Chapter 1
Lattices and the Space of Lattices

We start by recalling that a (continuous) *action* of a (topological) group G on a (topological) space X is a (continuous) map $G \times X \to X$, written $(g, x) \mapsto g{\cdot}x$, with the property that $g{\cdot}(h{\cdot}x) = (gh){\cdot}x$ and $e{\cdot}x = x$ for all $g, h \in G$ and $x \in X$, where e is the identity element of G. Furthermore, for any $x \in X$ the set

$$G{\cdot}x = \{g{\cdot}x \mid g \in G\}$$

is called the *G-orbit* of x and

$$\mathrm{Stab}_G(x) = \{g \in G \mid g{\cdot}x = x\}$$

is the *stabilizer subgroup* of x. There is a canonical isomorphism

$$G/\mathrm{Stab}_G(x) \ni g\,\mathrm{Stab}_G(x) \longmapsto g{\cdot}x \in G{\cdot}x$$

which we may refer to as the 'orbit stabilizer theorem'. The isomorphism carries the natural G-action by left multiplication on $G/\mathrm{Stab}_G(x)$ to the G-action on $G{\cdot}x \subseteq X$, but may or may not be a homeomorphism.

One of our interests in this volume is to study the relationship between orbits, orbit closures, and arithmetic properties of groups.

In this chapter we discuss discrete subgroups Γ of a locally compact σ-compact metric group G, the quotient space $X = \Gamma\backslash G$, which we will refer to as a locally homogeneous space, and the question of whether or not there is a G-invariant Borel probability measure on X. We finish by studying the central example X_d, the space of unimodular lattices in $\mathbb{R}^d$. In other words, we define the spaces (and the canonical measures) on which (or with respect to which) we will later discuss dynamical and arithmetic properties.

1.1 The Gauss Circle Problem

We start our discussions by outlining a lattice-point counting problem in the classical setting of the Gauss circle problem. This problem asks for the asymptotic count of the number of points in $\mathbb{Z}^2$ that lie within the disc of radius R.

Proposition 1.1. *For any $R > 0$ let*

$$N(R) = \big|\{n \in \mathbb{Z}^2 \mid \|n\| \leqslant R\}\big|,$$

where we write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^2$. Then

$$N(R) = \pi R^2 + \mathrm{O}(R).$$

The proof is highly geometric. Indeed, the main term πR^2 is the area of the 2-dimensional ball of radius R, and the error term is related to the area of an annulus, as indicated in Figure 1.1.

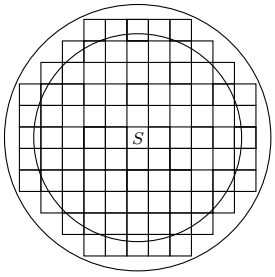

Fig. 1.1: Containing the error term for $N(R)$ inside an annulus.

Proof of Proposition 1.1. Consider the unit square $S = [-\frac{1}{2}, \frac{1}{2}) \times [-\frac{1}{2}, \frac{1}{2})$, which is a 'fundamental domain' for the 'lattice' $\mathbb{Z}^2$ in the group $\mathbb{R}^2$. Then, as indicated in Figure 1.1, we have

$$B_{R-\frac{1}{\sqrt{2}}}(0) \subseteq S + \{n \in \mathbb{Z}^2 \mid \|n\| \leqslant R\} \subseteq B_{R+\frac{1}{\sqrt{2}}}(0).$$

By taking areas, we conclude that

$$\left(R - \frac{1}{\sqrt{2}}\right)^2 \pi \leqslant N(R) \leqslant \left(R + \frac{1}{\sqrt{2}}\right)^2 \pi$$

as required. □

It is conjectured that[(1)]

$$N(R) = \pi R^2 + \mathrm{O}_\varepsilon\big(R^{\frac{1}{2}+\varepsilon}\big)$$

for all $\varepsilon > 0$. We refer to the paper of Ivić, Krätzel, Kühleitner and Nowak [72] for a survey of the many partial results towards this conjecture.

One of the success stories of homogeneous dynamics concerns other more delicate counting results. These are often obtained as corollaries of related equidistribution results, which in turn might be obtained by dynamical methods.

Roughly speaking, an equidistribution result for expanding circles inside the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ could help to improve the error term in Proposition 1.1 for the following reason. Near the circle of radius R a portion of the fundamental domain $n + S$ lies in the disc of radius R while the centre point $n \in \mathbb{Z}^2$ may or may not belong to it. If these two opposite cases arise with equal asymptotic frequency due to an equidistribution result then one might expect to improve the error term.[(2)]

We will however be interested in counting results in other spaces, as indicated at the end of the next section for example. The required equidistribution will then take place in 'quotients' that we will introduce in this chapter.

Exercise 1.2. Let $d \geqslant 2$. Prove that

$$N^*(R) = |\{n \in \mathbb{Z}^d \mid n \text{ is primitive and } \|n\| \leqslant R\}|$$

satisfies $N^*(R) = \big(\zeta(d)^{-1}V_d + o(1)\big)R^2$ as $R \to \infty$. Here V_d is the volume of the unit ball in $\mathbb{R}^d$ and $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ denotes the Riemann zeta function.

1.2 A Brief Review of $\mathbf{SL_2}(\mathbb{Z})\backslash \mathbf{SL_2}(\mathbb{R})$

We continue our introduction by motivating future discussions using a concrete visual setup. In the following sections and chapters we will prove generalizations of the facts presented here.

1.2.1 The Space

We recall (see, for example, [45, Ch. 9]) that the upper half-plane

$$\mathbb{H} = \{z = x + \mathrm{i}y \in \mathbb{C} \mid y = \Im(z) > 0\}$$

equipped with the Riemannian metric

$$\langle u, v\rangle_z = \frac{u \cdot v}{y^2}$$

for tangent vectors $(z, u), (z, v) \in \mathrm{T}_z\mathbb{H} = \{z\} \times \mathbb{C}$ is the *upper half-plane model* of the hyperbolic plane (where $u \cdot v$ denotes the inner product after identifying u and v with elements of $\mathbb{R}^2$). Moreover, the group $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathbb{H}$ transitively and isometrically via the Möbius transformation

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \mathbb{H} \ni z \longmapsto g{\cdot}z = \frac{az+b}{cz+d}. \tag{1.1}$$

The stabilizer of $\mathrm{i} \in \mathbb{H}$ is $\mathrm{SO}_2(\mathbb{R})$ so that

$$\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R}) \cong \mathbb{H}$$

under the map sending $g\,\mathrm{SO}_2(\mathbb{R})$ to $g{\cdot}\mathrm{i}$.

The action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ is differentiable, and so gives rise to a derived action on the tangent bundle $\mathrm{T}\mathbb{H} = \mathbb{H} \times \mathbb{C}$ by

$$\mathrm{D}\,g\colon (z,v) \longmapsto \left(g{\cdot}z, \frac{1}{(cz+d)^2}v\right),$$

where

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

and $\frac{1}{(cz+d)^2}$ is the complex derivative of $\mathbb{H} \ni z \mapsto g{\cdot}z$ as in (1.1). This action gives rise to the simply transitive action of

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

on the unit tangent bundle

$$\mathrm{T}^1\mathbb{H} = \big\{(z,v) \in \mathrm{T}\mathbb{H} \mid \|v\|_z^2 = \langle v, v\rangle_z = 1\big\},$$

so that

$$\mathrm{PSL}_2(\mathbb{R}) \cong \mathrm{T}^1\mathbb{H}.$$

This isomorphism may be chosen to send g to $\mathrm{D}\,g(\mathrm{i}, \uparrow)$, where we write $\uparrow$ for the upward pointing vector of hyperbolic length 1 at any $z \in \mathbb{H}$.

We recall that the hyperbolic plane has interesting and important geometric properties. For instance, geodesics (shortest paths connecting two points) follow straight vertical lines or half-circles intersecting the real line at a normal angle. For dynamical questions it is however too big. Instead we will always involve a discrete subgroup $\Gamma < \mathrm{PSL}_2(\mathbb{R})$ and use this to 'fold up' $\mathbb{H}$ and $\mathrm{T}^1\mathbb{H}$. Ideally one would want the quotient by the action of Γ to be compact, but this is too restrictive.

Let us highlight $\Gamma = \mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$ as an example of such a discrete subgroup. For $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ the shaded region E in Figure 1.2 is a fundamental region for the action of Γ on $\mathbb{H}$. By this we mean that

$$|E \cap \Gamma{\cdot}z| = 1$$

for every $z \in \mathbb{H}$. Strictly speaking we should describe carefully which parts of the boundary of the hyperbolic triangle shaded belong to the domain, but as the boundary is a nullset one usually ignores that issue—we will follow this tradition (see Exercise 1.4).

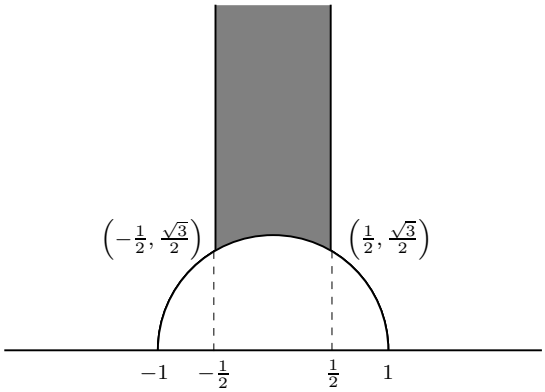

Fig. 1.2: A fundamental domain $E \subseteq \mathbb{H}$ for the action of $\mathrm{SL}_2(\mathbb{Z})$.

This shows that we can define a fundamental domain for the left action of $\mathrm{PSL}_2(\mathbb{Z})$ on

$$\mathrm{PSL}_2(\mathbb{R}) \cong \mathrm{T}^1\mathbb{H}$$

by taking all vectors (z, u) whose base point z lies in E, giving the set

$$F = \{g \in \mathrm{PSL}_2(\mathbb{R}) \mid \mathrm{D}\, g(\mathrm{i}, \uparrow) = (z, u) \text{ with } z \in E\}. \tag{1.2}$$

Once again, strictly speaking we should describe more carefully which vectors attached to points $z \in \partial E$ are allowed in F (see Exercise 1.4).

We claim that this argument shows that

$$\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R}) \cong \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$$

has finite volume. In order to see this, we recall some basic facts from [45, Ch. 9] (which we will prove in greater generality for $\mathrm{SL}_d(\mathbb{R})$ in Section 1.4.4):

- $\mathrm{SL}_2(\mathbb{R})$ is unimodular, meaning that there is a bi-invariant Haar measure on $\mathrm{SL}_2(\mathbb{R})$ (see Exercise 1.7).
- $\mathrm{SL}_2(\mathbb{R}) = NAK$ with[†]

$$N = U^- = \left\{ \begin{pmatrix} 1 & * \\ & 1 \end{pmatrix} \right\}, \qquad A = \left\{ \begin{pmatrix} a & \\ & a^{-1} \end{pmatrix} \,\middle|\, a > 0 \right\},$$

  and $K = \mathrm{SO}_2(\mathbb{R})$, in the sense that every $g \in \mathrm{SL}_2(\mathbb{R})$ can be written uniquely$^{(3)}$ as a product $g = nak$ with $n \in N$, $a \in A$ and $k \in K$.
- Let $B = NA = AN$ be the subgroup

[†] We sometimes indicate by $*$ any entry of a matrix which is only restricted to be a real number, and do not write entries that are zero.

$$B = \left\{ \begin{pmatrix} a & t \\ & a^{-1} \end{pmatrix} \;\middle|\; a > 0, t \in \mathbb{R} \right\}.$$

The Haar measure $m_{\mathrm{SL}_2(\mathbb{R})}$ decomposes in the coordinates $g = bk$, meaning that

$$m_{\mathrm{SL}_2(\mathbb{R})} \propto m_B \times m_K,$$

where $\propto$ denotes proportionality. The constant of proportionality only depends on the choice of the Haar measures.

- Moreover, the left Haar measure m_B decomposes in the coordinate system

$$b(x,y) = \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \begin{pmatrix} y^{1/2} & \\ & y^{-1/2} \end{pmatrix}$$

with $x \in \mathbb{R}$, $y > 0$, as

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

We also note that $b(x,y)\cdot\mathrm{i} = \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix}\cdot(\mathrm{i}y) = x + \mathrm{i}y$, and that the Haar measure m_B on B is identical to the hyperbolic area measure on $\mathbb{H}$ under the map $b(x,y) \mapsto b(x,y)\cdot\mathrm{i} = x + \mathrm{i}y$.

Combining these facts we get

$$m_{\mathrm{SL}_2(\mathbb{R})}(F) \leqslant \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{\frac{\sqrt{3}}{2}}^{\infty} \int_0^{\pi} \frac{1}{y^2}\,\mathrm{d}\theta\,\mathrm{d}y\,\mathrm{d}x < \infty.$$

The argument above also helps us to understand the space

$$\mathsf{X}_2 = \mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$$

globally: It is, apart from some difficulties arising from the distinguished points $\mathrm{i}, \frac{1}{2} + \frac{\sqrt{3}}{2}\mathrm{i} \in E$, the unit tangent bundle of the surface[†] $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. This surface may be thought of as being obtained by gluing the two vertical sides in Figure 1.2 together using the action of $\begin{pmatrix} 1 & \pm 1 \\ & 1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ and the third side to itself using the action of $\begin{pmatrix} & -1 \\ 1 & \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$. In particular, X_2 is non-compact; see Figure 1.3 and Exercise 1.6.

We note that $g \in \mathrm{SL}_2(\mathbb{R})$ acts on $x \in \mathsf{X}_2$ by setting $g\cdot x = xg^{-1}$. However, as we will discuss next, the geometric meaning of this action varies depending on the subgroup of $\mathrm{SL}_2(\mathbb{R})$ considered.

Exercise 1.3. Show that $K = \mathrm{SO}_2(\mathbb{R})$ is the stabilizer of $\mathrm{i} \in \mathbb{H}$. Moreover, its action on $\mathrm{T}_{\mathrm{i}}\mathbb{H}$ (by the derivative of the Möbius transformations) rotates the tangent vectors at 'double speed' clockwise. That is,

[†] Because of the distinguished points this surface is a good example of an *orbifold*, but not an example of a manifold.

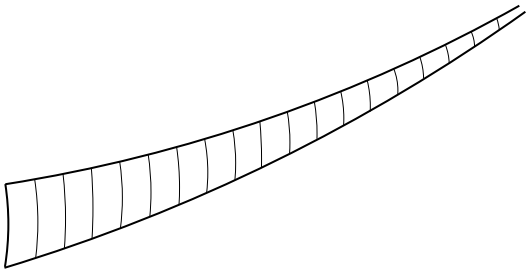

Fig. 1.3: Folding the hyperbolic triangle in Figure 1.2 creates a surface stretching off to infinity (the cusp) and with two exceptional points (with conical singularities).

$$k_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

applied to $(\mathrm{i}, v) \in \mathrm{T}_{\mathrm{i}}\mathbb{H}$ gives $(\mathrm{i}, \mathrm{e}^{-2\theta\mathrm{i}}v) \in \mathrm{T}_{\mathrm{i}}\mathbb{H}$.

Exercise 1.4. Let E be as in Figure 1.2.
(a) Use $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ to show that $\mathrm{SL}_2(\mathbb{Z})\bullet E$ is 'uniformly open', meaning that there exists some $\delta > 0$ such that $z \in \mathrm{SL}_2(\mathbb{Z})\bullet E$ implies that $B_\delta(z) \subseteq \mathrm{SL}_2(\mathbb{Z})\bullet E$. Conclude that $\mathrm{SL}_2(\mathbb{Z})\bullet E = \mathbb{H}$.
(b) Suppose that both z and $\gamma\bullet z$ lie in E for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. Show that either $\gamma = \pm I$ or $z \in \partial E$.
(c) Conclude that E can be modified (by defining which parts of the boundary of E should be included) to become a fundamental domain.
(d) Modify the definition of E in (1.2) at the points i and $\frac{1}{2} + \frac{\sqrt{3}}{2}\mathrm{i}$ so that F is indeed a fundamental domain for the left action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathrm{PSL}_2(\mathbb{R})$.

Exercise 1.5. Describe the orbit corresponding to the geodesic just on the left of the fundamental domain. That is, draw the continuation of the ray from ∞ to $-\frac{1}{2} + \frac{\sqrt{3}}{2}\mathrm{i}$ modulo $\mathrm{SL}_2(\mathbb{Z})$ as a subset of $\overline{E} \subseteq \mathbb{H}$.

Exercise 1.6. Show that the space $\mathrm{SL}_2(\mathbb{R})/\mathrm{SL}_2(\mathbb{Z}) \cong \{g\mathbb{Z}^2 \mid g \in \mathrm{SL}_2(\mathbb{R})\}$ can be identified with lattices $g\mathbb{Z}^2 \leqslant \mathbb{R}^2$ of covolume $\det g = 1$. Use the isomorphism with $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{T}^1\mathbb{H}$ discussed in this section to characterize compact subsets K of $\mathrm{SL}_2(\mathbb{R})/\mathrm{SL}_2(\mathbb{Z})$ in terms of elements of the lattices $g\mathbb{Z}^2$ for $g\,\mathrm{SL}_2(\mathbb{Z}) \in K$. More precisely, calculate the relationship between the shortest vector in $g\mathbb{Z}^2$ and the imaginary part of $g^{-1}\mathrm{i} \in \mathbb{H}$ under the assumption that the representative $g \in \mathrm{SL}_2(\mathbb{R})$ has been chosen with $g^{-1}\mathrm{i} \in E$ (with $E \subseteq \mathbb{H}$ as in Figure 1.2).

Exercise 1.7. Let $d \geqslant 2$. Show that

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

for any measurable $B \subseteq \mathrm{SL}_d(\mathbb{R})$ defines a (bi-invariant) Haar measure on the locally compact group

$$\mathrm{SL}_d(\mathbb{R}) = \big\{g \in \mathrm{Mat}_d(\mathbb{R}) \mid \det(g) = 1\big\},$$

which is called the *special linear group*, where $m_{\mathbb{R}^{d^2}}$ is the Lebesgue measure on the matrix algebra $\mathrm{Mat}_d(\mathbb{R})$ viewed as the vector space $\mathbb{R}^{d^2}$.

1.2.2 The Geodesic Flow—the Subgroup A

We recall that

$$a_t \colon \mathsf{X}_2 \ni x \longmapsto x \begin{pmatrix} \mathrm{e}^{t/2} & \\ & \mathrm{e}^{-t/2} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{-t/2} & \\ & \mathrm{e}^{t/2} \end{pmatrix} \cdot x$$

defines the geodesic flow on X_2 (see Exercise 1.8), whose orbits may also be described in the fundamental region as in Figure 1.4.

