Chapter 5
Horospherical Subgroups and Counting Results

The inheritance property of ergodicity of the Mautner phenomenon in Proposition 2.25 and in the general Theorem 2.55 (also see Exercise 2.24 and 2.45) established in Chapter 2 already gives the equidistribution of many orbits.

Indeed, if a simple Lie group G acts ergodically on (X, μ) and

$$\{g_t \mid t \in \mathbb{R}\} \subseteq G$$

is an unbounded one-parameter subgroup, then

$$\frac{1}{T}\int_0^T f(g_t \cdot x)\,\mathrm{d}t \longrightarrow \int_X f\,\mathrm{d}\mu$$

for μ-almost every $x \in X$, for any $f \in C_c(X)$ as $T \to \infty$. A point $x \in X$ with this property is called *generic* for μ and the one-parameter subgroup $\{g_t \mid t \in \mathbb{R}\}$.

In this chapter we start the discussion of unipotent dynamics by considering the case of horospherical actions. For those actions we will show 'unique ergodicity', and sometimes 'almost unique ergodicity', and we will understand precisely which points are generic for m_X. The method of proof also gives other equidistribution results of certain 'distorted orbits', which in turn can be used to prove asymptotic counting results. Hence in the second half of the chapter we will explain the set-up of Duke, Rudnick, and Sarnak and its dynamical interpretation by Eskin and McMullen.

5.1 Dynamics on Hyperbolic Surfaces

Let us start by discussing briefly the case of the geodesic flow and the horocycle flow on quotients of $\mathrm{SL}_2(\mathbb{R})$ as introduced in Section 1.2.

5.1.1 The Geodesic Flow

We note first that for the geodesic flow defined by the diagonal subgroup it is not possible to make a more general statement about the equidistribution of orbits by relaxing the requirement that the point be μ-typical. Indeed, in this case the flow is partially hyperbolic and as a result X contains many irregular orbits. As this result can be considered of negative type we will not prove it here, but refer to [45, Sec. 9.7.2] for a more detailed discussion of the case of the geodesic flow on the modular surface.

Example 5.1. For a compact quotient X of $\mathrm{SL}_2(\mathbb{R})$ by a uniform lattice as in Figure 5.1, the action of the one-parameter subgroup

$$A = \left\{ a_t = \begin{pmatrix} \mathrm{e}^{-t/2} & \\ & \mathrm{e}^{t/2} \end{pmatrix} \;\middle|\; t \in \mathbb{R} \right\} \tag{5.1}$$

has many orbits that, for example, stay on one side of the dotted line.[(27)]

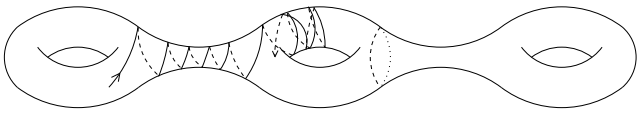


Fig. 5.1: There are many orbits under the action of A that stay on one side of the dotted line furthest to the right.

We also refer to Exercises 5.2–5.5 for the behaviour of the geodesic flow and higher dimensional analogues.

Exercise 5.2 (Anosov shadowing for $\mathrm{SL}_2(\mathbb{R})$). Let X be the quotient of $\mathrm{SL}_2(\mathbb{R})$ by a discrete subgroup $\Gamma < \mathrm{SL}_2(\mathbb{R})$.
(a) Let $x \in X$, $T > 0$, $\varepsilon > 0$ and $y \in X$ be chosen with $\mathsf{d}(a_T{\cdot}x, y) < \varepsilon$. Then there exists a point $z \in X$ with $\mathsf{d}(x, z) \ll \mathrm{e}^{-T}\varepsilon$ (and so $\mathsf{d}(a_t{\cdot}x, a_t{\cdot}z) \ll \varepsilon$ for $t \in [0, T]$) and $\mathsf{d}(a_t{\cdot}y, a_{T+t}{\cdot}z) \ll \varepsilon$ for all $t \geqslant 0$. Also show that there exists some δ with $|\delta| \ll \varepsilon$ such that $\mathsf{d}(a_{t+\delta}{\cdot}y, a_{T+t}{\cdot}z) \ll \mathrm{e}^{-t}$ for all $t \geqslant 0$.
(b) Assume now that X is compact (for example, as in Figure 5.1) and use (a) to construct non-periodic orbits as in Example 5.1.

Exercise 5.3 (Anosov closing for $\mathrm{SL}_2(\mathbb{R})$). Let X be as in Exercise 5.2. Let x in X and $T \geqslant 1$ be chosen so that $\mathsf{d}(a_T{\cdot}x, x) \leqslant \varepsilon < 1$. Show that there exists a point $z \in X$ which is periodic with period T_z satisfying

$$|T_z - T| \ll \varepsilon$$

and

$$\mathsf{d}(a_t{\cdot}x, a_t{\cdot}z) \ll \varepsilon$$

for all $t \in [0, T]$.

Exercise 5.4 (Anosov shadowing for G). Let G be a connected Lie group, let $\Gamma < G$ be a discrete subgroup, let $X = G/\Gamma$, and let $a \in G$ be such that Ad_a is diagonalizable with positive eigenvalues.
(a) Let $x \in X$, $N > 1$, $\varepsilon > 0$ and $y \in X$ be such that $\mathsf{d}(a^N \bullet x, y) < \varepsilon$. Then there exists a point $z \in X$, some $\lambda < 1$ (independent of x, y and Γ) with

$$\mathsf{d}(a^n \bullet x, a^n \bullet z) \ll \lambda^{N-n} \varepsilon$$

for $n = 0, \dots, N$ and

$$\mathsf{d}(a^{N+n} \bullet z, a^n \bullet y) \ll \varepsilon$$

for all $n \geqslant 0$.
(b) Assume that X has finite volume and a acts mixingly on X with respect to m_X. Construct non-periodic irregular orbits by iterating (a).

Exercise 5.5 (Anosov closing for $X = \mathrm{SL}_d(\mathbb{R})/\Gamma$). We let X be any quotient of the group $G = \mathrm{SL}_d(\mathbb{R})$ by a discrete subgroup $\Gamma < G$, and let A be the subgroup of G of positive diagonal matrices. Let $a \in A$ be a nontrivial element.
(a) Suppose that $x \in X$ and $N \geqslant 1$ are such that $\mathsf{d}(a^N, I) \geqslant 1$ but $\mathsf{d}(a^N \bullet x, x) \leqslant \varepsilon < 1$. Assume that ε is sufficiently small and that N is sufficiently large. Show that there exists some $z \in X$ and some $c \in \mathrm{SL}_d(\mathbb{R})$ with $ac = ca$, $\mathsf{d}(a^N, c) \ll \varepsilon$, $c \bullet z = z$ and

$$\mathsf{d}(a^n \bullet x, a^n \bullet z) \ll \varepsilon$$

for $n = 0, \dots, N$.
(b) Suppose that a is regular (that is, no two eigenvalues are the same) and X is compact. Show that z as in (a) is a periodic point for A.
(c) Suppose $d = 3$ and a is regular and does not have 1 as an eigenvalue, and

$$X = \mathsf{X}_3 = \mathrm{SL}_3(\mathbb{R})/\mathrm{SL}_3(\mathbb{Z}).$$

Show again that the point z as in (a) is periodic for A.
(d) Repeat (c) for

$$X = \mathsf{X}_d = \mathrm{SL}_d(\mathbb{R})/\mathrm{SL}_d(\mathbb{Z}),$$

assuming that $a \in A$ has the property that no product over a proper non-empty subset of the eigenvalues of a equals 1.
(e) In the setting of (b), (c), and of (d), show that periodic A-orbits are dense in X.
(f) Generalize the statement in (b) to any semisimple group.[†]

5.1.2 The Horocycle Flow

The discussion above for the geodesic flow is in stark contrast to the behaviour of horocycle orbits defined by the unipotent subgroup

$$U = G^+_{a_1} = \left\{ u_s = \begin{pmatrix} 1 & \\ s & 1 \end{pmatrix} \;\middle|\; s \in \mathbb{R} \right\}$$

[†] In that sense Poincaré recurrence can be used to construct anisotropic tori (see Section 9.3).

in the compact quotient X: The orbit of every point under this group action visits the right-hand side in Figure 5.1 at some point (indeed much more is true).

In fact Hedlund [69] showed in 1936 that the horocycle flow on any compact quotient of $\mathrm{SL}_2(\mathbb{R})$ is *minimal* (that is, has no nontrivial closed invariant subsets) and that Haar measure is ergodic. This was strengthened by Furstenberg [57] in 1972 and by Dani [17] in 1978, who showed the following theorems.

Theorem 5.6 (Unique ergodicity of horocycle flow). *If Γ is a uniform lattice in $\mathrm{SL}_2(\mathbb{R})$, then the horocycle flow (that is, the action of the subgroup U) is uniquely ergodic on the quotient X of $\mathrm{SL}_2(\mathbb{R})$ by Γ.*

Theorem 5.7 (Almost unique ergodicity of horocycle flow). *If $X = \mathsf{X}_2$ is the quotient of $\mathrm{SL}_2(\mathbb{R})$ defined by $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ (or another non-uniform lattice) then a probability measure m on X that is invariant and ergodic for the action of U is either*

- *the Haar measure m_X on X (inherited from the Haar measure $m_{\mathrm{SL}_2(\mathbb{R})}$) or*
- *a one-dimensional Lebesgue measure supported on a periodic orbit of the action for U.*

Moreover, both types of invariant measure indeed exist.

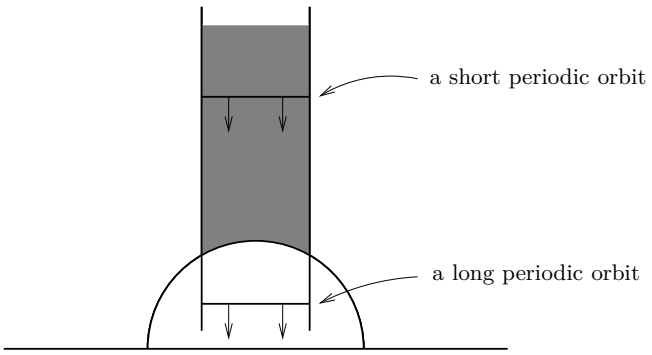


Fig. 5.2: In the standard fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$, the observed speed of a periodic horocycle orbit increases with the height, so the two different periodic orbits shown are of different lengths. The longer periodic orbit could also be drawn in the fundamental domain, but it would look very complicated.

Moreover, the tool discussed in the next section also gives the following theorem[†] of Sarnak [139] as well as Theorem 1.16 concerning expanding circles.

[†] Sarnak also gives an error rate in this equidistribution result—obtaining this (or even any) error estimate requires more sophisticated methods than those we will discuss here.

Theorem 5.8 (Equidistribution of long periodic horocycles). *Let X be a quotient of $\mathrm{SL}_2(\mathbb{R})$ by a non-uniform lattice and let A be the diagonal subgroup as in (5.1). Let $x \in X$ be a periodic orbit for the horocycle flow $U = G^+_{a_1}$ and let μ be the normalized Lebesgue measure on the one-dimensional orbit $U\bullet x$. Then the periodic orbit measures $(a_t)_*\mu$*

- *diverge for $t \to -\infty$ to infinity (in which case the periodic orbit $a_tU\bullet x$ becomes shorter and shorter) and*
- *equidistribute for $t \to \infty$ with respect to the Haar measure m_X (in which case the periodic orbit $a_tU\bullet x$ become longer and longer).*

We will prove Theorem 5.6 in Section 5.2.1 and Theorems 5.7 and 5.8 in Section 5.3.1.

5.2 The Banana Mixing Trick and Unique Ergodicity

We suppose in the following that G is a closed linear group and that the element $a \in G \leqslant \mathrm{SL}_d(\mathbb{R})$ only has real and positive eigenvalues. Let

$$G^+_a = \left\{ g \in G \;\middle|\; a^n g a^{-n} \longrightarrow I \text{ as } n \longrightarrow -\infty \right\}$$

be the unstable horospherical subgroup of a. The general method discussed below gives a way to classify the G^+_a-invariant ergodic probability measures on X. The method goes back to the PhD thesis of Margulis, who refers to this as the banana argument due to the shape of the sets involved.

Theorem 5.9 (Banana mixing argument for G^+_a). *Let $X = G\bullet x_0 \subseteq \mathsf{X}_d$ be a finite volume orbit for a closed connected subgroup $G \leqslant \mathrm{SL}_d(\mathbb{R})$. Let $a \in G$ only have real and positive eigenvalues, and suppose that a acts as a mixing transformation on X with respect to m_X. Let G^+_a be the unstable horospherical subgroup for a, and let B_0 be a neighbourhood of $I \in G^+_a$ with compact closure and a boundary of zero Haar measure. Let $f \in C_c(X)$ and $\varepsilon > 0$. Finally suppose that $K \subseteq X$ is a compact set such that $B_0 \ni u \mapsto u\bullet x$ is injective for any $x \in K$. Then there exists an integer N such that*

$$\left| \frac{1}{m_{G^+_a}(a^n B_0 a^{-n})} \int\limits_{a^n B_0 a^{-n}} f(u\bullet x)\, \mathrm{d}m_{G^+_a}(u) - \frac{1}{m_X(X)} \int\limits_X f \,\mathrm{d}m_X \right| < \varepsilon$$

for all $n \geqslant N$ whenever $a^{-n}\bullet x \in K$.

We will prove the theorem in Section 5.2.2.

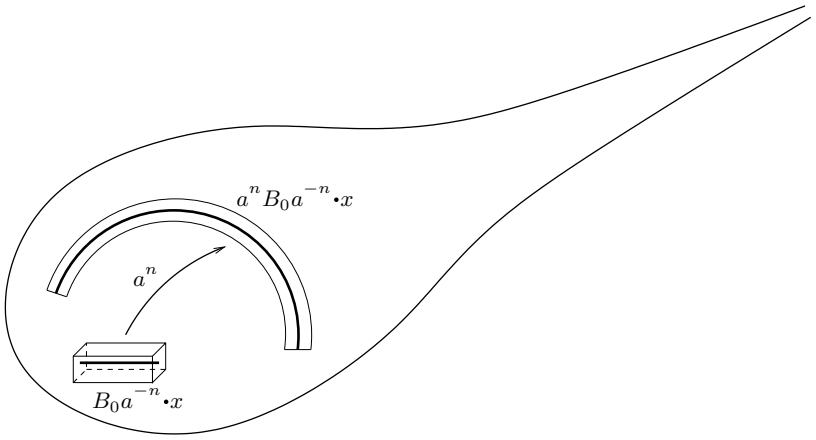


Fig. 5.3: A 'box' containing $B_0 a^{-n} \cdot x$ is mapped by a^n to a 'banana' that contains the much bigger set $a^n B_0 a^{-n} \cdot x$ in the direction of G_a^+, is about as thick as the original box in the direction of $C_G(a)$, but is much thinner in the direction of G_a^- (not drawn).

5.2.1 Unique Ergodicity on Compact Quotients

The following consequence of Theorem 5.9 is a generalization of Theorem 5.6.

Theorem 5.10 (Unique ergodicity of horospherical actions[28]). *Let G be a linear Lie group, $\Gamma < G$ be a uniform lattice, and let $a \in G$ have only real and positive eigenvalues. Suppose a acts mixingly on $X = G/\Gamma$. Then the action of G_a^+ on X is* uniquely ergodic*: m_X is the only G_a^+-invariant probability measure on X and every point $x \in X$ is generic for G_a^+ and m_X.*

Proof of Theorems 5.6 and 5.10. We note that compactness of X implies that $G_a^+ \ni u \mapsto u \cdot x \in X$ is injective for any $x \in X$. Indeed, if $u \cdot x = x$ for some $u \in G_a^+ \smallsetminus \{I\}$ and $x \in X$ then the injectivity radius at

$$a^{-n} \cdot x = (a^{-n} u a^n) a^{-n} \cdot x$$

would go to 0 for $n \to \infty$ and as a result contradict Lemma 1.17. Let $B_0 \subseteq G_a^+$ be as in Theorem 5.9,[†] set $K = X$ and $B_n = a^n B_0 a^{-n}$ for $n \in \mathbb{N}$. By Theorem 5.9 we have

$$\frac{1}{m_{G_a^+}(B_n)} \int_{B_n} f(u \cdot x) \, \mathrm{d}m_{G_a^+}(u) \longrightarrow \int_X f \, \mathrm{d}m_X$$

as $n \to \infty$ for any $f \in C(X)$ and any $x \in X$ (as the constraint $a^{-n} \cdot x \in K$ is meaningless for $K = X$).

[†] For example, $B_0 = B_r^{G_a^+}$ for $r > 0$ with $m_{G_a^+}\left(\left\{u \in G_a^+ \;\middle|\; \mathtt{d}_G(u, I) = r\right\}\right) = 0$

Now let μ be a G_a^+-invariant probability measure. Then

$$\int_X f \,\mathrm{d}\mu = \int_X \frac{1}{m_{G_a^+}(B_n)} \int_{B_n} f(u{\scriptstyle\bullet}x) \,\mathrm{d}m_{G_a^+}(u) \,\mathrm{d}\mu(x) \longrightarrow \int_X f \,\mathrm{d}m_X$$

as $n \to \infty$ by Fubini's theorem and dominated convergence. As this holds for any $f \in C(X)$ we deduce that $\mu = m_X$, as claimed. □

Notice that once unique ergodicity is proved then the pointwise everywhere convergence of the ergodic averages also follows for other Følner sets (see Exercises 5.11–5.12).

Exercise 5.11. Let $B_n = a^n B_0 a^{-n}$ be as in the proof of Theorem 5.10 with $m_{G_a^+}(\partial B_0) = 0$ for $n \geqslant 1$. Show that (B_n) is a Følner sequence in G_a^+, that is a sequence satisfying

$$\frac{m_{G_a^+}(B_n \triangle (KB_n))}{m_{G_a^+}(B_n)} \longrightarrow 0 \tag{5.2}$$

as $n \to \infty$ for every compact subset $K \subseteq G_a^+$.

Exercise 5.12. Let a and X be as in Theorem 5.10. Let (F_n) be any Følner sequence in G_a^+ satisfying (5.2) and show that

$$\frac{1}{m_{G_a^+}(F_n)} \int_{F_n} f(u{\scriptstyle\bullet}x) \,\mathrm{d}m_{G_a^+}(u) \longrightarrow \int_X f \,\mathrm{d}m_X$$

as $n \to \infty$, for any $f \in C(X)$ and any $x \in X$.

5.2.2 Proving the Banana Trick

For the proof of Theorem 5.9 we will use the 'stable parabolic subgroup'

$$P_a^- = \left\{ g \in G \;\middle|\; a^n g a^{-n} \text{ stays bounded as } n \to \infty \right\}$$

together with the unstable horospherical subgroup G_a^+.

