leqslant \frac{\delta}{2}$. Thus

$$m_{\mathbb{R}}(P_i) = (T_2 - T_1) \geqslant (T' - T_1) \geqslant |\alpha_2|^{-1}\frac{\delta}{4}.$$

On the other hand, $t \in B_i = [T_3, T_4]$ implies $|\alpha_1 + t\alpha_2| \leqslant \eta$, so

$$m_{\mathbb{R}}(B_i) = T_4 - T_3 < \frac{2\eta}{|\alpha_2|}.$$

Together, this gives the inequality (11.24).

For the final claim let $K = \Omega_{1/2}$ and apply the first statement of the theorem to $\varepsilon/2$ to define the set L. Now let $x \in X_2$, and notice that x is periodic for the horocycle flow if and only if the corresponding lattice $\Lambda \subseteq \mathbb{R}^2$ intersects the x-axis. Now assume that $x \in X_2$ is not periodic, so that Λ may have a very short primitive vector $v \in \Lambda$ that is not fixed. Therefore we may choose $T_1 > 0$ such that $u^-(T_1)v$ has norm between $\frac{1}{2}$ and 1. As in the argument above, this shows that $\pm u^-(T_1)v$ are the only non-trivial vectors of norm less than 1 for $u^-(T_1)\Lambda$, so that $xu^-(-T_1) \in K$. We choose T_x so that $T_1 = \frac{\varepsilon}{2}T_x$, and it is easy to check that equation (11.22) holds for all $T \geqslant T_x$. □

11.6.2 Extension to the General Case

Notice first that Theorem 11.32, in the case of quotients $\Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$ by lattices Γ in $\mathrm{PSL}_2(\mathbb{R})$, also implies the same statement for quotients $\Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ by lattices in $\mathrm{SL}_2(\mathbb{R})$. Indeed, if $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ is a lattice, then

$$\overline{\Gamma} = \Gamma \cdot \{\pm I\} \subseteq \mathrm{SL}_2(\mathbb{R})/\{\pm I\} = \mathrm{PSL}_2(\mathbb{R})$$

is also a lattice. Moreover, the natural map

$$\Gamma\backslash \mathrm{SL}_2(\mathbb{R}) \longrightarrow \overline{\Gamma}\backslash \mathrm{PSL}_2(\mathbb{R})$$

is proper (that is, every compact set $\overline{L} \subseteq \overline{\Gamma}\backslash \mathrm{PSL}_2(\mathbb{R})$ has a compact preimage L in $\Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. From this the earlier claim follows quickly.

With this reduction, we may use the geometry of $\mathbb{H}$ and Dirichlet domains. Indeed, we now state a description of $\overline{\Gamma}\backslash \mathrm{PSL}_2(\mathbb{R})$ which we essentially proved together with Lemma 11.29.

Proposition 11.34. *Let $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$ be a lattice. Then either Γ is uniform and $\Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$ is compact, or there exists a compact subset*

$$\Omega_{cp} \subseteq X = \Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$$

such that $X \smallsetminus \Omega_{cp}$ has finitely many connected components, each of which can be identified with a special finite volume subset of the infinite volume quotient

$$\mathscr{T} = \left\{ \begin{pmatrix} 1 & n \\ & 1 \end{pmatrix} \right\} \backslash \operatorname{PSL}_2(\mathbb{R}).$$

Indeed, for each component C of $X \smallsetminus \Omega_{cp}$ we find some $y_C > 0$ such that the subset $\mathscr{T}_C \subseteq \mathscr{T}$ is the image of $\{(z, v) \in \mathrm{T}^1\mathbb{H} \mid \Im(z) > y_C\}$, and the identification $\imath_C : C \to \mathscr{T}_C$ has the property that for a sufficiently small element $g \in \operatorname{PSL}_2(\mathbb{R})$ and any $x \in C$ with $x \bullet g \in C$ we have $\imath_C(x \bullet g) = \imath_C(x)g$.

Roughly speaking, the components of $X \smallsetminus \Omega_{cp}$ are the cusps of X.