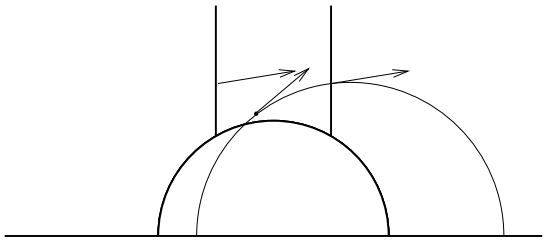

Fig. 1.4: The geodesic flow follows the circle determined by the arrow which intersects $\mathbb{R} \cup \{\infty\} = \partial\mathbb{H}$ normally, and is moved back to F via a Möbius transformation in $\mathrm{SL}_2(\mathbb{Z})$ once the orbit leaves F.

The diagonal subgroup

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

is also called a *Cartan subgroup*. We note that A acts ergodically on X_2 with respect to the Haar measure m_{X_2}, which we will also discuss from a more general point of view in Chapter 2.

There are many different types of A-orbits, which include the following:

- Divergent trajectories, for example the orbit $\mathrm{SL}_2(\mathbb{Z})A$ which corresponds to the vertical geodesic through $(\mathrm{i}, \uparrow)$ in $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{T}^1\mathbb{H}$.
- Compact trajectories, for example $\mathrm{SL}_2(\mathbb{Z})g_{\text{golden}}A$ is compact, where the matrix $g_{\text{golden}} \in K$ has the property[†] that

$$g_{\text{golden}}^{-1} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} g_{\text{golden}} = \begin{pmatrix} \frac{3+\sqrt{5}}{2} & \\ & \frac{3-\sqrt{5}}{2} \end{pmatrix} \in A.$$

Now notice that

[†] The eigenvalues of $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ are $\frac{3\pm\sqrt{5}}{2}$, and there is such a matrix $g_{\text{golden}} \in K$ because $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ is symmetric.

$$\mathrm{SL}_2(\mathbb{Z})g_{\text{golden}}\begin{pmatrix}\frac{3+\sqrt{5}}{2} & \\ & \frac{3-\sqrt{5}}{2}\end{pmatrix} = \mathrm{SL}_2(\mathbb{Z})\begin{pmatrix}1 & 1\\ 1 & 2\end{pmatrix} g_{\text{golden}} = \mathrm{SL}_2(\mathbb{Z})g_{\text{golden}}.$$

This identity shows that the orbit $\mathrm{SL}_2(\mathbb{Z})g_{\text{golden}}A$ is compact (see also Figure 1.5 in which $\lambda = \frac{1+\sqrt{5}}{2}$).

- The set of dense trajectories, which includes (but is much larger than) the set of equidistributed trajectories of typical points in $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$.
- Orbits that are neither dense nor closed.
- Orbits that exhibit completely different behaviour in the past and in the future.

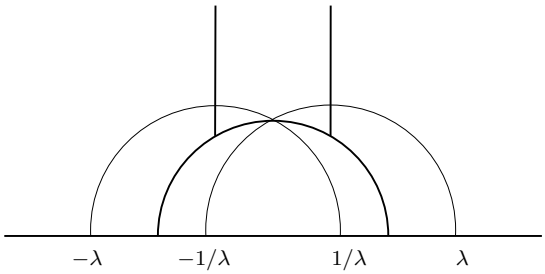

Fig. 1.5: The union of the two geodesics considered in X_2 with both directions allowed is a periodic A-orbit, and comprises the orbit $\mathrm{SL}_2(\mathbb{Z})g_{\text{golden}}A$.

Finally we would like to point out—in a sense to be made precise in Sections 3.1 and 3.6—that there is a correspondence between rational (or arithmetic) objects and closed A-orbits as in the first two types of A-orbit considered above (see Exercise 1.10 and 1.11).

Exercise 1.8. The *geodesic flow* on $\mathrm{T}^1\mathbb{H}$ is the action of $t \in \mathbb{R}$ defined in geometric terms as follows. Starting at a point $(z, v) \in \mathrm{T}^1\mathbb{H}$ draw the unique geodesic in $\mathbb{H}$ through z and tangent to v at z. Now follow this geodesic to a point at distance t (forward if $t > 0$ and backwards if $t < 0$). Let the image of (z, v) under the flow for time t be this point and the tangent vector of the geodesic at this point. Notice that for $(\mathrm{i}, \uparrow)$ this gives $(\mathrm{e}^t\mathrm{i}, \uparrow)$. Show that under the isomorphism $\mathrm{PSL}_2(\mathbb{R}) \ni g \mapsto Dg_\bullet(\mathrm{i}, \uparrow)$ the geodesic flow corresponds to right multiplication by a_t^{-1}.

Exercise 1.9. (a) Show that every geodesic on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ intersects the image of the geodesic segment from $-\frac{1}{2} + \frac{\sqrt{3}}{2}\mathrm{i}$ to $\frac{1}{2} + \frac{\sqrt{3}}{2}\mathrm{i}$.
(b) Show that every geodesic on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ intersects the periodic horocycle segment defined by $\{x + \mathrm{i} \mid x \in [-\frac{1}{2}, \frac{1}{2})\}$.

Exercise 1.10. Show that $\mathrm{SL}_2(\mathbb{Z})gA$ is a divergent trajectory ($A \ni a \mapsto \mathrm{SL}_2(\mathbb{Z})ga$ is a proper map) if and only if $ga \in \mathrm{SL}_2(\mathbb{Q})$ for some $a \in A$.

Exercise 1.11. Show that to any compact A-orbit in $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ one can attach a real quadratic number field K such that the length of the orbit is $\log|\xi|$, where ξ in $\mathcal{O}_K^*$ is a unit in the order $\mathcal{O}_K$ of K. Prove that there are only countably many such orbits.

1.2.3 The Horocycle Flow—the Subgroup $U^- = N$

We recall that the *(stable) horocycle flow* on X_2 is defined by the action

$$u_s \colon x \longmapsto x \begin{pmatrix} 1 & -s \\ & 1 \end{pmatrix} = \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} \cdot x$$

for $s \in \mathbb{R}$. Here the matrices

$$u_s = u_s^- = \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix}$$

for $s \in \mathbb{R}$ are unipotent (that is, only have 1 as an eigenvalue) and the corresponding subgroup

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

is precisely the *stable horospherical subgroup* of the geodesic flow, in the sense that

$$U^- = \left\{ g \in \mathrm{SL}_2(\mathbb{R}) \;\middle|\; a_t g a_t^{-1} \longrightarrow I \text{ as } t \longrightarrow \infty \right\}.$$

This implies that

$$\mathsf{d}_{\mathsf{X}_2}\left(a_t \cdot (x), a_t \cdot (u_s \cdot x)\right) \longrightarrow 0 \tag{1.3}$$

as $t \to \infty$ for any $x \in \mathsf{X}_2$ and $s \in \mathbb{R}$. We will define the metric $\mathsf{d}_{\mathsf{X}_2}$ and verify this claim in Section 1.3.

Geometrically, horocycle orbits can be described as circles in $\mathbb{H}$ touching the real axis with the arrows (that is, the tangent space component) normal to the circle pointing inwards or as horizontal lines with the arrows pointing upwards, as in Figure 1.6.

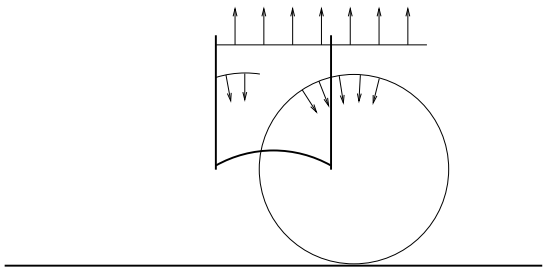

Fig. 1.6: The picture shows the two types of horocycle orbits; the orbits in X_2 can again be understood by using the appropriate Möbius transformation whenever the orbit leaves the fundamental domain.

We note that U^- also acts ergodically on X_2 with respect to the Haar measure m_{X_2} (see Chapter 2). However, unlike the case of A-orbits, the classification

of U^--orbits on X_2 is shorter (we will discuss this phenomenon again, and in particular we will prove the facts below in Chapter 5 and more general results in Chapter 6). The possibilities for U^--orbits are as follows:

- Compact trajectories, for example $\mathrm{SL}_2(\mathbb{Z})U^-$ is compact and corresponds to the horizontal orbit through $(\mathrm{i}, \uparrow) \in \mathrm{T}^1\mathbb{H}$.
- Dense trajectories, which are automatically also equidistributed with respect to m_{X_2} (both in their past and in their future).

This gives the complete list of types of U^--orbits (see Section 5.1), and once more gives substance to the claim that there is a correspondence between rational objects and closed orbits (see Exercise 1.13).

We also define

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

which we refer to as the *unstable horospherical subgroup*. The results above hold similarly for U^+, which is in fact conjugate to U^-.

The following should help explain the notation used for $U^{\pm}$.

Exercise 1.12. Show that conjugation by a_t normalizes the subgroups $U^{\pm}$ and changes the natural parameter in these groups by the factor $\mathrm{e}^{\pm t}$.

Exercise 1.13. Show that $\mathrm{SL}_2(\mathbb{Z})gU^-$ is compact if and only if $g(\infty) \in \mathbb{Q} \cup \{\infty\}$. Show that if $\mathrm{SL}_2(\mathbb{Z})gU^-$ is compact, then $\mathrm{SL}_2(\mathbb{Z})gU^- = \mathrm{SL}_2(\mathbb{Z})aU^-$ for some $a \in A$.

1.2.4 The Subgroups K and B

For $\mathrm{SL}_2(\mathbb{R})$ there are two more connected subgroups of importance (and up to conjugation this completes the list of connected subgroups), namely

- $K = \mathrm{SO}_2(\mathbb{R}) \subseteq \mathrm{SL}_2(\mathbb{R})$, and
- $B = U^-A = \left\{ \begin{pmatrix} \mathrm{e}^{-t/2} & s \\ & \mathrm{e}^{t/2} \end{pmatrix} \;\middle|\; s, t \in \mathbb{R} \right\}$

However, we note that for these two there is no correspondence between closed orbits and rational objects: For example, every K-orbit is compact since K itself is compact. On the other hand, every B-orbit is dense, independently of any rationality questions. In fact the latter follows from the properties of the horocycle flow. If xU^- is not periodic, then it is dense by the mentioned classification of U^--orbits in Section 1.2.3. If xU^- is periodic, then one can choose $a \in A$ so that $xaU^- \subseteq xB$ is a much longer periodic orbit. However, long periodic U^--orbits equidistribute in X_2 (see Sarnak [132] and Section 5.3.1).

This shows that the phenomenon of a correspondence between closed orbits and rational objects is more subtle. It can only hold in certain situations, which we will discuss starting in Chapter 3.

1.2.5 Intersections with Arithmetic Meaning

We wish to discuss a curious interplay between a geometric question formulated within $\mathsf{X}_2 = \mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ and arithmetic considerations. For this we define the cylinder

$$Y = \left\{\mathrm{SL}_2(\mathbb{Z})u_s^- a_\varepsilon^{-1} \mid s, \varepsilon \in [0,1]\right\}$$

using the directions in $B = U^-A$ as in Figure 1.7. We also define the closed loop

$$L = \left\{\mathrm{SL}_2(\mathbb{Z})u_{s_+}^+ \;\middle|\; s_+ \in [0,1]\right\}$$

using the third direction U^+ transverse to B as in Figure 1.7.

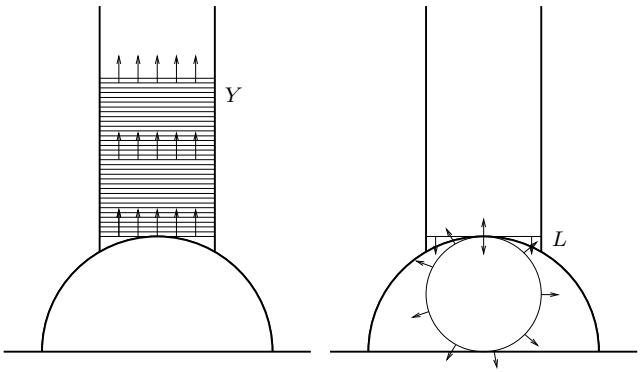

Fig. 1.7: The set Y on the left is a cylinder within the three-dimensional quotient X_2. The set L on the right can be drawn outside of F using a circle tangent to $\mathbb{R}$ at 0 or inside of F using a horizontal line with arrows pointing downwards. Note that L and Y intersect only at $(\mathrm{i}, \uparrow)$.

Applying a_t for a large $t \geqslant 0$ to L we obtain a new loop La_t^{-1} of length e^t, which will become more and more equidistributed in X_2 (as mentioned in Section 1.2.4; see Section 5.3.1). In particular, as Y is two-dimensional and transverse to La_t^{-1} within the three-dimensional space X_2 these two submanifolds have to intersect. That is, for every large enough t there exist $s^-, s^+, \varepsilon \in [0,1]$ so that

$$\mathrm{SL}_2(\mathbb{Z})u_{s^-}^- a_\varepsilon^{-1} = \mathrm{SL}_2(\mathbb{Z})u_{s^+}^+ a_t^{-1}.$$

Multiplying by $(u_{s^-}^- a_\varepsilon^{-1})^{-1}$ on the right we deduce that for infinitely many $t \geqslant 0$ there exist $s^-, s^+ \in [0,1]$ so that

$$u_{s^+}^+ a_t^{-1} (u_{s^-}^-)^{-1} = \gamma \in \mathrm{SL}_2(\mathbb{Z}). \tag{1.4}$$

We now calculate

$$u^+_{s^+} a_t^{-1} (u^-_{s^-})^{-1} = \begin{pmatrix} 1 & \\ s^+ & 1 \end{pmatrix} \begin{pmatrix} \mathrm{e}^{t/2} & \\ & \mathrm{e}^{-t/2} \end{pmatrix} \begin{pmatrix} 1 & -s^- \\ & 1 \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{t/2} & -\mathrm{e}^{t/2} s^- \\ \mathrm{e}^{t/2} s^+ & * \end{pmatrix}$$

where the remaining entry on the bottom right will not be important. By (1.4) this should be equal to

$$\begin{pmatrix} \mathrm{e}^{t/2} & -\mathrm{e}^{t/2} s^- \\ \mathrm{e}^{t/2} s^+ & * \end{pmatrix} = \gamma = \begin{pmatrix} a & -b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

for some $a, b, c, d \in \mathbb{Z}$. It follows that $t \geqslant 0$ should have the property that

$$\mathrm{e}^{t/2} = a \geqslant 1$$

is an integer. Moreover, $s^+ = \frac{c}{a}$, $s^- = \frac{b}{a}$, and

$$bc \equiv \det \gamma = 1 \pmod{a}. \tag{1.5}$$

Conversely, every tuple $(a, b, c) \in \mathbb{N}^3$ satisfying (1.5) will create an intersection of $\mathrm{SL}_2(\mathbb{Z})U^-$ and $\mathrm{SL}_2(\mathbb{Z})U^+ a_t^{-1}$ for $t = 2 \log a$.

Let us summarize and refine the calculation above. The one-dimensional loop $\mathrm{SL}_2(\mathbb{Z})U^-$ intersects the one-dimensional loop $\mathrm{SL}_2(\mathbb{Z})U^+ a_t^{-1}$ precisely when $t = 2 \log a$ for some $a \in \mathbb{N}$. Moreover, at those times there are precisely $\phi(a) = \left|(\mathbb{Z}/a\mathbb{Z})^\times\right|$ intersections and the natural coordinates with $\mathrm{SL}_2(\mathbb{Z})U^-$ and $\mathrm{SL}_2(\mathbb{Z})U^+ a_t^{-1}$ are rational with denominator a and numerators b and c satisfying $bc \equiv 1$ modulo a.

This interaction between geometric and dynamical properties on one hand and number theory on the other hand will be seen in many more instances in the course of our discussions.

Exercise 1.14. Verify the converse claim above, the precise description, and the count of the intersections.

1.2.6 Counting Points in $\Gamma \cdot \mathrm{i} \subseteq \mathbb{H}$

As indicated in Section 1.1, asymptotic counting results give rise to interesting applications of homogeneous dynamics. In this section we mention a particular case related to $\mathbb{H}$ equipped with the metric $\mathsf{d}_{\mathbb{H}}$ induced by the hyperbolic Riemannian metric.

Theorem 1.15 (Selberg). *We have*

$$\begin{aligned} N(R) &= \left|\{\gamma \cdot \mathrm{i} \mid \mathsf{d}_{\mathbb{H}}(\gamma \cdot \mathrm{i}, \mathrm{i}) < R, \gamma \in \mathrm{PSL}_2(\mathbb{R})\}\right| \\ &= \frac{\mathrm{vol}\left(B_R^{\mathbb{H}}(\mathrm{i})\right)}{2\,\mathrm{vol}\left(\mathrm{PSL}_2(\mathbb{R})\backslash\mathbb{H}\right)} + \mathrm{o}\big(\mathrm{vol}\big(B_R^{\mathbb{H}}(\mathrm{i})\big)\big) \end{aligned}$$

as $R \to \infty$.

Selberg [137] used a completely different (spectral) method to prove this theorem,[(4)] and obtains additional information about the error term. We will present an approach following Eskin and McMullen [51] that uses mixing and equidistribution following the set-up of Duke, Rudnick and Sarnak [37]. As we saw in Section 1.1, the simple argument for counting problems connected to the lattice $\mathbb{Z}^2 \leqslant \mathbb{R}^2$ that simply tiles the disc of radius R using translates of a fundamental domain will give a heuristic rationale for the main term. Moreover, in this case the error term was simply the area of an annulus. However, for the hyperbolic plane the area of the disc or radius R is asymptotic to $\pi e^{\mathbb{R}}$. This complicates the counting problem since the volume $\mathrm{vol}\left(B^{\mathbb{H}}_{R+c}(\mathrm{i}) \smallsetminus B^{\mathbb{H}}_{R-c}(\mathrm{i})\right)$ of an annulus is comparable in size to the volume $\mathrm{vol}\left(B^{\mathbb{H}}_{R}(\mathrm{i})\right)$ of the ball. In other words, the error term produced by the annulus has the same order of magnitude as the main term.

Another complication arises as for $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ the fundamental domain in Figure 1.2 is unbounded, so in order to use an annulus to capture all of them we should use $c = \infty$ (or at least some large value to capture most of the translates of the fundamental domain).

Because of this—a manifestation of the hyperbolic geometry at work here—the study of the boundary effects is much more important than it is in the case of $\mathbb{Z}^2 < \mathbb{R}^2$, where the volume of the annulus is asymptotically negligible in comparison with the volume of the ball.