Lemma 5.13 (A coordinate system). *Let $G \leqslant \mathrm{SL}_d(\mathbb{R})$ be a closed linear group and let $a \in G$ only have real and positive eigenvalues. Then P_a^- and G_a^+ are closed subgroups that together define a coordinate system in the following sense. The set $P_a^- G_a^+$ is open in G, the map $P_a^- \times G_a^+ \ni (h, u) \mapsto hu \in P_a^- G_a^+$ is a homeomorphism, and the Haar measure m_G restricted to $P_a^- G_a^+$ is proportional to the push-forward of the product of the Haar measures of P_a^- and G_a^+.*

Proof. By conjugating a, G, and its subgroups P_a^- and G_a^+ we may assume that $a = \mathrm{diag}(a_1, \ldots, a_d)$ is diagonal. With this P_a^- can be defined as a subgroup of the set of all $g = (g_{i,j}) \in G$ with $g_{i,j} = 0$ for all indices i, j with $a_i > a_j$. This shows that P_a^- is closed subgroup. Similarly G_a^+ is also a closed subgroup.

Moreover, from the definition it follows that the Lie algebra $\mathfrak{p}_a^-$ of P_a^- (respectively $\mathfrak{g}_a^+$ of G_a^+) is the direct sum of all eigenspaces in $\mathfrak{g}$ of Ad_a with eigenvalue less than or equal to 1 (respectively bigger than 1). This shows that the derivative of the map $\phi\colon P_a^- \times G_a^+ \ni (h,u) \mapsto hu \in P_a^- G_a^+$ at (I,I) is the linear isomorphism $\mathfrak{p}_a^- \times \mathfrak{g}_a^+ \ni (x,y) \mapsto x+y \in \mathfrak{g}$. By the inverse mapping theorem ϕ is locally a diffeomorphism.

To see that ϕ is injective let $(h_1,u_1),(h_2,u_2) \in P_a^- \times G_a^+$ satisfy $h_1u_1 = h_2u_2$. Then $g = h_2^{-1}h_1 = u_2u_1^{-1} \in P_a^- \cap G_a^+$ has the property that $a^n g a^{-n}$ remains bounded for $n \geqslant 0$ and converges to I as $n \to -\infty$. The latter implies that the matrix $g - I$ is zero or a sum of eigenvectors for conjugation by a with eigenvalues bigger than 1. Together with the behaviour for $n \geqslant 0$ this implies that $g = I$ and hence $(h_1,u_1) = (h_2,u_2)$.

Now let $O \subseteq P_a^- \times G_a^+$ be open and $(h,u) \in O$. Then $(h,I)^{-1}O(I,u)^{-1}$ is also open in $P_a^- \times G_a^+$. By the behaviour of ϕ near (I,I) obtained above we deduce that $\phi\big((h,I)^{-1}O(I,u)^{-1}\big) = h^{-1}\phi(O)u^{-1}$ is a neighbourhood of I, which shows that $\phi(O)$ is a neighbourhood of $hu = \phi\big((h,u)\big)$. It follows that $\phi(O) \subseteq G$ is open for any open subset $O \subseteq P_a^- \times G_a^+$, that $P_a^- G_a^+ \subseteq G$ is open, and that $\phi\colon P_a^- \times G_a^+ \to P_a^- G_a^+$ is a homeomorphism.

With this we may apply Lemma 1.58 and obtain that m_G restricted to $P_a^- G_a^+$ is proportional to the push-forward of the product of the (left) Haar measure on P_a^- and the (right) Haar measure on G_a^+. As G_a^+ is unipotent it is also unimodular and the lemma follows. □

The following upgrade (a fairly standard compactness argument) to the injectivity assumption for B_0 and K in Theorem 5.9 will be useful in the proof.

Lemma 5.14 (Upgrade to injectivity). *Let $K \subseteq X$ and $B_0 \subseteq G_a^+$ be compact sets for which $B_0 \ni u \mapsto u{\bullet}x$ is injective for all $x \in K$. Then there exists some $\delta = \delta(K,B_0) > 0$ such that*

$$B_\delta^{P_a^-} \times B_0 \ni (h,u) \longmapsto hu{\bullet}x$$

is injective for all $x \in K$.

Proof. If the conclusion of the lemma does not hold then there exist sequences $h_n \to I$ and $h_n' \to I$ as $n \to \infty$, $(u_n),(u_n')$ in B_0, and (x_n) in K with $(h_n,u_n) \neq (h_n',u_n')$ but $h_nu_n{\bullet}x_n = h_n'u_n'{\bullet}x_n$ for all $n \in \mathbb{N}$. As K and B_0 are assumed to be compact we may assume without loss of generality that $x_n \to x \in K$, $u_n \to u \in B_0$, and $u_n' \to u' \in B_0$ as $n \to \infty$. Together we obtain $u{\bullet}x = u'{\bullet}x$ which gives $u = u'$ by our assumption. Moreover, using

$$(\underbrace{h_nu_nu^{-1}}_{\to I})u{\bullet}x_n = h_nu_n{\bullet}x_n = h_n'u_n'{\bullet}x_n = \underbrace{h_n'u_n'u^{-1}}_{\to I}(u{\bullet}x_n)$$

as $n \to \infty$ and the injectivity radius at $u{\bullet}x_n \in B_0{\bullet}K$ we obtain

$$h_nu_nu = h_n'u_n'u^{-1}$$

for all sufficiently large n. Together with the properties of the local coordinate system $P_a^- G_a^+$ in Lemma 5.13 we deduce that $(h_n, u_n) = (h'_n, u'_n)$ for all sufficiently large n. This contradicts our choice of the sequences and proves the lemma. □

Proof of Theorem 5.9. Let us assume compatibility of the Haar measures in the sense that $m_X(\pi(B)) = m_G(B)$ for any injective Borel subset $B \subseteq G$ and that m_G restricted to $P_a^- G_a^+$ is equal to the product of the Haar measures $m_{P_a^-}$ and $m_{G_a^+}$.

We let $B_0 \subseteq G_a^+$ be a neighbourhood of the identity as in the theorem and define $B_n = a^{-n} B_0 a^n$ for $n \geqslant 1$. We suppose for now in addition that B_0 is compact and let $\delta(K, B_0)$ be as in Lemma 5.14.

Using continuity. Now fix a function $f \in C_c(X) \smallsetminus \{0\}$. By compactness of the support, f is uniformly continuous. So for $\varepsilon > 0$ there is a $\delta \in \big(0, \delta(K, B_0)\big)$ for which

$$\mathsf{d}_G(g, I) < \delta \implies |f(g{\cdot}y) - f(y)| < \varepsilon \tag{5.3}$$

for all $g \in G$ and $y \in X$, where d_G is a right-invariant metric on G (giving rise to the metric d on X). We choose a compact neighbourhood $V \subseteq B_\delta^{P_a^-}$ of the identity whose boundary has measure zero with

$$\mathsf{d}_G(a^n h a^{-n}, I) < \delta$$

for $h \in V$ and $n \geqslant 0$.

The banana trick. We now come to the heart of the argument involving the 'box' $VB_0a^{-n}{\cdot}x$ and the 'banana' $a^nVB_0a^{-n}{\cdot}x$ illustrated in Figure 5.3. Indeed

$$\frac{1}{m_{G_a^+}(B_n)} \int\limits_{B_n} f(u{\cdot}x)\, \mathrm{d}m_{G_a^+}(u)$$

is within ε of

$$\frac{1}{m_{P_a^-}(a^nVa^{-n}) m_{G_a^+}(B_n)} \int\limits_{a^nVa^{-n}B_n} \int f(gu{\cdot}x)\, \mathrm{d}m_{P_a^-}(g)\, \mathrm{d}m_{G_a^+}(u)$$

because of (5.3) applied for $y = u{\cdot}x$ and $g = a^n h a^{-n}$ with $h \in V$. Using the definition $B_n = a^n B_0 a^{-n} \subseteq G_a^+$ and Lemma 5.13, the latter may in turn be written as

$$\frac{1}{m_G(VB_0)} \int\limits_{VB_0} f(a^n g a^{-n}{\cdot}x)\, \mathrm{d}m_G(g), \tag{5.4}$$

since m_G is bi-invariant and on $P_a^- G_a^+$ the product of $m_{P_a^-}$ and $m_{G_a^-}$. Moreover, using the injectivity in Lemma 5.14 at $a^{-n}{\cdot}x \in K$ and the notation $y = ga^{-n}{\cdot}x$ with $g \in VB_0$ we see that (5.4) can also be written as

$$\frac{1}{m_G(VB_0)}\int_X f(a^n{\cdot}y)\mathbb{1}_{VB_0a^{-n}{\cdot}x}(y)\,\mathrm{d}m_X(y). \tag{5.5}$$

Using mixing. The expression in (5.5) is an inner product of f composed with a^n and a normalized characteristic function. Hence we would like to apply mixing of a to conclude that (5.5) is ε-close to $\int f\,\mathrm{d}m_X$ if n is large enough. However the characteristic function also depends on n, which in general would be an issue. Fortunately in our case $z = a^{-n}{\cdot}x \in K$, the map

$$K \ni z \longmapsto F_z = \mathbb{1}_{VB_0{\cdot}z} \in L^2_{m_X}(X)$$

is continuous, and the mixing property for f and these characteristic functions holds uniformly on the compact image (see below). So we indeed obtain for n large enough with $a^{-n}{\cdot}K$ that (5.5) is within $\mathrm{O}(\varepsilon)$ of $\int f\,\mathrm{d}m_X$.

Compactness in $L^2(X)$. Let $\eta > 0$ and recall that $m_G(\partial(VB_0)) = 0$. Using the fact that $(VB_0)^o$ is σ-compact there exists a compact subset $C \subseteq (VB_0)^o$ with $m_G(VB_0 \smallsetminus C) < \eta$. Now let $z' = g{\cdot}z, z \in K$ for g sufficiently close to I so that $Cg \subseteq VB_0$. With this we obtain

$$\begin{aligned}\|F_{z'} - F_z\|^2 &= \|F_{z'}\|^2 - 2\Re\langle F_z, F_{z'}\rangle + \|F_z\|^2\\ &= m_X(VB_0{\cdot}z') - 2m_X\big(VB_0g{\cdot}z \cap VB_0{\cdot}z\big) + m_X(VB_0{\cdot}z)\\ &\leqslant \big(2m_G(VB_0) - 2m_G(C)\big) < 2\eta.\end{aligned}$$

As $\eta > 0$ was arbitrary, this shows the continuity of $K \ni z \mapsto F_z \in L^2_{m_X}(X)$ claimed earlier. In particular, $\mathscr{F} = \{F_z \mid z \in K\} \subseteq L^2_{m_X}(X)$ is compact.

Uniform mixing. To prove the uniform mixing we use compactness of $\mathscr{F}$ and find a finite collection $z_1, \dots, z_J \in K$ so that for every $z \in K$ there exists some $j \in \{1, \dots, J\}$ with

$$\|F_z - F_{z_j}\| < \frac{\varepsilon m_G(VB_0)}{\|f\|_2}. \tag{5.6}$$

Applying mixing to f and F_{z_j} for $j = 1, \dots, J$ we may find N so that for $n \geqslant N$ we have

$$\left|\frac{1}{m_G(VB_0)}\langle f \circ a^n, F_{z_j}\rangle - \int f\,\mathrm{d}m_X\right| < \varepsilon. \tag{5.7}$$

However, this now implies for $z \in K$ and $j \in \{1, \dots, J\}$ satisfying (5.6) that

$$\begin{aligned}\left|\frac{1}{m_G(VB_0)}\langle f\circ a^n, F_z\rangle - \int f\,\mathrm{d}m_X\right| &\leqslant \left|\langle f\circ a^n, \frac{1}{m_G(VB_0)}(F_z - F_{z_j})\rangle\right|\\ &\quad + \left|\langle f\circ a^n, \frac{1}{m_G(VB_0)}F_{z_j}\rangle - \int f\,\mathrm{d}m_X\right|\\ &< 2\varepsilon\end{aligned}$$

by Cauchy–Schwarz, (5.6), and (5.7). This concludes the proof for compact sets $B_0 \subseteq G_a^+$.

FINDING A COMPACT B_0. If $B_0 \subseteq G_a^+$ is not compact, then we claim that there exists for any $\eta \in (0,1)$ a compact set $B_0' \subseteq B_0$ with $m_{G_a^+}(B_0 \smallsetminus B_o') < \eta m_{G_a^+}(B_0)$ and $m_{G_a^+}(\partial B_0') = 0$. The difference between the averages obtained using the sets $B_n = a^n B_0 a^{-n}$ and $B_n' = a^n B_0' a^-$ is then easily estimated. Indeed

$$\left| \frac{1}{m_{G_a^+}(B_n)} \int\limits_{B_n} f(u\cdot x)\,\mathrm{d}m_{G_a^+}(u) - \frac{1}{m_{G_a^+}(B_n)} \int\limits_{B_n'} f(u\cdot x)\,\mathrm{d}m_{G_a^+}(u) \right| \leqslant \frac{m_{G_a^+}(B_n \smallsetminus B_n')}{m_{G_a^+}(B_n)} \|f\|_\infty = \frac{m_{G_a^+}(B_0 \smallsetminus B_0')}{m_{G_a^+}(B_0)} \|f\|_\infty < \eta \|f\|_\infty$$

and

$$\begin{aligned} &\left| \frac{1}{m_{G_a^+}(B_n)} \int\limits_{B_n'} f(u\cdot x)\,\mathrm{d}m_{G_a^+}(u) - \frac{1}{m_{G_a^+}(B_n')} \int\limits_{B_n'} f(u\cdot x)\,\mathrm{d}m_{G_a^+}(u) \right| \\ &\qquad \leqslant \left| \frac{1}{m_{G_a^+}(B_n)} - \frac{1}{m_{G_a^+}(B_n')} \right| m_{G_a^+}(B_n') \|f\|_\infty \\ &\qquad = \frac{m_{G_a^+}(B_0 \smallsetminus B_0')}{m_{G_a^+}(B_0)} \|f\|_\infty < \eta \|f\|_\infty \end{aligned}$$

shows that the two averages differ by at most $2\eta\|f\|_\infty$. Using our discussion above for B_0 and setting $\eta = \frac{\varepsilon}{2\|f\|_\infty}$ gives the desired conclusion for B_0.

To prove the claim we recall that $m_{G_a^+}(\partial B_0) = 0$ and so we may assume that B_0 is open. It follows that B_0 is σ-compact and we can find a compact subset $C \subseteq B_0$ with $m_{G_a^+}(B_0 \smallsetminus C) < \eta m_{G_a^+}(B_0)$. It remains to ensure that the boundary is a null set. For this we note that for any $u_0 \in C$ and all but at most countably many radii $r > 0$ the boundary

$$\partial B_r(u_0) \subseteq \{u \in G_a^+ \mid \mathsf{d}(u, u_0) = r\}$$

is a null set. We choose $r(u_0) > 0$ small enough to ensure that in addition $\overline{B^{G_a^+}_{r(u_0)}(u_0)} \subseteq B_0$. The set $B_0' \subseteq B_0$ is then obtained as the closure of the union of a finite cover

$$B^{G_a^+}_{r(u_1)}(u_1) \cup \cdots \cup B^{G_a^+}_{r(u_k)}(u_k) \subseteq B_0$$

of C, completing the proof of the theorem. □

5.3 Almost Unique Ergodicity on Non-Compact Quotients with Finite Volume

We now explain, guided by examples, how the presence of a cusp (that is, the lack of compactness of the quotient) and the presence of horospherical invariant measures other than the Haar measure are related to each other.

5.3.1 Horocycle Action on Non-Compact Quotients

The following result is important for the study of the horocycle flow on quotients of $\mathrm{SL}_2(\mathbb{R})$ and holds much more generally (see also Exercise 1.39). We also allow $\mathrm{SL}_2(\mathbb{C})$ as this makes no difference to the argument.

Proposition 5.15 (Non-uniform lattices and unipotents). *Let $G = \mathrm{SL}_2(\mathbb{R})$ or $G = \mathrm{SL}_2(\mathbb{C})$. A lattice $\Gamma < G$ is non-uniform if and only if Γ contains non-trivial unipotent elements.*

For the proof we will need the following lemma which will help us to understand the small 'loops' for points in $X = G/\Gamma$. Here we say that $g \in G$ is a *loop* at $x \in X$ if $g{\cdot}x = x$.[(29)]

Lemma 5.16 (Zassenhaus neighbourhoods). *There exists a norm $\|\cdot\|$ on $\mathrm{Mat}_d(\mathbb{C})$ such that for $\mathcal{N} = \{g \in \mathrm{GL}_d(\mathbb{C}) \mid \|g - I\| < 1\}$ all nontrivial discrete subgroups $\Gamma < \mathrm{GL}_d(\mathbb{C})$ generated by $\Gamma \cap \mathcal{N}$ have nontrivial centre. Moreover, for $g, h \in \mathcal{N}$ we have*

$$\big\|[g,h] - I\big\| \leqslant \|g - I\|\|h - I\|. \tag{5.8}$$

Proof. Let $\|\cdot\|_{\mathrm{op}}$ be the operator norm on $\mathrm{Mat}_d(\mathbb{C})$ satisfying

$$\|uv\|_{\mathrm{op}} \leqslant \|u\|_{\mathrm{op}}\|v\|_{\mathrm{op}}$$

for any $u, v \in \mathrm{Mat}_d(\mathbb{C})$. Let $g, h \in \mathrm{Mat}_d(\mathbb{C})$ and define $u = g - I$ and $v = h - I$. We suppose that $\|u\|_{\mathrm{op}} < \frac{1}{2}$ and $\|v\|_{\mathrm{op}} < \frac{1}{2}$. Notice that this implies that the geometric series giving $g^{-1} = (I + u)^{-1}$ and $h^{-1} = (I + v)^{-1}$ converge and that $\|g^{-1}\|_{\mathrm{op}}, \|h^{-1}\|_{\mathrm{op}} < 2$. For the commutator $[g, h] = g^{-1}h^{-1}gh$ of g, h we then obtain

$$\begin{aligned}
[g,h] &= (I+u)^{-1}(I+v)^{-1}(I+u)(I+v)\\
&= (I+u)^{-1}(I+v)^{-1}(I+u+v) + \mathrm{O}(\|u\|_{\mathrm{op}}\|v\|_{\mathrm{op}})\\
&= (I+u)^{-1}\big((I+v)^{-1}(I+v) + (I+v)^{-1}u\big) + \mathrm{O}(\|u\|_{\mathrm{op}}\|v\|_{\mathrm{op}})\\
&= (I+u)^{-1}\big(I+u+\mathrm{O}(\|v\|_{\mathrm{op}})u\big) + \mathrm{O}(\|u\|_{\mathrm{op}}\|v\|_{\mathrm{op}})\\
&= I + \mathrm{O}(\|u\|_{\mathrm{op}}\|v\|_{\mathrm{op}}).
\end{aligned}$$

To summarize, we have shown that there exists a constant $c \geqslant 1$ such that

$$\|[g,h]-I\|_{\mathrm{op}} \leqslant c\|g-I\|_{\mathrm{op}}\|h-I\|_{\mathrm{op}} \tag{5.9}$$

for all g,h with $\|g-I\|_{\mathrm{op}}, \|h-I\|_{\mathrm{op}} < \frac{1}{2}$. We assume $c \geqslant 2$, define $\|g\| = c\|g\|_{\mathrm{op}}$, multiply (5.9) by c and obtain (5.8) for all g,h with $\|g-I\|, \|h-I\| < \frac{c}{2}$. We define $\mathcal{N}$ as in the lemma, which in particular ensures that $[g,h] \in \mathcal{N}$ for all $g,h \in \mathcal{N}$.

