Ratner's Theorems and the Oppenheim Conjecture

Senior thesis in Mathematics
submitted in partial fulfillment of the requirements
for the degree of Bachelor of Arts with Honors

Elena Yudovina
Adviser: Prof. Danijela Damjanovic

Department of Mathematics
Harvard University

1 Introduction

The field of dynamical systems studies the long-term behavior of a system that evolves under the repeated application of some transformation. For example, consider the solar system: its time evolution is approximated by Newtonian mechanics and laws of gravitation. The modern theory of dynamical systems originated at the end of the 19th century with an attempt to answer questions like, "What do the orbits in the solar system look like? (In particular, do they spiral into the Sun?)" The theory has since then developed into a broad field of mathematics with applications to meteorology, economics, astronomy, and other areas – including number theory.

In this thesis, I will focus on a fairly recent set of powerful theorems in the theory of dynamical systems, proved by Marina Ratner around 1990. The theorems, in full generality, concern Lie groups and the actions of their subgroups generated by unipotent elements. They can be thought of as a sweeping generalization of the observation that a line on a 2-dimensional torus is either closed (that is, it wraps around the torus a finite number of times) or dense. Ratner's theorems assert that, in general, the closure of an orbit will be a very nice topological set. We prove a special case of these theorems for the Lie group $SL_2(\mathbb{R})/SL_2(\mathbb{Z})$. We then present a surprising application of the theory of dynamical systems to number theory and the Oppenheim conjecture on the values of quadratic forms. The conjecture asserts that an indefinite quadratic form in $n \geq 3$ variables is either proportional to a form defined over $\mathbb{Z}$ or its values on $\mathbb{Z}^n$ are dense in $\mathbb{R}$; the surprising connection between this number-theoretic result and the theory of dynamical systems was realized by M. S. Ranghunathan in the 1980s. The Oppenheim conjecture was first proved by G. A. Margulis, who gave a partial proof of the Ranghunathan conjectures in 1989. Ratner's work in the early 1990s proved the Ranghunathan and Margulis conjectures in full generality.

The structure of this thesis is as follows. In Section 2 we present an introduction to the theory of dynamical systems and an overview of the key concept: ergodicity. In Section 3 we state some of Ratner's theorems on the orbits of dynamical systems under a unipotent flow; we then present a proof of the theorems for $SL_2(\mathbb{R})/SL_2(\mathbb{Z})$, which is much simpler than the proof in full generality. In Section 4 we state the Oppenheim conjecture on the values of quadratic forms and make some remarks about it. The connection between this number-theoretic result and the theory of dynamical systems is the main concern of Section 5, which derives the proof of the Oppenheim conjecture from Ratner's theorems for $SL_3(\mathbb{R})/SL_3(\mathbb{Z})$. We present two lines of proof – one using the theory of algebraic groups and another one closer to the original approach used by Margulis.

I would like to thank my adviser, Prof. Danijela Damjanovic, who was an endless source of help in uncovering and understanding the material presented in this thesis – especially in converting references of the form "See exercise 4.7b" into viable proofs. The Harvard University Department of Mathematics provided an excellent library, warm support, and highly suggestive deadlines. I owe a debt of gratitude to the generous souls who proofread this work in its more and less spell-checked stages. I must also thank my family and friends for putting up with increasingly disturbing displays of "Eep! Thesis!" for the past month. Last but not least, the spirit of this paragraph is due to an excellent essay by U. Eco, "How to Write and Introduction" (in: U. Eco, translated by W. Weaver, *How to Travel with a Salmon and other essays*, Harcourt, 1994).

Contents

2 Introduction to dynamical systems

In this section we present some background in the field of dynamical systems, which will be followed by a few examples. While some of the earlier examples can safely be omitted, the later ones present useful background on lattices and the geometry and dynamics of the complex upper half-plane; this material is used in the proof of Ratner's theorems for $SL_2(\mathbb{R})/SL_2(\mathbb{Z})$.

2.1 Measure-preserving transformations

Definition 2.1. A *discrete-time dynamical system* is a set X together with a transformation $T : X \to X$. A *continuous-time dynamical system* is a set X together with a one-parameter family of maps $T_t : X \to X$, $t \in \mathbb{R}_{\geq 0}$ that forms a semigroup: that is, $T_{s+t} = T_s \circ T_t$, and T_0 is the identity transformation. If t ranges over $\mathbb{R}$, the one-parameter family is called a *flow*, and T_t is invertible: $T_{-t} = T_t^{-1}$.

Definition 2.2. A non-empty family $\mathcal{F}$ of subsets of X is called a *σ-algebra* if $\mathcal{F}$ is closed under countable unions, countable intersections, and complements. A *measure* on $\mathcal{F}$ is a non-negative function $\mu : \mathcal{F} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ that is σ-additive: that is, for a countable collection of disjoint sets $A_i \in \mathcal{F}$ we have $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$.

In all that follows, we will be interested in complete measure spaces, in which all subsets of sets of measure 0 are measurable (and have measure 0). That is, $\mathcal{F} \subset \mathcal{P}(X)$ will contain every subset of every set of measure 0. In a complete measure space, we necessarily have $\emptyset \in \mathcal{F}$, and therefore $X \in \mathcal{F}$. We define μ to be a *probability measure on* X if $\mu(X) = 1$; we then refer to (X, μ) as a *probability space.*

We will use the terminology *for almost every* x or *almost everywhere* to mean "for all x except in a set of measure 0"; both of these are abbreviated *a.e.*

Definition 2.3. A transformation $T : (X, \mu) \to (Y, \nu)$ is *measurable* if $T^{-1}(U)$ is measurable for every measurable subset U of Y. It is *nonsingular* if $T^{-1}(U)$ has measure zero whenever U has measure zero. A measurable transformation $T : X \to Y$ is *measure-preserving* if $\mu(T^{-1}(U)) = \nu(U)$ for every measurable $U \subset Y$. Lastly, a flow T_t is *measurable* if the map $T : X \times \mathbb{R} \to X$, $(x, t) \mapsto T_t(x)$ is measurable, and T_t is measurable and nonsingular for every t. (The measure on $X \times \mathbb{R}$ is the product measure, corresponding to the Lebesgue measure on $\mathbb{R}$.) A measurable flow T_t is *measure-preserving* if each T_t is a measure-preserving transformation.

Remark 2.4. When we are speaking of dynamical systems, it often makes sense to speak of functions defined only a.e.. In particular, two systems $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{G}, \nu)$ are measure-isomorphic if there exists a measure-preserving bijection between subsets of full measure $X' \subset X$ and $Y' \subset Y$.

We will often be concerned with transformations $T : X \to X$, with a single measure μ on X. If $T : X \to X$ is μ-preserving, then μ is said to be *T-invariant.*

2.2 Recurrence

Theorem 2.5 (Poincaré Recurrence Theorem)**.** *Let T be a measure-preserving transformation on a probability space $(X, \mathcal{F}, \mu)$. If A is a measurable set, then for a.e. $x \in A$ there is some $n \in \mathbb{N}$ such that $T^n(x) \in A$; consequently, for a.e. $x \in A$ there are infinitely many $k \in \mathbb{N}$ such that $T^k(x) \in A$.*

Proof. Let $B = \{x \in A : T^k(x) \notin A \text{ for all } k \in \mathbb{N}\}$. Then B is measurable, all the preimages $T^{-k}(B)$ are disjoint, and they have the same measure as B. Since X has finite total measure, B must have measure 0, proving the first assertion of the theorem: a.e. point $x \in A$ must return to A. Of the points $x \in A$ returning to A, only a subset of measure zero fails to return to A for the second time, and only a subset of measure zero fails to return to A for the third time, and so on. In general, the set of x such that there are only finitely many values of k for which $T^k(x) \in A$ has measure zero, proving the second assertion of the theorem. □

2.3 Ergodicity, unique ergodicity

Definition 2.6. A transformation T on a probability space X is *ergodic* with respect to μ if for every T-invariant set A we have $\mu(A) = 0$ or $\mu(A) = 1$: that is, if $T^{-1}(A) = A$ then $\mu(A) = 0$ or $\mu(A) = 1$.

Proposition 2.7. *The following are equivalent:*

1. *T is ergodic.*

2. *For every essentially T-invariant set A we have $\mu(A) = 0$ or $\mu(A) = 1$. (A set A is essentially T-invariant if $\mu(T^{-1}(A)\Delta A) = 0$.)*

3. *Every essentially invariant function $f : X \to \mathbb{R}$ is constant a.e. (A function f is said to be invariant if $f(Tx) = f(x)$, and essentially invariant if this is true a.e.; if T is a flow, this must be true at each time t.)*

4. *For $p \in (0, \infty]$, every essentially invariant function $f \in L^p(X)$ is constant a.e..*

Proof. Clearly, $3 \implies 4$ and $2 \implies 1$; moreover, $4 \implies 2$ by considering the characteristic function of A.

To show $4 \implies 3$, let $M > 0$ and consider the function f_M which is equal to f if $|f(x)| \leq M$ and to 0 otherwise. Note that $f_M \in L^p$ (it is bounded), and is essentially invariant, so it must be constant a.e.; since this is true for all M, so must f.

To show that $1 \implies 2$, let A be essentially T-invariant, and consider $\bigcap_n T^{-n}A$: this set differs from A by a set of measure zero, and is strictly T-invariant; therefore, it must have measure 0 or 1, and A must have measure 0 or 1.

Finally, $2 \implies 4$ directly for characteristic functions, and therefore for simple functions; since every bounded function is uniformly approximated by simple functions, we have the result on L^∞, which suffices. □

Remark 2.8. Let us consider the example of X a compact quotient of $\mathbb{R}^n$ with the usual Lebesgue measure μ (μ is the unique translation-invariant measure on $\mathbb{R}^n$ normalized so that $\mu(X) = 1$). Suppose T is a measure-preserving action on $\mathbb{R}^n$, and the orbit Tx of $x \in \mathbb{R}^n$ is dense. We claim that T is ergodic.

Indeed, suppose $S \subset X$ is a T-invariant subset of positive measure: then for every $\epsilon > 0$ we can find a product of intervals $I = I_1 \times I_2 \times \ldots \times I_n$ such that S has density $\geq 1 - \epsilon$ in I (that is, $\mu(S \cap I)/\mu(I) \geq 1 - \epsilon$). By density, the orbit of Tx passes through I, so without loss of generality we may simply assume $x \in I$ to begin with. But, again by density, T-translates of I cover all of X, and therefore S has density $\geq 1 - \epsilon$ in all of X, i.e. has measure $\geq 1 - \epsilon$. Since this holds for every ϵ, we observe that $\mu(S) = 1$.

This observation – that a dense orbit implies ergodicity – will be useful to us when we analyze simple examples of dynamical systems.

We now define what it means for a transformation to be uniquely ergodic.

Definition 2.9. A transformation T on a compact metric space X is said to be *uniquely ergodic* if there is exactly one T-invariant probability measure μ on X.

Intuitively, the meaning of unique ergodicity is as follows: the properties that hold a.e. for ergodic transformations will hold everywhere for uniquely ergodic ones. For example, suppose we have an ergodic transformation T and a measurable set $A \subset X$. For almost every $x \in X$ it makes sense to speak of the proportion of time the orbit of x under the action of T spends in A (this proportion will be $\mu(A)$). If T were uniquely ergodic, this statement would be true for all $x \in X$.

Remark 2.10. It is true, although I won't prove it here, that there always exists at least one ergodic T-invariant probability measure μ on X. A proof may be found in [2, Section 4.6].

2.4 Ergodic theorems

Attached to a measure-preserving transformation $T : X \to X$ is an operator U_T on the space of measurable functions $f : X \to \mathbb{C}$: $U_T f = f \circ T$. U_T is sometimes referred to as the Koopman operator of T; it is linear and multiplicative (that is, $U_{S \circ T} = U_S \circ U_T$). Moreover, U_T is an isometry of $L^p(X)$ for every p, since T is measure-preserving.

The following theorem tells us that, if I have an isometry U (not necessarily ergodic!) of a Hilbert space H, and a function $f \in H$, it makes sense to speak of the time average of f under the action of U (it will be another function in H).

Theorem 2.11 (von Neumann Ergodic Theorem)**.** *Let U be an isometry of a separable Hilbert space H, and let P be orthogonal projection onto $I = \{f \in H : Uf = f\}$, the subspace of U-invariant elements of H. Then for every $f \in H$,*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i f = Pf.$$

Proof. The strategy is to show that under the operation described in the theorem, the component of f perpendicular to I gets reduced to zero.

Let $U_n = \frac{1}{n}\sum_{i=0}^{n-1} U^i$, and let $L = \{g - Ug : g \in H\}$. Observe that L and I are U-invariant, and I is closed. Now, for $f = g - Ug \in L$ we have $U_n(f) = \frac{1}{n}(g - U^n g)$, and since U is an isometry, $U_n f \to 0$. Note that the same will hold for $f \in \overline{L}$: if $\{f_k\} \to f \in \overline{L}$ then $\|U_n f\| \leq \|U_n f_k\| + \|U_n(f - f_k)\|$, where both terms approach zero. On the other hand, for $f \in I$, we of course have $U_n f = f$ for all n. It now suffices to show that $L \perp I$ and $H = \overline{L} \oplus I$.