To estimate these boundary effects we will need the following equidistribution result concerning large circles as illustrated in Figure 1.8 (which will be a consequence of the 'mixing' that will be discussed in Chapter 2).

Theorem 1.16 (Equidistribution of Large Circles). *For any point z in $\mathbb{H}$, the circles obtained by following geodesics from z in all directions for time t equidistribute in $\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{T}^1\mathbb{H}$.*

We will give proofs of generalizations of Theorems 1.15 and 1.16 as well as the details of the setup used by Duke–Rudnick–Sarnak in Chapter 5.

1.3 Discrete Subgroups and Lattices

We now start our formal discussions and introduce the quotient spaces we will mainly work with.

1.3.1 Metric, Topological, and Measurable Structure

In this section, we will always assume that G is a locally compact σ-compact metric group endowed with a left-invariant metric d_G giving rise to the topology of G. For example, d_G could be the metric derived from a Riemannian metric on

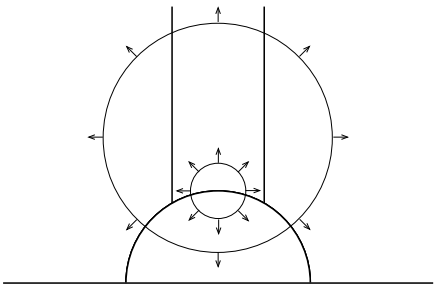

Fig. 1.8: Equidistribution of large circles in the modular surface becomes visible after the circle is moved into the fundamental domain using the isometries in Γ. The setup of Duke–Rudnick–Sarnak allows this equidistribution to be used to obtain a version of Selberg's counting result as in Theorem 1.15.

a connected Lie group G (see Exercise 1.21), but in fact any topological group with a countable basis for the topology has such a metric (see Lemma A.2). We note that the left-invariance of the metric implies that

$$\mathsf{d}_G(g,e) = \mathsf{d}_G(g^{-1}g, g^{-1}) = \mathsf{d}_G(g^{-1}, e)$$

for any $g \in G$. Write $B_r^G = B_r^G(e)$ for the metric open ball of radius r around the identity $e \in G$ so that the above shows precisely $\left(B_r^G\right)^{-1} = B_r^G$ for any $r > 0$. We also note that

$$B_{r_1}^G B_{r_2}^G \subseteq B_{r_1+r_2}^G \tag{1.6}$$

for $r_1, r_2 > 0$. To see this, let $g_1, g_2 \in G$ and notice that

$$\mathsf{d}_G(g_1g_2, e) = \mathsf{d}_G(g_2, g_1^{-1}) \leqslant \mathsf{d}_G(g_2, e) + \mathsf{d}_G(e, g_1^{-1}) = \mathsf{d}_G(g_1, e) + \mathsf{d}_G(g_2, e).$$

If Γ is a *discrete* subgroup (which means that e is an isolated point of Γ), then there is an induced metric on the right quotient space $X = \Gamma\backslash G$ defined by

$$\mathsf{d}_X(\Gamma g_1, \Gamma g_2) = \inf_{\gamma_1,\gamma_2\in\Gamma} \mathsf{d}_G(\gamma_1 g_1, \gamma_2 g_2) = \inf_{\gamma\in\Gamma} \mathsf{d}_G(\gamma g_1, g_2) \tag{1.7}$$

for any right cosets $\Gamma g_1, \Gamma g_2 \in X$, where both infima are minima if the metric is proper[†].

We note that $\mathsf{d}_X(\cdot,\cdot)$ indeed defines a metric on X, and that we will always use the topology induced by this metric. In particular, a sequence $\Gamma g_n \in X$ converges to Γg as $n \to \infty$ if and only if there exists a sequence $\gamma_n \in \Gamma$ such that $\gamma_n g_n \to g$ as $n \to \infty$.

We quickly verify the claim in 1.3, which shows that the orbit x_0U^- is the 'stable manifold' through $x_0 \in \mathsf{X}_2 = \mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R})$ for the geodesic flow. Indeed let $\Gamma = \mathrm{PSL}_2(\mathbb{Z}) < G = \mathrm{PSL}_2(\mathbb{R})$ and $x_0 = \Gamma g_0$ for $g_0 \in G$. Then we

[†] A metric is *proper* if any ball of finite radius has a compact closure.

have

$$\begin{aligned}\mathsf{d}_{\mathsf{X}_2}\big(a_t\cdot x_0, a_t u_s^- \cdot x_0\big) &= \mathsf{d}_{\mathsf{X}_2}\big(\Gamma g_0 a_t^{-1}, \Gamma g_0 u_{-s}^- a_t^{-1}\big)\\ &\leqslant \mathsf{d}_G\big(g_0 a_t^{-1}, g_0 u_{-s}^- a_t^{-1}\big)\\ &= \mathsf{d}_G\big(I, \underbrace{a_t u_{-s}^- a_t^{-1}}_{=u^-_{-\mathrm{e}^{-t}s}}\big) \longrightarrow 0\end{aligned}$$

as $t \to \infty$ as claimed.

Another consequence of the definition of this metric is that X and G are *locally isometric* in the following sense.

Lemma 1.17 (Injectivity radius). *Let Γ be a discrete subgroup in G (equipped with a left-invariant metric d_G as above). For any compact subset*

$$K \subseteq X = \Gamma\backslash G$$

there exists some $r = r(K) > 0$, called an injectivity radius on K*, with the property that for any $x_0 \in K$ the map*

$$B_r^G \ni g \longmapsto x_0 g \in B_r^X(x_0)$$

is an isometry between B_r^G and $B_r^X(x_0)$. If $K = \{x_0\}$ where $x_0 = \Gamma h$ for some $h \in G$, then

$$r = \tfrac{1}{4}\inf_{\gamma\in\Gamma\smallsetminus\{e\}} \mathsf{d}_G(h^{-1}\gamma h, e) \tag{1.8}$$

has this property.

Proof. We first show this locally, for $K = \{x_0\}$ where $x_0 = \Gamma h$. Let r be as in (1.8), which is positive since $h^{-1}\Gamma h$ is also a discrete subgroup. Then, for $g_1, g_2 \in B_r^G$,

$$\mathsf{d}_X(\Gamma h g_1, \Gamma h g_2) = \inf_{\gamma\in\Gamma} \mathsf{d}_G(h g_1, \gamma h g_2) = \inf_{\gamma\in\Gamma} \mathsf{d}_G(g_1, h^{-1}\gamma h g_2).$$

We wish to show that the infimum is achieved for $\gamma = e$. Suppose that $\gamma \in \Gamma$ has

$$\mathsf{d}_G(g_1, h^{-1}\gamma h g_2) \leqslant \mathsf{d}_G(g_1, g_2) < 2r.$$

Then

$$\mathsf{d}_G(h^{-1}\gamma h g_2, e) \leqslant \mathsf{d}_G(h^{-1}\gamma h g_2, g_1) + \mathsf{d}_G(g_1, e) < 3r$$

since $g_1 \in B_r^G$. Applying (1.6) we get

$$h^{-1}\gamma h = (h^{-1}\gamma h g_2) g_2^{-1} \in B_{3r}^G B_r^G \subseteq B_{4r}^G,$$

which implies that $\gamma = e$ by definition of r.

The lemma now follows by compactness of K. For x_0 and r as above it is easily checked that any $y \in B_{r/2}^X(x_0)$ satisfies the first claim of the proposition with r replaced by $r/2$. Hence K can be covered by balls so that on each ball there is a

uniform injectivity radius. Now take a finite subcover and the minimum of the associated injectivity radii. □

Notice that given an injectivity radius, any smaller number will also be an injectivity radius. We define *the maximal injectivity radius* r_{x_0} at $x_0 \in X$ as the supremum of the possible injectivity radii for the set $K = \{x_0\}$ (see also Exercise 1.25). If $x_0 = \Gamma h$ then

$$\tfrac{1}{4} \inf_{\gamma \in \Gamma \smallsetminus \{e\}} \mathsf{d}_G(h^{-1}\gamma h, e) \leqslant r_{x_0} \leqslant \inf_{\gamma \in \Gamma \smallsetminus \{e\}} \mathsf{d}_G(h^{-1}\gamma h, e) \tag{1.9}$$

by Lemma 1.17. We note that for the modular surface the maximal injectivity radius goes to 0 in the cusp.

We also define the natural quotient map

$$\begin{aligned} \pi_X \colon G &\longrightarrow X = \Gamma\backslash G \\ g &\longmapsto \Gamma g, \end{aligned}$$

and note that π_X is locally an isometry by left invariance of the metric and Lemma 1.17. Clearly $X = \Gamma\backslash G$ is a homogeneous space in the sense of algebra, but due to this local isometric property we will call X a *locally homogeneous space*.

One (rather abstract) way to understand the quotient space $X = \Gamma\backslash G$ may be to consider a subset $F \subseteq G$ for which the projection π_X, when restricted to F, is a bijection. This motivates the following definition.

Definition 1.18 (Fundamental domain). Let $\Gamma \leqslant G$ be a discrete subgroup. A *fundamental domain* $F \subseteq G$ for Γ is a measurable[†] set with the property that

$$G = \bigsqcup_{\gamma \in \Gamma} \gamma F,$$

(where $\bigsqcup$ denotes a disjoint union). Equivalently, $\pi_X|_F \colon F \to \Gamma\backslash G$ is a bijection. A measurable set $B \subseteq G$ will be called *injective (for Γ)* if $\pi_X|_B$ is an injective map, and *surjective (for Γ)* if $\pi_X(B) = \Gamma\backslash G$.

Example 1.19. The set $[0,1)^d \subseteq \mathbb{R}^d$ is a fundamental domain for the discrete subgroup $\Gamma = \mathbb{Z}^d \leqslant \mathbb{R}^d = G$.

The existence of a fundamental domain is a general property.

Lemma 1.20 (Existence of fundamental domains). *If Γ is a discrete subgroup of G and $B_{inj} \subseteq B_{surj} \subseteq G$ are measurable sets with B_{inj} injective and B_{surj} surjective, then there exists a fundamental domain F with $B_{inj} \subseteq F \subseteq B_{surj}$. Moreover, $\pi_X|_F \colon F \to X = \Gamma\backslash G$ is a bi-measurable[‡] bijection for any fundamental domain $F \subseteq G$.*

[†] Unless indicated otherwise, measurable always means Borel-measurable.

[‡] That is, both $\pi_X|_F$ and its inverse are measurable maps.

PROOF. Notice first that $\mathsf{d}_X(\pi_X(g_1),\pi_X(g_2)) \leqslant \mathsf{d}_G(g_1,g_2)$ for all $g_1,g_2 \in G$. Therefore, π_X is continuous (and hence measurable). Using the assumption that G is σ-compact and Lemma 1.17, we can find a sequence of sets (B_n) with $B_n = g_n B^G_{r_n}$ for $n \geqslant 1$ such that $\pi_X|_{B_n}$ is an isometry, and $G = \bigcup_{n=1}^\infty B_n$. It follows that for any Borel set $B \subseteq G$ the image $\pi_X(B \cap B_n)$ is measurable for all $n \geqslant 1$, and so $\pi_X(B)$ is measurable. This implies the final claim of the lemma.

Now let $B_{\mathrm{inj}} \subseteq B_{\mathrm{surj}} \subseteq G$ be as in the lemma. Define inductively the following measurable subsets of G:

$$\begin{aligned} F_0 &= B_{\mathrm{inj}}, \\ F_1 &= B_{\mathrm{surj}} \cap B_1 \smallsetminus \pi_X^{-1}\big(\pi_X(F_0)\big), \\ F_2 &= B_{\mathrm{surj}} \cap B_2 \smallsetminus \pi_X^{-1}\big(\pi_X(F_0 \cup F_1)\big), \end{aligned}$$

and so on. Then we will show that $F = \bigsqcup_{n=0}^\infty F_n$ satisfies all the claims of the lemma: Clearly F is measurable and $B_{\mathrm{inj}} \subseteq F \subseteq B_{\mathrm{surj}}$. If now $g \in G$ is arbitrary we need to show that $(\Gamma g) \cap F$ consists of a single element. If $\Gamma g = \pi_X^{-1}\big(\pi_X(g)\big)$ intersects B_{inj} nontrivially, then the intersection is a singleton by the assumption on B_{inj} and F_n will be disjoint to Γg for all $n \geqslant 1$ by construction. If Γg intersects B_{inj} trivially, then we choose $n \geqslant 1$ minimal such that Γg intersects $B_{\mathrm{surj}} \cap B_n$. By the properties of B_n this intersection is again a singleton, by minimality of n the point in the intersection also belongs to F_n, and Γg will intersect F_k trivially for $k > n$. Hence in all cases we conclude that $(\Gamma g) \cap F$ is a singleton, or equivalently F is a fundamental domain. □

In some special cases, for example $\mathbb{Z}^d < \mathbb{R}^d$, one can give concrete fundamental domains with better properties, where in particular the boundary of the fundamental domain consists of lower-dimensional objects. In those situations one could and should also ask about how the various pieces of the boundary are glued together under Γ. For instance, in the case of $\mathbb{Z}^d$ we know that opposite sides of $[0,1)^d$ are to be identified. Another such situation arose in the discussion in Section 1.2. As our goal is to consider more general quotients where this is typically not so easily done, we will not pursue this further.

Exercise 1.21. Let $\mathrm{GL}_d(\mathbb{R}) = \big\{g = (g_{i,j})_{i,j} \in \mathrm{Mat}_d(\mathbb{R}) \mid \det(g) \neq 0\big\}$, be the *general linear group* and let $G^o \leqslant \mathrm{GL}_d(\mathbb{R})$ be the connected component of the identity.
(a) For a continuous piecewise differentiable path $p\colon [0,1] \to G^o$ define its length by

$$L(p) = \int_0^1 \big\|p(t)^{-1}p'(t)\big\|\,\mathrm{d}t,$$

where $\|\cdot\|$ is a norm on $\mathrm{Mat}_d(\mathbb{R})$. Show that left translation does not change the length of a path.
(b) Define $\mathsf{d}_{G^o}(g_1,g_2)$ for $g_1,g_2 \in G^o$ to be the infimum of the lengths of all paths connecting g_1 and g_2. Show that d_{G^o} is a left-invariant metric giving the topology of G^o.
(c) Show that $G = \mathrm{GL}_d(\mathbb{R})$ embeds into $\mathrm{SL}_{d+1}(\mathbb{R}) < \mathrm{GL}_{d+1}(\mathbb{R})$. Conclude that G or any of its closed subgroups has a left-invariant metric giving its topology inherited from $\mathrm{Mat}_d(\mathbb{R})$.

Exercise 1.22. Show that a discrete subgroup $\Gamma < G$ is also closed.

Exercise 1.23. Let G be equipped with a left-invariant metric, and let Γ be a discrete subgroup of G. Show that

$$\mathsf{d}_X(x, xg) \leqslant \mathsf{d}_G(e, g)$$

for all $x \in X$ and $g \in G$, where as usual $X = \Gamma\backslash G$.

Exercise 1.24. Let $H < G$ be a closed subgroup. Imitate the definition in (1.7) to define a metric on $H\backslash G$. Show that $H\backslash G$ is locally compact and σ-compact (assuming, as always, that G is). Show that both G and $H\backslash G$ are complete as metric spaces.

Exercise 1.25. Show that the maximal injectivity radius as defined after Lemma 1.17 is indeed an injectivity radius. Show the upper bound in (1.9).

Exercise 1.26. Show that the topology induced by the metric $d_X(\cdot,\cdot)$ on $X = \Gamma\backslash G$ is the quotient topology of the topology on G for the natural map $\pi_X : G \to X$ (that is, the finest topology on X for which π_X is continuous).

1.3.2 Haar Measure and the Natural Action on the Quotient

Recall (see [45, Sec. 8.3] for an outline and [46, Sec. 10.1] or the monograph of Folland [52, Sec. 2.2] for a full proof) that any metric, σ-compact, locally compact group G has a (left) Haar measure m_G which is characterized (up to proportionality) by the properties

- $m_G(K) < \infty$ for any compact $K \subseteq G$;
- $m_G(O) > 0$ for any non-empty open set $O \subseteq G$;
- $m_G(gB) = m_G(B)$ for any $g \in G$ and measurable $B \subseteq G$.

Similarly there also exists a right Haar measure $m_G^{(r)}$ with the first two properties and invariance under right translation instead of left translation as above. For concrete examples it is often not so difficult to give an explicit description of the Haar measure, see Exercises 1.7 and 1.33.

Lemma 1.27 (Independence of choice of fundamental domain). *Let Γ be a discrete subgroup of G. Any two fundamental domains for Γ in G have the same left Haar measure. In fact, if $B_1, B_2 \subseteq G$ are injective sets for Γ with $\pi_X(B_1) = \pi_X(B_2)$ then*[†] *$m_G(B_1) = m_G(B_2)$.*

Alternatively we may phrase this lemma as follows. For any discrete subgroup $\Gamma < G$, the left Haar measure m_G induces a natural measure m_X on $X = \Gamma\backslash G$ such that

$$m_X(B) = m_G(\pi_X^{-1}(B) \cap F)$$

where $F \subseteq G$ is any fundamental domain for Γ in G. We also define the *covolume* $\mathrm{covol}_G(\Gamma)$ of Γ in G to be $m_G(F) = m_X(X)$.

[†] As the proof will show, we only need left-invariance of the measure under Γ. We will use this strengthening later.

Proof of Lemma 1.27. Suppose B_1 and B_2 are injective sets with

$$\pi_X(B_1) = \pi_X(B_2).$$

Then

$$B_1 = \bigsqcup_{\gamma\in\Gamma} B_1 \cap (\gamma B_2)$$

and

$$\bigsqcup_{\gamma\in\Gamma} \gamma^{-1}(B_1 \cap \gamma B_2) = \bigsqcup_{\gamma\in\Gamma} (\gamma B_1) \cap B_2 = B_2.$$

Note that the discrete subgroup $\Gamma < G$ must be countable as G is σ-compact. Therefore, we see that

$$m_G(B_1) = \sum_{\gamma\in\Gamma} m_G(B_1 \cap \gamma B_2) = \sum_{\gamma\in\Gamma} m_G\left(\gamma^{-1}B_1 \cap B_2\right) = m_G(B_2)$$

as required. □

Note that G acts naturally on $X = \Gamma\backslash G$ via right multiplication

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

for $x \in X$ and $g \in G$, and that this action satisfies

$$\pi_X(g_1g_2^{-1}) = \pi_X(g_1)g_2^{-1} = g_2{\cdot}\pi_X(g_1)$$

for all $g_1, g_2 \in G$. Also note that $g_2{\cdot}g_1 = g_1g_2^{-1}$ for $g_1 \in G$ is the natural action of $g_2 \in G$ on G on the right so that π_X satisfies the equivariance property $\pi_X(g_2{\cdot}g_1) = g_2{\cdot}\pi_X(g_1)$. We are interested in whether X supports a G-invariant probability measure, a property discussed in the next proposition and definition.

