Chapter 1
Lattices and the Space of Lattices

We recall that a (continuous) *action* of a (topological) group G on a (topological) space X is a (continuous) map $G \times X \to X$, written $(g, x) \mapsto g{\bullet}x$, with the property that $g{\bullet}(h{\bullet}x) = (gh){\bullet}x$ and $e{\bullet}x = x$ 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 $G{\bullet}x = \{g{\bullet}x \mid g \in G\}$ is called the G-*orbit* of x.

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 $\mathsf{X}_d = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. 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 Discrete Subgroups and Lattices

1.1.1 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 a connected Lie group G, 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

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

for any $g \in G$. Write $B_r^G = B_r^G(e)$ for the metric open ball of radius r around the identity $e \in G$. If Γ is a *discrete* subgroup (which means that e is an isolated point of Γ), then there is an induced metric on the quotient space $X = \Gamma\backslash G$ defined by

$$\mathsf{d}_X(\Gamma g_1, \Gamma g_2) = \inf_{\gamma_1,\gamma_2\in\Gamma} \mathsf{d}_G(\gamma_1 g_1, \gamma_2 g_2) = \inf_{\gamma\in\Gamma} \mathsf{d}_G(\gamma g_1, g_2) \tag{1.1}$$

for any $\Gamma g_1, \Gamma g_2 \in X$, where both infima are minima if the metric is proper[†]. As usual in geometry and number theory, we consider $\Gamma\backslash G$ instead of G/Γ; the latter is also often considered in dynamics. The two set-ups are equivalent via the bijection sending $\Gamma g \in \Gamma\backslash G$ to $g^{-1}\Gamma \in G/\Gamma$.

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 $n \to \infty$.

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

Lemma 1.1 (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

$$r = \tfrac{1}{4} \inf\nolimits_{\gamma\in\Gamma\smallsetminus\{I\}} \mathsf{d}_G(h^{-1}\gamma h, e) \tag{1.2}$$

has this property.

Proof. We first show this locally, for $K = \{x_0\}$ where $x_0 = \Gamma h$. Let r be as in (1.2), 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 $g_1 \in B_r^G$, and similarly

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

$$\begin{aligned}\mathsf{d}_G(h^{-1}\gamma h, e) &= \mathsf{d}_G(e, h^{-1}\gamma^{-1}h)\\ &\leqslant \mathsf{d}_G(e, g_2) + \mathsf{d}_G(g_2, h^{-1}\gamma^{-1}h)\\ &\leqslant r + \mathsf{d}_G(h^{-1}\gamma h g_2, e) < 4r.\end{aligned}$$

This 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 $y \in B^X_{r/2}(x_0)$ 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 $K = \{x_0\}$ as in the lemma (see also Exercise 1.1.3). If $x_0 = \Gamma h$ then

$$\tfrac{1}{4} \inf_{\gamma\in\Gamma} \mathsf{d}_G(h^{-1}\gamma h, e) \leqslant r_{x_0} \leqslant \inf_{\gamma\in\Gamma} \mathsf{d}_G(h^{-1}\gamma h, e) \tag{1.3}$$

by Lemma 1.1.

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 1.1. 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 1.2 (Fundamental domain). Let $\Gamma \leqslant G$ be a discrete subgroup. A *fundamental domain* $F \subseteq G$ 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 1.3. The set $[0,1)^d \subseteq \mathbb{R}^d$ is a fundamental domain for the discrete subgroup $\Gamma = \mathbb{Z}^d \leqslant \mathbb{R}^d = G$.

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

We will see more examples later, but the existence of a fundamental domain is a general property.

Lemma 1.4 (Existence of fundamental domains). *If Γ is a discrete subgroup of G and $B_{inj} \subseteq B_{surj} \subseteq G$ are injective (resp. surjective) sets, 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 $F \subseteq G$.*

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 1.1, 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:

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

and so on. Then $F = \bigsqcup_{n=0}^{\infty} F_n$ 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 $\Gamma g = \pi_X^{-1}\bigl(\pi_X(g)\bigr)$ 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 $\mathbb{Z}^d < \mathbb{R}^d$, we will be able to give very 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 will arise 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.

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

1.1.2 Haar Measure and the Natural Action on the Quotient

Recall (see [53, Sec. 8.3] for an outline and [54, Sec. 10.1] or the monograph of Folland [64, 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 Exercise 1.1.5 and Exercise 1.1.6.

Lemma 1.5 (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.

Proof of Lemma 1.5. 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

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

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

as required. □

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

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

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

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

for all $g_1, g_2 \in G$. Also note that $g_2\boldsymbol{\cdot}g_1 = g_1 g_2^{-1}$ 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 $\pi_X(g_2\boldsymbol{\cdot}g_1) = g_2\boldsymbol{\cdot}\pi_X(g_1)$. We are interested in whether X supports a G-invariant probability measure, a property discussed in the next proposition and definition.

Proposition 1.6 (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 $m_X(g\boldsymbol{\cdot}B) = m_X(B)$ for all measurable $B \subseteq X$ and all g in G;*
(b) *There is a fundamental domain F for $\Gamma \leqslant G$ with $m_G(F) < \infty$;*
(c) *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.*

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

Definition 1.7 (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 $m_X(X) = 1$. 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 $m_G(Bg^{-1}) = \mathrm{mod}(g)m_G(B)$ for all measurable $B \subseteq G$ and $g \in G$ (see [54, Sec. 10.1] for the details and references).

The modular character may also be defined using a right Haar measure $m_G^{(r)}$ via $m_G^{(r)}(gB) = \mathrm{mod}(g)m_G^{(r)}(B)$ 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 [53, Th. 2.21] and Exercise 1.1.7 for the proof.

Theorem 1.8 (Poincaré recurrence). *Let X be a locally compact metric space, and let μ be a Borel probability measure preserved by 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 1.6. We will start by proving that (a) $\implies$ (c). 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

$$\int f \,\mathrm{d}\mu = \int \sum_{\pi(g)=x} f(g) \,\mathrm{d}m_X(x) \tag{1.4}$$

for any $f \in C_c(G)$. Here the function defined by the sum

$$F\colon x = \Gamma g \longmapsto \sum_{\gamma\in\Gamma} f(\gamma g),$$

on the right-hand side belongs to $C_c(X)$ — indeed the sum vanishes if x does not lie in $\pi(\operatorname{supp} f)$, and for every given $g \in G$ (and also on any compact neighborhood of g) the sum can be identified with a sum over a finite subset of Γ, which implies continuity.