For $h \in L^{\perp}$, we have $0 = \langle h, g - Ug\rangle = \langle h - U^* h, g\rangle$ for all $g \in H$: therefore, $h = U^* h$, and hence $h = Uh$. The converse is true upon running the argument backwards. Therefore, $h \in I$, and $H = \overline{L} \oplus I$ as required. The theorem follows. □

As a corollary to the above, we have the following

Theorem 2.12. *Let T be a measure-preserving transformation on a probability space X. For $f \in L^2(X, \mu)$ set*

$$f_N^+(x) = \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x).$$

Then f_N^+ converges in L^2 to a T-invariant function $\overline{f}$. If T is invertible, then $f_N^- = \frac{1}{N}\sum_{n=0}^{N-1} f(T^{-n}x)$ also converges in L^2, to the same function $\overline{f}$.

Similarly, let T be a measure-preserving flow on a probability space X. For $f \in L^2(X, \mu)$ let

$$f_t^+ = \frac{1}{t}\int_0^t f(T_t(x))dt$$

and correspondingly for f_t^-. Then f_t^+ and f_t^- converge in L^2 to a T-invariant function $\overline{f}$.

The theorem asserts that for a measure-preserving transformation it makes sense (more or less – the convergence is in L^2 rather than pointwise) to speak of the time-average of $f(T^n(x))$ or $f(T_t(x))$. For example, if x is a particle and $f = \chi_A$, it "more or less" makes sense to say that x spends a certain proportion of its time in A. The Birkhoff ergodic theorem will make more precise the "more or less:" it can be replaced by "a.e.".

The independence of $\overline{f}$ on $n \to \infty$ or $n \to -\infty$ follows from the fact that the limit in the von Neumann theorem does not depend on whether we use U or U^{-1} as the isometry: the space $I = \{f \in H : Uf = f\}$ coincides with the space $\tilde{I} = \{f \in H : U^{-1}f = f\}$, so the limit Pf is the same in both cases.

Theorem 2.13 (Birkhoff Ergodic Theorem)**.** *Let T be a measure-preserving transformation in a probability space (X, μ), and let $f \in L^1(X, \mu)$. Then the limit*

$$\overline{f}(x) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$$

exists for a.e. $x \in X$, and moreover is μ-integrable and T-invariant, satisfying

$$\int_X \overline{f}(x)d\mu = \int_X f(x)d\mu.$$

If T is invertible, then $\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$ also converges a.e. to $\overline{f}$.

Similarly, for a measure-preserving flow T, $f_t^+(x)$ and $f_t^-(x)$ converge a.e. to the same μ-integrable and T-invariant function $\overline{f}$, and $\int_X \overline{f}(x)d\mu = \int_X f(x)d\mu$.

Proof. We will show that $\overline{f}$ exists by showing that $f^*(x) = \limsup_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$ and $f_*(x) = \liminf_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$ are equal a.e. Note that f^* and f_* are always defined. They are also T-invariant:

$$\frac{n+1}{n}\left(\frac{1}{n+1}\sum_{k=0}^{n} f(T^k x)\right) = \frac{1}{n}\sum_{k=0}^{n-1} f(T^k \circ Tx) + \frac{1}{n}f(x)$$

and by choosing the lim sup to come from the left-hand side or the right-hand side of the equation, we obtain both $f^* \geq f^* \circ T$ and $f^* \leq f^* \circ T$ (similarly for f_*).

Now, let $\alpha > \beta \in \mathbb{Q}$ and write

$$B_\alpha^\beta = \{x \in X : f_*(x) < \beta < \alpha < f^*(x)\}.$$

Clearly, it suffices to show that each B_β^α has measure 0. Our strategy in showing that $\mu(B_\alpha^\beta) = 0$ will be to show that $\int_{B_\alpha^\beta} f(x) \geq \alpha\mu(B_\alpha^\beta)$ and also $\int_{B_\alpha^\beta} f(x) \leq \beta\mu(B_\alpha^\beta)$: since $\alpha > \beta$, this is a contradiction unless $\mu(B_\alpha^\beta) = 0$.

Lemma 2.14 (Maximal inequality)**.** *Let f be a real-valued function and T a measure-preserving transformation. Let*

$$A = \{x \in X : f(x) + f(T(x)) + \ldots + f(T^k(x)) \geq 0 \textit{ for some } k \geq 0\}.$$

Then $\int_A f(x)d\mu \geq 0$.

Proof. Let $A_n = \{x \in X : \sum_{i=0}^{k} f(T^i(x)) \geq 0$ for some k, $0 \leq k < n\}$. Then $A_n \subset A_{n+1}$ and $A = \bigcup A_n$. By the dominated convergence theorem it suffices to show that $\int_{A_n} f(x)d\mu \geq 0$ for each n.

Define $f_0 = 0$, $f_1 = f$, and $f_k = f_{k-1} + U_T^{k-1}f = f_{k-1} + f \circ T^{k-1}$. Further define $F_n = \max_{1\leq k\leq n} f_k$. Then $A_n = \{x : F_n(x) \geq 0\}$.

Note that $F_n \geq f_k$ for $1 \leq k \leq n$, so $U_T F_n + f \geq U_T f_k + f = f_{k+1}$, and in particular $U_T F_n + f \geq \max_{2\leq k\leq n} f_k$. On A_n, we also have $F_n \geq 0 = f_0$, and consequently $U_T F_n + f \geq \max_{1\leq k\leq n} f_k = F_n$. Therefore,

$$\int_{A_n} f(x)d\mu \geq \int_{A_n} F_n(x)d\mu - \int_{A_n} U_T F_n(x)d\mu \geq \|F_n\|_1 - \|U_T F_n\|_1 \geq 0$$

(we are computing exactly $\|F_n\|_1$, since we are integrating over the entire set where it is nonzero, and we are not exceeding $\|U_T F_n\|_1$). The last inequality follows because U_T is composition with a measure-preserving transformation, and therefore is a positive operator of norm ≤ 1. □

Now we can approach the question of $\mu(B^\alpha_\beta)$. Note that $B^\alpha_\beta \subset B^\alpha = \{x : f^*(x) > \alpha\}$, or equivalently $\{x : f^*(x) - \alpha > 0\}$. By the above lemma, we can conclude that

$$\int_{B^\beta_\alpha} (f(x) - \alpha) d\mu \geq \int_{B^\beta_\alpha} (f(x) - \alpha) d\mu \geq 0.$$

(Note that if $f^*(x) < \beta$ then in particular $f(x) < \beta < \alpha$, so $B^\alpha \backslash B^\alpha_\beta$ makes a strictly negative contribution.) Equivalently, $\int_{B^\beta_\alpha} f(x) \geq \alpha\mu(B^\beta_\alpha)$. On the other hand, by considering $-f$ instead of f we also obtain $\int_{B^\beta_\alpha} f(x) \leq \beta\mu(B^\beta_\alpha)$. Since $\alpha > \beta$, this is a contradiction unless $\mu(B^\alpha_\beta) = 0$.

Therefore, f_* and f^* differ on a set of measure zero, i.e. $g_n = \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)$ converges a.e. to $\overline{f}(x)$. Note that this sequence also has a limit in L^1 by the previous theorem. Picking a subsequence, we see that the limit in L^1 can be made pointwise a.e., and we conclude that $\overline{f} \in L^1$.

Finally, it is easy to see that $\int_X f(x)d\mu = \int_X g_n(x)d\mu = \int_X \overline{f}(x)d\mu$. □

It is worth noting that $\int_X f(x)d\mu$ is the space average of f, and $\overline{f}$ is the time average of f. If T were ergodic, we would conclude that $\overline{f}$ was essentially constant: the time average of f at almost any particular point would be equal to the space average of f over all of X. In fact, we can phrase the notion of ergodicity in these terms:

Corollary 2.15. *A measure-preserving transformation T on a probability space (X, μ) is ergodic if and only if for every measurable set A, for a.e. $x \in X$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \chi_A(T^k(x)) = \mu(A),$$

i.e. if the proportion of time x spends in A is $\mu(A)$.

Both the Birkhoff ergodic theorem and the corollary above are true if we replace the words "a.e." by "everywhere," provided that instead of an ergodic transformation we have a uniquely ergodic transformation (or flow). To show this, let T be uniquely ergodic, and consider the Haar measure on the T-orbit of any point $x \in X$. That is, for any $E \subset X$ and f a function on X, we define

$$\int_X f d\nu = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x).$$

This is a T-invariant measure (by construction), so it must be the unique T-invariant measure on X. However, if μ is the measure on a single T-orbit, the statements follow immediately.

We derive another corollary of the Birkhoff ergodic theorem now:

Corollary 2.16. *Let $T_t : X \to X$ be a flow preserving an ergodic probability measure μ, and let $f \in L^1(\mu)$. For any $\epsilon > 0$ and $\delta > 0$ there exists $\tau_0 > 0$ and a set E with $\mu(E) < \epsilon$ such that for any $x \notin E$ and any $\tau > \tau_0$ we have*

$$\left| \frac{1}{\tau} \int_0^\tau f(\phi_t(x))dt - \int_X f d\mu \right| < \delta.$$

In other words, the average of f over the orbit of x converges to the average of f over X uniformly outside of a set of small measure.

Proof. Let E_n be the set of $x \in X$ such that for some $\tau > n$,

$$\left| \frac{1}{\tau} \int_0^\tau f(\phi_t(x))dt - \int_X f d\mu \right| \geq \delta.$$

By Birkhoff ergodic theorem, $\mu(\bigcap_n E_n) = 0$; thus for some n we have $\mu(E_n) < \epsilon$. Let $\tau_0 = n$ and $E = E_n$. □

2.5 Examples

2.5.1 Circle rotation

Consider the circle $S^1 = \mathbb{R}/\mathbb{Z}$. The circle inherits a translation-invariant measure μ from $\mathbb{R}$ ($\mu[a,b] = |b-a|$). We examine the circle rotation map R_α: $x \mapsto x + \alpha$; since μ is translation-invariant, R_α is measure-preserving.

If $\alpha = p/q$ is rational, then $(R_\alpha)^q$ is the identity map; therefore, R_α cannot be ergodic. For an example of an invariant set of measure 1/2, let B be the ball $B(0, \frac{1}{4q})$ of diameter $\frac{1}{2q}$, and consider the set $S = \bigcup_{n=0}^{q-1} B + \frac{n}{q}$. Then S has total measure 1/2, but is invariant under rotation by $1/q$ and therefore by p/q.

On the other hand, if α is irrational, then R_α is ergodic. By Remark 2.8, it suffices to show that the orbit of some (and, as it happens, every) point is dense.

We now show that if α is irrational, the orbit of every point is dense. Since the orbit of x is obtained from the orbit of 0 by translating by x, the two are dense or not simultaneously; we are thus reduced to showing that the orbit of 0, $\{n\alpha\}$, is dense mod1.

Consider $\{n\alpha \bmod 1\}$, and let $\epsilon > 0$ be arbitrary. By the pigeonhole principle, we must, for some $n \neq m$, have $|n\alpha - (m\alpha + k)| < \epsilon$ for some $k \in \mathbb{Z}$ (that is, $n\alpha \approx m\alpha \bmod 1$). Then $|(n-m)\alpha - k| < \epsilon$, i.e. $(n-m)\alpha$ is near 0 mod 1. Consequently, multiples of $(n-m)\alpha$ form an ϵ-net mod1. Since ϵ was arbitrary, the set of multiples of α is dense mod 1.

In fact, the usual Lebesgue measure is uniquely ergodic with respect to R_α. We will show that any measure that is invariant with respect to R_α must be invariant with respect to arbitrary translations, at which point it must be the unique translation-invariant measure. Let ν be a R_α-invariant measure, and let μ be the Lebesgue measure; we wish to show that $\mu = \nu$. Consider an interval I of μ-measure $1/n$. Now, as we showed above, some power of R_α is translation by $x \in (1/n, 2/n)$. This gives us $n/2$ disjoint translates of I, from which $\nu(I) < 2/n$; and consequently, for any Borel set E we have $\nu(E) < 2\mu(E)$.

Now, if we have an arbitrary interval $J = (a,b)$ and a translate of it $(a+c, b+c)$, then we can find some power of R_α that is approximately R_c: in particular, we can guarantee $\mu((a+c,b+c)\Delta R_\alpha^k(a,b)) < \epsilon$ for some k. But then

$$\nu((a+c,b+c)\Delta R_\alpha^k(a,b)) < 2\epsilon$$

and since ν is R_α-invariant, $|\nu(a+c,b+c) - \nu(a,b)| < 2\epsilon$. Since this holds for any c and any ϵ, we see that ν is the translation-invariant measure.

