Chapter 4
Quantitative Non-Divergence

In this chapter we will show that a unipotent trajectory cannot diverge to infinity in $\mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. In fact we will show that unipotent orbits have no 'escape of mass', which is also called 'quantitative non-divergence'. The former non-divergence result was shown by Margulis [126] in his work on the arithmeticity of lattices, and the quantitative refinement is due to Dani [21, 24, 25]. About 20 years later, the argument was further refined by Kleinbock and Margulis [101] and Kleinbock [97], and applied to various Diophantine problems. As a corollary we will also obtain a special case of the Borel–Harish-Chandra theorem [9]: $\mathbb{G}(\mathbb{Z})$ is a lattice in $\mathbb{G}(\mathbb{R})$ if $\mathbb{G}$ is a semi-simple algebraic group defined over $\mathbb{Q}$.

4.1 The Case of $\mathbf{SL_2}(\mathbb{Z})\backslash \mathbf{SL_2}(\mathbb{R})$.

We first describe a case that is both easy and familiar: horocycle orbits on

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

We refer to Section 1.2 or [53, Ch. 9] for the background and to [53, Ch. 11] for a more detailed proof.

4.1.1 A Topological Claim

In the hyperbolic description of X_2, the topological non-divergence claim is particularly easy to see.

Lemma 4.1 (Non-divergence for X_2). *For any $x \in \mathsf{X}_2$, the horocycle orbit $u_t \bullet x$ does not go to infinity as $t \to \infty$, nor as $t \to -\infty$.*

PROOF. Every $x \in \mathsf{X}_2$ corresponds to a point $(z, v) \in \mathrm{T}^1(\mathbb{H})$ with z chosen in the usual fundamental domain, which we denote by F, for $\mathrm{SL}_2(\mathbb{Z})$ in $\mathbb{H}$ (see Section 1.2). To prove the lemma we find for a given x a compact set K and a sequence $t_n \to \infty$ with $u_{t_n} \cdot x \in K$ for all $n \geqslant 1$. If x is periodic under the action of $\{u_t \mid t \in \mathbb{R}\}$ then the orbit is compact and we may take

$$K = \{u_t \cdot x \mid t \in \mathbb{R}\}$$

and obtain the claim trivially. Otherwise, we may take

$$K = \{(z, v) \mid z \in F, \Im(z) \leqslant 1\}.$$

Then it is easy to see (from the geometric picture of the horocycle flow) that there exists some $t_1 \geqslant 0$ with $u_{t_1} \cdot x \in K$, as illustrated in Figure 4.1. In fact,

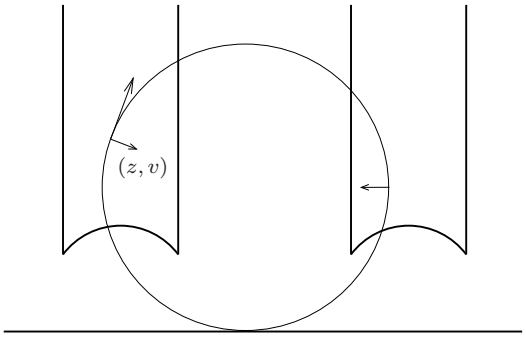

Fig. 4.1: A horocycle orbit returns to K.

the horocycle orbit is a circle touching $\mathbb{R}$. Hence it moves up and then down again, returning to K. Now consider the point $u_{t_1+1} \cdot x$, and apply the same argument to find some $t_2 \geqslant t_1 + 1$ with $u_{t_2} \cdot x \in K$. Repeating the argument proves the lemma by induction. □

4.1.2 Non-escape of Mass

While the topological statement in Lemma 4.1 above was easy to derive from the hyperbolic geometry of horocycle orbits, the quantitative claim is more difficult to see from this geometric picture. Hence we will switch the description and think of X_2 as the space of unimodular lattices in $\mathbb{R}^2$.

Proposition 4.2 (Quantitative non-divergence for X_2). *A point $x \in \mathsf{X}_2$ is either periodic for the horocycle flow or*[†] *has the property that there exists some $T_x \geqslant 0$ such that for all $\varepsilon > 0$ and all $T \geqslant T_x$ we have*

$$\frac{1}{T}\,|\{t \in [0,T] \mid u_t{\bullet}x \notin \mathsf{X}_2(\varepsilon)\}| \ll \varepsilon. \tag{4.1}$$

Here we are using $|A|$ as a shorthand for the Lebesgue measure of a subset $A \subseteq \mathbb{R}$, and the notation

$$\mathsf{X}_2(\varepsilon) = \{x \in \mathsf{X}_2 \mid \lambda_1(x) \geqslant \varepsilon\}$$

introduced in Section 1.3.3.

Proof of Proposition 4.2. Suppose that x is not periodic, and assume first that the lattice Λ_x associated to x has no vectors of length less than 1. Fix $T \geqslant 0$ and define, for every vector $v \in \Lambda_x \smallsetminus \{0\}$ a 'protecting' intervals

$$P_v = \{t \in [0,T] \mid \|vu_t^{-1}\| < 1\}.$$

Notice that if $v = (v_1, v_2)$, then

$$\begin{aligned} \|vu_{-t}\| &= \|(v_1, v_2 - tv_1)\| \\ &= \sqrt{v_1^2 + (v_2 - tv_1)^2}, \end{aligned} \tag{4.2}$$

and so P_v is a subinterval of $[0,T]$. If $v \in \Lambda_x \smallsetminus \{0\}$ is large enough (how large depends on T), then P_v is trivial. Hence there are only finitely many non-trivial intervals. As the unimodular lattice $\Lambda_x u_{-t}$ cannot contain two linearly independent vectors of length strictly less than 1, these intervals can only intersect if they are associated to linearly dependent vectors. To rule even this out, we choose within every Λ-rational line (that is, every line $\mathbb{R}v$ with $v \in \Lambda_x \smallsetminus \{0\}$) one and only one primitive vector in the lattice (that is, a vector $v \in \Lambda_x \smallsetminus \{0\}$ with $\mathbb{R}v \cap \Lambda_x = \mathbb{Z}v$). Let $v^{(1)}, \ldots, v^{(n)}$ be the resulting list of pairwise linearly independent primitive vectors, so that $P_i = P_{v^{(i)}}$ and

$$P_1 \sqcup \cdots \sqcup P_n = \{t \in [0,T] \mid \lambda_1(u_t{\bullet}x) < 1\}. \tag{4.3}$$

Now let $\varepsilon \geqslant 0$ and define the 'bad' set

$$B_i^\varepsilon = \{t \in [0,T] \mid \|v^{(i)}u_{-t}\| \leqslant \varepsilon\}$$

for $i = 1, \ldots, n$. We see that

$$B_1^\varepsilon \sqcup \cdots \sqcup B_n^\varepsilon = \{t \in [0,T] \mid \lambda_1(u_t{\bullet}x) \leqslant \varepsilon\}$$

[†] Note that the distinction of the two cases is absolutely necessary here: If $U{\bullet}x$ is a periodic orbit that is stuck high up in the cusp (equivalently a periodic orbit of short period), then the estimate (4.1) cannot hold uniformly for all $\varepsilon \leqslant 1$.

is precisely the set whose measure we wish to estimate. For this, we claim that

$$|B_i^\varepsilon| \ll \varepsilon\,|P_i| \tag{4.4}$$

for $i = 1, \dots, n$.

Summing this up, and using the disjointness in (4.3), the estimate (4.1) follows at once (for the case at hand, $\lambda_1(x) \geqslant 1$, and with $T_x = 0$).

To see the claim (4.4) we estimate both $|B_i^\varepsilon|$ and $|P_i|$ in terms of $|v^{(i)}|$. To simplify the notation we fix some $i \in \{1, \dots, n\}$ and drop the sub- and super-scripts. Notice first that we may assume $\varepsilon \leqslant \frac{1}{2}$ (for otherwise (4.4) is trivial) and hence $|v_1| \leqslant \frac{1}{2}$ (for otherwise B_i is empty and (4.4) is trivial). Thus, since $\lambda_1(x) \geqslant 1$ by definition and (4.2) we have $|v_2| \geqslant \sqrt{\frac{3}{4}}$ and

$$\begin{aligned} P_i &= \left\{t \in [0,T] \mid \|(v_1, v_2)u_t^{-1}\| < 1\right\} \\ &= \left\{t \in [0,T] \mid |v_2 - tv_1| < \sqrt{1-(v_1)^2}\right\} \\ &\supseteq \left\{t \in [0,T] \mid |v_2 - tv_1| < \sqrt{\tfrac{3}{4}}\right\}. \end{aligned}$$

On the other hand, we clearly have

$$B_i \subseteq \left\{t \in [0,T] \mid |v_2 - tv_1| \leqslant \varepsilon\right\}.$$

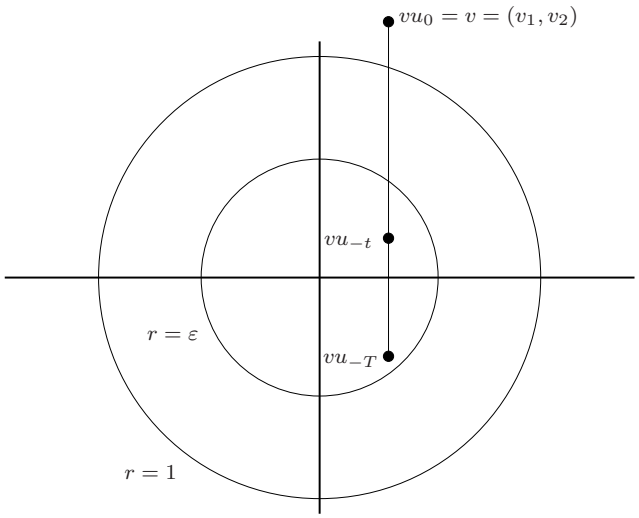

Fig. 4.2: The u_t-orbits of points $v \in \mathbb{R}^2$ travel at linear speed (determined by v_1). Thus the set B of bad times where $\|vu_t\| \leqslant \varepsilon$ is always a $\ll \varepsilon$-fraction of the protecting set P where $\|vu_t\| \leqslant 1$.