By invariance of μ 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 11, or rather their analogues for right Haar measures). By the construction above, $m_G^{(r)}$ is left-invariant under Γ. Finally, (1.4) extends using dominated and monotone convergence to any measurable non-negative function f on G. Applying this to $f = \mathbb{1}_F$ for a fundamental domain $F \subseteq G$ shows that $m_G^{(r)}(F) = 1$, hence (c).

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

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

By Lemma 1.5 (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 (c) are equivalent.

We also note that (b) $\Longrightarrow$ (c) 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 1.5, $m_G(F) = m_G(Fg) = m_G(F) \operatorname{mod}(g^{-1})$, so G is unimodular and (c) follows.

In the proof that (a) (or, equivalently, (c)) implies (b), we will again show that G is unimodular. Note that this implies that (b) and (c) are the same statement. Also note that by the equivalence of (a) and (c) above and the uniqueness of Haar measures we know that the measure m_X on X is derived (up to a scalar) from the right Haar measure $m_G^{(r)}$ on G restricted to a fundamental domain $F \subseteq G$. Let $B = B_r^G \subseteq G$ be a compact neighborhood of the identity e in G so that $r > 0$ is an injectivity radius at $\Gamma e \in X$ as in Lemma 1.1. Then $m_X(\pi_X(B)) = m_G^{(r)}(B)$ by (1.4) (for $\mu = m_G^{(r)}$ and the characteristic function of B). By the properties of the Haar measure we have also $m_X(\pi_X(B)) = m_G^{(r)}(B) > 0$.

Let now g be an element of G; we wish to show that $\operatorname{mod}(g) = 1$, and only know that g preserves a finite measure m_X on X (which we may assume without loss of generality to be a probability measure). By Poincaré recurrence (Theorem 1.8) there exists some $b \in B$ and sequences $(n_k), (\gamma_k), (b_k)$ with

$$n_k \nearrow \infty, \gamma_k \in \Gamma, b_k \in B$$

such that

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

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

$$\operatorname{mod}(\Gamma) = \{1\}$$

by (c), we see that

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

belongs to a compact neighborhood of $1 \in (0, \infty)$ for all $k \geqslant 1$. It follows that $\operatorname{mod}(g) = 1$, as required. □

Proposition 1.9 (Haar measure on $X = \Gamma \backslash G$). *Let G and Γ be as in Proposition* 1.6*, and suppose in addition that G is unimodular. Then the*

Haar measure m_G *on* G *induces a locally finite* G*-invariant measure* m_X*, also called the* Haar measure *on* $X = \Gamma\backslash G$*, such that*

$$\int_G f \, \mathrm{d}m_G = \int_X \sum_{\gamma\in\Gamma} f(\gamma g) \, \mathrm{d}m_X(\Gamma g) \tag{1.5}$$

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

The formula (1.5) 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).

Proof of Proposition 1.9. Since we assume that G is unimodular, the argument that (c) implies (a) in the proof of Proposition 1.6 can be used to define the measure m_X. Once again Lemma 1.5 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 definition, (1.5) holds for $f = \mathbb{1}_B$ if $B \subseteq F$ or if $B \subseteq \gamma F$ for some $\gamma \in \Gamma$. By linearity (1.5) also holds for any measurable $B \subseteq G$ and hence for any simple function. In particular, the sum on the right-hand side of (1.5) is a measurable function on X (or equivalently on F). The measurability of the sum and the equality of the integrals now extend by monotone convergence to show that (1.5) holds for any measurable non-negative function. □

Notice that Lemma 1.1 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 1.7.

Definition 1.10 (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 1.1 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 we will soon see, this would exclude some of the most natural lattices and their quotient spaces.

1.1.3 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 1.11 (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 (1.3).

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 there is some N_1 with

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

for $n \geqslant N_1$. Now remove the terms $x_2, \dots, x_{N_1-1}$ 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

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

and

$$\bigsqcup_{n=1}^\infty g_n B_{\varepsilon/2}^G$$

† In the remainder of the section we collect more fundamental results about locally homogeneous orbits, but the reader in a hurry could also move on to Section 1.2 and return to the material here later as needed.

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 1.1 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. □

1.1.4 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, which we will write as $(x, g) \mapsto g{\scriptstyle\bullet}x$ for $x \in X$ and $g \in G$, the *H-orbit* of $x \in X$ is the set

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

where

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

is the *stabilizer subgroup* of $x \in X$ and the isomorphisms are sending $h{\scriptstyle\bullet}x$ 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*, $\operatorname{vol}(H{\scriptstyle\bullet}x)$ 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 1.12 (Finite volume orbits are closed). *Let $\Gamma \leqslant G$ be a discrete subgroup, and let $H \leqslant G$ be a closed subgroup. Suppose that the point $x \in X = \Gamma\backslash G$ has a finite volume H-orbit. Then $xH \subseteq X$ is closed and the map from $\operatorname{Stab}_H(x)\backslash H$ to $\Gamma\backslash G$ is proper.*

We note that Corollary 1.12 can also be shown directly (see Figure 1.1 and Exercise 1.1.9). However, it is also a quick corollary of Proposition 1.11.

PROOF OF COROLLARY 1.12. We first show the last claim of the corollary. Suppose therefore that the sequence $\big(\operatorname{Stab}_H(x)h_n\big)$ has $\operatorname{Stab}_H(x)h_n \to \infty$ as $n \to \infty$ in $Y = \operatorname{Stab}_H(x)\backslash H$. Since Y has finite volume, we may apply Proposition 1.11 to H to see that there exists a sequence (λ_n) in $\operatorname{Stab}_H(x)$ such that $h_n^{-1}\lambda_n h_n \to e$ as $n \to \infty$. Let $g \in G$ such that $x = \Gamma g$ and hence also $\operatorname{Stab}_H(x) = g^{-1}\Gamma g \cap H$. Then $\lambda_n = g^{-1}\gamma_n g$ 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 $n \to \infty$. Again by Proposition 1.11 this shows that $\Gamma g h_n \to \infty$ in $X = \Gamma \backslash G$ as $n \to \infty$. Since $\bigl(\mathrm{Stab}_H(x) h_n\bigr)$ was an arbitrary seqeunce in Y going to infinity, the properness of the embedding map from Y to X follows.