Now let Γ be a nontrivial discrete subgroup generated by $\Gamma \cap \mathcal{N}$. Let

$$g \in \Gamma \cap \mathcal{N} \smallsetminus \{I\}$$

have the minimal distance to I with respect to the above norm. Then (5.8) shows that $\|[g,h]-I\| < \|g-I\|$ for all $h \in \Gamma \cap \mathcal{N}$. This forces $[g,h]=I$ and hence g belongs to the centre of $\langle \Gamma \cap \mathcal{N}\rangle = \Gamma$. □

The following special feature of SL_2 is particularly useful.

Lemma 5.17 (Centralizers). *Let $G = \mathrm{SL}_2(\mathbb{K})$ for $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Any two $g,h \in G$ that commute but do not belong to the centre define the same (automatically abelian) centralizers. If $g,h \in \mathrm{SL}_2(\mathbb{K})$ are close to the identity, then g and h commute if and only if $\log g$ and $\log h$ are linearly dependent over $\mathbb{K}$.*

Proof. It is sufficient to study the case $\mathbb{K} = \mathbb{C}$ and we will prove a version of the lemma for $\mathrm{Mat}_2(\mathbb{C})$. Assume first that $g = \mathrm{diag}(\alpha,\beta)$ for some $\alpha \neq \beta \in \mathbb{C}$. A simple calculation shows that g and some $h \in \mathrm{Mat}_2(\mathbb{C})$ commute if and only if h is also diagonal. By conjugation the first claim of the lemma follows in a more general form within $\mathrm{Mat}_2(\mathbb{C})$ if one of g or h is diagonalizable.

If g is not diagonalizable we may assume that $g = \lambda I + \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. For $h = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ we then have $gh = \begin{pmatrix} \lambda a + c & \lambda b + d \\ \lambda c & \lambda d \end{pmatrix}$ and $hg = \begin{pmatrix} \lambda a & \lambda b + a \\ \lambda c & \lambda d + c \end{pmatrix}$. If h commutes with g this gives $h = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}$. If h is not a scalar multiple of I we obtain that h once again has the same structure as g. Together with the above, this gives the first claim in the lemma for $\mathrm{Mat}_2(\mathbb{C})$.

For the second claim suppose that $g,h \in \mathrm{SL}_2(\mathbb{K}) \smallsetminus \{I\}$ commute and are close to the identity so that $u = \log g$ and $v = \log h$ are well-defined. Then $g = \exp(u)$ commutes with $\exp(su)$ for $s \in \mathbb{C}$ and $h = \exp(v)$ commutes with $\exp(tv)$ for t in $\mathbb{C}$. By the first part of the lemma the two one-parameter subgroups defined by $s \mapsto \exp(su)$ and $t \mapsto \exp(tv)$ commute. Moreover, this implies that $[u,v] = 0$. Now the first part of the proof implies that the traceless matrices $u, v \in \mathfrak{sl}_2(\mathbb{C})$ are multiples of each other. □

Proof of Proposition 5.15. Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ and let $\Gamma < G = \mathrm{SL}_2(\mathbb{K})$ be a lattice. Let $a_t = \mathrm{diag}\big(\mathrm{e}^{-\frac{t}{2}}, \mathrm{e}^{\frac{t}{2}}\big)$ as before and let $U = G^+_{a_1} < G$ the lower

unipotent subgroup. If Γ contains a nontrivial unipotent element γ then there exists some $g \in G$ so that $u = g\gamma g^{-1} \in U$. However, this implies that the point $g\Gamma$ satisfies $u{\cdot}g\Gamma = g\Gamma$. Applying a_{-t} gives $a_{-t}ua_t{\cdot}a_{-t}g\Gamma = a_{-t}g\Gamma$. As $u \neq I$ and $a_{-t}ua_t \to I$ as $t \to \infty$ it follows that $a_{-t}g\Gamma \to \infty$ as $t \to \infty$ by the divergence criterion in Proposition 1.35. Hence X is non-compact.

The converse is the more difficult direction. So suppose that Γ contains no nontrivial unipotent elements. Let $\mathcal{N}_0 \subseteq \mathrm{SL}_2(\mathbb{R})$ be an open neighbourhood of Γ so that the conclusions of Lemmas 5.16 and 5.17 hold on $\mathcal{N}_0$. Let $\mathcal{N}$ be an open neighbourhood with $\overline{\mathcal{N}} \subseteq \mathcal{N}_0$. By the divergence criterion in Proposition 1.35 the set

$$K = \big\{x \in X \mid \mathcal{N} \ni g \longmapsto g{\cdot}x \text{ is injective}\big\}$$

is compact. For $x_0 \in K$ there might be loops $g \in \mathcal{N}_0 \smallsetminus \mathcal{N}$ with $g{\cdot}x_0 = x_0$. Note that if $x_0 = g_0\Gamma$ then there exists some $\gamma \in \Gamma$ with $gg_0 = g_0\gamma$, which shows that the characteristic polynomial of the loop g is also the characteristic polynomial of γ (corresponding to the loop g at x_0). Similarly replacing x_0 by hx_0 for some $h \in G$ creates a loop hgh^{-1} at hx_0 with the same characteristic polynomial. A simple compactness argument now shows that varying $x_0 \in K$ and $g \in \mathcal{N}_0$ gives only a finite set $\mathscr{F} \subseteq \mathbb{K}[T]$ of characteristic polynomials of loops. As every such polynomial is also a characteristic polynomial of an element of Γ and we assume that Γ contains no nontrivial unipotent elements we deduce that $\mathscr{F}$ does not contain $(T-1)^2$. Hence there exists a neighbourhood O of $I \in \mathrm{SL}_2(\mathbb{K})$ so that $g \in O$ implies that the characteristic polynomial of g does not belong to $\mathscr{F}$.

We will show that $g \in O \smallsetminus \{I\}$ cannot appear as a loop of any $x_0 \in X$. By the divergence criterion in Proposition 1.35 this then shows that X must be compact. So suppose for the purposes of a contradiction that $g \in O \smallsetminus \{I\}$ is a loop at some $x_0 = g_0\Gamma$. Then $g_0^{-1}gg_0 \in \Gamma$ or equivalently $g \in \Lambda = g_0\Gamma g_0^{-1}$. As Λ is discrete we may apply Lemmas 5.16 and 5.17 and obtain that $\Lambda \cap \mathcal{N} \subseteq \exp(\mathbb{K}v)$ for some unit vector $v \in \mathrm{SL}_2(\mathbb{K})$. We may assume that $g = \exp(su) \in \Lambda \cap \mathcal{N} \smallsetminus \{I\}$ is a smallest element with $|s| < 1$.

If $v = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}$ has $c \neq 0$, we apply a_t with $t > 0$ to x_0. Simultaneously we conjugate the elements of $\Lambda \cap \mathcal{N}$ and the smallest loop $a_tga_{-t} = \exp(s \operatorname{Ad}_{a_t} u)$. Using the fact that c is expanded and continuity we find $t \in \mathbb{R}$ so that the loop a_tga_{-t} at $a_t{\cdot}x_0$ belongs to $\mathcal{N}_0 \smallsetminus \overline{\mathcal{N}}$. As this was the smallest loop, Lemmas 5.16 and 5.17 imply that $a_t{\cdot}x_0 \in K$. By our choice of O this gives a contradiction. If $c = 0$ but $b \neq 0$ we similarly apply a_t with $t < 0$. If $v = \begin{pmatrix} a & 0 \\ 0 & -a \end{pmatrix}$ we first apply $\begin{pmatrix} 1 & \\ 1 & 1 \end{pmatrix}$ to x_0, which replaces v by $\begin{pmatrix} a & \\ 2a & -a \end{pmatrix}$. After this we apply a_t for $t > 0$ and argue as above. □

With these preparations the remaining results from Section 5.1.2 on the horocycle flow follow quickly.

Proof of Theorem 5.7. Suppose that $\Gamma < \mathrm{SL}_2(\mathbb{R})$ is a non-uniform lattice. By Proposition 5.15 Γ contains a nontrivial unipotent element γ which is conjugated

to an element $g\gamma g^{-1} \in U$ for some $g \in \mathrm{SL}_2(\mathbb{R})$. However, this implies that $Ug\Gamma$ is a periodic orbit supporting a one-dimensional U-invariant Lebesgue measure.

We still have to show that m_X and the one-dimensional periodic orbit measure are the only U-invariant and ergodic probability measures on X. So assume that μ is a U-invariant and ergodic probability measure on X and let $x \in X$ be a generic point for μ. We consider the orbit $a_{-t}\cdot x$ for $t \geqslant 0$.

If x is periodic under U, then μ is the one-dimensional Lebesgue measure on $U\cdot x$ and $a_{-t}\cdot x \to \infty$ for $t \to \infty$. So suppose now that x is not periodic under U, let $t_0 > 0$ be arbitrary, $x' = a_{-t_0}\cdot x$, and $\mathcal{N} \subseteq \mathcal{N}_0$ be the neighbourhoods of I as in the proof of Proposition 5.15. Suppose x' has a loop in $\mathcal{N}$. As x is not periodic under U and a_{t_0} normalizes U we see that x' is also not periodic and so the loop cannot belong to U. We now argue along the lines of the proof of Proposition 5.15: Let $g = \exp(v)$ be a smallest loop at x with $v = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \neq 0$. If $b \neq 0$ then the smallest loop at $a_{-t}\cdot x$ eventually grows for $t > t_0$. If $b = 0$ and $a \neq 0$, then the smallest loop at $a_{-t}\cdot x$ will not grow but also will not go to zero. Finally, $b = 0$, $a = 0$, and $c \neq 0$ would mean that x' and hence also x are periodic for U. It follows that for any t_0 there exists $t \geqslant t_0$ so that a_{-t} belongs to a fixed compact subset $K \subseteq X$. This allows us to find a subsequence $t_n \to \infty$ with $a_{-t_n}\cdot x \in K$, apply Theorem 5.9, and obtain a subsequence of times for which the time average along $U\cdot x$ converges to $\int f \,\mathrm{d}m_X$ for all $f \in C_c(X)$. As the time average converges to $\int f \,\mathrm{d}\mu$ by the choice of x we obtain $\mu = m_X$. □

Proof of Theorem 5.8. Let x be periodic for U. Then the injectivity radius at any point in $a_t\cdot Ux$ goes to 0 for $t \to -\infty$, which shows that $a_t\cdot(Ux)$ diverges for $t \to -\infty$. For $t \to \infty$ we let $K = \{x\}$, let $B_0 \subseteq U$ be an interval so that $B_0 \ni s \mapsto u_s\cdot x \in Ux$ is bijective, and apply Theorem 5.9 to see that the normalized Lebesgue measure on the expanding orbits $a_t\cdot Ux$ equidistributes for $t \to \infty$ to the Haar measure m_X on X. □

5.3.2 Equidistribution for Non-Compact Quotients

Dani and Smillie showed in [26] that even for non-compact quotients of $\mathrm{SL}_2(\mathbb{R})$ a rather strong equidistribution theorem holds: A horocycle orbit is either periodic or it equidistributes with respect to the uniform measure m_X.

For higher dimensional non-compact quotients $X = G/\Gamma$ and their horospherical actions other possibilities can occur. For the following characterization of whether or not a horospherical orbit equidistributes we specialize to the case where the horospherical subgroup is abelian.

Theorem 5.18. *Let $G\cdot x_0 \subseteq \mathsf{X}_d = \mathrm{SL}_d(\mathbb{R})/\mathrm{SL}_d(\mathbb{Z})$ be a finite volume orbit for some closed connected subgroup $G \leqslant \mathrm{SL}_d(\mathbb{R})$ and some point $x_0 \in \mathsf{X}_d$. Let $a \in G$ have only real and positive eigenvalues so that the action of a is mixing with respect to $m_{G\cdot x_0}$. Let $U = G_a^+$ be the unstable horospherical subgroup of a and*

suppose that it is abelian. Let (F_n) be a Følner sequence in U containing I consisting of blocks whose sides are parallel to some fixed coordinate system spanned by eigenvectors for the conjugation map by a. Then for every $x \in G{\bullet}x_0$ the following are equivalent:

(1) *The U-orbit through x is equidistributed, meaning that*

$$\frac{1}{m_U(F_n)} \int_{F_n} f(u)\, \mathrm{d}m_U(u) \to \int_{\mathsf{X}_d} f \, \mathrm{d}m_{\mathsf{X}_d}$$

as $n \to \infty$ for any $f \in C_c(\mathsf{X}_d)$.

(2) *The orbit $U{\bullet}x$ is not contained in a closed orbit $L{\bullet}x$ for any proper connected subgroup $L < G$.*

If, in addition, $G = \mathrm{SL}_d(\mathbb{R})$ and $x = g\,\mathrm{SL}_d(\mathbb{Z})$ for some $g \in \mathrm{SL}_d(\mathbb{R})$ then we also have the equivalence to the next property.

(3) *There is no rational subspace $V \subseteq \mathbb{R}^d$ for which gV is fixed by U and expanded by a.*

Proof. We let $x \in G{\bullet}x_0$ be as in the theorem. If the U-orbit of x is contained in a closed orbit of a proper connected subgroup $L < G$ as in (2), then clearly we cannot have equidistribution of the U-orbit as in (1). This shows that (1) implies (2).

Assume now that the U-orbit $U{\bullet}x$ is not contained in a closed orbit $L{\bullet}x$ for any proper closed subgroup $L < G$ as in (2). Fix some $f \in C_c(\mathsf{X}_d)$ and $\varepsilon > 0$. We let $x_0 = g_0\,\mathrm{SL}_d(\mathbb{Z})$, let $\Lambda_0 = g_0\mathbb{Z}^d$ be the lattice corresponding to x_0, and define

$$\eta = \min\left\{ \mathrm{covol}(\Lambda_0 \cap V, V)^{1/\dim V} \;\middle|\; V \text{ is } \Lambda_0\text{-rational} \right\}. \tag{5.10}$$

By quantitative non-divergence for the action of $U = G_a^+$ there exists some compact set $K \subseteq \mathsf{X}_d$ with the property as in Proposition 4.13 with $\delta = \varepsilon$. We let B_0 be the symmetric cube (that is, centred at the origin) in

$$U = G_a^+ \cong \mathbb{R}^\ell$$

satisfying the injectivity requirement of Theorem 5.9 on K. Applying that theorem to f, B_0K, and ε, we find some $k \geqslant 1$ such that

$$\left| \frac{1}{m_{G_a^+}(a^k B_0 a^{-k})} \int\limits_{a^k B_0 a^{-k}} f(u{\bullet}y)\, \mathrm{d}m_{G_a^+}(u) - \int\limits_X f \, \mathrm{d}m_X \right| < \varepsilon \tag{5.11}$$

whenever $B_0 a^{-k}{\bullet}y$ intersects K non-trivially.

Now let $x' = a^{-k}{\bullet}x$ and notice that it may not belong to K. Since (F_n) is chosen to be a Følner sequence consisting of blocks, the same is true for $a^{-k}F_n a^k$. If $U = G_a^+$ fixes a $\Lambda_{x'}$-rational subspace V of covolume $< \eta^{\dim V}$, then we can define the subgroup

$$L = \mathrm{Stab}^1_G(V) = \{g \in G \mid gV = V \text{ and } g|_V \text{ has determinant } 1\} \lneq G$$

that does not contain a (by our definition of η in (5.10)). Exercise 3.12 shows that Lx' is closed, which shows that

$$a^k L{\cdot}x' = a^k L a^{-k}{\cdot}x$$

is a closed orbit of a proper subgroup which contradicts our assumption in (2).

It follows that U does not fix any Λ_x-rational subspaces V with covolume less than $\eta^{\dim V}$. Applying Proposition 4.13 we see now that for large enough n we have

$$\frac{1}{m_{G_a^+}\left(a^{-k}F_n a^k\right)} m_{G_a^+}\left(\left\{u \in a^{-k}F_n a^k \;\middle|\; u{\cdot}x' \notin K\right\}\right) < \varepsilon. \tag{5.12}$$

We now split $a^{-k}F_n a^k$ into translates $B_0 u_\ell$ for $\ell = 1, \dots, L$ of the cube B_0. Ignoring the effects of the boundary which contribute no more than $\mathrm{o}_f(1)$ to the ergodic average as $n \to \infty$, we now have

$$\begin{aligned}\frac{1}{m_{G_a^+}(F_n)} \int\limits_{F_n} f(u{\cdot}x)\,\mathrm{d}m_{G_a^+} \\ = \frac{1}{L}\sum_{\ell=1}^{L} \frac{1}{m_{G_a^+}(a^k B_0 a^{-k})} \int\limits_{a^k B_0 a^{-k}} f\left(u a^k u_\ell a^{-k}{\cdot}x\right)\,\mathrm{d}m_{G_a^+} + \mathrm{o}_f(1).\end{aligned}$$

For all those ℓ for which $B_0 u_\ell a^{-k}{\cdot}x$ intersects K the corresponding average is ε-close to $\int_X f\,\mathrm{d}m_X$ by (5.11). However, the number of boxes $B_0 u_\ell{\cdot}x'$ that do not intersect K is controlled by (5.12), and gives

$$\frac{1}{m_{G_a^+}(F_n)} \int\limits_{F_n} f(u{\cdot}x)\,\mathrm{d}m_{G_a^+}(u) = \int\limits_X f\,\mathrm{d}m_X + \mathrm{o}_f(1) + \mathrm{O}_f(\varepsilon)$$

for $n \to \infty$. As $\varepsilon > 0$ and $f \in C_c(X)$ were arbitrary, this shows (1).

Now suppose that $G = \mathrm{SL}_d(\mathbb{R})$. We note that (2) implies (3) by Exercise 3.12. It remains to show that (3) implies (1). For this let

$$a = \begin{pmatrix} \lambda^{-n} I_m & \\ & \lambda^m I_n \end{pmatrix} \in \mathrm{SL}_d(\mathbb{R})$$

for some $\lambda > 1$ so that

$$G_a^+ = \left\{ \begin{pmatrix} I_m & \\ * & I_n \end{pmatrix} \right\}$$

is indeed abelian (up to conjugation and the choice of m and n this is the only choice of a for which G_a^+ is abelian). Suppose now that $V \subseteq \mathbb{R}^d$ is a proper G_a^+-invariant subspace. Then either $V \subseteq \{0\}^m \times \mathbb{R}^n$ or V contains some $v = (v_m, v_n)$

with $v_m \in \mathbb{R}^m \smallsetminus \{0\}$ and $v_n \in \mathbb{R}^n$, which implies that $\{0\}^m \times \mathbb{R}^n \subseteq V$. In both cases $aV = V$ and the restriction of a to V has determinant bigger than 1. It follows that (3) is equivalent to the assumption that U does not fix any proper $g\mathbb{Z}^d$-rational subspace. By setting $\eta = 1$ the above argument now proves (1). □

In the exercises we outline how one can remove the assumptions on commutativity of U.