Since $v_2 - tv_1$ is linear in t with slope $-v_1$ as a function of t, it follows that either $B_i = \varnothing$ (this happens, for example, if $|v_1| > \varepsilon$) or

$$|P_i| \geqslant \left(\sqrt{\tfrac{3}{4}} - \tfrac{1}{2}\right) |v_1|^{-1}.$$

Here we subtract $\frac{1}{2}$ to also handle the case where $\|vu_{-T}\| < 1$. That is, the right end point of the interval $\left\{t \mid |v_2 - tv_1| < \sqrt{\frac{3}{4}}\right\}$ is to the right of $[0, T]$. Similarly we get

$$|B_i| \leqslant 2\varepsilon |v_1|^{-1}.$$

Therefore (4.4) (and so also (4.1)) follows for any $T > 0$ and any $x \in \mathsf{X}_2$ with $\lambda_1(x) \geqslant 1$.

If now $x_0 \in \mathsf{X}_2$ is non-periodic but otherwise arbitrary, then there exists some $T_0 > 0$ for which $x = u_{T_0} \cdot x_0$ has $\lambda_1(x) \geqslant 1$ by choosing T_0 such that the (unique up to sign) primitive vector $v_0 \in \Lambda_{x_0}$ with $\|v_0\| < 1$ has

$$\|v_0 u_{-T_0}\| = 1.$$

Let

$$\varepsilon_0 = \min_{t \in [0, T_0]} \lambda_1(u_t \cdot x),$$

and let T_x be chosen with

$$\frac{T_0}{T_x} \leqslant \varepsilon_0.$$

Now let $\varepsilon \in (0, 1]$ be arbitrary and $T \geqslant T_x$. If $\varepsilon < \varepsilon_0$, then

$$\{t \in [0, T] \mid u_t \cdot x \notin \mathsf{X}_2(\varepsilon)\} = \{t \in [T_0, T] \mid u_t \cdot x \notin \mathsf{X}_2(\varepsilon)\},$$

and applying the first case to $u_{T_0} \cdot x$ gives (4.1) in that case. If on the other hand $\varepsilon \geqslant \varepsilon_0$ then the first case applied to $u_{T_0} \cdot x$,

$$\{t \in [0, T] \mid u_t \cdot x \notin \mathsf{X}_2(\varepsilon)\} \subseteq [0, T_0] \cup \{t \in [T_0, T] \mid u_t \cdot x \notin \mathsf{X}_2(\varepsilon)\},$$

and $\frac{T_0}{T} \leqslant \varepsilon_0 \leqslant \varepsilon$ again gives (4.1), completing the proof. □

Corollary 4.3 (Non-escape of mass for X_2). *If $x \in \mathsf{X}_2$, then every weak*-limit of the collection of measures*

$$\left\{\frac{1}{T} \int_0^T (u_t)_* \delta_x \, \mathrm{d}t\right\}$$

is a probability measure on X_2.

Exercises for Section 4.1

Exercise 4.1.1. Prove Corollary 4.3.

4.2 The Case of $\mathsf{X}_3 = \mathrm{SL}_3(\mathbb{Z})\backslash \mathrm{SL}_3(\mathbb{R})$

The proof for the generalizations of Proposition 4.2 and Corollary 4.3 becomes significantly more involved for X_d with $d \geqslant 3$. We start with the case $d = 3$ because it is easier to envision and because it already contains all the main ingredients of the general case.

4.2.1 Non-Escape of Mass for Polynomial Trajectories

Even though we are primarily interested in unipotent trajectories, we will prove a more general claim allowing for general *polynomial orbits* of the shape

$$\mathrm{SL}_3(\mathbb{Z})p(t)$$

for $t \geqslant 0$ or for $t \in [0, T]$ for some $T \geqslant 0$, where

$$p\colon \mathbb{R} \to \mathrm{Mat}_3(\mathbb{R})$$

is a polynomial map taking values in $\mathrm{SL}_3(\mathbb{R})$. We say that p has degree no more than D if each matrix entry is a polynomial of degree no more than D. Notice that if $\{u_t \mid t \in \mathbb{R}\}$ is a one-parameter unipotent subgroup (of which there are precisely two up to conjugation in $\mathrm{SL}_3(\mathbb{R})$) with Lie algebra $\mathbb{R}v$ then $p(t) = gu_{-t} = g\exp(-tv)$ is a polynomial in t for any $g \in \mathrm{SL}_3(\mathbb{R})$. Hence a unipotent trajectory is also a polynomial trajectory. This generalization comes more or less for free in the sense that it does not complicate the proof significantly, while the generalization does have interesting consequences.

Much like a short periodic orbit for the horocycle flow on X_2, there is always the possibility that there are 'rational reasons' for a polynomial trajectory to remain stuck in the cusp in the following sense. There could be a vector $v \in \mathbb{Z}^3$ with $vp(t) = vp(0)$ for all t and with $\|vp(0)\|$ being small, or there could be a rational plane $V \subseteq \mathbb{R}^3$ with $Vp(t) = Vp(0)$ for all t such that the co-volume

of $Vp(0) \cap \mathbb{Z}^3$ is small[†], in which case there always exists for every $t \in \mathbb{R}$ a short vector in $(V \cap \mathbb{Z}^3)p(t)$, which may depend on t.

Similarly, a finite piece

$$\Gamma p(t)$$

for $t \in [0,T]$ of the trajectory would surely be entirely far out if there were a vector $v \in \mathbb{Z}^3$ with

$$\|vp(t)\| \leqslant \eta$$

for all $t \in [0,T]$, or if there is a rational plane $V \subseteq \mathbb{R}^3$ for which

$$\mathrm{vol}\big(Vp(t)/(V \cap \mathbb{Z}^3)p(t)\big) \leqslant \eta^2$$

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

As the last volume expression looks quite complicated but expresses the simple concept that we are studying the volume of the deformed plane with respect to the deformed lattice inside it, we now define some abbreviations for such expressions. For any $d \geqslant 2$ and any given discrete subgroup $\Lambda \leqslant \mathbb{R}^d$ (possibly of smaller rank) we write $\mathrm{covol}(\Lambda)$ as shorthand for the volume of $\mathbb{R}\Lambda/\Lambda$. Also if a polynomial $p(t) \in \mathrm{SL}_d(\mathbb{R})$ is given, we define for the study of the polynomial orbit $\mathbb{Z}^d p(t)$ the expression

$$\mathrm{covol}(V,t) = \mathrm{covol}\big((V \cap \mathbb{Z}^d)p(t)\big)$$

for any rational subspace $V \subseteq \mathbb{R}^d$.

To avoid the above mentioned 'rational constraints' for $d = 3$ we assume that there is some $\eta \leqslant 1$ such that

$$\sup_{t\in[0,T]} \|vp(t)\| \geqslant \eta \tag{4.5}$$

for all $v \in \mathbb{Z}^3 \smallsetminus \{0\}$, and

$$\sup_{t\in[0,T]} \mathrm{covol}\big((V \cap \mathbb{Z}^3)p(t)\big) \geqslant \eta^2 \tag{4.6}$$

[†] If $Vp(t) = Vp(0)$ for all $t \in \mathbb{R}$, then

$$\mathrm{vol}\big(Vp(t)/(\mathbb{Z}^3 \cap V)p(t)\big) = \mathrm{vol}\big(Vp(0)/(\mathbb{Z}^3 \cap V)p(0)\big)$$

for all $t \in \mathbb{R}$. If $p(t) = gu_{-t}$ is the parametrization of an orbit under a one-parameter unipotent subgroup, this is clear as the restriction of the unipotent subgroup to the invariant subspace Vg is again unipotent. In general, $\bigwedge^2 p(t)$ sends by assumption the line in $\bigwedge^2 \mathbb{R}^3$ corresponding to V to one and the same line for every t. If the co-volume of $(\mathbb{Z}^3 \cap V)p(t)$ inside $Vp(t)$ is not constant, or equivalently if $\bigwedge^2 p(t)$ applied to an element w of $\bigwedge^2 V \subseteq \bigwedge^2 \mathbb{R}^3$ is not constant, then $w \bigwedge^2 p(t) = (w \bigwedge^2 p(0))h(t)$ for a non-constant $\mathbb{R}$-valued polynomial $h(t)$. As $h(t)$ has a complex root, we get a contradiction to $\bigwedge^2 p(\mathbb{C}) \subseteq \mathrm{SL}(\bigwedge^2 \mathbb{C}^3)$.

for all rational planes $V \subseteq \mathbb{R}^3$. Using our abbreviation we could combine these two estimates into the assumption that

$$\sup_{t\in[0,T]} \operatorname{covol}(V,t) \geqslant \eta^{\dim V}$$

for any rational subspace $V \subseteq \mathbb{R}^3$. This unified treatment of all intermediate subspaces will be our view point in the general case, see Section 4.3, but will also play a role in the proof of the following theorem.[(16)]

Theorem 4.4 (Quantitative non-divergence for X_3). *Suppose that the piece $\Gamma p(t)$, $t \in [0,T]$ of a polynomial trajectory satisfies* (4.5) *and* (4.6) *for some $\eta \leqslant 1$. Then, for $\varepsilon \in (0,\eta]$,*

$$\frac{1}{T}\left|\{t \in [0,T] \mid \Gamma p(t) \notin \mathsf{X}_3(\varepsilon)\}\right| \ll_D \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{2D}}$$

where p is a polynomial of degree no more than D.