If now $xh_n \to z \in X$ as $n \to \infty$, then the sequence $\bigl(\mathrm{Stab}_H(x) h_n\bigr)$ in Y cannot go to infinity by the last paragraph. Choosing a subsequence (and re-labeling the sequence) we may assume that $\mathrm{Stab}_H(x) h_n \to \mathrm{Stab}_H(x) h$ as $n \to \infty$, which implies that $z = xh \in xH$. It follows that xH is closed. $\square$

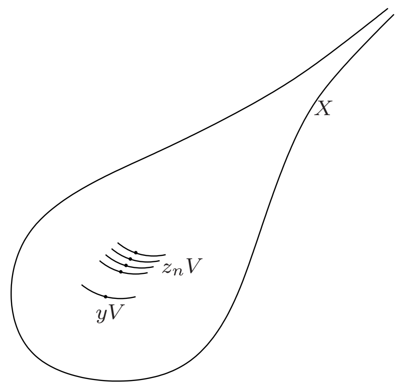

Fig. 1.1: We depict here an alternative to the proof of Corollary 1.12: By assuming (for the purposes of a contradiction) that the sets $z_n V \subseteq xH$ approach $yV \subseteq xH$ transverse to the orbit direction for a given neighbourhood V of $e \in H$, one can show that $\mathrm{vol}(xH) = \infty$.

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 1.6). 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 1.9 applied to $\mathrm{Stab}_H(x)\backslash H \cong xH$.

Proposition 1.13 (Closed orbits are embedded). *Let $\Gamma \leqslant G$ be a discrete subgroup and let $H \leqslant G$ be a closed subgroup. Suppose that the point $x \in X = \Gamma\backslash G$ has a closed H-orbit. Then $xH \subseteq X$ is* embedded*, meaning that the map $h \in \mathrm{Stab}_H(x)\backslash H \to xh \in xH$ is a homeomorphism. In particular,* vol_{xH} *is a locally finite measure on X.*

We note that for finite volume quotients the above also follows quickly from Corollary 1.12.

PROOF OF PROPOSITION 1.13. Clearly the map $\mathrm{Stab}_H(x)\backslash H \to xH \subseteq X$ is continuous, and we wish to show that its inverse is also continuous.

Replacing $x = \Gamma g$ and H simultaneously with Γ and gHg^{-1}, we may assume for simplicity that $x = \Gamma$ so that $\mathrm{Stab}_H(x) = \Gamma \cap H$.

By Exercise 1.1.2 (which also holds for G/H instead of $H\backslash G$) the quotient G/H is a locally compact complete metric space. We claim that our assumption that ΓH is closed in $\Gamma\backslash G$ also shows that ΓH is closed as a subset of G/H (this is actually an equivalence). Indeed, suppose that $(\gamma_n H)$ converges to gH in G/H. Then we can find a sequence (h_n) in H such that $\gamma_n h_n \to g \in G$ as $n \to \infty$, showing that $\Gamma h_n \to \Gamma g$. However, this implies by our assumption that $\Gamma g \in \Gamma H$, so that there is some $\gamma \in \Gamma$ and $h \in H$ with $g = \gamma h$. This shows that $gH = \gamma H \in \Gamma H$ as needed.

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

Now suppose that $\Gamma h_n \to \Gamma h$ as $n \to \infty$ in $\Gamma\backslash G$. Then there exists a sequence (γ_n) in Γ with $\gamma_n h_n \to h \in H$ as $n \to \infty$, which implies that

$$\gamma_n H \to H$$

as $n \to \infty$ in G/H. By the discreteness of $\Gamma H \subseteq G/H$, it follows that $\gamma_n \in H$ for large enough n, so that we also have

$$(\Gamma \cap H)\, h_n \longrightarrow (\Gamma \cap H)\, h$$

as $n \to \infty$ in $\Gamma \cap H\backslash H$.

For the last claim of the proposition notice that every compact set $K \subseteq X$ intersects xH in a compact set which has finite measure with respect to vol_{xH} (as $K \cap xH$ also corresponds to a compact set in $\mathrm{Stab}_H(x)\backslash H$). $\square$

Exercises for Section 1.1

Exercise 1.1.1. 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 1.1.2. Let $H < G$ be a closed subgroup. Imitate the definition in (1.1) 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 1.1.3. Show that the maximal injectivity radius as defined after Lemma 1.1 is indeed an injectivity radius. Show the upper bound in (1.3).

Exercise 1.1.4. 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 $\pi_X\colon G \to X$ (that is, the finest topology on X for which π_X is continuous).

Exercise 1.1.5. Show that the bi-invariant Haar measure $m_{\mathrm{GL}_d(\mathbb{R})}$ on the locally compact group

$$\mathrm{GL}_d(\mathbb{R}) = \bigl\{g = (g_{ij})_{i,j} \in \mathrm{Mat}_d(\mathbb{R}) \mid \det(g) \neq 0\bigr\},$$

which is called the *general linear group*, can be defined by the formula

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

Exercise 1.1.6. Let $d \geqslant 2$. Show that

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

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

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

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 $\mathbb{R}^{d^2}$.

Exercise 1.1.7. Show that Theorem 1.8 follows from the conventional formulation of Poincaré recurrence: if $(X, \mathscr{B}, \mu, T)$ is a measure-preserving system and $\mu(A) > 0$ then there is some $n \geqslant 1$ for which $\mu(A \cap T^{-n}A) > 0$ (see [53, Sec. 2.1]).

Exercise 1.1.8. Rephrase Proposition 1.11 as a compactness criterion characterizing compact subsets of $X = \Gamma\backslash G$ in terms of the injectivity radius.

Exercise 1.1.9. Prove Corollary 1.12 without using Proposition 1.11 by using Figure 1.1.

Exercise 1.1.10. Let $G < \mathrm{SL}_d(\mathbb{R})$ 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 [93]; also see Raghunathan [148, Ch. XI]. However, the proof for subgroups of the form $\Gamma = G \cap \mathrm{SL}_d(\mathbb{Z})$ is significantly easier.

Exercise 1.1.11. 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 [148, Remark 13.21] for the general case.

Exercise 1.1.12. 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 1.1.13. 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).