Proposition 1.28 (Finite volume quotients). *Let $\Gamma \leqslant G$ be a discrete subgroup. Then the following properties are equivalent:*

(a) *On $X = \Gamma\backslash G$ there exists a G-invariant probability measure, that is a probability measure m_X which satisfies $m_X(g{\cdot}B) = m_X(B)$ for all measurable $B \subseteq X$ and all g in G;*
(ã) *There is a fundamental domain $F \subseteq G$ which has finite right Haar measure $m_G^{(r)}(F) < \infty$ and $m_G^{(r)}$ is left Γ-invariant.*
(b) *There is a fundamental domain F for $\Gamma \leqslant G$ with $m_G(F) < \infty$;*

If any (and hence all) of these conditions hold, then G is unimodular *(that is, the Haar measure is bi-invariant).*

Definition 1.29 (Lattices). A discrete subgroup $\Gamma \leqslant G$ is called a *lattice* if $X = \Gamma\backslash G$ supports a G-invariant probability measure. In this case we also say that X has *finite volume*.

Given a fixed left Haar measure m_G on G, we can define the volume of X as $m_G(F)$ for any fundamental domain $F \subseteq G$ for Γ. Somewhat perversely, we will often normalize the Haar measure m_G to have $m_X(X) = m_G(F) = 1$. In the proof we will use the 'modular character' and the 'pigeonhole principle for ergodic theory'.

Right multiplication on G may not preserve the left Haar measure m_G. However, there is a continuous homomorphism, the *modular character*,

$$\mathrm{mod}\colon G \to \mathbb{R}_{>0}$$

with the property that $m_G(Bg) = \mathrm{mod}(g) m_G(B)$ for all measurable $B \subseteq G$ and $g \in G$ (see [46, Sec. 10.1] and [40, Sec. 1.2.3] for the details and references).

The modular character may also be defined using a right Haar measure $m_G^{(r)}$ via $m_G^{(r)}(g^{-1}B) = \mathrm{mod}(g) m_G^{(r)}(B)$ for all measurable $B \subseteq G$ and $g \in G$, and the left and right Haar measures may be normalized to have $m_G^{(r)}(B) = m_G(B^{-1})$ for any Borel set $B \subseteq G$, where $B^{-1} = \{g^{-1} \mid g \in B\}$.

The pigeonhole principle for ergodic theory is the *Poincaré recurrence theorem*, which may be formulated as follows in the metric setting. We refer to [45, Th. 2.21] and Exercise 1.34 for the proof.

Theorem 1.30 (Poincaré recurrence). *Let X be a locally compact σ-compact metric space, and let μ be a Borel probability measure invariant under a continuous map $T\colon X \to X$. Then for μ-almost every $x \in X$ there is a sequence $n_k \to \infty$ with $T^{n_k} x \to x$ as $k \to \infty$.*

Proof of Proposition 1.28. We will start by proving that (a) implies (ã). Suppose therefore that m_X is a probability measure on $X = \Gamma\backslash G$ invariant under the action of G on the right. Then we can define a measure μ on G via the Riesz representation theorem by letting

$$\int f \,\mathrm{d}\mu = \int \sum_{\pi_X(g)=x} f(g) \,\mathrm{d}m_X(x) \tag{1.10}$$

for any measurable $f \geqslant 0$. For $g \in G$ we may use Lemma 1.17 to find an injectivity radius $r > 0$, set $f = \mathbb{1}_{gB_r^G}$, and obtain $\mu\big(gB_r^G\big) \leqslant 1$. Therefore μ is locally finite.

By invariance of m_X under the action of G, we see that $\mu = m_G^{(r)}$ is a right Haar measure on G (the reader may check all the characterizing properties of Haar measures from page 23, or rather their analogues for right Haar measures). By the construction above, $m_G^{(r)}$ is left-invariant under Γ. Applying the identity (1.10) to $f = \mathbb{1}_F$ for a fundamental domain $F \subseteq G$ shows that $m_G^{(r)}(F) = 1$, and hence (ã).

Now suppose that (ã) holds, and let F be the fundamental domain. We define a measure m_X on X by

$$m_X(B) = \frac{1}{m_G^{(r)}(F)} m_G^{(r)}\left(F \cap \pi_X^{-1}(B)\right). \tag{1.11}$$

By Lemma 1.27 (and its footnote), this definition is independent of the particular fundamental domain used. Thus for $g \in G$ and $B \subseteq X$ we have

$$\begin{aligned} m_X(Bg) &= \frac{1}{m_G^{(r)}(F)} m_G^{(r)}\left(F \cap \pi_X^{-1}(Bg)\right) \\ &= \frac{1}{m_G^{(r)}(F)} m_G^{(r)}\left(F \cap \pi_X^{-1}(B)g\right) \\ &= \frac{1}{m_G^{(r)}(Fg^{-1})} m_G^{(r)}\left(Fg^{-1} \cap \pi_X^{-1}(B)\right) = m_X(B), \end{aligned}$$

since $Fg^{-1} \subseteq G$ is also a fundamental domain. This shows (a). It follows that (a) and ($\widetilde{\text{a}}$) are equivalent.

We also note that (b) implies ($\widetilde{\text{a}}$) rather quickly: If F is a fundamental domain with $m_G(F) < \infty$ and $g \in G$, then Fg is another fundamental domain. Therefore, by Lemma 1.27, $m_G(F) = m_G(Fg) = m_G(F)\bmod(g)$, so G is unimodular and ($\widetilde{\text{a}}$) follows.

In the proof that (a) (or, equivalently, ($\widetilde{\text{a}}$)) implies (b), we will again show that G is unimodular. So let m_X be a finite G-invariant measure on X. We may suppose that m_X is a probability measure. Let $r > 0$ be an injectivity radius at Γe and $B \subseteq B_r^G$ a compact neighbourhood of e. By invariance of m_X and transitivity of the G-action on X, we have $\operatorname{supp} m_X = X$ and so $m_X(\Gamma B) > 0$.

Let now g be an element of G; we wish to show that $\bmod(g) = 1$. By Poincaré recurrence (Theorem 1.30) there exists some $b \in B$ and sequences $(n_k), (\gamma_k), (b_k)$ with $n_k \nearrow \infty$ as $k \to \infty$, $\gamma_k \in \Gamma$ for all $k \geqslant 1$, and $b_k \in B$ for all $k \geqslant 1$ such that

$$bg^{-n_k} = \gamma_k b_k$$

for all $k \geqslant 1$. Applying the modular character, and noticing that

$$\bmod(\Gamma) = \{1\}$$

by ($\widetilde{\text{a}}$), we see that

$$\bmod(g)^{n_k} = \frac{\bmod(b)}{\bmod(b_k)}$$

belongs to a compact neighbourhood of $1 \in (0, \infty)$ for all $k \geqslant 1$. It follows that $\bmod(g) = 1$, as required. □

Proposition 1.31 (Haar measure on $X = \Gamma\backslash G$). *Let G be unimodular, Γ a discrete subgroup of G, and $X = G/\Gamma$. Then the Haar measure m_G on G induces a locally finite G-invariant measure m_X on X satisfying*

$$\int_G f \,\mathrm{d}m_G = \int_X \sum_{\gamma \in \Gamma} f(\gamma g)\,\mathrm{d}m_X(\Gamma g) \tag{1.12}$$

for all $f \in L^1_{m_G}(G)$.

The formula (1.12) is sometimes referred to as *folding* (if used from the left-hand side to the right-hand side), or *unfolding* (if used in the other direction). The measure m_X is called the *Haar measure*, the *uniform measure*, or the *volume measure* on X.

Proof of Proposition 1.31. Since we assume that G is unimodular, the argument that ($\widetilde{\text{a}}$) implies (a) in the proof of Proposition 1.28 can be used to define the measure m_X. Lemma 1.27 shows that m_X is independent of the choice of fundamental domain $F \subseteq G$ used in the definition, and shows that m_X is G-invariant. By the definition in (1.11), (1.12) holds for $f = \mathbb{1}_B$ if $B \subseteq \gamma F$ for some $\gamma \in \Gamma$. By linearity (1.12) also holds for any measurable $B \subseteq G$ and hence for any simple function. Finally, monotone convergence implies that (1.12) holds for any measurable non-negative function. □

Notice that Lemma 1.17 implies that any compact set $K_X \subseteq X$ is the image $K_X = \pi_X(K_G)$ of a compact set $K_G \subseteq G$. In particular, this implies that a compact quotient $\Gamma\backslash G$ is of finite volume in the sense of Definition 1.29.

Definition 1.32 (Uniform lattice). A discrete subgroup $\Gamma \leqslant G$ is called a (*co-compact* or) *uniform lattice* if the quotient space $X = \Gamma\backslash G$ is compact.

A consequence of this definition and Lemma 1.17 is that there is a choice of injectivity radius that is *uniform* across all of $\Gamma\backslash G$, which should help to explain the terminology of 'uniform lattice'. Roughly speaking, $\Gamma \leqslant G$ is a uniform lattice if the quotient space $\Gamma\backslash G$ is small topologically (compact) as well as measurably (of finite volume). At first sight, motivated by the abelian paradigm from $\mathbb{Z}^d \leqslant \mathbb{R}^d$, it seems reasonable to require that $\Gamma\backslash G$ should always be compact in defining a lattice. However, as discussed in Section 1.2, this would exclude some of the most natural lattices and their quotient spaces.

Exercise 1.33. Show that the bi-invariant Haar measure $m_{\mathrm{GL}_d(\mathbb{R})}$ on the locally compact group $\mathrm{GL}_d(\mathbb{R})$ can be defined by the formula

$$\mathrm{d}m_{\mathrm{GL}_d(\mathbb{R})}(g) = \frac{\prod_{i,j=1}^d \mathrm{d}g_{i,j}}{(\det g)^d}.$$

Exercise 1.34. Show that Theorem 1.30 follows from the usual formulation of Poincaré recurrence: If $(X, \mathscr{B}, \mu, T)$ is a measure-preserving system and $A \in \mathcal{B}$ has $\mu(A) > 0$ then there is some $n \geqslant 1$ for which $\mu(A \cap T^{-n}A) > 0$ (see [45, Sec. 2.1]).

1.3.3 Quotients Consisting of Left Cosets

As is common in geometry (see Section 1.2) and number theory we have so far considered quotients consisting of right cosets of the form $X = \Gamma\backslash G$. As discussed, in this setup one uses a left-invariant metric to define a metric on X and defines the action of G by $g \cdot x = xg^{-1}$ for $g \in G$ and $x \in X = \Gamma\backslash G$.

For the study of dynamical questions it is more natural or conventional to consider quotients of the form $X = G/\Gamma$ consisting of left cosets. Of course our discussion is equally valid for this setup. Here one would consider a right-invariant metric to define the metric on X and define the action of G on X by $g \cdot x = gx$ for $g \in G$ and $x \in X = G/\Gamma$.

It is easy to see that the map

$$\Gamma\backslash G \ni \Gamma g \longmapsto g^{-1}\Gamma \in G/\Gamma$$

gives a natural isomorphism between the two setups. We will make use of both of the two setups freely.

1.3.4 Divergence in the Quotient by a Lattice

In allowing non-compact quotients, it is natural to ask how compact subsets of $X = \Gamma\backslash G$ can be described or, equivalently, to characterize sequences (x_n) in X that *go to infinity* (that is, leave any compact subset of X).

Proposition 1.35 (Abstract divergence criterion). *Let $\Gamma < G$ be a lattice. Then the following properties of a sequence (x_n) in $X = \Gamma\backslash G$ are equivalent:*

(1) *$x_n \to \infty$ as $n \to \infty$, meaning that for any compact set $K \subseteq X$ there is some $N = N(K) \geqslant 1$ such that $n \geqslant N$ implies that $x_n \notin K$.*
(2) *The maximal injectivity radius at $x_n = \Gamma g_n$ goes to zero as $n \to \infty$. That is, there exists a sequence (γ_n) in $\Gamma \smallsetminus \{e\}$ such that $g_n^{-1}\gamma_n g_n \to e \in G$ as $n \to \infty$.*

Proof. We note that the two statements in (2) are equivalent due to (1.9).

Suppose that (1) holds, so that $x_n \to \infty$ as $n \to \infty$. We need to show that the maximal injectivity radius r_{x_n} at x_n goes to zero. So suppose the opposite, then we would have $r_{x_n} \geqslant \varepsilon > 0$ for some $\varepsilon > 0$ and infinitely many n, and by choosing this subsequnce we may assume without loss of generality that $r_{x_n} \geqslant \varepsilon > 0$ for all $n \geqslant 1$.

Decreasing ε if necessary, we may assume that $\overline{B_\varepsilon^G}$ is compact (since G is locally compact). Therefore, and by our assumption in (1) there is some N_1 with

$$x_n \notin x_1 \overline{B_\varepsilon^G}$$

for $n \geqslant N_1$. Now remove the terms $x_2, \ldots, x_{N_1 - 1}$ from the sequence. Similarly, there is an $N_2 \geqslant 1$ with

$$x_n \notin x_1 \overline{B_\varepsilon^G} \cup x_{N_1} \overline{B_\varepsilon^G}$$

for $n \geqslant N_2$. Repeating this process infinitely often, and renaming the thinned-out sequence remaining (x_n) again, we may assume without loss of generality that $\mathsf{d}(x_n, x_m) \geqslant \varepsilon$ for all $m \neq n$. This now gives a contradiction to the assumption that X has finite volume: If $x_n = \pi_X(g_n)$ then

$$X \supseteq \bigsqcup_{n=1}^{\infty} x_n B^G_{\varepsilon/2} = \Gamma \left(\bigsqcup_{n=1}^{\infty} g_n B^G_{\varepsilon/2} \right),$$

and

$$\bigsqcup_{n=1}^{\infty} g_n B^G_{\varepsilon/2}$$

is a disjoint union of infinite measure, and is an injective set.

Suppose now that (1) does not hold, so there exists some compact $K \subseteq X$ with $x_n \in K$ for infinitely many n. By Lemma 1.17 there exists an injectivity radius $r > 0$ on K and we see that $r_{x_n} \geqslant r$ for infinitely many n, so that (2) does not hold either. □

1.3.5 Orbits of Subgroups

In the following we will also be interested in orbits of subgroups $H \leqslant G$. Given an action of G on a space X the *H-orbit* of $x \in X$ is the set

$$H{\cdot}x = \{h{\cdot}x \mid h \in H\} \cong {}^{H}/\operatorname{Stab}_H(x) \cong \operatorname{Stab}_H(x)\backslash H,$$

where

$$\operatorname{Stab}_H(x) = \{h \in H \mid h{\cdot}x = x\}$$

is the *stabilizer subgroup* of $x \in X$ and the isomorphisms are sending $h{\cdot}x$ to $h \operatorname{Stab}_H(x)$ resp. to $\operatorname{Stab}_H(x)h^{-1}$. Note that if $X = \Gamma\backslash G$ and $x = \Gamma g$, then

$$\operatorname{Stab}_H(x) = H \cap g^{-1}\Gamma g$$

is a discrete subgroup of H. Fixing a Haar measure m_H on H we define the *volume of the H-orbit*, $\operatorname{vol}(H{\cdot}x)$ to be $m_H(F_H)$ where $F_H \subseteq H$ is a fundamental domain for $\operatorname{Stab}_H(x)$ in H.

Clearly if an H-orbit $xH \subseteq X = \Gamma\backslash G$ is compact, it is also closed. In fact the same conclusion can be reached for finite volume orbits.

Corollary 1.36 (Finite volume orbits are closed). *Let $\Gamma \leqslant G$ be a discrete subgroup, and let $H \leqslant G$ be a closed subgroup. For any $x \in X = \Gamma\backslash H$ the orbit map*

$$\operatorname{Stab}_H(x)\backslash H \ni \operatorname{Stab}_H(x)h \longmapsto xh \in X \tag{1.13}$$

is continuous. If xH has finite volume, then the orbit xH is closed in X and the orbit map (1.13) *is a proper homeomorphism.*

We note that Corollary 1.36 can also be shown directly (see Figure 1.9). However, it is also a quick corollary of Proposition 1.35.

Proof of Corollary 1.36. Let $x = \Gamma g$ and let $\Lambda = \operatorname{Stab}_H(x) = H \cap g^{-1}\Gamma g$. Suppose $\Lambda h_n \to \Lambda h$ in $\operatorname{Stab}_H(x)\backslash H$ as $n \to \infty$. Then there exists a sequence (γ_n)

in Γ so that $g^{-1}\gamma_n g h_n \to h$ as $n \to \infty$ in H (and so also in G), which already implies that $\Gamma g h_n \to \Gamma g h$ as $n \to \infty$ in X.

Next we show properness of the orbit map. Suppose therefore that $\Lambda h_n \to \infty$ as $n \to \infty$ for a sequence in $Y = \Lambda\backslash H$. Since Y has finite volume, we may apply Proposition 1.35 to H to see that there exists a sequence (λ_n) in Λ such that

$$h_n^{-1}\lambda_n h_n \longrightarrow e$$

as $n \to \infty$. Using $\mathrm{Stab}_H(x) = g^{-1}\Gamma g \cap H$ we have $\lambda_n = g^{-1}\gamma_n g$ for some sequence (γ_n) in Γ, and

$$h_n^{-1}g^{-1}\gamma_n g h_n = h_n^{-1}\lambda_n h_n \longrightarrow e$$

as $n \to \infty$. By Lemma 1.17 this shows that $\Gamma g h_n \to \infty$ in $X = \Gamma\backslash G$ as $n \to \infty$. Since $\big(\Lambda h_n\big)$ was an arbitrary sequence in Y going to infinity, the properness of the orbit map from Y to X follows.

Together the above shows that the orbit map in (1.13) extends continuously from the one-point compactification of $Y = \mathrm{Stab}_H(x)\backslash H$ to the one-point compactification of X by sending ∞ to ∞. In particular, $\overline{xH} = xH \cup \{\infty\}$, which shows that xH is closed in X. Moreover, as a continuous injective map has a continuous inverse, we see that the orbit map is a homeomorphism onto its image. □

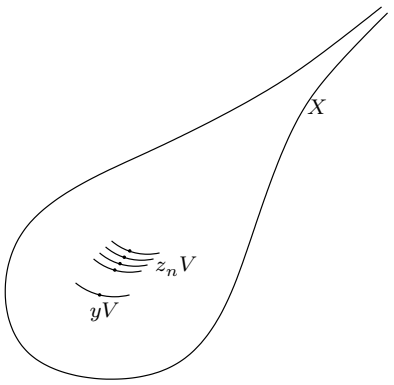

Fig. 1.9: We depict here an alternative to the proof of Corollary 1.36: By assuming (for the purposes of a contradiction) that the sets $z_nV \subseteq xH$ approach $yV \subseteq xH$ transverse to the orbit direction for a given neighbourhood V of $e \in H$, one can show that $\mathrm{vol}(xH) = \infty$.

