Zeitschrift: L'Enseignement Mathématique
Herausgeber: Commission Internationale de l'Enseignement Mathématique
Band: 58 (2012)

Artikel: The distribution of closed geodesics on the modular surface, and duke's theorem
Autor: Einsiedler, Manfred / Lindestrauss, Elon / Michel, Philippe
DOI: https://doi.org/10.5169/seals-515821

L'Enseignement Mathématique (2) 58 (2012), 249–313

THE DISTRIBUTION OF CLOSED GEODESICS ON THE MODULAR SURFACE, AND DUKE'S THEOREM

by Manfred EINSIEDLER [*], Elon LINDENSTRAUSS [‡], Philippe MICHEL [§] and Akshay VENKATESH [¶]

ABSTRACT. We give an ergodic theoretic proof of a theorem of Duke about equidistribution of closed geodesics on the modular surface. The proof is closely related to the work of Yu. Linnik and B. Skubenko, who in particular proved this equidistribution under an additional congruence assumption on the discriminant. We give a more conceptual treatment using entropy theory, and show how to use positivity of the discriminant as a substitute for Linnik's congruence condition.

CONTENTS

[*]) M. E. acknowledges the support by the Clay Mathematics Institute as a Research Scholar, by the NSF (grant 0554373) and the SNF (grant 200021-127145).

[‡]) E. L. acknowledges the support of NSF (grants DMS-0554345 and 0800345), the ISF (grant 983/09) and the European Research Council (Advanced Research Grant 267259).

[§]) Ph. M. was partially supported by the SNF (grant 200021-125291) and the European Research Council (Advanced Research Grant 228304).

[¶]) A. V. was supported by the Clay Mathematics Institute and by NSF Grant DMS-0903110.

1. Introduction

A non-zero integer d is called a *discriminant* if it can be represented in the form

$$d = b^2 - 4ac\,, \quad a, b, c \in \mathbf{Z}\,,$$

or equivalently if d is the discriminant of the binary quadratic form with integral entries

$$q(x, y) = ax^2 + bxy + cy^2. \tag{1.1}$$

It is easy to see that d is a discriminant if and only if $d \equiv 0, 1 \pmod 4$. A discriminant d is *fundamental* if d is either square-free (in which case d is congruent to 1 modulo 4) or $d/4$ is a square-free integer congruent to $2, 3 \pmod 4$. Equivalently: d is fundamental if it is the discriminant of the ring of integers of a quadratic field.

The study of integral binary quadratic forms goes back at least to the Greeks. Significant breakthroughs were accomplished by Gauss. In his *Disquisitiones Arithmeticae* he studied the set of $\mathrm{GL}_2(\mathbf{Z})$-orbits of such forms, where $\mathrm{GL}_2(\mathbf{Z})$ acts via the linear change of variables:

$$\gamma \,.\, q(x, y) = \frac{1}{\det(\gamma)} q((x, y)\gamma) = \frac{1}{\det(\gamma)} q(ux + wy, vx + zy)\,, \tag{1.2}$$

for $\gamma = \begin{pmatrix} u & v \\ w & z \end{pmatrix} \in \mathrm{GL}_2(\mathbf{Z})$. This action preserves the discriminant and Gauss proved that the set of $\mathrm{GL}_2(\mathbf{Z})$-orbits of integral binary quadratic forms of a given discriminant is finite, see [7, p. 128] for an accessible and more general treatment. Let

$$\begin{aligned} \mathrm{R}_{\mathrm{disc}}(d) &= \{q(x, y) = ax^2 + bxy + cy^2 :\ a, b, c \in \mathbf{Z},\ \mathrm{disc}(q) = d,\ \gcd(a, b, c) = 1\} \\ &\simeq \{(a, b, c) \in \mathbf{Z}^3 :\ \mathrm{disc}(a, b, c) = b^2 - 4ac = d,\ \gcd(a, b, c) = 1\} \end{aligned}$$

denote the set of forms of discriminant d with coprime coefficients, and let

$$[\mathrm{R}_{\mathrm{disc}}(d)] = \mathrm{GL}_2(\mathbf{Z}) \backslash \mathrm{R}_{\mathrm{disc}}(d)$$

be the set of orbits; its cardinality is the *class number* and is noted $h(d)$. Gauss also showed that the set $[\mathrm{R}_{\mathrm{disc}}(d)]$ could be given an additional structure of an abelian group (the law of composition of quadratic forms), leading to the notion of *class group* of quadratic forms of discriminant d. Nowadays these venerable and beautiful results are usually interpreted in terms of the theory of quadratic fields and ideal class groups. We will recall this connection below.

1.1 Linnik and Skubenko equidistribution theorems

In the late 50's, Linnik studied more refined properties of the set of representations $\mathrm{R}_{\mathrm{disc}}(d)$, in particular their distribution properties.

Let

$$V_{\mathrm{disc},\pm 1}(\mathbf{R}) = \{(a,b,c) \in \mathbf{R}^3 : \ b^2 - 4ac = \pm 1\};$$

this is a one-sheeted hyperboloid in the $+1$ case and a two-sheeted hyperboloid in the -1 case, and is identified with the set of real binary quadratic form with discriminant ± 1. In both cases $V_{\mathrm{disc},\pm 1}(\mathbf{R})$ is invariant under the natural action of $\mathrm{GL}_2(\mathbf{R})$ extending (1.2) and has one orbit.

The set of representation $\mathrm{R}_{\mathrm{disc}}(d)$ projects on $V_{\mathrm{disc},\pm 1}(\mathbf{R})$ (with $\pm 1 = \mathrm{sign}(d)$) by a homothety

$$|d|^{-1/2}\mathrm{R}_{\mathrm{disc}}(d) \subset V_{\mathrm{disc},\pm 1}(\mathbf{R}),$$

and Linnik studied how this set is distributed when $d \to \infty$. These hyperboloids carry a natural $\mathrm{GL}_2(\mathbf{R})$-invariant measure $\mu_{\mathrm{disc},\pm 1}$ defined, for any open set $\Omega \subset V_{\mathrm{disc},\pm 1}(\mathbf{R})$, as the Lebesgue measure in $\mathbf{R}^3$ of the solid cone emanating from the origin and ending at Ω, i.e.

$$\mu_{\mathrm{disc},\pm 1}(\Omega) = \mu_{\mathbf{R}^3}(\mathcal{C}(\Omega)),$$

where

$$\mathcal{C}(\Omega) = \{r.\mathbf{x} : \ \mathbf{x} \in \Omega, \ r \in [0,1]\}.$$

Using an original argument of ergodic theoretic flavor, Linnik [19, Chap. V] established the following equidistribution statement for *negative discriminants*.

THEOREM 1.1 (Linnik). *Let $p > 2$ be a fixed prime. As $d \to -\infty$ amongst the negative discriminants such that $\left(\frac{d}{p}\right) = 1$, the set*

$$|d|^{-1/2}\mathrm{R}_{\mathrm{disc}}(d) \subset V_{\mathrm{disc},-1}(\mathbf{R}),$$

becomes equidistributed with respect to $\mu_{\mathrm{disc},-1}$, in the following sense: for any two continuous compactly supported functions φ_1, φ_2 on $V_{\mathrm{disc},-1}(\mathbf{R})$ such that the integral $\mu_{\mathrm{disc},-1}(\varphi_2) \neq 0$ we have

$$\frac{\sum_{x \in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_1(|d|^{-1/2}x)}{\sum_{x \in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_2(|d|^{-1/2}x)} \to \frac{\mu_{\mathrm{disc},-1}(\varphi_1)}{\mu_{\mathrm{disc},-1}(\varphi_2)} \quad \text{as } d \to -\infty.$$

In particular, $\sum_{x \in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_2(|d|^{-1/2}x) \neq 0$ if d as above is large enough.

Building on Linnik's ergodic method Skubenko [24] (see also [19, Chap. VI.]) proved the analogous statement for *positive discriminants*:

THEOREM 1.2 (Skubenko). *Let $p > 2$ be a fixed prime. As $d \to +\infty$ amongst the positive discriminants such that $\left(\frac{d}{p}\right) = 1$, the set*

$$|d|^{-1/2}\mathrm{R}_{\mathrm{disc}}(d) \subset V_{\mathrm{disc},+1}(\mathbf{R}),$$

becomes equidistributed with respect to $\mu_{\mathrm{disc},+1}$, in the following sense: for any two continuous compactly supported functions φ_1, φ_2 on $V_{\mathrm{disc},+1}(\mathbf{R})$ such that the integral $\mu_{\mathrm{disc},+1}(\varphi_2) \neq 0$ we have

$$\frac{\sum_{x\in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_1(|d|^{-1/2}x)}{\sum_{x\in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_2(|d|^{-1/2}x)} \to \frac{\mu_{\mathrm{disc},+1}(\varphi_1)}{\mu_{\mathrm{disc},+1}(\varphi_2)} \quad \text{as } d \to +\infty.$$

In particular, $\sum_{x\in \mathrm{R}_{\mathrm{disc}}(d)} \varphi_2(|d|^{-1/2}x) \neq 0$ if d as above is large enough.

We refer to Figure 1 for an illustration of the case $d = 377$.

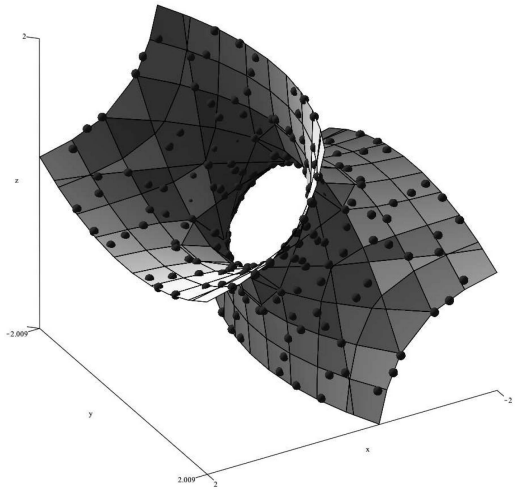

FIGURE 1
The distribution of $377^{-1/2}\mathrm{R}_{\mathrm{disc}}(377)$ viewed on the one-sheeted hyperboloid: $h(377) = 1$

The condition $\left(\frac{d}{p}\right) = 1$ for some fixed prime p is equivalent to the condition that

the fixed prime p splits in the quadratic field $\mathbf{Q}(\sqrt{d})$.

This condition (which we shall refer to as *Linnik's condition*) was an essential input for Linnik's ergodic method but, as was pointed out by Linnik himself, it should not be necessary for the equidistribution theorem to hold. It was only thirty years later that this condition was removed in the beautiful work of Duke [9].

1.2 DUKE'S THEOREM

A key point of Duke's approach is to reformulate the prior theorems in a dual form: in terms of equidistribution of "Heegner points" (for negative d) or of closed geodesics (for positive d) on the modular surface $Y_0(1) := \mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$.

Assuming that $d > 0$ is not a square, one associates to any $(a,b,c) \in \mathrm{R}_{\mathrm{disc}}(d)$ the geodesic corresponding to the geodesic semi-circle in the upper half-plane whose end points are

$$x_{a,b,c,\pm} = \frac{-b \pm \sqrt{d}}{2a}\,. \tag{1.3}$$

We lift this geodesic in the obvious way to the unit tangent bundle of $\mathbf{H}$ and then project it to a geodesic orbit on the unit tangent bundle $\mathbf{T}^1(Y_0(1))$. This geodesic orbit, which we denote by $\gamma_{[a,b,c]}$, is compact and depends only on the $\mathrm{SL}_2(\mathbf{Z})$-orbit of (a,b,c). We obtain in this way a collection of $h(d)$ closed geodesics

$$\mathcal{G}_d = \bigcup_{[a,b,c]} \gamma_{[a,b,c]} \subset \mathbf{T}^1(Y_0(1))\,,$$

see Figure 2 for the case $d = 377$. This collection of compact orbits of the geodesic flow then carries a natural probability measure invariant under the geodesic flow which we denote by μ_d. Let μ_{L} be the Liouville (Haar) probability measure on $\mathbf{T}^1(Y_0(1))$, then Duke's theorem (as extended by Chelluri [8] to the unit tangent bundle) gives the following:

THEOREM 1.3 (Duke). *As $d \to +\infty$ amongst the positive fundamental discriminants, the set $\mathcal{G}_d$ becomes equidistributed on the unit tangent bundle $\mathbf{T}^1(Y_0(1))$ with respect to the measure μ_{L}: for any continuous compactly supported function φ on $\mathbf{T}^1(Y_0(1))$,*

$$\int_{\mathcal{G}_d} \varphi(t) d\mu_d(t) \to \int_{\mathbf{T}^1(Y_0(1))} \varphi(u) d\mu_{\mathrm{L}}(u)\,.$$

The equivalence of the equidistribution statements in Theorem 1.2 and Theorem 1.3 will be explained in §2.4.

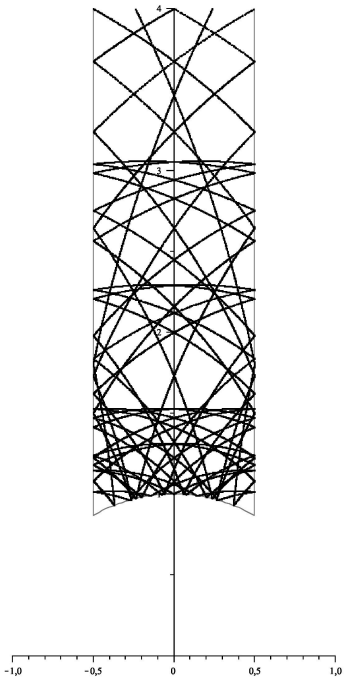

FIGURE 2
The distribution of $\mathcal{G}_{377}$ projected on the fundamental domain of $\mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$

The restriction to fundamental discriminants is not essential; indeed all the proofs extend to the general case, including the one we present here. Duke's proof is fundamentally different from Linnik's; it does not rely on ergodic theory but on harmonic analysis of the modular surface $\mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$, that is on the theory of automorphic forms supplemented by deep arguments from analytic number theory and in particular a breakthrough of Iwaniec [17].

In this paper we give a new proof of Duke's theorem in the case of positive discriminant. Our proof is strongly influenced by Linnik's ergodic method, and may be seen as a modern incarnation of Linnik's original ideas, and we use the positivity of the discriminant as a substitute to Linnik's condition that Skubenko relied on in his work.

There are two main ingredients in the proof:

1. *Linnik's Basic Lemma* — An upper bound on the number of nearby pairs of points in the projection of $\mathrm{R}_{\mathrm{disc}}(d)$ to $V_{\mathrm{disc},-1}(\mathbf{R})$ (as this set is infinite,

the quantity to be bounded needs some additional interpretation), which eventually reduces to an upper bound on the number of ways a given binary quadratic form can be represented by a ternary quadratic form.

2. The *uniqueness of measure of maximal entropy* for the flow corresponding to the one parameter group $a_t = \begin{pmatrix} e^t & \\ & e^{-t} \end{pmatrix}$ on $\mathrm{SL}_2(\mathbf{Z})\backslash \mathrm{SL}_2(\mathbf{R})$.

We have made an effort to present both of these main ingredients in a self-contained way, as each relies on some well-known results that are unfortunately well known in essentially disjoint circles of mathematicians.

The second of these two ingredients replaces a more explicit but less conceptual argument of Linnik and Skubenko. The uniqueness of the measure of maximal entropy for this action is well known (both in the cocompact and finite volume case) and in the cocompact case dates back to work of R. Bowen [4]. However the version we give here is new in that it allows us to control how much weight $\mathcal{G}_d$ gives to small neighborhoods of the cusp in $\mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$: essentially, we give a finitary version of the uniqueness of measure of maximal entropy in the *noncompact* quotient $\mathrm{SL}_2(\mathbf{Z})\backslash \mathrm{SL}_2(\mathbf{R})$. This finitary version is the content of Theorem 4.2, and involves a careful analysis of how much entropy can be carried by a_t-invariant measures that give disproportionately high weight to the cusp. A cleaner version of the relationship between entropy and mass in the cusp (although not directly applicable for our main purposes) is given in Theorem 5.1. We believe these results are of independent interest, and will likely have other applications; it also raises some interesting new questions (see e.g. [11]).

We mention that another modern exposition of Linnik's method in a similar context (distribution of integer points on spheres) by J. Ellenberg and two of us (Ph. M. and A. V.) has appeared already in [14]. In that work Linnik's Basic Lemma is again a central ingredient, complemented by a different argument to convert the upper bounds provided by the Basic Lemma to equidistribution (i.e. both upper and lower bounds on the number of points in specified regions). The reader may wish to compare these two complementary approaches.

1.3 NOTATION

We collect here some notation that is used throughout the paper:

The group $\mathrm{SL}_2(\mathbf{R})$ acts transitively on the upper half-plane model $\mathbf{H}$ of the hyperbolic plane by fractional linear transformations and the stabilizer of the point i is the compact subgroup $\mathrm{SO}_2(\mathbf{R})$. The resulting identification

$$\mathbf{H} \simeq \mathrm{SL}_2(\mathbf{R})/\mathrm{SO}_2(\mathbf{R})$$

descends to an identification of $\mathbf{H}$ with $\mathrm{PSL}_2(\mathbf{R})/\mathrm{PSO}_2(\mathbf{R})$; moreover the action of $\mathrm{PSL}_2(\mathbf{R})$ on the unit tangent bundle $\mathbf{H}$ is simply transitive. If we let $p \in T^1(\mathbf{H})$ be the tangent vector pointing up at i, then $g \mapsto gp$ gives an identification $\mathrm{PSL}_2(\mathbf{R}) \simeq T^1\mathbf{H}$. Taking the quotient by $\mathrm{PSL}_2(\mathbf{Z})$ we obtain an identification with the unit tangent bundle of the modular curve [1] $\mathrm{PSL}_2(\mathbf{Z})\backslash \mathrm{PSL}_2(\mathbf{R}) \simeq T^1(\mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H})$.

We shall make use of another identification of the quotient

$$\mathrm{PSL}_2(\mathbf{Z})\backslash \mathrm{PSL}_2(\mathbf{R}),$$

namely with the space of lattices in $\mathbf{R}^2$ up to homothety. Indeed, the space of lattices $\mathcal{L}_2(\mathbf{R})$ is identified with $\mathrm{GL}_2(\mathbf{Z})\backslash \mathrm{GL}_2(\mathbf{R})$ via $g \mapsto \mathbf{Z}^2.g$; the same map also identifies the space $[\mathcal{L}_2(\mathbf{R})]$ of lattices up to homothety with $\mathrm{PGL}_2(\mathbf{Z})\backslash \mathrm{PGL}_2(\mathbf{R})$ and the set $\mathcal{L}_2^{(1)}(\mathbf{R}) = X$ of lattices of covolume one with $\mathrm{SL}_2(\mathbf{Z})\backslash \mathrm{SL}_2(\mathbf{R}) = \mathrm{PSL}_2(\mathbf{Z})\backslash \mathrm{PSL}_2(\mathbf{R})$. Finally, the sets $[\mathcal{L}_2(\mathbf{R})]$ and $\mathcal{L}_2^{(1)}(\mathbf{R})$ are also identified via the map $[L] \mapsto \mathrm{vol}(L)^{-1/2}.L$.

Thus the following spaces are identified:

$$X \simeq \mathrm{PSL}_2(\mathbf{Z})\backslash \mathrm{PSL}_2(\mathbf{R}) \simeq T^1(\mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}) \simeq [\mathcal{L}_2(\mathbf{R})] \simeq \mathcal{L}_2^{(1)}(\mathbf{R}).$$

We take the following fundamental domain

$$\mathcal{S} = \{(z, v) \in \mathbf{H} \times S^1 : \ |\Re z| \le 1/2, \ |z| \ge 1\} \subset T^1(\mathbf{H}) \simeq \mathrm{PSL}_2(\mathbf{R})$$

for $\mathrm{PSL}_2(\mathbf{Z}) = \Gamma$.

Fix an arbitrary left-invariant Riemannian metric d on $\mathrm{PSL}_2(\mathbf{R})$. It descends to a metric on X, denoted d_X or simply d for short. Explicitly we have

$$d_X(\mathrm{PSL}_2(\mathbf{Z})g_1, \mathrm{PSL}_2(\mathbf{Z})g_2) = \min_{\gamma \in \mathrm{PSL}_2(\mathbf{Z})} d(g_1, \gamma g_2). \tag{1.4}$$

The geodesic curves on $T^1(\mathbf{H})$ — which in the upper half-plane are circles and lines intersecting the real axis in a normal angle — correspond to the orbits of the right A-orbits in $\mathrm{PSL}_2(\mathbf{R})$, where $A = \{a_t\}$ is the diagonal subgroup of $\mathrm{PSL}_2(\mathbf{R})$. By a slight abuse, we shall use A to refer to the diagonal subgroup of all three groups: $\mathrm{GL}_2(\mathbf{R}), \mathrm{PGL}_2(\mathbf{R})$ and $\mathrm{SL}_2(\mathbf{R})$.

ACKNOWLEDGEMENTS. The authors would like to thank Peter Sarnak for encouragement and many helpful conversations. A. V. would also like to thank Jordan Ellenberg for many discussions on the topic of quadratic forms. The authors also thank Menny Aka, Asaf Katz, Ilya Khayutin, Lior Rosenzweig for carefully going over a preliminary version of this paper.

[1] Actually the modular curve has singularities at the points i and $j = \frac{1+\sqrt{-3}}{2}$ owing to the fact that these points have non-trivial stabilizers in $\mathrm{PSL}_2(\mathbf{Z})$; we will ignore this minor point.

2. Representations by the discriminant, orbits and quadratic fields

In this section we explain in greater detail the relationship between Skubenko's equidistribution theorem and Duke's and connect these questions to the arithmetic of real quadratic fields. Along the way we will find a few equivalent ways in which to describe compact A-orbits in $\mathcal{G}_d$. Building on that we prove in §2.4 the equivalence between Skubenko's and Duke's formulations.

2.1 Overview of the bijections

Recall that we have previously associated to any element of $[\mathrm{R}_{\mathrm{disc}}(d)]$ — i.e. to any $\mathrm{GL}_2(\mathbf{Z})$-orbit in $\mathrm{R}_{\mathrm{disc}}(d)$ — a closed geodesic on $\mathrm{SL}_2(\mathbf{Z})\backslash\mathbf{H}$. On the other hand, as discussed in §1.3, a closed geodesic in $\mathcal{G}_d$ corresponds to a closed A-orbit on the space X.

Write $\mathcal{O}_d := \mathbf{Z}[\frac{d+\sqrt{d}}{2}]$ for the order of discriminant d.

We shall show below that the following sets are in natural bijection to each other:

(i) $[\mathrm{R}_{\mathrm{disc}}(d)]$, the set of $\mathrm{GL}_2(\mathbf{Z})$-orbits of primitive representations in $\mathrm{R}_{\mathrm{disc}}(d)$.

(ii) The set of $\mathrm{GL}_2(\mathbf{Z})$-conjugacy classes of ring embeddings $\iota\colon \mathcal{O}_d \hookrightarrow M_2(\mathbf{Z})$ which are *optimal*, i.e. for which the embedding cannot be extended to an embedding of a strictly bigger order $\mathcal{O} \gneq \mathcal{O}_d$ with image in $M_2(\mathbf{Z})$.

(iii) $\mathrm{Cl}(\mathcal{O}_d)$ = the set of $K^\times$-homothety classes of proper $\mathcal{O}_d$-ideals, where $K = \mathbf{Q}(\sqrt{d})$.

In the case of a fundamental discriminant the above objects and their bijections are a bit easier to explain. In fact, if d is a fundamental discriminant, then every representation is primitive, every embedding is optimal, and every $\mathcal{O}_d$-ideal is proper. In reading the remainder of the section the reader may first specialize to this case, or even continue reading with Section 3 and only refer to the portions of this section as needed for the remainder of the paper.

2.2 Discriminant and quadratic fields

We establish the bijections of §2.1.

Before beginning, we note that the sequence of maps

$$ax^2 + bxy + cy^2 \mapsto \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} \mapsto \begin{pmatrix} b & -2a \\ 2c & -b \end{pmatrix} \tag{2.1}$$

defines an isometry between the spaces of (real) binary quadratic forms, symmetric 2×2 real matrices and trace zero 2×2 real matrices, where each

of those is equipped with a quadratic form:

$$(\mathcal{Q}(\mathbf{R}^2), \text{disc}) \simeq (\text{Sym}_2(\mathbf{R}), -4\det) \simeq (M_2^0(\mathbf{R}), -\det).$$

The action of $\text{GL}_2(\mathbf{Z})$ in (1.2) is the restriction of the following action of $\text{GL}_2(\mathbf{R})$ on $\mathcal{Q}(\mathbf{R}^2)$:

$$g.q(x,y) = \frac{1}{\det(g)} q((x,y)g) = \frac{1}{\det(g)} q(ux+wy, vx+zy), \quad g = \begin{pmatrix} u & v \\ w & z \end{pmatrix},$$

which intertwines with the actions

$$g.(ax^2+bxy+cy^2) \longleftrightarrow \frac{1}{\det(g)} g \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} {}^t g \longleftrightarrow g \begin{pmatrix} b & -2a \\ 2c & -b \end{pmatrix} g^{-1}.$$

Observe that these actions factor through $\text{PGL}_2(\mathbf{R})$. They also induce an isomorphism between $\text{PGL}_2(\mathbf{Z})$ and the group of orthogonal transformations of $(\mathcal{Q}(\mathbf{R}^2), \text{disc})$ preserving the integral quadratic forms.

