Introduction to
Modern Mathematics
ALM 33, pp. 171–201

Homogeneous Dynamics: A Study Guide

Manfred Einsiedler* and Thomas Ward†

Abstract

These notes give a summary of the course "Ergodic Theory and Applications in Number Theory" at the Summer School in Modern Mathematics at Tsinghua University in Beijing, June 23–27 (2013). As in the summer school, we will need to be brief at times and refer to the references for a detailed treatment. Nonetheless we wish to survey some results and ideas on a close to geodesic journey from the most basic concepts of ergodic theory to some more sophisticated and recent results in homogeneous dynamics.

2000 Mathematics Subject Classification: 37A35, 37A45, 37D40, 11J25, 11J13.

Keywords and Phrases: Ergodic theory, homogeneous dynamics, nilflow.

1 Introduction

In these notes we will try and give an overview and road map to the area of homogeneous dynamics without becoming too diverted by details. For much of the background material we refer to [3] or [5], and for the more advanced material we refer to [4] or [7]. Sections 3–7 of these notes are essentially taken from a preliminary version of [4].

1.1 What is ergodic theory?

A vague answer to this question is that ergodic theory is *the study of dynamical systems from a probabilistic point of view.* In the next few sections we will expand on the meaning of these concepts. In the next section we will also introduce–at times in vague terms–some other key concepts for dynamical systems.

1.2 What is a dynamical system?

We will use more sophisticated settings later, but for now we define a dynamical system (X, T) to be a space X, usually equipped with a topology, together with a

*ETH Zürich, Departement Mathematik, Rämistrasse 101, 8092 Zürich, Switzerland. Email: manfred.einsiedler@math.ethz.ch

†Executive Office, Palatine Centre, Durham University, DH1 3LE, England. Email: t.b.ward@durham.ac.uk

map $T : X \to X$, usually assumed to be continuous. We will also refer to such a map as a transformation.

In a dynamical system (X, T) we will often be interested in the orbit

$$\mathsf{O}_T(x) = \{x, Tx, T^2x, \dots\}$$

of a point $x \in X$. If X is a topological space we can try to describe the closure $\overline{\mathsf{O}_T(x)}$ of the orbit of a point $x \in X$. It is also interesting to know whether the orbit spends more time in certain parts of the space, or whether it spreads out throughout X so as to spend a proportion of time in any set proportional to a natural notion of 'size' of the set. We will return to these questions later, and will make the latter concept–*equidistribution*–precise.

Example 1. Let $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, which we may also identify with $[0, 1)$ with the quotient topology inherited from $\mathbb{R}$. Thus, for example, $1 - \frac{1}{n} \to 0 \in [0, 1)$ as $n \to \infty$. For any $p \in \mathbb{Z}$ we define the map

$$\begin{aligned} T_p : \mathbb{T} &\longrightarrow \mathbb{T} \\ x &\longmapsto px \pmod{\mathbb{Z}}. \end{aligned}$$

The case $p = 10$ is particularly easy to describe. Given a number $x \in [0, 1)$ we may write its decimal expansion as

$$x = 0.a_1a_2 \cdots \in [0, 1),$$

with decimal digits $a_n \in \{0, 1, \dots, 9\}$ for all $n \geqslant 1$. Then we have

$$T_{10}(x) = 0.a_2a_3 \cdots \in [0, 1).$$

Thus if the orbit of $x \in [0, 1)$ under T_{10} is dense, then every digit from 0 to 9 will appear in the decimal expansion of x. In fact every finite block of digits, 2013 for example, has to appear infinitely often in the decimal expansion of such a point.

For this dynamical system it is clear that very different types of orbit are possible. It is easy to construct many decimal expansions that do not contain a given block like 2013. On the other hand it is also easy to find points with dense orbits. There are many ways to see this, here are two.

1. Enumerate all finite blocks of decimal digits in increasing order of length,

   $$0, 1, 2, \dots, 9, 00, 01, 02, \dots, 99, 000, \dots$$

   and then concatenate them to produce a number

   $$x = 0.012 \cdots 9000102 \cdots 99000001002 \cdots .$$

2. Enumerate the natural numbers in their natural order,

   $$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, \dots$$

   and then concatenate them to produce a number

   $$x = 0.1234567891011121314 \cdots .$$

Exercise 2. Show that $x \in [0,1)$ has a finite orbit under T_p for $p \geqslant 2$ if and only if $x \in \mathbb{Q}$. Describe which points are *periodic* (that is, have the property that there is some $n \geqslant 1$ for which $T_p^n(x) = x$).