Exercise 5.19. Let $G{\bullet}x_0 \subseteq \mathsf{X}_d = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$ be a finite volume orbit for some closed connected subgroup $G \leqslant \mathrm{SL}_d(\mathbb{R})$ and some point $x_0 \in \mathsf{X}_d$. Let $a \in G$ have only real and positive eigenvalues so that the action of a is mixing with respect to $m_{G{\bullet}x_0}$. Let $U = G_a^+$ be the unstable horospherical subgroup of a and let F_n be as in Exercise 4.14. Let $x \in G{\bullet}x_0$. Suppose that U does not fix any Λ_x-rational subspace which is not also fixed by G. Show that $F_n{\bullet}x$ equidistributes in $X = G{\bullet}x_0$ in the sense that

$$\frac{1}{m_U(F_n)} \int_{F_n} f(u{\bullet}x) dm_U(u) \longrightarrow \int_X f dm_X$$

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

5.4 The Counting Method of Duke–Rudnick–Sarnak and Eskin–McMullen

We return to the topic of Sections 1.1 and 1.2.6. In fact we wish to explain the work of Eskin and McMullen [52] who use mixing to establish asymptotic counting results in a more general context. For this (and in preparation for other special cases to be considered later) we describe in this section the general set-up for the work of Duke, Rudnick and Sarnak [38] (which is also used in the work of Eskin and McMullen) on how to relate a counting problem for points in Γ-orbits on $V = G/H$ to the equidistribution problem for 'translated' H-orbits of the form

$$gH{\bullet}\Gamma \subseteq X = G/\Gamma$$

for varying $gH \in V$.

In many cases (for example, in the context of 'affine symmetric spaces'), the methods of this chapter can be used to give the asymptotic of the counting for the number of integer points on varieties. In fact, suppose G and H consist of the $\mathbb{R}$-points of algebraic groups $\mathbb{G}$ and $\mathbb{H}$ defined over $\mathbb{Q}$ respectively, the quotient $V = G/H$ can be identified with the $\mathbb{R}$-points of an affine variety $\mathbb{V}$ defined over $\mathbb{Q}$, and $\mathbb{V}(\mathbb{Z})$ is non-empty. Then we get that $\mathbb{V}(\mathbb{Z})$ is a disjoint union

$$\mathbb{V}(\mathbb{Z}) = \bigsqcup_i \mathbb{G}(\mathbb{Z})v_i$$

of different $\Gamma = \mathbb{G}(\mathbb{Z})$-orbits. Frequently this is a finite union, and then one gets the asymptotic for $|\mathbb{V}(\mathbb{Z}) \cap B_t|$ by assembling the results for the individual counts $|\mathbb{G}(\mathbb{Z})v_i \cap B_t|$. We will discuss the details of such integer point counting

problems in special cases in the remaining sections of this chapter, and we refer to the papers of Duke, Rudnick and Sarnak and of Eskin and McMullen [38, 52] for a detailed discussion of the general problem of counting lattice points in 'affine symmetric spaces'.

5.4.1 Compatibility of all Haar Measures Involved

In order to state both method and result, we have to briefly describe the necessary compatibility of all the Haar measures involved. Let m_G be a Haar measure on a unimodular group G, and let $\Gamma < G$ be a lattice, on which we choose counting measure as the Haar measure. As we know m_G induces in a natural way a Haar measure m_X on $X = G/\Gamma$, giving total mass $m_X(X) = m_G(F)$ where $F \subseteq G$ is a fundamental domain for (the right action of) Γ.

Assume that $H < G$ is a closed unimodular subgroup with Haar measure m_H. Then (see Section A.2) we may define a locally finite measure $m_{G/H}$ with the following compatibility property, which is analogous to Fubini's theorem if G is thought of measurably as a product of H and G/H. If $f \in L^1_{m_G}(G)$ then the function F defined by the relation

$$F(gH) = \int_H f(gh)\,\mathrm{d}m_H(h) \tag{5.13}$$

exists for almost every $g \in G$, and the measure $m_{G/H}$ satisfies

$$\int_{G/H} F(gH)\,\mathrm{d}m_{G/H} = \int_G f\,\mathrm{d}m_G. \tag{5.14}$$

5.4.2 First step: Equidistribution gives an Averaged Counting Result

Let $\Gamma < G$ be a lattice, and assume that $H < G$ is a closed subgroup with the property that $\Gamma \cap H < H$ is also a lattice. Let $Y = H/\Gamma \cap H$ identified with the closed orbit $H{\cdot}\Gamma \subseteq X$, and let m_Y be the Haar measure on Y induced by the Haar measure m_H on H. We make the following[†] *equidistribution assumption:*

$$\text{the translated } H\text{-orbits } gH{\cdot}\Gamma \text{ equidistribute in } X = G/\Gamma \tag{5.15}$$

as $gH \to \infty$ in G/H. In other words the push-forward of $\frac{1}{m_Y(Y)}m_Y$ under g should converge to $\frac{1}{m_X(X)}m_X$ in the weak* topology as $gH \to \infty$.

The assumptions above already imply a weak* version of our desired counting result in the following sense. We let $\{B_t \mid t \geqslant 0\}$ be a collection of subsets of G/H

[†] Alternatively, we may just assume a form of 'equidistribution on average'—in a sense to be made clear in the proof.

each with finite Haar measure, and define for $t \geqslant 0$ a modified orbit-counting function $F_t \colon X \to \mathbb{R}_{\geqslant 0}$ by

$$F_t(g\Gamma) = \frac{1}{m_{G/H}(B_t)} \left| g\Gamma_{\bullet} H \cap B_t \right|, \tag{5.16}$$

which counts elements in B_t within the Γ-orbit of $H \in G/H$ translated by g.

Proposition 5.20 (Weak* Counting Result). *If $m_{G/H}(B_t) \to \infty$ as $t \to \infty$, then* (5.15) *implies the* weak*-convergence

$$F_t \, \mathrm{d}m_X \longrightarrow \frac{m_Y(Y)}{m_X(X)} \, \mathrm{d}m_X \tag{5.17}$$

as $t \to \infty$, where $Y = H/\Gamma \cap H$ and $X = G/\Gamma$.

5.4.3 Second step: Additional Geometric Assumptions imply the Counting Result

In order to be able to obtain the desired counting result from the averaged weak counting result above, we need to assume that the sets B_t are well behaved in a geometric manner.

Definition 5.21 (Geometric Assumption). A monotonically increasing family $\{B_t \mid t \geqslant 0\}$ of subsets of G/H is *well-rounded* if $m_{G/H}(B_t) \to \infty$ as $t \to \infty$, for every $\delta > 0$ there exists a neighbourhood U of $I \in G$ with

$$B_{t-\delta} \subseteq \bigcap_{g \in U} gB_t \subseteq B_t \subseteq \bigcup_{g \in U} gB_t \subseteq B_{t+\delta},$$

and furthermore for every $\varepsilon > 0$ there exists $\delta > 0$ with

$$\frac{m_{G/H}(B_{t+\delta})}{m_{G/H}(B_t)} < 1 + \varepsilon$$

for all $t \geqslant 0$.

Theorem 5.22 (Asymptotic Counting). *If $\Gamma < G$ and $\Gamma \cap H < H$ are lattices, the translated H-orbits equidistribute as assumed in* (5.15)*, and the family of sets $\{B_t\}$ is well-rounded as above, then we have the asymptotic*

$$\lim_{t \to \infty} \frac{1}{m_{G/H}(B_t)} \left| \Gamma_{\bullet} H \cap B_t \right| = \frac{m_Y(Y)}{m_X(X)} \tag{5.18}$$

for the orbit-point counting problem, where $Y = H/\Gamma \cap H$ and $X = G/\Gamma$.

We note that Selberg's Theorem 1.15 concerning $\mathrm{PSL}_2(\mathbb{Z})_{\bullet}\mathrm{i} \subseteq \mathbb{H}$ turns out to be a very special case of this setup.

5.4.4 Proofs

We now turn to considering the components of the outlined argument in greater detail and start by proving the equidistribution of expanding circles in Theorem 1.16 (leading to the instance of (5.15) needed for Selberg's theorem).

Proof of Theorem 1.16. The argument is similar to the banana mixing argument from Section 5.2 but easier. Let $\Gamma < \mathrm{SL}_2(\mathbb{R}) = G$ be a lattice, let $K = \mathrm{SO}_2(\mathbb{R})$, let $A = \{a_t \mid t \in \mathbb{R}\}$ be the diagonal subgroup, and let $N = G_a^-$ be the stable horocycle subgroup. By the Iwasawa decomposition in Proposition 1.55 we have $G = NAK$ and uniqueness of the corresponding decomposition. Moreover, by Lemma 1.58 the Haar measure on G is the direct product of the Haar measure m_{NA} on $B = NA$ and the Haar measure on K.

Let $x_0 \in X = G/\Gamma$, $f \in C_c(X)$, and $\varepsilon > 0$. Using a compactness argument as in Lemma 5.14 there exists a $\delta > 0$ such that the map

$$B_\delta^{NA} \times K\cdot x_0 \ni (h, k\cdot x_0) \longmapsto hk\cdot x_0 \in X$$

is injective. By shrinking δ we may assume that it satisfies the uniform continuity claim in (5.3) for f and ε. Using this and applying mixing for f and $\frac{1}{m_G(B_\delta^{NA}K\cdot x)}\mathbb{1}_{B_\delta^{NA}K\cdot x}$ leads to the desired estimate. □

Proof of the instance of (5.15) for Theorem 1.15. Let $H = K = \mathrm{SO}_2(\mathbb{R})$ and (g_n) in $G = \mathrm{SL}_2(\mathbb{R})$ so that $g_nK \to \infty$ as $n \to \infty$. Applying the Cartan decomposition to g_n we find $k_n \in \mathrm{SO}_2(\mathbb{R})$ and diagonal matrices a_{t_n} with $t_n \geqslant 0$ so that $g_nK = k_n a_{t_n} K$ for all $n \geqslant 1$. By choosing a subsequence we may assume that $k_n \to k$ as $n \to \infty$ for some $k \in K$.

Now fix $f \in C_c(X)$ and $\varepsilon > 0$. By uniform continuity we have

$$|f(k_n\cdot x) - f(k\cdot x)| < \varepsilon$$

for $x \in X$ and all sufficiently large n. Applying in addition the equidistribution of expanding circles in Theorem 1.16 to $f \circ k$ we have that

$$\left|\int_Y f(g_n\cdot y)\,\mathrm{d}m_Y(y) - \int_X f\,\mathrm{d}m_X\right| \leqslant \varepsilon + \left|\int_Y f(ka_{t_n}\cdot y)\,\mathrm{d}m_Y(y) - \int_Y f\,\mathrm{d}m_X\right| \leqslant 2\varepsilon$$

for all sufficiently large n. □

Combining Theorem 5.22 with the required version of (5.15) and using the fact that balls in $\mathbb{H}$ are well-rounded (see Exercise 5.24) then gives us Selberg's Theorem 1.15.

We return now to the general setup considered in Sections 5.4.1–5.4.3.

Proof of weak*-convergence in Proposition 5.20. We assume (5.15), or more precisely that the normalized translation

$$\frac{1}{m_Y(Y)} g_* m_Y$$

of the Haar measure m_Y on

$$Y = H/\Gamma \cap H \subseteq X = G/\Gamma$$

translated by $gH \in G/H$ converges to the normalized Haar measure

$$\frac{1}{m_X(X)} m_X$$

in the following averaged sense. For a test function $\alpha \in C_c(X)$ we require that

$$\begin{aligned} \frac{1}{m_Y(Y) m_{G/H}(B_t)} \int_{B_t} \int_Y \alpha(gh\Gamma) \, \mathrm{d}m_Y(h\Gamma) \, \mathrm{d}m_{G/H}(gH) \\ \longrightarrow \frac{1}{m_X(X)} \int \alpha \, \mathrm{d}m_X \end{aligned} \tag{5.19}$$

as $t \to \infty$. As $m_{G/H}$ is locally finite this is certainly satisfied if both

$$\frac{1}{m_Y(Y)} \int_Y \alpha(gh\Gamma) \, \mathrm{d}m_Y \longrightarrow \frac{1}{m_X(X)} \int \alpha \, \mathrm{d}m_X$$

as $gH \to \infty$ in G/H and

$$m_{G/H}(B_t) \longrightarrow \infty$$

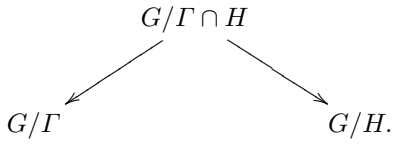
as $t \to \infty$, but (5.19) is a weaker requirement because of the additional averaging.

We wish to deduce from this assumption that

$$A_t^\alpha = \int_X F_t(x) \alpha(x) \, \mathrm{d}m_X \longrightarrow \frac{m_Y(Y)}{m_X(X)} \int_X \alpha \, \mathrm{d}m_X$$

as $t \to \infty$.

The proof is relatively short, and consists of an application of the following folding/unfolding trick (see also Proposition 1.31) using the spaces



By the definition of F_t in (5.16) we have

$$A_t^\alpha = \int_X F_t(x)\alpha(x)\,\mathrm{d}m_X = \frac{1}{m_{G/H}(B_t)} \int_{G/\Gamma} \sum_{\gamma\in\Gamma/\Gamma\cap H} \mathbb{1}_{B_t}(g\gamma\cdot H)\alpha(g\Gamma)\,\mathrm{d}m_X(g\Gamma),$$

in which the sum over $\gamma \in \Gamma/\Gamma\cap H$ denotes the sum over a list of representatives of the cosets of $\Gamma\cap H$ in Γ. Thus by using the compatibility of the Haar measures we get

$$\begin{aligned} A_t^\alpha &= \frac{1}{m_{G/H}(B_t)} \int_{G/\Gamma\cap H} \mathbb{1}_{B_t}(gH)\alpha(g\Gamma)\,\mathrm{d}m_{G/\Gamma\cap H}(g(\Gamma\cap H)) \\ &= \frac{1}{m_{G/H}(B_t)} \int_{G/H} \mathbb{1}_{B_t}(gH) \int_{H/\Gamma\cap H} \alpha(gh\Gamma)\,\mathrm{d}m_Y(h\Gamma)\,\mathrm{d}m_{G/H}(gH) \\ &= \frac{1}{m_{G/H}(B_t)} \int_{B_t}\int_Y \alpha(gh\Gamma)\,\mathrm{d}m_Y(h\Gamma)\,\mathrm{d}m_{G/H}(gH). \end{aligned}$$

For the first unfolding step note that if $F \subseteq G$ is a fundamental domain for Γ then

$$\bigsqcup_{\gamma\in\Gamma/\Gamma\cap H} F\gamma$$

is a fundamental domain for $\Gamma\cap H$. For the second, we use (5.13)–(5.14) and note that a fundamental domain $F \subseteq G$ for $\Gamma\cap H$ intersects any coset gH in the g-translate of a fundamental domain for $\Gamma\cap H$ in H.

Finally note that the last expression for A_t^α converges by our assumption in (5.19) to

$$\frac{m_Y(Y)}{m_X(X)} \int_X \alpha\,\mathrm{d}m_X$$

as $t \to \infty$. □

Proof of the pointwise count in Theorem 5.22. We now suppose that the weak*-convergence discussed above holds, and that the family of sets B_t is well-rounded as in Definition 5.21. From this we wish to derive the asymptotic

$$\frac{1}{m_{G/H}(B_t)} \left|(\Gamma\cdot H)\cap B_t\right| \longrightarrow \frac{m_Y(Y)}{m_X(X)}$$

as $t \to \infty$.

Let $\varepsilon > 0$ be arbitrary, and choose $\delta > 0$ so that

$$\frac{m_{G/H}(B_{t+\delta})}{m_{G/H}(B_t)} < 1 + \varepsilon$$

for all $t \geqslant 0$, and choose a symmetric neighbourhood $U = U^{-1} \subseteq G$ of $I \in G$ with

$$UB_t \subseteq B_{t+\delta}$$

for all t. Further let $\alpha \in C_c(X)$ be an approximate identity at the identity coset, in the sense that $\alpha \geqslant 0$, $\int_X \alpha \, \mathrm{d}m_X = 1$, and $\mathrm{supp}(\alpha) \subseteq U\Gamma$. Then we have for any $g \in U$ that

$$\begin{aligned} F_{t+\delta}(g) &= \frac{1}{m_{G/H}(B_{t+\delta})} \big| g\Gamma \cdot H \cap B_{t+\delta} \big| \\ &= \frac{1}{m_{G/H}(B_{t+\delta})} \Big| \Gamma \cdot H \cap \underbrace{g^{-1} B_{t+\delta}}_{\supseteq B_t} \Big| \\ &\geqslant \frac{m_{G/H}(B_t)}{m_{G/H}(B_{t+\delta})} \frac{1}{m_{G/H}(B_t)} \big| \Gamma \cdot H \cap B_t \big| \\ &\geqslant \frac{1}{1+\varepsilon} \frac{1}{m_{G/H}(B_t)} \big| \Gamma \cdot H \cap B_t \big|. \end{aligned}$$

Multiplying by α, integrating with respect to m_X and letting $t \to \infty$ gives

$$\limsup_{t\to\infty} \frac{1}{m_{G/H}(B_t)} \, |\Gamma \cdot H \cap B_t| \leqslant (1+\varepsilon) \frac{m_Y(Y)}{m_X(X)}.$$

The second inequality is derived in the same way (see Exercise 5.23). □

Exercise 5.23. Give a detailed argument to show that

$$\liminf_{t\to\infty} \frac{1}{m_{G/H}(B_t)} \, |\Gamma \cdot H \cap B_t| \geqslant (1+\varepsilon)^{-1} \frac{m_Y(Y)}{m_X(X)}.$$

To conclude the proof of Selberg's counting result in Theorem 1.15 the following is needed.

Exercise 5.24. Recall that the hyperbolic area of a ball $B_t^{\mathbb{H}}$ of radius $r \geqslant 0$ is $2\pi(\cosh R - 1)$ and use this to show that $B_t^{\mathbb{H}}$ is well-rounded in the sense of Definition 5.21,

5.5 Counting Integer Points on Quadratic Hypersurfaces

In this section we study our first class of examples of 'affine symmetric' varieties, namely the case of quadratic hypersurfaces. Let Q be a non-degenerate indefinite quadratic form with integer coefficients in $d \geqslant 3$ variables and $a \in \mathbb{Z}\smallsetminus\{0\}$. Then

$$\mathbb{V}(\mathbb{R}) = \{v \in \mathbb{R}^d \mid Q(v) = a\}$$

can be identified with $\mathbb{G}(\mathbb{R})/\mathbb{H}(\mathbb{R})$ for $\mathbb{G} = \mathrm{SO}_Q$ and $\mathbb{H} = \mathrm{Stab}_{\mathbb{G}}(v_0)$ for some $v_0 \in \mathbb{V}(\mathbb{R})$ as the $\mathbb{G}(\mathbb{R})$-action is transitive[†] by Witt's theorem.[(30)] Let us assume that $\mathbb{V}(\mathbb{Z})$ is non-empty and that $v_0 \in \mathbb{V}(\mathbb{Z})$. If now $\mathbb{H}(\mathbb{Z})$ is a lattice in $\mathbb{H}(\mathbb{R})$ (which is always the case for $d \geqslant 4$ and in many cases also for $d = 3$) then we can derive from the methods of the last section the asymptotics for the counting problem on $\mathbb{V}(\mathbb{Z})$.