2.5.2 Addition on a torus

Consider the 2-torus $\mathbb{T} = S^1 \times S^1$; it inherits a translation-invariant measure μ from the Lebesgue measure on $\mathbb{R}^2$. On $\mathbb{T}$, consider the flow $T_{a,b} : T_t(x, y) \mapsto (x + ta, y + tb)$. The $T_{a,b}$-orbit of a point is therefore the image of a line in $\mathbb{R}^2$ modulo $\mathbb{Z}^2$. Note that we can scale the pair (a, b) so that $a = 1$ without changing the orbit; we'll refer to the scaled flow by T_b.

If $b = p/q$ is rational, then the orbit of (x, y) is a closed circle: it will wind around the torus q times in one direction, and p times in the other. It is easy to see that in that case, T_b is not an ergodic flow: a thin strip around the orbit of a point will be T_b-invariant, but can have arbitrary measure.

On the other hand, if b is irrational, the orbit of $(0, 0)$ will be dense in the torus, and therefore T_b will be μ-ergodic. The orbit of $(0, 0)$ is the line $y = bx$. Showing that it passes through any open ball on the torus is equivalent to showing that this line passes through points of the form $(n_1, m_1 + \epsilon_1)$ and $(n_2 + \epsilon_2, m_2)$ for some integers n_i, m_i and arbitrarily small ϵ_i: if we show this, then linear combinations of these points will form an $\epsilon_1 \times \epsilon_2$-net on $\mathbb{T}$, and therefore the line will be dense. This, in turn, is equivalent to approximating the slope $b \notin \mathbb{Q}$ by fractions n_i/m_i to within ϵ/m_i, which was done in the previous example.

2.5.3 Geometry of the upper half-plane

We will now discuss the geometry of the complex upper half-plane, $\mathbb{H}$. The material in the next several paragraphs is standard, and some of the details are omitted; they can be found in [6, Chapter 5.4].

To give $\mathbb{H}$ the structure of a Riemannian manifold, we must specify a metric, that is, the inner product on the tangent bundle. For $z \in \mathbb{H}$ and $u_1 + iv_1$, $u_2 + iv_2 \in T_z\mathbb{H}$ we define the hyperbolic metric on $\mathbb{H}$ by

$$\langle u_1 + iv_1, u_2 + iv_2 \rangle_z = \operatorname{Re} \frac{(u_1 + iv_1)(u_2 - iv_2)}{(\operatorname{Im} z)^2} = \frac{u_1u_2 + v_1v_2}{(\operatorname{Im} z)^2}.$$

The verification that this is a metric is standard. First, a lemma that will let us establish what the geodesics on the upper half-plane are:

Lemma 2.17. *Let M be a Riemannian manifold, and let Γ be a group of isometries of M that is transitive on unit vectors: that is, $\forall v$, $\tilde{v}$ in the unit tangent bundle on M there exists $\phi_{v,\tilde{v}} \in \Gamma$ with $\phi_{v,\tilde{v}}(v) = \tilde{v}$. Let $\mathcal{C}$ be a nonempty family of unit-speed curves satisfying the following properties:*

1. *$\mathcal{C}$ is closed under the action of Γ: that is, for all $c \in \mathcal{C}$ and $\phi \in \Gamma$, the composition $\phi \circ c \in \mathcal{C}$;*

2. *Γ is transitive on $\mathcal{C}$: that is, $\forall c, \tilde{c} \in \mathcal{C}$ there exists $\phi_{c,\tilde{c}} \in \Gamma$ with $\phi_{c,\tilde{c}} \circ c = \tilde{c}$; and*

3. *$\mathcal{C}$ consists of the axes of Γ: that is, $\forall c \in \mathcal{C}$ there exists $\phi_c \in \Gamma$ such that c is the set of fixed points of ϕ_c.*

Then $\mathcal{C}$ is the family of (all) unit-speed geodesics on M.

Proof. First, we show that $\mathcal{C}$ contains all the geodesics. Let $v \in T_pM$ be a unit tangent vector to M at p. It determines a unique geodesic γ_v with $\dot{\gamma}_v(0) = v$. Now take some $c \in \mathcal{C}$, and let $\tilde{v} = \dot{c}(0)$. Consider the action of $\phi_{\tilde{v},v}$: it maps c to a curve $\tilde{c}$ that is tangent to γ_v at p. Now we look at $\phi_{\tilde{c}}$, the isometry that fixes $\tilde{c}$: it must map γ_v to a geodesic, but since it fixes $\tilde{c}$ it must also fix v, and consequently $\phi_{\tilde{c}} \circ \gamma_v$ and γ_v are tangent to each other at p. Two tangent geodesics must coincide, so γ_v is fixed by $\phi_{\tilde{c}}$, and consequently $\gamma_v = \tilde{c} \in \mathcal{C}$.

Finally, since $\mathcal{C}$ contains geodesics and Γ is transitive on it, every curve in $\mathcal{C}$ is the isometric image of a geodesic, and thus itself a geodesic. □

As a simple application, we can now characterize the geodesics on the standard 2-sphere: let $\mathcal{C}$ be the family of great circles parametrized with unit speed, and let Γ be the group generated by rotations and reflections in great circles. Checking the conditions of the lemma is easy; we conclude that $\mathcal{C}$ is exactly the group of unit-speed geodesics on the 2-sphere.

We now consider the isometries of $\mathbb{H}$. The projective special linear group $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\{\pm I\}$ acts on $\mathbb{H}$ by fractional linear transformations:

$$z \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{az+b}{cz+d}$$

For later discussion, it is more convenient to let this be a right action, and let later actions be left ones. The group $PSL_2(\mathbb{R})$ is generated by translations $z \mapsto z+b$ corresponding to $\begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}$; inversions $z \mapsto -1/z$ corresponding to $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$; and scaling $z \mapsto a^2 z$ corresponding to $\begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$. (All three of these clearly map points in $\mathbb{H}$ to $\mathbb{H}$.) Also, for $T \in PSL_2(\mathbb{R})$, we have

$$\operatorname{Im} T(z) = |T'(z)| \operatorname{Im}(z).$$

We only need to check this for the two generating transformations, since this formula respects composition. For $z \mapsto z+b$ and $z \mapsto a^2 z$ this is clear, and for $z \mapsto -1/z$ we have $\operatorname{Im}(-1/z) = \frac{\operatorname{Im}(z)}{|z|^2}$ as required.

We can now check that the action of $PGL_2(\mathbb{R})$ on $\mathbb{H}$ is isometric:

$$\langle T'(z)(u_1+iv_1), T'(z)(u_2+iv_2)\rangle_{T(z)} = \operatorname{Re} \frac{T'(z)(u_1+iv_1)\overline{T'(z)(u_2+iv_2)}}{(\operatorname{Im} T(z))^2} =$$
$$\operatorname{Re} \frac{(u_1+iv_1)(u_2-iv_2)}{(\operatorname{Im} z)^2} = \langle u_1+iv_1, u_2+iv_2\rangle$$

Note that the action of $PSL_2(\mathbb{R})$ on the unit tangent bundle of $\mathbb{H}$ is transitive. Indeed, any $z \in \mathbb{H}$ may be translated onto the positive imaginary axis $i\mathbb{R}_+$, and then scaled onto i. It remains to check that $PSL_2(\mathbb{R})$ is transitive on $T_i(\mathbb{H})$; the transformation $z \mapsto \frac{\cos(\theta/2)z+\sin(\theta/2)}{-\sin(\theta/2)z+\cos(\theta/2)}$, corresponding to the matrix $\begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix}$, sends $v \in T_i\mathbb{H}$ to $v(\cos\theta + i\sin\theta)$, i.e. rotates v by θ.

We are now in a position to classify the geodesics on $\mathbb{H}$.

Theorem 2.18. *The geodesics on $\mathbb{H}$ are vertical lines or semicircles with center on the real axis.*

Proof. To prove this theorem, we enlarge the group of isometries of $\mathbb{H}$ by reflection in the $i\mathbb{R}$ axis, $z \mapsto -\overline{z}$. The resulting group is, of course, still transitive on the unit tangent bundle. To show property 1, namely that our family of curves $\mathcal{C}$ is closed under the action of our isometry group Γ, we note that this is clear for $z \mapsto -\overline{z}$. For the Möbius transformations we will show transitivity first, and then observe that any Möbius transformation acting on $i\mathbb{R}_+$ sends it to an element of $\mathcal{C}$. Finally, transitivity together with the fact that $i\mathbb{R}_+$ is the set of fixed points of $z \mapsto -\overline{z}$ will show property 3.

We now show that any vertical line or semicircle with center on the real axis can be mapped to $i\mathbb{R}_+$. For a vertical line through $b \in \mathbb{R}$, the transformation $z \mapsto z - b$ works. For a semicircle through b, $b + a^2 \in \mathbb{R}$ we translate left by b and scale by a^{-2} to get a semicircle through 0 and 1. Now consider the map $z \mapsto z/(1-z)$. Its inverse is $z \mapsto z/(z+1)$, which sends $i\mathbb{R}_+$ onto the semicircle with endpoints 0 and 1, since

$$\left|\frac{it}{1+it} - \frac{1}{2}\right| = \left|\frac{2it - (1+it)}{2(1+it)}\right| = \frac{1}{2}$$

Therefore, we mapped our semicircle onto $i\mathbb{R}_+$ as required.

Finally, we must show that Möbius transformations map $i\mathbb{R}_+$ into $\mathcal{C}$. It suffices to check this for the generators of the group of Möbius transformations: $z \mapsto z + b$ sends $i\mathbb{R}_+$ to the vertical line through $b \in \mathbb{R}$; $z \mapsto az$ sends $i\mathbb{R}_+$ to itself, reparametrizing it along the way; and $z \mapsto -1/z$ sends $i\mathbb{R}_+$ to itself but with the opposite parametrization.

Thus, we are in a position to apply lemma 2.17 and conclude that the geodesics of $\mathbb{H}$ are the vertical lines and semicircles with center on the real axis. □

Not only is $PSL_2(\mathbb{R})$ transitive on $T^1\mathbb{H}$, but the action is free: that is, the transformation g mapping $v \in T_p\mathbb{H}$ to $v' \in T_p'\mathbb{H}$ is unique. Indeed, g must map the unique geodesic tangent to v at p to the unique geodesic tangent to v' at v. On the other hand, a Möbius transformation is uniquely determined by where it maps any three points; therefore, g is unique. We may therefore identify the unit tangent bundle of $\mathbb{H}$ with $PSL_2(\mathbb{R})$: we identify $v \in T_z\mathbb{H}$ with the (unique) transformation that sends the upwards unit vector at i to v. (The upwards unit vector at i is therefore identified with the identity matrix.)

2.5.4 Geodesic flow

We now consider the geodesic flow on the unit tangent bundle of $\mathbb{H}$, described as follows: for $v \in T_z\mathbb{H}$ we define $\phi_t(v) = \dot{\gamma}(t) \in T_{\gamma(z)}\mathbb{H}$, where γ is the unique unit-speed geodesic with $\dot{\gamma}(0) = v$. Under the identification of the unit tangent bundle with $PSL_2(\mathbb{R})$, we have the following

Lemma 2.19. *The geodesic flow on the unit tangent bundle of $\mathbb{H}$ corresponds to the flow on the group $PSL_2(\mathbb{R})$ given by left translation $g \mapsto h_t g$ with*

$$h_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/e} \end{pmatrix}$$

for $t \in \mathbb{R}$.

Proof. Let v be the unit upward vector at i: then, by definition, $\phi_t(v)$ is the unit (in the $\mathbb{H}$-metric) upward vector at the point z on the geodesic $i\mathbb{R}_+$ a distance t away from i. Now, I claim that $d(i, e^t i) = t$: indeed,

$$d(i, e^t i) = \int_{y=1}^{e^t} \frac{dy}{y} = t$$

The unit upward vector at $e^t i$ has the form $e^t v$; therefore, $\phi_t(v) = e^t v \in T_{e^t i}$.

On the other hand, $h_t(i, v) = (e^t i, e^t v)$. Consequently, the actions agree on this unit tangent vector.

Now, let $\zeta \in T_z\mathbb{H}$ be arbitrary, and let $g_{z\zeta} \in PSL_2(\mathbb{R})$ be such that $\zeta = vg_{z\zeta}$ (where v is still the unit upward vector at i). Since $g_{z\zeta}$ is an isometry on $\mathbb{H}$, we have

$$\phi_t(\zeta) = \phi_t(vg_{z\zeta}) = \phi_t(v)g_{z\zeta} = h_t v g_{z\zeta} = h_t\zeta,$$

which is the desired result. □

Now, let $v \in T_p\mathbb{H}$ and $w \in T_q\mathbb{H}$ be two vectors in the unit tangent bundle of $\mathbb{H}$. We would like to define the distance between them. Consider the unique (unit-speed) geodesic γ with $\gamma(0) = p$ and passing through q, and the vector field along γ having the same angle with γ' as the angle between v and $\gamma'(0)$. Define the angle between v and w to be the angle between the vector field at q and w. (We just described the process of parallel-translating v along γ to a tangent vector at q.) Now define

$$d(v, w) = \sqrt{(\angle(v, w))^2 + d(p, q)^2}.$$

This is the standard definition of distance on the unit tangent bundle (see, for example, [6, Section A.4]), and in particular does define a distance function.