1.2 A Brief Review of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$

1.2.1 The Space

We recall (see, for example, [53, 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 $(z, u), (z, v) \in \mathrm{T}_z\mathbb{H} = \{z\} \times \mathbb{C}$ 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

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \mapsto g \bullet z = \frac{az + b}{cz + d}. \tag{1.6}$$

The stabilizer of $\mathrm{i} \in \mathbb{H}$ is $\mathrm{SO}(2)$ so that

$$\mathrm{SL}_2(\mathbb{R}) / \mathrm{SO}(2) \cong \mathbb{H}$$

under the map sending $g\,\mathrm{SO}(2)$ to $g \bullet \mathrm{i}$.

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

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

where

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

This action gives rise to the simply transitive action of

$$\mathrm{PSL}_2(\mathbb{R})=\mathrm{SL}_2(\mathbb{R})/\{\pm 1\}$$

on the unit tangent bundle

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

so that

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

by sending g to $\mathrm{D}\,g(\mathrm{i},\uparrow)$, where we write $\uparrow$ for the upward pointing vector of length 1 at any $z\in\mathbb{H}$.

The shaded region E in Figure 1.2 is a fundamental region for the action of the discrete subgroup $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{H}$ (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 comply with this tradition), see Exercise 1.2.4.

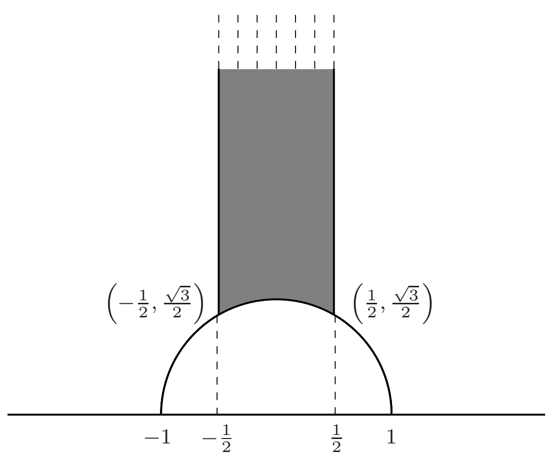

Fig. 1.2: A fundamental domain $E\subseteq\mathbb{H}$ for the action of $\mathrm{SL}_2(\mathbb{Z})$.

This shows that we can define a fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$ in

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

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

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

(Once again, strictly speaking we should describe more carefully which vectors attached to points $z \in \partial E$ are allowed in F.) Furthermore, we can lift the set $F \subseteq \mathrm{PSL}_2(\mathbb{R})$ to a surjective set $F \subseteq \mathrm{SL}_2(\mathbb{R})$ for $\mathrm{SL}_2(\mathbb{Z})$. 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 [53, Ch. 9] (which we will prove in greater generality for $\mathrm{SL}_d(\mathbb{R})$ in Section 1.3.4):

- $\mathrm{SL}_2(\mathbb{R})$ is unimodular (see Exercise 1.1.6).
- $\mathrm{SL}_2(\mathbb{R}) = NAK$ with†

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

  and $K = \mathrm{SO}(2)$, in the sense that every $g \in \mathrm{SL}_2(\mathbb{R})$ can be written uniquely(2) as a product $g = nak$ with $n \in N$, $a \in A$ and $k \in K$.
- Let $B = NA = AN$ be the subgroup $B = \left\{\begin{pmatrix} a & t \\ & a^{-1} \end{pmatrix} \mid a > 0, t \in \mathbb{R}\right\}$. 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 (with the constant of proportionality dependent only on the choices of 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

† 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.

$$m_{\mathrm{SL}_2(\mathbb{R})}(F) < \int_{-1/2}^{1/2} \int_{\sqrt{3}/2}^{\infty} \int_{0}^{2\pi} \frac{1}{y^2} \, \mathrm{d}\theta \, \mathrm{d}y \, \mathrm{d}x < \infty.$$

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.

1.2.2 The Geodesic Flow — the Subgroup A

We recall that

$$g_t \colon x \longmapsto x \begin{pmatrix} \mathrm{e}^{t/2} & \\ & \mathrm{e}^{-t/2} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{-t/2} & \\ & \mathrm{e}^{t/2} \end{pmatrix} \cdot x$$

defines the geodesic flow on X_2, whose orbits may also be described in the fundamental region as in Figure 1.3.

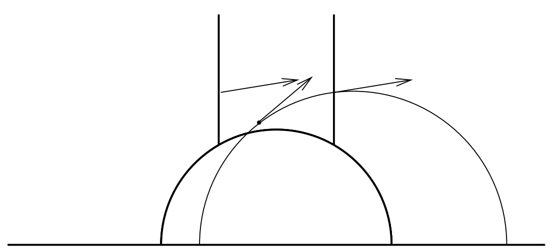

Fig. 1.3: The geodesic flow follows the circle determined by the arrow which intersects $\mathbb{R} \cup \{\infty\} = \partial\mathbb{H}$ normally, and is moved back to F via a Möbius transformation in $\mathrm{SL}_2(\mathbb{Z})$ once the orbit leaves F.

The diagonal subgroup

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

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

is also called the *torus* or *Cartan subgroup*. We recall that A acts ergodically on X_2 with respect to the Haar measure m_{X_2} (see [53, Sec. 9.5]; we will also discuss this 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 $\mathrm{SL}_2(\mathbb{Z})g_{\mathrm{golden}}A$ is compact, where the matrix $g_{\mathrm{golden}} \in K$ has the property† that

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

  Now notice that

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

  This identity shows that the orbit $\mathrm{SL}_2(\mathbb{Z})g_{\mathrm{golden}}A$ is compact (see also Figure 1.4 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.

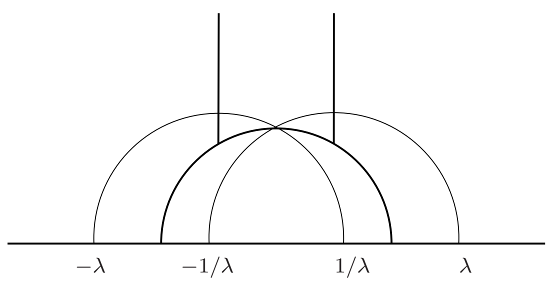

Fig. 1.4: The union of the two geodesics considered in X_2 with both directions allowed is a periodic A-orbit, and comprises the orbit $\mathrm{SL}_2(\mathbb{Z})g_{\mathrm{golden}}A$.