Remark 4.5. (1) The alternating tensor product $\bigwedge^2(\mathbb{R}^3)$ may be identified with $\mathbb{R}^3$ by choosing (for example) the basis $e_2 \wedge e_3$, $e_3 \wedge e_1$ and $e_1 \wedge e_2$ where as usual e_1, e_2, e_3 is the standard basis of $\mathbb{R}^3$. This way the map

$$(v,w) \in \mathbb{R}^3 \times \mathbb{R}^3 \to v \wedge w \in \bigwedge^2 \mathbb{R}^3$$

is identified with the exterior product

$$(v,w) \in \mathbb{R}^3 \times \mathbb{R}^3 \to v \times w \in \mathbb{R}^3.$$

The linear map

$$\bigwedge^2 p(t) \colon \bigwedge^2(\mathbb{R}^3) \longrightarrow \bigwedge^2(\mathbb{R}^3)$$

is then the linear map with

$$e_i \wedge e_j \longmapsto (e_i p(t)) \wedge (e_j p(t))$$

for $1 \leqslant i,j \leqslant 3$. It is again a polynomial (of at most doubled degree) with values in $\mathrm{SL}\left(\bigwedge^2(\mathbb{R}^3)\right)$. Moreover, note that the co-volume of $\mathbb{Z}v_1 + \mathbb{Z}v_2$ equals the area of the parallelogram spanned by v_1 and v_2, or equivalently the length of $v_1 \wedge v_2$ (identified with the exterior product $v_1 \times v_2$).

(2) As we will see in the course of the proof, the exponent $2D$ can be replaced by any $D' \geqslant 1$ with the property that $\|vp(t)\|^2$ for $v \in \mathbb{R}^3$ and $\|w \bigwedge^2 p(t)\|^2$ for $w \in \bigwedge^2(\mathbb{R}^3)$ are polynomials of degree no more than $2D'$. Notice that the choice $D' = 2D$ has this property. In the case of the orbit of the one-parameter unipotent subgroup given by

$$p(t) = g \begin{pmatrix} 1 & -t & \\ & 1 & \\ & & 1 \end{pmatrix},$$

we may take $D' = 1$, while for that defined by

$$p(t) = g \begin{pmatrix} 1 & -t & \frac{1}{2}t^2 \\ & 1 & -t \\ & & 1 \end{pmatrix}$$

we may take $D' = 2$.

(3) There are two ways in which one can establish the assumptions (4.5) and (4.6), and both are important in applying Theorem 4.4.

(a) Given p and T, one can find $\eta > 0$ with the desired property, for example by taking

$$\eta = \min\left\{\lambda_1(\mathbb{Z}^3 p(0)), \sqrt{\alpha_2(\mathbb{Z}^3 p(0))}\right\}.$$

(b) Given p such that $vp(t)$ is non-constant for any $v \in \mathbb{Z}^3$ and also $Vp(t)$ is a non-constant subspace for any rational plane $V \subseteq \mathbb{R}^2$, one can find some $T_0 > 0$ such that for $T \geqslant T_0$ we can use $\eta = 1$. In fact, there are only finitely many vectors $v \in \mathbb{Z}^3$ with $\|vp(0)\| \leqslant 1$, and for each of them $vp(t)$ is non-constant and hence there must be some T_0 such that (4.5) holds for $T \geqslant T_0$ and $\eta = 1$. The argument to establish (4.6) is similar.

4.2.2 A Lemma About Polynomials

We now prove a lemma which replaces the argument involving the linear function $v_2 - tv_1$ in the proof of Proposition 4.2 (see in particular Figure 4.1).

Lemma 4.6 (Small values of polynomials). *Let $p \in \mathbb{R}[t]$ be a polynomial of degree L, and fix $T > 0$. Then for every $\varepsilon > 0$,*

$$\frac{1}{T}\left|\left\{t \in [0,T] \mid |p(t)| < \varepsilon \|p\|_{T,\infty}\right\}\right| \ll_L \varepsilon^{1/L}, \tag{4.7}$$

where

$$\|p\|_{T,\infty} = \sup_{t\in[0,T]} |p(t)|.$$

The situation is illustrated in Figure 4.3 for the polynomial $p(t) = t^4$.

The main property of polynomials that will be used in the proof of Theorem 4.4 is Lemma 4.6. A function or family of functions $p\colon [0,T] \to \mathbb{R}$ is[†] *polynomial-like of degree no more than L*, or simply is *of degree no more*

[†] The more common, but less informative, terminology is (C,α)-good, where $\alpha = \frac{1}{L}$ and C is the implied constant.

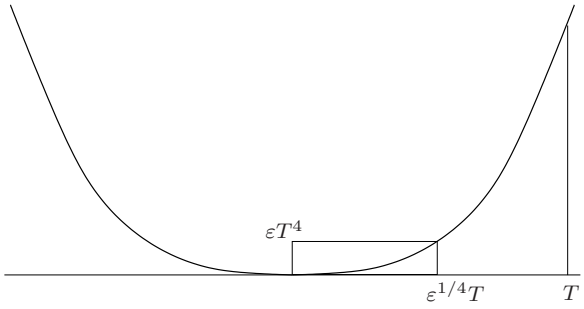

Fig. 4.3: The graph of $p(t) = t^L$ for $L = 4$ shows that the left-hand side of (4.7) can indeed be of the size $\varepsilon^{1/L}$.

than L if p satisfies the conclusion of Lemma 4.6, and the implied constant does not depend on the particular function p if a whole family of such functions is being considered. We will not pursue this generality here, and instead refer to the papers of Kleinbock and Margulis [101] and of Kleinbock [97].

Proof of Lemma 4.6. By induction we may assume that the lemma already holds for all polynomials of degree less than L. The claim of the lemma is invariant under the following transformations:

- Replacing p by $\frac{1}{\|p\|_{T,\infty}} p$.
- Replacing T by 1 and at the same time $p(t)$ by $p(tT)$ for $t \in [0,1]$.

Thus we may assume without loss of generality that $T = 1$ and $\|p\|_{1,\infty} = 1$. Let $a_1, \dots, a_r \in \mathbb{R}$ and $z_1, \overline{z_1}, \dots, z_s, \overline{z_s} \in \mathbb{C} \smallsetminus \mathbb{R}$ be the list of real zeros and pairs of complex conjugate zeros of p, listed with multiplicity so that

$$r + 2s = L.$$

Let $b \in \mathbb{R}$ be the leading coefficient of p, so that

$$p(t) = b(t - a_1) \cdots (t - a_r)(t - z_1)(t - \overline{z_1}) \cdots (t - z_s)(t - \overline{z_s}).$$

Suppose first that $|a_1| \geqslant 2$. Then we have

$$1 \ll \left| \frac{t - a_1}{a_1} \right| \ll 1$$

for all $t \in [0,1]$. Hence the claim for p is equivalent to the claim for the polynomial

$$\tilde{p}(t) = ba_1(t - a_2) \cdots (t - a_r)(t - z_1)(t - \overline{z_1}) \cdots (t - z_s)(t - \overline{z_s})$$

of degree $L-1$ (with different multiplicative constants). Similarly, if $|z_1| \geqslant 2$, then

$$1 \ll \left| \frac{(t-z_1)(t-\overline{z_1})}{z_1\overline{z_1}} \right| \ll 1$$

for all $t \in [0,1]$, and we may reduce the claim to a polynomial of degree $L-2$.

Thus we may assume that

$$|a_1|, \dots, |a_r|, |z_1|, \dots, |z_s| \leqslant 2.$$

Now for $t \in [0,1]$ we have

$$|t - a_i| \leqslant 3 \text{ and } |t - z_j| \leqslant 3$$

for $i = 1, \dots, r$ and $j = 1, \dots, s$. It follows that

$$1 = \|p\|_{1,\infty} \leqslant |b| 3^L. \tag{4.8}$$

Suppose now

$$|\{t \in [0,1] \mid |q(t)| < \varepsilon\}| \ll_L \varepsilon^{1/L} \tag{4.9}$$

holds for all $\varepsilon > 0$ and the polynomial

$$q(t) = (t-a_1)\cdots(t-a_r)(t-z_1)(t-\overline{z_1})\cdots(t-z_s)(t-\overline{z_s}) = \frac{1}{b}p(t).$$

Then, since $|p(t)| < \varepsilon$ implies $|q(t)| < \frac{\varepsilon}{b}$, we get from (4.8)–(4.9) that

$$|\{t \in [0,T] \mid |p(t)| < \varepsilon\}| \ll_L \left(\frac{\varepsilon}{b}\right)^{1/L} \leqslant 3\varepsilon^{1/L}$$

and so the lemma.