Clearly if we are interested in finding finite volume H-orbits (that will carry finite H-invariant measures), then we need to restrict to unimodular subgroups $H \leqslant G$ (by Proposition 1.28). If H is unimodular (and, as before, we have fixed some Haar measure m_H) then the *volume measure* vol_{xH} on the H-orbit

is defined by

$$\mathrm{vol}_{xH}(B) = m_H\left(\{h \in F \mid xh \in B\}\right)$$

where $F \subseteq H$ is a fundamental domain for $\mathrm{Stab}_H(x)$ in H. This measure may be finite or infinite (and in the latter case it may be locally finite considered on X or not), but is always invariant under the right action of H due to Proposition 1.31 applied to $\mathrm{Stab}_H(x)\backslash H \cong xH$.

Proposition 1.37 (Closed orbits are embedded). *Let $\Gamma \leqslant G$ be a discrete subgroup and let $H \leqslant G$ be a closed subgroup. Suppose that $x \in X = \Gamma\backslash G$ has a closed H-orbit. Then $xH \subseteq X$ is* embedded*, meaning that the map*

$$\mathrm{Stab}_H(x)\backslash H \ni \mathrm{Stab}_H(x)h \longmapsto xh \in xH \tag{1.14}$$

is a homeomorphism. In particular, if H is unimodular then vol_{xH} is a locally finite measure on X.

We postpone the proof to the end of the next subsection.

1.3.6 Duality

Let $H < G$ be again a closed subgroup. Using a left-invariant metric d_G we can define a metric on the quotient $H\backslash G$ by

$$\mathsf{d}_{H\backslash G}(Hg_1, Hg_2) = \inf_{h_1,h_2\in H} \mathsf{d}_G(h_1g_1, h_2g_2) = \inf_{h\in H} \mathsf{d}_G(hg_1, g_2) \tag{1.15}$$

for $Hg_1, Hg_2 \in H\backslash G$ as in (1.7). We note that unless H is discrete there is no notion of injectivity radius. However, assuming that G is locally compact and σ-compact, the quotient is locally compact, σ-compact, and complete. We leave these claims as an exercise (see Exercise 1.24). Using again the 'inverse isomorphism'

$$H\backslash G \ni Hg \longmapsto g^{-1}H \in G/H,$$

this also applies to G/H.

Now let $\Gamma < G$ be a discrete (or, more generally, a closed) subgroup and let $H < G$ be a closed subgroup. In many ways the dynamics of H acting on $\Gamma\backslash G$ is strongly related to the dynamics of Γ on G/H. We only start this line of thought with the following topological observation.

Proposition 1.38 (Topological duality). *Let $\Gamma, H < G$ be closed subgroups and let $g_0 \in G$. Then the following are equivalent:*

(1) *The H-orbit of $\Gamma g_0 \in \Gamma\backslash G$ is closed.*
(2) *The set $\Gamma g_0 H \subseteq G$ is closed.*
(3) *The Γ-orbit of $g_0H \in G/H$ is closed.*

If Γ is discrete and the above holds true, then in fact the Γ-orbit of $g_0H \in G/H$ is also a discrete subset of G/H.

Proof. We start the proof with a more general statement. Let $Y = G/H$ and write $\pi_Y \colon G \ni g \mapsto gH \in G/H$ for the canonical projection map. Now let $B \subseteq G$ be a union of left cosets so that $B = \pi_Y^{-1}\pi_Y(B)$. We claim that B is closed as a subset of G if and only if $\pi_Y(B) = \{bh \mid b \in B\}$ is closed as a subset of $Y = G/H$.

As π_Y is continuous we see that $\pi_Y(B)$ being closed implies that B is closed.

So suppose now that B is closed and a sequence (b_n) in B and $g \in G$ have the property that $b_n H \to gH \in \overline{\pi_Y(B)}$ as $n \to \infty$. By definition of the metric d_Y in (1.15) this implies that there exists some (h_n) in H with $b_n h_n \to g$ as $n \to \infty$ in G. By assumption $b_n h_n \in B$ for all $n \in \mathbb{N}$ and hence $g \in \overline{B} = B$, which gives $gH \in \pi_Y(B)$. It follows that $\pi_Y(B)$ is closed.

The claim applied to $B = \Gamma g_0 H$ implies the equivalence of (2) and (3). However, applying the inverse isomorphism $\Gamma g \to g^{-1}\Gamma$ shows that the claim holds equally well for the quotient $\Gamma\backslash G$ and subsets of G that are unions of right Γ-cosets. This gives the equivalence of (1) and (2).

Suppose Γ is discrete. It remains to show that the Γ-orbit of $y_0 = g_0 H$ is a discrete subset of $Y = G/H$. If the orbit is not discrete, then we may choose a sequence (η_n) in Γ so that $\eta_n y_0 \to gH$ as $n \to \infty$ for some g in G, but $\eta_n y_0 \neq gH$ for $n \geqslant 1$. Then $gH = \eta y_0$ for some $\eta \in \Gamma$ as the Γ-orbit is closed. Multiplying on the left by $\gamma\eta^{-1}$ for an arbitrary $\gamma \in \Gamma$ gives a sequence in $\Gamma y_0 \subseteq Y$ with limit γy_0 such that the limit is not achieved in the sequence. This shows that any element of Γy_0 is an accumulation point of Γy_0 (that is, Γy_0 is a closed *perfect subset*[5] of $Y = G/H$). As Γ is countable (since G is σ-compact) we can write $\Gamma y_0 = \{\gamma_1 y_0, \gamma_2 y_0, \dots\}$. Now $O_n = \Gamma y_0 \smallsetminus \{\gamma_n y_0\}$ is an open dense subset of Γy_0 for any $n \geqslant 1$, which implies by the Baire category theorem that $\bigcap_{n\geqslant 1} O_n$ must be dense in Γy_0. This gives a contradiction as the intersection is empty. □

Proof of Proposition 1.37. By Corollary 1.36 the map in (1.14) is continuous. We wish to show that its inverse is also continuous.

Let $x = \Gamma g$ for $g \in G$. Now suppose that $\Gamma g h_n \to \Gamma g h$ in $\Gamma\backslash G$ as $n \to \infty$. Then there exists a sequence (γ_n) in Γ with

$$\gamma_n g h_n \longrightarrow gh \in gH \tag{1.16}$$

as $n \to \infty$, which implies that

$$\gamma_n gH \longrightarrow gH$$

in G/H as $n \to \infty$. By the discreteness of the Γ-orbit of gH in Proposition 1.38, it follows that $\gamma_n gH = gH$ for large enough n. Equivalently, $g^{-1}\gamma_n g \in \mathrm{Stab}_H(x)$ for large enough $n \in \mathbb{N}$ and

$$\mathrm{Stab}_H(x) h_n \longrightarrow \mathrm{Stab}_H(x) h$$

as $n \to \infty$ in $\mathrm{Stab}_H(x)\backslash H$ by (1.16).

For the last claim of the proposition let $K \subseteq X$ be a compact subset so that $K \cap xH$ is compact also. By the first part of the proposition $K \cap xH$ corresponds to a compact subset in $\mathrm{Stab}_H(x)\backslash H$ and so has finite measure. □

Exercise 1.39. Let $G \subseteq \mathrm{SL}_d(\mathbb{R})$ be a closed linear group, and let

$$\Gamma = G \cap \mathrm{SL}_d(\mathbb{Z}) < G$$

be a non-uniform lattice in G. Show that Γ must contain a unipotent matrix (that is, a matrix for which 1 is the only eigenvalue). We note that this is true in general, as conjectured by Selberg and proved by Každan and Margulis [78]; also see Raghunathan [121, Ch. XI]. However, the proof for subgroups of the form $\Gamma = G \cap \mathrm{SL}_d(\mathbb{Z})$ is significantly easier.

Exercise 1.40. Let $\Gamma < G$ be a uniform lattice in a connected σ-compact locally compact group G equipped with a proper left-invariant metric. Show that Γ is finitely generated. This again holds more generally, but for connected groups and for compact quotients the proof is straightforward; we refer to Raghunathan [121, Remark 13.21] for the general case.

Exercise 1.41. Let $\Gamma < G$ be a discrete subgroup, let $x \in X = \Gamma\backslash G$, and let H_1, H_2 be two closed subgroups of G for which xH_1 and xH_2 are closed orbits. Prove that

$$x(H_1 \cap H_2) \subseteq (xH_1) \cap (xH_2)$$

is a closed orbit.

Exercise 1.42. Let $\Gamma < G$ be a discrete, and $H < G$ a closed, subgroup of G. Recall that a dynamical system is called *topologically transitive* if there exists a dense orbit, and is called *minimal* if every orbit is dense. Show that the action of H on $\Gamma\backslash G$ is topologically transitive (or minimal) if and only if the action of Γ on G/H is topologically transitive (or minimal).

1.4 The Space X_d of Lattices in $\mathbb{R}^d$

In this section we will introduce the most important locally homogeneous space for ergodic theory and its connections to number theory, namely the space X_d consisting of all lattices in $\mathbb{R}^d$ with covolume one. Moreover, this space will give rise to other arithmetical quotients by looking at orbits of subgroups of $\mathrm{SL}_d(\mathbb{R})$ on X_d. Such orbits will be discussed starting in Chapter 3.

1.4.1 Basic Definitions

Let $d \in \mathbb{N}$. A lattice in $\mathbb{R}^d$ in the sense of Definition 1.29 has the form $\Lambda = g\mathbb{Z}^d$ for some $g \in \mathrm{GL}_d(\mathbb{R})$ (see Exercise 1.43). A fundamental domain for Λ is given by the parallelepiped $g[0,1)^d$ which is spanned by the column vectors of g, and has Lebesgue measure $|\det g|$. A lattice $\Lambda \subseteq \mathbb{R}^d$ is called *unimodular* if the covolume is 1. The space of all unimodular lattices in $\mathbb{R}^d$—the *moduli space of lattices*—is therefore

$$\mathsf{X}_d = \{g\mathbb{Z}^d \mid g \in \mathrm{SL}_d(\mathbb{R})\}, \tag{1.17}$$

which is the orbit of $\mathbb{Z}^d$ under the action of $\mathrm{SL}_d(\mathbb{R})$ on the subsets of $\mathbb{R}^d$: For $B \subseteq \mathbb{R}^d$ and $g \in \mathrm{SL}_d(\mathbb{R})$ the action sends (g, B) to

$$gB = \{gv \mid v \in B\}.$$

Notice that

$$\mathrm{Stab}_{\mathrm{SL}_d(\mathbb{R})}(\mathbb{Z}^d) = \mathrm{SL}_d(\mathbb{Z}),$$

so that

$$\mathsf{X}_d = \mathrm{SL}_d(\mathbb{R})/\mathrm{SL}_d(\mathbb{Z}) \tag{1.18}$$

where $g\,\mathrm{SL}_d(\mathbb{Z})$ corresponds to the lattice $g\mathbb{Z}^d$. We will think of this isomorphism as an equality. In particular, the topology, the action of $G = \mathrm{SL}_d(\mathbb{R})$, and the Haar measure on X_d are as discussed in Section 1.3. To understand X_d better, we need to develop a better understanding of lattices in $\mathbb{R}^d$.

Thinking of $\mathbb{R}^d$ as the space of column vectors leads naturally to the quotient $\mathrm{SL}_d(\mathbb{R})/\mathrm{SL}_d(\mathbb{Z})$ consisting of left cosets. If one worked instead with the row vectors, this would lead naturally to the quotient consisting of right cosets.

To obtain the natural isomorphism indicated by (1.17)–(1.18) we study lattices in $\mathbb{R}^d$. However, for most of our discussion we could equally well study lattices in a d-dimensional Euclidean vector space.

Exercise 1.43. Check that any lattice in $\mathbb{R}^d$ (in the sense of Definition 1.29) is indeed of the form $g\mathbb{Z}^d$ for some $g \in \mathrm{GL}_d(\mathbb{R})$. Also show that for $v_1, \dots, v_d \in \mathbb{R}^d$ either

$$\Lambda = \mathbb{Z}v_1 + \cdots + \mathbb{Z}v_d$$

is a lattice, or for every $\varepsilon > 0$ there exists a non-zero integer vector $(n_1, \dots, n_d) \in \mathbb{Z}^d$ with

$$\|n_1 v_1 + \cdots + n_d v_d\| < \varepsilon.$$

1.4.2 Geometry of Numbers

The next result will be almost immediate from the abstract results in Section 1.4.1. It is a weak form of a classical result due to Minkowski in 1896 (see [111] for a modern reprinting).

Theorem 1.44 (Minkowski's first theorem). *If $\Lambda \subseteq \mathbb{R}^d$ is a lattice of covolume V, then there exists a non-zero vector in Λ of length $\ll \sqrt[d]{V}$, with the implicit constant depending only on d.*

Recall that $f \ll g$ if there is a constant $C > 0$ with $f \leqslant Cg$, and $f \asymp g$ if $f \ll g$ and $g \ll f$; where the constant depends on other parameters these will often appear as subscripts as, for example, in the obvious bound

$$|\Lambda \cap B_1^{\mathbb{R}^d}(0)| \ll_\Lambda 1.$$

Since we will not be varying d throughout any of our discussions, we will not indicate dependencies on d in this way. We use this notation here as the par-

ticular value of the constants appearing in Theorems 1.44 and 1.45 will not be important for our purposes.

PROOF OF THEOREM 1.44. Choose $r_d > 0$ so that $B_{r_d}^{\mathbb{R}^d}(0)$ has Lebesgue measure 2 (any measure exceeding 1 will do). Then $\sqrt[d]{V}B_{r_d}^{\mathbb{R}^d}(0)$ has measure $2V$, and so cannot be an injective domain in the sense of Definition 1.18. It follows that there must exist $x_1 \neq x_2$ in $\sqrt[d]{V}B_{r_d}^{\mathbb{R}^d}(0)$ with $x_1 - x_2 = \lambda \in \Lambda \smallsetminus \{0\}$ of length $\|\lambda\| \leqslant 2r_d\sqrt[d]{V}$. □

A typical goal of *lattice reduction theory* is to develop algorithms that start with a set of generators of a lattice and efficiently produce a different set of generators that are short and almost orthogonal. We note that the three attributes of efficiency, shortness, and close to orthogonality are in tension—and hence the subject is an intricate one.(6) The minima defined below are sometimes referred to as *Minkowski's successive minima*.

Theorem 1.45 (Successive minima). *Let $\Lambda \subseteq \mathbb{R}^d$ be a lattice. We define the successive minima of Λ by*

$$\lambda_k(\Lambda) = \min\{r \mid \Lambda \text{ contains } k \text{ linearly independent vectors of norm} \leqslant r\}$$

for $k = 1, \ldots, d$. Then

$$\lambda_1(\Lambda)\cdots\lambda_d(\Lambda) \asymp \operatorname{covol}(\Lambda). \tag{1.19}$$

Moreover, if

$$\alpha_k(\Lambda) = \min\{\operatorname{covol}(\Lambda \cap V) \mid V \subseteq \mathbb{R}^d \text{ is a subspace of dimension } k\},$$

then

$$\alpha_k(\Lambda) \asymp \lambda_1(\Lambda)\cdots\lambda_k(\Lambda)$$

for $1 \leqslant k \leqslant d$.

To envision the successive minimas $\lambda_1(\Lambda), \ldots, \lambda_d(\Lambda)$ consider the ball $B_r^{\mathbb{R}^d}$. For $r > 0$ close to 0 we have $\overline{B_r^{\mathbb{R}^d}} \cap \Lambda = \{0\}$. By increasing r we may find more and more linearly independent vectors. The successive minima record those radii for which the dimension of the linear hull of the intersection increases.

For a subspace $V \subseteq \mathbb{R}^d$ there are two possibilities: Either $V \cap \Lambda$ spans V or it does not. In the first case $\Lambda \cap V$ is a lattice in V, we say that V is *Λ-rational*, and the covolume $\operatorname{covol}_V(\Lambda \cap V)$ of $\Lambda \cap V$ in V is finite. In the second case, we write $\operatorname{covol}_V(\Lambda \cap V) = \infty$. Strictly speaking we have to mention how we are normalizing the Haar measures of the different subspaces $V \subseteq \mathbb{R}^d$. However, we do this as one would expect: The Euclidean norm on $\mathbb{R}^d$ induces a Euclidean norm on V by restriction which in turn induces the Haar measure on V such that a unit cube in V has volume one. We note that the minimum in the definition of $\alpha_k(\Lambda)$ is indeed achieved for any k, see Exercise 1.49.

The proof of Theorem 1.45 is geometric, and relies on starting with a shortest vector (of size $\lambda_1(\Lambda)$) and then extending it with other vectors, chosen to be almost orthogonal to obtain a basis of $\mathbb{R}^d$.

Proof of Theorem 1.45. We use induction on the dimension d to prove (1.19). For $d = 1$ it is clear that $\Lambda = \mathbb{R}v_1$ for some $v_1 \in \mathbb{R}\smallsetminus\{0\}$ and

$$\lambda_1(\Lambda) = \|v_1\| = \operatorname{covol}(\Lambda).$$

Assume therefore that (1.19) holds for $d-1$, and let $\Lambda \subseteq \mathbb{R}^d$ be a lattice. It is clear by definition that

$$\lambda_1(\Lambda) \leqslant \lambda_2(\Lambda) \leqslant \cdots \leqslant \lambda_d(\Lambda).$$

Pick a vector $v_1 \in \Lambda$ of length $\lambda_1(\Lambda)$, and define $W = (\mathbb{R}v_1)^\perp \subseteq \mathbb{R}^d$. Also let $\pi\colon \mathbb{R}^d \to W$ be the orthogonal projection along $\mathbb{R}v_1$ onto W.

First preparatory step: Discreteness. We claim that $\Lambda_W = \pi(\Lambda) \subseteq W$ is a discrete subgroup in W with the property that all of its non-zero vectors have length $\gg \lambda_1(\Lambda)$, or in symbols that $\lambda_1(\Lambda_W) \gg \lambda_1(\Lambda)$.