† The eigenvalues of $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ are $\frac{3\pm\sqrt{5}}{2}$, and there is such a matrix $g_{\mathrm{golden}} \in K$ because $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$ is symmetric.

Finally we would like to point out — in a sense to be made precise in Sections 3.1 and 3.5 — 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 1.2.6 and 1.2.7).

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

$$h_s \colon x \longmapsto x \begin{pmatrix} 1 & -s \\ & 1 \end{pmatrix} = u(s) \bullet x$$

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

$$\begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} = u(s)$$

are unipotent (that is, only have 1 as an eigenvalue) and the corresponding subgroup

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

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

$$U^- = \left\{ g \in \mathrm{SL}_2(\mathbb{R}) \mid \begin{pmatrix} \mathrm{e}^{-t/2} & \\ & \mathrm{e}^{t/2} \end{pmatrix} g \begin{pmatrix} \mathrm{e}^{t/2} & \\ & \mathrm{e}^{-t/2} \end{pmatrix} \to I_2 \text{ as } t \to \infty \right\}.$$

This implies that

$$\mathsf{d}\left(g_t(x), g_t(u(s) \bullet x)\right) \to 0$$

as $t \to \infty$ for any $x \in \mathsf{X}_2$ and $s \in \mathbb{R}$, see Exercise 1.1.1.

Geometrically, the horocycle orbits $U^- \bullet x = xU^-$ can be described as circles 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 1.5.

We recall that U^- also acts ergodically on X_2 with respect to the Haar measure m_{X_2} (see [53, Sec. 11.3] and 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 are as follows:

- Compact trajectories, for example $\mathrm{SL}_2(\mathbb{Z})U^-$ is compact and corresponds to the horizontal orbit through $(\mathrm{i}, \uparrow) \in \mathrm{T}^1\mathbb{H}$.

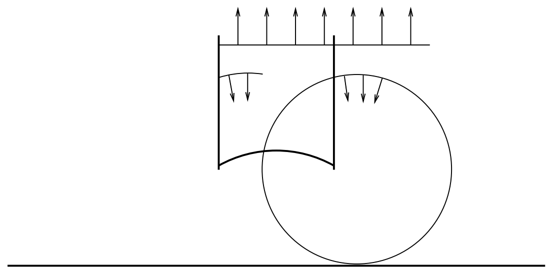

Fig. 1.5: The picture shows the two types of horocycle orbits; the orbits in X_2 can again be understood by using the appropriate Möbius transformation whenever the orbit leaves the fundamental domain.

- Dense trajectories, which are automatically also equidistributed with respect to m_{X_2}.

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 1.2.8).

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

- $K = \mathrm{SO}(2) \subseteq \mathrm{SL}_2(\mathbb{R})$, and
- $B = U^-A = \left\{ \begin{pmatrix} a & s \\ & a^{-1} \end{pmatrix} \mid a > 0, s \in \mathbb{R} \right\}$

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^- is a much longer periodic orbit. However, long periodic U^--orbits equidistribute in X_2 (see Sarnak [161] 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 in Chapters 3 and 4.

Exercises for Section 1.2

Exercise 1.2.1. Show that the action of $K = \mathrm{SO}(2)$ on $\mathrm{T}_{\mathrm{i}}\mathbb{H}$ rotates the tangent vectors at 'double speed'. That is,

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

applied to $(\mathrm{i}, v) \in \mathrm{T}_1\mathbb{H}$ gives $(\mathrm{i}, \mathrm{e}^{-2\theta\mathrm{i}}v) \in \mathrm{T}_1\mathbb{H}$.

Exercise 1.2.2. 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 1.2.3. (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 be $\{x + \mathrm{i} \mid x \in [-\frac{1}{2}, \frac{1}{2})\}$.

Exercise 1.2.4. Let E be as in Figure 1.2.

(1) 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 $\mathrm{SL}_2(\mathbb{Z})\bullet E = \mathbb{H}$.
(2) Suppose that both z and $\gamma\bullet z$ lie in E for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. Show that either $\gamma = \pm e$ or $z \in \partial E$.
(3) Conclude that E can be modified (by defining which parts of the boundary of E should be included) to become a fundamental domain.

Exercise 1.2.5. Show that $\mathrm{SL}_2(\mathbb{R})$ is generated by the unipotent subgroups

$$\left\{\begin{pmatrix} 1 & * \\ & 1 \end{pmatrix}\right\} \text{ and } \left\{\begin{pmatrix} 1 & \\ * & 1 \end{pmatrix}\right\}.$$

Exercise 1.2.6. 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 1.2.7. 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.

Exercise 1.2.8. 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 any other compact orbit is of the form $\mathrm{SL}_2(\mathbb{Z})gaU^-$ for some $a \in A$.

Exercise 1.2.9. Show that $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R}) \cong \{\mathbb{Z}^2 g \mid g \in \mathrm{SL}_2(\mathbb{R})\}$ can be identified with lattices $\mathbb{Z}^2 g \subseteq \mathbb{R}^2$ of co-volume $\det g = 1$. 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 $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ in terms of elements of the lattices $\mathbb{Z}^2 g$ for $\mathrm{SL}_2(\mathbb{Z})g \in K$. More precisely, calculate the relationship between the shortest vector $ng \in \mathbb{Z}^2 g$ and the imaginary part of $g\mathrm{i} \in \mathbb{H}$ under the assumption that the representative $g \in \mathrm{SL}_2(\mathbb{R})$ has been chosen with $g\mathrm{i} \in E$ (with $E \subseteq \mathbb{H}$ as in Figure 1.2).

1.3 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

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

which gives 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.