It remains to prove (4.9). Suppose that $t \in [0,1]$ has distance at least $\varepsilon^{1/L}$ from any of the zeros

$$a_1, \dots, a_r, z_1, \overline{z_1}, \dots, z_s, \overline{z_s}.$$

Then clearly $|q(t)| \geqslant \varepsilon$. On the other hand the elements $t \in [0,1]$ with distance less than $\varepsilon^{1/L}$ from a_i (respectively z_i) lie in an interval of length at most $2\varepsilon^{1/L}$. This gives (4.9), with $2(r+s) \leqslant 2L$ as the implied constant. As discussed above, the lemma follows. □

4.2.3 Protection Arising From a Flag

The most important feature that makes the proof of Proposition 4.2 easier than the case of $\mathrm{SL}_3(\mathbb{R})$ considered here is the fact that a unimodular lattice $\Lambda \leqslant \mathbb{R}^2$ cannot have two linearly independent vectors of length less than one. This gave automatic 'protection' from short vectors: if there is a Λ-

primitive vector of length less than one, and this vector is not tiny, then no tiny non-zero vector can exist in Λ. Using this we defined protecting intervals which were automatically disjoint.

This property of only one short vector is manifestly false for unimodular lattices in $\mathbb{R}^3$. For example, the lattice

$$\Lambda_n = \tfrac{1}{n}\mathbb{Z}e_1 + \tfrac{1}{n}\mathbb{Z}e_2 + n^2\mathbb{Z}e_3$$

is unimodular for any $n \geqslant 1$, and contains two linearly independent vectors of length $\frac{1}{n}$. What we need to discuss in order to get a similar protection phenomenon in $\mathbb{R}^3$ are flags.

A *flag* in $\mathbb{R}^d$ is a collection comprising a line

$$V_1 = \mathbb{R}v_1,$$

a plane

$$V_2 = \mathbb{R}v_1 + \mathbb{R}v_2 \supseteq V_1,$$

and so on up to a hyperplane

$$V_{d-1} = V_{d-2} + \mathbb{R}v_{d-1} \supseteq V_{d-2}.$$

We also write $V_0 = \{0\}$ and $V_d = \mathbb{R}^d$.

Lemma 4.7 (Protection coming from flags). *Let $\Lambda \leqslant \mathbb{R}^d$ be a unimodular lattice, and let*

$$V_0 = \{0\} \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_d = \mathbb{R}^d$$

be a flag of Λ-rational subspaces. Then

$$\lambda_1(\Lambda) \geqslant \min_{i=1,\dots,d} \left\{ \frac{\operatorname{covol}(\Lambda \cap V_i)}{\operatorname{covol}(\Lambda \cap V_{i-1})} \right\},$$

where $\operatorname{covol}(\{0\}) = \operatorname{covol}(\Lambda) = 1$.

This gives the desired protection in the following sense, illustrated for the case $d = 3$: If $\operatorname{covol}(\Lambda \cap V_1)$ is of size roughly ε, and $\operatorname{covol}(\Lambda \cap V_2)$ is of size roughly ε^2, then Λ does not contain vectors that are much shorter than ε.

Proof of Lemma 4.7. Let $v \in \Lambda$ be chosen with norm $\|v\| = \lambda_1(\Lambda)$. If v does not lie in V_{d-1}, then the co-volume of

$$\Lambda \cap V_{d-1} + \mathbb{Z}v \subseteq \mathbb{R}^d$$

is equal to $\operatorname{covol}(\Lambda \cap V_{d-1}) \cdot \|\pi(v)\|$, where $\pi\colon \mathbb{R}^d \to V_{d-1}^{\perp}$ is the orthogonal projection. In particular, since the co-volume of $\Lambda \supseteq \Lambda \cap V_{d-1} + \mathbb{Z}v$ is 1, we have

$$\begin{aligned} 1 = \mathrm{covol}(\Lambda) &\leqslant \mathrm{covol}(\Lambda \cap V_{d-1} + \mathbb{Z}v) \\ &= \mathrm{covol}(\Lambda \cap V_{d-1})\|\pi(v)\| \\ &\leqslant \mathrm{covol}(\Lambda \cap V_{d-1})\|v\|, \end{aligned}$$

which implies the lemma in the case $v \notin V_{d-1}$.

Suppose now that $v \in \Lambda \cap V_{i+1}$ but $v \notin V_i$ for some $i \in \{1, \dots, d\}$. As before,

$$\begin{aligned} \mathrm{covol}(\Lambda \cap V_{i+1}) &\leqslant \mathrm{covol}(\Lambda \cap V_i + \mathbb{Z}v) \\ &= \mathrm{covol}(\Lambda \cap V_i)\|\pi(v)\| \\ &\leqslant \mathrm{covol}(\Lambda \cap V_i)\|v\|, \end{aligned}$$

where π is the appropriate projection, and the lemma follows at once. □

To handle the lack of disjointness of the protecting intervals for individual vectors or subspaces we are also going to use a simple[(17)] *covering lemma*.

Lemma 4.8 (A covering lemma on intervals). *Let $I \subseteq \mathbb{R}$ be a compact interval, and let $P_1, \dots, P_K \subseteq I$ be a finite collection of compact sub-intervals. Then there exists a subcollection of these intervals $P_{j(1)}, \dots, P_{j(k)}$ which are* nearly disjoint *in the sense that*

$$\sum_{\ell=1}^{k} \mathbb{1}_{P_{j(\ell)}} \leqslant 2 \tag{4.10}$$

while still having the same union,

$$\bigcup_{\ell=1}^{k} P_{j(\ell)} = \bigcup_{n=1}^{K} P_n. \tag{4.11}$$

Moreover, none of the selected intervals $P_{j(\ell)}$ is strictly contained in any of the other intervals $P_1, \dots, P_K$.

Proof. Let

$$U = \bigcup_{n=1}^{K} P_n,$$

and note that we may remove from the finite collection of intervals any interval P_n which is properly contained in any other interval of the list without affecting the set U. Below we construct our subcollection of intervals from the remaining ones with a simple greedy algorithm, which will immediately imply the last claim in the lemma.

Let $P_n = [a_n, b_n]$ for $n = 1, \dots, K$, and let $d_0 = \min U$. Among all intervals P_n with $a_n = d_0$, there is one with maximal b_n. We select this interval first, writing d_1 for its endpoint so

$$P_{j(1)} = [d_0, d_1].$$

If d_1 is an interior point of U then we list all intervals containing d_1 and choose once again an interval whose right-hand end point is largest among all those containing d_1. Formally,

$$\begin{aligned} d_1 &= \max\{b_n \mid a_n = d_0\}, \\ d_2 &= \max\{b_n \mid a_n \leqslant d_1 < b_n\} > d_1, \end{aligned}$$

and

$$\begin{aligned} P_{j(1)} &= [d_0, d_1], \\ P_{j(2)} &= P_n = [a_n, d_2], \end{aligned}$$

where we pick some n as in the definition of d_2.

If d_2 is again an interior point of U we continue in the same way, until we come to a boundary point of U. At this stage we may restart the process, with the next biggest element of U, if there is one, or finish the process of selecting intervals if U has no remaining points to the right of the right end point of the last selected interval.

By construction, at each stage of the selection of the intervals

$$P_{j(1)}, \dots, P_{j(\ell)}$$

we have

$$\bigcup_{r=1}^{\ell} P_{j(r)} = U \cap (-\infty, b_{j(\ell)}],$$

giving (4.11). Moreover, the only selected interval other than $P_{j(1)}$ that can intersect non-trivially with $P_{j(1)}$ is $P_{j(2)}$, since if

$$P_{j(1)} \cap P_{j(\ell)} \neq \varnothing$$

for some $\ell \geqslant 2$ then $a_{j(\ell)} \leqslant d_1$, d_1 is an interior point of U, and

$$b_{j(\ell)} > b_{j(1)} = d_1$$

(since $\ell \geqslant 2$). Hence $b_{j(\ell)} \leqslant d_2$ by our choice of d_2, which forces ℓ to be 2. Repeating this argument gives (4.10) as required. □

4.2.4 Non-Divergence for X_3 — Obtaining Protecting Flags

In the course of the proof we will treat 1- and 2-dimensional subspaces on the same footing, so we will use the notation V uniformly for both from now on.

Proof of Theorem 4.4. Assume that $p\colon [0,T] \to \mathrm{SL}_3(\mathbb{R})$ has the property that

$$\mathrm{covol}\,(V,t)^2$$

is polynomial of degree no more than $2D'$ for every rational subspace $V \subseteq \mathbb{R}^3$. Furthermore, let $\eta \leqslant 1$ satisfy (4.5) and (4.6), and fix $\varepsilon \in (0,\eta]$. We note that $D' = 2D$ satisfies this assumption for D as in the theorem (see also Remark 4.5).

First stage protection intervals: Notice that there are only finitely many rational subspaces $V \subseteq \mathbb{R}^3$ for which

$$\mathrm{covol}\,(V,t) \leqslant \eta^{\dim V}$$

for some $t \in [0,T]$. For each of those subspaces V we define the intervals $P_{V,i}$ for $i = 1,\dots,\ell_V$ to be the set of maximal subintervals† of

$$\{t \in [0,T] \mid \mathrm{covol}\,(V,t) \leqslant \eta^{\dim V}\}.$$

Notice that by maximality of the subintervals and the assumptions (4.5) and (4.6) we have $\mathrm{covol}(V,t) = \eta^{\dim V}$ for at least one of the endpoints of each of the intervals $P_{V,i}$. In particular

$$\sup_{t\in P_{V,i}} \mathrm{covol}\,(V,t) = \eta^{\dim V}. \tag{4.12}$$

This defines a collection of closed intervals $P_{V,i}$ where we vary both V and i. Applying Lemma 4.8 to this collection and the interval $[0,T]$, we obtain a nearly disjoint subcollection

$$P_1,\dots,P_m$$

of these intervals with

$$\bigcup_V \bigcup_i P_{V,i} = \bigcup_{r=1}^{m} P_r$$

and with

$$\sum_{r=1}^{m} \mathbb{1}_{P_r} \leqslant 2.$$

† As each such subinterval accounts for two roots of the polynomial equation

$$\mathrm{covol}\,(V,t)^2 = \eta^{2\dim V},$$

there can be at most D such intervals.