Now consider the unit upward vector at i and at $x+i$ for some $x \in \mathbb{R}$. Their orbits under the geodesic flow will be $t \mapsto ie^t$ and $t \mapsto x + ie^t$. The hyperbolic distance between these two points is easily seen to be bounded by xe^{-t} by considering the horizontal line segment joining ie^t and $x + ie^t$; moreover, the angle between v and w is readily seen to be $2\tan^{-1}(x/(2e^t))$, and in particular is also $\leq xe^{-t}$. Therefore, the distance between the tangent vectors to these geodesics is bounded by $\sqrt{2}xe^{-t}$: the orbits of the upward vertical unit vectors at all points $x \in \mathbb{R} + i$ are positively asymptotic to that of i (and to each other).

Applying $z \mapsto -1/z$, we see that the orbits of the outward unit normals to the circle of (Euclidean) radius $1/2$ centered at $i/2$ are negatively asymptotic to that of i (and to each other). This brings us to the concept of horocycles.

2.5.5 Horocycle flow

Definition 2.20. A *horocycle* on $\mathbb{H}$ is either a circle tangent to $\mathbb{R}$ at x or a horizontal line $\mathbb{R} + ir = \{t + ir | t \in \mathbb{R}\}$. In the first case, we say that the horocycle is centered at x; in the second case, we say that it is centered at ∞.

All horocycles are isometric to the line $\mathbb{R} + i$. Indeed, for horocycles $H = \mathbb{R} + ri$ the isometry $T(z) = z/r$ suffices; for a horocycle centered at $x \in \mathbb{R}$ of Euclidean diameter r, take

$T_1(z) = z - x$, $T_2(z) = z/r$ (applying $T_2 \circ T_1$ gets us a horocycle centered at 0 of Euclidean diameter 1), and $T_3(z) = -1/z$. Then $T = T_3 \circ T_2 \circ T_1$ maps our horocycle isometrically onto $\mathbb{R} + i$.

The horocycle flow on the unit tangent bundle is described as follows: for $v \in T_z\mathbb{H}$ there exists a unique horocycle passing through z whose inward normal is v. (We define "inward" to be "up" for the horocycle at ∞; with this definition, each horocycle rests at the $+\infty$ end of the geodesic tangent to v.) The action of ψ_s is to move z to a point s units away on the horocycle, and parallel-transport the unit tangent vector to an inward normal at that point. Equivalently, for v an upward normal to a point $z \in \mathbb{R} + i$, we define $\psi_s(v)$ to be the upward normal to the point $z + s \in \mathbb{R} + i$, noting that on the line $\mathbb{R} + i$ the hyperbolic metric coincides with the Euclidean one.

From this description, we conclude that under the identification of the unit tangent bundle of $\mathbb{H}$ with $PSL_2(\mathbb{R})$ the horocycle flow is given by the left action of

$$u_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}, \quad s \in \mathbb{R}$$

2.6 More on flows on the upper half-plane

2.6.1 As actions on $PSL_2(\mathbb{R})/\Gamma$

Let Γ be a discrete subgroup of $PSL_2(\mathbb{R})$ with finite covolume, i.e. a lattice. Then Γ acts freely and discontinuously on $\mathbb{H}$, and therefore is the group of deck transformations for $\mathbb{H}$ regarded as a covering space of $\mathbb{H}/\Gamma$. The identification of the unit tangent space of $\mathbb{H}$ with $PSL_2(\mathbb{R})$ induces an identification of the unit tangent space of $\mathbb{H}/\Gamma$ with $PSL_2(\mathbb{R})/\Gamma$; in this identification, the geodesic flow corresponds to the action of $h_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$, $t \in \mathbb{R}$, and the horocycle flow corresponds to the action of $u_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$, $t \in \mathbb{R}$.

2.6.2 Ergodicity of the geodesic and horocycle flows

We use the following notation: let

$$N^+ = \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} | x \in \mathbb{R} \right\}; \quad A = \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} | a > 0 \right\}; \quad N^- = \left\{ \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} | x \in \mathbb{R} \right\}$$

Then N^+, N^-, and A together generate all of $PSL_2(\mathbb{R})$. Recall that A is the set of matrices in the geodesic flow.

We will show that the geodesic flow on $\mathbb{H}/\Gamma$ is ergodic for every lattice Γ in $\mathbb{H}$. To do this, we will show that if a function f is invariant under the action of A, then it must be invariant under the action of all of $PSL_2(\mathbb{R})$, and then if f is in L^2, it must be constant. This will establish ergodicity.

We observe the following

Lemma 2.21. *For $g = g_a \in A$ and $h \in N^+$ if $a < 1$, or $h \in N^-$ if $a > 1$, we have*

$$\lim_{n \to \infty} g^n h g^{-n} = e.$$

That is, conjugation by g contracts the relevant h.

The proof is by direct computation:

$$\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & a^2x \\ 0 & 1 \end{pmatrix}$$

so

$$\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}^n \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}^{-n} = \begin{pmatrix} 1 & a^{2n}x \\ 0 & 1 \end{pmatrix}.$$

Also,

$$\begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ a^{-2n}x & 1 \end{pmatrix}.$$

Now, we can use the following result due to Mautner:

Lemma 2.22 (Mautner Lemma)**.** *Consider a unitary representation on $PSL_2(\mathbb{R})/\Gamma$ as above, and suppose $g, h \in SL_2(\mathbb{R})$ satisfying $\lim_{n\to\infty} g^n h g^{-n} = 1$. Then all $f \in L^2(PSL_2(\mathbb{R})/\Gamma)$ that are invariant under the action of g are also invariant under the action of h.*

Proof. We use the associated operators of these actions instead. Note that

$$\|T_h f - f\| = \|T_h T_{g^{-n}} f - T_{g^{-n}} f\| = \|T_{g^n} T_h T_{g^{-n}} f - f\|$$

We may let $n \to \infty$ inside the norm, concluding that $\|T_h f - f\| = 0$ and f is invariant under h. □

Ergodicity of the geodesic flow on $PSL_2(\mathbb{R})/\Gamma$. If $T_a f = f$ for every $a \in A$, then combining Lemma 2.21 with the Mautner lemma we derive that $T_g f = f$ for every $g \in PSL_2(\mathbb{R})$. For $f \in L^2$ this means that f is essentially constant, and transformation by A is ergodic as required. □

We can also show that the horocycle flow is ergodic:

Ergodicity of the horocycle flow on $PSL_2(\mathbb{R})/\Gamma$. Let $f \in L^2(PSL_2(\mathbb{R})/\Gamma)$ be invariant under N^+. We will show that f must be invariant under A as well, and then proceed as in the proof of the ergodicity of the horocycle flow.

For each $g \in SL_2(\mathbb{R})$ define $\phi(g) = \langle T_g f, f \rangle$. Since f is invariant under all of N^+, the operator ϕ is constant on every double coset $N^+ g N^+$.

Now let $\lambda_n \to 0$, $\lambda_n \neq 0$ for every n, and consider $g_n = \begin{pmatrix} 0 & \lambda_n^{-1} \\ \lambda_n & 0 \end{pmatrix}$. Let $a = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha^{-1} \end{pmatrix} \in A$ be arbitrary; then

$$\begin{pmatrix} 1 & \alpha\lambda_n^{-1} \\ 0 & 1 \end{pmatrix} g_n \begin{pmatrix} 1 & \alpha^{-1}\lambda_n^{-1} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ \lambda_n & \alpha^{-1} \end{pmatrix}.$$

That is, $\phi(a) = \lim_{n\to\infty} \phi(g_n)$. Since g_n does not depend on a, we conclude that ϕ is constant on A; therefore, $\langle T_a f, f \rangle = \langle f, f \rangle$, and by Cauchy-Schwarz it must be that a acts on f by multiplication by some constant $\chi(a)$. However, it is easy to see that this constant must be 1: that is, f is invariant under A.

We can now follow the same logic as for the geodesic flow. □

This proof relied essentially on the interaction between the geodesic and the horocycle flow: the geodesic flow contracts one direction of the horocycle flow and expands the other direction.

2.6.3 More on the geodesic flow on compact surfaces

The dynamics of the geodesic flow on compact hyperbolic surfaces is quite complicated. As an example, we prove the following theorem. It is not the main thrust of this thesis; however, it is an important result in the theory of dynamical systems on compact hyperbolic surfaces.

Theorem 2.23. *Let Γ be a discrete group of fixed-point-free isometries of $\mathbb{H}$ such that $M = \mathbb{H}/\Gamma$ is compact. Then the periodic orbits of the geodesic flow on the unit tangent bundle of M are dense in the unit tangent bundle of M.*

Proof. Our strategy is as follows: let v be a unit tangent vector of M, and let w be some lift of it to $\mathbb{H}$. Let c, $\tilde{c}$ be the $t = +\infty$ and $t = -\infty$ endpoints of the geodesic on $\mathbb{H}$ tangent to w. We will find an element $\gamma \in \Gamma$ such that the endpoints of its axis lie in small neighborhoods of c and $\tilde{c}$; and then among the tangent vectors to this axis, one will be close to w. This axis projects to M and is the desired closed geodesic.

Throughout this discussion, I will be working on $\mathbb{H}$, ignoring the issue of the point at ∞ (and the possibility that it is one of c and $\tilde{c}$). We can always use Möbius transformations to move the point at ∞ away from c, $\tilde{c}$.

Note that for any $\epsilon > 0$ there exists a $\delta > 0$ such that if $\operatorname{Im} z < \delta$ then any two geodesics through z of Euclidean length greater than ϵ have a mutual angle of at most $\pi/4$. Let U and V be Euclidean δ-neighborhoods of c and $\tilde{c}$.

We are now interested in the compact space $\mathbb{H}/\Gamma$. Let D be a Dirichlet domain

$$D = D_p = \{x \in \mathbb{H} : d(x, p) < d(x, \gamma p)\ \forall \gamma \in \Gamma\};$$

note that D is compact, $\gamma(D_p) = D_{\gamma p}$, and the interiors of D_p and γD_p are disjoint when $\gamma \neq 1$ because Γ is discrete. Lastly, there are only finitely many γ_i such that $D_p \cap D_{\gamma_i p} \neq \emptyset$, and the boundary of D_p is formed by finitely many geodesic segments (points equidistant from p and $\gamma_i p$). In short, D is a particularly nice fundamental domain for $\mathbb{H}/\Gamma$.

Pick a Dirichlet domain containing the lift w of v, and call it (by abuse of notation) D. We now consider the sequence of images of D that intersect our chosen geodesic c. There are images $D_1 \subset U$ and $D_2 \subset V$, and $\gamma \in \Gamma$ preserving the ordering of the sequence of images of D such that $\gamma(D_1) = D_2$. Now, for any $z \in D_1$, we have $\gamma(z) \in D_2$. Since most geodesics through z are contained in U, and the same is true for $\gamma(z)$ and V, we can (after finding a D_1 of smaller Euclidean size if necessary) find a geodesic $\kappa \subset U$ through z such that $\gamma(\kappa) \subset V$. Since γ preserved order, it must map the region bounded by κ and $\mathbb{R}$ to the region bounded outside of $\gamma(\kappa)$, which in particular contains the complement of V. We therefore conclude that γ has two fixed points on $\mathbb{R}$: one inside U and another inside V. Consequently, the set of fixed points of γ is uniformly close to c; in particular, it has a tangent vector close to $w = \dot{c}(0)$ (which can get arbitrarily closer as δ shrinks). □

2.6.4 $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$ as the space of unimodular lattices

Definition 2.24. A subset L of $\mathbb{R}^n$ is called a *lattice* if it is a module of rank n whose basis is a basis for $\mathbb{R}^n$. Equivalently, L is a discrete subgroup of $\mathbb{R}^n$ that spans $\mathbb{R}^n$ as a subset of an $\mathbb{R}$-vector space.

Every lattice thus has a basis $v_1, \dots, v_n$ of linearly independent vectors. We derive from this two corollaries. First, $GL_n(\mathbb{R})$ acts on the space of lattices as follows: g acts on L by applying g to each element of L, in particular applying g to the $\mathbb{Z}$-basis of L. Second, every lattice L is an image of the standard lattice $\mathbb{Z}^n$ under some transformation in $GL_n(\mathbb{R})$ (the columns of g are given by the basis vectors of L).

Definition 2.25. A lattice L in $\mathbb{R}^n$ is *unimodular* if the covolume of L, that is, the volume of the compact set $\mathbb{R}^n/L$, is 1.