1.3.1 Basic Definitions

A lattice in $\mathbb{R}^d$ in the sense of Definition 1.7 has the form $\Lambda = \mathbb{Z}^d g$ for some $g \in \mathrm{GL}_d(\mathbb{R})$ (see Exercise 1.3.1). A fundamental domain for Λ is given by the parallelepiped $[0,1)^d g$ which is spanned by the row vectors of g, and has Lebesgue measure $|\det g|$. This measure is also called the *covolume* $\mathrm{covol}(\Lambda)$ of Λ. A lattice $\Lambda \subseteq \mathbb{R}^d$ is called *unimodular* if the co-volume is 1. The space of all unimodular lattices in $\mathbb{R}^d$ — the *moduli space of lattices* — is therefore

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

which is the orbit of $\mathbb{Z}^d$ under the right 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 right action sends (g, B) to $Bg = \{vg : v \in B\}$. Notice that

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

so that

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

where $\mathrm{SL}_d(\mathbb{Z})g$ corresponds to the lattice $\mathbb{Z}^d g$. 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 1.1. To understand X_d better, we need to develop a better understanding of lattices in $\mathbb{R}^d$.

1.3.2 Geometry of Numbers

The next result will be almost immediate from the abstract results in Section 1.3.1. It is a weak form of a classical result due to Minkowski in 1896 (see [135] for a modern reprinting).

Theorem 1.14 (Minkowski's first theorem). *If $\Lambda \subseteq \mathbb{R}^d$ is a lattice of co-volume 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 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 particular value of the constants appearing in Theorems 1.14 and 1.15 will not be important for our purposes.

Proof of Theorem 1.14. 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 1.2. 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. We refer to the monographs of Cassels [13] or Gruber and Lekkerkerker [73] for thorough accounts of the topic and its history. For our purposes the following result, a consequence of the reduction algorithm of Korkine and Zolotareff [105, 106, 107], will suffice. The minima defined below are sometimes referred to as *Minkowski's successive minima*.

Theorem 1.15 (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

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

Moreover, if

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

then

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

for $1 \leqslant k \leqslant d$.

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 co-volume $\operatorname{covol}(\Lambda \cap V)$ of $\Lambda \cap V$ in V is finite. In the second case, we write $\operatorname{covol}(\Lambda \cap V) = \infty$. 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.

The proof of Theorem 1.15 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 $\mathbb{R}^d$. We note that the minimum in the definition of $\alpha_k(\Lambda)$ is indeed achieved for any k, see Exercise 1.3.4.

Proof of Theorem 1.15. We use induction on the dimension d. For $d = 1$ (and so also $k = 1$), it is clear that

$$\lambda_1(\Lambda) = \alpha_1(\Lambda) = \operatorname{covol}(\Lambda).$$

Assume therefore that the theorem 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 $\pi : \mathbb{R}^d \to W$ be the orthogonal projection along $\mathbb{R}v_1$ onto W. We claim that $\Lambda_W = \pi(\Lambda) \subseteq W$ is a discrete subgroup in W such that all of its nonzero 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.

Next we claim that Λ_W is a lattice and that

$$\lambda_k(\Lambda_W) \asymp \lambda_{k+1}(\Lambda) \tag{1.7}$$

for $k = 1, \dots, d-1$. 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 exists 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

$$\operatorname{covol}(\Lambda) = \lambda_1(\Lambda)\operatorname{covol}(\Lambda_W). \tag{1.8}$$

This shows that Λ_W is a lattice in W.

Let $k \in \{1, \dots, d-1\}$. 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 $\lambda_1(\Lambda) \ll \lambda_1(\Lambda_W) \leqslant \lambda_k(\Lambda_W)$.

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 (1.8) this gives $\operatorname{covol}(\Lambda) \asymp \lambda_1(\Lambda) \cdots \lambda_d(\Lambda)$ as claimed in the theorem.

To see the last statement in the theorem, we proceed similarly. If $v_j \in \Lambda$ has norm $\lambda_j(\Lambda)$ for $j = 1, \dots, k$, $v_1, \dots, v_k$ are linearly independent (over $\mathbb{R}$), and $V = \mathbb{R}v_1 + \cdots + \mathbb{R}v_k$ 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 $V \subseteq \mathbb{R}^n$ has dimension k and is Λ-rational, then we may apply the above to the lattice $\Lambda \cap V$ in V to get

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

which 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 1.16 (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 Λ 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 1.16 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 1.3.5 for a simple counterexample.

Proof of Corollary 1.16. 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 1.15. Recall that these assumptions lead to (1.7). 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 1.15. 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

$$\|v_2\| \asymp \|w_1\| \asymp \lambda_2(\Lambda),$$

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. □

1.3.3 Mahler's Compactness Criterion

The space $\mathsf{X}_d = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$ 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 $\Lambda_n = \mathbb{Z}^d g_n \to \mathbb{Z}^d g$ as $n \to \infty$. Then (after replacing g_n with $\gamma_n g_n$ 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. (1.1) on page 8 and the following discussion). Thus we can write $g_n = g h_n$ with $h_n \to I_d$ as $n \to \infty$, which implies that $\lambda_1(\mathbb{Z}^d g_n) \to \lambda_1(\mathbb{Z}^d g) > 0$ as $n \to \infty$ (see Exercise 1.3.2).

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 1.17 (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*

$$\Lambda \in B \implies \lambda_1(\Lambda) \geqslant \delta. \tag{1.9}$$

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 (1.9) will also be described by saying that elements of B do not contain any non-trivial δ-short vectors. An equivalent formulation of Theorem 1.17 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 1.6, in part to express the meaning of Theorem 1.17.

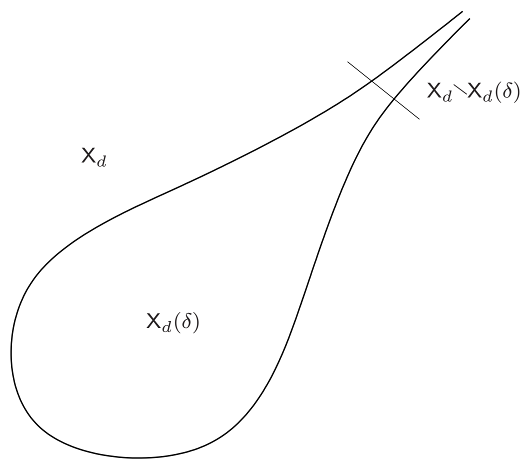

Fig. 1.6: A compact subset of X_d is contained in $\mathsf{X}_d(\delta) = \{\Lambda \in \mathsf{X}_d \mid \lambda_1(\Lambda) \geqslant \delta\}$ for some $\delta > 0$. The non-compact part $\mathsf{X}_d \smallsetminus \mathsf{X}_d(\delta)$, loosely referred to as a *cusp*, is depicted as a thin set to indicate the finite total volume (see Theorem 1.18). For $d > 2$ the geometry of the cusp is much more complicated than the cusp in the $d = 2$ case.