Let d be a discriminant which is not a perfect square; let $(a,b,c) \in \text{R}_{\text{disc}}(d)$ be a representation, and let

$$m = m_{a,b,c} = \begin{pmatrix} b & -2a \\ 2c & -b \end{pmatrix} \tag{2.2}$$

be the trace zero matrix associated to it via the map (2.1). Since

$$m^2 = d \cdot \text{Id}$$

this defines an embedding of the quadratic field (d is not a square) $K = \mathbf{Q}(\sqrt{d})$ into $M_2(\mathbf{Q})$

$$\iota_m: \begin{array}{ccc} K & \hookrightarrow & M_2(\mathbf{Q}) \\ u+v\sqrt{d} & \mapsto & u\,\text{Id} + v.m \end{array}$$

2.2.1 REPRESENTATIONS AND OPTIMAL EMBEDDING. The integrality properties of this embedding are measured by considering

$$\mathcal{O}_m := \iota_m^{-1}(M_2(\mathbf{Z}))$$

which is an order in K. Let us identify which order: Note that $\mathcal{O}_{\lambda.m} = \mathcal{O}_m$ for any $\lambda \in \mathbf{Q}^\times$. Hence if $b^2 - 4ac = d$ for $a,b,c \in \mathbf{Z}$ we may write

$$(a,b,c) = f(a',b',c')$$

with $f \in \mathbf{Z}$ and $a',b',c' \in \mathbf{Z}$ coprime integers satisfying

$$\text{disc}(a',b',c') = d' = d/f^2.$$

This reduces the discussion to the case where (a,b,c) is a *primitive representation* of d (a representation with coprime entries).

Assuming that (a,b,c) is primitive, one sees quickly that

$$\mathcal{O}_m = \mathcal{O}_d = \mathbf{Z}[\tfrac{d+\sqrt{d}}{2}] \tag{2.3}$$

is the order of *discriminant* d. If (2.3) holds, we say that ι_m defines an optimal embedding of $\mathcal{O}_d$ into $M_2(\mathbf{Z})$. We obtain in that way a bijection between

the set of $\mathrm{GL}_2(\mathbf{Z})$-orbits of primitive representations $[\mathrm{R}_{\mathrm{disc}}(d)]$

and

the set of $\mathrm{GL}_2(\mathbf{Z})$-conjugacy classes of optimal embeddings $\iota\colon \mathcal{O}_d \hookrightarrow M_2(\mathbf{Z})$.

2.2.2 EMBEDDINGS AND IDEAL CLASSES. Let us recall that a lattice $I \subset K$ is a *proper* $\mathcal{O}_d$*-ideal*, iff

$$\mathcal{O}_I := \{\lambda \in K :\ \lambda.I \subset I\} = \mathcal{O}_d\,.$$

Then there is a bijection between

the set of $\mathrm{GL}_2(\mathbf{Z})$-conjugacy classes of optimal embeddings of $\mathcal{O}_d$

and the set of *proper ideal classes* of $\mathcal{O}_d$

$\mathrm{Cl}(\mathcal{O}_d) =$ the set of $K^\times$-homothety classes of proper $\mathcal{O}_d$-ideals.

This bijection goes as follows [18]: Given a proper $\mathcal{O}_d$-ideal $I \subset K$, one may choose a $\mathbf{Z}$-basis $I = \mathbf{Z}.\alpha + \mathbf{Z}.\beta$ which gives an identification

$$\theta\colon \begin{array}{ccc} I & \to & \mathbf{Z}^2 \\ u\alpha + v\beta & \mapsto & (u,v) \end{array}$$

This identification induces the embedding

$$\iota\colon K \hookrightarrow M_2(\mathbf{Q})$$

defined by

$$\iota(\lambda)(u,v) = \theta(\lambda.(u\alpha + v\beta))$$

(or in other terms, such that $\theta(\lambda.x) = \theta(x)\iota(\lambda)$).

Since $\mathcal{O}_d.I \subset I$, one has $\iota(\mathcal{O}_d)\mathbf{Z}^2 \subset \mathbf{Z}^2$, that is $\iota(\mathcal{O}_d) \subset M_2(\mathbf{Z})$ and the fact that I is a proper $\mathcal{O}_d$-ideal is equivalent to the fact that ι is an optimal embedding of $\mathcal{O}_d$. If we replace the $\mathbf{Z}$-basis (α,β) by another basis (α',β') then ι is replaced by a $\mathrm{GL}_2(\mathbf{Z})$-conjugate. Finally if I is replaced by an ideal in the same class $I' = \lambda.I$, $\lambda \in K^\times$, then the corresponding $\mathrm{GL}_2(\mathbf{Z})$-conjugacy classes coincide: $[\iota_{I'}] = [\iota_I]$.

The inverse of the map

$$[I] \mapsto [\iota_I]$$

is as follows: given an optimal embedding $\iota\colon K \hookrightarrow M_2(\mathbf{Q})$ of $\mathcal{O}_d$, let $e_1 = (1,0) \in \mathbf{Z}^2$ be the first vector of the standard basis[2]) of $\mathbf{Z}^2$, then the map

$$\theta\colon \begin{array}{ccc} K & \to & \mathbf{Q}^2 \\ \lambda & \mapsto & e_1.\iota(\lambda) \end{array}$$

is an isomorphism of $\mathbf{Q}$-vector spaces; next define the lattice $I = \theta^{-1}(\mathbf{Z}^2)$ in K which is invariant under multiplication by $\mathcal{O}_d$. In other words, I is an $\mathcal{O}_d$-ideal and I being proper is equivalent to ι being optimal.

2.2.3 THE PICARD GROUP OF THE ORDER $\mathcal{O}_d$. We now recall the definition and basic properties of the Picard group for an order $\mathcal{O}_d$ in a quadratic field.

The product of two $\mathcal{O}_d$-ideals I and J gives another $\mathcal{O}_d$-ideal

$$I \cdot J = \{\lambda\lambda' : \lambda \in I, \lambda' \in J\};$$

and clearly this operation respects the equivalence relation introduced above on $\mathcal{O}_d$-ideals. An $\mathcal{O}_d$-ideal I is *invertible* if there is some $\mathcal{O}_d$-ideal J so that $I \cdot J = \mathcal{O}_d$. An $\mathcal{O}_d$-ideal I is *locally principal* if for any prime p,

$$I_p := I \otimes_{\mathbf{Z}} \mathbf{Z}_p = \lambda_p(\mathcal{O}_d)_p\,,$$

where $(\mathcal{O}_d)_p = \mathcal{O}_d \otimes_{\mathbf{Z}} \mathbf{Z}_p$ and λ_p is an element of $(K \otimes_{\mathbf{Q}} \mathbf{Q}_p)^\times$. Both properties depend only on the ideal class $[I]$ and not on I itself.

For general orders $\mathcal{O}$ in number fields and $\mathcal{O}$-ideals I, one has the following implications:

$$I \text{ is locally principal} \Longrightarrow I \text{ is invertible} \Longrightarrow I \text{ is proper}.$$

We shall make use of the following property of orders in quadratic number fields:

PROPOSITION 2.1. *For the orders $\mathcal{O}_d$ in quadratic number fields the inverse implication*

$$I \text{ is proper} \Longrightarrow I \text{ is locally principal}$$

holds for $\mathcal{O}_d$-ideals I. In particular, the set of proper ideal classes $\mathrm{Cl}(\mathcal{O}_d)$, *endowed with the composition law induced by forming the product of two lattices, has the structure of an abelian group.*

---

[2]) We could have chosen any primitive vector in $\mathbf{Z}^2$.

This nice special feature of quadratic orders comes from the fact that in the quadratic case, orders are always *monogenic* (i.e. of the form $\mathcal{O} = \mathbf{Z}[x]$).

Proof. Recall that $\mathcal{O}_d = \mathbf{Z}[x]$ for $x = \frac{d+\sqrt{d}}{2}$. Assume now that I is a proper $\mathcal{O}_d$-ideal and consider the 2-dimensional $\mathbf{F}_p$-vector space $I_p/pI_p \simeq I/pI$. The natural map

$$(\mathcal{O}_d)_p/p(\mathcal{O}_d)_p \mapsto \mathrm{End}_{\mathbf{F}_p}(I_p/pI_p)$$

is injective. To see this, suppose that $\lambda \in (\mathcal{O}_d)_p$ acts trivially on I_p/pI_p. Then $\lambda I_p \subset pI_p$ and $\frac{\lambda}{p} I_p \subset I_p$ and so $\frac{\lambda}{p} \in \mathcal{O}_p$ as required. It follows that $\bar{x}$ the image of x in $\mathrm{End}_{\mathbf{F}_p}(I_p/pI_p)$ has a minimal polynomial of degree 2 and that I_p/pI_p is a cyclic $\mathbf{F}_p[\bar{x}]$-module. So there exist $\lambda_p \in I_p$ such that $I_p = \lambda_p(\mathcal{O}_d)_p + pI_p$ which implies that

$$\begin{aligned} I_p &= \lambda_p(\mathcal{O}_d)_p + p(\lambda_p(\mathcal{O}_d)_p + pI_p) = \\ &= \lambda_p(\mathcal{O}_d)_p + p^2 I_p = \lambda_p(\mathcal{O}_d)_p + p^3 I_p = \ldots = \lambda_p(\mathcal{O}_d)_p . \end{aligned}$$ □

2.3 Interpretation in terms of lattices

Let us verify that the various descriptions of $\mathcal{G}_d$ are equivalent:
Given $(a,b,c) \in R_{\mathrm{disc}}(d)$, put

$$h_{a,b,c} = \begin{pmatrix} b+\sqrt{d} & b-\sqrt{d} \\ 2c & 2c \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in \mathrm{SL}_2(\mathbf{Z}) .$$

Then $wh_{a,b,c}$ maps $\{\infty, 0\}$ to $\frac{-b \pm \sqrt{d}}{2a}$. Therefore, the geodesic $\gamma_{[a,b,c]}$ on $\mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}$ associated to (a,b,c) after Equation (1.3) is:

$$\gamma_{[a,b,c]} = wh_{a,b,c}.(0,\infty) ,$$

where $(0,\infty)$ is the geodesic on $\mathbf{H}$ joining 0 and ∞. Now $(0,\infty)$ corresponds, in the realization $T^1(\mathbf{H})$, to the A-orbit of the identity in $\mathrm{SL}_2(\mathbf{R})$; therefore $\gamma_{[a,b,c]}$ corresponds to $\mathrm{SL}_2(\mathbf{Z}) \cdot wh_{a,b,c}A = \mathrm{SL}_2(\mathbf{Z}) \cdot h_{a,b,c}A$, or equivalently the lattices of the form $\mathbf{Z}^2 \cdot h_{a,b,c}a_t \subset \mathcal{L}_2^{(1)}$ $(a_t \in A)$. Now one calculates

$$\frac{1}{\det(h_{a,b,c})} h_{a,b,c} \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} {}^t h_{a,b,c} = \frac{1}{\sqrt{d}} \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix} ,$$

which shows that in a particular basis of $\mathbf{Z}^2 h_{a,b,c}$ the quadratic form $q_0(x,y) = xy$ takes the shape as in (2.4) below.

Since A is the stabilizer subgroup of q_0, we have verified that $\gamma_{[a,b,c]}$ corresponds to:

The set of homothety classes of lattices L, such that the restriction of the quadratic form $q_0(x,y) = xy$ to L, expressed in terms of a basis α, β of L, takes the form

$$q_0(u\alpha + v\beta) = \mathrm{vol}(L)\frac{au^2 + buv + cv^2}{d^{1/2}}. \tag{2.4}$$

Note that the particular quadratic form $\frac{au^2+buv+cv^2}{\sqrt{d}}$ is not canonically attached to the lattice L because of the different choices of a basis.

Set $m_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and ι_0 to be the embedding $\iota_0 \colon K \hookrightarrow \mathrm{Diag}_2(\mathbf{R}) \subset M_2(\mathbf{R})$ obtained by mapping $\sqrt{d}$ to $d^{1/2}m_0$ and θ_0 be the linear embedding $\theta_0 \colon K \hookrightarrow \mathbf{R}^2$ given by

$$\theta_0(\lambda) = (1,1)\iota_0(\lambda), \quad \text{i.e.} \quad \theta_0(u + v\sqrt{d}) = (u + v|d|^{1/2}, u - v|d|^{1/2}).$$

Now let us verify, as asserted in §2.1, that the A-orbit of $\theta_0(I)$ belongs to $\mathcal{G}_d$, for any proper $\mathcal{O}_d$-ideal I. (We do not verify the more precise assertion that this is exactly the element of $\mathcal{G}_d$ that corresponds to the class of I under the bijection $\mathrm{Cl}(\mathcal{O}_d) \leftrightarrow [\mathrm{R}_{\mathrm{disc}}(d)]$.) We need to verify (according to (2.4)) that $\lambda \in I \mapsto \frac{q_0(\theta_0(\lambda))}{\mathrm{vol}(\theta_0(I))}\sqrt{d}$ is a quadratic form of discriminant d. But $q_0(\theta_0(\lambda)) = \mathbf{N}_{K/\mathbf{Q}}(\lambda)$ is the norm; and for any ideal $I \subset K$ we have $\mathrm{vol}(\theta_0(I)) = |d|^{1/2}\mathbf{N}(I)$. Here we have defined the *norm* $\mathbf{N}(I)$ of an ideal (relative to $\mathcal{O}_d$) by the ratio of indexes

$$\mathbf{N}(I) = \frac{(\mathcal{O}_d : \mathcal{O}_d \cap I)}{(I : \mathcal{O}_d \cap I)}.$$

Now, for any ideal I, the map $x \in I \mapsto \frac{\mathbf{N}_{K/\mathbf{Q}}(x)}{\mathbf{N}(I)}$ is easily verified to be an integer quadratic form of discriminant d, as desired.

2.4 A DUALITY PRINCIPLE

Our goal now is to show that the equidistribution statements of Skubenko's theorem and of Duke's theorem are equivalent.

The discussion which follows is valid in great generality; but we will consider only $G = \mathrm{PGL}_2(\mathbf{R})$, $\Gamma = \mathrm{PGL}_2(\mathbf{Z})$, and the diagonal torus A in G.

Since $\mathrm{PGL}_2(\mathbf{R})$ is identified with $\mathrm{SO}_{\mathrm{disc}}(\mathbf{R})$, it acts transitively on $V_{\mathrm{disc},+1}(\mathbf{R})$ (by Witt's theorem) and equals the $\mathrm{PGL}_2(\mathbf{R})$-orbit of (say)

$q_0(x,y) = xy$; equivalently $V_{\mathrm{disc},+1}(\mathbf{R})$ is identified with the $\mathrm{PGL}_2(\mathbf{R})$-conjugacy class of the matrix m_0 which has A as its stabilizer subgroup in G. Hence

$$V_{\mathrm{disc},+1}(\mathbf{R}) = \mathrm{PGL}_2(\mathbf{R}).q_0 \simeq \mathrm{PGL}_2(\mathbf{R}).m_0 \simeq \mathrm{PGL}_2(\mathbf{R})/A\,.$$

2.4.1 DUALITY BETWEEN ORBITS. It follows from the previous discussion that each representation $(a,b,c) \in \mathrm{R}_{\mathrm{disc}}(d)$ is identified with some class $g_{a,b,c}A/A \in G/A$ or what is the same to an orbit $g_{a,b,c}A \subset G$ for some $g_{a,b,c} \in G$ such that

$$g_{a,b,c}.q_0 = |d|^{-1/2}(a,b,c)\,, \quad q_0 = (0,1,0)\,.$$

As we have seen Γ acts on $\mathrm{R}_{\mathrm{disc}}(d)$ and the latter decomposes into a finite disjoint union of Γ-orbits, setting

$$[a,b,c] = \Gamma\backslash\Gamma(a,b,c) \in [\mathrm{R}_{\mathrm{disc}}(d)]\,,$$

for the orbit of (a,b,c), one has

$$\mathrm{R}_{\mathrm{disc}}(d) = \bigsqcup_{[a,b,c]\in[\mathrm{R}_{\mathrm{disc}}(d)]} \Gamma.(a,b,c)\,.$$

Hence $|d|^{-1/2}.\mathrm{R}_{\mathrm{disc}}(d)$ is identified with the collection of Γ-orbits

$$\bigsqcup_{[a,b,c]\in[\mathrm{R}_{\mathrm{disc}}(d)]} \Gamma g_{a,b,c}A/A \subset G/A\,;$$

thus the problem of the distribution of $|d|^{-1/2}.\mathrm{R}_{\mathrm{disc}}(d)$ inside $V_{\mathrm{disc},+1}(\mathbf{R})$ is a problem about the distribution of a collection of Γ-orbits inside the quotient space G/A.

There is an almost tautological equivalence between (left) Γ-orbits on G/A and (right) A-orbits on $\Gamma\backslash G$ given by

$$\Gamma gA/A \longleftrightarrow \Gamma gA \longleftrightarrow \Gamma\backslash\Gamma gA\,. \tag{2.5}$$

This duality induces a close relationship between the study of the distribution of $|d|^{-1/2}.\mathrm{R}_{\mathrm{disc}}(d)$ inside $V_{\mathrm{disc},+1}(\mathbf{R})$ and the distribution of the collection of right A-orbits

$$\mathcal{G}_d = \bigcup_{[a,b,c]\in[\mathrm{R}_{\mathrm{disc}}(d)]} x_{[a,b,c]}A \subset \Gamma\backslash G$$

inside the homogeneous space $\Gamma\backslash G$, with

$$x_{[a,b,c]} = \Gamma\backslash\Gamma g_{a,b,c}\,. \tag{2.6}$$

This is the "duality principle" alluded to at the beginning of this section. Let us make this principle a bit more precise by identifying the orbits in question:

Assuming that $(a,b,c) \in \mathrm{R}_{\mathrm{disc}}(d)$ is primitive; one has

$$x_{[a,b,c]}A = \Gamma\backslash\Gamma g_{a,b,c}A = \Gamma\backslash\Gamma A_{a,b,c}g_{a,b,c}\,,$$

where

$$A_{a,b,c} = g_{a,b,c}Hg_{a,b,c}^{-1} = \mathrm{stab}_{(a,b,c)}(G)$$

is the stabilizer of (a,b,c) in G. That group is the group of real points of a $\mathbf{Q}$-algebraic group, which we will denote by $\mathbf{T}_{a,b,c}$, namely the image in PGL_2 of the centralizer Z_m of

$$m = m_{a,b,c} = \begin{pmatrix} b & 2c \\ -2a & -b \end{pmatrix}.$$

In terms of the embedding $\iota = \iota_{m_{a,b,c}} \colon K \hookrightarrow M_2(\mathbf{Q})$, one has

$$Z_m(\mathbf{Q}) = \iota(K^\times),$$

and

$$\mathbf{T}(\mathbf{Q}) = \iota(K^\times)/\mathbf{Q}^\times\,\mathrm{Id}\,, \qquad A_{a,b,c} = \mathbf{T}_{a,b,c}(\mathbf{R}) = \iota(K\otimes\mathbf{R})^\times/\mathbf{R}^\times\,\mathrm{Id}\,,$$

and (since $M_2(\mathbf{Z}) \cap \iota(K) = \iota(\mathcal{O}_d)$),

$$\Gamma_{a,b,c} := \Gamma \cap A_{a,b,c} = \iota(\mathcal{O}_d^\times)/\{\pm\mathrm{Id}\}.$$

Alternatively, let ι_0 denote the (real) embedding

$$\iota_0 \colon \begin{array}{ccc} K & \hookrightarrow & M_2(\mathbf{R}) \\ u + v\sqrt{d} & \mapsto & u\,\mathrm{Id} + v.d^{1/2}m_0 \end{array}$$

obtained by conjugating ι_m with $g_{a,b,c}^{-1}$, we have

$$\iota_0(K\otimes_{\mathbf{Q}}\mathbf{R})^\times/\mathbf{R}^\times\,\mathrm{Id} = A$$

and

$$\Gamma'_{a,b,c} := g_{a,b,c}^{-1}\Gamma g_{a,b,c} \cap A = \iota_0(\mathcal{O}_d^\times)/\{\pm\mathrm{Id}\}$$

so that we have homeomorphisms

$$x_{[a,b,c]}A = \Gamma\backslash g_{a,b,c}A \simeq g_{a,b,c}^{-1}\Gamma g_{a,b,c} \cap A\backslash A = \iota_0(K\otimes\mathbf{R})^\times/\mathbf{R}^\times\iota_0(\mathcal{O}_d^\times)\,. \tag{2.7}$$

By Dirichlet's unit theorem, $\iota_0(K\otimes\mathbf{R})^\times/\mathbf{R}^\times\iota_0(\mathcal{O}_d^\times)$ is compact hence $x_{[a,b,c]}A$ is compact and since $[\mathrm{R}_{\mathrm{disc}}(d)]$ is finite we obtain:

THEOREM 2.2. *The union of A-orbits $\mathcal{G}_d$ is compact.*

2.4.2 DUALITY BETWEEN MEASURES. To consider equidistribution problems, one needs to refine the correspondence (2.5) at the level of measures. Roughly speaking, the choice of the counting measure μ_Γ on Γ and of the left-invariant Haar measure μ_A on[3]) A define a measure-theoretic version of the correspondence (2.5):

FACT. There exist homeomorphisms between the following spaces of Radon measures (relative to the weak-* topology):

$$\begin{array}{ccccc} \text{left } \Gamma\text{-invariant} & & \text{left } \Gamma\text{, right } A\text{-invariant} & & \text{right } A\text{-invariant} \\ \text{Radon measures} & \longleftrightarrow & \text{Radon measures} & \longleftrightarrow & \text{Radon measures} \\ \lambda \text{ on } G/A & & \rho \text{ on } G & & \nu \text{ on } \Gamma\backslash G\,. \end{array} \tag{2.8}$$

These homeomorphisms are characterized by the identities: for any $\varphi \in \mathcal{C}_c(G)$, one has

$$\lambda(\varphi_A) = \rho(\varphi) = \nu(\varphi_\Gamma)\,,$$

where

$$\varphi_A(g) := \int_A \varphi(gh) d\mu_A(h)\,, \quad \varphi_\Gamma(g) = \sum_{\gamma\in\Gamma} \varphi(\gamma.g)\,.$$

See for instance [2, §8.1] for a proof of that fact. We work out this correspondence in specific cases:

- ρ is a Haar measure μ_G on G, which is G-biinvariant as G is unimodular. The correspondence (2.8) yield the quotient measures $\nu = \mu_{\Gamma\backslash G}$ on $\Gamma\backslash G$, and $\lambda = \mu_{G/A} \propto \mu_{\mathrm{disc},\pm 1}$ on G/A. The former measure ν is finite (i.e. Γ is a lattice in G) and we may adjust μ_G so that $\mu_{\Gamma\backslash G}$ is a probability measure.
- The sum λ_d of Dirac measures on G/A given by

$$\begin{aligned} \lambda_d &= \sum_{(a,b,c)\in \mathrm{R}_{\mathrm{disc}}(d)} \delta_{g_{a,b,c}A/A} = \sum_{[a,b,c]} \sum_{g\in\Gamma.g_{a,b,c}} \delta_{gA/A} \\ &= \sum_{[a,b,c]} \sum_{\gamma\in\Gamma/\Gamma_{a,b,c}} \delta_{\gamma g_{a,b,c}A/A}\,. \end{aligned}$$

PROPOSITION. *The measure ν_d on $\Gamma\backslash G$ corresponding to λ_d under* (2.8) *is the sum of the push forwards of the Haar measure μ_A over the set of A-orbits $x_{[a,b,c]}A$, $[a,b,c] \in [\mathrm{R}_{\mathrm{disc}}(d)]$.*

---

[3]) Note that A is unimodular.

Indeed, set $\lambda_{[a,b,c]} = \sum_{\gamma\in\Gamma/\Gamma_{a,b,c}} \delta_{\gamma g_{a,b,c}A/A}$. Then if S denotes a fundamental domain in A for $\Gamma'_{a,b,c}$

$$\begin{aligned}\lambda_{[a,b,c]}(\varphi_A) &= \sum_{\gamma\in\Gamma/\Gamma_{a,b,c}} \int_A \varphi(\gamma g_{a,b,c}h)dh = \sum_{\gamma\in\Gamma}\int_S \varphi(\gamma g_{a,b,c}h)dh \\ &= \int_{\Gamma'_{a,b,c}\backslash A} \varphi_\Gamma(g_{a,b,c}h)dh = \int_{x_{[a,b,c]}A} \varphi_\Gamma(h)dh\,,\end{aligned}$$

hence the measure on $\Gamma\backslash G$ corresponding to $\lambda_{[a,b,c]}$ is given by the push forward of the Haar measure μ_A to the periodic A-orbit $x_{[a,b,c]}A$, and the proposition follows.