Let $\mathcal{L}_n$ denote the space of unimodular lattices in $\mathbb{R}^n$. $G = SL_n(\mathbb{R})$ acts on $\mathcal{L}_n$, because the covolume of gL is the covolume of L multiplied by the determinant of g. The action is transitive: every unimodular lattice is the mage of $\mathbb{Z}^n$ under some transformation in $SL_n(\mathbb{R})$. (The element of $GL_n(\mathbb{R})$ that effects the transformation must have determinant 1, since both lattices are unimodular.)

Any endomorphism of a lattice must send the basis vectors to some $\mathbb{Z}$-linear combinations of them; that is, a lattice endomorphism may be represented by a matrix with integer entries. The endomorphism is invertible if the matrix is invertible. Since the matrix determinant is an integer, this can happen only if the determinant is ± 1. Conversely, the expansion by minors formula shows that if M is a matrix with integer coefficients and $\det M = \pm 1$, then M is invertible over $\mathbb{Z}$, and thus represents an automorphism of a lattice.

We conclude that the space of all unimodular lattices in $\mathbb{R}^n$ can be identified with $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$. This geometric observation will be useful for us in the discussion of Ratner's theorems.

3 Ratner's theorems for $SL_2(\mathbb{R})/SL_2(\mathbb{Z})$

Ratner's theorems, in full generality, assert the following:

Theorem 3.1 (Ratner's measure classification theorem)**.** *Let G be a connected Lie group, and let Γ be a lattice in G. Let U be a connected Lie subgroup of G generated by one-parameter unipotent groups. Then any ergodic U-invariant measure μ on G/Γ is algebraic; that is, there exists $\overline{x} \in G/\Gamma$ and a subgroup $F \subset G$ containing U (and generated by unipotents) such that $F\overline{x}$ is closed and μ is the F-invariant probability measure on $F\overline{x}$.*

Theorem 3.2 (Ratner's orbit closure theorem)**.** *Let G be a connected Lie group, and let Γ be a lattice in G. Let H be a connected Lie subgroup of G generated by one-parameter unipotent groups. Then for any $x \in G/\Gamma$ there exists a closed connected subgroup $P \supset H$ such that $\overline{Hx} = Px$ and Px admits a P-invariant probability measure.*

There are several other theorems included in the term "Ratner's theorems," including several quantitative results. These will not be discussed in any degree of detail in this thesis. This ensemble of theorems, proved by M. Ratner in 1990s, fully settled the Ranghunathan conjectures on the orbits of unipotent flows.

A good introductory example to Ratner's theorems is the case of a line on a torus, Section 2.5.2: Ratner's orbit closure theorem can be thought of as a sweeping generalization of the result that a line on a torus is either closed or dense.

We will restrict our attention to $G = SL_2(\mathbb{R})$ and $\Gamma = SL_2(\mathbb{Z})$. In spirit the proof is similar for $SL_2(\mathbb{R})/\Gamma$ for an arbitrary lattice $\Gamma \subset SL_2(\mathbb{R})$; however, the arguments are simpler and more explicit in this particular case. The arguments for a general Lie group are mostly present for the case of $G = SL_3(\mathbb{R})$; they are significantly more complicated than the case of $SL_2(\mathbb{R})$, and therefore not addressed here.

3.1 Preliminaries

Recall that a lattice L in $\mathbb{R}^n$ is *unimodular* if the covolume of L, that is, the volume of $\mathbb{R}^n/L$, is 1. Let $\mathcal{L}_n$ denote the space of unimodular lattices in $\mathbb{R}^n$. $G = SL_n(\mathbb{R})$ acts on $\mathcal{L}_n$ by acting on the $\mathbb{Z}$-basis of the lattice. The action is transitive, and the stabilizer of the standard lattice $\mathbb{Z}^n$ is $\Gamma = SL(n, \mathbb{Z})$. Therefore, $\mathcal{L}_n$ is identified with G/Γ. A right-invariant metric on G will descend to G/Γ (and thus to $\mathcal{L}_n$).

For $\epsilon > 0$, let $\mathcal{L}_n(\epsilon) \subset \mathcal{L}_n$ denote the set of lattices whose shortest non-zero vector has length at least ϵ.

Theorem 3.3 (Mahler compactness)**.** *For any $\epsilon > 0$ the set $\mathcal{L}_n(\epsilon)$ is compact.*

A proof of this deep result will appear at the end of this section.

We will now consider $n = 2$. Given a pair of vectors v_1, v_2 such that $\begin{pmatrix} v_1 & v_2 \end{pmatrix} \in G$ (that is, $\det \begin{pmatrix} v_1 & v_2 \end{pmatrix} = 1$), we can find a unique rotation matrix $k \in K = SO_2(\mathbb{R})$ so that kv_1 is pointing along the positive x-axis and kv_2 is in the upper half-plane. The map $(v_1\ v_2) \mapsto kv_2$ gives an identification of $K\backslash G$ with the complex upper half-plane. G (and in particular $\Gamma \subset G$) acts on $K\backslash G$ by multiplication on the right; this is a variant of the usual action by fractional linear transformations. Section 2.5.3 and a few following ones have more details on the geometry of this construction; we take the quotient $K\backslash G$ instead of G here because we are interested in $\mathbb{H}$ rather than its unit tangent bundle.

We recall a few things about the geodesic and horocycle flows. Let

$$u_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \quad a_t = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \quad v_t = \begin{pmatrix} t & 0 \\ t & 1 \end{pmatrix}$$

and let $U = \{u_t\}_{t\in\mathbb{R}}$, $A = \{a_t\}_{t\in\mathbb{R}}$, $V = \{v_t\}_{t\in\mathbb{R}}$. The action of U on $G = SL_2(\mathbb{R})$ by left multiplication is the horocycle flow, and the action of A on $SL_2(\mathbb{R})$ the geodesic flow. It is worth repeating that that U, A, V are subsets of $G = SL_2(\mathbb{R})$, and therefore act on $\mathcal{L}_2$. The basic commutation relations are

$$a_t u_s a_t^{-1} = u_{e^{2t}s} \quad a_t v_s a_t^{-1} = v_{e^{-2t}s};$$

that is, conjugation by a_t for $t > 0$ contracts V and expands U. Projecting the orbits of the geodesic flow to $K\backslash G$ gives vertical lines or semicircles orthogonal to the x-axis; projecting the orbits of the horocycle flow gives horizontal lines or circles tangent to the x-axis.

Finally, we define flowboxes, which will be our basic open sets of interest. Let $W_+ \subset U$, $W_- \subset V$, $W_0 \subset A$ be (images of) open intervals containing 0 (i.e. the identity matrix). A *flowbox* is a subset of G of the form $W_+W_0W_-g$ for some $g \in G$; it is an open set containing g. (Recall that right multiplication by g is an isometry, so the flow box is isometric to $W_+W_0W_-$.)

Now we obtain a few results in the spirit of Margulis's lemma (Lemma 5.13), but simpler. They are specific to dimension 2.

Lemma 3.4. *There exists an absolute constant $\epsilon > 0$ such that the following holds: let $L \in \mathcal{L}_2$ be a unimodular lattice, then L does not contain two linearly independent vectors each of length less than ϵ.*

Proof. Let v_1 be the shortest vector in L, and let v_2 be the shortest vector linearly independent of v_1. Then the covolume of L is $\leq \|v_1\|\|v_2\|$, which therefore must be at least 1. Thus, we may choose $\epsilon = 1$. □

Lemma 3.5. *Let $L \subset \mathcal{L}_2$ be a unimodular lattice. If L does not contain a horizontal vector, then there exists $t \geq 0$ such that $a_t^{-1}L \subset \mathcal{L}_2(\epsilon)$.*

Consequently, there exists a sequence of $t_n \to \infty$ such that $a_{t_n}^{-1}L \subset \mathcal{L}_2(\epsilon)$.

Proof. Suppose L does not contain a horizontal vector, and $L \notin \mathcal{L}_2(\epsilon)$. Then L contains a vector v of norm less than ϵ, which is not horizontal. Note that a_t^{-1} stretches the second coordinate of v, so in particular there exists a smallest $t_0 > 0$ such that $\|a_{t_0}^{-1}v\| = \epsilon$. Since for $t \in [0, t_0)$, the lattice $a_t^{-1}L$ contains no vectors shorter than ϵ except $a_t^{-1}v$ and possibly multiples of it, we derive that $a_{t_0}^{-1}L \in \mathcal{L}_2(\epsilon)$. □

Proof of Mahler compactness. The proof is not entirely self-contained; more details can be found in [1, Chapter V.3]. While the proof of the theorem is not entirely in line with the main thrust of the argument, we do rely heavily on Mahler compactness in our treatment of lattices, and a sketch of the proof is appropriate.

We use the Iwasawa decomposition of $G = SL_n(\mathbb{R})$,

$$G = KAN$$

where $K = SO_n(\mathbb{R})$, $A \subset SL_n(\mathbb{R})$ is the subgroup of diagonal matrices with positive entries, and N is the subgroup of unipotent upper triangular matrices. In dimension 2 this decomposition has the following interpretation: identify $SL_2(\mathbb{R})$ with the unit tangent bundle to the complex upper half-plane $\mathbb{H}$; now for any vector v we draw the horocycle H whose inward normal is v, find the geodesic normal to H that passes through i, and rotate upon arrival until v matches up with the unit upward vector at i. (This process was also described in Section 2.5.3.)

It can be shown that there exist real numbers t, u such that

$$G = \Sigma_{t,u}\Gamma, \quad \Sigma_{t,u} = KA_tN_u$$

where $A_t = \{a \in A : a_i/a_{i+1} \leq t\}$, and $N_u = \{n \in N : |n_{ij}| \leq u\}$. (For $n = 2$ it is not hard to see, by mapping $SL_2(\mathbb{R})$ onto the complex plane, that $G = \Sigma_{2/\sqrt{3},1/2}\Gamma$.) Consequently, $L_n(\epsilon) = M'\mathbb{Z}^n$ for some subset M' of $\Sigma_{s,t}$; and it suffices to show that M' is compact.

Since the N-component of $\Sigma_{s,t}$ is compact already, we will focus our attention on the A-component. We will show that the matrix entries $(a_g)_i$ of the A-component of all $g \in M'$ can be bounded by $0 < \alpha \leq (a_g)_i \leq \beta$ for some constants α, β; this is sufficient.

Indeed, $(a_g)_1 = \|g(e_1)\| \geq \epsilon$ for all $g \in M'$; the lower bound now follows from $a_g \in A_t$. On the other hand, given a lower bound on the matrix entries, the condition $\det a_g = 1$ forces an upper bound as well. □

3.2 Measure classification

We are now ready to classify U-invariant measures on $\mathcal{L}_2 = SL_2(\mathbb{R})/SL_2(\mathbb{Z})$.

Lemma 3.6. *For $L \in \mathcal{L}_2$, the U-orbit of L is closed if and only if L contains a horizontal vector.*

Proof. Note that the action of U preserves the y-components of vectors, and fixes horizontal vectors. Therefore, if $v \in L$ is a horizontal vector, then v is contained in u_tL for all t, and therefore is contained in $\overline{U(L)}$. Now, let a matrix for L be $\begin{pmatrix} a & c \\ 0 & d \end{pmatrix}$ containing the fundamental horizontal vector $\begin{pmatrix} a \\ 0 \end{pmatrix}$. Then all vectors in $U(L)$ will have y-components that are multiples of d, and in particular the horizontal vectors in $U(L)$ will be the same as those in L. Consequently, the matrix of any lattice L' in $\overline{U(L)}$ can be written as $\begin{pmatrix} a & c' \\ 0 & d' \end{pmatrix}$. Note that the covolume of the lattice is $|ad|$, and therefore $d' = \pm d$; without loss of generality, let $d' = d$. We finally observe that $c' = c + td$ for some t since $d \neq 0$ (otherwise L is not a lattice); therefore, $L' = u_t(L)$.

On the other hand, suppose L does not contain a horizontal vector; then it is generated by two vectors whose y-coordinates are incommensurable. In particular, L contains vectors whose y-coordinates are arbitrarily close to 0. Let $v_n \in L$ be primitive vectors satisfying

$0 < (v_n)_y < \frac{1}{n}$. Pick t such that $u_n = u_t v_n = \begin{pmatrix} 1 \\ (v_n)_y \end{pmatrix} \in u_t(L)$; and find a second vector generating the lattice $u_t(L)$. It can be chosen so that its x-coordinate is in $[0,1]$; the y-coordinate must be approximately 1 because the covolume of the lattice is 1. Letting $(v_2)_n$ be the sequence of such second vectors, we note that all the $(v_2)_n$ are contained in a compact set, and therefore have a converging sequence. Then the sequence of pairs $(u_{n_k}, (v_2)_{n_k})$ converges to some pair of generators for a lattice with the first vector horizontal: that is, $U(L) \neq \overline{U(L)}$ in this case. □

Now, any closed U-orbit supports a U-invariant probability measure. Moreover, we have the Haar measure ν on $\mathcal{L}_2 = G/\Gamma$, normalized so that $\nu(\mathcal{L}_2) = 1$; this ν is ergodic for both the horocycle and the geodesic flows. Ratner's measure classification theorem asserts that these are the only U-invariant ergodic probability measures on $\mathcal{L}_2$.