PROOF OF THEOREM 1.17. We have already mentioned that λ_1 is a continuous function on X_d (see Exercise 1.3.2). 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 $(\mathbb{Z}^d g_n)$ in $\mathsf{X}_d(\delta)$ be any sequence. Then, by Corollary 1.16, the lattice $\mathbb{Z}^d g_n$ has a basis $v_1^{(n)}, \dots, v_d^{(n)}$ with

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

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

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

as $k \to \infty$, so that $\mathrm{SL}_d(\mathbb{Z})g_{n_k} \to \mathrm{SL}_d(\mathbb{Z})g$ as required. □

1.3.4 X_d has Finite Volume

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

Theorem 1.18 (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 1.16 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 $\pi(F) = \mathsf{X}_d$ 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 NAK decomposition).

Proposition 1.19 (Iwasawa decomposition). *Let* $K = \mathrm{SO}(d)$ *and*

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

where

$$U = N = \left\{ \begin{pmatrix} 1 & & & \\ u_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ u_{d1} & u_{d2} & \cdots & 1 \end{pmatrix} \right\}$$

and

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

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

Proof. This is the Gram–Schmidt procedure[(3)] in disguise. Let

$$g = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix},$$

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

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

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

$$\widetilde{w_2} = u_{21} w_1 + w_2$$

with $u_{21} \in \mathbb{R}$ 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

$$\widetilde{w_d} = u_{d1} w_1 + u_{d2} w_2 + \dots + w_d$$

with $u_{d1}, u_{d2}, \dots, u_{d(d-1)} \in \mathbb{R}$ 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

$$u = \begin{pmatrix} 1 & & & \\ u_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ u_{d1} & u_{d2} & \cdots & 1 \end{pmatrix}$$

and

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

then

$$ug = \begin{pmatrix} w_1 \\ \widetilde{w_2} \\ \vdots \\ \widetilde{w_d} \end{pmatrix}, a^{-1} ug = \begin{pmatrix} w_1' \\ \vdots \\ w_d' \end{pmatrix} = k.$$

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 $g = u^{-1}ak$.

To see that this decomposition is unique, notice that B is a subgroup with $B \cap K = \{I_d\}$ so that $b_1k_1 = b_2k_2$ implies $b_2^{-1}b_1 = k_2k_1^{-1} = I_d$. Similarly, $A \cap U = \{I_d\}$, and the proposition follows. □

Our geometric arguments in the proof of Theorem 1.15 and Corollary 1.16 are closely related to the Gram–Schmidt procedure used in Proposition 1.19. Combining these gives the next result.

Definition 1.20 (Siegel domain for X_d). A set of the form

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

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

$$U_s = \left\{ \begin{pmatrix} 1 & & & \\ u_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ u_{d1} & u_{d2} & \cdots & 1 \end{pmatrix} \mid |u_{ij}| \leqslant s \right\},$$

and

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

is called a *Siegel domain*.

We note that U_s is a compact subset of the lower unipotent subgroup but A_t is a non-compact subset of the diagonal subgroup.

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 1.21 (Surjectivity of Siegel domains). *There exists some t_0 such that for $t \leqslant t_0$ and $s \geqslant \frac{1}{2}$ the Siegel domain $\Sigma_{s,t}$ is surjective (that is, $\pi(\Sigma_{s,t}) = \mathsf{X}_d$).*

A more careful analysis of the proof shows that $t_0 = \frac{\sqrt{3}}{2}$ suffices in any dimension; see also Exercise 1.3.10 which can also be used to prove this claim.

PROOF OF COROLLARY 1.21. Let $\Lambda \in \mathsf{X}_d$ be a unimodular lattice, and let $w_1, \dots, w_d$ be the $\mathbb{Z}$-basis as in Corollary 1.16. Replacing w_d by $-w_d$ if necessary, we may assume that $\det(g) = 1$, where

$$g = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}.$$

Now apply the Gram–Schmidt procedure as in the proof of Proposition 1.19 to g. By Corollary 1.16 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 and $t \leqslant t_0$ accordingly gives

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

Therefore $\Lambda = \mathbb{Z}^d g$ and $g = uak$ with $u \in U$ and $k \in K$. Notice that by replacing g by $u_{\mathbb{Z}} g$ with $u_{\mathbb{Z}} \in U(\mathbb{Z}) = U \cap \operatorname{Mat}_d(\mathbb{Z})$ we can easily ensure that $u_{(i+1)i} \in [-\frac{1}{2}, \frac{1}{2})$. Having achieved this we may use another $u_{\mathbb{Z}} \in U(\mathbb{Z})$ with $(u_{\mathbb{Z}})_{(i+1)i} = 0$ for $i = 1, \ldots, d-1$, which makes it easy to calculate the next off-diagonal of $u_{\mathbb{Z}} u$ as follows:

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

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

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

for some $u \in U_{1/2}$, $a \in A_t$, 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 1.22 (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 neighborhood 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 [53, Lem. 11.31], [54, Lem. 10.57], and Knapp [103] for the proof. The above lemma is useful for us because of the following.

Lemma 1.23. $\mathrm{SL}_d(\mathbb{R})$ *is unimodular.*

As an alternative to Exercise 1.1.6 (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 $\mathbb{K}$, generalizing Exercise 1.2.5.

Lemma 1.24 (Unipotent Generation). *Over any field* $\mathbb{K}$*, the special linear group* $\mathrm{SL}_d(\mathbb{K})$ *is generated by the elementary unipotent subgroups*

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

with $i \neq j$ *and* E_{ij} *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 $U_{ij}(\mathbb{R})$ and $U_{ij}(\mathbb{C})$ 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 1.24. 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 $u_{ij}(t) \in U_{ij}(\mathbb{K})$ 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 $g_{12} \neq 0$ with a suitable row operation, then use another row operation to ensure that $g_{11} = 1$. 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 1.23. Recall the unipotent subgroups

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

for $i \neq j$ from Lemma 1.24. Let $a \in A$ be any diagonal matrix, and notice that $au_{ij}(t)a^{-1} = u_{ij}(\frac{a_i}{a_j}t)$ for $t \in \mathbb{R}$. Therefore, the commutator satisfies

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

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

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

contains U_{ij} for all $i \neq j$. By Lemma 1.24 it follows that

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

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

To complete the proof of Theorem 1.18, it remains to show the following lemma.