Let

$$\mathrm{vol}(\mathcal{G}_d) := \nu_d(\mathcal{G}_d) = \sum_{[a,b,c]} \mathrm{vol}(x_{[a,b,c]}A)$$

denote the total volume of this (finite) collection of (compact) A-orbits. From (2.7) we see that the various orbits associated to primitive representations of d have the same volume, namely with the correct normalization of the Haar measure of A

$$\mathrm{vol}(x_{[a,b,c]}A) = \mathrm{vol}(\mathbf{R}^\times \iota_0(\mathcal{O}_d^\times)\backslash A) = \mathrm{Reg}(\mathcal{O}_d)\,,$$

where $\mathrm{Reg}(\mathcal{O}_d)$ is the *regulator* of $\mathcal{O}_d$. Therefore,

$$\mathrm{vol}(\mathcal{G}_d) = |\operatorname{Pic}(\mathcal{O}_d)|\ \mathrm{Reg}(\mathcal{O}_d)\,.$$

If $d = \mathrm{disc}(\mathcal{O}_K)$ is a fundamental discriminant, the *Dirichlet class number formula* gives

$$\mathrm{vol}(\mathcal{G}_d) = |\operatorname{Pic}(\mathcal{O}_d)|\ \mathrm{Reg}(\mathcal{O}_d) = \lambda|d|^{1/2}L\big(\big(\tfrac{d}{\cdot}\big),1\big)\,,$$

where λ is some absolute constant, $(\frac{d}{\cdot})$ is the *Kronecker symbol* and $L((\frac{d}{\cdot}),s)$ its associated L-function. Then by Siegel's theorem $L((\frac{d}{\cdot}),1) = |d|^{o(1)}$ as $d\to\infty$ so that

$$\mathrm{vol}(\mathcal{G}_d) = |d|^{1/2+o(1)}\,. \tag{2.9}$$

If $d = d'f^2$ with d' a fundamental discriminant

$$\frac{|\operatorname{Pic}(\mathcal{O}_d)|\ \mathrm{Reg}(\mathcal{O}_d)}{|\operatorname{Pic}(\mathcal{O}_{d'})|\ \mathrm{Reg}(\mathcal{O}_{d'})} = f\prod_{p|f}\left(1-p^{-1}\left(\frac{d'}{p}\right)\right)$$

which shows again that $|\operatorname{Pic}(\mathcal{O}_d)|\,\mathrm{Reg}(\mathcal{O}_d) = |d|^{1/2+o(1)}$ and hence (2.9) holds in general (cf. e.g. [10, Sect. 9.6]). We let

$$\mu_d := \frac{1}{\mathrm{vol}(\mathcal{G}_d)}\nu_d\,.$$

This is an A-invariant probability measure on $\Gamma\backslash G$ and the above discussion shows that Skubenko's Theorem 1.2 follows from the following:

THEOREM 2.3. *As $d \to \infty$ amongst the non-square discriminants, the sequence of measures μ_d weak-* converge to the probability measure $\mu_{\Gamma\backslash G}$, i.e. for any $\varphi_\Gamma \in \mathcal{C}_c(\Gamma\backslash G)$, one has*

$$\mu_d(\varphi_\Gamma) = \frac{1}{\mathrm{vol}(\mathcal{G}_d)} \sum_{[a,b,c]} \int_{x_{[a,b,c]}A} \varphi_\Gamma(h)dh \to \mu_{\Gamma\backslash G}(\varphi_\Gamma)\,.$$

Indeed any continuous compactly supported function on G/A is of the form φ_A for $\varphi \in \mathcal{C}_c(G)$, hence by Theorem 2.3

$$\begin{aligned}\lambda_d(\varphi_A) = \nu_d(\varphi_\Gamma) &= \mathrm{vol}(\mathcal{G}_d)\mu_d(\varphi_\Gamma) \\ &= \mathrm{vol}(\mathcal{G}_d)(\mu_{\Gamma\backslash G}(\varphi_\Gamma) + o(1)) = \mathrm{vol}(\mathcal{G}_d)(\mu_{G/A}(\varphi_A) + o(1))\,.\end{aligned}$$

3. SPACING PROPERTIES OF TORUS ORBITS

In this section, we show that the various distinct orbits $x_{[a,b,c]}A \subset \mathcal{G}_d$ are in a suitable sense *well spaced* from each other; the main result is Proposition 3.6. Recall that

$$\mathcal{G}_d = \bigsqcup_{[a,b,c]\in[\mathrm{R}_{\mathrm{disc}}(d)]} x_{[a,b,c]}A\,,$$

where $x_{[a,b,c]}$ is defined in (2.6).

3.1 IDEAL CLASSES ARE CONTROLLING THE TIME SPENT NEAR THE CUSP

The space X is not compact and this is measured through a *height function* (normalized to be invariant under scaling) given, for $L = \mathbf{Z}^2.g \subset \mathbf{R}^2$ a lattice, by

$$\mathrm{ht}(L) = \Big(\frac{\min_{x\in L-\{0\}} \|x\|}{\mathrm{vol}(L)^{1/2}}\Big)^{-1} = \Big(\frac{\min_{x\in\mathbf{Z}^2-\{0\}} \|xg\|}{|\det(g)|^{1/2}}\Big)^{-1},$$

where $\|.\|$ denote the Euclidean norm. This continuous function is proper. Indeed, if $x \in X$ and $(z,v) \in \mathcal{S}$ any representative, then the height $\mathrm{ht}(x)$ and the imaginary part $\Im(z)$ satisfy $\Im(z) = \mathrm{ht}(x)^2$. For any $H > 1$ let $X_{\geq H}$ denote the set of all $x \in X$ with $\mathrm{ht}(x) \geq H$.

In this section we evaluate explicitly how big the height of a lattice in $\mathcal{G}_d$ could be.

PROPOSITION 3.1. *Suppose the proper integral ideal $J \subset \mathcal{O}_d$ corresponds to $[a,b,c] \in \mathrm{R}_{\mathrm{disc}}(d)$ under the bijection of §2.1. Then $x_{[a,b,c]}A \cap X_{\geq H}$ is nonempty if and only if J^{-1} is equivalent to an ideal $I \subset \mathcal{O}_d$ of norm $\leq \frac{1}{2}H^{-2}d^{1/2}$. Moreover, this defines a bijection between connected components of $\mathcal{G}_d \cap X_{\geq H}$ and proper $\mathcal{O}_d$-ideals $I \subset \mathcal{O}_d$ of norm $\leq \frac{1}{2}H^{-2}d^{1/2}$.*

Even though the above does not control escape of mass for μ_d as $d \to \infty$ it does give an upper bound for $\mu_d(X_{\geq H})$, see Proposition 3.3, which we will use in our proof of Duke's theorem. Note that Proposition 2.1 guarantees that there is an inverse J^{-1} to the proper ideal J.

REMARK 3.2. Applying this result to $H = d^{1/4}$ we see that $\mathcal{G}_d \cap X_{\geq d^{1/4}}$ is empty (as there are no ideals of norm < 1). This implies that $\mathcal{G}_d$ is pre-compact.

Proof. Note that, if we identify $x \in X$ with a lattice L of covolume 1, then $xA \cap X_{\geq H}$ is nonempty if and only if there is some nonzero vector $(u,v) \in L$ with $|uv| \leq \frac{1}{2}H^{-2}$.

Therefore (using the explicit bijection of §2.1) the A-orbit defined by J intersects $X_{\geq H}$, if and only if J contains an element λ with

$$|\mathbf{N}(\lambda)| \leq \frac{1}{2}H^{-2}\mathbf{N}(J)d^{\frac{1}{2}}.$$

Recall that $\mathbf{N}(J^{-1}) = \mathbf{N}(J)^{-1}$ by standard properties of the norm. It follows that the A-orbit defined by J intersects $X_{\geq H}$ if and only if $\mathbf{N}(\lambda J^{-1}) \leq \frac{1}{2}H^{-2}d^{\frac{1}{2}}$ for some $\lambda \in J$ (so that $\lambda J^{-1} \subset \mathcal{O}_d$).

Finally, notice that for $H > 1$ there is, in a lattice $L' \in X_{\geq H}$, up to sign, only one primitive nonzero vector of length $\leq H^{-1}\operatorname{vol}(L')^{1/2}$ (which is a simple volume computation). Therefore, fixing J, in the above argument, a connected component of $\theta_0(J).A \cap X_{\geq H}$ corresponds to a unique primitive element $\lambda \in J$ with $|\mathbf{N}(\lambda)| \leq \frac{1}{2}H^{-2}\mathbf{N}(J)d^{\frac{1}{2}}$ (up to sign) and we can associate to this connected component the ideal $I = \lambda J^{-1} \subset \mathcal{O}_d$ of norm $\leq \frac{1}{2}H^{-2}d^{\frac{1}{2}}$. □

PROPOSITION 3.3. *There is "not too much mass high in the cusp" in the sense that*

$$\mu_d(X_{\geq H}) \ll_\varepsilon d^\varepsilon H^{-2}$$

for all $\varepsilon > 0$ and $H \geq 1$.

Note that to make this estimate useful, we will set later $H = d^\varepsilon$ for some $\varepsilon > 0$.

Proof. We note first that in any orbit in $\mathcal{G}_d$ the maximal height achieved is $\leq d^{\frac{1}{4}}$ (see Remark 3.2). This implies that for $H > 1$ any connected component of $\mathcal{G}_d \cap X_{\geq H}$ has length $\ll \log(d)$. Indeed such a component corresponds (in the upper half-plane model) to the segment of some oriented geodesic circle (i.e. a half-circle centered on the real line) made of those points which have imaginary part between H and $d^{1/4}$: the hyperbolic length of such a segment is bounded by $\ll \log(d^{\frac{1}{4}}/H)$.

Therefore, by Proposition 3.1

$$\mathrm{vol}(\mathcal{G}_d \cap X_{\geq H}) \ll \log(d) N_{\leq H}(d)\,,$$

where $N_{\leq H}(d)$ is the number of proper ideals $I \subset \mathcal{O}_d$ of norm $\mathbf{N}(I) \leq \frac{1}{2}H^{-2}d^{\frac{1}{2}}$. Recall that for any $n \in \mathbf{N}$ the number of proper ideals in $\mathcal{O}_d$ of norm equal to n is bounded by the number of divisors of n and so by $\ll_\varepsilon n^\varepsilon$. By summing over all $1 \leq n \leq \frac{1}{2}H^{-2}d^{\frac{1}{2}}$ we get that $N_{\leq H}(d) \ll_\varepsilon (H^{-2}d^{\frac{1}{2}})^{1+\varepsilon}$. Together with (2.9) this proves the proposition. □

3.2 Linnik's Basic Lemma and representing binary quadratic forms by ternary forms

Following Linnik we will derive the "Basic Lemma" from representation numbers of quadratic forms: Let q, Q be two integral non-degenerate quadratic forms on $\mathbf{Z}^m$ and $\mathbf{Z}^n$ respectively. Assuming that $m \leq n$, *a representation* of q by Q is an isometric embedding of quadratic lattices

$$\iota\colon (\mathbf{Z}^m, q) \hookrightarrow (\mathbf{Z}^n, Q)$$

in other terms a $\mathbf{Z}$-linear map $\iota\colon \mathbf{Z}^m \to \mathbf{Z}^n$ such that for $\mathbf{x} \in \mathbf{Z}^m$

$$Q(\iota(\mathbf{x})) = q(\mathbf{x})\,.$$

For instance a representation $\mathbf{x} \in \mathbf{Z}^n$ of an integer $d \in \mathbf{Z}$ by a quadratic form Q on $\mathbf{Z}^n$ may be viewed as the isometric embedding

$$\iota_{\mathbf{x}}\colon \begin{array}{ccc} (\mathbf{Z}, dx^2) & \hookrightarrow & (\mathbf{Z}^n, Q) \\ n & \mapsto & n\mathbf{x} \end{array}\,.$$

Let $\mathrm{R}_Q(q)$ be the set of such representations: the group $\Gamma = \mathrm{SO}_Q(\mathbf{Z})$ acts on $\mathrm{R}_Q(q)$ (for $\gamma \in \Gamma$, $\gamma.\iota = \gamma \circ \iota$) and the quotient $\Gamma\backslash\mathrm{R}_Q(q)$ is finite.

We are interested here in evaluating $|\Gamma\backslash\mathrm{R}_Q(q)|$ in the codimension one case (i.e. when $n - m = 1$). More precisely, we will need to show that, in this case, $|\Gamma\backslash\mathrm{R}_Q(q)|$ is rather small. The simplest evidence comes from the case $m = 1$, $n = 2$: the representations of an integer by a binary quadratic form.

For instance it is well known that for $d \neq 0$ the number of integral solutions to $xy = d$ (i.e. the number of divisors of d) is bounded by $O_\varepsilon(d^\varepsilon)$. Similarly the number of representations of an integer as a sum of two squares satisfies the same bound; indeed, for any binary integral quadratic form Q one has $|\Gamma\backslash \mathrm{R}_Q(d)| \ll_q |d|^\varepsilon$ for any $\varepsilon > 0$. The following is a version of this claim for $m = 2$, $n = 3$, where in the case of non-fundamental discriminants the estimate is not as strong.

PROPOSITION 3.4. *Let Q be an integral ternary quadratic form, and let*

$$q(x, y) = ax^2 + bxy + cy^2$$

an integral binary quadratic form, both supposed non-degenerate. Assume that $f^2 | \gcd(a, b, c)$ is the greatest common square divisor of a, b, c. Then the number N of embeddings of $(\mathbf{Z}^2, q)$ into $(\mathbf{Z}^3, Q)$, modulo the action of $\mathrm{SO}_Q(\mathbf{Z})$, is $\ll_{Q,\varepsilon} f \max(|a|, |b|, |c|)^\varepsilon$.

When $Q = x^2 + y^2 + z^2$ is the "sum of three squares" quadratic form such a bound is a consequence of an explicit formula on the number of representations due to Venkov [25] (assuming a square-free). This bound was later generalized by Pall [21, Thm. 5]. We provide a self-contained treatment in Appendix A. Let

$$\begin{aligned}\langle (a, b, c), (a', b', c') \rangle_{\mathrm{disc}} &= \mathrm{disc}(a + a', b + b', c + c') - \mathrm{disc}(a, b, c) - \mathrm{disc}(a', b', c') \\ &= 2bb' - 4ac' - 4a'c\end{aligned}$$

be the *polarization inner product* associated with the quadratic form disc. We will apply Proposition 3.4 to the pair

$$Q = \mathrm{disc}\,, \qquad q(x, y) = dx^2 + \ell xy + dy^2\,,$$

and note that $q(x, y)$ is non-degenerate if and only if $\ell \neq \pm 2d$. Hence we obtain:

COROLLARY 3.5. *Let $\Gamma = \mathrm{SO}_{\mathrm{disc}}(\mathbf{Z})$. Then for any two integers d, ℓ with $\ell \neq \pm 2d$, the number of Γ-orbits on pairs*

$$\big\{((a, b, c), (a', b', c')) \in \mathbf{Z}^3 \times \mathbf{Z}^3 : \\ \mathrm{disc}(a, b, c) = \mathrm{disc}(a', b', c') = d, \quad \langle (a, b, c), (a', b', c') \rangle_{\mathrm{disc}} = \ell \big\}$$

is $\ll_\varepsilon f(\max(|d|, |\ell|))^\varepsilon$, where f^2 is the largest square factor of $\gcd(d, \ell)$.

We now translate the information obtained about quadratic forms above to Linnik's Basic Lemma, which we phrase in the geometric context. This falls short from equidistribution but will suffice as the arithmetic input to the ergodic arguments later.

PROPOSITION 3.6 (Basic Lemma). *We have*

$$\mu_d \times \mu_d\{(x,y) \in X^2_{\leq H} : d_X(x,y) \leq \delta\} \ll_\varepsilon H^4\delta^3 d^\varepsilon$$

whenever $d^{-\frac{1}{4}} \leq \delta \leq \frac{1}{3}H^{-2}$ *and* $\varepsilon > 0$.

Note that the exponent 3 of δ^3 is optimal, and suggests that μ_d is 3-dimensional in the appropriate scale. The trivial exponent is 1, which follows from A-invariance of μ_d.

Proof. We start by indicating the relationship between δ-close tuples in $(\mathcal{G}_d \cap X_{\leq H})^2$ and the representation of the binary quadratic form $q(x,y) = dx^2 + \ell xy + dy^2$ by the discriminant ternary quadratic form disc.

From (1.4), $g_1, g_2 \in \mathrm{PSL}_2(\mathbf{R})$ are such that $x_i = \Gamma g_i \in \mathcal{G}_d \cap X_{\leq H}$ for $i = 1,2$ and $d_X(x_1,x_2) < \delta$, then we may assume

$$g_1 \in \mathcal{S}\,, \quad g_2 \in \mathcal{S}', \quad \Gamma g_1 \in X_{\leq H} \quad \text{and} \quad d(g_1,g_2) < \delta\,, \tag{3.1}$$

where $\mathcal{S}'$ is some slightly bigger set containing the fundamental domain $\mathcal{S}$ in its interior. For concreteness we take

$$\mathcal{S}' = \{(z,v) \in \mathbf{H} \times S^1 : \quad |\Re z| \leq 1, \quad \Im z \geq 1/2\}.$$

This clearly shows that the matrix entries of both g_i are *controlled*, i.e. $\|g_i\| \ll H$ where

$$\|g\| = \mathrm{tr}(g^t g)^{1/2}.$$

Moreover, we may associate to g_i the primitive integral quadratic form,

$$q_i(x,y) = \sqrt{d}[g_i.q_0](x,y) = a_i x^2 + b_i xy + c_i y^2, \quad b_i^2 - 4a_i c_i = d, \quad \gcd(a_i,b_i,c_i) = 1\,.$$

We have to consider two different possible cases. Either $q_1 = q_2$ (i.e. $g_2 \in g_1 A$) or $q_1 \neq q_2$.

The total mass for the first case is easy to estimate by $\ll_\epsilon d^{1/2+\epsilon}\delta$ before normalization by the total volume, which gives after the normalization that

$$\mu_d \times \mu_d\{(\Gamma g_1, \Gamma g_1 h) \in X^2_{\leq H} : h \in A, d(\mathrm{Id},h) \leq \delta\} \ll_\varepsilon \delta d^{-1/2} d^\varepsilon \leq \delta^3 d^\varepsilon$$

since $d^{-1/4} \leq \delta$.

Henceforth we assume $q_1 \neq q_2$. Since $\|g_i\| \ll H$, we have

$$\max(|a_i|, |b_i|, |c_i|) \ll d^{1/2} H^2. \tag{3.2}$$

Also by assumption $g_2 = g_1 h$ with $d(h, \mathrm{Id}) < \delta$. This shows that $q_2 = \sqrt{d} g_1.(h.q_0)$ where $\|h.q_0 - q_0\| \ll \delta$. Therefore,

$$\max(|a_1 - a_2|, |b_1 - b_2|, |c_1 - c_2|) \ll d^{1/2} H^2 \delta. \tag{3.3}$$

We now define

$$q(u, v) = \mathrm{disc}\big(u(a_1, b_1, c_1) + v(a_2, b_2, c_2)\big) = du^2 + \ell uv + dv^2.$$

From the bound (3.3) on the difference of the vectors we know

$$|q(1, -1)| = |2d - \ell| \ll dH^4 \delta^2.$$

In order to apply Corollary 3.5 on q, we need to check that q is not degenerate, i.e. that $\ell \neq \pm 2d$. Indeed, if $\ell = \pm 2d$ then

$$d(a_2 \mp a_1)^2 = q(a_2, -a_1) = \mathrm{disc}\big(a_2(a_1, b_1, c_1) - a_1(a_2, b_2, c_2)\big) = (a_2 b_1 - a_1 b_2)^2,$$

which contradicts the assumption that d is *not* a perfect square. Therefore $\ell \neq \pm 2d$. In this case we may apply Corollary 3.5 to obtain the bound

$$N_{\ell,d} = \big|\mathrm{SO}_{\mathrm{disc}}(\mathbf{Z}) \backslash \{(\mathbf{Z}^2, dx^2 + \ell xy + dy^2) \hookrightarrow (\mathbf{Z}^3, \mathrm{disc})\}\big| \ll f \max(|d|, |\ell|)^{\varepsilon}$$

on the number $N_{\ell,d}$ of inequivalent ways in which the quadratic form $dx^2 + \ell xy + dy^2$ can be represented, where $f^2 | \gcd(d, \ell)$ is the greatest square divisor. Note that the group $\mathrm{SO}_{\mathrm{disc}}$ is rationally equivalent to PGL_2, and so up to isogeny rationally equivalent to SL_2. Therefore, $\mathrm{SO}_{\mathrm{disc}}(\mathbf{Z})$ is commensurable to the image of $\Gamma = \mathrm{SL}_2(\mathbf{Z})$ and we may also use Γ instead of $\mathrm{SO}_{\mathrm{disc}}(\mathbf{Z})$ in the above estimate.

Let

$$\Gamma(q_1^{(1)}, q_2^{(1)}), \ldots, \Gamma(q_1^{(k)}, q_2^{(k)})$$

be a complete list of diagonal Γ-orbits of pairs of quadratic forms which can be written as

$$q_i^{(j)}(x, y) = \sqrt{d} g_i^{(j)}.q_0(x, y)$$

with $g_1^{(j)}, g_2^{(j)}$ satisfying (3.1)

The number k of these diagonal Γ-orbits of quadratic forms is bounded by

$$\begin{aligned} k &\leq \sum_{\ell=2d-L}^{2d+L} N_{\ell,d} = \sum_{f^2 | d} \sum_{\substack{|2d-\ell| \leq L \\ f^2 | \ell,\ \ell \neq \pm 2d}}{}' N_{\ell,d} \\ &\ll_{\varepsilon} \sum_{f^2 | d} \sum_{\substack{|2d-\ell| \leq L \\ f^2 | \ell,\ \ell \neq \pm 2d}}{}' f\, d^{\epsilon} \ll_{\epsilon} \sum_{f^2 | d} f\, \frac{d^{1+\epsilon} H^4 \delta^2}{f^2} \ll_{\varepsilon} d^{1+2\varepsilon} \delta^2 H^4, \end{aligned}$$

where $L \ll dH^4\delta^2$ and $\sum'$ denotes a sum over ℓ for which $\frac{(d,\ell)}{f^2}$ is square-free.

We claim that for $q_1^{(j)} \neq q_2^{(j)}$ we have

$$d(g_1^{(j)}a_t, g_2^{(j)}A) \gg d^{-1}. \tag{3.4}$$

Indeed suppose $d(g_1^{(j)}a_t, g_2^{(j)}a_{t'}) \leq cd^{-1}$ (for some constant c determined in a moment). Then we may find some $\gamma \in \Gamma$ with $\gamma g_1^{(j)}a_t \in \mathcal{S}$, which also implies $\gamma g_2^{(j)}a_{t'} \in \mathcal{S}'$. By Remark 3.2 we have $\mathcal{G}_d \subset X_{\leq H'}$ for $H' = d^{1/4}$. Hence by choosing c appropriately the upper bound in (3.3) (applied for $H' = d^{1/4}$ and $\delta = cd^{-1}$) is less than one, which gives a contradiction.

Writing $g_2 = g_1 \exp v$ for some $v = v^- + v^+ + v_A \in \mathfrak{sl}_2(\mathbf{R})$, with v^-, v^+, v_A eigenvectors of Ad_{a_t} with eigenvalues $e^{-t}, e^t, 1$ respectively, the estimate (3.4) implies that both $\|v^-\|, \|v^+\| \gg d^{-1}$. It follows that for any j the inequality

$$d(g_1^{(j)}a_t, g_2^{(j)}A) < 1 \tag{3.5}$$

can hold only for t in some interval I_j of length $\ll \log d$.

CLAIM. For each pair $(g_1^{(j)}, g_2^{(j)})$ there is an interval $I_j \subset \mathbf{R}$ of length $\ll_\varepsilon d^\varepsilon$ with the following property:

If $(x_1, x_2) \in (\mathcal{G}_d \cap X_{\leq H})^2$ with $d(x_1, x_2) < \delta$ have representatives (g_1, g_2) satisfying (3.1) for which the associated forms $q_i = \sqrt{d}g_i.q_0$ are different, then $x_1 = \Gamma g_1^{(j)}a_t$ for some j and some $t \in I_j$.

Indeed, $(\gamma.q_1, \gamma.q_2) = (q_1^{(j)}, q_2^{(j)})$ for some $\gamma \in \Gamma$ and some $j \in [1,k]$ and so $g_1 = \gamma^{-1}g_i^{(j)}a_t$ resp. $g_2 \in \gamma^{-1}g_2^{(j)}A$. By assumption on g_1, g_2 we have $d(g_1^{(j)}a_t, g_2^{(j)}A) < \delta$.

Using the claim and a fixed Haar measure of A (i.e. before normalization) we get that the measure of the collection of points $(x_1, x_2) \in (\mathcal{G}_d \cap X_{\leq H})^2$, which can be represented as $x_i = \Gamma g_i$ with g_i as in (3.1) and for which the associated quadratic forms are different, is

$$\ll \sum_{j=1}^{k} |I_j|\delta \ll_\epsilon d^\varepsilon \delta k \ll_\varepsilon d^{1+2\varepsilon}H^4\delta^3.$$

Therefore, by dividing the above by the total volume of $(\mathcal{G}_d)^2$, the claim (together with the analysis of the case $q_1 = q_2$) implies the proposition.