Theorem 3.7 (Measure classification, dimension 2)**.** *Let μ be an ergodic U-invariant probability measure on $\mathcal{L}_2$. Then either μ is supported on a closed orbit, or μ is the Haar measure ν.*

Proof. Let $\mathcal{L}_2' \subset \mathcal{L}_2$ denote the U-invariant set of lattices which contain a horizontal vector.

If μ is an ergodic U-invariant measure on $\mathcal{L}_2$, then either $\mu(\mathcal{L}_2') = 0$ or $\mu(\mathcal{L}_2') = 1$. We first show that if the latter holds, then μ is supported on a single closed orbit.

Parametrize orbits in $\mathcal{L}_2'$ by a, the length of the primitive horizontal vector. Let S_n be the set of lattices with a-coordinate in $[n, n+1)$; by ergodicity, $\mu(S_n)$ must be equal to 1 for some unique n. By further partitioning that interval, we will determine a unique value of a such that μ is supported on that particular orbit.

We now restrict our attention to the case of $\mu(\mathcal{L}_2') = 0$, and seek to show that $\mu = \nu$. Let $f : \mathcal{L}_2 \to \mathbb{R}$ be a compactly supported continuous function, and let $\epsilon > 0$. By uniform continuity of f we may find neighborhoods of the identity $W_0' \subset A$, $W_-' \subset V$ such that for $a \in W_0'$, $v \in W_-'$, and any $L \in \mathcal{L}_2$,

$$|f(vaL) - f(L)| < \epsilon/3 \quad (\forall L \in \mathcal{L}_2). \tag{1}$$

Let $W_+ \subset U$, $W_0 \subset A$, and $W_- \subset V$ be small enough neighborhoods of the identity such that for all $g \in G$ with $\pi(g) \in \mathcal{L}_2(\epsilon)$ the restriction of π to the flowbox $W_0 W_0 W_+ g$ is injective. (Here, π is the natural projection $G \to G/\Gamma = \mathcal{L}_2$; such a flowbox exists because $\mathcal{L}_2(\epsilon)$ is compact.) Restricting to a smaller flowbox if necessary, we may assume $W_- \subset W_-'$ and $W_0 \subset W_0'$. Let $\delta = \nu(W_- W_0 W_+)$ be the Lebesgue measure of the flowbox.

Applying a corollary of the Birkhoff ergodic theorem (Lemma 2.16) to the Lebesgue measure ν (for which the unipotent flow is ergodic), there exists a set $E \subset \mathcal{L}_2$ with $\nu(E) < \delta$ and $T_1 > 0$ such that for any interval I containing the origin of length $|I| \geq T_1$ and any lattice $L' \notin E$,

$$\left| \frac{1}{|I|} \int_I f(u_t L') dt - \int_{\mathcal{L}_2} f d\nu \right| < \epsilon/3 \quad (\forall L' \notin E). \tag{2}$$

On the other hand, applying the Birkhoff ergodic theorem (Theorem 2.13) to our measure μ, we have for μ-almost every $L \in \mathcal{L}_2$ and for some $T_2 > 0$ that, for all intervals I containing the origin and of length $|I| \geq T_2$,

$$\left| \frac{1}{|I|} \int_I f(u_t L) dt - \int_{\mathcal{L}_2} f d\mu \right| < \epsilon/3 \quad \text{(for } \mu\text{-almost every } L \in \mathcal{L}_2\text{)}. \tag{3}$$

We may assume that L does not contain horizontal vectors, since $\mu(\mathcal{L}_2') = 0$; and therefore by Lemma 3.5, we can construct a sequence $t_n \to \infty$ such that $a_{t_n}^{-1} L \in \mathcal{L}_2(\epsilon)$. Now, for any t_n as above, consider

$$Q = Q(L) = a_t W_- W_0 W_+ a_t^{-1} L = (a_t W_- a_t^{-1}) W_0 (a_t W_+ a_t^{-1}) L;$$

for large t, Q is long in the U direction and short in the A and V directions. The set Q is a copy of a flowbox in $\mathcal{L}_2$ containing L, and $\nu(Q) = \delta$.

Consider the foliation of Q by the orbits of U. For $\tilde{L} \in Q$ let $I(\tilde{L})$ be the connected component containing the origin of the set $\{t \in \mathbb{R} : u_t \tilde{L} \in Q\}$. Note that the length of $I(\tilde{L})$ is just the length of W_+ multiplied by e^{2t_n}, and in particular is independent of the choice of $\tilde{L} \in Q$. For all large enough t_n we therefore have $\left| I(\tilde{L}) \right| \geq \max(T_1, T_2)$. Now, $a_t W_- a_t^{-1} \subset W_- \subset W_-'$ and $W_0 \subset W_0'$, so applying equation (1) for any $\tilde{L} \in Q$ we have

$$\left| \frac{1}{\left| I(\tilde{L}) \right|} \int_{I(\tilde{L})} f(u_t \tilde{L}) dt - \frac{1}{|I(L)|} \int_{I(L)} f(u_t L) dt \right| < \epsilon/3 \quad (\forall \tilde{L} \in Q(L),\ L \in \mathcal{L}_2 \setminus \mathcal{L}_2'); \tag{4}$$

that is, the integral of f over each U-orbit in the foliation of Q is nearly the same, provided the above construction makes sense (that is, L does not contain any horizontal vectors).

Since $\mu(Q) > \mu(E)$, we may pick a lattice $\tilde{L} \in Q \cap E^c$; then putting together equations (2), (3), and (4) we observe that

$$\left| \int_{\mathcal{L}_2} f d\mu - \int_{\mathcal{L}_2} f d\nu \right| < \epsilon, \tag{5}$$

that is, μ is the Haar measure. □

The key property of U used in this proof is that the one-parameter unipotent subgroup U is contracted by the one-parameter diagonal subgroup, that is, that U is horospherical. This is special to $SL_2(\mathbb{R})$. This led to an early generalization of this proof to the case of a horospherical flow in higher dimensions; however, to get the result for any unipotent one-parameter flow in higher dimensions, other properties of U are needed.

3.3 Orbit closures

Theorem 3.8 (Orbit closures, dimension 2). *Let $L \in \mathcal{L}_2 = SL_2(\mathbb{R})/SL_2(\mathbb{Z})$. Then the U-orbit of L is either closed or dense.*

Proof. Suppose UL is not closed; by Lemma 3.6, this means that $L \notin \mathcal{L}_2'$. We wish to show that UL passes though every open set $\tilde{O} \subset SL_2(\mathbb{R})/SL_2(\mathbb{Z})$.

Find an open subset O of a compact subset C of $\tilde{O}$ (we like working with functions of compact support, so all of $\tilde{O}$ might be too large for us). Let f be a uniformly continuous, nonnegative function supported on C and equal to 1 on O; then $0 < \nu(O) \leq \int_{\mathcal{L}_2} f d\nu \leq \nu(\tilde{O})$. That is, we approximate the characteristic function of O by a uniformly continuous function of compact support. Let $\epsilon < \nu(O)$.

Since our U-orbit is not closed, it is the orbit of some lattice $L \notin \mathcal{L}_2'$. Let the sequence $t_n \to \infty$, the flowbox Q, and the exceptional set E be as in the proof of measure classification above. Since $\mu(Q) > \mu(E)$, for a large enough t_n we can put together (2) and (4) to find an interval I such that

$$\left| \frac{1}{|I|} \int_I f(u_t L) dt - \int_{\mathcal{L}_2} f d\nu \right| < \epsilon.$$

However, this is only possible if $f(u_t L)$ actually visits $O \subset \tilde{O}$; since $\tilde{O}$ was arbitrary, we conclude that the U-orbit of L is dense. □

4 Oppenheim conjecture

The statement of the Oppenheim conjecture is as follows:

Theorem 4.1 (Oppenheim's conjecture)**.** *Let Q be a nondegenerate indefinite quadratic form in $n \geq 3$ variables. Then either Q is proportional to a form with integer coefficients, or $Q(\mathbb{Z}^n)$ is dense in $\mathbb{R}$.*

4.1 Some history

The original statement conjectured by Oppenheim in 1929 is that for a nondegenerate, indefinite, irrational quadratic form Q in $n \geq 5$ variables there exist integers $x_1, \ldots, x_n$ such that $|Q(x_1, \ldots, x_n)| < \epsilon$. It was later extended to $n \geq 3$ by Davenport, and strengthened by Oppenheim to a statement that $0 < |Q(x_1, \ldots, x_n)| < \epsilon$; the stronger conjecture can be shown to imply that $Q(\mathbb{Z}^n)$ is dense in $\mathbb{R}$.

The conjecture was proved completely by Margulis around 1987. Prior to the attack by dynamical systems, the conjecture was studied by the methods of analytic number theory, using circle methods, but with unsatisfactory results.

Margulis's proof relies on the Ranghunathan conjectures on the closures of unipotent orbits. The proof of these conjectures in full generality is the set of Ratner's theorems on unipotent flows; Margulis settled certain special cases, which sufficed for the application to the Oppenheim conjecture. The connection between orbits of unipotent flows and the Oppenheim conjecture was already made by Ranghunathan in the mid-seventies.

Much of Margulis's work was done in collaboration with S. Dani, who proved cases of the orbit closures theorem for horospherical flows; the proof we give in dimension 2 is similar in spirit to those proofs.

4.2 Some remarks about the theorem

Remark 4.2 (The necessity of the conditions)**.** If Q is definite, the image of $\mathbb{Z}^n$ is confined to $\mathbb{R}_{\geq 0}$, and in fact is a lattice.

The requirement of $n \geq 3$ variables is also necessary. Indeed, let α be a real algebraic number of degree 2; then it is well-known that $|\alpha - p/q| \geq C/q^2$ for some constant C and all rationals p/q. Consequently, the quadratic form $Q(x, y) = y^2 - \alpha^2 x^2$ has the property that

$$|Q(x, y)| = \left|x^2(y/x - \alpha)(y/x + \alpha)\right| \geq C\,|\alpha|$$

and 0 is an isolated point in the image of Q.

Finally, if Q is degenerate, then after a suitable change of coordinates it is isomorphic to an $(n-1)$-form; and since the requirement of $n \geq 3$ variables is necessary, so is nondegeneracy.

Remark 4.3. The theorem is true if we replace $\mathbb{Z}^n$ by the set of primitive vectors (a vector $p = (p_1, \ldots, p_n)$ is primitive if $\gcd(p_1, \ldots, p_n) = 1$).

Remark 4.4. The general case of Oppenheim's conjecture can be reduced to the case of $n = 3$ variables. The argument is somewhat tedious, but straightforward; see [1, Section VI.2].

Remark 4.5. Quantitative versions of Margulis's and Ratner's work let one derive quantitative versions of the Oppenheim conjecture as well: that is, it is possible to give some asymptotics for the integer points $(x_1, \ldots, x_n) \in \mathbb{Z}^n$ on which the quadratic form Q takes small values. For example, for all a, b, T we might count

$$N_{a,b}(T) = \#\{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{Z}^n | \, |\mathbf{x}| < T, a < Q(\mathbf{x}) < b\};$$

the qualitative conjecture implies that $N_{a,b}(T) \to \infty$ as $T \to \infty$, and we might ask how quickly this happens. Lower bounds on $N_{a,b}$ were proved by Dani and Margulis in 1993 by giving a strengthened, uniform version of Ratner's theorem on the equidistribution of orbits of unipotent flows. Upper bounds proved more difficult, but sharp estimates were given by Margulis, Eskin, and Mozes (1995-2005).

5 A proof of the Oppenheim conjecture

In this section, our main concern is the connection between the number theory of the Oppenheim conjecture and the theory of dynamical systems in Ratner's theorems. The connection comes via the group $SO(Q) \subset SL_n(\mathbb{R})$, the group of transformations that leaves the quadratic form Q invariant. The theory of dynamical systems will let us show that $\overline{SO(Q)}$ must be "nice", and we'll see that this leaves only two options for it – corresponding to Q rational or $Q(\mathbb{Z}^n)$ dense in $\mathbb{R}$.

5.1 Some definitions

Definition 5.1. If Q is a quadratic form in n variables, the *(special) orthogonal group of* Q is

$$SO(Q) = \{h \in SL_n(\mathbb{R}) \,|\, Q(vh) = Q(v) \; \forall v \in \mathbb{R}^n\}.$$

We will let $SO(Q)^0$ be the connected component of the identity in $SO(Q)$.

Since every indefinite quadratic form has signature $(2,1)$ or $(1,2)$, and the two cases differ from each other only by an overall sign, we will let Q_0 denote the standard quadratic form of signature $(2,1)$: that is, $Q_0(x_1, x_2, x_3) = x_1^2 + x_2^2 - x_3^2$. Then our arbitrary quadratic form Q is conjugate to $\pm Q_0$. We will let $H = SO(Q_0)^0$ stand for the connected component of the identity in the special orthogonal group of Q_0.