Exercise 5.25. Show that $\mathbb{G}(\mathbb{R})$ acts transitively on $\mathbb{V}(\mathbb{R})$. Show also that $\mathbb{G}(\mathbb{R})^o$ acts transitively on every connected component of $\mathbb{V}(\mathbb{R})$.

5.5.1 The Asymptotic Counting Result

We wish to discuss the counting result now in greater detail. For the following calculations it is convenient to sometimes fix a particular quadratic form. So we set $Q_0(x_1, \dots x_p, y_1, \dots, y_q) = x_1^2 + \cdots + x_p^2 - y_1^2 - \cdots - y_q^2$ with $p, q \geqslant 1$, and $d = p + q \geqslant 3$, while Q denotes a general non-degenerate quadratic form of signature (p, q).

Corollary 5.26 (Counting on quadratic hypersurfaces). *Let $a \in \mathbb{Z}\smallsetminus\{0\}$, let $\mathbb{V} = \{v \mid Q(v) = a\}$ and assume that $\mathbb{V}(\mathbb{Z})$ is non-empty. Suppose furthermore that either $d \geqslant 4$, that $d = 3$ and $0 \notin Q(\mathbb{Q}^3\smallsetminus\{0\})$, or that (p, q) is $(2, 1)$, $Q = Q_0$, and a is not a square in $\mathbb{Z}$. Define*

$$B_R^V = \big\{v \in \mathbb{V}(\mathbb{R}) \;\big|\; \|v\| \leqslant R\big\}$$

for a suitable Euclidean norm $\|\cdot\|$ on $\mathbb{R}^d$. Then there exists constants $c > 0$ and $c' > 0$ such that

$$\big|\mathbb{V}(\mathbb{Z}) \cap B_R^V\big| \sim c\,\mathrm{vol}_V\big(B_R^V\big) \sim c' R^{d-2}$$

as $R \to \infty$, where vol_V denotes the G-invariant Haar measure on $V = \mathbb{V}(\mathbb{R})$.

As we will see, the constants above can be expressed using a, the volumes of the associated homogeneous spaces that arise in the proof, and the 'geometry' of Q.

[†] We will see throughout this section enough elements of $\mathbb{G}(\mathbb{R})$ to derive this transitivity directly.

We note that we define (and normalize) the Haar measure vol_V on $V(\mathbb{R})$ by the Lebesgue measure in $\mathbb{R}^d$ using the formula

$$\mathrm{vol}_V(B) = m_{\mathbb{R}^d}(\{tv \mid t \in [0,1], v \in B\}) \tag{5.20}$$

for any measurable $B \subseteq \mathbb{V}(\mathbb{R})$.

For the following proof we further define $\mathbb{G}$ and let $G = \mathbb{G}(\mathbb{R})$ or $G = \mathbb{G}(\mathbb{R})^o$ and $\Gamma = \mathrm{SO}_Q(\mathbb{Z}) \cap G$. We note that G^o is a simple real Lie group except in the case $p = q = 2$, in which case G^o is semisimple without compact factors instead.

5.5.2 Reduction to orbit counting problems

Recall from the beginning of the section that $V = \mathbb{V}(\mathbb{R})$ is a single $\mathbb{G}(\mathbb{R})$-orbit. We start with a fundamental finite decomposition result due to Borel and Harish-Chandra [9].

Proposition 5.27 (Borel–Harish-Chandra). *The integer points of $\mathbb{V}$ are a finite disjoint union*

$$\mathbb{V}(\mathbb{Z}) = \bigsqcup_i \mathbb{G}(\mathbb{Z})v_i \tag{5.21}$$

of $\mathbb{G}(\mathbb{Z})$-orbits.

Since the connected component $\mathbb{G}(\mathbb{R})^o$ has finite index in $\mathbb{G}(\mathbb{R})$, a version of (5.21) also holds for $\mathbb{G}(\mathbb{Z}) \cap G^o$ instead of $\mathbb{G}(\mathbb{Z})$.

Proof of Proposition 5.27. We let $G = \mathrm{SO}_Q(\mathbb{R})$ and $\Gamma = \mathrm{SO}_Q(\mathbb{Z})$, choose one $v_0 \in \mathbb{V}(\mathbb{Z})$, and set $H = \mathrm{Stab}_G(v_0)$. We associate to any point

$$v = g^{-1}\cdot v_0 \in \mathbb{V}(\mathbb{Z})$$

with $g \in G$ the orbit $Hg\cdot\Gamma \subseteq X$. We will show below that there exists some $\delta > 0$ such that any H-orbit associated to some $v \in \mathbb{V}(\mathbb{Z})$ intersects the compact set $\mathsf{X}_d(\delta)$. This implies the proposition. Indeed, as $X = G\cdot \mathrm{SL}_d(\mathbb{Z}) \subseteq \mathsf{X}_d$ is closed,

$$X \cap \mathsf{X}_d(\delta) = B\cdot\Gamma$$

is also compact and is the image of a compact set $B \subseteq G$. If now $hg\gamma \in B$ for $v = g^{-1}\cdot v_0 \in V(\mathbb{Z})$, $g \in G$, $h \in H$, and $\gamma \in \Gamma$, then $\gamma^{-1}\cdot v = \gamma^{-1}g^{-1}h^{-1}\cdot v_0$ belongs to the finite set $F = \mathbb{V}(\mathbb{Z}) \cap B^{-1}\cdot v_0$. Hence

$$\mathbb{V}(\mathbb{Z}) = \bigcup_{w \in F} \Gamma\cdot w$$

is a finite union of orbits as desired.

We define the inner product

$$\langle w, v\rangle_Q = \tfrac{1}{2}\big(Q(w+v) - Q(w) - Q(v)\big)$$

associated to Q and note that $\langle \mathbb{Z}^d, v\rangle_Q \subseteq \frac{1}{2}\mathbb{Z}$ is a subgroup containing a. Let us also write

$$v_Q^\perp = \{w \mid \langle w, v\rangle_Q = 0\}$$

for the orthogonal complement of $v \in \mathbb{R}^d$ with respect to $\langle\cdot,\cdot\rangle_Q$.

Index claim. Suppose therefore that $v = g^{-1}\cdot v_0 \in \mathbb{V}(\mathbb{Z})$, $g \in G$, and $\Lambda = g\mathbb{Z}^d$. Then $v_0 = g\cdot v \in \Lambda$ and the orthogonal complement defined using Q is Λ-rational, meaning that $v_0^\perp \cap \Lambda = g(v^\perp \cap \mathbb{Z}^d)$ spans $v_0^\perp$. We claim that

$$\mathbb{Z}v_0 + v_0^\perp \cap \Lambda$$

has index no more than $2|a|$ within Λ. By applying g^{-1} we may instead consider $\mathbb{Z}v + v^\perp \cap \mathbb{Z}^d < \mathbb{Z}^d$. Let $w \in \mathbb{Z}^d$ be chosen with $k = \langle w, v\rangle_Q \in \frac{1}{2}\mathbb{Z}$ minimal (and hence with $a \in \mathbb{Z}k$). Subtracting from a given element of $\mathbb{Z}^d$ a multiple of w allows us to reduce to an element in $v^\perp$, which shows that $\mathbb{Z}w + v^\perp \cap \mathbb{Z}^d = \mathbb{Z}^d$. Moreover

$$\langle \tfrac{a}{k}w - v, v\rangle_Q = \tfrac{a}{k}k - Q(v) = 0,$$

which shows that $\frac{a}{k}w \in \mathbb{Z}v + v^\perp \cap \mathbb{Z}^d$. As $\left|\frac{a}{k}\right| \leqslant 2a$, this proves the claim.

Reduction to complement. As our goal is to show a uniform lower bound for the norms of the non-zero vectors in $\Lambda = hg\mathbb{Z}^d$ for some $h \in H$ the index claim above is helpful. Indeed the vectors in $\Lambda \cap \mathbb{R}v_0$ cannot be close to zero as their image under Q belongs to $\mathbb{Z}\smallsetminus\{0\}$. Suppose now that we can find a uniform $\delta_0 \in (0,1]$ so that for any $v \in V(\mathbb{Z})$ there exists some $h \in H$ so that all non-zero vectors in $v_0^\perp \cap \Lambda = hg(v^\perp \cap \mathbb{Z}^d)$ have norm at least δ_0. Then all non-zero vectors in $hg(\mathbb{Z}v + v^\perp \cap \mathbb{Z}^d)$ have norm $\gg \delta_0$ and by the index claim the same applies to $hg\mathbb{Z}^d$ (where the implicit constant depends on the splitting $\mathbb{R}v_0 + v_0^\perp$ and on a).

Let $\Lambda_\perp = g(v^\perp \cap \mathbb{Z}^d)$. We consider $\Lambda_\perp$ as a lattice in $v_0^\perp \cong \mathbb{R}^{d-1}$ and equip $\mathbb{R}^{d-1}$ with the quadratic form $Q_\perp$ obtained by restricting Q to $v_0^\perp$.

A definite complement. Suppose now that Q restricted to $v_0^\perp$ is definite. In this case all non-zero vectors in $v_0^\perp \cap g\mathbb{Z}^d = g(v^\perp \cap \mathbb{Z}^d)$ have image under Q in $\mathbb{Z}\smallsetminus\{0\}$ and hence cannot be close to zero. This establishes the desired claim.

The three-dimensional case. Suppose now that $d = 3$ and the orthogonal complement is indefinite of signature $(1,1)$. In this case there exists a basis of $v_0^\perp$ with respect to which $Q_\perp$ is in the coordinates of this basis a multiple of the quadratic form x_1x_2. Moreover, $\mathrm{SO}_{Q_\perp}(\mathbb{R})$ is in this basis the diagonal subgroup. If $\Lambda_\perp$ has a primitive short vector $w_1 \in \Lambda$ then $Q_\perp(w_1) = 0$ (as it is small and belongs to $\mathbb{Z}$), w_1 belongs to one of the coordinate axes, and an element of H can be used to map w_1 to an element of norm one. Summarizing, either $g(v^\perp \cap \mathbb{Z}^d)$ already has no short vector or we can find $h \in H$ so that $hg(v^\perp \cap \mathbb{Z}^d)$ contains an element w_1 of norm one with $Q(w_1) = 0$. In the latter case $hg(v^\perp \cap \mathbb{Z}^d)$ contains no non-zero short vectors. Indeed, suppose otherwise and that $w_2 \in hg(v^\perp \cap \mathbb{Z}^d)$

is short. Then $Q(w_2) = 0$ as it belongs to $\mathbb{Z}$ and $\langle w_1, w_2\rangle_Q = 0$ as it belongs to $\frac{1}{2}\mathbb{Z}$. In the basis $v_0, w_1, w_2 \in \mathbb{R}^3$ the quadratic form Q therefore has the matrix representation

$$\begin{pmatrix} a & * & * \\ * & 0 & 0 \\ * & 0 & 0 \end{pmatrix} \tag{5.22}$$

which is a contradiction to Q being non-degenerate.

Higher dimensions. It remains to consider the case where $d \geqslant 4$ and Q restricted to $v_0^\perp$ is indefinite. In this case

$$H \cong \mathrm{SO}_{Q_\perp}(\mathbb{R}) < \mathrm{SL}_{d-1}(\mathbb{R})$$

is simple or semi-simple and generated by unipotent one-parameter subgroups, and the standard representation of H on $\mathbb{R}^{d-1}$ is irreducible. To see the latter, suppose $W < \mathbb{R}^{d-1}$ is a nontrivial subspace. If $w_0 \in W$ then $\mathfrak{h} = \operatorname{Lie} H$ satisfies $\mathfrak{h}w_0 = w_0^\perp \subseteq W$ (see Exercise 5.28). If $Q_\perp(w_0) \neq 0$, then $w_0^\perp$ is a hyperplane not containing w_0 and we obtain $W = \mathbb{R}^{d-1}$. If $Q_\perp(w_0) = 0$, then $w_0^\perp$ contains w_0 but also contains a vector w_1 with $Q_\perp(w_1) \neq 0$, leading to $W = \mathbb{R}^{d-1}$ once more. Indeed, if $Q_\perp$ were to vanish on $w_0^\perp$ then $\langle\cdot,\cdot\rangle_{Q_\perp}$ would also vanish on this hyperplane and we would again obtain a contradiction (similar to (5.22)) by considering the matrix representation of $Q_\perp$. We conclude that if $W \leqslant \mathbb{R}^{d-1}$ is non-trivial and invariant under H, then $W = \mathbb{R}^{d-1}$.

We claim that this implies that there exists some $h \in H^o$ so that

$$\lambda_1(h\Lambda_\perp) \asymp \cdots \asymp \lambda_{d-1}(h\Lambda_\perp).$$

To see this we rescale $\Lambda_\perp$ to an element $\Lambda_\perp^1 \in \mathsf{X}_{d-1}$ and apply quantitative non-divergence (Theorem 4.11) with $\eta = 1$ and find h within a suitable one-parameter unipotent subgroup $U < H^o$ (see below). It now follows that $h\Lambda_\perp$ has no short vector, for otherwise it would have a $\mathbb{Z}$-basis of short vectors and $Q_\perp = 0$ would again lead to a contradiction of Q being non-degenerate.

To find a choice of U for which we can indeed set $\eta = 1$ we start with a one-parameter unipotent subgroup $U_0 < H^o$ that is not contained in a proper normal subgroup of H^o. For a given non-trivial proper subspace V we consider the variety

$$\mathbb{S}_V = \big\{ g \in \mathbb{G} \;\big|\; V \text{ is invariant under } gU_0g^{-1} \big\}.$$

As $\mathbb{G}$ is irreducible, we see that either $\mathbb{S}_V = \mathbb{G}$ or all irreducible components of $\mathbb{S}_V$ have dimension strictly less than $\dim \mathbb{G}$. In the former case V would be invariant under the normal subgroup generated by gU_0g^{-1} for $g \in \mathbb{G}$. However, by our choice of U_0 and because $\mathbb{G}$ acts irreducibly we see that this is only possible for $V = \{0\}$ or for $V = \mathbb{R}^{d-1}$. Finally note that there are only finitely many non-trivial subspaces V with covolume less than 1. Picking $g \in G$ outside the union of the associated varieties we see that for $U = gU_0g^{-1}$ none of these

subspaces is invariant. Picking $T > 0$ large enough allows us to ensure that $\eta = 1$ satisfies the assumptions of Theorem 4.11. □

Exercise 5.28. Let Q be a non-degenerate quadratic form in d variables. Let $\mathfrak{g} = \operatorname{Lie} G$ for $G = \mathrm{SO}_Q(\mathbb{R})$. Show that $\mathfrak{g}v = v^\perp$ for any $v \in \mathbb{R}^d$, where $v^\perp$ is defined using the inner product $\langle \cdot, \cdot \rangle_Q$ associated to Q.

5.5.3 Finite volume assumptions

The standing assumptions in Section 5.4 were that $\Gamma < G$ and $\Gamma \cap H < H$ are both lattices. We now check these assumptions in the setting of Corollary 5.26.

Since $\mathbb{G} = \mathrm{SO}_Q$ is, for $d \geqslant 3$, a semisimple algebraic group defined over $\mathbb{Q}$ it follows by Theorem 4.18 that $\mathbb{G}(\mathbb{Z})$ is a lattice in $\mathbb{G}(\mathbb{R})$. This also implies that the subgroup $\Gamma = \mathbb{G}(\mathbb{Z}) \cap G$ is a lattice in $G = \mathrm{SO}_Q(\mathbb{R})^o$.

If $d \geqslant 4$ then $\mathbb{H} = \mathrm{Stab}_{\mathbb{G}}(v_0)$ is again a semisimple algebraic group defined over $\mathbb{Q}$ (since $a \neq 0$ it is simply the orthogonal group of the non-degenerate quadratic form on the orthogonal complement). Hence in that case $\mathbb{H}(\mathbb{Z})$ is a lattice in $\mathbb{H}(\mathbb{R})$ which once more implies that $\Gamma \cap H$ is a lattice in the group $H = \mathbb{H}(\mathbb{R}) \cap G$.

In the remaining case where $d = 3$, we see that $\mathbb{H}$ is either SO_2 or $\mathrm{SO}_{1,1}$. In the former case there is nothing to prove. If $0 \notin Q(\mathbb{Q}^3 \smallsetminus \{0\})$ then the same is true for the restriction to the orthogonal complement and we may apply Proposition 3.2.

So suppose now $Q = Q_0$ and that $a \in \mathbb{Z}$ is not a square. Then $Q_0|_{v_0^\perp}$ is a rational non-degenerate binary quadratic form. We claim that Proposition 3.2 applies to this restriction, which implies once more that $\mathbb{H}(\mathbb{Z})$ is a lattice in $\mathbb{H}(\mathbb{R})$. To see this we suppose for the purposes of a contradiction that $Q_0(w_1) = 0$ for some $w_1 \in \mathbb{Q}^3 \cap v_0^\perp$. Let $w_2 \in \mathbb{Q}^3 \cap v_0^\perp$ be linearly independent to w_1 so that $Q_0(xw_1 + yw_2) = 2bxy + cy^2$ for some $b, c \in \mathbb{Q}$. In the basis v_0, w_1, w_2 the quadratic form Q_0 has the matrix representation

$$\begin{pmatrix} a & & \\ & 0 & b \\ & b & c \end{pmatrix}$$

with determinant $-ab^2$. The determinant of the matrix representation of a quadratic form changes by the square of the determinant of a coordinate change matrix. In the standard basis the determinant is -1, which implies that $b \neq 0$ and that a must be a square, which contradicts our assumption.

5.5.4 Proving the Equidistribution

The main dynamical assumption in Section 5.4 (and Section 5.4.2 in particular) is the equidistribution of $gH\cdot\Gamma$ in $X = G/\Gamma$ as $gH \to \infty$ in G/H. We claim that this follows in the context of this section once again from the same 'mixing argument' that was used in Section 5.2. We will not repeat this argument in detail, but will provide the technical input that reduces this repetition of the mixing argument into a straightforward exercise.

What is needed in order to do this is an analogue of the local coordinate system $P_a^- G_a^+$ from Section 5.2.2 and the Iwasawa decomposition from Section 1.2.6. As this part of the argument only concerns $G = \mathrm{SO}_Q(\mathbb{R})^o$ (a real Lie group) and its subgroup H we suppose for simplicity that:

- $Q = Q_0$,
- $G = G_0 = \mathrm{SO}_{p,q}(\mathbb{R})^o$,
- $v_0 = e_1$ (by rescaling v_0, swapping p and q if necessary), and
- $H = H_0 = \mathrm{Stab}_G(e_1)$.