4. An ergodic theoretic proof of Duke's theorem

4.1 Entropy and the unique measure of maximal entropy

A basic underlying concept in our proof is that of entropy. We recall that if $\mathcal{P}$ is a finite partition of the probability space (X,ν), the *entropy* of $\mathcal{P}$ is defined as

$$H_\nu(\mathcal{P}) := \sum_{S\in\mathcal{P}} -\nu(S)\log\nu(S)\,.$$

It is clear that $H_\nu(\mathcal{P}) = H_\nu(T^{-1}\mathcal{P})$ if $T\colon X\to X$ preserves ν — below we will use this fact without explicit reference. We note for future reference that entropy is controlled by an L^2-norm

$$H_\nu(\mathcal{P}) \geq -\log\left(\sum_{S\in\mathcal{P}}\nu(S)^2\right) \tag{4.1}$$

as one easily sees from convexity of the logarithm map. Moreover, entropy has the following basic *subadditivity property*: if $\mathcal{P}_1,\mathcal{P}_2$ are two partitions, then

$$H_\nu(\mathcal{P}_1\vee\mathcal{P}_2) \leq H_\nu(\mathcal{P}_1)+H_\nu(\mathcal{P}_2)\,, \tag{4.2}$$

where $\vee$ denotes common refinement.

If T is a measure-preserving transformation of (X,ν), then the *measure theoretic entropy* of T is defined as:

$$h_\nu(T) = \sup_{\mathcal{P}}\lim_{n\to\infty}\frac{H_\nu(\mathcal{P}\vee T^{-1}\mathcal{P}\vee\cdots\vee T^{-(n-1)}\mathcal{P})}{n}\,, \tag{4.3}$$

where the supremum is taken over all finite partitions of X. We also note that the limit in the definition exists and is equal to the infimum because the sequence

$$a_n = H_\nu(\mathcal{P}\vee T^{-1}\mathcal{P}\vee\cdots\vee T^{-(n-1)}\mathcal{P})$$

is subadditive (i.e. $a_{n+m}\leq a_n+a_m$).

A key role in our argument is played by the fact that the uniform measure on $\Gamma\backslash \mathrm{SL}_2(\mathbf{R})$ for any lattice Γ can be distinguished using entropy, as it is the *unique* measure of maximal entropy:

THEOREM 4.1. *Let $X=\Gamma\backslash\mathrm{SL}_2(\mathbf{R})$ be a quotient by a lattice $\Gamma<\mathrm{SL}_2(\mathbf{R})$, and let T denote the time-one-map of the geodesic flow, i.e. right translation*

$$T(x) = x\begin{pmatrix} e^{1/2} & 0\\ 0 & e^{-1/2}\end{pmatrix}.$$

Then for any invariant measure ν the entropy satisfies $h_\nu(T) \le 1$ where equality holds if and only if $\nu = \mu_X$ is the $\mathrm{SL}_2(\mathbf{R})$-invariant probability measure on X.

The inequality $h_\nu(T) \le 1$ is not hard and can be proved in many ways. Identifying the uniform measure as the unique measure where this maximum is attained is somewhat more delicate. We give a self-contained treatment in Appendix B.

4.2 Proof of Duke's theorem, an outline

Let $T\colon X \to X$ denote the time-one-map of the geodesic flow as in Theorem 4.1. Recall that

$$U^- = \left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbf{R} \right\} \quad \text{resp.} \quad U^+ = \left\{ \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} : t \in \mathbf{R} \right\}$$

are the *stable*, resp. *unstable* horocycle subgroups. The orbits of these two subgroups give the foliation into stable and unstable manifolds in the following sense. If $u = u(t) \in U^-$, then the distance between $T^n(x)$ and $T^n(xu)$ converges rapidly to zero:

$$\begin{aligned} d(T^n(x), T^n(xu)) &= d\left(x \begin{pmatrix} e^{n/2} & 0 \\ 0 & e^{-n/2} \end{pmatrix}, xu \begin{pmatrix} e^{n/2} & 0 \\ 0 & e^{-n/2} \end{pmatrix} \right) \\ &\le d\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} e^{-n/2} & 0 \\ 0 & e^{n/2} \end{pmatrix} u \begin{pmatrix} e^{n/2} & 0 \\ 0 & e^{-n/2} \end{pmatrix} \right) \\ &= d\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & e^{-n}t \\ 0 & 1 \end{pmatrix} \right). \end{aligned}$$

To give an outline of our argument, it is perhaps preferable to simplify the situation. In our proof, the *noncompact* nature of our space X is a significant complication, so instead of considering the quotient $\mathrm{SL}_2(\mathbf{Z})\backslash \mathrm{SL}_2(\mathbf{R})$ for the purposes of this outline let us consider a compact quotient $\widehat{X} = \Gamma\backslash \mathrm{SL}_2(\mathbf{R})$ on which we have a sequence of T-invariant probability measures μ_d satisfying the following simplified version of the conclusion of Proposition 3.6:

$$\mu_d \times \mu_d\{(x,y) \in \widehat{X}^2 : d_{\widehat{X}}(x,y) \le \delta\} \ll_\varepsilon \delta^3 d^\varepsilon \qquad \text{for } \delta > d^{-1/4}. \tag{4.4}$$

Let $r > 0$ be an *injectivity radius* of $\widehat{X}$ so that for any $x \in \widehat{X}$ the map $B_r^G(e) \to \widehat{X}$ sending g to xg is injective (with $G = \mathrm{SL}_2(\mathbf{R})$, and B_r^G denoting a ball of radius r in G). Also assume $\eta < \frac{1}{e} r$ is small enough so that $B_\eta^G(e)$ is an injective image under the exponential map of a neighborhood of 0 in the Lie algebra.

Let $\mathcal{P}$ be a finite measurable partition all of whose elements have "diameter smaller than η", i.e. if x and $y = xg$ with $g \in B_r^G$ belong to the same element of $\mathcal{P}$, then $g \in B_\eta^G$. Assume that the same holds as well for $T^i(x)$ and $T^i(y)$ for $i = -N, \dots, 0, 1, \dots, N$. Then $d(T(x), T(y)) < \eta$ and $d(e, a^{-1}ga) < r$ so that $a^{-1}ga \in B_\eta^G(e)$. Repeating, this implies that

$$g \in B_N = \bigcap_{n=-N}^{N} \begin{pmatrix} e^{1/2} & \\ & e^{-1/2} \end{pmatrix}^{-n} B_\eta^G(e) \begin{pmatrix} e^{1/2} & \\ & e^{-1/2} \end{pmatrix}^{n}.$$

We define a *Bowen N-ball* to be the translate xB_N for some $x \in X$.

Notice that the set B_N is "tube-like": it has width at most $e^{-N}\eta$ along the stable and unstable directions, but is of length η in the direction A of the geodesic flow. The above shows that every element of the partition

$$\mathcal{P}^{[-N,N]} = \bigvee_{n=-N}^{N} T^{-n}\mathcal{P} \tag{4.5}$$

is contained in a *single* Bowen N-ball. Together we conclude that

$$\bigcup_{S \in \mathcal{P}^{[-N,N]}} S \times S \subset \bigcup_{i=1}^{k} \{(x, ya_i) : d(x, y) < re^{-N}\},$$

where $k \ll e^N$ and $a_1, \dots, a_k \in B_r^A(1)$ are chosen to be *δ-dense* — that is to say, the union of the δ-neighborhoods around a_i cover $B_r^A(1)$.

Together with (4.4) this shows that

$$\sum_{S \in \mathcal{P}^{[-N,N]}} \mu_d(S)^2 \ll_\varepsilon e^{-2N} d^\varepsilon$$

whenever $\delta = \eta e^{-N} \geq d^{-\frac{1}{4}}$ or equivalently $N \leq \frac{1}{4}\log d + \log r$. We choose $N = \lfloor \frac{1}{5}\log d \rfloor$ (the "extra space" will be useful in suppressing a d^ε). Using (4.1) we have

$$H_{\mu_d}\left(\mathcal{P}^{[-N,N]}\right) \geq (2 - 6\varepsilon)N$$

for large enough d.

In this statement we cannot yet let $d \to \infty$ to get a statement about a weak* limit μ, because N is a function of d, and so the size of $\mathcal{P}^{[-N,N]}$ increases with d. Thus let $N_0 \geq 1$ be any fixed integer: $[-N, N]$ can be covered by $\lceil \frac{N}{N_0} \rceil$ many translates of $[-N_0, N_0]$. This in turn shows that $\mathcal{P}^{[-N,N]}$ can be obtained as a refinement of the $\lceil \frac{N}{N_0} \rceil$ partitions

$$\mathcal{P}^{[-N,-N+2N_0]}, \mathcal{P}^{[-N+2N_0,-N+4N_0]}, \dots$$

(in the obvious generalization of the notation (4.5)). By subadditivity (4.2) (and invariance) this implies

$$H_{\mu_d}\left(\mathcal{P}^{[-N_0,N_0]}\right) \geq (2-7\varepsilon)N_0$$

for large enough d. By choosing the original partition $\mathcal{P}$ such that $\mu(\partial S)=0$ for all $S\in\mathcal{P}$ and some weak* limit μ of the sequence μ_d we can now take the limit as $d\to\infty$ to obtain

$$H_{\mu}\left(\mathcal{P}^{[-N_0,N_0]}\right) \geq (2-7\varepsilon)N_0 \quad \text{for all } \varepsilon>0 \text{ and } N_0\geq 1,$$

i.e. that $h_\mu(T)\geq 1$. Theorem 4.1 can now be invoked to show that μ must be the $\mathrm{SL}_2(\mathbf{R})$-invariant measure on X.

We remark that the analysis above works only in the cocompact case; for e.g. $\Gamma=\mathrm{SL}_2(\mathbf{Z})$, there is no global injectivity radius; and no matter how fine one takes the partition $\mathcal{P}$, to cover a single atom of the partition $\mathcal{P}^{[-N,N]}$ one typically needs exponentially many Bowen N-balls.

4.3 Proof of Duke's theorem, controlling the time spent near the cusp

Passing from the cocompact to the nonuniform case raises two difficulties:

(i) Why is such a weak* limit a probability measure (indeed, why cannot such a sequence of measures μ_d converge to the zero measure)?

(ii) The proof outline presented in §4.2 used heavily the relation between Bowen N-balls and atoms of the partition $\mathcal{P}^{[-N,N]}$ for a finite partition $\mathcal{P}$. How can we adapt this argument to the nonuniform situation where in general many Bowen N-balls are needed to cover a partition element $S\in\mathcal{P}^{[-N,N]}$?

It turns out that these two difficulties are not unrelated, and to handle them one needs to control the time an orbit spends in the neighborhood of the cusp, so that this problem is related to *controlling the escape of mass*. What is needed is the following finitary version of the uniqueness of measure of maximal entropy:

Theorem 4.2. *Suppose μ_i is a sequence of A-invariant measures on X, and suppose there is a constant $r>0$ and a sequence $\delta_i\to 0$ such that for all sufficiently small $\varepsilon>0$ the "heights" $H_i=\delta_i^{-\varepsilon}$ satisfy*

(1) $\mu_i(X_{\geq H_i})\to 0$, *as* $i\to\infty$;

(2) $\mu_i\times\mu_i(\{(x,y)\in X_{\leq H_i}\times X_{\leq H_i} : d(x,y)<\delta_i\}\ll_\varepsilon \delta_i^{3-5\varepsilon}$.

Then $\mu_i\to\mu_X$, the $\mathrm{SL}_2(\mathbf{R})$-invariant measure on X, as $i\to\infty$.

Clearly, this, Proposition 3.3, and Proposition 3.6 with $\delta = d^{-\frac{1}{4}}$ are sufficient to prove Duke's theorem. Apart from the ideas already discussed in the last section, the main additional step is:

PROPOSITION 4.3. *Fix a height $M \geq 1$. Let $N \geq 1$ and consider a subset $V \subset [-N, N]$. Then the set*

$$Z(V) = \Big\{ x \in T^N X_{<M} \cap T^{-N} X_{<M} : \text{ for all } n \in [-N,N] \text{ we have } T^n(x) \in X_{\geq M} \Leftrightarrow n \in V \Big\}$$

can be covered by $\ll_M e^{2N - \frac{1}{2}|V|}$ Bowen N-balls. Moreover, $Z(V)$ is nonempty for only $\ll_M e^{\frac{2 \log\log M}{\log M} N}$ different sets $V \subset [-N, N]$.

In words, $Z(V)$ is the set of points $x \in X$ so that the trajectory $T^{-N}x$, $T^{-N+1}x, \dots, T^N x$ between times $-N$ and N begins and ends below height M and are above height M precisely at the time specified by the set V. So the content of the proposition is that orbits that spend a lot of time in a neighborhood of the cusp in fact can be covered by relatively few tube-like sets. Later we will turn this into the statement that those orbits have relatively little mass.

Note that as the size of V grows the number of Bowen N-balls needed to cover $Z(V)$ decreases, though even if $V = [-N-1, N+1]$ it is still exponential — indeed $\asymp e^N$, which is essentially the square root of the estimate we get for $V = \varnothing$.

We defer the proof of Proposition 4.3 to the next section. A purely ergodic theoretic formulation of this phenomenon is that a lot of mass near the cusp for an invariant probability measure results in a significantly smaller entropy for the geodesic flow. We will give such a formulation in Theorem 5.1; it implies in particular that:

Given a sequence of T-invariant probability measures μ_i with entropies $h_{\mu_i}(T) \geq c$, any weak limit μ satisfies $\mu(X) \geq 2c - 1$.*

We will discuss in Remark 5.2 why $c = 1/2$ is the critical point for this phenomenon.

4.4 CONTROLLING ESCAPE OF MASS, AND MAXIMAL ENTROPY

We proceed to the proof of Theorem 4.2, and start by showing that mass cannot escape, using assumption (2). We will use (1) of that theorem which

gives a mild control on how fast mass could possibly escape to be able to apply the covering argument in Proposition 4.3. That (2) can replace entropy in that argument is not surprising since we have already seen in Section 4.2 a relationship between this assumption and entropy.

LEMMA 4.4. *Let μ_i be a sequence of T-invariant measures as in Theorem* 4.2. *Let μ be a weak* limit of any subsequence of μ_i. Then*

$$\mu(X_{<M}) \geq 1 - \frac{2\log\log M}{\log M}$$

for every sufficiently large M, and so μ is a probability measure.

Proof. Fix some $\kappa > \frac{2\log\log M}{\log M}$. We will show that $\mu(X_{<M}) \geq 1-\kappa$.

We set $N_i = \lceil -\log \delta_i \rceil$ and $H_i = \delta_i^{-\epsilon}$ for some $\epsilon > 0$ determined below (more precisely: before the final displayed equation of this proof) in terms of κ. Notice that a geodesic trajectory of a point $x \in X_{\leq H_i}$ will visit $X_{<M}$ in less than $2\log H_i - 2\log M \leq 2\epsilon N_i$ steps either in the future or in the past. Hence

$$\bigcup_{n=-\lfloor 2\epsilon N_i\rfloor}^{\lfloor 2\epsilon N_i\rfloor} T^{-n} X_{<M} \supset X_{\leq H_i}$$

and so this union contains most of the μ_i-mass according to the assumption (1) of Theorem 4.2.

Let $N_i' = N_i + \lfloor 2\epsilon N_i \rfloor$. Then $T^{N_i'} X_{\leq H_i} \cap T^{-N_i'} X_{\leq H_i}$ is contained in the union of $\ll (\epsilon N_i)^2$ many sets of the form $T^{N_i'+n_-} X_{<M} \cap T^{-N_i'+n_+} X_{<M}$ where $|n_-|, |n_+| \leq 2\epsilon N_i$. We apply this to the set

$$X_\kappa = \left\{ x \in T^{N_i'} X_{\leq H_i} \cap T^{-N_i'} X_{\leq H_i} : \frac{1}{2N_i'+1} \sum_{n=-N_i'}^{N_i'} 1_{X_{\geq M}}(T^n x) > \kappa \right\}$$

consisting of points that spend an unexpected high portion of $[-N_i', N_i']$ above M.

We wish to estimate $\mu_i(X_\kappa)$.The set X_κ is also a union of sets of the form

$$Z' = X_\kappa \cap T^{N_i'+n_-} X_{<M} \cap T^{-N_i'+n_+} X_{<M}$$

with n_-, n_+ as before. It suffices to estimate $\mu_i(Z')$ for some fixed n_-, n_+. Replacing Z' by an appropriate shift $Z := T^k Z'$ we may consider instead $Z \subset T^N X_{<M} \cap T^{-N} X_{<M}$ where $N \in [N_i, N_i + 4\epsilon N_i]$. Adjusting the condition on the "average time spent above M" appropriately,

$$Z \subseteq \left\{ x \in T^N X_{<M} \cap T^{-N} X_{<M} : \frac{1}{2N+1} \sum_{n=-N}^{N} 1_{X_{\geq M}}(T^n x) > \kappa - O(\epsilon) \right\}.$$

To the right-hand set we apply Proposition 4.3; which shows that Z is covered by

$$\ell \ll_M e^{\frac{2\log\log M}{\log M}N} e^{2N-(\kappa-O(\epsilon))N} \le e^{2N_i+\frac{2\log\log M}{\log M}N_i-\kappa N_i+O(\epsilon)N_i}$$

many Bowen N-balls. Because $N \ge N_i$, we may also cover Z by ℓ many Bowen N_i-balls $S_1, \dots, S_\ell$.

Since Bowen N_i-balls have thickness $\le e^{-N_i} \le \delta_i$ along stable and unstable horocycle directions and thickness $\ll 1$ along A, we get that

$$\bigcup_{j=1}^{\ell} S_j \times S_j \subset \bigcup_{j=1}^{k} \{(x, ya_j) : d(x,y) < \delta_i\},$$

where $k \ll e^{N_i}$ and $a_j \in B_1^A$ are δ_i-dense. This remains true if we make the sets S_j disjoint by replacing S_2 by $S_2' = S_2 \smallsetminus S_1$, S_3 by $S_3' = S_3 \smallsetminus (S_1 \cup S_2), \dots$. By our assumption (2) we now get

$$\sum_{j=1}^{\ell} \mu_i(S_j')^2 \ll_\epsilon \delta_i^{3-5\epsilon} k \ll e^{-2N_i+5\epsilon N_i}.$$

Therefore, by Cauchy-Schwarz

$$\mu_i(Z) \le \sum_{j=1}^{\ell} \mu_i(S_j') \le \left(\sum_{j=1}^{\ell} \mu(S_j')^2\right)^{1/2} \ell^{1/2} \ll_{\epsilon,M} e^{\frac{\log\log M}{\log M}N_i - \frac{1}{2}\kappa N_i + O(\epsilon)N_i}.$$

Going through all possibilities for n_-, n_+ (of which there are $\ll e^{\epsilon N_i}$ many) this implies

$$\mu_i(X_\kappa) \ll_{\epsilon,M} e^{\left(\frac{\log\log M}{\log M} - \frac{1}{2}\kappa + O(\epsilon)\right)N_i}.$$

Given that we assume $\kappa > \frac{2\log\log M}{\log M}$ we can choose $\epsilon > 0$ small enough such that the exponent in the above expression is negative so that the measure goes to zero for $i \to \infty$ (since $N_i \to \infty$). By definition of X_κ we have

$$\mu_i(X_{\ge M}) = \int 1_{X_{\ge M}} \,\mathrm{d}\mu_i = \int \frac{1}{2N_i'+1} \sum_{n=-N_i'}^{N_i'} 1_{X_{\ge M}} \,\mathrm{d}\mu_i \le \kappa + \mu_i(X_\kappa) + 2\mu_i(X_{\ge H_i}),$$

which when $i \to \infty$ implies that $\mu(X_{<M}) \ge 1-\kappa$ for any $\kappa > \frac{2\log\log M}{\log M}$. This gives the lemma. □

We indicated in Section 4.2 how the elements of the refinement $\bigvee_{n=-N}^{N} T^{-n}\mathcal{P}$ are related to Bowen N-balls; but that analysis fails in the noncompact case, when trajectories visit the cusp. We now discuss the general case.

LEMMA 4.5. *For every $M > 1$ there exists a finite partition $\mathcal{P}$ of X such that for every $\kappa \in (0,1)$ and every N, "most elements of the refinement $\bigvee_{n=-N}^{N} T^{-n}\mathcal{P}$ are controlled by Bowen N-balls" in the following sense: there exists a set $X' \subset X$ so that*

- *X' is a union of $S_1, \dots, S_\ell \in \bigvee_{n=-N}^{N} T^{-n}\mathcal{P}$;*
- *each such S_j is contained in a union of at most $3^{\kappa(2N+1)}$ many Bowen N-balls;*
- *$\mu(X') \geq 1 - 2\mu(X_{\geq M})\kappa^{-1}$ for every invariant probability measure μ.*

For a given μ the choice of $\mathcal{P}$ can be made such that the boundaries of all sets of $\mathcal{P}$ have zero measure.

Proof. We define $\mathcal{P} = \{Q, P_1, \dots, P_k\}$ where $Q = X_{\geq M}$ and $\{P_1, \dots, P_k\}$ is a measurable partition of $X_{<M}$ whose elements have diameter less than η, where η is small enough in comparison to the injectivity radius of $X_{<M}$ (in the same sense as in the discussion in Section 4.2).

Note that the boundary of Q is a null set for every probability measure μ that is invariant under the geodesic flow. This is because every trajectory hits the boundary of Q in a countable set. Also, given μ we can find for every point $x \in X_{<M}$ an $\epsilon < \eta/2$ so that the boundary has measure zero. Applying compactness we construct $P_1, \dots, P_k$ from the algebra generated by finitely many such balls.

We claim that $S \in \mathcal{P}_N = \bigvee_{n=-N}^{N} T^{-n}\mathcal{P}$ has the property that any two points $x, y \in S$ satisfy

$$T^n x \in X_{<M} \Leftrightarrow T^n y \in X_{<M} \quad \text{for } n \in [-N, N] \quad \text{and}$$
$$d(T^n x, T^n y) < \eta \quad \text{whenever} \quad T^n x, T^n y \in X_{<M} \quad \text{and} \quad n \in [-N, N]\,.$$

Therefore, the *average* $f(x) = \frac{1}{2N+1} \sum_{n=-N}^{N} 1_{X_{\geq M}(T^n x)}$ is constant on sets of $\mathcal{P}_N$. We define

$$X' = \{x \in T^{-N} X_{<M} : f(x) \leq \kappa\}\,.$$

If μ is an invariant probability measure, invariance implies $\int f(x)\,\mathrm{d}\mu = \mu(X_{\geq M})$ and so $\mu(\{x : f(x) > \kappa\}) \leq \mu(X_{\geq M})\kappa^{-1}$. Therefore, X' has measure $\mu(X') \geq 1 - \mu(X_{\geq M}) - \mu(X_{\geq M})\kappa^{-1}$.

Consider now an element $S \in \mathcal{P}_N$ with $S \subset X'$. After taking the image of S under T^N we have for any $x, y \in S' = T^N S$ that

$$x \in X_{<M}\,, \qquad \frac{1}{2N+1} \sum_{n=0}^{2N} 1_{X_{\geq M}}(T^n x) \leq \kappa \quad \text{and} \tag{4.6}$$
$$d(T^n x, T^n y) < \eta \quad \text{whenever} \quad T^n x, T^n y \in X_{<M} \quad \text{and} \quad n \in [0, 2N]\,.$$

Let $V = \{n \in [0,2N] : T^n S' \subset X_{\geq M}\}$. We can now show inductively that for every $n \in [0,2N]$ the set S' is contained in a union of $3^{|[0,n-1]\cap V|}$ many sets of the form

$$xB^{U^+}_{2\eta e^{-n}} B^{U^-A}_{2\eta}, \qquad \text{where} \qquad x \in S'.$$

We will refer to these sets as *forward Bowen n-balls* and to x as its *center*. For $n = 0$ there is nothing to show (for notice that we allowed a bigger radius in the subgroups U^+ and U^-A). Suppose the claim holds for some n and let $x \in S'$ be a center of one of the forward Bowen n-balls. If $T^{n+1}x \in X_{<M}$ then $T^{n+1}S' \subset P_i$ for $i \geq 1$ and it follows easily that any point $y = xu^+g \in S'$ with $u^+ \in B^{U^+}_{2\eta e^{-n}}$ and $g \in B^{U^-A}_{2\eta}$ satisfies $u^+ \in B^{U^+}_{2\eta e^{-(n+1)}}$ (assuming again that η is small enough in comparison with the injectivity radius). If $T^{n+1}x \in X_{\geq M}$ then we can cover the forward Bowen n-ball by 3 forward Bowen $(n+1)$-balls.

Recall that for $S \subset X'$ we have $|V| \leq \kappa N$ and so by taking the preimages of $S' = T^N S$ and the forward Bowen $2N$-balls obtained the lemma follows. □

To prove Theorem 4.2 it remains to establish the following lemma and combine it with Lemma 4.4 and Theorem 4.1.

LEMMA 4.6. *A weak* limit μ of a subsequence of the invariant probability measures μ_i as in Theorem* 4.2 *has maximal entropy $h_\mu(T) = 1$.*

Proof. Let $\mathcal{P}$ be as in Lemma 4.5. Set $N_i = \lceil -\log \delta_i \rceil$ and define

$$\mathcal{P}_{N_i} = \bigvee_{n=-N_i}^{N_i} T^{-n}\mathcal{P}.$$

We wish to show that $H_{\mu_i}(\mathcal{P}_{N_i})$ is large by using Lemma 4.5 and assumption (2). Let $\kappa = \mu(X_{\geq M})^{1/2}$ for some weak* limit μ and define X_i as in Lemma 4.5 using $N = N_i$.