To see the claim, assume for the purpose of a contradiction that

$$w = \pi(v) \in \Lambda_W\smallsetminus\{0\}$$

has length less than $\frac{\sqrt{3}}{2}\|v_1\|$. Here $v = w + tv_1 \in \Lambda$ for some $t \in \mathbb{R}$, and we may assume (by replacing $v \in \Lambda$ with $v + nv_1 \in \Lambda$ for a suitable $n \in \mathbb{Z}$) that $t \in [-\frac{1}{2}, \frac{1}{2})$. However, since v_1 and w are orthogonal by construction, this implies that

$$\|v\|^2 = \|w\|^2 + t^2\|v_1\|^2 < \tfrac{3}{4}\|v_1\|^2 + \tfrac{1}{4}\|v_1\|^2 = \|v_1\|^2,$$

which contradicts the choice of v_1 as a non-zero vector in Λ of smallest length.

Second preparatory step: Lattice Property. Next we claim that Λ_W is a lattice. To see this, consider a fundamental domain F_W for Λ_W inside W. Then $F = [0,1)v_1 + F_W$ is a fundamental domain for Λ. Indeed, for any $x \in \mathbb{R}^d$ there is a unique $w \in \Lambda_W = \pi(\Lambda)$ with

$$y = \pi(x) - w \in F_W.$$

Choosing $v \in \Lambda$ with $\pi(v) = w$, this shows that $x - v - y \in \mathbb{R}v_1$, and there exists a unique $n \in \mathbb{Z}$ and $t \in [0,1)$ with $x - v - nv_1 = tv_1 + y \in F$. Using Fubini's theorem we get

$$\operatorname{covol}(\Lambda) = \lambda_1(\Lambda)\operatorname{covol}(\Lambda_W). \tag{1.20}$$

This shows that Λ_W is a lattice in W.

Third preparatory step: Relating the Successive Minimas. As our last preparation for the induction step we show that

$$\lambda_k(\Lambda_W) \asymp \lambda_{k+1}(\Lambda) \tag{1.21}$$

for $k = 1, \dots, d-1$. Given $k+1$ linearly independent vectors of length less than $\lambda_{k+1}(\Lambda)$, we may replace one of them by v_1 (of norm $\lambda_1(\Lambda)$) and assume that these vectors are given by $v_1, v_2, \dots, v_{k+1} \in \Lambda$. In particular,

$$\pi(v_2), \dots, \pi(v_{k+1}) \in \Lambda_W$$

are linearly independent and also have length no more than $\lambda_{k+1}(\Lambda)$. Hence

$$\lambda_k(\Lambda_W) \leqslant \lambda_{k+1}(\Lambda)$$

for any $k = 1, \dots, d-1$. On the other hand, assume that

$$w_1 = \pi(v_2), \dots, w_k = \pi(v_{k+1}) \in \Lambda_W$$

are linearly independent of length no more than $\lambda_k(\Lambda_W)$. As above, we may assume $v_{j+1} = w_j + t_j v_1 \in \Lambda$ with $t_j \in [-\frac{1}{2}, \frac{1}{2})$ for $j = 1, \dots, k$, and so

$$\|v_{j+1}\| \ll \lambda_k(\Lambda_W) + \lambda_1(\Lambda) \ll \lambda_k(\Lambda_W),$$

since $\lambda_1(\Lambda) \ll \lambda_1(\Lambda_W) \leqslant \lambda_k(\Lambda_W)$.

Concluding the Induction. By the inductive assumption and the statement above, we get that

$$\operatorname{covol}(\Lambda_W) \asymp \lambda_1(\Lambda_W) \cdots \lambda_{d-1}(\Lambda_W) \asymp \lambda_2(\Lambda) \cdots \lambda_d(\Lambda).$$

Together with (1.20) this gives $\operatorname{covol}(\Lambda) \asymp \lambda_1(\Lambda) \cdots \lambda_d(\Lambda)$ as claimed in (1.19).

The Second type of Minimas. To see the last statement in the theorem, we let $v_1, \dots, v_k$ be linearly independent vectors satisfying $\|v_j\| \leqslant \lambda_j(\Lambda)$ for $j = 1, \dots, k$. We define the subspace $V = \mathbb{R}v_1 + \cdots + \mathbb{R}v_k$. Then

$$\operatorname{covol}(\Lambda \cap V) \leqslant \operatorname{covol}(\mathbb{Z}v_1 + \cdots + \mathbb{Z}v_k) \leqslant \|v_1\| \cdots \|v_k\| = \lambda_1(\Lambda) \cdots \lambda_k(\Lambda),$$

and so $\alpha_k(\Lambda) \leqslant \lambda_1(\Lambda) \cdots \lambda_k(\Lambda)$. Indeed, the first inequality holds as $\Lambda \cap V$ may have more lattice elements than $\mathbb{Z}v_1 + \cdots + \mathbb{Z}v_k \subseteq \Lambda \cap V$, and the second follows as the volume of a parallelepiped is less than the product of the lengths of its sides.

On the other hand, if $V \subseteq \mathbb{R}^d$ has dimension k and is Λ-rational, then we may apply (1.19) to the lattice $\Lambda \cap V$ in V to get

$$\operatorname{covol}(\Lambda \cap V) \asymp \lambda_1(\Lambda \cap V) \cdots \lambda_k(\Lambda \cap V) \geqslant \lambda_1(\Lambda) \cdots \lambda_k(\Lambda).$$

This shows that $\alpha_k(\Lambda) \gg \lambda_1(\Lambda) \cdots \lambda_k(\Lambda)$ and proves the theorem. □

Using the same inductive argument (by projection to the orthogonal complement of the shortest vector) we also get the following.

Corollary 1.46 (Basis of a lattice). *Let $\Lambda \subseteq \mathbb{R}^d$ be a lattice. Then there is a $\mathbb{Z}$-basis $v_1, \dots, v_d \in \Lambda$ of $\Lambda = \mathbb{Z}v_1 + \cdots + \mathbb{Z}v_d$ such that*

$$\|v_1\| = \lambda_1(\Lambda), \|v_2\| \asymp \lambda_2(\Lambda), \dots, \|v_d\| \asymp \lambda_d(\Lambda).$$

Moreover, the projection $\pi_k(v_k)$ of v_k onto the orthogonal complement of

$$\mathbb{R}v_1 + \cdots + \mathbb{R}v_{k-1}$$

has

$$\|\pi_k(v_k)\| \asymp \lambda_k(\Lambda) \asymp \|v_k\|$$

for $k = 2, \dots, d$

Corollary 1.46 may seem obvious, but our intuition about lattices does not extend to higher dimensions without some additional complexities. In particular, it is not true that there always exists a $\mathbb{Z}$-basis $v_1, \dots, v_d$ for a lattice with

$$\|v_1\| = \lambda_1(\Lambda), \|v_2\| = \lambda_2(\Lambda), \dots, \|v_d\| = \lambda_d(\Lambda),$$

see Exercise 1.50 for a simple counterexample.

Proof of Corollary 1.46. Assume the corollary for dimension $(d-1)$, and define $W = (\mathbb{R}v_1)^\perp$, $\pi = \pi_1$, and $\Lambda_W = \pi(\Lambda)$ as in the proof of Theorem 1.45. Recall that these assumptions lead to (1.21). By assumption, Λ_W has a $\mathbb{Z}$-basis $w_1 = \pi(v_2), \dots, w_{d-1} = \pi(v_d)$ satisfying all the claims. Once more we may assume that $v_k = w_{k-1} + t_k v_1 \in \Lambda$ with $t_k \in [-\frac{1}{2}, \frac{1}{2})$ so that $\|v_k\| \ll \lambda_k(\Lambda)$ as in the proof of Theorem 1.45. Using the inductive hypothesis, it follows that $v_1, \dots, v_d \in \Lambda$ is a $\mathbb{Z}$-basis of Λ with $\|v_1\| = \lambda_1(\Lambda)$, and $\|v_k\| \asymp \lambda_k(\Lambda)$ for $k = 2, \dots, d$.

For the last claim in the corollary, recall that we already showed that

$$\|v_2\| \asymp \|w_1\| = \lambda_1(\Lambda_W) \asymp \lambda_2(\Lambda),$$

which is the claim for $k = 2$. For $k > 2$, notice that $\pi_k \pi = \pi_k$ is (when restricted to W) also the orthogonal projection $\pi_{W,k-1}$ in W onto the orthogonal complement of $\mathbb{R}w_1 + \cdots + \mathbb{R}w_{k-2}$. Therefore, the inductive assumption applies to give

$$\|\pi_k(v_k)\| = \|\pi_{W,k-1}(w_{k-1})\| \asymp \lambda_{k-1}(\Lambda_W) \asymp \lambda_k(\Lambda) \asymp \|v_k\|,$$

which proves the corollary. □

Exercise 1.47. (1) Show that $\lambda_1(hg\mathbb{Z}^d) \leqslant \|h\|\lambda_1(g\mathbb{Z}^d)$ for $g, h \in \mathrm{GL}_d(\mathbb{R})$, where $\|\cdot\|$ denotes the operator norm.
(2) Conclude that $\lambda_1 \colon \mathsf{X}_d \to (0, \infty)$ is continuous.
(3) Generalize (2) to λ_k for $1 \leqslant k < d$.

Exercise 1.48. Suppose that $\Lambda_n = g_n\mathbb{Z}^d \to \Lambda = g\mathbb{Z}$ as $n \to \infty$ in the sense of the quotient X_d and its metric defined by (1.7). Show that

$$\Lambda = \big\{u \in \mathbb{R}^d \mid \text{there exists } v_n \in \Lambda_n \text{ with } \lim_{n\to\infty} v_n = u\big\}$$

and conclude once more that $\lambda_1 \colon \mathsf{X}_d \to (0, \infty)$ is continuous.

Exercise 1.49. Show that the minimum in the definition of $\alpha_k(\Lambda)$ in Theorem 1.45 is indeed achieved.

Exercise 1.50. Let $d \geqslant 5$. Let $\Lambda = \mathbb{Z}^{d-1} \times \{0\} + \mathbb{Z}v$ where $v = (\frac{1}{2}, \dots, \frac{1}{2})$. Show that

$$\lambda_1 = \cdots = \lambda_d = 1,$$

that $\operatorname{covol}(\Lambda) = \frac{1}{2}$, and that there does not exist a basis of Λ consisting of vectors of length 1.

1.4.3 Mahler's Compactness Criterion

The space $\mathsf{X}_d = \mathrm{SL}_d(\mathbb{R})/\mathrm{SL}_d(\mathbb{Z})$ cannot be compact for $d \geqslant 2$, since X_d is the space of unimodular lattices, and it is possible to degenerate a sequence of lattices. For example, the sequence of unimodular lattices (Λ_n) defined by

$$\Lambda_n = (\tfrac{1}{n}\mathbb{Z}) \times (n\mathbb{Z}) \times \mathbb{Z}^{d-2}$$

has no subsequence converging to a unimodular lattice. Indeed, if we were to assign a limit to this sequence, then we could only have

$$\Lambda_n \longrightarrow \mathbb{R} \times \{0\} \times \mathbb{Z}^{d-2}$$

as $n \to \infty$, so the putative 'limit' is not discrete and does not span $\mathbb{R}^d$.

More generally, any sequence (Λ_n) of unimodular lattices containing vectors with length converging to 0 (that is, with $\lambda_1(\Lambda_n) \to 0$ as $n \to \infty$) cannot converge in X_d. To see this concretely, suppose that $\Lambda_n = g_n\mathbb{Z}^d \to g\mathbb{Z}^d$ as $n \to \infty$. Then (after replacing g_n with $g_n\gamma_n$ for a suitable choice of $\gamma_n \in \mathrm{SL}_d(\mathbb{Z})$ if necessary) we can assume that $g_n \to g$ as $n \to \infty$ in the topology of $\mathrm{SL}_d(\mathbb{R})$ (cf. (1.7) on page 19 and the following discussion). Thus we can write $g_n = h_n g$ with $h_n \to I$ as $n \to \infty$, which implies that $\lambda_1(g_n\mathbb{Z}^d) \to \lambda_1(g\mathbb{Z}^d) > 0$ as $n \to \infty$ (see Exercise 1.47).

A reasonable guess is that the argument above is the only way in which the non-compactness of X_d comes about (that is, a sequence (Λ_n) of lattices with no convergent subsequence has $\lambda_1(\Lambda_n) \to 0$ as $n \to \infty$; equivalently any closed subset of X_d on which λ_1 has a positive lower bound—a 'uniformly discrete' set of lattices—is pre-compact).

Theorem 1.51 (Mahler's compactness criterion). *A subset $B \subseteq \mathsf{X}_d$ has compact closure if and only if there exists some $\delta > 0$ for which*

$$\Lambda \in B \implies \lambda_1(\Lambda) \geqslant \delta. \tag{1.22}$$

That is, B is compact if and only if it is closed and uniformly discrete.

Because of this result, it will be convenient to define the subset

$$\mathsf{X}_d(\delta) = \{\Lambda \in \mathsf{X}_d \mid \lambda_1(\Lambda) \geqslant \delta\}$$

for any $\delta > 0$. The condition in (1.22) will also be described by saying that elements of B do not contain any non-trivial δ-short vectors. An equivalent formulation of Theorem 1.51 is to say that a set $B \subseteq \mathsf{X}_d$ of unimodular lattices is compact if and only if it is closed and the *height* function defined by

$$\mathrm{ht}(\Lambda) = \frac{1}{\lambda_1(\Lambda)}$$

is bounded on B. Even though it is difficult to depict X_d on paper (for example, X_3 is topologically an 8-dimensional space), it is conventionally depicted as in Figure 1.10, in part to express the meaning of Theorem 1.51.

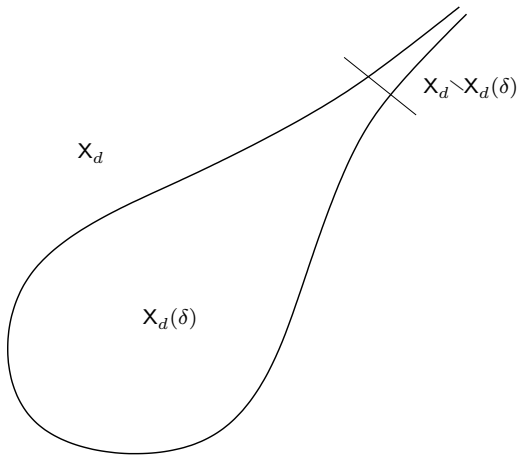

Fig. 1.10: A compact subset of X_d is contained in $\mathsf{X}_d(\delta) = \{\Lambda \in \mathsf{X}_d \mid \lambda_1(\Lambda) \geqslant \delta\}$ for some $\delta > 0$. The non-compact part $\mathsf{X}_d \smallsetminus \mathsf{X}_d(\delta)$, loosely referred to as a *cusp*, is depicted as a thin set to indicate the finite total volume (see Theorem 1.54). For $d > 2$ the geometry of the cusp is much more complicated than the cusp in the $d = 2$ case.

Proof of Theorem 1.51. We have already mentioned that λ_1 is a continuous function on X_d (see Exercise 1.47). Since λ_1 only achieves positive values, it follows that a compact subset of X_d must lie in $\mathsf{X}_d(\delta)$ for some $\delta > 0$. It remains to prove that $\mathsf{X}_d(\delta)$ is itself compact. Let $\left(g_n\mathbb{Z}^d\right)$ in $\mathsf{X}_d(\delta)$ be any sequence. Then, by Corollary 1.46, the lattice $g_n\mathbb{Z}^d$ has a $\mathbb{Z}$-basis $v_1^{(n)}, \ldots, v_d^{(n)}$ with

$$\delta \leqslant \lambda_1(g_n\mathbb{Z}^d) = \|v_1^{(n)}\| \ll \|v_2^{(n)}\| \ll \cdots \ll \|v_d^{(n)}\|$$

and

$$\|v_1^{(n)}\| \cdots \|v_d^{(n)}\| \ll 1,$$

which implies that

$$\|v_i^{(n)}\| \ll \delta^{-(d-1)}$$

for $i = 1, \dots, d$. As this change of basis of $g_n\mathbb{Z}^d$ corresponds to multiplication on the right of g_n by some $\gamma_n \in \mathrm{SL}_d(\mathbb{Z})$, we deduce that the entries of the matrix $g_n\gamma_n$ are all $\ll \delta^{-(d-1)}$. Thus there is a convergent subsequence

$$g_{n_k}\gamma_{n_k} \longrightarrow g$$

as $k \to \infty$ within $\mathrm{SL}_d(\mathbb{R}) \subseteq \mathrm{Mat}_d(\mathbb{R})$. It follows that $g_{n_k}\,\mathrm{SL}_d(\mathbb{Z}) \to g\,\mathrm{SL}_d(\mathbb{Z})$ as $k \to \infty$, as required. □

Exercise 1.52. Can Mahler's compactness criterion also be phrased in terms of λ_d, or in terms of λ_j for $2 \leqslant j < d$?

Exercise 1.53. Define for every $\Lambda \in \mathsf{X}_d$ the *covering radius* by

$$\rho(\Lambda) = \inf\left(\left\{r > 0 \mid \Lambda + B_r^{\mathbb{R}^d} = \mathbb{R}^d\right\}\right) > 0,$$

and show that $\rho\colon \mathsf{X}_d \to [0, \infty)$ is a proper continuous function. (Here it is necessary to include 0 in the range in order to give 'proper' the correct meaning.)

1.4.4 X_d has Finite Volume

Write π for the canonical quotient map $\pi\colon \mathrm{SL}_d(\mathbb{R}) \to \mathsf{X}_d$.

Theorem 1.54 (X_d has finite volume). $\mathrm{SL}_d(\mathbb{Z})$ *is a lattice in* $\mathrm{SL}_d(\mathbb{R})$.

We will prove the theorem by showing that Corollary 1.46 gives a surjective set of finite Haar measure—that is, a measurable set $F \subseteq \mathrm{SL}_d(\mathbb{R})$ (called a *Siegel domain*) with $\pi_{\mathsf{X}_d}(F) = \mathsf{X}_d$ and

$$m_{\mathrm{SL}_d(\mathbb{R})}(F) < \infty.$$

The fact that $m_{\mathrm{SL}_d(\mathbb{R})}(F)$ is finite is essentially a calculation, but is considerably helped by the *Iwasawa decomposition* (this is also referred to as the KAN decomposition).

Proposition 1.55 (Iwasawa decomposition). *Let* $K = \mathrm{SO}_d(\mathbb{R})$ *and*

$$B = AU = \left\{ \begin{pmatrix} a_1 & * & \cdots & * \\ & a_2 & \cdots & * \\ & & \ddots & \vdots \\ & & & a_d \end{pmatrix} \;\middle|\; a_1, \dots, a_d > 0, a_1 \cdots a_d = 1 \right\},$$

where

$$U = N = \left\{ \begin{pmatrix} 1 & u_{1,2} & \cdots & u_{1,d} \\ & 1 & \cdots & u_{2,d} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \mid u_{1,2}, \dots, u_{d-1,d} \in \mathbb{R} \right\}$$

and

$$A = \left\{ \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix} \;\middle|\; a_1, \dots, a_d > 0, a_1 \cdots a_d = 1 \right\}.$$

Then $\mathrm{SL}_d(\mathbb{R}) = KB = KAU$ *in the sense that for every* $g \in \mathrm{SL}_d(\mathbb{R})$ *there are unique matrices* $k \in K$, $a \in A$, $u \in U$ *with* $g = kau$.