We start by defining a one-parameter diagonalizable subgroup

$$A_0 = \left\{ a_s = \begin{pmatrix} \cosh s & 0 & \sinh s & 0 \\ 0 & I_{p-1} & 0 & 0 \\ \sinh s & 0 & \cosh s & 0 \\ 0 & 0 & 0 & I_{q-1} \end{pmatrix} \;\middle|\; s \in \mathbb{R} \right\},$$

and the compact subgroup

$$K_0 = \big(\mathrm{SO}_p(\mathbb{R}) \times \mathrm{SO}_q(\mathbb{R})\big) \cap G_0.$$

The next lemma is not yet the analogous decomposition we are seeking, but is needed nonetheless.

Lemma 5.29 (A first group decomposition). *We have $G_0 = K_0 A_0 H_0$.*

Proof. Let $g \in G_0$ be an arbitrary element, and define

$$v = ge_1 = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

for $w_1 \in \mathbb{R}^p$ and $w_2 \in \mathbb{R}^q$. Then

$$Q_0(e_1) = Q_0(v) = \|w_1\|^2 - \|w_2\|^2 = 1.$$

If $p = 1$ the first coordinate of $v = gv_0$ must be positive because $G = \mathrm{SO}_{Q_0}(\mathbb{R})^o$. Hence, in any case there exists some $k \in K_0$ such that

$$kv = \|w_1\| e_1 \pm \|w_2\| e_{p+1}.$$

Let $s \in \mathbb{R}$ be chosen so that $\cosh s = \|w_1\|$ and $\sinh s = \pm\|w_2\|$. Then

$$kv = kgv_0 = a_s v_0,$$

equivalently $a_{-s}kg = h \in H_0$, or

$$g = k^{-1}a_s h \in K_0 A_0 H_0$$

as required. □

As K_0 is compact, the requirement that $g_n H_0 \to \infty$ in G_0/H_0 is equivalent to $g_n = k_n a_{s_n} h_n$ with $a_{s_n} \to \infty$ as $n \to \infty$. Furthermore, as in the proof of (5.15) on page 199, the sequence (k_n) has no effect on the desired equidistribution claim. Thus we can simply assume that $g_n = a_{s_n}$ with $s_n \to \infty$ or $s_n \to -\infty$ as $n \to \infty$. Below we will assume that $s_n \to \infty$ as $n \to \infty$ (the other case is similar). We now define the local coordinate system that is needed in the proof of the equidistribution statement.

Lemma 5.30 (A coordinate system). *The stable horospherical subgroup $G^-_{a_1}$ has the property that $G^-_{a_1}A_0H_0$ is open and that the product map provides a coordinate system on the image.*

Proof. For the proof it is convenient to switch to the Lie algebra. The Lie algebra element corresponding to A_0 is

$$h = \begin{pmatrix} 0_1 & 0 & 1 & 0 \\ 0 & 0_{p-1} & 0 & 0 \\ 1 & 0 & 0_1 & 0 \\ 0 & 0 & 0 & 0_{q-1} \end{pmatrix},$$

where (for example) the second 0 represents $(p-1)$ zeros in a row. We claim that the Lie algebra of G_{a_1} contains

$$W = \begin{pmatrix} 0_1 & -w_1^{\mathrm{t}} & 0 & w_2^{\mathrm{t}} \\ w_1 & 0_{p-1} & w_1 & 0 \\ 0 & w_1^{\mathrm{t}} & 0_1 & -w_2^{\mathrm{t}} \\ w_2 & 0 & w_2 & 0_{q-1} \end{pmatrix} \tag{5.23}$$

for all $w_1 \in \mathbb{R}^{p-1}$ and $w_2 \in \mathbb{R}^{q-1}$. This requires two calculations, as follows. Since $WJ + JW^{\mathrm{t}} = 0$ for the companion matrix

$$J = \begin{pmatrix} I_p & \\ & -I_q \end{pmatrix},$$

all elements W of the form (5.23) belong to the Lie algebra of G. Moreover, since

$$
\begin{aligned}
[h,W] &= hW - Wh \\
&= \begin{pmatrix} 0&0&1&0\\ 0&0&0&0\\ 1&0&0&0\\ 0&0&0&0 \end{pmatrix} \begin{pmatrix} 0 & -w_1^{\mathrm{t}} & 0 & w_2^{\mathrm{t}} \\ w_1 & 0 & w_1 & 0 \\ 0 & w_1^{\mathrm{t}} & 0 & -w_2^{\mathrm{t}} \\ w_2 & 0 & w_2 & 0 \end{pmatrix} - \begin{pmatrix} 0 & -w_1^{\mathrm{t}} & 0 & w_2^{\mathrm{t}} \\ w_1 & 0 & w_1 & 0 \\ 0 & w_1^{\mathrm{t}} & 0 & -w_2^{\mathrm{t}} \\ w_2 & 0 & w_2 & 0 \end{pmatrix} \begin{pmatrix} 0&0&1&0\\ 0&0&0&0\\ 1&0&0&0\\ 0&0&0&0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & w_1^{\mathrm{t}} & 0 & -w_2^{\mathrm{t}} \\ 0 & 0 & 0 & 0 \\ 0 & -w_1^{\mathrm{t}} & 0 & w_2^{\mathrm{t}} \\ 0 & 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0&0&0&0\\ w_1&0&w_1&0\\ 0&0&0&0\\ w_2&0&w_2&0 \end{pmatrix} \\
&= -W,
\end{aligned}
$$

it follows that W belongs to the Lie algebra of $G_{a_1}^-$ (where we have dropped the dimension subscripts for brevity).

We recall that $\dim \mathrm{SO}_d = \frac{d(d-1)}{2}$, which follows by induction and also holds for other signatures in dimension d. In particular,

$$\dim C_{G_0}(A_0) \geqslant \dim A_0 + \dim \mathrm{SO}_{p-1,q-1} = 1 + \frac{(d-2)(d-3)}{2}.$$

The elements in (5.23) give a subspace of dimension $(d-2)$, and it follows that $\dim G_{a_1}^- \geqslant d-2$. Flipping the sign of $v_0 = e_1$ belongs to $\mathrm{O}_{p,q}(\mathbb{R})$ and conjugation by it maps a_s to a_{-s} and $G_{a_1}^-$ to $G_{a_1}^+$, so that $\dim G_{a_1}^+ \geqslant d-2$. Taking the sums we have

$$1 + \frac{(d-1)(d-3)}{2} + 2(d-2) = \frac{2 + d^2 - 5d + 6 + 4d - 8}{2} = \frac{d(d-1)}{2}.$$

Hence our inequalities were equalities and the elements in (5.23) comprise the full Lie algebra of $G_{a_1}^-$.

If now $P = G_{a_1}^- A_0$ and $\mathfrak{p} = \mathrm{Lie}(P)$ is its Lie algebra, then we see from the inverse function theorem that Pe_1 must contain a neighbourhood of $e_1 \in V$, since $\mathfrak{p}e_1$ contains $\{0\} \times \mathbb{R}^{p+q-1}$ (which coincides with the tangent space at the point $e_1 \in V$).

It follows that if g is sufficiently close to the identity, then $ge_1 = ua_s e_1 \in Pe_1$ for some $ua_s \in P$, which gives $(ua_s)^{-1} g = h \in H_0$ and $g = ua_s h$ as required.

Suppose now that

$$u_1 a_{s_1} h_1 = u_2 a_{s_2} h_2$$

for $u_1, u_2 \in G_{a_1}^-$, $s_1, s_2 \in \mathbb{R}$, and $h_1, h_2 \in H_0$. Then

$$a_{s_2}^{-1} u_2^{-1} u_1 a_{s_1} = h_2 h_1^{-1} \in H_0$$

fixes e_1. Note that $\|a_t e_1\|^2 = \cosh 2t$ and that for $u = a_{s_2}^{-1} u_2^{-1} u_1 a_{s_2}$ we have $\|a_t u a_{-t}\| \to 1$ for $t \to \infty$. Using these facts after applying a_t to

$$ua_{s_1-s_2}e_1 = a_{s_2}^{-1}u_2^{-1}u_1a_{s_1}e_1 = e_1$$

we obtain $s_1 = s_2$. If now $u \neq I$ but $ue_1 = e_1$, then $u^n e_1 = e_1$ for all $n \in \mathbb{Z}$. However, u belongs to a one-parameter unipotent subgroup of G^-_{0,a_1} which than has to fix e_1 as well (since $\mathbb{Z}$ is Zariski dense in $\mathbb{R}$). Taking the derivative gives an element W in the Lie algebra of G^-_{0,a_1} with $We_1 = 0$, which contradicts (5.23). That the image is open now follows as in Lemma 5.13. □

We return to the general case of a non-degenerate indefinite quadratic form Q with signature (p, q) and integer coefficients, $a \in \mathbb{Z}\smallsetminus\{0\}$, $\mathbb{V} = \{v \mid Q(v) = 0\}$, and $v_0 \in \mathbb{V}(\mathbb{Z})$ as in Corollary 5.26. For the asymptotic counting of points in $\Gamma\cdot v_0$ we need the following equidistribution result.

Theorem 5.31 (Translated orbits). *Let $H = \mathrm{Stab}_G(v_0)$. Then the orbit $gH\cdot\Gamma$ equidistributes in the space $X = G/\Gamma$ as $gH \to \infty$ in G/H.*

Sketch of Proof. By our discussion in Section 3.1 and Theorem 3.5 in particular $G = \mathrm{SO}_Q(\mathbb{R})^o$ and $G_0 = \mathrm{SO}_{p,q}(\mathbb{R})^o$ are conjugate by a coordinate change in $\mathrm{SL}_d(\mathbb{R})$. We may moreover assume that $v_0 \in \mathbb{V}(\mathbb{Z})$ is mapped to a multiple of e_1 and hence H is conjugated to H_0. We define the subgroups $A, K < G$ by this conjugation using $A_0, K_0 < G_0$.

We may assume that $g = a_s$ with $s \to \infty$. We set

$$P^- = G^-_{a_1}A$$

and deduce from Lemma 5.30 and Lemma 1.58 that the Haar measure on m_G restricted to P^-H equals the direct product of the Haar measures on P^- and on H respectively.

Assume at first that $H\cdot\Gamma \subseteq X$ is compact. Then there exists some uniform injectivity radius $\delta > 0$ for all points in $H\cdot\Gamma$. Let $B = B_\delta^{P^-}$ be the corresponding neighbourhood of the identity in P^-, and set $T = BH\cdot\Gamma$, which we should think of as a *tubular neighbourhood* of $H\cdot\Gamma \subseteq X$ as in Figure 5.4.

Now pick some $f \in C_c(X)$, some $\varepsilon > 0$, ensuring that $\delta > 0$ is sufficiently small to work for f and ε, and apply mixing (Theorem 2.41) to obtain the desired contradiction for f, up to a precision controlled by ε.

If $H\cdot\Gamma$ is not compact, then the outline above needs to be adjusted (for otherwise, the failure of injectivity in a cusp makes the proof break down). Fortunately this case is not difficult either. Let $\kappa > 0$ be arbitrarily small and let $S \subseteq H\cdot\Gamma$ be a compact set of measure

$$\frac{1}{m_{H\cdot\Gamma}(H\cdot\Gamma)}m_{H\cdot\Gamma}(S) > 1 - \kappa.$$

Now apply the argument for the compact case above, with $BH\cdot\Gamma$ replaced by BS to obtain the desired conclusion up to a precision $\varepsilon + \|f\|_\infty\kappa$. □

Exercise 5.32. a) Upgrade the sketch of proof of Theorem 5.31 to a real proof.

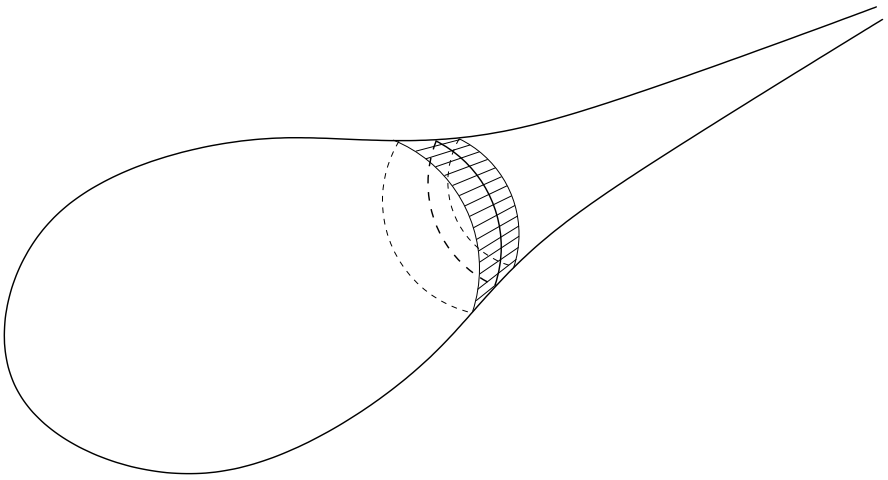


Fig. 5.4: The shaded region depicts the tubular neighbourhood T of the orbit $H\cdot\Gamma$ in the 'centre of T'.

b) The previous exercise notwithstanding, notice that the sketch proof applies a result without the right hypotheses. If $p = q = 2$, then G is not simple but only semisimple, and so the Howe–Moore theorem in the form of Theorem 2.41 cannot be applied. However, Theorem 2.44 does apply in this case. Decide whether or not this makes a difference to Theorem 5.31.

5.5.5 The Asymptotics of $\mathrm{vol}_V(B_R^V)$ and Well-Roundedness

Recall from (5.20) that we can define the Haar measure on V using the (G-invariant) Lebesgue measure on $\mathbb{R}^d$. By a linear coordinate change and by choosing the norm accordingly we may assume that $Q = Q_0$. Using this, we get the following result.

Lemma 5.33. $\mathrm{vol}_V\big(B_R^V\big) \sim \dfrac{\mathrm{vol}(\mathbb{S}^{p-1})\,\mathrm{vol}(\mathbb{S}^{q-1})}{d(d-2)}|a|^{\frac{d}{2}}\Big(\dfrac{R}{\sqrt{2}}\Big)^{d-2}$ *as* $R \to \infty$.

Proof. We will assume that $a = 1$ (which by a scaling argument allows the case $a > 0$ to be deduced; the case $a < 0$ then follows by swapping p and q) and that $R > 1$. Choose $S > 0$ with $R^2 = \cosh 2S$ so that $\cosh^2 S + \sinh^2 S = R^2$. Let $W_+ \subseteq \mathbb{R}^{p-1}$ be open and Jordan measurable and let $\phi_+ \colon W_+ \to \mathbb{S}^{p-1}$ be a smooth parameterization[†] of $\mathbb{S}^{p-1}$ up to a set of measure 0. Similarly, let $W_- \subseteq \mathbb{R}^{q-1}$ and $\phi_- \colon W_- \to \mathbb{S}^{q-1}$ be a smooth paramaterization of $\mathbb{S}^{q-1}$. Therefore the G-invariant volume $\mathrm{vol}_V(B_R^V)$ is given by

[†] For example, using generalized spherical coordinates.

$$
\begin{aligned}
&m_{\mathbb{R}^d}\left(\left\{ t\begin{pmatrix}\cosh(s)\phi_+(w_+)\\ \sinh(s)\phi_-(w_-)\end{pmatrix} \;\middle|\; t\in[0,1], s\in[0,S], w_+\in W_+, w_-\in W_-\right\}\right)\\
&\quad=\int_0^1\int_0^S\int_{W_+}\int_{W_-}\Bigg|\det\Bigg(\begin{pmatrix}\cosh(s)\phi_+(w_+)\\ \sinh(s)\phi_-(w_-)\end{pmatrix},\begin{pmatrix}t\sinh(s)\phi_+(w_+)\\ t\cosh(s)\phi_-(w_-)\end{pmatrix},\\
&\qquad\qquad\begin{pmatrix}t\cosh(s)\,\mathrm{D}_{w_+}\,\phi_+(w_+)\\ 0\end{pmatrix},\begin{pmatrix}0\\ t\sinh(s)\,\mathrm{D}_{w_-}\,\phi_-(w_-)\end{pmatrix}\Bigg)\Bigg|\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\mathrm{d}w_-\,\mathrm{d}w_+\,\mathrm{d}s\,\mathrm{d}t\\
&\quad=\int_0^1\int_0^S\int_{W_+}\int_{W_-}t^{p+q-1}\cosh(s)^{p-1}\sinh(s)^{q-1}\cdot\\
&\qquad\Bigg|\det\Bigg(\begin{pmatrix}\cosh(s)\phi_+(w_+)\\ \sinh(s)\phi_-(w_-)\end{pmatrix},\begin{pmatrix}\sinh(s)\phi_+(w_+)\\ \cosh(s)\phi_-(w_-)\end{pmatrix},\\
&\qquad\qquad\begin{pmatrix}\mathrm{D}_{w_+}\,\phi_+(w_+)\\ 0\end{pmatrix},\begin{pmatrix}0\\ \mathrm{D}_{w_-}\,\phi_-(w_-)\end{pmatrix}\Bigg)\Bigg|\,\mathrm{d}w_-\,\mathrm{d}w_+\,\mathrm{d}s\,\mathrm{d}t,
\end{aligned}
$$

where $\mathrm{D}_{w_+}\,\phi_+$ and $\mathrm{D}_{w_-}\,\phi_-$ are the total derivatives of ϕ_+ and of ϕ_- respectively. Using multilinearity and anti-symmetry in the first two vectors allows us to split the determinant above into three factors. This gives for $\mathrm{vol}_V(B_R^V)$ that it is equal to $\frac{1}{d}=\int_0^1 t^{d-1}\,\mathrm{d}t$ multiplied by

$$
\begin{aligned}
&\int_0^S\int_{W_+}\int_{W_-}(\cosh s)^{p-1}(\sinh s)^{q-1}\det\begin{pmatrix}\cosh s & \sinh s\\ \sinh s & \cosh s\end{pmatrix}\\
&\qquad\cdot\left|\det\left(\begin{pmatrix}\phi_+(w_+)\\ 0\end{pmatrix},\begin{pmatrix}0\\ \phi_-(w_-)\end{pmatrix},\begin{pmatrix}\mathrm{D}_{w_+}\phi_+(w_+)\\ 0\end{pmatrix},\begin{pmatrix}0\\ \mathrm{D}_{w_-}\phi_-(w_-)\end{pmatrix}\right)\right|\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\mathrm{d}w_+\,\mathrm{d}w_-\,\mathrm{d}s\\
&\quad=\int_{W_+}\Big|\det\Big(\phi_+(w_+),\mathrm{D}_{w_+}\,\phi_+(w_+)\Big)\Big|\mathrm{d}w_+\\
&\qquad\qquad\cdot\int_{W_-}\Big|\det\Big(\phi_-(w_-),\mathrm{D}_{w_-}\,\phi_-(w_-)\Big)\Big|\mathrm{d}w_-\cdot\int_0^S\cosh(s)^{p-1}\sinh(s)^{q-1}\,\mathrm{d}s\\
&\quad=\mathrm{vol}\big(\mathbb{S}^{p-1}\big)\,\mathrm{vol}\big(\mathbb{S}^{q-1}\big)\int_0^S\left(\frac{1}{2^{p+q-2}}\mathrm{e}^{s(p+q-2)}+\mathrm{O}(\mathrm{e}^{s(p+q-4)})\right)\mathrm{d}s\\
&\quad=\mathrm{vol}\big(\mathbb{S}^{p-1}\big)\,\mathrm{vol}\big(\mathbb{S}^{q-1}\big)\frac{1}{2^{d-2}}\frac{1}{d-2}\mathrm{e}^{(d-2)S}+\mathrm{O}\left(\mathrm{e}^{(d-4)S}\right).
\end{aligned}
$$

Recall that $R^2=\cosh 2S\sim\frac{1}{2}\mathrm{e}^{2S}$, so that $\mathrm{e}^S\sim\sqrt{2}R$. This gives

$$\operatorname{vol}_V\left(B_R^V\right) \sim \frac{\operatorname{vol}\left(\mathbb{S}^{p-1}\right)\operatorname{vol}\left(\mathbb{S}^{q-1}\right)}{d(d-2)}\left(\frac{R}{\sqrt{2}}\right)^{d-2}.$$

□

Lemma 5.34. *The sets* $B_t = B_{\mathrm{e}^t}^V = \left\{v \in V \;\middle|\; \|v\| \leqslant R = \mathrm{e}^t\right\}$ *for* $t \geqslant 0$ *are well-rounded in the sense of Definition* 5.21.