Proof of Proposition 11.34. We will leave some of the details of the proof as an exercise, but indicate how the proof of Lemma 11.29 applies to Proposition 11.34. We defined a cusp of $X = \Gamma \backslash \operatorname{PSL}_2(\mathbb{R})$ to be an equivalence class of boundary vertices of a Dirichlet domain D for Γ, and showed in the proof of Lemma 11.29 that for every boundary vertex r there exists a non-trivial unipotent element of Γ fixing the boundary vertex r. Without loss of generality, we may assume that $r = \infty$. While proving this we used several other elements $\gamma_i \in \Gamma$, and the respective images $\gamma_i(D)$ of the Dirichlet domain. At the end of this argument we had obtained the copies $\gamma_1(D), \gamma_1\gamma_2(D), \ldots, \gamma_1\gamma_2 \cdots \gamma_n(D)$ of D which were all located to the right of D meeting each other full-edge to full-edge on vertical geodesics. Let

$$F = D \cup \gamma_1(D) \cup \gamma_1\gamma_2(D) \cup \cdots \cup \gamma_1\gamma_2 \cdots \gamma_n(D)$$

be the union of these domains. By construction, the quotient map

$$\mathrm{T}^1\mathbb{H} \to \Gamma \backslash \operatorname{PSL}_2(\mathbb{R})$$

restricted to $\{(z, v) \in \mathrm{T}^1\mathbb{H} \mid z \in F, \Im(z) > y_\infty\}$ is injective if $y_\infty > 0$ is chosen sufficiently large. Denote this injective image by $C_\infty \subseteq X$. By applying the Möbius transformation $\begin{pmatrix} a & \\ & a^{-1} \end{pmatrix}$ for some $a > 0$ we can ensure that the unipotent element $\gamma = \gamma_1 \cdots \gamma_n$ has the form $\begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$.

If necessary we may repeat this construction for the remaining cusps, and can ensure disjointness by choosing the associated parameters y sufficiently big. After completion of this argument for every cusp D, the pre-images of the various sets give a set whose image in X has compact closure. □

Sketch Proof of Theorem 11.32 for General Γ. As mentioned before, it is enough to treat the case of a lattice $\Gamma \subseteq \operatorname{PSL}_2(\mathbb{R})$. Applying Proposition 11.34, we can find a compact subset $\Omega_{cp} \subseteq X = \Gamma \backslash \operatorname{PSL}_2(\mathbb{R})$ and finitely many tube-like sets $C_1, \ldots, C_\ell$ which together give a partition

$$X = \Omega_{cp} \sqcup C_1 \sqcup \cdots \sqcup C_\ell.$$

Enlarging Ω_{cp} if necessary, we may assume that $K \subseteq \Omega_{cp}$ and that each C_i is naturally identified with a tube

$$\left\{ \begin{pmatrix} 1 & n \\ & 1 \end{pmatrix} \mid n \in \mathbb{Z} \right\} \setminus \left\{ (z, v) \in \mathrm{T}^1\mathbb{H} \mid \Im(z) > y_i \right\} = \mathscr{T}(y_i)$$

with $y_i > 1$. We claim that for each C_i there is a subset $L_i \subseteq C_i$ with compact closure such that

$$L = \Omega_{cp} \sqcup L_1 \sqcup \cdots \sqcup L_\ell$$

satisfies the first statement of Theorem 11.32.

Let us now describe the form of non-divergence we will prove (more precisely, that we have already proved) for the tube-like sets $\mathscr{T}(y)$ with $y > 1$. There exists a compact set $L \subseteq \overline{\mathscr{T}(y)}$ such that for any $x \in \partial\mathscr{T}(y)$ and any $T > 0$, either $h(T)\cdot x \notin \overline{\mathscr{T}(y)}$ or

$$m_{\mathbb{R}} \left(\{ t \in [0, T] \mid h(t)\cdot x \notin L \} \right) < \varepsilon T. \tag{11.25}$$

Notice that $x \in \overline{\mathscr{T}(y)}$ and $h(T)\cdot x \in \overline{\mathscr{T}(y)}$ implies that $h(t)\cdot x \in \overline{\mathscr{T}(y)}$ for t in $[0, T]$, because of the geometry of horocycle orbits in $\mathrm{T}^1\mathbb{H}$.