For any $S \in \mathcal{P}_{N_i}$ with $S \subset X_i$ there exists a cover of S consisting of $\leq 3^{\kappa(2N_i+1)}$ many Bowen N_i-balls; so there is a partition $\mathcal{R}(S)$ of S into $\leq 3^{\kappa(2N_i+1)}$ sets, each a subset of a Bowen N_i-ball. We define the partition $\mathcal{Q}_i$ as the partition consisting of all $S \in \mathcal{P}_{N_i}$ with $S \subset X \smallsetminus X_i$ and all elements of $\mathcal{R}(S)$ for any $S \subset X_i$. It follows that

$$H_{\mu_i}(\mathcal{Q}_i|\mathcal{P}_{N_i}) = \sum_{S\in\mathcal{P}_{N_i},\, S\subset X_i} \mu_i(S) H_{\mu_i|_S}(\mathcal{Q}_i) \leq \kappa(2N_i+1)\log 3. \tag{4.7}$$

Also since $\mathcal{Q}_i$ is a finer partition than $\mathcal{P}_{N_i}$ we have

$$H_{\mu_i}(\mathcal{Q}_i) = H_{\mu_i}(\mathcal{Q}_i \vee \mathcal{P}_{N_i}) = H_{\mu_i}(\mathcal{P}_{N_i}) + H_{\mu_i}(\mathcal{Q}_i|\mathcal{P}_{N_i})\,, \tag{4.8}$$

which together with (4.7) indicates that we wish to show that $H_{\mu_i}(\mathcal{Q}_i)$ is large.

Here we will use the assumption (2) from Theorem 4.2; but the elements of $\mathcal{Q}_i$ that lie outside X_i can be irregularly shaped, requiring a further estimate:

$$H_{\mu_i}(\mathcal{Q}_i) \geq H_{\mu_i}(\mathcal{Q}_i|\{X_i, X \smallsetminus X_i\}) \geq \mu_i(X_i) H_{\mu_i|_{X_i}}(\mathcal{Q}_i)\,. \tag{4.9}$$

Using (4.1) for the restriction $\mu_i|_{X_i}$ we see that

$$H_{\mu_i|_{X_i}}(\mathcal{Q}_i) \geq -\log \sum_{S\in\mathcal{Q}_i,\, S\subset X_i} \left(\frac{\mu(S)}{\mu(X_i)}\right)^2. \tag{4.10}$$

By construction of $\mathcal{Q}_i$ every $S \in \mathcal{Q}_i$ with $S \subset X_i$ is a subset of a Bowen N_i-ball. Proceeding as in Section 4.2 it follows that

$$\bigcup_{S\in\mathcal{Q}_i,\, S\subset X_i} S\times S \subset \bigcup_{i=1}^{k} \{(x, ya_i) : d(x,y) < \delta_i\}\,,$$

where $k \ll e^{N_i}$ and $a_1, \dots, a_k \in B_r^A(1)$ are chosen to be δ_i-dense. Together with assumption (2) of Theorem 4.2 this shows

$$\sum_{S\in\mathcal{Q}_i,\, S\subset X_i} \mu_i(S)^2 \ll_\varepsilon \delta_i^{3-5\varepsilon} e^{N_i} \ll e^{(-2+5\varepsilon)N_i}.$$

Let C_ϵ be the implicit constant here, that is to say,

$$\sum_{S\in\mathcal{Q}_i,\, S\subset X_i} \mu_i(S)^2 \leq C_\epsilon e^{-(2+5\varepsilon)N_i}.$$

Then, taking into account (4.9)–(4.10),

$$H_{\mu_i}(\mathcal{Q}_i) \geq 2\mu_i(X_i)\log\mu_i(X_i) - \mu_i(X_i)\log C_\epsilon + \mu_i(X_i)(2-5\epsilon)N_i\,.$$

Here the first two terms are bounded, so for large enough i

$$\begin{aligned} H_{\mu_i}(\mathcal{Q}_i) &\geq \mu_i(X_i)(2-6\epsilon)N_i \\ &\geq (1 - 2\kappa^{-1}\mu_i(X_{\geq M}))(2-6\epsilon)N_i\,, \end{aligned}$$

where we also used the estimate for X_i in Lemma 4.5. Combining this with (4.8) and (4.7) we get

$$H_{\mu_i}\Big(\bigvee_{n=-N_i}^{N_i} T^{-n}\mathcal{P}\Big) \geq \big(1 - 2\kappa^{-1}\mu_i(X_{\geq M})\big)(2-6\epsilon)N_i - O(\kappa N_i)\,.$$

Now fix some integer $N_0 \geq 1$. Using subadditivity of entropy we have for any large enough i that

$$H_{\mu_i}\Big(\bigvee_{n=-N_0}^{N_0} T^{-n}\mathcal{P}\Big) \geq \big(1 - 2\kappa^{-1}\mu_i(X_{\geq M})\big)(2-6\epsilon)N_0 - O(\kappa N_0) - \epsilon N_0\,.$$

This is now a statement involving only finitely many test functions, namely the characteristic functions of all elements of $\bigvee_{n=-N_0}^{N_0} T^{-n}\mathcal{P}$ and of $X_{\geq M}$. Since there is no escape of mass by Lemma 4.4 and since we can assume without loss of generality that all boundaries have zero measure for the weak* limit μ by Lemma 4.5, we get the same estimate for μ. Dividing by $2N_0$ and letting N_0 now go to infinity we arrive at

$$h_\mu(T) \geq \big(1 - 2\mu(X_{\geq M})^{1/2}\big)(1-3\epsilon) - O(\mu(X_{\geq M})^{1/2}) - \epsilon$$

for any $M \geq 1$ and $\epsilon > 0$.

Since $\mu(X_{\geq M})$ can be made arbitrarily small, it follows that $h_\mu(T) \geq 1$, i.e. T has maximal entropy. □

5. Trajectories spending time high in the cusp, and a proof of Proposition 4.3

Apart from the characterization of the Haar measure as the unique measure of maximal entropy in Theorem 4.1, the main technical estimate needed to prove Theorem 4.2 is Proposition 4.3. We recall that this proposition states that the set

$$Z(V) = \Big\{x \in T^N X_{<M} \cap T^{-N} X_{<M} :\ \text{for all } n \in [-N,N] \text{ we have } T^n(x) \in X_{\geq M} \Leftrightarrow n \in V\Big\}$$

can be covered by $\ll_M e^{2N - \frac{1}{2}|V|}$ Bowen N-balls.

In addition to proving this, we shall also prove here the promised purely ergodic formulation of "high entropy inhibits escape of mass", namely:

THEOREM 5.1. *Let T be the time-one-map for the geodesic flow. There exists some M_0 with the property that*

$$h_\mu(T) \leq 1 + \frac{\log\log M}{\log M} - \frac{\mu(X_{\geq M})}{2}$$

for any probability measure μ *on* $X = \mathrm{SL}(2,\mathbf{Z})\backslash \mathrm{SL}(2,\mathbf{R})$ *invariant under the geodesic flow and any* $M \geq M_0$. *In particular, for a sequence of* T*-invariant probability measures* μ_i *with entropies* $h_{\mu_i}(T) \geq c$, *any weak** *limit* μ *satisfies* $\mu(X) \geq 2c - 1$.

REMARK 5.2. Roughly speaking $1/2$ is the critical point for Theorem 5.1 because the "upward" and "downward" parts of a trajectory, that goes high in the cusp, are strongly related to each other. In fact, in the case of a p-adic flow this phenomenon is easy to explain.

We consider another dynamical system of similar flavor: here the space will be [4])

$$Y = \mathrm{PGL}_2(\mathbf{Z}[1/p])\backslash \mathrm{PGL}_2(\mathbf{R}) \times \mathrm{PGL}_2(\mathbf{Q}_p)$$

and the action will be by multiplication on the right of the $\mathrm{PGL}_2(\mathbf{Q}_p)$-component by $a_p = \begin{pmatrix} p & \\ & 1 \end{pmatrix}$. Let $M < \mathrm{PGL}_2(\mathbf{R}) \times \mathrm{PGL}_2(\mathbf{Q}_p)$ be the product of $\mathrm{PO}_2(\mathbf{R})$ and the group of diagonal matrices in $\mathrm{PGL}_2(\mathbf{Z}_p)$. There is a natural right M-invariant projection $\pi\colon Y \to \mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}$, and on this latter space we have the *Hecke correspondence* which attaches to a point $\dot{z} \in \mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}$ a set $T_p(\dot{z})$ of $p+1$ new points, namely if $z \in \mathbf{H}$ is a representative of $\dot{z}$ then

$$T_p(\dot{z}) = \mathrm{PSL}_2(\mathbf{Z})\backslash \{pz, z/p, (z+1)/p, \ldots, (z+p-1)/p\}. \tag{5.1}$$

The space Y/M can be identified with the set of bi-infinite sequences $\ldots, y_{-1}, y_0, y_1, \ldots$ with $y_i \in T_p(y_{i-1}) \smallsetminus \{y_{i-2}\}$, and under this identification multiplication by a_p in the p-direction becomes simply the *shift action*. This in particular shows that multiplication by a_p on Y/M (or, with a bit more effort on Y) has entropy $\leq \log p$, and just like in our case this maximum is attained for the Haar measure on Y. From (5.1) it is clear that if $y \in \mathrm{PSL}(2,\mathbf{Z})\backslash\mathbf{H}$ is high up in the cusp, precisely 1 of its T_p-points will be higher in the cusp, and p of these points would be lower than y in the cusp. Therefore if $\ldots y_{-1}, y_0, y_1, \ldots$ is a sequence of points of $\mathrm{PSL}(2,\mathbf{Z})\backslash\mathbf{H}$ as above and if y_k are high up in the cusp for some contiguous range of k's, say $n \leq k \leq m$, then in this range given the value of y_k there is only one possible way of choosing y_{k+1} so that it is higher than y_k. Since by assumption $y_{k+2} \neq y_k$ once y_{k+1} is lower than y_k, the point y_{k+2} must be lower than y_{k+1}. Hence if y_{k+1} is lower than y_k for some k in the above range, then $y_{k'+1}$ must be lower then $y_{k'}$ for all k' in the range $k \leq k' \leq m$. From the above discussion it follows that while the trajectory is high up in the cusp, we have a choice of which subsequent point to choose *only* half of the time, whence the factor $\frac{1}{2}$.

[4]) For technical reasons, it is preferable to use PGL_2 here rather than SL_2.

5.1 Proof of Proposition 4.3: the number of possible sets V

The easiest part of Proposition 4.3 is the final assertion, i.e. if we write

$$Q_{M,N} = \bigvee_{n=-N}^{N} T^{-n}\{X_{\geq M}, X_{<M}\},$$

then the above partition $Q_{M,N}$ has $\ll_M e^{\frac{2\log\log M}{\log M}N}$ many elements.

We make use of the fundamental domain $\mathcal{S} \subset \mathrm{PSL}_2(\mathbf{R})$ from §1.3; the geodesic flow X corresponds to following the geodesic determined by (z, v) until the boundary of the fundamental region is reached, at which point one applies either $\begin{pmatrix} 1 & \pm 1 \\ & 1 \end{pmatrix}$ to shift the geodesic horizontally or $\begin{pmatrix} & -1 \\ 1 & \end{pmatrix}$ to reflect on the bottom boundary of the fundamental region.

The basic point in the proof is that if $x \in X$ satisfies $\mathrm{ht}(x) \geq M$, then $\mathrm{ht}(T^n x) \geq 1$ so long as $n < \lfloor 2\log M \rfloor$, i.e. one needs at least $\lfloor 2\log M \rfloor$ steps to reach points of height less than 1.

Therefore, in a time interval of length $2\lfloor 2\log M \rfloor$ there can be only one stretch of times for which the points on the orbit are of height at least M. In other words, the possible starting and end points of that time interval completely determine an element of $Q_{M,\lfloor 2\log M\rfloor}$ which therefore has at most $\ll \log^2 M$, say $\leq c_0 \log^2 M$, many elements. To obtain the final assertion of Proposition 4.3, we note that $Q_{M,N}$ can be obtained by taking refinements of $\lfloor \frac{2N+1}{2\lfloor 2\log M\rfloor + 1} \rfloor \leq \frac{2N+1}{4\log M - 1}$ many images and pre-images of $Q_{M,\lfloor 2\log M\rfloor}$ and at most $2\lfloor 2\log M \rfloor$ many of $\{X_{\geq M}, X_{<M}\}$. We get that $Q_{M,N}$ has size $\ll_M (c_0 \log^2 M)^{\frac{2N}{4\log M - 1}}$, which is at most $e^{\frac{2\log\log M}{\log M}N}$ once M is large enough.

5.2 Proof of Proposition 4.3: covering $Z(V)$ by Bowen balls

Write $a = \begin{pmatrix} e^{\frac{1}{2}} & 0 \\ 0 & e^{-\frac{1}{2}} \end{pmatrix}$, so that $T(x) = xa$. Since $X_{<M}$ has compact closure, it suffices to restrict ourselves to a neighborhood O of a point $x_0 \in X_{<M}$. By taking the image under T^N it also suffices to study the forward orbit as follows. We will show that for the set $V \subset [0, N-1]$ picked, the set

$$Z_O^+ = \Big\{ x \in O \cap T^{-N} X_{<M} : \\ \text{for all } n \in [0, N-1] \text{ we have } T^n(x) \in X_{\geq M} \Leftrightarrow n \in V \Big\}$$

can be covered by $\ll_M 2^{N-\frac{1}{2}|V|}$ forward Bowen N-balls xB_N^+, where

$$B_N^+ = \bigcap_{n=0}^{N-1} a^{-n} B_\eta^G a^n.$$

We may assume that the neighborhood we will consider is of the form

$$O = x_0 B_{\eta/2}^{U^+} B_{\eta/2}^{U^- A},$$

where B_r^H denotes the r-ball of the identity in a subgroup $H < \mathrm{SL}_2(\mathbf{R})$, A denotes the diagonal subgroup, and U^+ resp. U^- denote the unstable and stable horocyclic subgroups as in Section 4.2.

Notice that by applying T^n to O we get a neighborhood of $T^n(x_0)$ for which the U^+-part is e^n times as big while the second part is still contained in $B_{\eta/2}^{U^- A}$. By breaking the U^+-part into $\lceil e^n \rceil$ sets of the form $u_i^+ B_{\eta/2}^{U^+}$ for various $u_i^+ \in U^+$ we can write $T_2^n(O)$ as a union of $\lceil e^n \rceil$ sets of the form

$$T^n(x_0) u_i^+ B_{\eta/2}^{U^+} a^{-n} B_{\eta/2}^{U^- A} a^n,$$

i.e. we obtain neighborhoods of similar shape. If we take the preimage under T^n of this set, we obtain a set contained in the forward Bowen n-ball $T^{-n}(T^n(x_0)u_i^+)B_n^+$. We will be iterating this procedure, but by using the information that the orbit has to stay above height M for a long time we will be able to cut down on the number of $u_i^+ \in U^+$ needed to cover Z_O^+.

In the proof of the claim we will use a partition of $[0,N]$ into subintervals of two types according to the set V. Notice that as in the proof of §5.1, we can assume that V itself consists of intervals that are separated by $2\lfloor 2\log M \rfloor$. For otherwise the set Z_O^+ is empty since no orbit under T can leave $X_{\geq M}$ and return to it in a shorter amount of time. We enlarge every such subinterval of V by $\lfloor 2\log M \rfloor$ on both sides to obtain the first type of disjoint intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k$. At the end points 0 and N we have required that $x, T^N(x) \in X_{<M}$ for all $x \in Z_O^+$. For this reason we can assume without loss of generality that all of these intervals are contained in $[0,N]$. (If this is not the case, we can enlarge the interval $[0,N]$ accordingly and absorb the change of the desired upper estimate in the multiplicative constant that depends on M alone.) The remainder of $[0,N]$ we collect into the intervals $\mathcal{J}_1, \ldots, \mathcal{J}_\ell$.

We will go through the time intervals $\mathcal{I}_i$ and $\mathcal{J}_j$ in their respective order inside $[0,N]$. At each stage we will divide any of the sets obtained earlier into $\lceil e^{|\mathcal{I}_i|} \rceil$ — or $\lceil e^{|\mathcal{J}_j|} \rceil$ — many sets, and in the case of $\mathcal{I}_i$ show that we do not have to keep all of them. More precisely, we assume inductively that for some $K \leq N$ we have $[0,K] = \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_i \cup \mathcal{J}_1 \cup \ldots \cup \mathcal{J}_j$ and that all

points in Z_O^+ can be covered by

$$\leq 2e^{|\mathcal{J}_1|+\cdots+|\mathcal{J}_j|+i\lfloor 2\log_2 M\rfloor+\frac{1}{2}(|\mathcal{I}_1|+\cdots+|\mathcal{I}_i|)}$$

many preimages under T^K of sets of the form

$$T_2^K(x_0)u^+B_{\eta/2}^{U^+}a^{-K}B_{\eta/2}^{U^-A}a^K. \tag{5.2}$$

Note that for $K=N$ this gives the lemma since by construction $|\mathcal{I}_1|+\cdots+|\mathcal{I}_k| = 2k\lfloor 2\log M\rfloor + |V|$.

For the inductive step it will be useful to assume a slightly stronger inductive assumption, namely that the multiplicative factor 2 is only allowed if $[0,K]$ ends with the interval $\mathcal{J}_j$. Therefore, notice that if the next interval is $\mathcal{J}_{j+1}$ (i.e. $[0,K]$ ends with $\mathcal{I}_i$) then there is not much to show. In that case we keep all of the $\lceil e^{|\mathcal{J}_{j+1}|}\rceil \leq 2e^{|\mathcal{J}_{j+1}|}$-many Bowen balls constructed above and obtain the claim.

So assume now that the next time interval is $\mathcal{I}_{i+1}=[K+1,K+S]$. Here we will make use of the geometry of geodesics that visit $X_{\geq M}$ during that subinterval. Pick one of the sets (5.2) obtained in the earlier step and denote it by Y. By definition of Z_O^+ we are only interested in points $y\in Y$ which satisfy

$$T^n(y)\in X_{\geq M} \Leftrightarrow K+n\in V,$$

or equivalently

$$\begin{aligned}\operatorname{ht}(y),\operatorname{ht}\big(T(y)\big),\dots,\operatorname{ht}\big(T^{\lfloor 2\log M\rfloor}(y)\big) &< M,\\ \operatorname{ht}\big(T^{\lfloor 2\log M\rfloor+1}(y)\big),\dots,\operatorname{ht}\big(T^{S-\lfloor 2\log M\rfloor}(y)\big) &\geq M,\\ \operatorname{ht}\big(T^{S-\lfloor 2\log M\rfloor+1}(y)\big),\dots,\operatorname{ht}\big(T^{S}(y)\big) &< M.\end{aligned}$$

If there is no such point in Y there is nothing to show. So suppose $y,y'\in Y$ are such points. We will use the above restrictions on the heights to show that if

$$y=T_2^K(x_0)u^+u^+(t)v \quad\text{and}\quad y'=T_2^K(x_0)u^+u^+(t')v' \tag{5.3}$$

for $u^+(t),u^+(t')\in B_{\eta/2}^{U^+}$ and v,v' in the conjugate of $B_{\eta/2}^{U^-A}$, then $|t-t'|\ll 2^{-S/2}$. We can draw the geodesic orbits defined by y and y' in the upper half model of the hyperbolic plane and relate the conditions on y,y' to geometric information about these geodesics. We choose the lifting of the paths in such a way that the time interval for which the height is above M becomes the part of the geodesic where the imaginary part is above M^2.

For the translation of the properties we will use the following observation: For two points $z_1, z_2 \in \mathbf{H}$ on a geodesic line that are either both on the upwards part or both on the downwards part of the corresponding semi-circle their hyperbolic distance satisfies

$$|\log \mathrm{Im}(z_1) - \log \mathrm{Im}(z_2)| \leq d(z_1, z_2) \leq |\log \mathrm{Im}(z_1) - \log \mathrm{Im}(z_2)| + 1\,. \tag{5.4}$$

The lower bound actually gives the shortest distance between points with imaginary part $\mathrm{Im}(z_1)$ and points with imaginary part $\mathrm{Im}(z_2)$. The upper bound gives the length of a path that first connects the point lower down, say z_1, to the point z' immediately above with imaginary part $\mathrm{Im}(z_2)$ and then moves horizontally to a point that is $\mathrm{Im}(z_2)$ far to the left or right of z' towards z_2. For two points z_1, z_2 on the upwards or downwards part of a semi-circle this path actually goes through z_2.

Applying the lower bound in (5.4) to the points corresponding to

$$y \quad \text{and} \quad T_2^{\lfloor 2\log M\rfloor+1}(y)$$

whose hyperbolic distance is $\lfloor 2\log M\rfloor + 1$ we see that $\mathrm{Im}(y) \gg 1$ (where in a slight abuse of notation we identify y with the lifted point in $\mathbf{H}$). Similarly, we get from the upper bound for y and $T_2^{\lfloor 2\log M\rfloor}(y)$ that $\mathrm{Im}(y) \ll 1$. Similar estimates hold for $T_2^S(y), y'$ and $T_2^S(y')$.

We assume that the points y, y' are lifted in such a way that $\Re(y) \in [-1/2, 1/2]$ and such that y' is close to y. Denote by $\alpha_-, \alpha_+ \in \mathbf{R}$ the backwards and forwards limit points of the geodesic defined by y on the boundary of $\mathbf{H}$ and similarly by α'_-, α'_+ the endpoints of the geodesic for y'. Then $|\alpha_-| < 2 + \frac{1}{2}$ since the lifting of the point y was chosen such that the times of height $\geq M$ in X correspond to imaginary part $\geq M^2$. For y' this implies for small enough η that $|\alpha'_-| < 3$.

Let $R = \frac{1}{2}|\alpha_+ - \alpha_-|$ be the radius of the half-circle defined by y and define R' similarly for y'. Then the above shows $R \ll |\alpha_+| \ll R$ once M and so R are large enough to make α_- negligible in comparison to α^+. Similarly $R' \ll |\alpha'_+| \ll R'$.

Applying (5.4) twice, once for y and the point z on the same geodesic with imaginary part R, and once for z and $T_2^S(y)$ we get

$$|S - 2\log R| \ll 1 \quad \text{and similarly} \quad |S - 2\log R'| \ll 1\,. \tag{5.5}$$

Therefore, $R \ll R' \ll R$ and so $|\alpha_+| \ll |\alpha'_+| \ll |\alpha_+|$.

Suppose $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbf{R})$ defines $y = T_2^K(x_0)u^+u^+(t)v$ in the sense that the natural action of g maps the upwards vector at i to the vector

associated to y for the lifting considered above. Then $\alpha_+ = g(\infty) = \frac{a}{c}$ and $\alpha_- = g(0) = \frac{b}{d}$. Similarly, suppose g' defines $y' = T_2^K(x_0)u^+u^+(t')v'$ such that $\alpha'_+ = g'(\infty)$. Using this notation we summarize what we already know about these matrices

$$\begin{aligned} \max(|a|,|b|,|c|,|d|) &\ll 1\,, \\ R \ll |\alpha_+| = |\frac{a}{c}| &\ll R\,, \\ R \ll |\alpha'_+| &\ll R\,, \text{ and} \\ |\alpha_-| = |\frac{b}{d}| &\ll 1\,. \end{aligned} \tag{5.6}$$

Here the first estimate follows since we know roughly the position of the lift corresponding to y which means that g belongs to a compact subset of $\mathrm{SL}(2,\mathbf{R})$. We claim the above implies that

$$1 \ll |d|\,, \quad 1 \ll |a|\,, \quad \text{and} \quad |c| \ll |a|R^{-1} \ll R^{-1}. \tag{5.7}$$

The first estimate follows since $|b| \ll |d|$ by the last estimate in (5.6) and since $g \in \mathrm{SL}(2,\mathbf{R})$ belongs to a compact subset so that not both b and d are small. The second claim follows similarly from the second estimate in (5.6).

To simplify the following calculation we would like to remove the elements v, v' (as in (5.3)) from our consideration — but to do this we need to see how this affects the above statements. Recall first that $v, v' \in B_\eta^{U^-A}$ and so $v(\infty) = v'(\infty) = \infty$. Therefore, the first three estimates above remain unaffected when changing g resp. g' on the right by $v^{-1}, (v')^{-1}$. Moreover, we have $|v^{-1}(0)| \ll \eta$ and so for small enough η that $1 \ll |d| \ll |cv^{-1}(0)+d|$ which implies $|gv^{-1}(0)| \ll 1$. In other words, none of the estimates in (5.6) are affected (apart from possibly the values of the implicit constants) by the proposed transition from g to gv^{-1} resp. g' to $g'(v')^{-1}$ and we can assume $v = v' = e$.