Example 3. A different, and much better behaved, dynamical system on the circle $\mathbb{T}$ is given by the *circle rotation*

$$\begin{aligned} R_\alpha : \mathbb{T} &\longrightarrow \mathbb{T} \\ x &\longmapsto x + \alpha. \end{aligned}$$

If $\alpha = \frac{p}{q} \in \mathbb{Q}$ (written in lowest terms) then $R_\alpha^q = I$ and all orbits are periodic (with q elements). As this behavior is too simple, we will always assume that α is not rational, in which case R_α is called an *irrational rotation*.

Exercise 4. Show that the orbit of every point under an irrational rotation is dense in the circle.

The sense in which Example 3 is *better behaved* than Example 1 is an instance of the important notion of *measure rigidity*, which will emerge as a central theme of these notes. For the moment, notice that the orbits of the map R_α for a fixed α are all very similar–for rational α they are finite sets of fixed cardinality, and for irrational α they are dense. Moreover, for fixed α any orbit may be obtained from any other orbit by a simple rotation.

Example 5. Let

$$\begin{aligned} M_z : \mathbb{C} &\longrightarrow \mathbb{C} \\ w &\longmapsto zw \end{aligned}$$

be the multiplication by z map on the complex plane. If $|z| < 1$ then every orbit converges to $0 \in \mathbb{C}$. If $|z| > 1$ then the orbit of every point apart from 0 diverges to infinity. If $|z| = 1$ then every orbit stays on a circle. Thus for most choices of the parameter z this map does not have interesting orbits, and for $|z| = 1$ the orbits remain on a circle of fixed radius, and the map M_z restricted to each such circle is isomorphic to a rotation on $\mathbb{T}$.

One way to view the system $(\mathbb{C}, M_z)$ for $|z| = 1$ is as a combination of constituent systems, each of which is a copy of a circle rotation, and it is simpler to study its constituent systems. However, in general we cannot expect such a geometrically straightforward decomposition of a dynamical system into constituent parts which cannot readily be decomposed further.

1.3 What is the probabilistic point of view?

A deeper examination of maps like T_p from Example 1 suggests that the possible behaviors of orbits means that the study of orbits and their closures is in general impossibly difficult. As a result one is led to the question of whether it is more straightforward to describe the orbits of *typical* points. One of the most satisfactory answers to this question is given by the pointwise ergodic theorem of Birkhoff

from 1931. In order to state this, we need the following compatibility notion that picks out those probability measures on the underlying space of a dynamical system that respect the dynamics.

Definition 6. Let (X, T) be a dynamical system, and assume that X is equipped with a σ-algebra $\mathcal{B}$ of measurable subsets[1]. A probability measure μ defined on $\mathcal{B}$ is called *T-invariant* (or T is called *μ-preserving*) if

$$\mu(T^{-1}B) = \mu(B)$$

for all $B \in \mathcal{B}$.

Theorem 7 (Birkhoff, Pointwise Ergodic Theorem)**.** *Let (X, T, μ) be a measure-preserving dynamical system, and let $f \in L^1(X, \mu)$. Then*

$$\frac{1}{N}\sum_{n=0}^{N-1} f({}^n x) \longrightarrow f^*(x)$$

in $L^1(X, \mu)$ and μ-almost everywhere, where $f^* \in L^1(X, \mu)$ is a T-invariant function.

The difficult part of this theorem is the convergence almost everywhere (hence the name pointwise). The convergence in $L^1(X, \mu)$ with respect to the L^1 norm is easier, and can be derived from von Neumann's mean ergodic theorem established in the next exercise. While the behavior of T on individual points is described by the map T itself, it is convenient to study the behavior of T on functions via the associated operator U_T defined by $U_T(f)(x) = f(Tx)$.

Exercise 8. Let (X, T, μ) be a measure-preserving dynamical system, and suppose for simplicity that T is invertible.

(a) Show that $U_T : L^2(X, \mu) \to L^2(X, \mu)$ is *unitary*, meaning that

$$\langle U_T f, U_T g\rangle = \langle f, g\rangle$$

for all $f, g \in L^2(X, \mu)$.