Applying the above claim for each of the sets C_i, we obtain the partition

$$L = \Omega_{cp} \sqcup L_1 \sqcup \cdots \sqcup L_\ell.$$

If $x \in K \subseteq \Omega_{cp}$ and $T > 0$, then the interval $[0, T]$ is naturally decomposed into subintervals $I = [T_0, T_1]$ of times where

$$\begin{aligned} &h(t)\cdot x \in \Omega_{cp} \text{ for all } t \in I^o, \text{ or} \\ &h(t)\cdot x \in C_1 \text{ for all } t \in I^o, \text{ or} \\ &\qquad \vdots \\ &h(t)\cdot x \in C_\ell \text{ for all } t \in I^o. \end{aligned}$$

For a subinterval of the first type there is nothing to show, while for an interval of the other types we can apply the claim in equation (11.25) to the initial point in $\mathscr{T}_i$ corresponding to $xu^-(-T_0) \in \partial C_i$ and time $T_1 - T_0$. Adding these estimates together gives the result.

To see that we already know equation (11.25), notice that $\mathscr{T}(y)$ maps injectively to $X_2 = \mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$ under the natural quotient map π. Apply Theorem 11.32 to the image K of $\partial\mathscr{T}(y)$ in X_2, to obtain a compact subset in X_2. Let $L \subseteq \mathscr{T}(y)$ be its pre-image under the quotient map, which has compact closure in $\mathscr{T}$.

Choose $x \in \partial\mathscr{T}(y)$ and $T > 0$ such that $h(T)\cdot x \in \overline{\mathscr{T}(y)}$, then $h(t)\cdot x \in L$ if and only if $h(t)\cdot\pi(x) \in \pi(L)$ for any $t \in [0, T]$, which gives the claim.

The last statement in the theorem also follows from the special case of $X_2 = \mathrm{SL}_2(\mathbb{Z}) \backslash \mathrm{SL}_2(\mathbb{R})$ in the same way. □

Exercises for Section 11.6

Exercise 11.6.1. As discussed above, $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ can be identified both with the unit tangent bundle of the modular surface (and so with vectors attached to points in the fundamental domain F shown in Figure 9.5 on p. 303), and with the space of unimodular lattices in $\mathbb{R}^2$. Make this correspondence explicit, and in particular explain the meaning of the imaginary coordinate in terms of properties of the associated lattice. Use this to deduce an independent proof of Theorem 11.33 in the case $d = 2$.

Exercise 11.6.2. In this exercise we show how special algebraic groups defined over $\mathbb{Q}$ give rise to closed orbits in $\mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$. Let

$$\rho : \mathrm{SL}_n(\mathbb{R}) \hookrightarrow \mathrm{SL}_N(\mathbb{R})$$

be a linear representation such that for $1 \leqslant i, j \leqslant N$ the (i, j)th matrix entry of $\rho(g)$ is a polynomial in the matrix entries of g with rational coefficients (that is, ρ is an algebraic representation defined over $\mathbb{Q}$). Let $v \in \mathbb{Q}^N$ be a vector, and define*

$$\mathbb{G} = \{g \in \mathrm{SL}_n \mid \rho(g)v = v\}$$

to be the stabilizer of v. Prove that the orbit $\mathrm{SL}_n(\mathbb{Z})\mathbb{G}(\mathbb{R})$ in $\mathrm{SL}_n(\mathbb{Z})\backslash \mathrm{SL}_n(\mathbb{R})$ under the group $\mathbb{G}(\mathbb{R})$ is closed.

Exercise 11.6.3. In this exercise we show how special quaternion division algebras over $\mathbb{Q}$ give rise to uniform (and, by definition, also) arithmetic lattices in $\mathrm{SL}_2(\mathbb{R})$. A quaternion division algebra over $\mathbb{Q}$ is an algebra D over the field $\mathbb{Q}$ such that D has dimension four over $\mathbb{Q}$, D has a unit 1_D, $\mathbb{Q}$ (identified with $\mathbb{Q} \cdot 1_D$) is the center of D, and every non-zero element of D has a multiplicative inverse. The best-known example is the algebra of rational Hamiltonian quaternions defined by $\mathbb{Q} + \mathbb{Q}\mathrm{i} + \mathbb{Q}\mathrm{j} + \mathbb{Q}\mathrm{k}$ with $\mathrm{i}^2 = \mathrm{j}^2 = \mathrm{k}^2 = -1$ and $\mathrm{ij} = -\mathrm{ji} = \mathrm{k} \ldots$: We will not be able to use this particular algebra to construct a lattice in $\mathrm{SL}_2(\mathbb{R})$, so we begin by constructing a different quaternion algebra, and then show how it gives rise to a lattice.
(a) Show that for $a, b, c, d \in \mathbb{Z}$ the sum $a^2 + b^2 + c^2 + d^2$ is not divisible by 8 unless a, b, c, d are all even.
(b) Use (a) to show that