Comparing the definitions of y and y' we get $g' = gu^+(t)^{-1}u^+(t')$. Therefore,

$$\alpha'_+ = g'(\infty) = \big(gu^+(t'-t)\big)(\infty) = \frac{\frac{a}{t'-t} + b}{\frac{c}{t'-t} + d} = \frac{a + b(t'-t)}{c + d(t'-t)}\,.$$

Since $1 \ll |a|$, $u^+(t), u^+(t') \in B_{\eta/2}^{U^+}$, and so $|t'-t| \ll \eta$ we can simplify the numerator and obtain together with the third estimate in (5.6) that for small enough $\eta > 0$

$$R \ll \big|\frac{a}{c + d(t'-t)}\big| \ll R\,,$$

or equivalently

$$R^{-1} \ll |c + d(t' - t)| \ll R^{-1}.$$

Since $|c| \ll R^{-1}$ and $1 \ll |d|$ by (5.7) this implies the estimate $|t'-t| \ll R^{-1}$. Now recall from (5.5) that $e^{S/2} \ll R$, so that we get the desired $|t'-t| \ll e^{-S/2}$.

Recall next that in the current time interval $\mathcal{I}_{i+1}$ we divide $B^{U^+}_{\eta/2}$ into $\lceil e^S \rceil$ balls of the form $B^{U^+}_{e^{-s}\eta/2}$. Since all points y' that belong to $Y \cap T^K(Z_O^+)$ satisfy the estimate $|t'-t| \ll e^{-S/2}$ we see that only $\ll e^S e^{-S/2} = e^{S/2}$ of these balls can (after the correct thickening along AU^-) contain an element of $Y \cap T^K(Z_O^+)$. This implies the inductive claim if we assume M sufficiently large that the upper bound we got is strictly bounded from above by $\frac{1}{2}e^{\lfloor 2\log M \rfloor + S/2}$.

This concludes the proof of Proposition 4.3. □

5.3 Entropy and covers; proof of Theorem 5.1

For the proof of Theorem 5.1 we need to relate entropy and covers via Bowen balls. For this we need the following (well-known) result, which is proved in Appendix B below (for cocompact Γ it follows from Brin and A. Katok [5]).

Lemma 5.3. *Let μ be an A-invariant measure on $X = \Gamma\backslash \mathrm{SL}(2,\mathbf{R})$. For any $N \geq 1$ and $\epsilon > 0$ let $BC(N,\epsilon)$ be the minimal number of Bowen N-balls needed to cover any subset of X of measure bigger than $1-\epsilon$. Then*

$$h_\mu(T) \leq \lim_{\epsilon\to 0} \liminf_{N\to\infty} \frac{\log BC(N,\epsilon)}{2N},$$

where T is the time-one-map of the geodesic flow.

Proof of Theorem 5.1. Note first that it suffices to consider *ergodic* measures. For if μ is not ergodic, we can write μ as an integral of its ergodic components $\mu = \int \mu_t \,\mathrm{d}\tau(t)$ for some probability space (T,τ). Therefore, $\mu(X_{\geq M}) = \int \mu_t(X_{\geq M})\,\mathrm{d}\tau(t)$ but also $h_\mu(T) = \int h_{\mu_t}(T)\,\mathrm{d}\tau(t)$ by [26, Thm. 8.4], so that the desired estimate follows from the ergodic case.

Suppose μ is ergodic. To apply Lemma 5.3 we need to show that most of X can be covered by not too many Bowen N-balls. Once $M > 3$ we have that every T-orbit visits $X_{<M}$, and so $\mu(X_{<M}) > 0$. By the ergodic theorem there exists for every $\epsilon > 0$ some $K \geq 1$ such that

$$Y = \bigcup_{k=0}^{K-1} T^{-k} X_{<M} \quad \text{satisfies} \quad \mu(Y) > 1 - \epsilon\,.$$

Moreover, also by the ergodic theorem

$$\frac{1}{2N+1}\sum_{n=-N}^{N} 1_{X_{\geq M}}(T^n(x)) \to \mu(X_{\geq M})$$

as $N \to \infty$ for a.e. $x \in X$. So for large enough N the average on the left will be bigger than $\kappa = \mu(X_{\geq M}) - \epsilon$ for any $x \in X_1$ and some subset $X_1 \subset X$ of measure $\mu(X_1) > 1 - \epsilon$. Clearly for any N the set

$$Z = X_1 \cap T^N Y \cap T^{-N} Y$$

has measure bigger than $1-3\epsilon$. Recall that we are interested in the asymptotics of the minimal number of Bowen N-balls needed to cover Z. Here $N \to \infty$ while ϵ, and so also K, remains fixed. Since we can decompose Z into K^2 many sets of the form

$$Z' = X_1 \cap T^{N-k_1} X_{<M} \cap T^{-N-k_2} X_{<M}\,,$$

it suffices to cover these, and for simplicity of notation we assume $k_1 = k_2 = 0$. Next we split Z' into the sets $Z(V)$ as in Proposition 4.3 for the various subsets $V \subset [-N, N]$. §5.1 shows that we need at most $\ll_M e^{\frac{2\log\log M}{\log M}N}$ many of these. Moreover, by our assumption on X_1 we only need to look at sets $V \subset [-N, N]$ with $|V| \geq \kappa(2N+1)$. Therefore, Proposition 4.3 gives that each of those sets $Z(V)$ can be covered by $\ll_M e^{(1-\frac{\kappa}{2})2N}$ many Bowen N-balls. Together we see that Z can be covered by $\ll_{M,K} e^{\frac{2\log\log M}{\log M}N + (1-\frac{\kappa}{2})2N}$ Bowen N-balls. Applying Lemma 5.3 we arrive at

$$h_\mu(T) \leq 1 + \frac{\log\log M}{\log M} - \frac{\mu(X_{\geq M}) - \epsilon}{2}$$

for any $\epsilon > 0$, which proves the theorem. □

A. Representations of binary quadratic forms by ternary forms

In this section we establish Proposition 3.4:

PROPOSITION. *Let Q be an non-degenerate, integral*[5]*) ternary quadratic form on $\mathbf{Z}^3$, and let*

$$q(x, y) = a_1 x^2 + a_2 xy + a_3 y^2$$

[5]) I.e. $Q(\mathbf{Z}^3) \subset \mathbf{Z}$

be a non-degenerate binary quadratic form on $\mathbf{Z}^2$. *Let* f^2 *be the greatest square dividing* $\gcd(a_1, a_2, a_3)$. *Then the number* $N(q)$ *of embeddings of* $(\mathbf{Z}^2, q)$ *into* $(\mathbf{Z}^3, Q)$, *modulo the action of* $\mathrm{SO}_Q(\mathbf{Z})$, *is* $\ll_{Q,\epsilon} f \max(|a_1|, |a_2|, |a_3|)^\epsilon$.

We recall that an *embedding* of $(\mathbf{Z}^2, q)$ into $(\mathbf{Z}^3, Q)$ is a linear map $\iota\colon \mathbf{Z}^2 \to \mathbf{Z}^3$ with the property that $Q(\iota(\mathbf{x})) = q(\mathbf{x})$. Such a proposition was established for the first time by Venkov for $Q = x^2 + y^2 + z^2$ and extended by Pall to other quadratic forms [25, 21]. The proposition can be deduced from *Siegel's mass formula*; here we present a direct and elementary argument inspired by the adelic proof of Siegel's mass formula as outlined by Tamagawa (cf. Weil's paper [27]).

REMARK A.1.

– One may wonder what the dependency on Q in the above bound looks like; this is for instance important to obtain equidistribution results when Q is allowed to vary (see for instance [14, Thm. 1.8]). In the case where Q is a multiple of the norm form on a lattice in the space of trace zero elements of a quaternion algebra whose associated order is an *Eichler order*, it can be shown that the dependency is of the shape $\ll_\varepsilon |\mathrm{disc}(Q)|^{1/2+\varepsilon} \dots$. It seem plausible that this holds in general

– The argument provides, in fact, an upper bound for the sum over a set of representatives $Q_i,\ i = 1, \dots, g$ of the genus classes of Q, of the number of embeddings of $(\mathbf{Z}^2, q)$ into $(\mathbf{Z}^3, Q_i)$ modulo $\mathrm{SO}_{Q_i}(\mathbf{Z})$.

– Finally it is easy to see that this argument carries over without significant changes to quadratic forms defined over a general number field.

A.1 REDUCTION TO LOCAL COUNTING PROBLEMS

Fix an embedding $\iota\colon (\mathbf{Z}^2, q) \hookrightarrow (\mathbf{Z}^3, Q)$ and let

$$L := \iota(\mathbf{Z}^2)$$

be its image (if no such embedding exists, we are obviously done). Then any other embedding ι' is (by *Witt's theorem*; see [22, IV.1.5, Theorem 3]) of the form $g \circ \iota$, with $g \in \mathrm{SO}_Q(\mathbf{Q})$. The stabilizer of ι inside $\mathrm{SO}_Q(\mathbf{Q})$ is trivial, for any isometry fixing L pointwise would need to map $L^\perp$ to itself and so must be multiplication by ± 1 on $L^\perp$; the condition of determinant 1 forces it to be the identity. The number of embeddings $N(L)$ (up to the action of $\mathrm{SO}_Q(\mathbf{Z})$) is therefore precisely the number of cosets $\dot{g} \in \mathrm{SO}_Q(\mathbf{Z}) \backslash \mathrm{SO}_Q(\mathbf{Q})$ so that $gL \subset \mathbf{Z}^3$.

Given a rational lattice $\Lambda \subset \mathbf{Q}^3$, for any prime p we denote by

$$\Lambda_p = \Lambda \otimes_{\mathbf{Z}} \mathbf{Z}_p$$

its closure inside $\mathbf{Q}_p^3$. Let us recall that the map

$$\Lambda \mapsto (\Lambda_p)_p$$

is a bijection between the set of lattices in $\mathbf{Q}^3$ and the set of sequences of lattices indexed by the primes $(\Lambda_p)_p$, $\Lambda_p \subset \mathbf{Q}_p^3$ such that $\Lambda_p = \mathbf{Z}_p^3$ for a.e. p. Write $K_p = \mathrm{SO}_Q(\mathbf{Z}_p)$ for the stabilizer of $\mathbf{Z}_p^3$ inside $\mathrm{SO}_Q(\mathbf{Q}_p)$ and let

$$\mathrm{SO}_Q(\mathbf{A}_f) = \{g_f = (g_p)_p \ : \quad g_p \in \mathrm{SO}_Q(\mathbf{Q}_p)\,, \quad g_p \in \mathrm{SO}_Q(\mathbf{Z}_p) \quad \text{for a.e. } p\}\,;$$

the above bijection induces an action of $\mathrm{SO}_Q(\mathbf{A}_f)$ on the set of rational lattices:

$$g_f.\Lambda \leftrightarrow g_f.(\Lambda_p)_p := (g_p\Lambda_p)_p\,.$$

REMARK A.2. The group $\mathrm{SO}_Q(\mathbf{A}_f)$ is the group of *finite adèles* of SO_Q. The $\mathrm{SO}_Q(\mathbf{A}_f)$-orbit of a lattice $\Lambda \in \mathbf{Q}^3$ under this action is called the *Q-genus* of Λ. We will not need much of this terminology or discuss further properties of adelic groups here.

The group $\mathrm{SO}_Q(\mathbf{Q})$ embeds diagonally into $\mathrm{SO}_Q(\mathbf{A}_f)$. Now the stabilizer of $\mathbf{Z}^3$ in $\mathrm{SO}_Q(\mathbf{A}_f)$ is $K_f = \prod_p \mathrm{SO}_Q(\mathbf{Z}_p)$ and since $K_f \cap \mathrm{SO}_Q(\mathbf{Q}) = \mathrm{SO}_Q(\mathbf{Z})$, $\mathrm{SO}_Q(\mathbf{Z})\backslash \mathrm{SO}_Q(\mathbf{Q})$ injects into $K_f \backslash \mathrm{SO}_Q(\mathbf{A}_f)$.

Consequently, letting $L_p = L \otimes_{\mathbf{Z}} \mathbf{Z}_p$ be the closure of L inside $\mathbf{Z}_p^3$, we have

$$\begin{aligned} N(L) &\le |\{g_f \in K_f \backslash \mathrm{SO}_Q(\mathbf{A}_f) : g_f.L \subset \mathbf{Z}^3\}| \\ &\le \prod_p |\{g_p \in \mathrm{SO}_Q(\mathbf{Z}_p)\backslash \mathrm{SO}_Q(\mathbf{Q}_p) : g_p.L_p \subset \mathbf{Z}_p^3\}| \\ &= \prod_p |\{g_p \in \mathrm{SO}_Q(\mathbf{Q}_p)/\mathrm{SO}_Q(\mathbf{Z}_p) : L_p \subset g_p\mathbf{Z}_p^3\}| = \prod_p N(L_p)\,, \end{aligned}$$

with

$$N(L_p) = |\{g_p \in \mathrm{SO}_Q(\mathbf{Q}_p)/K_p : L_p \subset g_p\mathbf{Z}_p^3\}| = |\{\Lambda \in \mathrm{SO}_Q(\mathbf{Q}_p).\mathbf{Z}_p^3 : L_p \subset \Lambda\}|$$

being the number of lattices in $\mathbf{Q}_p^3$, within the Q-isometry class of $\mathbf{Z}_p^3$ that contain L_p. We have proven that

$$N(L) \le \prod_p N(L_p)\,,$$

and thus have reduced our counting problem to a collection of local counting problems (as we will see below $N(L_p) = 1$ for a.e. p); a more careful analysis of what we have said so far is very closely related to the proof of the mass formula. In the present paper, however, we need only upper bounds.

A.2 The anisotropic case and a reduction step

We first introduce some notations. We denote by

$$\langle \mathbf{x}, \mathbf{x}' \rangle = Q(\mathbf{x} + \mathbf{x}') - Q(\mathbf{x}) - Q(\mathbf{x}')$$

the bilinear form associated with Q; so $\langle \mathbf{x}, \mathbf{x} \rangle = 2Q(\mathbf{x})$. The *discriminant of* Q is set to be

$$\operatorname{disc}(Q) = \det(\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_{i,j \le 3}$$

for $\{x_1, x_2, x_3\}$ any basis of $\mathbf{Z}^3$. Since Q is integral $\langle \mathbf{Z}^3, \mathbf{Z}^3 \rangle \subset \mathbf{Z}$, so $\operatorname{disc}(Q)$ is a non-zero integer.

We notice first that if Q does not represent 0 nontrivially over $\mathbf{Q}_p$ (i.e. is *anisotropic over* $\mathbf{Q}_p$), then $\mathrm{SO}_Q(\mathbf{Q}_p)$ is compact and

$$N(L_p) \le [\mathrm{SO}_Q(\mathbf{Q}_p) : \mathrm{SO}_Q(\mathbf{Z}_p)] \ll_Q 1\,. \tag{A.1}$$

This (favorable) situation can occur only if p divides $\operatorname{disc}(Q)$.

We suppose now that Q is isotropic over $\mathbf{Q}_p$ for some prime $p \mid 2\operatorname{disc}(Q)$, we will reduce the problem of bounding $N(L_p)$ to the case where the integral quadratic form is given by $Q(x, y, z) = xy + z^2$. We note that $\operatorname{disc}(xy + z^2) = 2$. This reduction comes with the cost that we also have to replace q by a different quadratic form $q' = up^{m_p} q$ with $u \in \mathbf{Z}_p^*$ and $m_p \ge 0$. However, we only have to make this change for $p \mid 2\operatorname{disc}(Q)$ and m_p will only depend on Q. Using these facts we will see in Subsection A.7 that the bound for the number of representations of q' by $xy + z^2$ will suffice for the proof of Proposition 3.4.

We claim that there exists a basis of $\mathbf{Q}_p^3$ over $\mathbf{Q}_p$ so that the quadratic form Q with respect to the coordinates of this basis has the form $up^{-\ell}(xy + z^2)$ for some $u \in \mathbf{Z}_p^*$ and $\ell \in \{0, 1\}$. Indeed as Q is isotropic, there exists a hyperbolic plane in $\mathbf{Q}_p^3$. Complementing the basis of the hyperbolic plane with a vector of the orthogonal complement we arrive at a basis so that Q has the form $xy + up^{-\ell} z^2$ with $u \in \mathbf{Z}_p^*$ and $\ell \in \mathbf{Z}$. If necessary we may replace the last basis vector by a multiple and can ensure that $\ell \in \{0, 1\}$. Similarly we may divide the first basis vector by $up^{-\ell}$ and arrive at the claim.

Let Λ be the $\mathbf{Z}_p$-lattice in $\mathbf{Q}_p^3$ spanned by the above basis. There exists some k (depending only on Λ) so that $p^k \mathbf{Z}_p^3 \subset \Lambda$. Let $\iota\colon (\mathbf{Z}_p^2, q) \to (\mathbf{Z}_p^3, Q)$

be an embedding of q. Then $p^k\iota\colon (\mathbf{Z}_p^2, p^{2k}q) \to (\Lambda, Q)$ and finally

$$p^k\iota\colon (\mathbf{Z}_p^2, u^{-1}p^{2k+\ell}q) \to (\Lambda, u^{-1}p^{\ell}Q) \simeq (\mathbf{Z}_p^3, xy+z^2)$$

are also embeddings of quadratic lattices. We set $m_p = 2k+\ell$ and $q' = u^{-1}p^{m_p}q$ and obtain that there is an injection from the set of embeddings $\iota\colon (\mathbf{Z}_p^2, q) \to (\mathbf{Z}_p^3, Q)$ to the embeddings $\iota'\colon (\mathbf{Z}_p^2, q') \to (\mathbf{Z}_p^3, xy+z^2)$.

A.3 The case of an unramified lattice

The previous section reduces the proof of Proposition 3.4 to the problem of finding an upper bound for $N(L_p)$ where we may assume that either $p \nmid 2\operatorname{disc}(Q)$ or that $Q(x,y,z) = xy+z^2$. This will be done in the following two local counting lemmas which depend on whether $p=2$ or $p>2$:

Recall that for $p>2$ any quadratic form q on some rank two $\mathbf{Z}_p$-lattice L taking value in $\mathbf{Z}_p$ may be written, in a suitable basis, in the form

$$q(xe_1+ye_2) = up^a x^2 + vp^b y^2, \quad u,v \in \mathbf{Z}_p^\times,\ 0 \le a \le b \in \mathbf{Z}_{\ge 0}\,. \tag{A.2}$$

To see this take an element $e_1 \in L$ such that the valuation of $q(e_1)$ is minimal and then take the orthogonal complement of e_1, cf. [7, Sect. 8.3]. We shall call the integers $a \le b$ the *invariants* of the quadratic form (e.g. the invariants of x^2+5y^2 over $\mathbf{Z}_5$ are $(0,1)$). This is a kind of p-adic analogue of the notion of successive minima. The invariants determine the quadratic form over $\mathbf{Z}_p$ — up to isometry — up to $O(1)$ possibilities. We will prove the following lemma.

LEMMA A.3. *Let $p>2$, let Q be an isotropic quadratic form over $\mathbf{Q}_p^3$ so that $p \nmid \operatorname{disc}(Q)$. Let $L \subset \Lambda$ be a rank two sublattice such that $Q_{|L}$ has invariants (a,b), then*

$$N(L;\Lambda) := |\{\Lambda' \in \mathrm{SO}_Q(\mathbf{Q}_p).\Lambda : L \subset \Lambda'\}| \ll (b+1)^2 p^{\lfloor a/2 \rfloor}\,,$$

where the implied constant is absolute. Moreover, if $(a,b) = (0,0)$, $N(L;\Lambda) = 1$.

In the 2-adic case, any quadratic form q on some rank two $\mathbf{Z}_2$-lattice L taking value in $\mathbf{Z}_2$ may be written, in a suitable basis either (cf. [16, Lemma 2.1] and [7, Sect. 8.4]) in the form

$$q(xe_1+ye_2) = u2^a x^2 + v2^b y^2, \quad u,v \in \mathbf{Z}_2^\times,\ 0 \le a \le b \in \mathbf{Z}_{\ge 0}\,, \tag{A.3}$$

or in the form

$$q(xe_1+ye_2) = u2^b x^2 + w2^a xy + v2^b y^2, \quad u,v,w \in \mathbf{Z}_2^\times,\ 0 \le a \le b \in \mathbf{Z}_{\ge 0}\,. \tag{A.4}$$

In both cases we will refer to $a \leq b$ once more as *the invariants of* q. We have the following lemma.

LEMMA A.4. *Consider* $Q(x,y,z) = xy + z^2$ *as a quadratic form over* $\mathbf{Q}_2^3$, *let* $\Lambda \subset \mathbf{Q}_2^3$ *be a lattice satisfying* $Q(\Lambda) \subset \mathbf{Z}_2$ *and which is maximal for this property. Let* $L \subset \Lambda$ *be a rank* 2*-sublattice such that* $Q_{|L}$ *has invariants* (a,b), *then*

$$N(L;\Lambda) \ll (b+1)^2 2^{\lfloor a/2 \rfloor},$$

where the implied constant is absolute.

The proof of these two lemmas will use a geometric interpretation of the quotient $\mathrm{SO}_Q(\mathbf{Q}_p^3)/\mathrm{SO}_Q(\Lambda)$.

A.4 THE BRUHAT-TITS TREE

Let Q be an isotropic quadratic form such that $p \nmid \mathrm{disc}(Q)$ or $Q(x,y,z) = xy + z^2$. Note that $\Lambda_0 = \mathbf{Z}_p^3$ has the property that $Q(\Lambda_0) \subset \mathbf{Z}_p$ and that Λ_0 is maximal for this property. We set

$$\mathcal{T}_Q = \mathrm{SO}_Q(\mathbf{Q}_p^3)\Lambda_0 \simeq \mathrm{SO}_Q(\mathbf{Q}_p^3)/K_p\,.$$

Even though this will not be used here, let us also mention that $\mathcal{T}_Q$ is the set of all lattices Λ in $\mathbf{Q}_p^3$ such that

$$Q(\Lambda) \subset \mathbf{Z}_p$$

and which are maximal for this property (see [15, Cor. 4.17]).

We will need that $\mathcal{T}_Q$ has the structure of a $(p+1)$-regular tree (see [6]) in which Λ, Λ' are *adjacent* if and only if $\Lambda \cap \Lambda'$ has index p in Λ (or equivalently in Λ'). More generally, the *distance* $d(\Lambda,\Lambda')$ between two vertices Λ, Λ' satisfies

$$p^{d(\Lambda,\Lambda')} = [\Lambda : \Lambda \cap \Lambda'] = [\Lambda' : \Lambda \cap \Lambda']\,,$$

and the geodesic between Λ and Λ' consists of all $\Lambda'' \in \mathcal{T}_Q$ satisfying $\Lambda \cap \Lambda' \subset \Lambda''$.

Let us describe the adjacency structure on $\mathcal{T}_Q$ more explicitly using the quadratic structure. Given any lattice $\Lambda \in \mathcal{T}_Q$, and any *primitive* $\mathbf{v} \in \Lambda$ (i.e. $\mathbf{v} \notin p\Lambda$) for which $\overline{\mathbf{v}} = \mathbf{v} + p\Lambda \in \Lambda/(p\Lambda)$ is isotropic over $\mathbf{F}_p$ (i.e. $p \mid Q(\mathbf{v})$) we can define a lattice $\Lambda_{\overline{\mathbf{v}}} \in \mathcal{T}_Q$, which only depends on the line through $\overline{\mathbf{v}}$, as follows. Since

$$Q(a\mathbf{v} + \mathbf{z}) = a^2 Q(\mathbf{v}) + Q(\mathbf{z}) + a\langle \mathbf{z}, \mathbf{v}\rangle \in \mathbf{Z}_p \tag{A.5}$$

and since the linear form $\langle \cdot, \overline{\mathbf{v}} \rangle$ is not zero (even for $p = 2$), we may modify $\mathbf{v}$ by some element $p\mathbf{z}_0 \in p\Lambda$ to ensure that $p^2 \mid Q(\mathbf{v}+p\mathbf{z}_0)$. Here the element $\mathbf{z}_0$ is uniquely determined by $\mathbf{v}$ up to $\{\mathbf{z} \in \Lambda : \langle \mathbf{z}, \mathbf{v} \rangle \equiv 0 \mod p\}$. Therefore, the lattice

$$\Lambda_{\overline{\mathbf{v}}} := \frac{1}{p}\mathbf{Z}_p(\mathbf{v} + p\mathbf{z}_0) + \{\mathbf{z} \in \Lambda : \langle \mathbf{z}, \mathbf{v} \rangle \equiv 0 \mod p\}$$

depends only on $\overline{\mathbf{v}}$, indeed only on the line through $\overline{\mathbf{v}}$. Using (A.5) we see quickly that $Q(\Lambda_{\overline{\mathbf{v}}}) \subset \mathbf{Z}_p$. Below we will always assume that $p^2 \mid Q(\mathbf{v})$ and set $\mathbf{z}_0 = 0$.

Under our assumptions on Q this lattice $\Lambda_{\overline{\mathbf{v}}} \in \mathcal{T}_Q$ is a neighbor of Λ, and there are exactly $p + 1 = |\mathbf{P}^1(\mathbf{F}_p)|$ such lines, and thus every neighbor arises.