(b) Show that[2]

$$\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n = \frac{1}{N}\sum_{n=0}^{N-1} U_T^n f \longrightarrow P(f)$$

for every $f \in L^2(X, \mu)$, where $P(f)$ denotes the orthogonal projection of f onto the subspace

$$\mathcal{I}_T = \{f \in L^2(X, \mu) \mid U_T f = f\}.$$

[1] In most situations we will encounter, X will be a topological space and $\mathcal{B}$ will be the Borel σ-algebra, which is defined to be the smallest σ-algebra that contains all the open sets.

[2] Consider the cases $f \in \mathcal{I}_T$ and $f \in \mathcal{I}_T^\perp$ separately, and then use the orthogonal decomposition. For the latter subspace consider the functions $f = g - U_T(g)$ and show that these are dense in $\mathcal{I}_T^\perp$.

As in Exercise 8, the function f^* in Theorem 7 should be thought of as the 'orthogonal projection' of $f \in L^1(X,\mu)$ to the space

$$\{g \in L^1(X,\mu) \mid U_T g = g\}.$$

This projection (which *a priori* does not make sense because $L^1(X,\mu)$ does not have the geometry of a Hilbert space) is called the *conditional expectation* and is denoted by $E(f\big|\mathcal{E})$ where

$$\mathcal{E} = \{B \in \mathcal{B} \mid T^{-1}B = B\}$$

is the σ-algebra of T-invariant sets. We refer to [3, Chap. 5] for the details.

Definition 9. Let μ be a T-invariant probability measure on X. We say that μ is *ergodic* if any set $B \in \mathcal{B}$ with $T^{-1}B = B$ has $\mu(B) \in \{0,1\}$. Equivalently, μ is ergodic if $U_T f^* = f^*$ for some $f^* \in L^2(X,\mu)$ implies that $f^*(x) = \int_X f^* \,\mathrm{d}\mu$ for μ-almost every x (that is, f^* is equal to a constant μ-almost everywhere).

The equivalence of these two definitions of ergodicity is relatively easy and may be found in [3, Chap. 2]. Notice that the first formulation of ergodicity expresses the idea that (X,T,μ) cannot be decomposed into invariant subsets that are non-trivial with respect to μ. For this reason ergodicity has also been called *indecomposability*. With this notion of indecomposability we can give a strengthening of the conclusion of Theorem 7 as follows.

Corollary 10. *Let $T : X \to X$ be a continuous transformation of a σ-compact metric space. If μ is a T-invariant and ergodic probability measure on X, then μ-almost every point $x \in X$ is* generic *for T.*

Here $x \in X$ is said to be generic if

$$\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) \longrightarrow \int_X f \,\mathrm{d}\mu$$

for all $f \in C_c(X)$.

We note that Corollary 10 is not a completely trivial consequence of Theorem 7 because there are uncountably many functions in $C_c(X)$, each of which potentially requires a null set of badly-behaved points to be avoided. The single null set arising in Corollary 10 does not depend on the function f. The main idea in the proof is to get around this by taking advantage of the fact that $C_c(X)$ is *separable*, meaning that it has a countable subset that is dense with respect to the supremum norm. This observation together with a relatively easy approximation argument gives the corollary (see [3, Chap. 4] for the details).

Example 11. The Lebesgue measure λ on $[0,1) \cong \mathbb{T}$ is invariant under the map T_p for all $\geqslant 1$. One way to see this is to note that for $0 \leqslant a < b < 1$ we have

$$T_p^{-1}\left([a,b)\right) = [\tfrac{a}{p}, \tfrac{b}{p}) \sqcup [\tfrac{a+1}{p}, \tfrac{b+1}{p}) \sqcup \cdots \sqcup [\tfrac{a+p-1}{p}, \tfrac{b+p-1}{p})$$

is a disjoint union, so

$$\lambda\left(T_p^{-1}([a,b))\right) = \lambda\left([a,b)\right) = b - a.$$

Using this, one can show by an approximation argument that we also have

$$\lambda(T^{-1}B) = \lambda(B)$$

for all Borel sets $B \subset [0,1)$ since the intervals of the form $[a,b)$ generate the Borel σ-algebra.