$$D = \left\{ \begin{pmatrix} a + \sqrt{7}b & c + \sqrt{7}d \\ -(c - \sqrt{7}d) & a - \sqrt{7}b \end{pmatrix} \mid a, b, c, d \in \mathbb{Q} \right\}$$

is a quaternion division algebra over $\mathbb{Q}$.
(c) Show that $D \otimes_{\mathbb{R}} \mathbb{R} = \mathrm{Mat}_{22}(\mathbb{R})$ (that is, show that D is $\mathbb{R}$-split).

* Here we intentionally omit the field or ring after $\mathbb{G}$ and SL_n, but for any of the choices $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ we obtain definitions of $\mathbb{G}(\mathbb{Z}), \mathbb{G}(\mathbb{Q}), \mathbb{G}(\mathbb{R})$ by requiring that the elements belong to $\mathrm{SL}_n(\mathbb{Z}), \mathrm{SL}_n(\mathbb{Q}), \mathrm{SL}_n(\mathbb{R})$ respectively.

(d) Embed D into $\mathrm{Mat}_{44}(\mathbb{Q})$ by using the basis

$$1_D = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}, \mathrm{i}_D = \begin{pmatrix} \sqrt{7} & \\ & -\sqrt{7} \end{pmatrix}, \mathrm{j}_D = \begin{pmatrix} & 1 \\ -1 & \end{pmatrix}, \mathrm{k}_D = \begin{pmatrix} & -\sqrt{7} \\ \sqrt{7} & \end{pmatrix}$$

of D, and identifying $d \in D$ with the linear map $D \ni f \mapsto df \in D$. The image of D is a four-dimensional linear subspace W of $\mathrm{Mat}_{44}(\mathbb{Q})$ defined by 12 linear equations. On W the original coordinates of a, b, c, d with respect to the chosen basis can be represented by rational linear combinations of the matrix entries of $g \in W$.
(e) Show that

$$\mathbb{G} = \{g \in \mathrm{SL}_4 \mid g \in W, a^2 - 7b^2 + c^2 - 7d^2 = 1\}$$

has $\mathbb{G}(\mathbb{R}) \cong \mathrm{SL}_2(\mathbb{R})$. Here $\mathbb{G}(\mathbb{R})$ is defined to consist of all $g \in W$ (that is, g satisfying the 12 linear equations) for which the pre-image in D satisfies

$$a^2 - 7b^2 + c^2 - 7d^2 = 1.$$

(f) Show that $\mathrm{SL}_4(\mathbb{Z})\mathbb{G}(\mathbb{R})$ is closed by using Exercise 11.6.2.
(g) Show that $\mathrm{SL}_4(\mathbb{Z})\mathbb{G}(\mathbb{R}) \cong \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is compact by using the Mahler compactness criterion (Theorem 11.33) on $\mathrm{SL}_4(\mathbb{Z})\backslash \mathrm{SL}_4(\mathbb{R})$.

11.7 Equidistribution of Horocycle Orbits

We are now ready to prove the theorem promised earlier, due to Dani and Smillie [63]. The argument used below is due to Ratner [306, Lem. 2.1].

Theorem 11.35. *Let $\Gamma \subseteq \mathrm{SL}_2(\mathbb{R})$ be a lattice, and let $x \in X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$. Then either x is periodic for the horocycle flow (that is, $h(t)\cdot x = x$ for some $t > 0$), or the horocycle orbit of x is equidistributed with respect to the Haar measure m_X of X:*

$$\frac{1}{T}\int_0^T f(h(t)\cdot x)\,\mathrm{d}t \longrightarrow \int f\,\mathrm{d}m_X$$

as $T \to \infty$.

If X is compact, then there are no periodic orbits by Lemma 11.28, and the statement of Theorem 11.35 is already known by Theorem 11.27.