We write V_r for the subspace that gave rise to the interval $P_r = P_{V_r,i_r}$ for some $i_r \in \{1, \dots, \ell_{V_r}\}$. As this subspace alone does not give protection (since Lemma 4.7 needs a complete flag and we only have one subspace), we need to do another search for a compatible subspace as follows.

Second stage protection intervals: Suppose first that V_r for

$$r \in \{1, \dots, m\}$$

is a line. Consider now the intervals

$$P_{V,i} \cap P_r$$

for all rational planes $V \subseteq \mathbb{R}^3$ that are compatible with V_r, in the sense that $V_r \subseteq V$. Now apply the covering lemma on P_r to this collection to obtain nearly disjoint subintervals

$$P_{r,1}, \dots, P_{r,n(r)} \subseteq P_r$$

with

$$\bigcup_{\substack{V_r \subseteq V, \\ \text{plane}}} \bigcup_i P_{V,i} \cap P_r = \bigcup_{s=1}^{n(r)} P_{r,s} \cap P_r.$$

Similarly, if V_r for $r \in \{1, \dots, m\}$ is a plane, then we obtain nearly disjoint subintervals

$$P_{r,1}, \dots, P_{r,n(r)} \subseteq P_r$$

defined by compatible rational lines $V \subseteq V_r$ with

$$\bigcup_{\substack{V \subseteq V_r, \\ \text{line}}} \bigcup_i P_{V,i} \cap P_r \subseteq \bigcup_{s=1}^{n(r)} P_{r,s} \cap P_r.$$

In both cases $n(r) = 0$ is possible.

Just as we denote by V_r the subspace that gave rise to the interval P_r, we also write $V_{r,s}$ for the subspace giving rise to $P_{r,s}$ and $i_{r,s}$ for the corresponding index so that $P_{r,s} = P_{V_{r,s},i_{r,s}} \cap P_r$.

By construction V_r and $V_{r,s}$ are compatible (that is, they define a complete flag in $\mathbb{R}^3$) for all r and s. We will show that the intervals

$$P_1, \dots, P_m, P_{1,1}, \dots, P_{1,n(1)}, \dots, P_{m,1}, \dots, P_{m,n(m)}$$

together give the desired protection.

Bad subsets: We now define for $\varepsilon > 0$ the associated bad subsets of the intervals above:

$$\mathrm{Bad}\,(r,\varepsilon) = \left\{t \in P_r \mid \mathrm{covol}\,(V_r,t) \leqslant \varepsilon\eta^{\dim V_r - 1}\right\},$$
$$\mathrm{Bad}\,(r,s,\varepsilon) = \left\{t \in P_{r,s} \mid \mathrm{covol}\,(V_{r,s},t) \leqslant \varepsilon\eta^{\dim V_{r,s}-1}\right\}$$

and the union

$$\mathrm{Bad}(\varepsilon) = \bigcup_{r=1}^{m}\left(\mathrm{Bad}\,(r,\varepsilon) \cup \bigcup_{s=1}^{n(r)} \mathrm{Bad}\,(r,s,\varepsilon)\right).$$

Estimate of bad subset: We now apply Lemma 4.6 to the polynomial[†]

$$\mathrm{covol}\,(V_r,t)^2$$

of degree no larger than $2D'$ on the interval P_r, with supremum norm $\eta^{2\dim V_r}$ by (4.12). This gives

$$|\mathrm{Bad}\,(r,\varepsilon)| \ll \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D'}} |P_r|\,, \tag{4.13}$$

by definition of $\mathrm{Bad}\,(r,\varepsilon)$.

To prove the same for $\mathrm{Bad}(r,s,\varepsilon)$ we need to show an analogue of (4.12) for $P_{r,s} = P_{V_{r,s},i_{r,s}} \cap P_r$. For this, notice that by Lemma 4.8 (from the first application that gave rise to $P_1,\ldots,P_r,\ldots,P_m$) none of the intervals $P_{V_{r,s},i_{r,s}}$ can contain P_r properly — let us refer to this as the *non-containment*. If both end points t of $P_{V_{r,s},i_{r,s}}$ satisfy $\mathrm{covol}\,(V_{r,s},t) = \eta^{\dim V_{r,s}}$ (because they are in $(0,T)$, for example) then (due to the non-containment) one of them must be in P_r, and so

$$\sup_{t\in P_{r,s}} \mathrm{covol}\,(V_{r,s},t) = \eta^{\dim V_{r,s}}. \tag{4.14}$$

If, on the other hand, we have $\mathrm{covol}\,(V_{r,s},t) < \eta^{\dim V_{r,s}}$ for one of the endpoints of $P_{V_{r,s},i_{r,s}}$ (this endpoint would have to be 0 or T), then the other will have to be in P_r (due to the non-containment) and we again get (4.14). Therefore, using Lemma 4.6 together with (4.14) as a replacement for (4.12) gives as before

$$|\mathrm{Bad}\,(r,s,\varepsilon)| \ll \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D'}} |P_{r,s}|\,. \tag{4.15}$$

Since the intervals $P_{r,s} \subseteq P_r$ are all nearly disjoint we get

$$\sum_{s=1}^{n(r)} |P_{r,s}| \leqslant 2\,|P_r| \ll |P_r|\,. \tag{4.16}$$

[†] Note that $\mathrm{covol}\,(V_r,t)$ is in general not a polynomial, but $\mathrm{covol}\,(V_r,t)^2$ is.

Thus we may take the union and use (4.13), (4.15) and (4.16) to obtain the estimate

$$\begin{aligned}|\mathrm{Bad}(\varepsilon)| &\leqslant \sum_{r=1}^{m}\left(|\mathrm{Bad}\,(r,\varepsilon)| + \sum_{s=1}^{n(r)}|\mathrm{Bad}(r,s,\varepsilon)|\right)\\ &\ll \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D'}}\sum_{r=1}^{m}|P_r|\\ &\leqslant 2\left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D'}}T,\end{aligned}$$

since the intervals

$$P_1,\ldots,P_m \subseteq [0,T]$$

are nearly disjoint.

Protection: We now show that

$$\{t\in[0,T]\mid \mathrm{SL}_3(\mathbb{Z})p(t)\notin \mathsf{X}_3(\varepsilon)\}\subseteq \mathrm{Bad}(\varepsilon), \tag{4.17}$$

for all $\varepsilon\leqslant\eta$, so that the estimate above then implies the theorem.

Suppose therefore that $t\in[0,T]$ has the property that $\mathbb{Z}^3p(t)$ contains an ε-short vector $vp(t)$. Since $\varepsilon\leqslant\eta$, this shows that t belongs to one of the protecting intervals defined by $V=\mathbb{R}v$. Hence we must have $t\in P_r$ for some $r\in\{1,\ldots,m\}$ by choice of these intervals.

If $V=V_r$ then we have $t\in\mathrm{Bad}\,(r,\varepsilon)\subseteq\mathrm{Bad}(\varepsilon)$. If V_r is a line but $V\neq V_r$, then $V+V_r$ is a subspace compatible with V_r and

$$\mathrm{covol}\,(V+V_r,t)\leqslant \mathrm{covol}\,(V,t)\,\mathrm{covol}\,(V_r,t)\leqslant\varepsilon\eta\leqslant\eta^2.$$

Therefore $t\in P_r\cap P_{V+V_r,i}$ (for some i) and so $t\in P_{r,s}$ for some s in $\{1,\ldots,n(r)\}$, by construction. We have obtained a complete flag: $V_r\subseteq V_{r,s}$ with

$$t\in P_r\cap P_{r,s}.$$

Suppose now that V_r is a plane. Recall that

$$\mathrm{covol}(V,t)\leqslant\varepsilon$$

and

$$\mathrm{covol}(V_r,t)\leqslant\eta^2.$$

We may assume that $V\subseteq V_r$. For if $V+V_r=\mathbb{R}^3$, $\eta\leqslant 1$ and $\varepsilon<1$ (which we may assume) then we get a contradiction to the unimodularity of the three-dimensional lattice. Therefore, $t\in P_r\cap P_{V,i}$ for some i and so there

must exist some $s \in \{1, \dots, n(r)\}$ with $t \in P_{r,s}$. Once more we have obtained a complete flag: $V_{r,s} \subseteq V_r$ with $t \in P_r \cap P_{r,s}$.

Hence it remains to consider the case $t \in P_r$ and $t \in P_{r,s}$. Let us also assume, for the purposes of a contradiction, that

$$t \notin \mathrm{Bad}\,(r, \varepsilon) \cup \mathrm{Bad}\,(r, s, \varepsilon)\,.$$

Hence

$$\varepsilon\eta^{\dim V_r - 1} \leqslant \mathrm{covol}(V_r, t) \leqslant \eta^{\dim V_r}$$

and

$$\varepsilon\eta^{\dim V_{r,s} - 1} \leqslant \mathrm{covol}(V_{r,s}, t) \leqslant \eta^{\dim V_{r,s}},$$

and together V_r and $V_{r,s}$ define a flag in $\mathbb{R}^3$. Lemma 4.7 may now be applied to show that

$$\lambda_1\left(\mathbb{Z}^3 p(t)\right) \geqslant \min\left(\varepsilon, \varepsilon, \tfrac{1}{\eta^2}\right) = \varepsilon,$$

in contradiction to the assumption on t. This proves the claim (4.17), and hence the theorem. □

4.3 The General Case of $\mathbf{X_d = SL_d(\mathbb{Z})\backslash SL_d(\mathbb{R})}$

Let us now state and prove the general version of the non-divergence theorem (using the abbreviations and tools introduced in the last section).