We will use also the following simple facts:

(1) For an isotropic $\overline{\mathbf{v}}$ we have

$$\Lambda \cap \Lambda_{\overline{\mathbf{v}}} = \mathbf{Z}_p\mathbf{v} + \{\mathbf{z} \in \Lambda : \langle \mathbf{v}, \mathbf{z} \rangle \equiv 0 \mod p\}.$$

(2) For $\overline{\mathbf{v}}, \overline{\mathbf{v}}'$ generating distinct isotropic lines the intersection

$$\Lambda_{\overline{\mathbf{v}}} \cap \Lambda_{\overline{\mathbf{v}}'} = \{\mathbf{z} \in \Lambda : \langle \mathbf{v}, \mathbf{z} \rangle \equiv \langle \mathbf{v}', \mathbf{z} \rangle \equiv 0 \mod p\} = \mathbf{Z}_p\mathbf{w} + p\Lambda$$

is the preimage in Λ of the orthogonal subspace $(\mathbf{F}_p\overline{\mathbf{v}} + \mathbf{F}_p\overline{\mathbf{v}}')^{\perp} \subset \mathbf{F}_p^3$.

(3) Given three isotropic vectors $\overline{\mathbf{v}}, \overline{\mathbf{v}}', \overline{\mathbf{v}}''$ generating distinct lines and assuming $p > 2$ we have

$$\Lambda_{\overline{\mathbf{v}}} \cap \Lambda_{\overline{\mathbf{v}}'} \cap \Lambda_{\overline{\mathbf{v}}''} = p\Lambda.$$

One establishes also the following generalization:

PROPOSITION A.5. *Let Λ lie at the mid-point of the geodesic between Λ' and Λ'' (i.e. there is $n \geq 1$ such that $d(\Lambda, \Lambda') = d(\Lambda, \Lambda'') = n$, $d(\Lambda', \Lambda'') = 2n$). There exists a primitive $\mathbf{v} \in \Lambda$ so that $Q(\mathbf{v}) \equiv 0(p^n)$ and $\mathbf{w} \in \Lambda$ with $Q(\mathbf{w}) \not\equiv 0(p)$ and $\langle \mathbf{v}, \mathbf{w} \rangle \equiv 0(p^n)$ so that*

$$\Lambda \cap \Lambda' = \{\mathbf{z} \in \Lambda : \ \langle \mathbf{z}, \mathbf{v} \rangle \equiv 0(p^n)\} = \mathbf{Z}_p\mathbf{v} + \mathbf{Z}_p\mathbf{w} + p^n\Lambda$$

and

$$\Lambda' \cap \Lambda'' = \mathbf{Z}_p\mathbf{w} + p^n\Lambda$$

is the preimage of the non-isotropic line defined by w under the projection $\Lambda \mapsto \Lambda/p^n\Lambda$. Moreover, for $m \leq n$, let Λ'_m be the lattice on the segment $[\Lambda, \Lambda']$ at distance m from Λ, then

$$\Lambda \cap \Lambda'_m = \{\mathbf{z} \in \Lambda : \ \langle \mathbf{z}, \mathbf{v} \rangle \equiv 0(p^m)\} = \mathbf{Z}_p\mathbf{v} + \mathbf{Z}_p\mathbf{w} + p^m\Lambda \supset \Lambda \cap \Lambda'.$$

A.5 Proof of Lemma A.3

Let $p > 2$ and Q be as in the lemma. Define

$$\mathcal{R}(L) := \{\Lambda \in \mathcal{T}_Q \,:\, L \subset \Lambda\} \subset \mathcal{T}_Q\,, \qquad N(L) = |\mathcal{R}(L)|\,.$$

In the notation of Lemma A.3, $N(L) = N(L;\Lambda)$ for any $\Lambda \in \mathcal{T}_Q$.

We start by remarking that $\mathcal{R}(L)$ is connected: if Λ, Λ' both contain L, then $L \subset \Lambda \cap \Lambda' \subset \Lambda''$ for any Λ'' on the geodesic path between Λ and Λ'.

Let q be as in (A.2). Suppose $\mathcal{R}(L)$ is non-empty and let $\iota\colon (\mathbf{Z}_p^2, q) \hookrightarrow (\Lambda, Q)$ be an isometric embedding with image $L = \iota(\mathbf{Z}_p^2)$ and let $e_1 = \iota(1,0)$, $e_2 = \iota(0,1)$ so

$$Q(e_1) = up^a, \quad Q(e_2) = vp^b, \quad \langle e_1, e_2\rangle = 0\,.$$

A.5.1 The case $(a,b) = (0,0)$. We show $\mathcal{R}(L) = \{\Lambda\}$. If not, L is also contained in a neighbor $\Lambda_{\overline{\mathbf{v}}}$ of Λ. However, the induced quadratic form on the span of $\bar{e}_1, \bar{e}_2$ is nondegenerate, so this span cannot be $\overline{\mathbf{v}}^{\perp}$ for an isotropic $\overline{\mathbf{v}} \in \Lambda/p\Lambda$. So $N(L) = 1$.

A.5.2 The case $a = 0$, $b \geq 1$. Suppose that $N(L) > 1$. Then there is an isotropic $\overline{\mathbf{v}}$ so that $\overline{e_1}$ belongs to $\overline{\mathbf{v}}^{\perp}$. This shows that $e_1^{\perp}$ is a hyperbolic plane (first modulo p, and then since $p \neq 2$ also on $\mathbf{Q}_p^3$).

In other words, $e_1^{\perp} \cap \Lambda$ is a rank two lattice generated by two isotropic vectors $\mathbf{v}, \mathbf{v}'$ (which are liftings of isotropic vectors generating $\overline{e_1}^{\perp}$) and then, there are exactly two neighboring lattices containing e_1, namely $\Lambda_{\overline{\mathbf{v}}}$ and $\Lambda_{\overline{\mathbf{v}}'}$; that there are at most two follows from Fact (A.4). Pursuing this reasoning, we see that the only lattices that can contain e_1 are the lattices

$$\Lambda_n := \mathbf{Z}_p p^{-n}\mathbf{v} + \mathbf{Z}_p e_1 + \mathbf{Z}_p p^n \mathbf{v}', \quad n \in \mathbf{Z}$$

(which is a geodesic in the tree determined by e_1).

Let us now see that for $n > b$, $\Lambda_{\pm 2n}$ does not contain e_2, which will show that $N(L) \leq 4b + 3$. Suppose $e_2 \in \Lambda_n$, then

$$e_2 \in \Lambda \cap \Lambda_{2n} = \mathbf{Z}_p e_1 + p^n \Lambda_n$$

write $e_2 = \alpha e_1 + \mathbf{z}$, $\alpha \in \mathbf{Z}_p$, $\mathbf{z} \in p^n\Lambda_n$ we obtain

$$\langle e_1, e_2\rangle = 0 \equiv \alpha \pmod{p^n}\,, \quad Q(e_2) = vp^b \equiv \alpha^2 \equiv 0 \pmod{p^n}\,.$$

This is a contradiction for $n > b$.

A.5.3 THE CASE $a = 1$. We show $N(L) \le 2$: Suppose that $L \subset \Lambda_{\overline{\mathbf{v}}}$ for some $\overline{\mathbf{v}}$. Since $\overline{e}_1 \in \Lambda/p\Lambda$ is a non-zero isotropic vector contained in $\overline{\mathbf{v}}^{\perp}$ it has to be a multiple of $\overline{\mathbf{v}}$. By symmetry between Λ and $\Lambda_{\overline{\mathbf{v}}}$, this also shows that Λ is the only neighbor of $\Lambda_{\overline{\mathbf{v}}}$ which contains L. Since $\mathcal{R}(L)$ is a connected subset of the tree, this shows that $N(L) \le 2$ as claimed.

A.5.4 THE CASE $a \ge 2$. Let

$$L_1 := \mathbf{Z}_p e_1' + \mathbf{Z}_p e_2, \quad L_2 := \mathbf{Z}_p e_1 + \mathbf{Z}_p e_2', \quad e_i' = e_i/p, \quad i = 1,2, \quad L_1 + L_2 = \frac{1}{p} L;$$

these are rank two lattices containing L, on which Q is $\mathbf{Z}_p$-valued with respective invariants $(a-2,b)$, $(a,b-2)$ and $(a-2,b-2)$. We will show that either $N(L) = 1$ or

$$\mathcal{R}(L) \subset \mathcal{R}(L_1) \cup \mathcal{R}(L_2) \cup \bigcup_{\Lambda' \in \mathcal{R}(\frac{1}{p}L)} B(\Lambda', 1)\,, \tag{A.6}$$

where $B(\Lambda', d) = \{\Lambda'' \in \mathcal{T}_Q,\ d(\Lambda', \Lambda'') \le d\}$ is the ball in the tree of radius d centered at Λ'; it has cardinality $1 + (p+1)\frac{p^d - 1}{p-1} \le (1 + \frac{3}{p})p^d$.

Here is the proof of (A.6). Let $\Lambda \in \mathcal{R}(L)$. If $e_1 \in p\Lambda$ or $e_2 \in p\Lambda$, then $\Lambda \in \mathcal{R}(L_1) \cup \mathcal{R}(L_2)$. So suppose now $e_1, e_2 \in \Lambda$ are both primitive vectors. By assumption, we have for $i = 1,2$ (since $Q(e_i) \equiv 0 \pmod p$) that $\overline{e_i}$ is a non-zero isotropic vector. Since $\langle e_1, e_2 \rangle = 0$, $\overline{e_1}$ and $\overline{e_2}$ have to be co-linear; otherwise the induced form on the reduction $\overline{\Lambda}$ would be identically zero on a plane. Now $\Lambda_{\overline{e}_1}$ contains both L_1 and L_2; so it belongs to $\mathcal{R}(\frac{1}{p}L)$. Thus Λ is at distance at most 1 from $\mathcal{R}(\frac{1}{p}L)$.

Let us now see how to conclude the proof of Lemma A.3: for $r, s \in \mathbf{N}$, let

$$L_{r,s} := \mathbf{Z}_p p^{-r} e_1 + \mathbf{Z}_p p^{-s} e_2\,.$$

Q takes integral values on $L_{r,s}$ for $r \le \lfloor a/2 \rfloor$, $s \le \lfloor b/2 \rfloor$. In this notation (A.6) states

$$\mathcal{R}(L_{0,0}) \subset \mathcal{R}(L_{1,0}) \cup \mathcal{R}(L_{0,1}) \cup \bigcup_{\Lambda' \in \mathcal{R}(L_{1,1})} B(\Lambda', 1)\,.$$

We can now apply (A.6) again to each of the terms on the right. With each application r or s or both increase by 1. In the latter case we obtain that the previous lattice $\Lambda' \in \mathcal{R}(L_{r,s})$ (to which (A.6) was applied) is at distance 1 from the new lattice $\Lambda'' \in \mathcal{R}(L_{r+1,s+1})$. Also note that in the latter case both a and b are reduced by 2, so that this case can only happen $\le \lfloor a/2 \rfloor$ many

times. Therefore, induction on $a+b$ shows that

$$\mathcal{R}(L) = \mathcal{R}(L_{0,0}) \subset \bigcup \big\{ B(L_{\lfloor a/2 \rfloor, s}, \lfloor a/2 \rfloor), B(L_{r, \lfloor b/2 \rfloor}, \lfloor a/2 \rfloor) : \ 0 \le r, s \le \lfloor b/2 \rfloor \big\}.$$

Each $L' = L_{\lfloor a/2 \rfloor, s}$ resp. $L' = L_{r, \lfloor b/2 \rfloor}$ has invariants $(0, b')$ or $(1, b')$ with $b' \le b$ and by the previous sections $N(L') = O(b+1)$ in all cases. Consequently

$$N(L) \ll \sum_{L'} \sum_{\Lambda' \in \mathcal{R}(L')} |B(\Lambda', \lfloor a/2 \rfloor)| \ll (b+1)^2 p^{\lfloor a/2 \rfloor}. \qquad \square$$

A.6 Proof of Lemma A.4

Recall that we assume that $Q(x,y,z) = xy + z^2$. Note that $(1,0,0), (0,1,0)$ and $(-1,1,1)$ are three isotropic vectors that are linearly independent modulo 2, which define the neighbors of $\mathbf{Z}_2^3$. For every pair f_1, f_2 of these vectors we can find a third vector $f_3 \in \mathbf{Z}_2^3$ so that $Q(xf_1 + yf_2 + zf_3) = xy + z^2$. Of the four non-zero non-isotropic vectors modulo 2 the vector $k = (0,0,1)$ is special, it is the only element in the kernel of $\langle \cdot, \cdot \rangle$ modulo 2 and also satisfies $k \equiv f_3$ modulo 2 for any basis (f_1, f_2, f_3) as above. Below we will always use the letter $\overline{k}$ to denote the corresponding element in the lattice $\Lambda / 2\Lambda$.

A.6.1 The diagonal case (A.3). Suppose that in a suitable basis q takes the form (A.3). This situation is similar to the proof of Lemma A.3. We only discuss the details where the two proofs differ.

A.6.2 The case $(a,b) = (0,0)$. We claim that $\Lambda \in \mathcal{R}(L)$ has at most one neighbor in $\mathcal{R}(L)$. If one of $\overline{e_1}$ or $\overline{e_2}$ is not equal to $\overline{k}$, then we claim that $\mathcal{R}(L)$ contains at most one neighbor of Λ. To see this suppose $\overline{e_1} \neq \overline{k}$ and $L \subset \Lambda_{\overline{v}} \cap \Lambda_{\overline{v}'}$. Then by Fact (2), L is contained modulo 2 in the common kernel of $\langle \cdot, \overline{v} \rangle$ and $\langle \cdot, \overline{v}' \rangle$, which is one-dimensional and actually equal to the span of $\overline{k}$ — a contradiction. Therefore, $L \subset \Lambda \cap \Lambda_{\overline{v}}$ for at most one neighbor $\Lambda_{\overline{v}}$ as claimed.

So suppose $\overline{e_1} = \overline{e_2} = \overline{k}$ and $w \in \Lambda$ is such that $Q\big(xe_1 + y(e_1 + 2w)\big) = ux^2 + vy^2$ as in (A.3). Since we also have

$$Q\big(xe_1 + y(e_1 + 2w)\big) = x^2 Q(e_1) + y^2 Q(e_1 + 2w) + xy(2Q(e_1) + 2\langle e_1, w \rangle)$$

and $2 \mid \langle e_1, w \rangle$, it follows that $Q\big(xe_1 + y(e_1 + 2w)\big)$ is not as in (A.3). So we have seen that in all possible cases we have at most one neighbor of Λ in $\mathcal{R}(L)$. However, this shows $N(L) \le 2$ for $(a,b) = (0,0)$.

A.6.3 THE CASE $a = 0$ AND $b \geq 1$. We claim that the main difference between the case of $p = 2$ and $p > 2$ lies in this case. Here we will see that $\mathcal{R}(L)$ is only contained in the set of elements at distance one to points on a geodesic. This is caused by the fact that if $\overline{e_1} = \overline{k}$ and $\overline{e_2} = 0$, then $\mathcal{R}(L)$ contains all neighbors of Λ due to Fact (1) and since $\overline{k}$ is orthogonal to all three nonzero isotropic vectors in $\Lambda/2\Lambda$.

On the other hand, we have already seen above (in the case $a = 0, b = 0$) that if $\overline{e_1} \neq \overline{k}$ then only one neighbor of Λ can be in $\mathcal{R}(L)$. To prove that $\mathcal{R}(L)$ consists of points at distance one from a geodesic we only have to show that if $\overline{e_1} = \overline{k}$, then for at least one neighbor Λ' of Λ we have $\overline{e_1} \neq \overline{k'}$ where $\overline{k'} \in \Lambda'/2\Lambda'$ is the corresponding special vector for Λ'. This follows if we can find some vector $w \in \Lambda'$ with $\langle \overline{e_1}, \overline{w} \rangle \neq 0$.

To see this we simplify the notation and assume without loss of generality $\Lambda = \mathbf{Z}_2^3$. Let $e_1 = (\alpha, \beta, \gamma)$ so that $\langle e_1, (1,0,0) \rangle = \beta$, $\langle e_1, (0,1,0) \rangle = \alpha$, and $\langle e_1, (0,0,1) \rangle = 2\gamma$. Since $\overline{e_1} \neq 0$, one quickly sees that one of these inner products is not divisible by 4. Without loss of generality we may assume $4 \nmid \beta$. Now consider the neighbor $\Lambda' = \frac{1}{2}\mathbf{Z}_2 \times 2\mathbf{Z}_2 \times \mathbf{Z}_2$ of Λ. Then $w = (\frac{1}{2}, 0, 0) \in \Lambda'$ satisfies $\langle e_1, w \rangle = \frac{1}{2}\beta \not\equiv 0 \pmod 2$. Hence as claimed, $\overline{e_1} \neq \overline{k'}$ and so only one neighbor of Λ', namely Λ itself, can belong to $\mathcal{R}(L)$.

It follows that there exists a line segment $I \subset \mathcal{R}(L)$ in a geodesic in $\mathcal{T}(Q)$ so that $\mathcal{R}(L) \subset \bigcup_{\Lambda \in I} B(\Lambda, 1)$. Arguing as in Subsection A.5.2 we can bound the length of I in terms of b and obtain $N(L) \leq 3(4b + 3)$.

A.6.4 THE CASE $a \geq 1$. The arguments for $p > 2$ carry over to the remaining cases.

A.6.5 THE NON-DIAGONAL CASE (A.4). So suppose now q is represented by the lattice $L = \mathbf{Z}_2 e_1 + \mathbf{Z}_2 e_2 \subset \Lambda$ with

$$Q(e_1) = u2^b, \quad Q(e_2) = v2^b, \quad \langle e_1, e_2 \rangle = w2^a, \quad u, v, w \in \mathbf{Z}_2^\times, \quad 0 \leq a \leq b\,.$$

A.6.6 THE CASE $a = 0$. If $(a, b) = (0, 0)$, then $\overline{e_1}$ and $\overline{e_2}$ are linearly independent in $\Lambda/2\Lambda$ since otherwise $w = \langle e_1, e_2 \rangle \equiv 0 \pmod 2$. Also note that the plane generated by $\overline{e_1}$ and $\overline{e_2}$ does not contain any isotropic vector. However, this implies that e_1, e_2 cannot be both contained in any $\Lambda_{\overline{\mathbf{v}}}$ for then $\overline{\mathbf{v}}^\perp$ would contain $\overline{e_1}, \overline{e_2}, \overline{\mathbf{v}}$ three linearly independent vectors.

If now $(a, b) = (0, b \geq 1)$, $\overline{e_1}$ and $\overline{e_2}$ are two linearly independent isotropic vectors and so e_1 can only be contained in $\Lambda_{\overline{e_1}}$. Similarly, e_2 is only contained in $\Lambda_{\overline{e_2}}$. So L cannot be contained in any neighbor of Λ.

In conclusion for $a = 0$ we have

$$N(L) = 1\,.$$

A.6.7 THE CASE $a = 1$. In that case at least one of the vectors $\overline{e_1}$ and $\overline{e_2}$ must be a non-zero isotropic vector, for otherwise $a \geq 2$. Suppose $\overline{e_1} \neq 0$. Then $\overline{e_1} \in \Lambda_{\overline{\mathbf{v}}}$ only for $\overline{e_1} = \overline{\mathbf{v}}$. Therefore, L can only have one neighbor in $\mathcal{R}(L)$ and so $N(L) \leq 2$.

A.6.8 THE CASE $a \geq 2$. We consider again the 2 rank two lattices

$$L_1 := \mathbf{Z}_2 e_1' + \mathbf{Z}_2 e_2\,, \quad L_2 := \mathbf{Z}_2 e_1 + \mathbf{Z}_2 e_2'\,, \quad e_i' = e_i/2$$

which contain L and on which Q is $\mathbf{Z}_2$-valued:

$$Q(e_1') = u2^{b-2}, \quad Q(e_2') = v2^{b-2}, \quad \langle e_1', e_2\rangle = \langle e_1, e_2'\rangle = w2^{a-1}\,.$$

We describe now the type and the invariants of L_1 — by symmetry L_2 behaves the same way.

If $a = b$ we may solve the equation in $\beta \in \mathbf{Z}_2^\times$

$$0 = \langle e_2 + \beta e_1', e_1'\rangle = w2^{a-1} + \beta u 2^{b-1}$$

and so $Q_{|L_1}$ is of diagonal form (A.3) in the basis $\{e_2 + \beta e_1', e_1'\}$. Furthermore, since

$$\langle e_2 + \beta e_1', e_2 + \beta e_1'\rangle = 2Q(e_2 + \beta e_1') = v2^{b+1} + \beta w 2^{a-1}$$

it has invariants $(a-2, b-2)$.

If $a < b$, take $\beta = 2^{b-a}$: in the basis $\{e_2 + \beta e_1', e_1'\}$, $Q_{|L_1}$ takes the non-diagonal form (A.4) with $(a', b') = (a-1, b-2)$. Finally $Q_{|L_1+L_2} = Q_{|L/2}$ takes the form (A.4) with $(a'', b'') = (a-2, b-2)$.

We then conclude exactly as in Subsection A.5.4 by proving that either $N(L) = 1$ or (A.6) holds. This implies once more the desired bound.

A.7 PROOF OF PROPOSITION 3.4

We now show how the previous subsections combine to the proof of Proposition 3.4.

Recall that we are bounding the number of representations $N(L)$ of the quadratic form $q(x, y) = a_1 x^2 + a_2 xy + a_3 y^2$ by the ternary quadratic form Q up to $\mathrm{SO}_Q(\mathbf{Z})$. For any p let us write a_p and b_p for the invariants of q over $\mathbf{Z}_p$ as in Section A.3. Let $f^2 | \gcd(a_1, a_2, a_3)$ be the greatest common square divisor of the coefficients of q. Then $a = v_p(f)$.

By the discussion in Sections A.1-A.2 we know that

$$N(L) \ll \prod_{\substack{p \\ Q \ p\text{-isotropic}}} N(L_p)\,.$$

Also recall from Section A.2 that for bounding $N(L_p)$ for $p|\operatorname{disc}(Q)$ we may replace Q by $xy+z^2$ and q by a fixed multiple q' of q, where the factor only depends on Q. From this we see that Lemmas A.3-A.4 also hold for $p|\operatorname{disc}(Q)$ for q and Q, except that the implicit constant depends for those primes also on Q.

Notice that for any prime $p>2$ we have $a_p+b_p=v_p(\operatorname{disc}(q))$ and $a_p=v_p(\gcd(a_1,a_2,a_3))$. For $p=2$ we have $v_2(\operatorname{disc}(q))=a+b+2$ in the diagonal case and $v_2(\operatorname{disc}(q))=2a$ in the non-diagonal case. Also let $c\geq 1$ be the implied constant in Lemma A.3. Together with Lemmas A.3-A.4 this gives

$$N(L) \ll \prod_{p|2\operatorname{disc}(q)} c\big(v_p(\operatorname{disc}(q))+1\big)^2 p^{v_p(f)} \ll_\epsilon f \max(a_1,a_2,a_3)^\epsilon\,,$$

as desired. □

B. Entropy, Bowen balls and uniqueness of measure of maximal entropy

B.1 Statement of main results

We recall some notations: we work in the space $X=\Gamma\backslash G$ with $G=\mathrm{SL}_2(\mathbf{R})$, and let T denote the time-one-map of the geodesic flow, i.e. the map

$$T\colon x\mapsto xa \qquad \text{with} \qquad a=\begin{pmatrix} e^{1/2} & 0 \\ 0 & e^{-1/2}\end{pmatrix}.$$

We define a *Bowen (N,η)-ball* in this space to be any set of the form $xB_{N,\eta}$ with $x\in X$ and

$$B_{N,\eta}=\bigcap_{n=-N}^{N} a^{-n}B_\eta^G(e)a^n$$

(in the sections above η remained fixed and was omitted from the notations, but here it will be convenient to be able to use Bowen balls of varying η).

If Γ is cocompact, for all η sufficiently small, the Bowen (N,η)-ball $xB_{N,\eta}$ coincides with the set

$$xB_{N,\eta}=\{y : d(T^n(x),T^n(y))<\eta \quad \text{for all } -N\leq n\leq N\}\,.$$

This is *not* true any more for noncompact quotients, where in general the right-hand side can be significantly bigger than the left-hand side which is the source of some complications.

The following theorem was proved for compact quotients by Bowen in [4]. It is certainly well known also in the finite volume case, and proofs using leafwise measures can be found e.g. [20, Prop. 9.6] and the more recent lecture notes [12, Thm. 7.9].

THEOREM B.1. *Let $X = \Gamma\backslash \mathrm{SL}_2(\mathbf{R})$ and $T: X \to X$ be as above. Then for any T-invariant probability measure ν the entropy satisfies $h_\nu(T) \leq 1$. Moreover, equality holds if and only if $\nu = \mu_X$ is the $\mathrm{SL}_2(\mathbf{R})$-invariant probability measure on X.*

We give here a direct proof not using leafwise measures, based on Lemma B.2 (which is identical to Lemma 5.3 and was needed for the proofs in §4), in the spirit of Bowen's proof (that in turn was inspired by a proof by Adler and Weiss [1] of the uniqueness of measure of maximal entropy in *irreducible shifts of finite type*).