Proof. Suppose that the point $x_0 \in X$ is not periodic for the horocycle flow, let $T_n \nearrow \infty$ be any sequence, and define a sequence (μ_n) of probability measures by

$$\int f\,\mathrm{d}\mu_n = \frac{1}{T_n}\int_0^{T_n} f(h(t)\cdot x)\,\mathrm{d}t$$

for $f \in C_c(X)$. By passing to a subsequence, we may assume that $\mu_n \to \mu$ in the weak*-topology, and by Theorem 11.32 the limit μ is a probability measure. To show the theorem we need to show that $\mu = m_X$.

By the version of Theorem 4.1 for flows (which is contained in the proof of Theorem 8.10), we know that μ is invariant under the horocycle flow. By the ergodic decomposition theorem* (Theorem 6.2) extended to flows (which is contained in the more general Theorem 8.20) we can write

$$\mu = \int \mu_y \, \mathrm{d}\nu(y)$$

as a generalized convex combination of probability measures that are invariant and ergodic under the horocycle flow. By Theorem 11.27, each such measure is either the Haar measure m_X or is the Lebesgue measure on a periodic orbit of a point $y = h(t_y)\cdot y$ with $t_y > 0$. Therefore, to show that $\mu = m_X$ we only have to show

$$\mu\left(\{y \in X \mid h(t_y)\cdot y = y \text{ for some } t_y > 0\}\right) = 0. \tag{11.26}$$

Let us point out a few complications at this point. The set of periodic points as in equation (11.26) is dense (in fact long periodic orbits become equidistributed by Exercise 11.5.2). Moreover, to show that $\mu(B) = 0$ for a given measurable set B it is not sufficient to show that $\mu_i(B) \to 0$ as $i \to \infty$†. However, if B is compact then, in order to prove that $\mu(B) = 0$, it is enough to find, for every $\varepsilon > 0$, an open set $O \supseteq B$ with

$$\limsup_{i\to\infty} \mu_i(O) \leqslant \varepsilon.$$

This criterion for vanishing of $\mu(B)$ follows easily from the definition of the weak*-topology and the existence of a continuous function $f \in C_c(X)$ with $\chi_B \leqslant f \leqslant \chi_O$. In order to apply this criterion, we need to write the set of periodic points as a countable union of compact sets, for which we have already developed all the necessary tools. By Lemma 11.29 there are finitely many one-parameter families of periodic orbits (one for each cusp). Fix a periodic point $x \in X$, and restrict the parameter $t \in \mathbb{R}$ from Lemma 11.29 to a compact set $I \subseteq \mathbb{R}$. Allowing $s \in [0, t_x]$ to vary, we obtain a compact set $B = \{xu^-(s)a(-t) \mid s \in [0, t_x], t \in I\}$ comprising periodic orbits. Varying x through a finite list and increasing I, we can write the set of periodic points as in equation (11.26) as a countable union of such compact sets B. Thus it is sufficient to show that $\mu(B) = 0$ for one such set B. Now fix $\varepsilon > 0$ and let $L = L(\varepsilon) \subseteq X$ be a compact set constructed as in the final conclusion of Theorem 11.32. Recall that the period of $xa(-t)$ with respect to the horocycle

* By using Exercise 11.5.3, this dependence on material from Chapter 6 can be avoided.

† For example, if $B = X \smallsetminus x_0U^-$ then $\mu_i(B) = 0$ for all $i \geqslant 1$, but we cannot have $\mu(B) = 0$ as $x_0U^- \cong U^-$ cannot have a U^--invariant probability measure.

flow is $\mathrm{e}^{-2t}t_x$ if $t_x > 0$ is the period of x, and so $xa(-t)$ diverges to one of the cusps as $t \to \infty$ by Lemma 11.29. Since I is compact, there exists some $t_\varepsilon \in \mathbb{R}$ with $Ba(-t_\varepsilon) \subseteq X \smallsetminus L$. We define an open set $O = (X \smallsetminus L)\, a(t_\varepsilon) \supseteq B$ and claim that $\mu_i(O) \leqslant \varepsilon$ for all large enough i. However,