Remark 5.2. We do not lose much generality by working with H rather than the entire special orthogonal group. $SO(Q_0)$ has only two connected components: thus, H has index 2 in $SO(Q_0)$. A proof of this classical result may be found in [8].

5.2 Outline of proof

In this section, we present a schematic proof of the Oppenheim conjecture using Ratner's theorems. The supplementary lemmas follow the main body of the argument.

We will exploit the fact that $SO(Q)$ is large, and *a priori* $SO(Q)\mathbb{Z}^n$ is much larger than $\mathbb{Z}^n$. Ratner's theorem will let us quantify this: either $SO(Q)\mathbb{Z}^3$ is dense in $\mathbb{R}^3$ or (after a few more arguments) Q must be rational. Recall the precise statement of Ratner's theorem:

Theorem 5.3 (Ratner's orbit closures theorem)**.** *Let G be a connected Lie group, and let Γ be a lattice in G. Let H be a connected Lie subgroup of G generated by unipotent one-parameter groups. Then for any $x \in G/\Gamma$ there exists a closed connected subgroup $P \subset G$ containing H such that $\overline{Hx} = Px$ and Px admits a P-invariant probability measure.*

The dichotomy in the statement of the Oppenheim conjecture results from the fact that the H in question is a maximal connected subgroup of G, and therefore there are only two possible choices for P; namely, Hx is either dense or closed, corresponding to the cases of $Q(\mathbb{Z}^n)$ dense in $\mathbb{R}$ and Q proportional to an integer form respectively.

Proof of the Oppenheim conjecture. Let $g_Q \in SL_3(\mathbb{R})$ and $\lambda \in \mathbb{R}^\times$ be such that $Q = \lambda Q_0 \circ g_Q$. In that case, $SO(Q)^0 = g_Q H g_Q^{-1}$.

Now, $H = SO(Q_0)^0 \cong SL_2(\mathbb{R})$ is generated by unipotent elements (Lemma 5.10), and $SL_3(\mathbb{Z})$ is a lattice in $SL_3(\mathbb{R})$, so we can apply Ratner's Orbit Closure Theorem (Theorem 5.3) to obtain the following:

There exists a closed, connected subgroup $P \subset SL_3(\mathbb{R})$ such that

- $H \subset P$;
- $\overline{Hg_Q} = \overline{Pg_Q}$;
- and there is an P-invariant probability measure on Pg_Q.

Lemma 5.11 shows that there are only two possibilities for a closed, connected subgroup of $SL_3(\mathbb{R})$ containing $H = SO(Q_0)$: namely, $S = H$ or $S = SL_3(\mathbb{R})$. We consider these two cases separately.

Case 1. Assume $S = SL_3(\mathbb{R})$. In that case, $SL_3(\mathbb{Z})gH$ is dense in $SL_3(\mathbb{R})$. Therefore,

$$\begin{aligned}
Q(\mathbb{Z}^3) &= Q_0(\mathbb{Z}^3 g_Q) && \text{(definition of } g_Q\text{)} \\
&= Q_0(\mathbb{Z}^3 SL_3(\mathbb{Z}) g_Q) && (\mathbb{Z}^3 = \mathbb{Z}^3 SL_3(\mathbb{Z})) \\
&= Q_0(\mathbb{Z}^3 SL_3(\mathbb{Z}) g_Q H) && (H = SO(Q_0)^0) \\
\text{is dense in } & Q_0(\mathbb{Z}^3 G) && (Q \text{ is continuous}) \\
&= Q_0(\mathbb{R}^3 - \{0\}) && (vG = \mathbb{R}^3 - \{0\} \text{ for nonzero } v) \\
&= \mathbb{R}.
\end{aligned}$$

That is, since Q is indefinite, it must map $\mathbb{R}^3$ onto $\mathbb{R}$, and we concluded that $SO(Q)\mathbb{Z}^3$ is dense in $\mathbb{R}^3$ – so its image is dense in $\mathbb{R}$.

Case 2. Assume $S = H = SO(Q_0)$. This is the degenerate case, where Q is a scalar multiple of a form with integer coefficients. We present two proofs of this: the first one relying on the theory of algebraic groups, and the second on Margulis's lemma (Lemma 5.13) in analysis. The algebraic approach is more concise, but uses fairly deep results from the theory of algebraic groups; the analytic approach is closer to the argument used by Margulis in his original 1987 proof of the Oppenheim conjecture.

5.2.1 Algebraic proof

If $S = H$, then the orbit $g_Q H = g_Q S$ has a finite H-invariant measure. Therefore, $\Gamma_{g_Q} = \Gamma \cap (g_Q H g_Q^{-1}) = SL_3(\mathbb{Z}) \cap (g_Q H g_Q^{-1})$ is a lattice in $g_Q H g_Q^{-1} = SO(Q)^0$. Since H is generated by unipotents (Lemma 5.10), Borel density theorem (Theorem 5.5) implies that $SO(Q)^0$ is contained in the Zariski closure of Γ_{g_Q}; and since $\Gamma_{g_Q} \subset \Gamma = SL_3(\mathbb{Z})$, we conclude that $SO(Q)^0$ is defined over $\mathbb{Q}$ (Lemma 5.7). Consequently, up to a scalar multiple, Q has integer coefficients (Lemma 5.8).

5.2.2 Analytic proof

This proof is closer to the original approach used by Margulis in his 1987 proof of the Oppenheim Conjecture, and relies on deep statements about the behavior of unipotent flows. The estimates derived by Margulis are weaker than Ratner's general estimates (especially the more quantitative ones); some of the spirit of the original argument is given in this section, although we avoid presenting the proof in full generality.

If $S = H$ then $g_Q H$ is closed, and therefore so is the H-orbit orbit of $g_Q SL_3(\mathbb{Z})$ in $G/\Gamma = SL_3(\mathbb{R})/SL_3(\mathbb{Z})$. Let $x = g_Q SO_3(\mathbb{Z}) \in G/\Gamma$ and $x_0 = SO_3(\mathbb{Z}) \in G/\Gamma$. Then $SO(Q)x_0 = g_Q^{-1}Hx$ is also closed. Let $\Delta = SO_3(\mathbb{Z}) \cap SO(Q)$.

Our strategy will be to show that there exist real symmetric 3×3 matrices S satisfying $\gamma^t S \gamma = S$ for all $\gamma \in \Delta$, and that all such matrices correspond to quadratic forms that are proportional to Q. Since this system of equations for S is defined over the integers, if it has some solution, it will have a rational solution – yielding a rational quadratic form proportional to Q.

In terms of quadratic forms, $\gamma^t S \gamma = S$ means $\Delta \subset SO(Q')$ for a quadratic form Q'. Existence of such a Q', therefore, is trivial: $\Delta \subset SO(Q)$. The difficult part will be to show that if $\Delta \subset SO(Q')$ then Q and Q' are proportional.

Now, $SO(Q)^0$ is similar to $SO(Q)$ (as follows from Remark 5.2, it is an index-2 subgroup), but it is generated by unipotent one-parameter subgroups (Lemma 5.10). We will show that $\Delta \subset SO(Q')$ implies that all unipotent 1-parameter subgroups of $SO(Q)$ are contained in $SO(Q')$, and hence $SO(Q)^0 \subset SO(Q')$.

Fix a point $p \in \mathbb{R}^3$, and consider $f_p : SO(Q) \to \mathbb{R}$, $g \mapsto Q'(g^{-1}p)$. If $\Delta \subset SO(Q')$, then f_p factors through Δ to a continuous function

$$\tilde{f}_p : SO(Q)/\Delta \to \mathbb{R}.$$

Now, let $\{u(t)\}_{t\in\mathbb{R}} \subset SO(Q)$ be a unipotent one-parameter subgroup. The function $q : \mathbb{R} \to \mathbb{R}$, $t \mapsto f_p(u(t))$ is polynomial in t, since the entries of $u(t)$ are polynomial in t (Lemma 5.14).

We now invoke Margulis's Lemma (Lemma 5.13) to produce $K \subset G/\Gamma$ compact such that the set $\{t \geq 0 : u(t)x_0 \in K\}$ is unbounded: that is, a compact set to which the $u(t)$-orbit of x_0 returns infinitely often.

Since $SO(Q)x_0$ is closed, the map $\phi : SO(Q)/\Delta \to SO(Q)x_0$ via $g\Delta \mapsto gx_0$ is a homeomorphism, and $K' = \phi^{-1}(K)$ is compact. Therefore, $\tilde{f}_p(K')$ is a compact subset of $\mathbb{R}$. On the other hand, K was chosen so that $\{t \in \mathbb{R} : q(t) \in f_p(K')\}$ is unbounded, implying that q is the constant polynomial.

That is, $f_p(u(t))$ is constant, and $Q'(u(t)p) = Q'(p)$ for all $t \in \mathbb{R}$. Since p was arbitrary, this holds at every $p \in \mathbb{R}^3$, implying that $\{u(t)\}_{t\in\mathbb{R}} \subset SO(Q')$.

We can therefore conclude that if $\Delta \subset SO(Q')$ then $SO(Q)^0 \subset SO(Q')$.

Now, let σ and σ' be the symmetric matrices corresponding to Q and Q' respectively. We have for all $h \in SO(Q)^0$,

$$h\sigma'\sigma^{-1}h^{-1} = (h\sigma' h^t)((h^{-1})^t \sigma^{-1} h^{-1}) = \sigma'\sigma^{-1}.$$

Now, $H = SO(Q_0)^0$ is centralized only by scalars (Lemma 5.12), and the same holds for $SO(Q)^0$ since it is conjugate to H. Therefore, $\sigma\sigma^{-1}$ is a scalar, i.e. the two matrices are proportional. This concludes the proof that Q is proportional to a rational matrix. □

5.3 Lemmas for the algebraic approach

Definition 5.4. A subset $H \subset SL_l(\mathbb{R})$ is *Zariski closed* if there exists a subset $S \subset \mathbb{R}[x_{1,1}, \ldots, x_{l,l}]$ such that $H = \{g \in SL_l(\mathbb{R}) \,|\, Q(g) = 0 \ \forall Q \in S\}$, where we understand

$Q(g)$ to denote the value obtained by substituting the matrix entries $g_{i,j}$ into the variables $x_{i,j}$. That is, H is Zariski closed "if the matrix entries are characterized by polynomials".

For $H \subset SL_l(\mathbb{R})$, let $\overline{\overline{H}}$ denote the Zariski closure of H, that is, the unique smallest Zariski closed set containing H.

Lemma 5.5 (Borel Density Theorem)**.** *Let $G \subset SL_l(\mathbb{R})$ be a closed subgroup, and let Γ be a lattice in G. Then the Zariski closure $\overline{\overline{\Gamma}}$ of Γ contains every unipotent element of G.*

Before we prove this, we introduce another lemma:

Lemma 5.6. *Let $g \in SL_m(\mathbb{R})$ be unipotent, and let μ be a g-invariant probability measure on $P\mathbb{R}^{m-1} = (\mathbb{R}^m)^\times/\mathbb{R}^\times$. Then μ is supported on the set of fixed points of g.*

Proof. Let $T = g - I$ (then T is nilpotent), and let $v \in (\mathbb{R}^m)^\times$. Let r be such that $vT^r \neq 0$ but $vT^{r+1} = 0$. Then $gT^r v = T^r v$, so $[T^r v]$ is a fixed point of g.

On the other hand, it is easy to see $g^n[v] \to [T^r v]$ as $n \to \infty$. By Poincaré recurrence (Theorem 2.5), for μ-every $[v] \in P\mathbb{R}^{m-1}$ there exists a sequence $n_k \to \infty$ such that $g^{n_k}[v] \to [v]$. Since we know that $g^n[v]$ converges to a fixed point of g as $n \to \infty$, we conclude that μ-every point is a fixed point, i.e. μ is supported on the set of fixed points of g. □

Proof of the Borel Density Theorem. By Chevalley's theorem (which will not be proved here, and is standard material – see, for example, [9], or a good text on number theory), there exists a polynomial homomorphism $\rho : SL(l, \mathbb{R}) \to SL(m, \mathbb{R})$ for some m, and a vector $[v] \in P\mathbb{R}^{m-1}$, such that

$$\overline{\overline{\Gamma}} = \{g \in SL(l, \mathbb{R}) \mid [v]\rho(g) = [v]\}.$$

Therefore, ρ induces a well-defined map on $G/\Gamma \to P\mathbb{R}^{m-1}$:

$$\overline{\rho}(g\Gamma) = \rho(g)[v].$$

Let $g \in G$ be unipotent, and let μ_0 be a G-invariant probability measure on G/Γ. This is pushed to a $\rho(G)$-invariant measure on $P\mathbb{R}^{m-1}$ defined by $\mu(A) = \mu_0(\overline{\rho}^{-1}(A))$; and since $\rho(g)$ is unipotent, by the preceding lemma, μ is supported on the set of fixed points of $\rho(g)$. However, it is not hard to show that $[v]$ lies in the support of μ; and therefore $\rho(g)$ must fix $[v]$, from which $g \in \overline{\overline{\Gamma}}$. □

Lemma 5.7. *Let C be a subset of $SL_l(\mathbb{Q})$; then $\overline{\overline{C}}$ is defined over $\mathbb{Q}$.*

Proof. Suppose $\overline{\overline{C}}$ is defined by $S \subset P^d$, where P^d is the set of all polynomials of degree $\leq d$. Now, the subspace $\{Q \subset P^d : Q(C) = 0\}$ is defined by linear equations with rational coefficients; and therefore it is spanned by some rational vectors, which therefore determine the set S. □

Lemma 5.8. *For a nondegenerate quadratic form Q, $SO(Q)$ is defined over $\mathbb{Q}$ if and only if Q is proportional to a form with rational coefficients.*

Proof. If Q is a rational form, then $SO(Q)$ is quite apparently defined over $\mathbb{Q}$; note that $SO(Q)$ does not depend on the scaling of Q.