Theorem 4.9 (Quantitative non-divergence for X_d by Margulis, Dani and Kleinbock[18]). *Suppose that*

$$p\colon \mathbb{R} \to \mathrm{SL}_d(\mathbb{R})$$

is a polynomial and $T > 0$ is such that

$$\sup_{t \in [0,T]} \mathrm{covol}(V, t) \geqslant \eta^{\dim V} \tag{4.18}$$

for some $\eta \in (0, 1]$ and all rational subspaces $V \subseteq \mathbb{R}^d$. Assume furthermore that $2D$ is an upper bound for the degrees of $\mathrm{covol}(V, t)^2$ for all rational subspaces $V \subseteq \mathbb{R}^d$. Then, for $\varepsilon \in (0, \eta]$,

$$\frac{1}{T}\left|\{t \in [0, T] \mid \Gamma p(t) \notin \mathsf{X}_d(\varepsilon)\}\right| \ll_{d,D} \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}}. \tag{4.19}$$

Proof. The proof comprises the following steps:

(1) iterated construction of protecting intervals and partial flags;
(2) definition and estimate of the bad subsets;

(3) reaching the conclusion by combining the established properties.

Inductive step to construct protecting intervals: Suppose we are given an interval $I \subseteq \mathbb{R}$ and a 'partial flag'

$$\mathscr{F} = \left\{\{0\} \subsetneq V_1 \subsetneq V_2 \subsetneq \cdots \subsetneq V_k \subsetneq \mathbb{R}^d\right\}$$

of rational subspaces of $\mathbb{R}^d$ with $0 \leqslant k < d-1$ such that

$$\sup_{t\in I} \operatorname{covol}(V_j, t) \leqslant \eta^{\dim V_j}$$

for $j = 1, \ldots, k$ and

$$\sup_{t\in I} \operatorname{covol}(V, t) \geqslant \eta^{\dim V} \tag{4.20}$$

for any rational subspace $V \leqslant \mathbb{R}^d$ that is compatible with $\mathscr{F}$. This means that $V \notin \mathscr{F}$ and $\mathscr{F} \cup \{V\}$ is again a partial flag or a flag. Here we say that V is compatible with the partial flag. Initially we have

$$\mathscr{F} = \{\{0\} \subsetneq \mathbb{R}^d\}, I = [0, T], k = 0,$$

and (4.20) is precisely the assumption (4.18) in Theorem 4.9.

Now consider all rational subspaces $V \leqslant \mathbb{R}^d$ that are compatible with the partial flag $\mathscr{F}$. For each such subspace split

$$\{t \in I \mid \operatorname{covol}(V, t) \leqslant \eta^{\dim V}\}$$

into its connected components, giving rise to subintervals

$$P_{V,1}, \ldots, P_{V,\ell_V}.$$

Varying both V and the second index, we may apply Lemma 4.8 to obtain a finite nearly disjoint subcollection

$$P_1, \ldots, P_m$$

of these intervals with the same union, so

$$\bigcup_{\substack{V \text{ compatible} \\ \text{with } \mathscr{F}}} \{t \in I \mid \operatorname{covol}(V, t) \leqslant \eta^{\dim V}\} = \bigcup_{r=1}^{m} P_r$$

and and union is nearly disjoint so that

$$\sum_{r=1}^{m} |P_r| \leqslant 2|I|.$$

Let us write V_r for the subspace that gave rise to the interval $P_r = P_{V_r,i_r}$ for some i_r.

On each of those sub-intervals P_r we have the new (maybe partial or complete) flag

$$\mathscr{F} \cup \{V_r\}$$

with

$$\sup_{t\in P_r} \mathrm{covol}(V,t) \leqslant \eta^{\dim V}$$

for all $V \in \mathscr{F} \cup \{V_r\}$. Now let V be either V_r or a rational subspace that is compatible with $\mathscr{F} \cup \{V_r\}$. In particular, V is compatible with $\mathscr{F}$ and so was considered in the construction of the subintervals

$$P_1, \dots, P_m \subseteq I.$$

By Lemma 4.8 this shows that P_r is not strictly contained in any of the subintervals

$$P_{V,1}, \dots, P_{V,\ell_V}$$

defined by V, so that (by the same argument that lead to (4.14))

$$\sup_{t\in P_r} \mathrm{covol}(V,t) \geqslant \eta^{\dim V}$$

respectively

$$\sup_{t\in P_r} \mathrm{covol}(V_r,t) \geqslant \eta^{\dim V_r}. \tag{4.21}$$

Iterating the construction: As hinted at before, we start the iterative construction with $I = [0,T]$, $\mathscr{F} = \{\{0\} \subsetneq \mathbb{R}^d\}$, and $k = 0$. By (4.18) the inductive hypothesis is satisfied, and the inductive step above defines intervals

$$P_1, \dots, P_r$$

and subspaces

$$V_1, \dots, V_r.$$

On each of the intervals P_{i_1} for $i_1 = 1, \dots, r$, the partial flag

$$\mathscr{F}_{i_1} = \{\{0\} \subsetneq V_{i_1} \subsetneq \mathbb{R}^d\}$$

satisfies the inductive hypothesis so that the inductive step can be repeated, giving rise to intervals

$$P_{i_1,i_2} \subseteq P_{i_1}$$

and partial flags

$$\mathscr{F}_{i_1,i_2} = \mathscr{F}_{i_1} \cup \{V_{i_1,i_2}\}$$

for a compatible subspace V_{i_2}. In general, let us write

$$\overline{\imath} = (i_1, \ldots, i_k)$$

for the multi-index arising,

$$P_{\overline{\imath}} = P_{i_1,\ldots,i_k}$$

for the intervals arising, and

$$\mathscr{F}_{\overline{\imath}} = \mathscr{F}_{i_1,\ldots,i_k}$$

for the flags or partial flags arising. The construction stops when, for a given interval $P_{\overline{\imath}}$ and partial or complete flag $\mathscr{F}_{\overline{\imath}}$ there is no compatible rational subspace V for which

$$\{t \in P_{\overline{\imath}} \mid \operatorname{covol}(V, t) < \eta^{\dim V}\}$$

is non-empty, and certainly stops if $\mathscr{F}_{\overline{\imath}}$ is a (complete) flag. This may be thought of as a finite graded tree labeled by the intervals and the flags or partial flags, as illustrated in Figure 4.4.

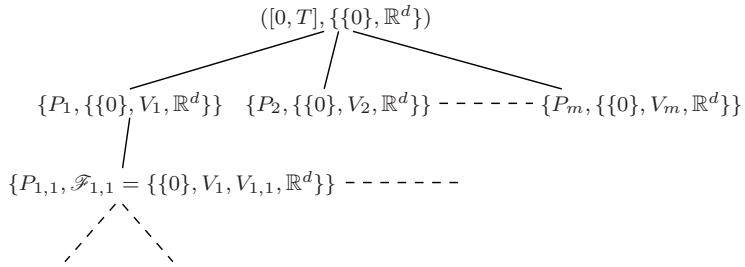

Fig. 4.4: Inductive construction of the intervals and flags.

Definition of Bad subsets: For any $(P_{\overline{\imath}}, \mathscr{F}_{\overline{\imath}})$ as constructed above, we define the following bad subset

$$\operatorname{Bad}(\overline{\imath}, \varepsilon) = \{t \in P_{\overline{\imath}} \mid \operatorname{covol}(V_{\overline{\imath}}, t) \leqslant \varepsilon \eta^{\dim V}\}.$$

Taking the union we define

$$\operatorname{Bad}(\varepsilon) = \bigcup_{\substack{\overline{\imath}=(i_1,\ldots,i_k),\\ k\geqslant 1}} \operatorname{Bad}(\overline{\imath}, \varepsilon)$$

Estimate for Bad subset: Applying Lemma 4.6 to the interval $P_{\overline{\imath}}$ and the polynomial $\operatorname{covol}(V_{\overline{\imath}}, t)^2$ (using (4.21), and the definition of $\operatorname{Bad}(\overline{\imath}, \varepsilon)$), we get

$$|\operatorname{Bad}(\overline{\imath}, \varepsilon)| \ll_D \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} |P_{\overline{\imath}}|\,. \tag{4.22}$$

We now have to induct backwards to obtain the desired estimate for $\mathrm{Bad}(\varepsilon)$. In fact we claim that

$$\left| \bigcup_{(j_1,\dots,j_s)} \mathrm{Bad}\left((\overline{\imath}, j_1, \dots, j_s), \varepsilon\right) \right| \ll_{d,D} \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} |P_{\overline{\imath}}| \,. \tag{4.23}$$