Proof. This is the Gram–Schmidt procedure in disguise.(7) Let

$$g = (w_1, \dots, w_d),$$

where $w_1, \dots, w_d \in \mathbb{R}^d$ are the column vectors of g. We apply the Gram–Schmidt procedure to define

$$w_1' = \frac{1}{a_1} w_1$$

with $a_1 = \|w_1\| > 0$,

$$\widetilde{w_2} = u_{1,2} w_1 + w_2$$

with $u_{1,2} \in \mathbb{R}$ such that $\widetilde{w_2} \perp w_1$, and

$$w_2' = \frac{1}{a_2} \widetilde{w_2}$$

with $a_2 = \|\widetilde{w_2}\| > 0$ (by linear independence of w_1 and w_2). We continue this until

$$\widetilde{w_d} = u_{1,d} w_1 + u_{2,d} w_2 + \cdots + w_d$$

with $u_{1,d}, u_{2,d}, \dots, u_{d-1,d} \in \mathbb{R}$ such that

$$\widetilde{w_d} \perp w_1, \dots, w_{d-1}$$

(or, equivalently, $\widetilde{w_d} \perp w_1', \dots, w_{d-1}'$) and

$$w_d' = \frac{1}{a_d} w_d^{(1)}$$

with $a_d = \|\widetilde{w_d}\| > 0$ (again by linear independence). This has the following effect. If

$$u = \begin{pmatrix} 1 & u_{1,2} & \cdots & u_{1,d} \\ & 1 & \cdots & u_{2,d} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}$$

and

$$a = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix}$$

then

$$gu = (w_1, \widetilde{w_2}, \ldots, \widetilde{w_d})$$

and

$$gua^{-1} = (w_1', \ldots, w_d') = k.$$

By construction k has orthogonal rows, so that $\det(k) = \pm 1$. However,

$$\det(g) = 1 = \det(u)$$

and $\det(a) > 0$ which gives $\det(a) = 1 = \det(k)$. This shows the existence of the claimed $u \in U, a \in A$, and $k \in K$ with $g = kau^{-1}$.

To see that this decomposition is unique, first notice that B is a subgroup with $B \cap K = \{I\}$ so that $k_1 b_1 = k_2 b_2$ implies $k_2^{-1} k_1 = b_2 b_1^{-1} = I$. Similarly, we have $A \cap U = \{I\}$, and the proposition follows. □

Our geometric arguments in the proof of Theorem 1.45 and Corollary 1.46 are closely related to the Gram–Schmidt procedure used in Proposition 1.55. Combining these gives the next result.

Definition 1.56 (Siegel domain for X_d). A set of the form

$$\Sigma_{s,t} = K A_t U_s$$

where $s > 0$, $t > 0$,

$$U_s = \left\{ \begin{pmatrix} 1 & u_{1,2} & \cdots & u_{1,d} \\ & 1 & \cdots & u_{2,d} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \;\middle|\; |u_{i,j}| \leqslant s \right\},$$

and

$$A_t = \left\{ \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix} \;\middle|\; \tfrac{a_{i+1}}{a_i} \geqslant t \text{ for } i = 1, \ldots, d-1 \right\},$$

is called a *Siegel domain*.

Notice that U_s is a compact subset of the upper unipotent subgroup but A_t is a non-compact subset of the diagonal subgroup. We refer to Figure 1.11 in the case $d = 2$ (where it can be drawn). The purpose of Siegel domains is to avoid discussing the precise nature of a fundamental domain, which for $d \geqslant 3$ would require us to deal with a set in at least eight (or five if we choose to ignore K) dimensions.

The next result could again be attributed to Korkine and Zolotareff, while Siegel extended constructions of this sort to all classical non-compact simple groups.

Corollary 1.57 (Surjectivity of Siegel domains). *There exists some t_0 such that the Siegel domain $\Sigma_{\frac{1}{2},t_0}$ is surjective (that is, the image $\pi_{\mathsf{X}_d}(\Sigma_{\frac{1}{2},t_0})$ is X_d).*

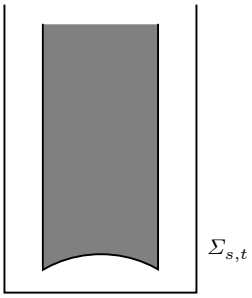

Fig. 1.11: For $d = 2$ a Siegel domain $\Sigma_{s,t}$ contains the standard fundamental domain precisely if $|s| \geqslant \frac{1}{2}$ and $t \leqslant \frac{\sqrt{3}}{2}$.

A more careful analysis of the proof shows that $t_0 = \frac{\sqrt{3}}{2}$ suffices in any dimension; see also Exercise 1.64 which can also be used to prove this claim.

Proof of Corollary 1.57. Let $\Lambda \in \mathsf{X}_d$ be a unimodular lattice, and let $w_1, \ldots, w_d$ be the $\mathbb{Z}$-basis as in Corollary 1.46. Replacing w_d by $-w_d$ if necessary, we may assume that $\det(g) = 1$, where $g = (w_1, \ldots, w_d)$. Now apply the Gram–Schmidt procedure as in the proof of Proposition 1.55 to g. By the second part of Corollary 1.46 we get

$$\begin{aligned} a_1 &= \|w_1\| = \lambda_1(\Lambda) \\ a_2 &= \|\widetilde{w_2}\| \asymp \lambda_2(\Lambda) \\ &\vdots \\ a_d &= \|\widetilde{w_d}\| \asymp \lambda_d(\Lambda) \end{aligned}$$

which satisfy

$$\frac{a_{i+1}}{a_i} \gg \frac{\lambda_{i+1}(\Lambda)}{\lambda_i(\Lambda)} \geqslant 1$$

for $i = 1, \ldots, d-1$. Choosing t_0 accordingly gives

$$a = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix} \in A_{t_0}.$$

Therefore $\Lambda = g\mathbb{Z}^d$ and $g = kau$ with $u \in U$ and $k \in K$. Notice that by replacing g by $gu_{\mathbb{Z}}$ with $u_{\mathbb{Z}} \in U(\mathbb{Z}) = U \cap \mathrm{Mat}_d(\mathbb{Z})$ we only replace u by $uu_{\mathbb{Z}}$. Moreover, $(uu_{\mathbb{Z}})_{i,i+1} = u_{i,i+1} + (u_{\mathbb{Z}})_{i,i+1}$ for $i = 1, \ldots, d-1$. Hence using this replacement for a suitable $u_{\mathbb{Z}}$ we can ensure that $u_{i(i+1)} \in [-\frac{1}{2}, \frac{1}{2})$. Having achieved this we may use another $u_{\mathbb{Z}} \in U(\mathbb{Z})$ with $(u_{\mathbb{Z}})_{i(i+1)} = 0$ for $i = 1, \ldots, d-1$,

which makes it easy to calculate the next off-diagonal of $uu_{\mathbb{Z}}$ as follows:

$$\begin{aligned}(uu_{\mathbb{Z}})_{i,i+2} &= u_{i,i+2} + u_{i,i+1}(u_{\mathbb{Z}})_{i+1,i+2} + (u_{\mathbb{Z}})_{i,i+2} \\ &= u_{i,i+2} + 0 + (u_{\mathbb{Z}})_{i,i+2}\end{aligned}$$

for any $i = 1, \dots, d-2$. Therefore, we can modify u by some $u_{\mathbb{Z}}$ as above to ensure that $u_{i(i+2)}$ lies in $[-\frac{1}{2}, \frac{1}{2})$ for $i = 1, \dots, d-2$. Proceeding by induction gives

$$\Lambda = g\mathbb{Z}^d = kau\mathbb{Z}^d$$

for some $u \in U_{1/2}$, $a \in A_{t_0}$, and $k \in K$. □

It remains to show that the Haar measure of the Siegel domains is finite. For this the Iwasawa decomposition also helps us to understand the Haar measure $m_{\mathrm{SL}_d(\mathbb{R})}$ as a result of the following general fact about locally compact groups.

Lemma 1.58 (Decomposition of Haar measure). *Let G be a unimodular, metric, σ-compact, locally compact group. Let $S, T \subseteq G$ be closed subgroups with $S \cap T = \{I\}$ and with the property that $m_G(ST) > 0$ (for example, because ST contains an open neighbourhood of I). Then*

$$m_G|_{ST} \propto \phi_* \left(m_S \times m_T^{(r)} \right),$$

where $\phi\colon S \times T \to G$ is the product map $\phi\colon (s,t) \mapsto st$.

We refer to [45, Lem. 11.31], [46, Lem. 10.57], and Knapp [87] for the proof. The above lemma is useful for us because of the following.

Lemma 1.59. $\mathrm{SL}_d(\mathbb{R})$ *is unimodular.*

As an alternative to Exercise 1.7 (which is quite special but gives the above lemma) we start with a general lemma about the structure of $\mathrm{SL}_d(\mathbb{K})$ over any field $\mathbb{K}$.

Lemma 1.60 (Unipotent Generation). *Over any field $\mathbb{K}$, the special linear group $\mathrm{SL}_d(\mathbb{K})$ is generated by the elementary unipotent subgroups*

$$U_{i,j}(\mathbb{K}) = \{u_{i,j}(t) = I + tE_{i,j} \mid t \in \mathbb{K}\}$$

with $i \neq j$ and $E_{i,j}$ being the elementary matrix with (i,j)th entry 1 and all other entries 0.

For $\mathbb{K} = \mathbb{R}$ (and for $\mathbb{K} = \mathbb{C}$), this implies that $\mathrm{SL}_d(\mathbb{R})$ (and $\mathrm{SL}_d(\mathbb{C})$) are connected as topological spaces, because each subgroup $U_{i,j}(\mathbb{R})$ and $U_{i,j}(\mathbb{C})$ is connected. In particular, this shows that $\mathrm{SL}_d(\mathbb{R})$ carries a left-invariant Riemannian metric, and by restriction of this metric to any closed subgroup of $\mathrm{SL}_d(\mathbb{R})$ (which may be connected or not) one has a left-invariant metric on the subgroup (which induces the locally compact, σ-compact, induced topology).

Outline proof of Lemma 1.60. Notice that for $i \neq j$ the row (and column) operation of adding t times the jth row to the ith row (or t times the ith column to the jth column) corresponds to multiplication by the elements $u_{i,j}(t) \in U_{i,j}(\mathbb{K})$ on the left (resp. right) of a given matrix $g \in \mathrm{SL}_d(\mathbb{K})$. This restricted Gaussian elimination can be used to reduce the matrix g to the identity. To do this we may first ensure that $g_{1,2} \neq 0$ with a suitable row operation, then use another row operation to ensure that $g_{1,1} = 1$. Then suitable row and column operations can be used to obtain $g_{1i} = 0 = g_{i1}$ for $i > 1$, and we may then continue by induction. At the last step the fact that $\det(g) = 1$ is needed to ensure that the diagonal matrix produced is in fact the identity. This can be used to express g as a finite product of elementary unipotent matrices. □

Proof of Lemma 1.59. Recall the unipotent subgroups

$$U_{i,j} = \{u_{i,j}(t) = I + tE_{i,j} \mid t \in \mathbb{R}\}$$

for $i \neq j$ from Lemma 1.60. Let $a \in A$ be any diagonal matrix, and notice that $au_{i,j}(t)a^{-1} = u_{i,j}(\frac{a_i}{a_j}t)$ for $t \in \mathbb{R}$. Therefore, the commutator satisfies

$$[a, u_{i,j}(t)] = a^{-1}u_{i,j}(-t)au_{i,j}(t) = u_{i,j}((1 - \tfrac{a_j}{a_i})t).$$

Choosing $a \in A$ correctly, it follows that the commutator group

$$[\mathrm{SL}_d(\mathbb{R}), \mathrm{SL}_d(\mathbb{R})]$$

contains $U_{i,j}$ for all $i \neq j$. By Lemma 1.60 it follows that

$$[\mathrm{SL}_d(\mathbb{R}), \mathrm{SL}_d(\mathbb{R})] = \mathrm{SL}_d(\mathbb{R}).$$

Since the modular character $\mathrm{mod}\colon \mathrm{SL}_d(\mathbb{R}) \to \mathbb{R}_{>0}$ is a homomorphism to an abelian group it follows that $\mathrm{mod}(\mathrm{SL}_d(\mathbb{R})) = \{1\}$, proving the lemma. □

We will also need the following lemma.

Lemma 1.61. *The right Haar measure $m_B^{(r)}$ on $B = AU$ can be defined by*

$$\mathrm{d}m_B^{(r)}(au) = \rho(a)\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u),$$

where

$$\rho\left(\begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix}\right) = \prod_{i<j} \frac{a_i}{a_j} \tag{1.23}$$

and we use the coordinate system $au \in B$ for $a \in A$ and $u \in U$.

Proof. We note first that the Haar measure m_U on U can be defined using the Lebesgue measure on the unconstrained coefficients

$$u_{1,2}, \dots, u_{1,d}, u_{2,3}, \dots, u_{2,d}, \dots, u_{d-1,d}$$

of $u \in U$. Moreover, U is in fact unimodular by Exercise 1.63. Let $f \geqslant 0$ be a measurable function on B and $\widetilde{u} \in U$. Then

$$\int_{AU} f(a \underbrace{u\widetilde{u}}_{u'})\rho(a)\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u) = \int_{AU} f(au')\rho(a)\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u')$$

since m_U is right-invariant. Now let $\widetilde{a} \in A$ and calculate

$$\begin{aligned}
&\int_{AU} f(au\widetilde{a})\rho(a)\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u) \\
&\qquad = \int_{AU} f(\underbrace{a\widetilde{a}}_{a'}(\widetilde{a}^{-1}u\widetilde{a}))\rho(\underbrace{a\widetilde{a}}_{a'})\rho(\widetilde{a})^{-1}\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u) \\
&\qquad = \int_{AU} f(a'(\underbrace{\widetilde{a}^{-1}u\widetilde{a}}_{u'}))\rho(a')\,\mathrm{d}m_A(a') \underbrace{\prod_{i<j} \frac{\widetilde{a}_j}{\widetilde{a}_i}}_{=\rho(\widetilde{a})^{-1}}\,\mathrm{d}m_U(u)
\end{aligned}$$

since ρ is a homomorphism and m_A is invariant. We also note that

$$\widetilde{a}^{-1}u\widetilde{a} = \begin{pmatrix}
1 & \frac{\widetilde{a}_2}{\widetilde{a}_1}u_{1,2} & \cdots & \cdots & \frac{\widetilde{a}_d}{\widetilde{a}_1}u_{1,d} \\
 & 1 & \frac{\widetilde{a}_3}{\widetilde{a}_2}u_{2,3} & \cdots & \frac{\widetilde{a}_d}{\widetilde{a}_2}u_{2,d} \\
 & & \ddots & & \vdots \\
 & & & 1 & \frac{\widetilde{a}_{d-1}}{\widetilde{a}_d}u_{d-1,d} \\
 & & & & 1
\end{pmatrix}.$$

Using the fact that $\mathrm{d}m_U$ is the Lebesgue measure, we can make the linear substitution $u' = \widetilde{a}^{-1}u\widetilde{a}$ (or, equivalently, $u'_{i,j} = \frac{\widetilde{a}_j}{\widetilde{a}_i}u_{i,j}$ for $i < j$) and see that $\rho(\widetilde{a})^{-1}$ is precisely the Jacobian for this substitution. It follows that

$$\int_{AU} f(au\widetilde{a})\rho(a)\,\mathrm{d}m_A(a)\,\mathrm{d}m_U(u) = \int_{AU} f(a'u')\rho(a')\,\mathrm{d}m_A(a')\,\mathrm{d}m_U(u').$$

Together with the above identity for right translation by $\widetilde{u}$, this proves the lemma. □

To complete the proof of Theorem 1.54, it remains to show the following lemma.

Lemma 1.62. *For any $s > 0$ and $t > 0$, we have $m_{\mathrm{SL}_d(\mathbb{R})}\left(\Sigma_{s,t}\right) < \infty$.*

Proof. Using Lemma 1.58 for $G = \mathrm{SL}_d(\mathbb{R})$, $S = K$, and $T = B$ we see that K can be ignored and we have to calculate $m_B^{(r)}(A_t U_s)$, where as usual $m_B^{(r)}$ denotes the right Haar measure on B. We have $\mathrm{d}m_B^{(r)} = \rho(a)\,\mathrm{d}m_A \times \mathrm{d}m_U$ by

Lemma 1.61, where ρ is given by (1.23). Using this, we get

$$m_B^{(r)}(A_t U_s) \ll \underbrace{m_U(U_s)}_{<\infty} \int_{A_t} \rho(a)\, \mathrm{d}m_A(a),$$

and so the problem is reduced to the integral over A_t.

Using the relations

$$\frac{a_i}{a_j} = \frac{a_i}{a_{i+1}} \cdots \frac{a_{i-1}}{a_j} = \prod_{k=i}^{j-1} \frac{a_k}{a_{k+1}}$$

for $i < j$, we also obtain the formula

$$\rho\left(\begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix}\right) = \prod_{i<j} \frac{a_i}{a_j} = \prod_{k=1}^{d-1} \left(\frac{a_k}{a_{k+1}}\right)^{r_k} = \prod_{k=1}^{d-1} \left(\frac{a_{k+1}}{a_k}\right)^{-r_k} \tag{1.24}$$

for some integers $r_k > 0$. Here $r_k = (d-k)k$ equals the number of tuples of indices (i,j) with $i \leqslant k < j$, but the exact form of $r_k > 0$ does not matter at this point.

Next notice that

$$A \ni a = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_d \end{pmatrix} \longmapsto (y_1, \ldots, y_{d-1}) = \left(\log \tfrac{a_2}{a_1}, \ldots, \log \tfrac{a_d}{a_{d-1}}\right) \in \mathbb{R}^{d-1} \tag{1.25}$$

is an isomorphism of topological groups which maps A_t to $[\log t, \infty)^{d-1}$, so that

$$\int_{A_t} \rho(a)\, \mathrm{d}m_A(a) \propto \prod_{k=1}^{d-1} \int_{\log t}^{\infty} \mathrm{e}^{-r_k y_k}\, \mathrm{d}y_k < \infty$$

as claimed. □

The proof presented above is usually referred to as the *reduction theory* of SL_d, and this generalizes to other algebraic groups by a theorem of Borel and Harish–Chandra [8] (see Siegel [146]). In Chapter 4 we will give a second proof which will also lead to the general result for other groups in Chapter 7.