Conversely, given $SO(Q)$ defined over $\mathbb{Q}$, Q is uniquely determined up to scalar multiplication. Consider an automorphism ϕ of $\mathbb{R}/\mathbb{Q}$, and notice that $SO(\phi Q) = \phi SO(Q) = SO(Q)$; that is, ϕ must send Q to a scalar multiple of itself. Now scale Q so that it has one rational coordinate; that coordinate will be fixed by ϕ, and therefore the scalar multiple must in fact be 1. That is, the scaled Q is invariant under all the automorphisms of $\mathbb{R}/\mathbb{Q}$, and consequently Q is proportional to a form with rational coefficients. □

5.4 Lemmas for the analytic approach

Lemma 5.9. *The Lie algebra of $SO(Q_0)$ is*

$$\mathfrak{so}(Q_0) = \left\{ X_{a,b,c} = \begin{pmatrix} 0 & a & b \\ a & 0 & c \\ b & -c & 0 \end{pmatrix} | a, b, c \in \mathbb{R} \right\}$$

Proof. This is verified by direct computation. Let $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^t$, then $Q_0(x) = x_1^2 + x_2^2 - x_3^2$. Now, the Lie algebra consists of matrices M such that $Q_0((I + \epsilon M)\mathbf{x}) - Q_0(\mathbf{x}) = O(\epsilon^2)$. Writing this out explicitly, if $M = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$ then

$$Q_0((I+\epsilon M)\mathbf{x}) - Q_0(\mathbf{x}) = 2\epsilon((ax_1+bx_2+cx_3)x_1+(dx_1+ex_2+fx_3)x_2-(gx_1+hx_2+ix_3)x_3)+O(\epsilon^2),$$

from which we conclude that $a = e = i = 0$, and $b = d$, $c = g$, and $f = -h$. Therefore, the matrices have the desired form.

We compute the Lie brackets:

$$\begin{aligned} [X_{0,0,1}, X_{0,1,0}] &= X_{1,0,0} \\ [X_{1,0,0}, X_{0,1,0}] &= X_{0,0,1} \\ [X_{1,0,1}, X_{0,0,1}] &= X_{0,1,0} \end{aligned}$$

Note that over $\mathbb{C}$ we can let $u = X_{i,1,0}$, $v = X_{-i,1,0}$, and $h = X_{0,0,2i}$; these clearly generate the same Lie algebra, and satisfy the commutation relations $[u, v] = h$, $[u, h] = -2u$ and $[v, h] = 2v$, the generating equations for $\mathfrak{sl}_2(\mathbb{C})$. □

Lemma 5.10. $H = SO(2, 1)^0$ *is isomorphic to* $SL_2(\mathbb{R})$ *and generated by unipotent elements.*

Proof. We observe that the determinant on $\mathfrak{sl}(2, \mathbb{R})$ has signature $(2, 1)$, and the adjoint representation $\mathrm{Ad}_{SL(2,\mathbb{R})}$ maps $SL(2, \mathbb{R})$ into $SO(\det)$. From this we conclude that $SL(2, \mathbb{R})$ is locally isomorphic to $SO(2, 1)^0$, and since $SL(2, \mathbb{R})$ is generated by unipotents and $SO(2, 1)^0$ is connected, we conclude that $SO(2, 1)^0$ is generated by unipotents.

Equivalently, we could conclude this from the direct characterization of $\mathfrak{so}(2, 1)$ above: since the Lie algebra is isomorphic to $\mathfrak{sl}_2(\mathbb{R})$ and generated by nilpotents, the connected component of the identity of the Lie group is generated by unipotents. □

Lemma 5.11. *$SO(2,1)^0$ is a maximal connected subgroup of $SL_3(\mathbb{C})$. Equivalently, the Lie algebra $\mathfrak{so}(2,1)$ is a maximal subalgebra of the Lie algebra $\mathfrak{sl}_3(\mathbb{C})$.*

Proof. The equivalence is a standard theorem in Lie theory; see [4, Theorem 2.1].

We now use the characterization of representations of $\mathfrak{sl}_2(\mathbb{C})$; see [8, Chapter 4.2]. Namely, a representation V of $\mathfrak{sl}_2(\mathbb{C})$ over $\mathbb{C}$ has a basis $\beta_1, \ldots, \beta_j$ such that

$$\begin{aligned} &h\beta_n = (j-2n)v_n, && n = 0, 1, \ldots, j; \\ &u\beta_j = 0, \quad u\beta_n = v_{n+1}, && n = 0, 1, \ldots, j-1; \\ &v\beta_0 = 0, \quad v\beta_n = n(j-n+1)\beta_{n-1}, && n = 1, \ldots, j. \end{aligned}$$

Here, u, v, h are the generators of $\mathfrak{sl}_2(\mathbb{C})$ satisfying $[u,h] = 2u$, $[v,h] = -2v$, and $[u,v] = h$; we abuse notation to let multiplication denote the action of the elements of $\mathfrak{sl}_2(\mathbb{C})$ on V.

From this it is easy to see that no proper $\mathfrak{sl}_2(\mathbb{R})$-invariant subspace of V contains $\ker u$. By the lemma above we know that $\mathfrak{so}(2,1)$ is isomorphic to $\mathfrak{sl}_2(\mathbb{C})$, and consequently the same is true of $\mathfrak{so}(2,1)$.

Now, suppose $\mathfrak{so}(2,1) \subsetneq \mathfrak{h} \subsetneq \mathfrak{sl}_3(\mathbb{C})$. Use the adjoint representation, and note that $u = X_{i,1,0} \in \mathfrak{so}(2,1)$ has a kernel of dimension 2 in $\mathfrak{sl}_3(\mathbb{C})$: by direct computation, elements of the kernel have the form $\begin{pmatrix} 0 & a & -ia \\ a & b & -ib \\ -ia & -ib & -b \end{pmatrix}$. In particular, $X_{i,1,0}$ is in the kernel of u. Since no proper $\mathfrak{sl}_2(\mathbb{C})$-invariant subspace of any $\mathfrak{sl}_2(\mathbb{C})$-module can contain the kernel of u, we see that a proper containment of Lie algebras

$$\mathfrak{so}(2,1) \subsetneq \mathfrak{h} \subsetneq \mathfrak{sl}_3(\mathbb{C})$$

would imply a proper containment of kernels

$$\big(\mathfrak{so}(2,1) \cap \ker u\big) \subsetneq \big(\mathfrak{h} \cap \ker u\big) \subsetneq \big(\mathfrak{sl}_3(\mathbb{C}) \cap \ker u\big).$$

However, the first of these spaces has dimension 1 and the last one has dimension 2: therefore, there is nothing properly contained between them, and we conclude that in fact, we must have had $\mathfrak{h} = \mathfrak{so}(2,1)$ or $\mathfrak{h} = \mathfrak{sl}_3(\mathbb{C})$. □

Lemma 5.12. *The centralizer of $SO(Q_0)$ in $GL_3(\mathbb{R})$ consists of multiples of the identity.*

Proof. We use the computation of the Lie algebra $\mathfrak{so}(2,1)$ of $SO(Q_0) = SO(2,1)$ given above. If a matrix $A \in GL_3(\mathbb{R})$ commutes with all the matrices from H, then it must commute with all the matrices from $\mathfrak{so}(2,1)$, and in particular with the diagonal matrix $X_{1,0,0}$ with diagonal entries 1, 0, -1. Therefore, A must be a diagonal matrix, and since it also commutes with $X_{0,1,0}$ it must be a scalar. □

Lemma 5.13 (Margulis's Lemma). *Let $n \geq 2$. Let $\{u_t\}_{t\in\mathbb{R}}$ be a unipotent one-parameter subgroup of $SL_n(\mathbb{R})$, and let $x \in SL_n(\mathbb{R})/SL_n(\mathbb{Z})$. Then there exists a compact set $K \subset SL_n(\mathbb{R})/SL_n(\mathbb{Z})$ such that $\{t \geq 0 | u_t x \in K\}$ is unbounded: that is, $u_t x$ does not tend to infinity as $t \to \infty$.*

Proof. In general, this is a rather difficult result. We present a proof in dimension 2; the case of dimension 3 is more similar to the general case, and consequently harder. In dimension 2, however, we can use Lemma 3.4: there exists a universal constant $\epsilon > 0$ such that no unimodular lattice in $\mathbb{R}$ contains two linearly independent vectors of length $< \epsilon$.

By Mahler's compactness criterion (Theorem 3.3), the set of lattices that contain no vectors shorter than ϵ is compact. Thus, if the lemma were false, we would conclude that $\forall \epsilon$ $\exists t_0 > 0$ such that $\forall t \geq t_0$, the lattice $u_t \Lambda$ contains a vector of length $< \epsilon$.

Notice that for any fixed $v \in \Lambda$ we have $\|u_t v\| \to \infty$ as $t \to \infty$ (u_t acts by shifting the x-coordinate of v and fixing the y-coordinate). Consequently, it is not the case that there exists a single vector $v \in \Lambda$ such that $u_t v$ (or $u_t \lambda v$) has has length $< \epsilon$ for all $t \geq t_0$.

Therefore, it must be the case that at some time t the shortest vector, say $u_t v$, has length $< \epsilon$; and at some slightly later time t' the shortest vector will become the image of some $w \neq \lambda v$, and will have $u_{t'} w$ of length $< \epsilon$. However, $u_{t'} v$ (or $u_{t'}(\lambda v)$ for some λv in the lattice) must still have length $< \epsilon$ at time t' by continuity – a contradiction, since $u_t \Lambda$ is unimodular. □

Note that this proof is very specific to dimension 2. In dimensions 3 and higher, Margulis's lemma becomes much more complicated; a full proof may be found in [1].

Lemma 5.14. *Let $\{u_t\}_{t \in \mathbb{R}}$ be a unipotent one-parameter subgroup of $\mathbb{R}^n$. Then the matrix entries of u_t are polynomial in t.*

Proof. If $\{u_t\}$ is a unipotent one-parameter subgroup then $u_t = \exp(tA)$ for some nilpotent A: say $A^n = 0$. Therefore, the nth t-derivative of u_t is $A^n u_t = 0$. Consequently, the matrix entries are polynomials of degree at most $n - 1$. □

6 Further reading

Much of the general dynamical systems background in Section 2 follows [2]; the additional background on the geodesic and horocycle flows comes comes from [6], [1]. The general outline of Section 3 comes from [3], although many of the proofs here are more complete. (We use [7] to fill in the details.) Historical notes on the Oppenheim conjecture follow [5]; the section on its proof is based on [9] for the algebraic approach and [1] for the analytic one. The exposition here fills in many of the the details that [9] relegates to exercises.

References

[1] M. Bekka, M. Mayer, *Ergodic Theory and Topological Dynamics of Group Actions on Homogeneous Spaces*, Cambridge University Press 2000.

[2] M. Brin, G. Stuck, *Introduction to Dynamical Systems*, Cambridge University Press 2002.

[3] A. Eskin and D. Kleinbock, *Unipotent Flows and Applications: Lecture Notes for Clay Institute Summer School*, June 11 2007.

[4] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*, American Mathematical Society Graduate Studies in Mathematics v.34, 2001.

[5] L. Ji, *A Summary of the Work of Gregory Margulis*, Pure and Applied Mathematics Quaterly v.34 (2008), n0.1, 1-69.

[6] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.

[7] M. Ratner, *Raghunathan's conjectures for $SL(2, \mathbb{R})$*, Israel J. Math. 80 (1992), no. 1-2, 131.

[8] V.S. Varadarajan, *Lie groups, Lie algebras, and their representations*, Springer Graduate Texts in Mathematics v.102, 2001.

[9] D. Witte-Morris, *Ratner's Theorems on Unipotent Flows*, University of Chicago Press 2005.