If $\{P_{\overline{\imath}}, \mathscr{F}_{\overline{\imath}}\}$ is a bottom leaf of the tree in Figure 4.4, then this is the same bound as (4.22). If, on the other hand, it is not then we may assume that (4.23) already holds for $(\overline{\imath}, j_1)$ for all $j_1 = 1, 2, \dots$. Therefore

$$\begin{aligned}\left| \bigcup_{(j_1,\dots,j_s)} \mathrm{Bad}\left((\overline{\imath}, j_1, \dots, j_s)\varepsilon\right) \right| &\leqslant |\mathrm{Bad}(\overline{\imath}, \varepsilon)| + \sum_{j_1} \left| \bigcup_{(j_2,\dots,j_s)} \mathrm{Bad}\left((\overline{\imath}, j_1, \dots, j_s), \varepsilon\right) \right| \\ &\ll_{d,D} \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} |P_{\overline{\imath}}| + \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} \sum_{j_1} |P_{\overline{\imath},j_1}|\end{aligned}$$

by (4.22) for $\mathrm{Bad}(\overline{\imath}, \varepsilon)$ and the inductive hypothesis. Since the intervals

$$P_{\overline{\imath},1}, \dots, P_{\overline{\imath},m} \subseteq P_{\overline{\imath}}$$

are nearly disjoint we also have

$$\sum_{j_1} |P_{\overline{\imath},j_1}| \ll |P_{\overline{\imath}}|,$$

which concludes the inductive step. For $\overline{\imath} = \varnothing$ (the root at the top of the graded tree) this shows

$$|\mathrm{Bad}(\varepsilon)| \ll_{d,D} \left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} T. \tag{4.24}$$

Conclusion of the argument: It remains to show that

$$\{t \in [0,T] \mid \Gamma p(t) \notin \mathsf{X}_d(\varepsilon)\} \subseteq \mathrm{Bad}(\varepsilon), \tag{4.25}$$

since (4.24) then proves the theorem. Suppose therefore that

$$\Gamma p(t) \notin \mathsf{X}_d(\varepsilon),$$

or equivalently that there exists some vector $w \in \mathbb{Z}^d \smallsetminus \{0\}$ with $\|wp(t)\| < \varepsilon$. Since $\varepsilon \leqslant \eta$ we have $t \in P_{W,j}$ for $W = \mathbb{R}w$ and some j. Hence t lies in P_{i_1} for some i_1. If $t \in \mathrm{Bad}(i_1, \varepsilon) \subseteq \mathrm{Bad}(\varepsilon)$ then we have shown (4.25) for this value of t. So we may assume that $t \notin \mathrm{Bad}(i_1, \varepsilon)$. For the sake of the induction to come we continue the argument in greater generality.

Suppose we have shown (or rather, reduced the problem to the case) $t \in P_{\overline{\imath}}$ but $\varepsilon\eta^{\dim V-1} < \operatorname{covol}(V,t) \leqslant \eta^{\dim V}$ for all $V \in \mathscr{F}_{\overline{\imath}}$. Write

$$\mathscr{F}_{\overline{\imath}} = \{V_1 \subsetneq V_2 \subsetneq \cdots \subsetneq V_k\}$$

and assume that $a \in \{1,\dots,k\}$ is maximal with respect to the property

$$W = \mathbb{R}w \not\subseteq V_a.$$

This implies that

$$\operatorname{covol}(V_a + W, t) \leqslant \eta^{\dim V_a}\varepsilon \leqslant \eta^{\dim V_a}$$

and so $V_a + W \notin \mathscr{F}_{\overline{\imath}}$ and $V_a + W$ is compatible with $\mathscr{F}_{\overline{\imath}}$ (since it contains V_a and is contained in V_{a+1}). In other words, $\mathscr{F}_{\overline{\imath}}$ is not a complete flag and t belongs to one of the intervals defined by $V_a + W$, so that $t \in P_{(\overline{\imath}_1, i_{k+1})}$ for some i_{k+1}. If $t \in \operatorname{Bad}(\overline{\imath}, i_{k+1}, \varepsilon)$ then we are again done. That is, we have the same situation as before and can repeat the argument.

The iterative argument above will only stop when t lies in $\operatorname{Bad}(\varepsilon)$. Since every time the argument repeats we know that we only can have had a partial flag at the last stage, it can take at most d iterations to reach the conclusion. □

Corollary 4.10 (Non-escape of mass for X_d). *If $x \in \mathsf{X}_d$ and*

$$\{u_t \mid t \in \mathbb{R}\} < \mathrm{SL}_d(\mathbb{R})$$

is a one-parameter unipotent subgroup, then every weak-limit of the collection of measures*

$$\left\{\tfrac{1}{T}\int_0^T (u_t)_*\delta_x \,\mathrm{d}t \mid T > 0\right\}$$

is a probability measure on X_d.

Proof. Let $\Lambda_x < \mathbb{R}^d$ be the lattice corresponding to $x \in \mathsf{X}_d$, and define

$$\eta = \min\left\{\sqrt[k]{\alpha_k(\Lambda_x)} \mid 1 \leqslant k \leqslant d\right\}.$$

Fix an arbitrary $\varepsilon \in (0,\eta]$ and choose some $f \in C_c(\mathsf{X}_d)$ with

$$\mathbb{1}_{\mathsf{X}_d(\varepsilon)} \leqslant f \leqslant \mathbb{1} = \mathbb{1}_{\mathsf{X}_d}.$$

By Theorem 4.9 we have

$$1 - c\left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} \leqslant \frac{1}{T}\int_0^T f\left(u_t \cdot x\right) \,\mathrm{d}t \leqslant 1$$

for some constant $c = c_{d,D}$. Now choose a weak*-convergent subsequence of the measures

$$\frac{1}{T}\int_0^T (u_t)_* \, \delta_x \, \mathrm{d}t$$

to obtain the bound

$$1 - c\left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}} \leqslant \int_{\mathsf{X}_d} f \, \mathrm{d}\mu$$

for the limit measure μ. Since $f \leqslant 1$ this shows that

$$\mu\left(\mathsf{X}_d\right) \geqslant 1 - c\left(\frac{\varepsilon}{\eta}\right)^{\frac{1}{D}}.$$

As $\varepsilon \in (0, \eta]$ was arbitrary, the corollary follows. □

4.4 Closed Orbits (often) Have Finite Volume

In this section we return to the discussion of orbits $H{\cdot}x$ for a connected subgroup $H < \mathrm{SL}_d(\mathbb{R})$ and point $x \in \mathsf{X}_d = \mathrm{SL}_d(\mathbb{Z})\backslash \mathrm{SL}_d(\mathbb{R})$. Recall that H is called *semi-simple* if its Lie algebra is semi-simple, and that this implies that H is an almost direct product of normal simple subgroups (which may be compact or non-compact; see Section 2.2). We say that the subgroup H is *unipotent* if H can be conjugated into the strict upper-triangular subgroup

$$N = \left\{ \begin{pmatrix} 1 & * & \cdots & * \\ & 1 & * & * \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right\},$$

which implies that its Lie algebra is nilpotent. We note that semi-simple subgroups without compact factors and connected unipotent subgroups are both *unipotently generated*, meaning that it is generated by finitely many one-parameter unipotent subgrous. For these subgroups we can give another connection between the property of having a closed orbit and the property of having an orbit of finite volume. In fact we will prove a partial converse to Corollary 1.12.

Theorem 4.11 (Borel–Harish-Chandra theorem, Part I). *Let $x \in \mathsf{X}_d$, and let $H < \mathrm{SL}_d(\mathbb{R})$ be a connected subgroup which is semi-simple or unipotently generated. If the orbit $H{\cdot}x$ is closed, then it has finite volume*[†]*. In the case H is a connected unipotent subgroup, the orbit is compact.*

[†] That is, the orbit supports a finite H-invariant measure.

We refer to Exercise 4.4.2 for an immediate corollary (which is the standard way of phrasing the Borel–Harish-Chandra theorem) and to Section 7.4 for the general case of the theorem (which requires a few more definitions from the theory of algebraic groups).

Proof of Theorem 4.11 for semi-simple subgroups. Let

$$H = H_1 \cdots H_\ell H_{\text{compact}}$$

be the almost direct product of simple non-compact normal factors $H_1, \dots, H_\ell$ and a compact normal semi-simple subgroup H_{compact} as in Section 2.2. Now choose, for each H_i, a non-trivial unipotent one-parameter subgroup

$$U_i = \{u_i(t) \mid t \in \mathbb{R}\}$$

and define the diagonally embedded unipotent subgroup

$$U = \{u_1(t)u_2(t) \cdots u_n(t) \mid t \in \mathbb{R}\}.$$

By Proposition 2.11, this subgroup $U \leqslant H$ satisfies the following form of the Mautner phenomenon: If H acts unitarily on a Hilbert space[†] $\mathcal{H}$ and a vector is fixed by U, then the same vector is fixed by $H_1 \cdots H_\ell$.