Proof. Let $\delta > 0$ and choose a neighbourhood U of $I \in G$ such that

$$\max\left(\|g\|, \|g^{-1}\|\right) < \mathrm{e}^{\delta}$$

for all $g \in G$. Let $v \in B_{t-\delta}$, so that $\|v\| \leqslant \mathrm{e}^{t-\delta}$, $\|g^{-1}v\| < \mathrm{e}^t$, and so $v \in gB_t$ for $g \in U$. This gives

$$B_{t-\delta} \subseteq \bigcap_{g\in V} gB_t.$$

Similarly we see that

$$\bigcup_{g\in U} gB_t \subseteq B_{t+\delta}.$$

Finally, we have

$$\frac{\operatorname{vol}_V(B_{t+\delta})}{\operatorname{vol}_V(B_t)} \sim \frac{\mathrm{e}^{(d-2)(t+\delta)}}{\mathrm{e}^{(d-2)t}} = \mathrm{e}^{(d-2)\delta} < 1+\varepsilon$$

for small enough δ. Therefore

$$\frac{\operatorname{vol}_V(B_{t+\delta})}{\operatorname{vol}_V(B_t)} < 1+\varepsilon \tag{5.24}$$

for all $t > T$. From the proof of Lemma 5.33, we also see that $\operatorname{vol}_V(B_t)$ depends continuously on t, so we can make δ even smaller if necessary to ensure that (5.24) also holds for all $t \in [0, T]$. □

5.5.6 Conclusion

Proof of Corollary 5.26. Let Q, $d = p+q$, $a \in \mathbb{Z}\smallsetminus\{0\}$, and

$$\mathbb{V} = \{v \mid Q(v) = 0\}$$

be as in the statement of the corollary. Let $G = \mathrm{SO}_Q(\mathbb{R})^o$ and $\Gamma = \mathrm{SO}_Q(\mathbb{Z}) \cap G$. By Proposition 5.27 we know that $\mathbb{V}(\mathbb{Z})$ is a finite union of disjoint Γ-orbits, say

$$\mathbb{V}(\mathbb{Z}) = \bigsqcup_{j=1}^{J} \Gamma \cdot v_j. \tag{5.25}$$

We use a linear coordinate change in $\mathrm{SL}_d(\mathbb{R})$ to map Q to a multiple of Q_0 and a fixed vector $v_0 \in \mathbb{V}(\mathbb{R})$ to a multiple of e_1. We also use this to define the balls B_t as in Section 5.5.5. In particular we have by Lemma 5.34 that the family of sets B_t for $t \geqslant 0$ are well-rounded.

Now fix some $j \in \{1, \dots, J\}$ and define $H_j = \mathrm{Stab}_G(v_j)$. The discussion in Section 5.5.3 shows that $\Gamma \cap H_j$ is a lattice in H_j. By our discussion in Section 5.5.4 and Theorem 5.31 in particular the translated H-orbits $gH \cdot \Gamma$ equidistribute in $X = G/\Gamma$ as $gH_j \to \infty$ in G/H_j.

This allows us to apply Theorem 5.22. Assuming that the Haar measures on H and G are chosen to be compatible with the Haar measure m_V on $V = \mathbb{V}(\mathbb{R})$, it follows that

$$\lim_{t\to\infty} \frac{1}{m_V(B_t)} \big|\Gamma \cdot v_j \cap B_t\big| = \frac{m_{H_j/\Gamma\cap H_j}(H_j/\Gamma \cap H_j)}{m_{G/\Gamma}(G/\Gamma)}.$$

Taking the union in (5.25), this implies that

$$\lim_{t\to\infty} \frac{1}{m_V(B_t)} \big|\mathbb{V}(\mathbb{Z}) \cap B_t\big| = \sum_{j=1}^{J} \frac{m_{H_j/\Gamma\cap H_j}(H_j/\Gamma \cap H_j)}{m_{G/\Gamma}(G/\Gamma)}.$$

Together with Lemma 5.33 this gives the corollary. □

5.6 Counting Integer Matrices with Given Determinant

In this section we want to apply the results of Section 5.4 to prove the following corollary.

Corollary 5.35. *Let $d \in \{2, 3\}$ and $a \in \mathbb{Z} \smallsetminus \{0\}$. Then there exists a positive constant c_a such that*

$$\big|\{M \in \mathrm{Mat}_d(\mathbb{Z}) \mid \det M = a \text{ and } \|M\| \leqslant R\}\big| \sim c_a R^{d(d-1)}.$$

We note that this implies in particular the asymptotic counting result for the lattice elements of $\mathrm{SL}_d(\mathbb{Z})$.

5.6.1 Reduction to orbit counting problems

We note that the asymptotic counting problem in Corollary 5.35 for a and $-a$ are equivalent (by simply changing the sign of one column). So let us assume

that $a > 0$. We define the set

$$V(\mathbb{R}) = \left\{ M \in \mathrm{Mat}_d(\mathbb{R}) \mid \det(M) = a \right\} \cong \mathrm{SL}_d(\mathbb{R}) \times \mathrm{SL}_d(\mathbb{R}) / \Delta_{\mathrm{SL}_d(\mathbb{R})},$$

where $\Delta_{\mathrm{SL}_d(\mathbb{R})} = \{(g, g) \mid g \in \mathrm{SL}_d(\mathbb{R}\}$. In fact $G = \mathrm{SL}_d(\mathbb{R}) \times \mathrm{SL}_d(\mathbb{R})$ acts on matrices $M \in V$ via

$$(g_1, g_2) \cdot M = g_1 M g_2^{-1}.$$

Then, if $M_0 = \sqrt[d]{a} I$,

$$\mathrm{Stab}_G(M_0) = \Delta_{\mathrm{SL}_d(\mathbb{R})}$$

and transitivity is easy.

Next we show that

$$V(\mathbb{Z}) = \{ M \in \mathrm{Mat}_d(\mathbb{Z}) \mid \det M = a \}$$

is a finite union of $\Gamma = \mathrm{SL}_d(\mathbb{Z}) \times \mathrm{SL}_d(\mathbb{Z})$-orbits. Let $M \in V(\mathbb{Z})$ be arbitrary. Applying elements of Γ to M correspond to certain types of row and column operations on M. Indeed using the elementary unipotent matrices of $\mathrm{SL}_d(\mathbb{Z})$ we can add any multiple of a row (or column) to any other row (or column). Similarly we may permute rows and columns (potentially switching the sign of one of them).

These steps allow a type of Euclidean algorithm: Assuming that the top left corner is already the smallest non-zero entry of absolute value we may either reduce the remaining entries on the first row and column to zero or produce a smaller entry. Hence eventually we create a block matrix with a non-zero entry in the top left corner, zeroes on the remainder of the first row and column, and some matrix in the lower right block. We may repeat this procedure and arrive at a diagonal integer matrix. As there are only finitely many integer diagonal matrices with determinant equal to a the result follows.

As $V(\mathbb{Z})$ is a finite union of Γ-orbits it suffices to establish the counting result for each individual Γ-orbit.

5.6.2 Finite volume assumptions

We now check the standing assumptions of Section 5.4 that both homogeneous spaces appearing there have finite volume.

By Theorem 1.54 $\Gamma = \mathrm{SL}_d(\mathbb{Z}) \times \mathrm{SL}_d(\mathbb{Z})$ is a lattice in $G = \mathrm{SL}_d(\mathbb{R}) \times \mathrm{SL}_d(\mathbb{R})$.

By the argument in the previous section it suffices to study the counting problem for $\Gamma M = \mathrm{SL}_d(\mathbb{Z}) M \, \mathrm{SL}_d(\mathbb{Z})$ where $M \in V(\mathbb{Z})$ is a diagonal matrix. This defines our second group

$$H = \mathrm{Stab}_G(M) = \{(g_1, g_2) \mid g_1, g_2 \in \mathrm{SL}_d(\mathbb{R}) \text{ and } g_2 = M^{-1} g_1 M\},$$

which clearly is isomorphic to $\mathrm{SL}_d(\mathbb{R})$. In this isomorphism $\Gamma \cap H$ corresponds to $\{g \in \mathrm{SL}_d(\mathbb{Z}) \mid M^{-1} g M \in \mathrm{SL}_d(\mathbb{Z})\}$. The latter is a finite index subgroup

of $\mathrm{SL}_d(\mathbb{Z})$ (one way to see this is to note that it contains the congruence subgroup $\{g \in \mathrm{SL}_d(\mathbb{Z}) \mid a \text{ divides } g - I\}$), which implies that $H/(\Gamma \cap H)$ has finite volume as required.

5.6.3 Proving the Equidistribution

The main dynamical assumption in Section 5.4 (see Section 5.4.2) is the equidistribution of $gH\Gamma$ in $X = G/\Gamma$ as gH goes to infinity in G/H. The argument is similar to that used in Section 5.2. We will not repeat the 'mixing argument', but will instead discuss the technical requirements that make it work.

For these preparations we set $H_0 = \Delta_{\mathrm{SL}_2(\mathbb{R})}$.

Lemma 5.36 (A first group decomposition). *Let $A \leqslant \mathrm{SL}_d(\mathbb{R})$ denote the full positive diagonal subgroup, and define*

$$\widetilde{A} = \{(a, a^{-1}) \mid a \in A\},$$

$$K = \mathrm{SO}_d(\mathbb{R}) \times \mathrm{SO}_d(\mathbb{R}).$$

Then $G = K\widetilde{A}H_0$. Moreover, we have $G = K\widetilde{A}^+H_0$, where $\widetilde{A}^+$ consists of all $(a, a^{-1}) \in \widetilde{A}$ with the diagonal entries of a monotonically increasing.

Proof. Multiplying $(g_1, g_2) \in G$ on the right by $(g_2^{-1}, g_2^{-1}) \in H_0$ we see that $(g_1, g_2)H = (g, I)H$ for $g = g_1 g_2^{-1} \in \mathrm{SL}_d(\mathbb{R})$. Let $g = k_1 a k_2$ be a KAK decomposition of g, and let $a_1 \in A$ be a square root of a. Then

$$(g_1, g_2) \in (g, I)H_0 = (k_1 a_1^2 k_2^{-1}, I)H_0 = (k_1, k_2)(a_1, a_1^{-1})H_0 \subseteq K\widetilde{A}H_0$$

as required. Finally recall that in the KAK decomposition of g we may assume that the diagonal entries of the element of A are monotonically increasing. □

If we now consider a sequence $(g_n H_0)$ going to infinity in G/H_0, then it is clear that we may replace g_n by $k_n\widetilde{a_n} \in K\widetilde{A}^+$. As K is compact, we may suppress the elements $k_n \in K$ by using compactness and considering them as part of the function f (as in the proof of (5.15) on page 199) and consider simply the case

$$\widetilde{a_n}H_0 \longrightarrow \infty$$

as $n \to \infty$ in G/H_0, with $\widetilde{a_n} \in \widetilde{A}$.

Lemma 5.37 (A coordinate system). *Let $N \leqslant \mathrm{SL}_d(\mathbb{R})$ be the upper-triangular unipotent subgroup, and let*

$$\widetilde{N} = \big\{(n_1, n_2^{\mathrm{t}}) \mid n_1, n_2 \in N\big\}.$$

Then $\widetilde{N}\widetilde{A}H_0$ is open and the product map provides a coordinate system.

PROOF. We have

$$\begin{aligned}\mathrm{Lie}(H_0) &= \{(v,v) \mid v \in \mathfrak{sl}_d(\mathbb{R})\},\\ \mathrm{Lie}(\widetilde{A}) &= \{(h,-h) \mid h \text{ diagonal}, \mathrm{tr}(h)=0\},\end{aligned}$$

and $\mathrm{Lie}(\widetilde{N})$ is the direct product of the upper and lower nilpotent triangular Lie subalgebras of $\mathfrak{sl}_d(\mathbb{R})$. It is easy to see that these subspaces are transversal, and their dimension sums to the dimension of the Lie algebra of G. The lemma follows by the inverse function theorem and the fact that $(\widetilde{N}\widetilde{A}) \cap H_0 = \{I\}$; see also the proof of Lemma 5.13. □

Theorem 5.38. *$gH_\bullet\Gamma$ equidistributes in $X = G/\Gamma$ as $gH \to \infty$ in G/H.*

SKETCH OF PROOF. Here $H = \mathrm{Stab}_G(M)$ for a matrix $M \in V(\mathbb{Z})$. As we may assume that $a > 0$, we see that $a^{\frac{1}{d}}M \in \mathrm{SL}_d(\mathbb{R})$ which implies that H is conjugate to $H_0 = \Delta_{\mathrm{SL}_d(\mathbb{R})}$ via the element

$$\widetilde{g} = (I, a^{-\frac{1}{d}}M).$$

It follows that

$$H\Gamma = \widetilde{g}^{-1}H_0\widetilde{g}\Gamma,$$

and it is enough to show that $gH_0\widetilde{g}\Gamma$ equidistributes as $gH_0 \to \infty$ in G/H_0. By Lemma 5.37 we may safely assume that $g = k\widetilde{a}$ with $\widetilde{a} \in \widetilde{A}$, and even that $g = \widetilde{a} \in \widetilde{A}^+$.

Let $\varepsilon > 0$ be arbitrarily small, and choose a compact subset $S \subseteq H_0\widetilde{g}\Gamma$ such that

$$m_{H_0\widetilde{g}\Gamma}(S) > (1-\varepsilon)m_{H_0\widetilde{g}\Gamma}(H_0\widetilde{g}\Gamma).$$

Fix $f \in C_c(X)$ and choose $\delta > 0$ smaller than the injectivity radius on S and small enough to ensure that

$$\mathsf{d}(x_1,x_2) < \delta \implies |f(x_1) - f(x_2)| < \varepsilon.$$

Set $P = \widetilde{N}\widetilde{A}$ and let $B = B_\delta^P$ be the δ-neighbourhood of $I \in P$. Now replace $H_0\widetilde{g}\Gamma$ first by S and then by BS, use the mixing property (Theorem 2.44), use the fact that $g_n = \widetilde{a_n} \in \widetilde{A}^+$ does not expand N, and deduce the proof of the theorem. □

5.6.4 The Asymptotics of $\mathrm{vol}_V(B_R^V)$ and Well-Roundedness

Clearly for any $a > 0$ there is a bijection

$$V = \{M \in \mathrm{Mat}_d(\mathbb{R}) \mid \det M = a\} \longleftrightarrow \mathrm{SL}_d(\mathbb{R}),$$

obtained by multiplying by $a^{-\frac{1}{d}}$. Thus it is sufficient to study the volume of 'balls' in $\mathrm{SL}_d(\mathbb{R})$.

Proposition 5.39 (Asymptotics of balls in $\mathrm{SL}_d(\mathbb{R})$). *The asymptotic growth in the volume of balls in $\mathrm{SL}_d(\mathbb{R})$ for $d \in \{2,3\}$ has the form*

$$m_{\mathrm{SL}_d(\mathbb{R})}\big(\{g \in \mathrm{SL}_d(\mathbb{R}) \mid \|g\| < R\}\big) \sim c_d R^{d(d-1)}$$

for some $c_d > 0$. Moreover, the sets $B_t = \{g \in \mathrm{SL}_d(\mathbb{R}) \mid \|g\| < \mathrm{e}^t\}$ are well-rounded.

We normalize the Haar measure on $\mathrm{SL}_d(\mathbb{R})$ by giving the definition

$$m_{\mathrm{SL}_d(\mathbb{R})}(B) = m_{\mathbb{R}^{d^2}}(\{tb \mid t \in [0,1], b \in B\})$$

for any measurable $B \subseteq \mathrm{SL}_d(\mathbb{R})$.

Proof of Proposition 5.39 for $d = 2$. We note that this case follows quickly from hyperbolic geometry and the connection between the Haar measures on $\mathrm{SL}_2(\mathbb{R})$ and $\mathbb{H}$. However, as the proof of Proposition 5.39 for $d = 3$ below uses a more complicated version of the following calculation, we give an independent argument also for $d = 2$.