LEMMA B.2. *Let μ be an A-invariant measure on $X = \Gamma\backslash \mathrm{SL}(2,\mathbf{R})$. Fix $\eta > 0$ and $\epsilon \in (0,1)$. For any $N \geq 1$ we let $BC_\eta(N,\epsilon)$ be the minimal number of Bowen (N,η)-balls needed to cover any subset of X of measure bigger than $1-\epsilon$. Then*

$$h_\mu(T) \leq \lim_{\epsilon\to 0} \liminf_{N\to\infty} \frac{\log BC_\eta(N,\epsilon)}{2N}. \tag{B.1}$$

It is easy to see that for any $\eta, \eta' > 0$ a Bowen (N,η)-ball can be covered by $\ll 1$ Bowen (N,η')-balls. Therefore,

$$\liminf_{N\to\infty} \log BC_\eta(N,\epsilon)/2N \tag{B.2}$$

is independent of η. One can show that if μ is ergodic, equality holds in (B.1), and moreover that the quantity in (B.2) is independent of ϵ. If μ is not ergodic, then in general equality in (B.1) fails: in this case $h_\mu(T)$ is the average of the entropy of the ergodic components of μ and the right-hand side of (B.1) gives the essential supremum of the entropies of the ergodic components of μ. We shall not need either fact (nor will we prove them), but will use the following related estimates for μ ergodic:

LEMMA B.3. *Assume that μ is in addition ergodic for T. Then for any sufficiently small η (depending only on X) and for any $\epsilon \in (0,1)$ and any large enough N (depending on μ, ϵ), for any $\epsilon_1 \in (0,\epsilon)$, if k is sufficiently large (depending on $\epsilon_1, \epsilon, N, \mu, \eta$) then*

$$\log BC_\eta(kN, \epsilon_1) \le k(1-2\epsilon)\log BC_\eta(N,\epsilon) + 4\epsilon Nk + qk\,.$$

Here q is some absolute constant.

For our proof of Theorem B.1 it is crucial that the second error term (qk) does *not* depend on N. Roughly speaking the lemma says, if we manage to cover some set of measure bigger than $1-\epsilon$ by relatively few Bowen (N,η)-balls, then a set of size $1-\epsilon'$ can also be covered by relatively few Bowen (Nk,η)-balls.

The reader may wish to look now at the proof of Theorem B.1 in Subsection B.4 to see how the above two lemmas are used in combination to imply the uniqueness of the measure of maximal entropy.

B.2 PROOF OF LEMMA B.2

In the proof we will need the notion of relative entropy for partitions: For two partitions $\mathcal{P} = \{S_1, \dots, S_\ell\}$ and $\mathcal{Q} = \{Q_1, \dots, Q_m\}$ of a probability space (X, μ) the *relative entropy of $\mathcal{P}$ given $\mathcal{Q}$* is defined by

$$H_\mu(\mathcal{P}|\mathcal{Q}) = -\sum_{i,j} \mu(S_i \cap Q_j) \log \frac{\mu(S_i \cap Q_j)}{\mu(Q_j)}\,,$$

and it is easy to see that it gives the following *additivity of entropy*

$$H_\mu(\mathcal{P} \vee \mathcal{Q}) = H_\mu(\mathcal{Q}) + H_\mu(\mathcal{P}|\mathcal{Q})\,. \tag{B.3}$$

We should also use the notation $\mathcal{P}(x)$ to denote the elements of the finite or countable partition $\mathcal{P}$ containing x.

Proof. Let $\mathcal{P} = \{Q, S_1, \dots, S_\ell\}$ be a finite partition where Q is the only unbounded set, all boundaries ∂S_i are null sets which satisfy additionally

$$\mu((\partial S_i)B_\kappa^G) < C\kappa$$

for some constant $C > 0$ and all $\kappa > 0$, and finally $h_\mu(T,\mathcal{P}) > h_\mu(T) - \delta$. Here

$$h_\mu(T,\mathcal{P}) = \lim_{N\to\infty} \frac{H_\mu(\mathcal{P}^{[-N,N]})}{2N+1}$$

is the expression over which one needs to takes the supremum to define $h_\mu(T)$. Such a partition exists since (i) by the general theory of entropy $h_\mu(T)$ can be

approximated by $h_\mu(T,\mathcal{P})$ once $\mathcal{P}$ is a sufficiently fine partition, and (ii) one can find for every $x \in X$ arbitrary small $r > 0$ for which $\mu\big((\partial B_r(x))B_\kappa^G\big) < C\kappa$ for all $\kappa > 0$ (since for every x the function $r \mapsto \mu(B_r(x))$ is monotone increasing hence differentiable for a.e. r).

We claim that for most points $x \in X$ (we shall quantify this presently) it holds that

$$\mathcal{P}^{[-N,N]}(x) \supset xB_{N,2\eta'} \qquad \text{for } \eta' = \eta N^{-2}, \tag{B.4}$$

hence if $y \in xB_{N,\eta'}$, then $yB_{N,\eta'} \subset \mathcal{P}^{[-N,N]}(x)$. To show this, suppose $y = xh \notin \mathcal{P}^{[-N,N]}(x)$ for $h \in B_{N,\eta'}$. Then for some n with $|n| \le N$ the elements

$$xa^n \quad \text{and} \quad xha^n$$

belong to different elements of $\mathcal{P}$. It follows that at least one of the elements xa^n belong to $(\partial P)B_{2\eta'}^G$ for some $P \in \mathcal{P}$, $|n| \le N$. Therefore, x belongs to

$$\bigcup_{n=-N}^{N} T^n \bigcup_{S\in\mathcal{P}} (\partial S)B_{2\eta'}^G \tag{B.5}$$

which has measure less than $2(2N+1)\ell C\eta N^{-2} \ll N^{-1}$. This proves the above claim.

Roughly speaking $B_{N,\eta}$ has length η in the direction of A and ηe^{-N} along stable and unstable horocycle directions while $B_{N,\eta'}$ has ηN^{-2} and $\eta N^{-2}e^{-N}$ instead. From this one can easily deduce that one needs at most $\ll N^6$ many translates of $B_{N,\eta'}$ to cover $B_{N,\eta}$. Choose $f > \lim_{\epsilon\to 0} \liminf_{N\to\infty} \frac{\log BC(N,\epsilon)}{2N}$. Then for any $\epsilon > 0$, there is some large $N \ge 1$ depending on ϵ such that the measure of the set in (B.5) is less than ϵ, and moreover such that $1-\epsilon$ of the space can be covered by less than e^{2Nf} many translates of the set $B_{N,\eta'}$.

Say $y_1B_{N,\eta'}, \dots, y_kB_{N,\eta'}$ (with $k \le e^{2Nf}$) cover $X_1 \subset X$ with $\mu(X_1) \ge 1-\epsilon$. If $x \in X_1$ is not in the union in (B.5). Since $x \in y_jB_{N,\eta'}$ for some j, it follows from (B.4) that $y_jB_{N,\eta'} \subset \mathcal{P}^{[-N,N]}(y_j)$. In other words, it follows that $1-2\epsilon$ of the space can be covered by e^{2Nf} elements of the partition $\mathcal{P}^{[-N,N]}$.

Let P be the union of these partition elements and let $\mathcal{P} = \{P, X\smallsetminus P\} \subset \sigma(\mathcal{P})$ be the associated partition. Write $\mu_B = (\mu(B))^{-1}\mu|_B$ for the normalized restriction of the measure μ to a Borel set B. It follows now from (B.3) that

$$\begin{aligned} H_\mu(\mathcal{P}^{[-N,N]}) &= H_\mu(\mathcal{P}) + H_\mu(\mathcal{P}^{[-N,N]}|\mathcal{P}) \\ &= H_\mu(\mathcal{P}) + \mu(P)H_{\mu_P}(\mathcal{P}^{[-N,N]}) + \mu(X\smallsetminus P)H_{\mu_{X\smallsetminus P}}(\mathcal{P}^{[-N,N]}) \\ &\le \log 2 + 2Nf + 4\epsilon N\ell \end{aligned}$$

since the entropy of a partition with K elements is at most $\log K$. For $N \to \infty$ this shows that

$$h_\mu(T) - \delta < h_\mu(T, \mathcal{P}) \le f + 2\epsilon\ell\,,$$

which implies the lemma since δ and ϵ were arbitrary. (Note that ℓ depends on δ but not on ϵ.) $\square$

B.3 Proof of Lemma B.3

We shall say a Bowen ball $yB_{N,\eta}$ is *injective* if the map $g \mapsto yg$ is injective on $B_{N,\eta}$. Let $\eta_0 > 0$ be such that $2\eta_0$ is smaller than the length of any closed geodesic in X. An easy compactness argument shows that if $\eta \le \eta_0$ for any compact $F \subset X$ there is a N_0 so that if $N > N_0$ and $y \in F$ the Bowen ball $yB_{N,\eta}$ is injective. In the proof we shall also make use of shifted $(s,t;\eta)$-Bowen balls — sets of the form $yB_{s,t;\eta}$ where $B_{s,t;\eta} := \bigcap_{i=s}^{t} a^i B_\eta^G a^{-i}$ and $(s,t;\eta)$ sub-Bowen balls which are simply sets of the form yB for some $B \subset B_{s,t;\eta}$. A shifted $(s,t;\eta)$-Bowen ball $yB_{s,t;\eta}$ (respectively, a $(s,t;\eta)$ sub-Bowen ball yB) is injective if the map $g \mapsto yg$ is injective on $B_{s,t;\eta}$ (or B). We note the following important properties of shifted Bowen balls:

(Bowen-1) For any $s \le t \le r$, the intersection of an injective $(s,t;\eta)$ sub-Bowen ball with an injective $(t,r;\eta)$ sub-Bowen ball can be covered by at most q injective $(s,r;\eta)$ sub-Bowen balls;

(Bowen-2) For any $s \le t \le r$, an injective $(s,t;\eta)$ sub-Bowen ball can be covered by at most qe^{r-t} injective $(s,r;\eta)$ sub-Bowen balls.

Proof of claims. Both claims can easily be reduced to their special cases where $t = 0$ and where we only consider Bowen balls of the form $gB_{s,r;\eta}$ in G instead of injective sub-Bowen balls in X.

For the proof of (Bowen-1) notice that there exists some $C > 0$ so that

$$g_1 B_{s,0;\eta} \subset g_1 B^{U^+}_{C\eta} B^{U^-}_{C\eta e^s} B^A_{C\eta}\,, \tag{B.6}$$

where B_r^H denotes the r-ball around the identity in a subgroup $H \subset \mathrm{SL}_2(\mathbf{R})$. Similarly,

$$g_2 B_{0,r;\eta} \subset g_2 B^{U^+}_{Ce^{-r}\eta} B^{U^-}_{C\eta} B^A_{C\eta}\,. \tag{B.7}$$

We can now decompose each of the balls appearing on the right-hand side of (B.6)–(B.7) into $\ll 1$ many balls with certain smaller radius and obtain that $g_1 B_{s,0;\eta} \cap g_2 B_{0,r;\eta}$ is the union of $\ll 1$ many sets of the form

$$O = (g_1 u_1^+ B^{U^+}_{\eta/8} u_1^- B^{U^-}_{\eta e^s/8} a_1 B^A_{\eta/8}) \cap (g_2 u_2^+ B^{U^+}_{\eta e^{-r}/8} u_2^- B^{U^-}_{\eta/8} a_2 B^A_{\eta/8})\,,$$

where $u_1^+ \in B^{U^+}_{C\eta}$, $u_2^+ \in B^{U^+}_{C\eta e^{-r}}$, $u_1^- \in B^{U^-}_{C\eta e^s}$, $u_2^- \in B^{U^-}_{C\eta}$, $a_1, a_2 \in B^A_{C\eta}$. If $g \in O$ and η_0 is sufficiently small so that conjugation by an element of distance $C\eta_0$ does not increase the distance to the identity significantly, it follows that $O \subset gB_{(s,r;\eta)}$ which proves the first claim.

The second claim follows similarly by splitting the set $B_{s,0;\eta}$ as in (B.6) into $\ll e^r$ many sets of the form

$$O = g_1 u_1^+ B^{U+}_{\eta e^{-r}/8} u_1^- B^{U^-}_{\eta e^s/8} a_1 B^A_{\eta/8}$$

with $u^+ \in B^{U^+}_{\ll \eta}$ and $u^- \in B^{U^-}_{\ll \eta e^s}$, and showing that for $g \in O$ we have $O \subset gB_{s,r;\eta}$. □

Proof of Lemma B.3. Let $\eta \in (0, \eta_0)$ where η_0 is as defined above, and let M be sufficiently large so that $\mu(X_{\le M}) > 1 - \epsilon/2$ and similarly choose M_1 so that $\mu(X_{\le M_1}) > 1 - \epsilon_1/2$. We require that N be sufficiently large so that any (N,η)-Bowen ball $yB_{N,\eta}$ intersecting $X_{\le M}$ is injective, and we choose k_1 so that a similar statement holds for any $(k_1 N, \eta)$-Bowen ball intersecting $X_{\le M_1}$.

Let Ξ be a collection of (N,η)-Bowen balls of cardinality $BC_\eta(N,\epsilon)$ covering a subset of X with μ-measure at least $1-\epsilon$. Then

$$\Xi' = \{B \in \Xi : B \cap X_{\le M} \neq \varnothing\}$$

has $\mu\left(\bigcup_{B\in\Xi'} B\right) \ge 1 - \frac{3\epsilon}{2}$. Let $Y = \bigcup_{B\in\Xi'} B$. By the pointwise ergodic theorem, there is a $k_2 \ge k_1$ and a subset $Y_1 \subset X_{\le M_1}$ of μ-measure $\ge 1 - \frac{3\epsilon_1}{4}$ so that points in Y_1 spend most of their time in Y in the following sense:

$$\frac{1}{2n}\sum_{s=-n}^{n-1} 1_Y(T^s(y)) > 1 - 2\epsilon \qquad \text{for all } n \ge k_2 N \text{ and } y \in Y_1. \tag{B.8}$$

To complete the proof of Lemma B.3 we will show that for any $k \ge k_3$ there is a collection Ξ_1 of (kN,η)-Bowen balls covering Y_1 of cardinality

$$|\Xi_1| \ll N(2q)^k BC_\eta(N,\epsilon)^{k(1-2\epsilon)} e^{(4\epsilon k+4)N}.$$

Let c be the implied multiplicative constant. Then for large enough q' (depending only on q and some absolute constants above) we have $cN(2q)^k e^{4N} \le e^{q'k}$ for all sufficiently large k (where the bound is allowed to depend on N). Therefore, the existence of Ξ_1 as above will establish the lemma.

Fix $k \ge k_2$ and let $y \in Y_1$. We partition the finite orbit

$$\{T^{-kN}(y), \dots, T^{kN-1}(y)\}$$

into the $2N$ finite orbits of the form

$$\{T^{-kN+\ell}(y), T^{(-k+2)N+\ell}(y), \dots, T^{(k-2)N+\ell}(y)\}$$

for $\ell \in \{0, \dots, 2N-1\}$. By equation (B.8) there must for any $y \in Y_1$ exist some $\ell(y) \in \{0, \dots, 2N-1\}$ so that

$$\frac{1}{k}\sum_{s=0}^{k-1} 1_Y(T^{(-k+2s)N+\ell(y)}(y)) \geq 1 - 2\epsilon\,.$$

Let $L = \lceil (1-2\epsilon)k \rceil$. It follows that there are $0 \leq t_1 < t_2 \cdots < t_L < k$ with $T^{(-k+2t_i)N+\ell(y)}(y) \in Y$. Furthermore, there exist injective (N,η)-Bowen balls $B_1, \dots, B_L \in \Xi$ so that

$$y \in \bigcap_{i=1}^{L} T^{-(-k+2t_i)N-\ell(y)} B_i\,.$$

Recall that Ξ has $BC_\eta(N,\epsilon)$ many elements. We now apply the properties (Bowen-1) and (Bowen-2), and we conclude that the set of all $y \in Y_1$ with a given value of $\ell(y)$ and $t_1, \dots, t_L$ can be covered by

$$\ll BC_\eta(N,\epsilon)^{k(1-2\epsilon)+1} e^{4Nk\epsilon+2N} q^{k+1}$$

injective (kN,η)-Bowen balls. Since there are at most $2N2^k$ choices of $\ell(y)$ and $t_1, \dots, t_L$ we are done. □

B.4 Proof of Theorem B.1

We begin with the observation that the $\mathrm{SL}(2,\mathbf{R})$-invariant measure μ_X on X achieves the upper bounds stated on the entropy, and moreover is ergodic under T. Let $\nu \neq \mu_X$ be another T-invariant probability measure and without loss of generality we may assume that ν is singular with respect to μ_X (which is the case e.g. if ν is also ergodic), and let η_0 be as in the proof of Lemma B.3.

Let f be a nonnegative, continuous, compactly supported function so that

$$\int f\,d\mu_X < \int_0^1 dt \int f(xa_t)\,d\nu\,, \tag{B.9}$$

R some real number strictly between the left-hand side and right-hand side of (B.9) and set

$$Y_T = \left\{ x : \frac{1}{T}\int_0^T f(xa_t)dt > R \right\}.$$

By construction Y_T is compact, and (for $\epsilon > 0$ arbitrary) by the pointwise ergodic theorem if T is large enough $\mu_X(Y_T) < \epsilon$ and $\nu(Y_T) > 1-\epsilon$. In fact,

if T is large enough, for any sufficiently large N it holds that

$$\mu_X(Y_T B_{N,\eta_0}) < 2\epsilon\,. \tag{B.10}$$

Fix such a T, and choose N so that (B.10) holds and moreover any (N,η_0)-Bowen ball intersecting Y_T is injective.

Now choose a maximal collection of disjoint $(N,\eta_0/2)$-Bowen balls intersecting Y_T. Each of these balls has μ_X-volume $\gg_{\eta_0} e^{-2N}$ (the implicit constant is independent of ϵ and N). In view of (B.10), it follows that the cardinality of this collection is $\ll_{\eta_0} \epsilon e^{2N}$, and by maximality the corresponding collection of (N,η_0)-Bowen balls covers Y_T. As $\nu(Y_T) > 1-\epsilon$ we obtain $BC_{\eta_0}(N,\epsilon,\nu) \ll_{\eta_0} \epsilon e^{2N}$ (note that since we are simultaneously discussing two measures we have added ν to the notation $BC(\cdot)$).

Roughly speaking the above upper bound should lead to $h_\nu(T) < 1$ by using Lemma B.2: most of the space with respect to ν is covered by relatively few, namely $\le C\epsilon e^{2N}$, Bowen (N,η)-balls. However, as (B.1) first takes the limit as $N \to \infty$ this inequality does not directly imply $h_\nu(T) < 1$. To overcome this we introduce an $\epsilon' \in (0,\epsilon)$ and will use Lemma B.3 to obtain the bound on the covering number for ϵ' and kN. Indeed applying Lemma B.3 we conclude that for any $\epsilon' \in (0,\epsilon)$ if k is sufficiently large

$$\begin{aligned}\log BC_{\eta_0}(kN,\epsilon',\nu) &\le k(1-2\epsilon)(2N+\log(C\epsilon)) + 4\epsilon kN + qk\\ &\le k(1-2\epsilon)2N + \frac{1}{2}k\log(C\epsilon) + 4\epsilon kN + qk = 2Nk + \Big(q+\frac{1}{2}\log(C\epsilon)\Big)k\,,\end{aligned}$$

where we also assumed $\epsilon < 1/4$ and $C\epsilon < 1$. Hence we obtain for any $\epsilon' \in (0,\epsilon)$ that

$$\liminf_{k\to\infty} \frac{1}{2kN}\log BC_{\eta_0}(kN,\epsilon',\nu) \le 1 + \frac{2q+\log(C\epsilon)}{4N}\,.$$

However, for sufficiently small ϵ the right-hand side is < 1. Hence by Lemma B.2 we get $h_\nu(T) < 1$. Therefore, m_X is the only probability measure on X with $h_{m_X}(T) \ge 1$.

REFERENCES

[1] ADLER, R. L. and B. WEISS. Entropy, a complete metric invariant for automorphisms of the torus. *Proc. Nat. Acad. Sci. U. S. A. 57* (1967), 1573–1576.

[2] BENOIST, Y. and H. OH. Equidistribution of rational matrices in their conjugacy classes. *Geom. Funct. Anal. 17* (2007), 1–32.

[3] BOREL, A. Some finiteness properties of adele groups over number fields. *Inst. Hautes Études Sci. Publ. Math. 16* (1963), 5–30.

[4] BOWEN, R. Maximizing entropy for a hyperbolic flow. *Math. Systems Theory 7* (1973), 300–303.

[5] BRIN, M. and A. KATOK. On local entropy. In: *Geometric Dynamics (Rio de Janeiro, 1981)*, 30–38. Lecture Notes in Mathematics *1007*. Springer, Berlin, 1983.

[6] BRUHAT, F. et J. TITS. Schémas en groupes et immeubles des groupes classiques sur un corps local. II. Groupes unitaires. *Bull. Soc. Math. France 115* (1987), 141–195.

[7] CASSELS, J. W. S. *Rational Quadratic Forms.* London Mathematical Society Monographs *13*. Academic Press, Inc., London-New York, 1978.

[8] CHELLURI, TH. Equidistribution of roots of quadratic congruences. Ph.D. Thesis Rutgers The State University of New Jersey, New Brunswick, 2004.

[9] DUKE, W. Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math. 92* (1988), 73–90.

[10] EDWARDS, H. M. *Fermat's Last Theorem. A Genetic Introduction to Algebraic Number Theory.* Graduate Texts in Mathematics *50*. Springer-Verlag, New York-Berlin, 1977.

[11] EINSIEDLER, M. and S. KADYROV. Entropy and escape of mass for $\mathrm{SL}_3(\mathbf{Z})\backslash \mathrm{SL}_3(\mathbf{R})$. *Israel J. Math. 190* (2012), 253–288.

[12] EINSIEDLER, M. and E. LINDENSTRAUSS. Diagonal actions on locally homogeneous spaces. In: *Homogeneous Flows, Moduli Spaces and Arithmetic*, 155–241. Clay Math. Proc. *10*. Amer. Math. Soc., Providence, RI, 2010.

[13] EINSIEDLER, M., E. LINDENSTRAUSS, PH. MICHEL and A. VENKATESH. Distribution of periodic torus orbits on homogeneous spaces. *Duke Math. J. 148* (2009), 119–174.

[14] ELLENBERG, J., PH. MICHEL and A. VENKATESH. Linnik's ergodic method and the distribution of integer points on spheres. In: *Proceedings of the 2012 International Math.Colloquium held at the TIFR Mumbay*, to appear; arXiv: 1001.0897 (2010).

[15] GOLDMAN, O. and N. IWAHORI. The space of $\mathfrak{p}$-adic norms. *Acta Math. 109* (1963), 137–177.

[16] HANKE, J. Local densities and explicit bounds for representability by a quadratic form. *Duke Math. J. 124* (2004), 351–388.

[17] IWANIEC, H. Fourier coefficients of modular forms of half-integral weight. *Invent. Math. 87* (1987), 385–401.

[18] LATIMER, C. G. and C. C. MACDUFFEE. A correspondence between classes of ideals and classes of matrices. *Ann. of Math. (2) 34* (1933), 313–316.

[19] LINNIK, YU. V. *Ergodic Properties of Algebraic Fields.* Translated from the Russian by M. S. Keane. Ergebnisse der Mathematik und ihrer Grenzgebiete *45*. Springer-Verlag, New York Inc., New York, 1968. Russian original, Izdat. Leningrad. Univ., Leningrad, 1967.

[20] MARGULIS, G. A. and G. M. TOMANOV. Invariant measures for actions of unipotent groups over local fields on homogeneous spaces. *Invent. Math. 116* (1994), 347–392.

[21] PALL, G. Representation by quadratic forms. *Canadian J. Math. 1* (1949), 344–364.

[22] SERRE, J.-P. *A Course in Arithmetic*. Translated from the French. Graduate Texts in Mathematics *7*. Springer-Verlag, New York-Heidelberg, 1973,

[23] —— *Trees*. Translated from the French original by John Stillwell. Corrected 2nd printing of the 1980 English translation. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2003. (French original: *Arbres, amalgames,* SL_2, Astérisque *46*. Soc. Math. France, Paris, 1977.)

[24] SKUBENKO, B. F. The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems. *Izv. Akad. Nauk SSSR Ser. Mat. 26* (1962), 721–752.

[25] VENKOV, B. A. Über die Klassenanzahl positiver binärer quadratischer Formen. *Math. Z. 33* (1931), 350–374.

[26] WALTERS, P. *An Introduction to Ergodic Theory*. Graduate Texts in Mathematics *79*. Springer-Verlag, New York-Berlin, 1982.

[27] WEIL, A. Sur la formule de Siegel dans la théorie des groupes classiques. *Acta Math. 113* (1965), 1–87.

(Reçu le 7 mars 2011)

Manfred Einsiedler

Department of Mathematics
ETH Zürich
Rämistrasse 101
CH-8092 Zürich
Switzerland
e-mail: manfred.einsiedler@math.ethz.ch

Philippe Michel

EPF Lausanne
SB-IMB-TAN, Station 8
CH-1015 Lausanne
Switzerland
e-mail: philippe.michel@epfl.ch

Elon Lindenstrauss

The Einstein Institute of Mathematics
Edmond J. Safra Campus
Givat Ram
The Hebrew University of Jerusalem
Jerusalem, 91904
Israel
e-mail: elon@math.huji.ac.il

Akshay Venkatesh

Stanford University
Department of Mathematics
Building 380
Stanford, CA 94305
U. S. A.
e-mail: akshay@math.stanford.edu