Now choose a compact set $K \subseteq H{\cdot}x$ of positive volume with respect to the H-invariant Haar measure $m_{H{\cdot}x}$ on the orbit $H{\cdot}x \subseteq \mathsf{X}_d$ (as in Proposition 1.9 applied to $\mathrm{Stab}_H(x)\backslash H$). Since $K \subseteq \mathsf{X}_d$ is compact, we can find some $\eta \in (0, 1]$ such that

$$\alpha_k(\Lambda_x) \geqslant \eta^k$$

for $k = 1, \dots, d$ and any $x \in K$. Now apply Theorem 4.9 to find some $\varepsilon \in (0, \eta]$ with

$$\frac{1}{T} \left|\{t \in [0, T] \mid u_t{\cdot}x \notin \mathsf{X}_d(\varepsilon)\}\right| < \frac{1}{2} \tag{4.26}$$

for all $T > 0$. Since $H_{\text{compact}} \subseteq H$ is compact and $H{\cdot}x$ is closed, we have that (see Proposition 1.13)

$$\mathsf{X}_d(\varepsilon)H_{\text{compact}} \cap H{\cdot}x$$

is a compact subset of the orbit $H{\cdot}x$, and so

$$f = \mathbb{1}_{\mathsf{X}_d(\varepsilon)H_{\text{compact}}} \in L^2(H{\cdot}x, m_{H{\cdot}x})$$

is square-integrable with respect to $m_{H{\cdot}x}$. We define

$$\underline{f}(y) = \liminf_{n\to\infty} \frac{1}{n} \int_0^n f(u_t{\cdot}y)\, \mathrm{d}t.$$

Notice that

[†] In this instance, the Hilbert space will be $L^2(H{\cdot}x, m_{H{\cdot}x})$.

$$\left\| \frac{1}{n} \int_0^n f(u_t \cdot x)\, \mathrm{d}t \right\|^2_{L^2(m_{H\cdot x})} = \frac{1}{n^2} \int_0^n \int_0^n \underbrace{\int_{H\cdot x} f(u_{t_1}\cdot y) f(u_{t_2}\cdot y)\, \mathrm{d}m_{H\cdot x}(y)}_{\leqslant \|f\|^2_{L^2(m_{H\cdot x})}}$$
$$\leqslant \|f\|^2_{L^2(m_{H\cdot x})}.$$

Hence, by Fatou's lemma, we get

$$\begin{aligned} \|\underline{f}\|^2_{L^2(m_{H\cdot x})} &= \int_{H\cdot x} \liminf_{n\to\infty} \left(\frac{1}{n} \int_0^n f(u_t\cdot y)\, \mathrm{d}t \right)^2 \mathrm{d}m_{H\cdot x}(y) \\ &\leqslant \liminf_{n\to\infty} \int_{H\cdot x} \left(\frac{1}{n} \int_0^n f(u_t\cdot y)\, \mathrm{d}t \right)^2 \mathrm{d}m_{H\cdot x}(y) \\ &\leqslant \|f\|^2_{L^2(m_{H\cdot x})} < \infty, \end{aligned}$$

or equivalently $\underline{f} \in L^2(m_{H\cdot x})$.

We can now finish the proof quite quickly. Since $f \in L^2(m_{H\cdot x})$ is u_t-invariant for all $t \in \mathbb{R}$ by construction, it is also $H_1 \cdots H_\ell$-invariant by the Mautner phenomenon (Proposition 2.11). Furthermore, $f = \mathbb{1}_{\mathsf{X}_d(\varepsilon) H_{\mathrm{compact}}}$ is invariant under H_{compact} by definition. Since u_t commutes with H_{compact}, it follows that $\underline{f}$ is also invariant under H_{compact}. Since $H = H_1 \cdots H_\ell H_{\mathrm{compact}}$ and $\underline{f} \in L^2(m_{H\cdot x})$ we see that $\underline{f} \equiv c$ is equal $m_{H\cdot x}$-almost everywhere to some constant c. By definition and (4.26) we have $c \geqslant \frac{1}{2}$ and so

$$c^2 m_{H\cdot x}(H\cdot x) = \|\underline{f}\|^2_{L^2(m_{H\cdot x})} < \infty$$

implies that $H\cdot x$ has finite volume. □

Proof of Theorem 4.11 for unipotent subgroups. In the proof of Theorem 4.11 for the semi-simple case it was convenient that we could find one one-parameter unipotent subgroup that satisfied the hypothesis of the Mautner phenomenon for 'most' of H. In the general case, we have instead to use finitely many one-parameter unipotent subgroups $U_j = \{u_j(t) \mid t \in \mathbb{R}\}$ for $j = 1, \ldots n$ that together generate H.

Let $K \subseteq H\cdot x$ be a compact set. Then, finding first $\eta > 0$ and then $\varepsilon \in (0, \eta]$ as above, there exists a compact subset $L \subseteq H\cdot x$ (where $L = \mathsf{X}_d(\varepsilon) \cap H\cdot x$, relying on the assumption that $H\cdot x$ is closed) such that

$$\frac{1}{T} \left| \{ t \in [0,T] \mid u_1(t)\cdot y \notin L \} \right| < \frac{1}{2} \tag{4.27}$$

for all $y \in K$. Now let $f = \mathbb{1}_L \in L^2(m_{H\cdot x})$ and

$$f_1(y) = \underline{f}(y) = \liminf_{n\to\infty} \frac{1}{n} \int_0^n f(u_1(t)\cdot y)\, \mathrm{d}t$$

so that $f_1 \in L^2(m_{H\cdot x})$, f_1 is U_1-invariant, and $f_1(y) \geqslant \frac{1}{2}$ for all $y \in K$.

Suppose now that for $j \leqslant n$ we have already shown that for any compact set $K \subseteq H{\cdot}x$ there exists some $f_j \in L^2(m_{H{\cdot}x})$ which is U_1-invariant, U_2-invariant, and so on up to U_j-invariant, and satisfies $f_j(y) \geqslant (\frac{1}{2})^j$ for all y in K. If $j = n$, then the function is H-invariant and the theorem follows as before.

So suppose that $j < n$ and let $K \subseteq H{\cdot}x$ be a compact subset. Now choose $L \subseteq H{\cdot}x$ as in (4.27) but for $u_j(t)$ instead of $u_1(t)$. Next apply the inductive hypothesis to L to find a function $f_j \in L^2(m_{H{\cdot}x})$ which is invariant under $U_1, U_2, \ldots, U_j$ and satisfies $f_j(y) \geqslant (\frac{1}{2})^j$ for all $y \in L$. We define

$$f_{j+1}(y) = \underline{f_j}(y) = \liminf_{n\to\infty} \frac{1}{n} \int_0^n f_j(u_t{\cdot}y)\,\mathrm{d}t.$$

By construction of f_j, L, and f_{j+1} we know that $f_{j+1} \in L^2(m_{H{\cdot}x})$, that f_{j+1} is U_{j+1}-invariant, and that $f_{j+1}(y) \geqslant (\frac{1}{2})^{j+1}$ for all $y \in K$. However, at first sight it may not be clear why f_{j+1} is still invariant under U_i for $i = 1, \ldots, j$ (since U_i may not commute with U_{j+1}). Here the Mautner phenomenon comes to the rescue. In fact, by Theorem 2.15, f_j is actually invariant under a normal subgroup $N \lhd H$ containing $U_1, \ldots, U_j$. Therefore

$$u_{j+1}(t)u_i(s) = n_{s,t}u_{j+1}(t)$$

for all $i = 1, \ldots, j$, and $s, t \in \mathbb{R}$, and some $n_{s,t} \in N$. This shows that

$$f_j(u_{j+1}(t)u_i(s){\cdot}y) = f_j(n_{s,t}u_{j+1}(t){\cdot}y) = f_j(u_{j+1}(t){\cdot}y)$$

for $m_{H{\cdot}x}$-almost every y. Integrating over $t \in [0, n]$ and taking the limit infimum as in the definition of f_{j+1}, we get

$$f_{j+1}(u_i(s){\cdot}y) = f_{j+1}(y).$$

This concludes the induction and so also the proof of the first statement Theorem 4.11 for unipotently generated subgroups.

It remains to show that $H{\cdot}x$ is compact if H is unipotent. Note that by assumption, H can be conjugated into the upper triangular unipotent subgroup. On the upper triangular unipotent subgroup, the logarithm map is a polynomial with a polynomial inverse. This implies that H consists of the image of the Lie algebra of H. Hence we see that a unipotent connected subgroup H consists of the $\mathbb{R}$-points $H = \mathbb{H}(\mathbb{R})$ of an algebraic subgroup $\mathbb{H}$ over $\mathbb{R}$. If $x = \mathrm{SL}_d(\mathbb{Z})g$, then we may conjugate H by $g \in G$ and assume without loss of generality that $x = \mathrm{SL}_d(\mathbb{Z})$. Then the Borel density theorem (Theorem 3.30) implies that the intersection $H \cap \mathrm{SL}_d(\mathbb{Z})$ is Zariski dense in $\mathbb{H}$, which in turn implies that $\mathbb{H}$ is an algebraic group over $\mathbb{Q}$. Hence the Lie algebra of H is a rational subspace of $\mathfrak{sl}_d(\mathbb{R})$, and by Theorem 3.9 we see that $H{\cdot}\,\mathrm{SL}_d(\mathbb{Z})$ is compact. □

Exercises for Section 4.4

Exercise 4.4.1. Let Q be a real non-degenerate quadratic form of signature (p, q) in $d \geqslant 3$ variables with $p \geqslant q \geqslant 1$. Suppose that the orbit $\mathrm{SL}_d(\mathbb{Z})\,\mathrm{SO}(Q)(\mathbb{R})$ is closed. Show that a multiple of Q has integer coefficients.

Exercise 4.4.2. [(19)] Let $\mathbb{G} < \mathrm{SL}_d$ be a semi-simple or unipotent algebraic group defined over $\mathbb{Q}$. Show that $\mathbb{G}(\mathbb{Z}) = \mathbb{G}(\mathbb{R}) \cap \mathrm{SL}_d(\mathbb{Z})$ is a lattice in $\mathbb{G}(\mathbb{R})$.

Notes to Chapter 4

[(16)](Page 136) This result, or rather its higher-dimensional counterpart in Section 4.3, has a long history; see Margulis [124], [125]; Dani [21], [25]; Kleinbock and Margulis [101]; Kleinbock [97].
[(17)](Page 141) This is a simple special case of the Besicovitch covering lemma (see [5]).
[(18)](Page 147) As mentioned before, this result has a long history; see Margulis [124], [125]; Dani [21], [25]; Kleinbock and Margulis [101]; Kleinbock [97].
[(19)](Page 157) This is a special case of the Borel–Harish-Chandra theorem [9].