$$\begin{aligned}\mu_i(O) &= \frac{1}{T_i} m_{\mathbb{R}} \left(\{s \in [0, T_i] \mid h(s)\cdot x \in O = (X \smallsetminus L)a(t_\varepsilon)\}\right) \\ &= \frac{1}{T_i} m_{\mathbb{R}} \left(\left\{s \in [0, T_i] \mid xa(-t_\varepsilon)u^-(-\mathrm{e}^{-2t}s) \notin L\right\}\right) \\ &= \frac{1}{\mathrm{e}^{-2t_\varepsilon} T_i} m_{\mathbb{R}} \left(\left\{s \in [0, e^{-2t_\varepsilon} T_i] \mid xa(-t_\varepsilon)u^-(s) \notin L\right\}\right)\end{aligned}$$

is the expression discussed in equation (11.22) for the initial point $xa(-t_\varepsilon)$. Since $\mathrm{e}^{-t_\varepsilon} T_i \to \infty$ as $i \to \infty$, and x (equivalently, $xa(-t_\varepsilon)$) is not periodic by assumption, the final statement in Theorem 11.32 shows that $\mu_i(O) \leqslant \varepsilon$ for large enough i. This shows that $\mu(B) = 0$ and hence equation (11.26), which gives the theorem. □

As discussed in Chapter 1, Raghunathan and Dani formulated far-reaching conjectures. Firstly Raghunathan conjectured that orbit closures under unipotent actions are always orbits of closed subgroups, generalizing the classification of orbits for the horocycle flow (that is, the statement that a horocycle orbit is either dense or periodic, which follows trivially from the stronger equidistribution result in Theorem 11.35). Dani conjectured measure rigidity of unipotent flows other than the horocycle flow (which is handled in Theorem 11.27). Ratner proved these conjectures in full, and in addition proved the generalization of equidistribution for horocycle orbits (Theorem 11.35) to other unipotent flows, leading to numerous applications especially in number theory.

Notes to Chapter 11

[93] (Page 341) Some of the striking ergodic and dynamical properties satisfied by the horocycle flow are the following. It is a highly non-trivial example of a flow that is mixing of all orders, as shown by Marcus [244]; Ratner has classified joinings between horocycle flows [301], measurable factors of horocycle flows [300], and many other measurable rigidity properties (see Ratner's survey article [302]); Marcus has also shown a form of rigidity for topological conjugacy of horocycle flows [245].

[94] (Page 342) The material in this section follows the book of Bekka and Mayer [21, Chap. 2] closely.

[95] (Page 355) This *Klein model* was used by both Klein [198] and Poincaré; the upper half-plane model is usually called the *Poincaré model* because of the influential paper [288] exploring its properties. Both models were used earlier by Beltrami [23] in order to show that hyperbolic geometry is as consistent as Euclidean geometry.

[96] (Page 355) The uniformization theorem generalizes the Riemann mapping theorem, which states that if U is a simply connected open proper subset of $\mathbb{C}$, then

there exists a bijective holomorphic map from U onto the open unit disk. This was stated in Riemann's thesis [310], and the first complete proof was given by Carathéodory [48]. The uniformization theorem was finally proved by Koebe [207] and by Poincaré [290]; a convenient account may be found in the monograph of Farkas and Kra [89, Chap. IV] or Stillwell [355, Chap. 5].

[97] (Page 357) The work of Mautner clarified earlier results of Gel′fand and Fomin [112]. A useful overview of its use in ergodic theory may be found in Starkov's monograph [352].

[98] (Page 366) A different approach comes from work of Ryzhikov, using joinings. For the case of horocycle flows, Ryzhikov [328] showed mixing of all orders as a consequence of a new criterion for mixing of all orders, giving a new proof of the theorem of Marcus [244] (this is also described in the article of Thouvenot [361]).

[99] (Page 371) For a compact quotient, Hedlund showed that the horocycle flow is minimal [145] (every point has a dense orbit), and this was later generalized by Veech to nilpotent flows on semi-simple Lie groups [369].

[100] (Page 377) There are more general results concerning decompositions of Haar measure in locally compact groups – see Knapp [204, Sect. 8.3].

[101] (Page 381) The original work on non-divergence is due to Margulis [246]; subsequent refinements include work of Dani [62] and Kleinbock and Margulis [199] on quantitative statements; Kleinbock and Tomanov [200] on extensions to the S-arithmetic setting; Ghosh [114] on the positive characteristic case.