Exercise 1.63. Show that U is unimodular and that the Haar measure on U can be defined by

$$\mathrm{d}m_U(u) = \mathrm{d}u_{1,2} \cdots \mathrm{d}u_{1,d}\, \mathrm{d}u_{2,3} \cdots \mathrm{d}u_{2,d} \cdots \mathrm{d}u_{d-1,d},$$

where $u_{1,2}, \ldots, u_{1,d}, u_{2,3}, \ldots, u_{2,d}, \ldots, u_{d-1,d}$ are the unconstrained coefficients of the matrix u.

Exercise 1.64 (LLL algorithm[(8)]). In this exercise a different proof of Corollary 1.57 will be given (which will not use Minkowski's theorem on successive minimas). For this let $v_1, \ldots, v_d$ be an ordered basis of a unimodular lattice $\Lambda < \mathbb{R}^d$. For every $i = 1, \ldots, d$ define v_i^* to be

the projection of v_i onto the orthogonal complement of the linear span of $v_1, \dots, v_{i-1}$. Recall that $\|v_i^*\|$ is the ith diagonal entry of the A-component of the decomposition of the matrix g whose rows consist of $v_1, \dots, v_d$. We may assume that we have $\det g = 1$.

The basis is called *semi-reduced* if all linear coefficients of $v_i - v_i^*$, when expressed as a linear combination of $v_1, \dots, v_{i-1}$, are in $[-\frac{1}{2}, \frac{1}{2})$ (that is, the U-part of g in the Iwasawa decomposition belongs to $U_{1/2}$).

The basis is called *t-reduced* (for some fixed $t > 0$) if it is semi-reduced and if $\frac{\|v_{i+1}^*\|}{\|v_i^*\|} \geqslant t$ for $i = 1, \dots, d-1$ (that is, the A-part of g in the NAK-decomposition belongs to A_t).

Prove that the following algorithm terminates for every fixed $t < \frac{\sqrt{3}}{2}$ with a t-reduced ordered basis of Λ.
(a) Check if the ordered basis is semi-reduced. If not perform a simple change of basis (using only a change of basis in $U \cap \mathrm{SL}_d(\mathbb{Z})$) and produce a new ordered basis which is semi-reduced.
(b) Check if the basis is t-reduced. If so, the algorithm terminates.
(c) So assume that the ordered basis is not t-reduced but is semi-reduced. Then there exists a smallest i for which $\frac{\|v_{i+1}^*\|}{\|v_i^*\|} < t$. Now replace the basis with the new basis where the order of v_i and v_{i+1} is reversed (but all other basis elements retain their place), and start the algorithm from the beginning.

For the proof you may find useful the function θ of the ordered basis defined by

$$\theta(v_1, \dots, v_d) = \prod_{i=1}^{d} \mathrm{covol}(\mathbb{Z}v_1 + \cdots \mathbb{Z}v_i).$$

1.4.5 The Siegel Transform*

†We fix $d \geqslant 2$ and define for $f \in C_c(\mathbb{R}^d)$ its *Siegel transform* by

$$\widetilde{f} \colon \mathsf{X}_d \ni \Lambda \longmapsto \sum_{v \in \Lambda \smallsetminus \{0\}} f(v). \tag{1.26}$$

Note that $\left|\Lambda \cap \left(B_1^{\mathrm{SL}_d(\mathbb{R})} \cdot \operatorname{supp} f\right)\right| < \infty$, which shows that the sum defining $\widetilde{f}(g\Lambda)$ involves only finitely many summands depending on $\Lambda \in \mathsf{X}_d$ but independent of $g \in B_1^{\mathrm{SL}_d(\mathbb{R})}$. This implies that $\widetilde{f}(\Lambda)$ is well-defined for every $\Lambda \in \mathsf{X}_d$ and that $\widetilde{f} \in C(\mathsf{X}_d)$.

Theorem 1.65 (Siegel formula). *The Siegel transform satisfies*

$$\frac{1}{m_{\mathsf{X}_d}(\mathsf{X}_d)} \int_{\mathsf{X}_d} \widetilde{f}(x) \, \mathrm{d}m_{\mathsf{X}_d}(x) = \int_{\mathbb{R}^d} f(v) \, \mathrm{d}v$$

for all $f \in C_c(\mathbb{R}^d)$.

The first step towards the proof of Theorem 1.65 is to show that $\widetilde{f}$ is integrable with respect to m_{X_d}. For this the following upper bound in terms of the successive minima $\lambda_1, \dots, \lambda_d \colon \mathsf{X}_d \to (0, \infty)$ from Theorem 1.45 will be useful.

† This section will not be needed later, which we indicate with the asterisk in the title.

Lemma 1.66 (Upper bound). *For $f \in C_c(\mathbb{R}^d)$ and $r > 0$ with $\operatorname{supp} f \subseteq B_r^{\mathbb{R}^d}$ we have*

$$\left|\widetilde{f}\right| \ll \|f\|_\infty \max_{k=1,\dots,d} \frac{r^k}{\lambda_1 \cdots \lambda_k}.$$

Proof. Let V be the linear hull of $\Lambda \cap B_r^{\mathbb{R}^d}$ and $k = \dim V$. Note that this means $\lambda_k(\Lambda) < r$ but $\lambda_{k+1}(\Lambda) \geqslant r$. We apply Corollary 1.46 to the lattice $\Lambda \cap V$ inside V and obtain a $\mathbb{Z}$-basis $v_1, \dots, v_k$ of $\Lambda \cap V$ with

$$\|v_j\| \asymp \lambda_j(\Lambda \cap V) = \lambda_j(\Lambda) \leqslant r$$

for $j = 1, \dots, k$. Let $F = \sum_{j=1}^k [0,1)v_j$ be a fundamental domain for $\Lambda \cap V < V$. The k-dimensional volume of $F \subseteq V$ satisfies

$$\operatorname{vol}_V(F) \asymp \lambda_1(\Lambda \cap V) \cdots \lambda_k(\Lambda \cap V) = \lambda_1(\Lambda) \cdots \lambda_k(\Lambda)$$

by the second part of Corollary 1.46. For any $v \in \Lambda \cap B_r^{\mathbb{R}^d}$ this implies that

$$v + F \subseteq V \cap B_R^{\mathbb{R}^d}$$

with $R \ll r$. Therefore

$$\left|\Lambda \cap B_r^{\mathbb{R}^d}\right| \operatorname{vol}_V(F) \leqslant \operatorname{vol}_V\left(V \cap B_R^{\mathbb{R}^d}\right) \asymp R^k,$$

which gives

$$\left|\Lambda \cap B_r^{\mathbb{R}^d}\right| \ll \frac{r^k}{\lambda_1(\Lambda) \cdots \lambda_k(\Lambda)}.$$

Together with the definition of $\widetilde{f}$ in (1.26) this gives the lemma. □

Lemma 1.67 (Integrability). *For $k = 1, \dots, d$ the functions*

$$\frac{1}{\lambda_1 \cdots \lambda_k} \colon \mathsf{X}_d \longrightarrow (0, \infty)$$

are integrable with respect to m_{X_d}. In particular, $\widetilde{f}$ is integrable for any function $f \in C_c(\mathbb{R}^d)$.

For the proof we will reuse and extend ideas from the proof of Theorem 1.54.

Proof of Lemma 1.67. Note that for $k = d$ we have $\lambda_1 \cdots \lambda_d \asymp 1$ by Theorem 1.45 and hence the lemma reduces to Theorem 1.54. So we now suppose that $k \in \{1, \dots, d-1\}$. By Corollary 1.57 there exists $t_0 > 0$ so that the Siegel domain $\Sigma_{\frac{1}{2},t_0} = KA_{t_0}U_{\frac{1}{2}}$ is surjective. The functions $\lambda_1, \dots, \lambda_d$ are K-invariant and hence it suffices (by Lemma 1.58) once more to consider integrals over $A_{t_0}U_{\frac{1}{2}}$ with respect to $m_B^{(r)}$. As in the proof of Corollary 1.57 the diagonal entries $a_1, \dots, a_d$ of $a \in A_{t_0}$ satisfy $a_j \asymp \lambda_j(au\mathbb{Z}^d)$ for $j = 1, \dots, d$ and $u \in U_{\frac{1}{2}}$. Therefore we obtain

$$\int_{\mathsf{X}_d} \frac{1}{\lambda_1 \cdots \lambda_k} \,\mathrm{d}m_{\mathsf{X}_d} \ll \int_{A_t} \frac{1}{a_1 \cdots a_k} \rho(a) \,\mathrm{d}m_A(a).$$

Next we wish to use the isomorphism (1.25) between A and $\mathbb{R}^{d-1}$. In order to do this we need to express the product $a_1 \cdots a_k$ of the first k diagonal entries of a in A in terms of $y_j = \log \frac{a_{j+1}}{a_j}$ for $j = 1, \ldots, d-1$. Using the relation

$$a_1 \cdots a_d = \det a = 1$$

we have

$$\begin{aligned} a_1^d \cdots a_k^d &= (a_1 \cdots a_k)^{d-k} (a_{k+1} \cdots a_d)^{-k} \\ &= \left(\tfrac{a_1}{a_2}\right)^{d-k} \left(\tfrac{a_2}{a_3}\right)^{2(d-k)} \cdots \left(\tfrac{a_k}{a_{k+1}}\right)^{k(d-k)} \\ &\qquad \times \left(\tfrac{a_{k+1}}{a_{k+2}}\right)^{k(d-k)-k} \left(\tfrac{a_{k+2}}{a_{k+3}}\right)^{k(d-k)-2k} \cdots \left(\tfrac{a_{d-1}}{a_d}\right)^{k(d-k)-(d-k-1)k} \end{aligned}$$

and so

$$\frac{1}{a_1 \cdots a_k} = \exp\Big(\sum_{j=1}^{k} j\big(1 - \tfrac{k}{d}\big) y_j\Big) \times \exp\Big(\sum_{j=k+1}^{d-1} \big(k - j\tfrac{k}{d}\big) y_j\Big).$$

Together with the formula (1.24) for ρ and $r_j = j(d-j)$ for $j = 1, \ldots, d-1$ (as mentioned there) we obtain

$$\begin{aligned} \int_{\mathsf{X}_d} \frac{1}{\lambda_1 \cdots \lambda_k} \,\mathrm{d}m_{\mathsf{X}_d} \ll \prod_{j=1}^{k} \int_t^\infty \exp\Big(\big(j(1 - \tfrac{k}{d}) - j(d-j)\big) y_j\Big) \,\mathrm{d}y_j \\ \times \prod_{j=k+1}^{d-1} \int_t^\infty \exp\Big(\big(k - j\tfrac{k}{d} - j(d-j)\big) y_j\Big) \,\mathrm{d}y_j. \end{aligned}$$

For $j = 1, \ldots, k$ the exponent is negative because

$$\big(1 - \tfrac{k}{d}\big) < 1 \leqslant d - j.$$

For $j = k+1, \ldots, d-1$ we also have

$$k - j\tfrac{k}{d} - j(d-j) = \big(\tfrac{k}{d} - j\big)\big(d - j\big) < 0.$$

It follows that the integrals are all finite.

The final claim of the lemma follows from the first part and Lemma 1.66. □

With these preparations we are now ready to prove Siegel's formula.

Proof of Theorem 1.65. By Lemma 1.67 the linear functional

$$\ell \colon C_c(\mathbb{R}^d) \longmapsto \int_{\mathsf{X}_d} \widetilde{f} \,\mathrm{d}m_{\mathsf{X}_d}$$

is well-defined. Moreover, $f \geqslant 0$ implies $\widetilde{f} \geqslant 0$ and hence $\ell(\widetilde{f}) \geqslant 0$. In other words, ℓ is a positive linear functional on $C_c(\mathbb{R}^d)$. By the Riesz representation theorem (see [46, Th. 7.44]) there exists a uniquely determined positive locally finite measure μ on $\mathbb{R}^d$ so that

$$\int_{\mathsf{X}_d} \widetilde{f}\, \mathrm{d}m_{\mathsf{X}_d} = \int_{\mathbb{R}^d} f\, \mathrm{d}\mu. \tag{1.27}$$

Moreover, for $g \in \mathrm{SL}_d(\mathbb{R})$ we have

$$\widetilde{f}(g\Lambda) = \sum_{v \in \Lambda \smallsetminus \{0\}} f(gv) = \widetilde{f \circ g}(\Lambda)$$

which implies that

$$\int_{\mathbb{R}^d} f \circ g\, \mathrm{d}\mu = \int_{\mathsf{X}_d} \widetilde{f \circ g}\, \mathrm{d}m_{\mathsf{X}_d} = \int_{\mathsf{X}_d} \widetilde{f} \circ g\, \mathrm{d}m_{\mathsf{X}_d} = \int_{\mathsf{X}_d} \widetilde{f}\, \mathrm{d}m_{\mathsf{X}_d} = \int_{\mathbb{R}^d} f\, \mathrm{d}\mu$$

by invariance of m_{X_d}. It follows that μ is invariant under the action of $\mathrm{SL}_d(\mathbb{R})$ on $\mathbb{R}^d$.

The action of $\mathrm{SL}_d(\mathbb{R})$ on $\mathbb{R}^d$ has only two orbits, namely the fixed point $\{0\}$ and $\mathbb{R}^d \smallsetminus \{0\} = \mathrm{SL}_d(\mathbb{R}) \cdot e_1$. Uniqueness of invariant measures on homogeneous space (see Appendix C) therefore implies that

$$\mu = c_0 \delta_0 + c m_{\mathbb{R}^d} \tag{1.28}$$

for constants $c_0, c \geqslant 0$. We will show that $c_0 = 0$ and $c = m_{\mathsf{X}_d}(\mathsf{X}_d)$, which by (1.27) gives the theorem.

Notice that (1.27) also holds for $f_r = \mathbb{1}_{B_r^{\mathbb{R}^d}}$ for any $r > 0$ by monotone convergence. For $r \searrow 0$ we have $\widetilde{f}_r \searrow 0$ (since the origin is not part of the sum defining $\widetilde{f}$ in (1.26)). This already implies that

$$c_0 = \lim_{r \searrow 0} \int_{\mathbb{R}^d} f_r\, \mathrm{d}\mu = \lim_{r \searrow 0} \int_{\mathsf{X}_d} \widetilde{f}_r\, \mathrm{d}m_{\mathsf{X}_d} = 0$$

by dominated convergence.

To calculate c we consider the normalized function $\frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})} f_r$ as $r \to \infty$. We claim that

$$\frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})} \widetilde{f}_r \longrightarrow \mathbb{1} \tag{1.29}$$

as $r \to \infty$. To see this fix $\Lambda \in \mathsf{X}_d$ and a bounded fundamental domain F for Λ, and let $s = \sup_{v \in F} \|v\|$. Then

$$\left| \Lambda \cap B_r^{\mathbb{R}^d} \right| = m_{\mathbb{R}^d}\left(\bigsqcup_{v \in \Lambda \cap B_r^{\mathbb{R}^d}} (v + F) \right) \leqslant m_{\mathbb{R}^d}\left(B_{r+s}^{\mathbb{R}^d} \right) = m_{\mathbb{R}^d}\left(B_1^{\mathbb{R}^d} \right) (r+s)^d$$

and

$$\left|\Lambda \cap B_r^{\mathbb{R}^d}\right| = m_{\mathbb{R}^d}\left(\bigsqcup_{v\in\Lambda\cap B_r^{\mathbb{R}^d}} (v+F)\right) \geqslant m_{\mathbb{R}^d}\left(B_{r-s}^{\mathbb{R}^d}\right) = m_{\mathbb{R}^d}\left(B_1^{\mathbb{R}^d}\right)(r-s)^d,$$

which together already imply the claim (1.29). Moreover, Lemma 1.66 applied to $\frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})} f_r$ gives

$$\frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})}\widetilde{f}_r \ll \max_{k=1,\dots,d} \frac{1}{\lambda_1\lambda_2\cdots\lambda_k}$$

for all $r \geqslant 1$. By Lemma 1.67 the upper bound is integrable and so we may use dominated convergence in (1.29). Together with (1.27) and (1.28) this gives

$$c = \frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})}\int_{\mathbb{R}^d} f_r \,\mathrm{d}\mu = \int_{\mathsf{X}_d} \frac{1}{m_{\mathbb{R}^d}(B_r^{\mathbb{R}^d})}\widetilde{f}_r \,\mathrm{d}m_{\mathsf{X}_d} \longrightarrow \int_{\mathsf{X}_d} \mathbb{1}\,\mathrm{d}m_{\mathsf{X}_d} = m_{\mathsf{X}_d}(\mathsf{X}_d)$$

as $r \to \infty$. □

Notes to Chapter 1

(1) (Page 6) The error term $N(R) - \pi R^2$ was shown to be bounded above by $2\sqrt{2}\pi R$ by Gauss. Hardy [65] and Landau [94] found a *lower* bound for the error by showing that the error is not $\mathrm{o}(R^{\frac{1}{2}}(\log R)^{\frac{1}{4}})$. It is conjectured that the upper bound is $\mathrm{O}_\varepsilon(R^{\frac{1}{2}+\varepsilon})$. The power of R must be at least $\frac{1}{2}$ by the lower bound of Hardy and Landau, and has been shown to be less than or equal to $\frac{131}{208}$ by Huxley [71].

(2) (Page 7) An account of this argument may be found in the authors' notes [47].

(3) (Page 9) This is a simple instance of the more general Iwasawa decomposition of a connected real semi-simple Lie group; see the original paper of Iwasawa [74] or Knapp's monograph [87] for an account.

(4) (Page 18) For the history and primary references of these developments we refer to the paper of Phillips and Rudnick [119].

(5) (Page 32) In fact any perfect Polish space allows an embedding of the middle-third Cantor set into it, so in particular such a space has the cardinality of the continuum. We refer to Kechris [79, Sec. 6.A].

(6) (Page 35) We refer to the monographs of Cassels [11] or Gruber and Lekkerkerker [64] for thorough accounts of the topic and its history. For our purposes Theorem 1.45, a consequence of the reduction algorithm of Korkine and Zolotareff [89, 90, 91], will suffice.

(7) (Page 42) This method was presented by E. Schmidt [133, Sec. 3, p. 442], and he pointed out that essentially the same method was used earlier by Gram [63]; the modern view is that the methods differ, and that the Gram form was used earlier by Laplace [96, p. 497ff.] in a different setting.

(8) (Page 48) This is based on the so-called LLL algorithm of A. K. Lenstra, H. W. Lenstra, Jr., and Lovász [97].