Lemma 1.25. *For any $s > 0$ and $t > 0$, we have $m_{\mathrm{SL}_d(\mathbb{R})}(\Sigma_{s,t}) < \infty$.*

Proof. Using Lemma 1.22 for $G = \mathrm{SL}_d(\mathbb{R})$, $S = B$, and $T = K$ we see that K can be ignored and we have to calculate $m_B(U_s A_t)$ (where as usual m_B denotes the left Haar measure on B). Note that $B = UA$ is not unimodular so that we cannot apply Lemma 1.22 again (indeed, applying it erroneously would not give the desired result). On the other hand, U and A are unimodular (see Exercise 1.3.8). Furthermore, the left Haar measure on B is given by a density function $\rho(a)$ with respect to $m_U \times m_A$ (using the coordinate system arising from $B = UA$). In fact

$$\mathrm{d}m_B \propto \rho(a)\,\mathrm{d}m_U \times \mathrm{d}m_A, \tag{1.10}$$

where

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

Using the fact that the Haar measure on U is simply the Lebesgue measure (in the coordinate system implied by the way we write down these matrices) and that A normalizes U, the relation in (1.10) can be checked directly (see Exercise 1.3.9).

Using this, we get

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

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

Using the relations

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

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

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

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

Next notice that

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

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 [9] (see Siegel [174]). In Chapter 4 we will give a second proof which will also lead to the general result for other groups in Chapter 7.

Exercises for Section 1.3

Exercise 1.3.1. Check that any lattice in $\mathbb{R}^d$ (in the sense of Definition 1.7) is indeed of the form $\mathbb{Z}^d g$ 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 $\mathbf{n} \in \mathbb{Z}^d$ with

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

Exercise 1.3.2. (1) Show that $\lambda_1(\mathbb{Z}^d gh) \leqslant \lambda_1(\mathbb{Z}^d g)\|h\|$ for $g, h \in \mathrm{GL}_d(\mathbb{R})$, where $\|\cdot\|$ denotes the operator norm.
(2) Conclude that $\lambda_1 : \mathsf{X}_d \to (0, \infty)$ is continuous.
(3) Generalize (2) to λ_k for $1 \leqslant k < d$.

Exercise 1.3.3. Suppose that $\Lambda_n = \mathbb{Z}^d g_n \to \Lambda = \mathbb{Z}g$ as $n \to \infty$ in the sense of the quotient X_d and its metric defined by (1.1). Show that

[†] The symbol $\propto$ denotes proportionality, and here the constant of proportionality depends on the choices of the Haar measures on A and on $\mathbb{R}^{d-1}$.

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

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

Exercise 1.3.4. Show that the minimum in the definition of $\alpha_k(\Lambda)$ in Theorem 1.15 is indeed achieved.

Exercise 1.3.5. 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.

Exercise 1.3.6. 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 1.3.7. Define for every $\Lambda \in \mathsf{X}_d$ the *covering radius* by

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

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

Exercise 1.3.8. Prove that U and A are unimodular (and describe their Haar measures).

Exercise 1.3.9. Let $B = UA$, m_B, m_U, m_A, and ρ be as in the proof of Lemma 1.25. Let $f \geqslant 0$ be any measurable function on B, and fix some $b \in B$. Using Fubini's theorem and substitution prove that

$$\int_B f(bua)\rho(a)\,\mathrm{d}m_U(u)\,\mathrm{d}m_A(a) = \int_B f(ua)\rho(a)\,\mathrm{d}m_U(u)\,\mathrm{d}m_A(a),$$

first for $b = u_0 \in U$ and then for $b \in A$. Deduce that (1.10) holds.

Exercise 1.3.10 (LLL algorithm$^{(4)}$). In this exercise a different proof of Corollary 1.21 will be given (which will not use Minkowski's theorem on successive minimas). For this let $v_1, \dots, v_d$ be an ordered basis of a unimodular lattice $\Lambda < \mathbb{R}^d$. For every $i = 1, \dots, d$ 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 NAK-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 N-part of g in the NAK-decomposition belongs to $U_{\frac{1}{2}}$).

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.

(1) 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.
(2) Check if the basis is t-reduced. If so, the algorithm terminates.

(3) 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

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

Exercise 1.3.11 (Siegel formula). For any $f \in C_c(\mathbb{R}^d)$ we define the Siegel transform at $x \in \mathsf{X}_d$ by

$$S_f(x) = \sum_{v \in \Lambda_x \smallsetminus \{0\}} f(v),$$

where $\Lambda_x = \mathbb{Z}^d g$ denotes the lattice corresponding to $x = \mathrm{SL}_d(\mathbb{R})g$. In this exercise we wish to show that there exists some $c > 0$ (depending on the choice of Haar measures) such that $\int_{\mathsf{X}_d} S_f \,\mathrm{d}m_{\mathsf{X}_d} = c \int_{\mathbb{R}^d} f(t) \,\mathrm{d}t$ for all $f \in C_c(\mathbb{R}^d)$.

(1) Show that $\int_{\mathsf{X}_d} S_f \,\mathrm{d}m_{\mathsf{X}_d} < \infty$.
(2) Show that the positive measure μ on $\mathbb{R}^d$ defined by Riesz representation and the functional $f \mapsto \int_{\mathsf{X}_d} S_f \,\mathrm{d}m_{\mathsf{X}_d}$ satisfies $\mu(\{0\}) = 0$.
(3) Show that μ is $\mathrm{SL}_d(\mathbb{R})$-invariant and conclude the claim.

Notes to Chapter 1

(1)(Page 19) 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 [94, Sec. 6.A].

(2)(Page 23) This is a simple instance of the more general Iwasawa decomposition of a connected real semi-simple Lie group [85] (see also [103]).

(3)(Page 36) This method was presented by E. Schmidt [162, Sec. 3, p. 442], and he pointed out that essentially the same method was used earlier by Gram [72]; the modern view is that the methods differ, and that the Gram form was used earlier by Laplace [112, p. 497ff.] in a different setting.

(4)(Page 43) This is based on the so-called LLL algorithm of A. K. Lenstra, H. W. Lenstra, Jr., and Lovász [117].