We define $B_R = \{g \in \mathrm{SL}_2(\mathbb{R}) \mid \|g\| \leqslant R\}$. We are going to use the KAK decomposition for

$$g = k_\phi \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k_\psi \in \mathrm{SL}_2(\mathbb{R}) \tag{5.26}$$

with $r \geqslant 1$, $\phi \in [0, 2\pi)$, and $\psi \in [0, \pi)$. Using the definition of the Haar measure on $\mathrm{SL}_2(\mathbb{R})$ via the Lebesgue measure on $\mathrm{Mat}_2(\mathbb{R})$ we have

$$\begin{aligned} m_{\mathrm{SL}_2(\mathbb{R})}(B_R) &= m_{\mathbb{R}^4}(\{tg \mid g \in B_R, t \in [0,1]\}) \\ &= \int_0^1 \int_1^{R_0} \int_0^{2\pi} \int_0^{\pi} \Bigg| \det\Bigg(k_\phi \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k_\psi, tk_\phi \begin{pmatrix} 1 & \\ & -r^{-2} \end{pmatrix} k_\psi, \\ &\qquad tk_\phi \begin{pmatrix} & 1 \\ -1 & \end{pmatrix} \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} k_\psi, tk_\phi \begin{pmatrix} r & \\ & r^{-1} \end{pmatrix} \begin{pmatrix} & 1 \\ -1 & \end{pmatrix} k_\psi \Bigg) \Bigg| \\ &\qquad\qquad \mathrm{d}\psi\, \mathrm{d}\phi\, \mathrm{d}r\, \mathrm{d}t. \end{aligned}$$

Here the parameter $R_0 \geqslant 1$ is chosen so that $\sqrt{R_0^2 + R_0^{-2}} = R$, and the 2×2 matrices in the determinant above are the partial derivatives of the parameterization in (5.26), and these should be converted into ordinary 4-dimensional vectors before the determinant is taken. This calculation leads to

$$m_{\mathrm{SL}_2(\mathbb{R})}(B_R) = \int_0^1 t^3\, \mathrm{d}t \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \mathrm{d}\psi \int_1^{R_0} \left| \det \begin{pmatrix} r & 1 \\ r^{-1} & -r^{-2} \end{pmatrix} \det \begin{pmatrix} -r & -r^{-1} \\ r^{-1} & r \end{pmatrix} \right| \mathrm{d}r$$

by taking the factor t out of the determinant, noticing that the matrices k_ϕ, k_ψ on the left (respectively, right) appear in each matrix and do not affect the total determinant, and by splitting the resulting determinant into the determinant of the diagonal (respectively, the determinant of the off-diagonal) entries. The first three integrals define a constant c_2 and we obtain the asymptotics

$$\begin{aligned} m_{\mathrm{SL}_2(\mathbb{R})}(B_R) &\sim c_2 \int_1^{R_0} 2r^{-1}\big(r^2 - r^{-2}\big)\,\mathrm{d}r \\ &\sim c_2\big(R_0^2 - 1\big) - c_2\big(R_0^{-2} - 1\big) \sim c_2 R_0^2. \end{aligned}$$

The well-roundedness now follows in the same way as in Lemma 5.34. □

Proof of Proposition 5.39 for $d = 3$. We set $K = \mathrm{SO}_3(\mathbb{R})$ and let

$$W \ni \phi \longmapsto k_\phi \in K$$

be a piecewise smooth parameterization of a conull set in K on a Jordan measurable subset $W \subseteq \mathbb{R}^3$ and write

$$a_{r_1,r_2} = \begin{pmatrix} r_1 & & \\ & r_2 & \\ & & \frac{1}{r_1 r_2} \end{pmatrix}$$

for $r_1, r_2 > 0$. Let us write $\approx$ for asymptotic up to a positive multiple. As in the case $d = 2$, we have

$$m_{\mathrm{SL}_3(\mathbb{R})}\left(B_R\right) = m_{\mathbb{R}^9}\left(\{tg \mid g \in B_R, t \in [0,1]\}\right)$$

and we can calculate the latter by the Cartan decomposition and substitution. The integral with respect to the variable $t \in [0,1]$ produces the factor $\frac{1}{9}$ and hence we are interested in the remaining eight-dimensional integral

$$\begin{aligned} \int_{S_R}\int_W\int_W \det\Big(&k_\phi a_{r_1,r_2} k_\psi^{\mathrm{t}}, k_\phi \partial_{r_1} a_{r_1,r_2} k_\psi^{\mathrm{t}}, k_\phi \partial_{r_2} a_{r_1,r_2} k_\psi^{\mathrm{t}}, \\ &\big(D_\phi k_\phi\big) a_{r_1,r_2} k_\psi^{\mathrm{t}}, k_\phi a_{r_1,r_2}\big(D_\psi k_\psi^{\mathrm{t}}\big)\Big)\,\mathrm{d}\psi\,\mathrm{d}\phi\,\mathrm{d}(r_1,r_2), \end{aligned}$$

where

$$S_R = \left\{(r_1,r_2) \in \mathbb{R}^2 \;\middle|\; r_1 \geqslant r_2 \geqslant \tfrac{1}{r_1 r_2} > 0 \text{ and } \sqrt{r_1^2 + r_2^2 + \tfrac{1}{r_1^2 r_2^2}} \leqslant R\right\}.$$

Just as in the case $d = 2$, the main interest arises from the integration over $(r_1, r_2) \in S_R$. There are, however, some differences between the two cases. Firstly, the domain S_R is more complicated, and for part of the calculation we will slightly simplify this domain by using a different set $\widetilde{S_R}$. Secondly, the

paramaterization of K does not have a constant Jacobian.[†] As we do not care about scale factors like $m_K(K)$ we instead concentrate on the distortion of measure in moving from $K \times S_R \times K$ to $\mathbb{R}^9$. We do this by working with the left-invariant vector fields defined by

$$\mathbf{b}_1 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \mathbf{b}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$$

for $\phi \in W$ instead of the partial derivatives with respect to ϕ_1, ϕ_2, ϕ_3 (and similarly ψ_1, ψ_2, ψ_3). This makes the determinant function invariant under ϕ (respectively, under ψ) since each of the matrices in the determinant again has k_ϕ on the left and k_ψ^{t} on the right. This gives

$$m_{\mathrm{SL}_3(\mathbb{R})}(B_R) \approx \int_{S_R} \big|\det\big(a_{r_1,r_2}, \partial_{r_1} a_{r_1,r_2}, \partial_{r_2} a_{r_1,r_2}, \mathbf{b}_1 a_{r_1,r_2}, \mathbf{b}_2 a_{r_1,r_2}, \mathbf{b}_3 a_{r_1,r_2}, a_{r_1,r_2}\mathbf{b}_1, a_{r_1,r_2}\mathbf{b}_2, a_{r_1,r_2}\mathbf{b}_3\big)\big| \, \mathrm{d}r_1 \, \mathrm{d}r_2.$$

As before, each matrix a_{r_1,r_2} and so on should be thought of as a 9-dimensional vector, so that we can take the determinant of the resulting 9×9 matrix. The matrix has block form, with

$$a_{r_1,r_2},\ \partial_{r_1} a_{r_1,r_2} = \begin{pmatrix} 1 & & \\ & 0 & \\ & & -\frac{1}{r_1^2 r_2} \end{pmatrix},\ \partial_{r_2} a_{r_1,r_2} = \begin{pmatrix} 0 & & \\ & 1 & \\ & & -\frac{1}{r_1 r_2^2} \end{pmatrix}$$

forming one 3×3 block, and

$$\mathbf{b}_1 a_{r_1,r_2} = \begin{pmatrix} 0 & r_2 & 0 \\ -r_1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ a_{r_1,r_2}\mathbf{b}_1 = \begin{pmatrix} 0 & r_1 & 0 \\ -r_2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

respectively

$$\mathbf{b}_2 a_{r_1,r_2} = \begin{pmatrix} 0 & 0 & \frac{1}{r_1 r_2} \\ 0 & 0 & 0 \\ -r_1 & 0 & 0 \end{pmatrix},\ a_{r_1,r_2}\mathbf{b}_2 = \begin{pmatrix} 0 & 0 & r_1 \\ 0 & 0 & 0 \\ -\frac{1}{r_1 r_2} & 0 & 0 \end{pmatrix},$$

respectively

$$\mathbf{b}_3 a_{r_1,r_2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1}{r_1 r_2} \\ 0 & -r_2 & 0 \end{pmatrix},\ a_{r_1,r_2}\mathbf{b}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & r_2 \\ 0 & -\frac{1}{r_1 r_2} & 0 \end{pmatrix}$$

[†] For a convenient reference for the change of variables formula involved here we refer to Lee [99].

the other three 2×2 blocks. This allows us to calculate the determinant without too much pain, giving

$$m_{\mathrm{SL}_3(\mathbb{R})}(B_R) \approx \int_{S_R} \left(\frac{2}{r_1 r_2}\right)\left(r_1^2 - r_2^2\right)\Bigg(\underbrace{r_1^2 - \frac{1}{r_1^2 r_2^2}}_{\sim r_1^2}\Bigg)\Bigg(\underbrace{r_2^2 - \frac{1}{r_1^2 r_2^2}}_{(*)}\Bigg)\,\mathrm{d}r_1\,\mathrm{d}r_2.$$

Notice that $\frac{1}{r_1 r_2} \to 0$ as $R \to \infty$ and $r_1 \to \infty$, so we may replace $r_1^2 - \frac{1}{r_1^2 r_2^2}$ by r_1^2 without affecting the asymptotic behaviour. This argument does not apply in the same way to the last term $(*)$.

As mentioned earlier, the original domain of integration is difficult to work with, so we instead consider the set

$$\widetilde{S_R} = \left\{(r_1, r_2) \in \mathbb{R}^2 \;\middle|\; r_1 \geqslant r_2 \geqslant \tfrac{1}{r_1 r_2} > 0 \text{ and } \sqrt{r_1^2 + r_2^2} \leqslant R\right\} \supseteq S_R$$

(which is only slightly less annoying). For any $\varepsilon > 0$ the portion of $\widetilde{S_R}$ on which $\frac{1}{r_1^2 r_2^2} \geqslant \varepsilon r_2^2$ is negligible as $R \to \infty$. Hence we may simplify $(*)$ to r_2^2. Denote by $\widetilde{B_R} \supseteq B_R$ the set corresponding to $\widetilde{S_R}$, and calculate

$$m_{\mathrm{SL}_3(\mathbb{R})}(\widetilde{B_R}) \approx \int_{\widetilde{S_R}} \frac{1}{r_1 r_2}\left(r_1^2 - r_2^2\right) r_1^2 r_2^2\,\mathrm{d}r_1\,\mathrm{d}r_2 = \int_{\widetilde{S_R}} \left(r_1^3 r_2 - r_1 r_2^3\right)\mathrm{d}r_1\,\mathrm{d}r_2.$$

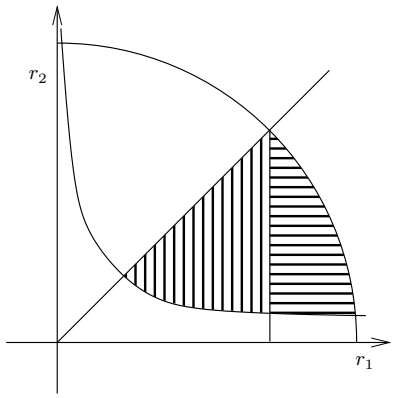


Fig. 5.5: Splitting the region $\widetilde{S_R}$.

Splitting $\widetilde{S_R}$ into two regions depending on whether $r_1 \leqslant \frac{R}{\sqrt{2}}$ or $r_1 \geqslant \frac{R}{\sqrt{2}}$ (see Figure 5.5) this integral can be evaluated in an elementary way. This leads to the asymptotics

$$m_{\mathrm{SL}_3(\mathbb{R})}\big(\widetilde{B_R}\big) \approx R^6.$$

Recall that $B_R \subseteq \widetilde{B_R}$. However, for any $\kappa > 0$ we also have $\widetilde{B}_{R-\kappa} \subseteq B_R$ for all sufficiently large R. Therefore, $m_{\mathrm{SL}_3(\mathbb{R})}(B_R)$ has the same asymptotics. Well-roundedness follows once more from this, giving the proposition. □

Corollary 5.35 now follows once more by combining the discussions of this section with those of Section 5.4.

5.7 Computing the Volume of X_d

In this section we will describe a method for calculating $\mathrm{vol}(\mathsf{X}_d)$ without actually finding a fundamental domain for $\mathrm{SL}_d(\mathbb{Z}) < \mathrm{SL}_d(\mathbb{R})$. Of course the answer depends on a normalization of the Haar measure on $\mathrm{SL}_d(\mathbb{R})$. Both the normalization and the method to find the volume work inductively.

Theorem 5.40 (Volume of X_{d+1}). *For $d \geqslant 1$ define the subgroup*

$$H_d = \left\{ \begin{pmatrix} 1 & w^{\mathrm{t}} \\ 0 & g \end{pmatrix} \;\middle|\; g \in \mathrm{SL}_d(\mathbb{R}), w \in \mathbb{R}^d \right\} < \mathrm{SL}_{d+1}(\mathbb{R}).$$

Assume that $m_{\mathrm{SL}_d(\mathbb{R})}$ has been defined by induction as follows. We start from the normalization $m_{\mathrm{SL}_1(\mathbb{R})}(\{I\}) = 1$. Using the Lebesgue measure on $\mathbb{R}^d$ (and Lemma 1.58) this defines a normalization of the Haar measure

$$m_{H_d} = m_{\mathrm{SL}_d(\mathbb{R})} \times m_{\mathbb{R}^d}.$$

Now use the identification $V = \mathrm{SL}_{d+1}(\mathbb{R})/H_d \cong \mathbb{R}^{d+1} \smallsetminus \{0\}$ to normalize $m_{\mathrm{SL}_{d+1}(\mathbb{R})}$ to also be compatible with the Lebesgue measure on $\mathbb{R}^{d+1}$. Then

$$\mathrm{vol}(\mathsf{X}_{d+1}) = \zeta(d+1)\zeta(d)\cdots\zeta(2).$$

The standing assumption in Section 5.4 were that

$$\mathrm{vol}\big(G/\Gamma\big)$$

and

$$\mathrm{vol}\big(H/\Gamma \cap H\big)$$

were both finite. This follows in the case at hand for $G = \mathrm{SL}_{d+1}(\mathbb{R})$ and $H = H_d$ from Theorem 1.54 and the fact that

$$\left\{ \begin{pmatrix} 1 & w^{\mathrm{t}} \\ & 1 \end{pmatrix} \;\middle|\; w \in \mathbb{R}^d \right\}$$

intersects $\mathrm{SL}_{d+1}(\mathbb{Z})$ in a lattice with covolume 1.

The dynamical assumption of equidistribution of $gH{\cdot}\Gamma$ in

$$X = \mathsf{X}_{d+1} = G/\Gamma$$

for $G = \mathrm{SL}_{d+1}(\mathbb{R})$, $\Gamma = \mathrm{SL}_{d+1}(\mathbb{Z})$ and $gH \to \infty$ in G/H is easy to establish—where it is true. In order to do this, we again need to exploit two decompositions of G (see Exercise 5.44).

Lemma 5.41 (A first group decomposition). *Write*

$$K = \mathrm{SO}_{d+1}(\mathbb{R})$$

and

$$A = \left\{ \begin{pmatrix} a & \\ & a^{-\frac{1}{d}} I \end{pmatrix} \;\middle|\; a > 0 \right\}.$$

Then $G = KAH$.

Lemma 5.42 (A coordinate system). *Let*

$$a_1 = \begin{pmatrix} \mathrm{e} & \\ & \mathrm{e}^{-\frac{1}{d}} I \end{pmatrix} \in A$$

and

$$G^-_{a_1} = \left\{ \begin{pmatrix} 1 & \\ v & 1 \end{pmatrix} \;\middle|\; v \in \mathbb{R}^d \right\}.$$

Then $G^-_{a_1} AH$ is open and the product map from $G^-_{a_1} \times A \times H$ gives a homeomorphism to the image $G^-_{a_1} AH$.

Theorem 5.43 (Translated orbits). *The orbit $gH{\bullet}\Gamma$ equidistributes on average as $gH \to \infty$ in G/H.*

Outline of proof. By Lemma 5.41 it is enough to consider the case $gH = a_t H$ where

$$a_t = \begin{pmatrix} \mathrm{e}^t & \\ & \mathrm{e}^{-\frac{t}{d}} I \end{pmatrix} \in A$$

with $|t| \to \infty$.

Notice that taking $t \leqslant 0$ in a_t cforresponds to non-zero elements in the unit ball of $\mathbb{R}^{d+1}$, and that the unit ball has finite Haar measure on

$$G/H \cong \mathbb{R}^{d+1} \smallsetminus \{0\}$$

(since the Haar measure coincides with the restriction of the Lebesgue measure). As a result we may ignore the case $t \to -\infty$ in the equidistribution claim sought.

The remaining case $t \to \infty$ may be carried out as in Theorem 5.38. □

The geometric hypothesis that the sets

$$B_t = B^{\mathbb{R}^d}_{\mathrm{e}^t} = \left\{ v \in \mathbb{R}^d \;\middle|\; \|v\| \leqslant \mathrm{e}^t \right\} \tag{5.27}$$

are well-rounded for $t \geqslant 0$ is easy to check (see Exercise 5.46).

Proof of Theorem 5.40. For $d = 2$ we have

$$H = \left\{ \begin{pmatrix} 1 & * \\ & 1 \end{pmatrix} \right\}$$

and $\mathrm{vol}(H/\Gamma \cap H) = 1$. Furthermore, notice that

$$\mathrm{SL}_2(\mathbb{Z}) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbb{Z}^2_{\mathrm{prim}} = \left\{ n \in \mathbb{Z}^2 \;\middle|\; \gcd(n) = 1 \right\}.$$

Hence

$$\lim_{R\to\infty} \frac{\left|\mathbb{Z}^2_{\mathrm{prim}} \cap B_R^{\mathbb{R}^2}\right|}{\pi R^2} = \frac{1}{m_{\mathsf{X}_2}(\mathsf{X}_2)}$$

by (5.18). By Exercise 1.2 we also know that

$$\lim_{R\to\infty} \frac{\left|\mathbb{Z}^2_{\mathrm{prim}} \cap B_R^{\mathbb{R}^2}\right|}{\pi R^2} = \frac{1}{\zeta(2)},$$

which implies that $m_{\mathsf{X}_2}(\mathsf{X}_2) = \zeta(2) = \dfrac{\pi^2}{6}$.

For $d \geqslant 2$ we have

$$\mathrm{vol}\left(H/\Gamma \cap H\right) = \mathrm{vol}(\mathsf{X}_d)$$

and

$$\mathrm{SL}_d(\mathbb{Z})e_1 = \mathbb{Z}^d_{\mathrm{prim}} = \left\{ n \in \mathbb{Z}^d \;\middle|\; \gcd(n) = 1 \right\}. \tag{5.28}$$

Combining (5.18) and Exercise 1.2 we get once more

$$\lim_{R\to\infty} \frac{\left|\mathbb{Z}^d_{\mathrm{prim}} \cap B_R^{\mathbb{R}^d}\right|}{V_d R_d} = \frac{1}{\zeta(d)} = \frac{\mathrm{vol}(\mathsf{X}_d)}{\mathrm{vol}(\mathsf{X}_{d+1})}$$

which gives the theorem. □

We leave the details of the arguments above to the exercises below.

Exercise 5.44. Prove Lemmas 5.41 and 5.42.

Exercise 5.45. Prove Theorem 5.43.

Exercise 5.46. Prove that the sets $B_t = B_{\mathrm{e}^t}^{\mathbb{R}^d} = \left\{ v \in \mathbb{R}^d \;\middle|\; \|v\| \leqslant \mathrm{e}^t \right\}$ are well-rounded (see (5.27)).

Exercise 5.47. Prove (5.28) for $d \geqslant 2$.

Exercise 5.48. Prove that

$$N_R = \left| \left\{ W \leqslant \mathbb{R}^d \;\middle|\; \dim(W) = m, W \text{ is rational, and } \mathrm{covol}(W \cap \mathbb{Z}^d) \leqslant R \right\} \right|$$

has an asymptotic of the form $N_R \sim cR^d$.

Notes to Chapter 5

[27](Page 182) We refer to [45, Ex. 3.3.1, 9.6.3] for one example of such a construction. McMullen [112] gives explicit constructions of bounded geodesics of arbitrary length associated to elements of any given quadratic field, and relates the construction to continued fractions.

[28](Page 186) This is an example of a circle of results developed among others by Dani [19, 23] and Veech [161].

[29](Page 192) Results of this sort in greater generality are also known as the Každan–Margulis lemma. Various versions may be found in the original lecture notes of Zassenhaus [169], the monograph of Raghunathan [128], and for an accessible modern treatment we refer to Winkelmann [167].

[30](Page 205) This was shown by Witt [168], and a modern treatment may be found in the monograph of Elman, Karpenko and Merkurjev [50].