The measure λ is also ergodic with respect to T_p for $p \geqslant 2$. One way to see this is to use Fourier analysis for functions on the circle as follows. Assume that $f \in L^2(\mathbb{T},\lambda)$ is invariant under T_p, and write

$$f(x) = \sum_{n\in\mathbb{Z}} c_n \mathrm{e}^{2\pi i n x}$$

for the Fourier expansion of f at $x \in \mathbb{T}$. Thus, working in L^2, we have

$$f(x) = \sum_{n\in\mathbb{Z}} c_n \mathrm{e}^{2\pi i n x} = f(T_p x) = \sum_{n\in\mathbb{Z}} c_n \mathrm{e}^{2\pi i n p x}.$$

This equality in L^2 means that the Fourier coefficients of both functions must coincide, so

$$c_n = c_{np} = c_{np^2} = \cdots$$

for all $n \in \mathbb{Z}$. On the other hand

$$\sum_{n\in\mathbb{Z}} |c_n|^2 = \|f\|_2 < \infty$$

by the Plancherel theorem. It follows that $|c_n| \to 0$ as $n \to \infty$ so, in particular, $c_n = 0$ for $n \neq 0$. Thus $f = c_0$ is a constant as required.

Applying Theorem 7 to the map T_{10} we can now deduce that λ-almost every $x \in \mathbb{T}$ has the property that the decimal block of digits 2013 appears in the decimal expansion of x not only infinitely often, but does so with the asymptotic frequency $\frac{1}{10^4}$.

Exercise 12. Show that the Lebesgue measure λ on $[0,1) \cong \mathbb{T}$ is invariant and ergodic for any irrational rotation $R_\alpha : \mathbb{T} \to \mathbb{T}$.

Using the exercise above we prove an even stronger statement about irrational translations.

Corollary 13 (Unique ergodicity for irrational rotations)**.** *For $\alpha \notin \mathbb{Q}$, every point $x \in \mathbb{T}$ is generic for R_α.*

Proof. Fix a function $f \in C(\mathbb{T})$ and some $\varepsilon > 0$. Since f is uniformly continuous, there exists some $\delta > 0$ such that[3]

$$\mathsf{d}(x,y) < \delta \implies |f(x) - f(y)| < \varepsilon$$

[3]We have not formally defined the translation-invariant metric d, but the reader should find it easy to construct this as the metric inherited from the usual one on $\mathbb{R}$.

for all $x, y \in \mathbb{T}$. Given $x \in \mathbb{T}$, we can use Exercise 12 and Corollary 10, almost every $y \in \mathbb{T}$ is generic for R_α. In particular, there exists some generic point y with $\mathsf{d}(x, y) < \delta$. Then

$$\mathsf{d}(R_\alpha^n x, R_\alpha^n y) < \delta$$

for all $n \geqslant 0$, and so

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} f(R_\alpha^n x) - \frac{1}{N} \sum_{n=0}^{N-1} f(R_\alpha^n y) \right| < \varepsilon,$$

which for large enough N gives

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} f(R_\alpha^n x) - \int_0^1 f \, \mathrm{d}\lambda \right| < 2\varepsilon$$

as required. □

Exercise 14. Prove Corollary 13 directly, using the characters $x \mapsto \mathrm{e}^{2\pi \mathrm{i} x n}$ for $x \in \mathbb{T}$ and $n \in \mathbb{Z}$.

Corollary 15. *There exists some $n \geqslant 1$ (in fact, there are infinitely many) such that the decimal expansion of 2^n starts with the digits*[4] 2013*.*

Exercise 16. Prove Corollary 15 using Corollary 13 (use the fact that if the rotation parameter $\alpha = \log_{10} 2$ then $\log_{10} 2^n = n\alpha$ and consider what the latter identity means modulo 1).

Exercise 17. Generalize[5] the discussion above to show that there exists some n for which the decimal expansion of both 2^n and of 3^n start with the digits 2013.

We end this introduction to the basics of ergodic theory with a short proof of the pointwise ergodic theorem, due to Hasselblatt and Katok [5].

Proof of Theorem 7. Let $f^* = E(f \big| \mathcal{E}_T)$, which as mentioned above should be thought of as the projection of f onto the space of T-invariant functions. If μ is ergodic, then $f^* = \int_X f \, \mathrm{d}\mu$.

We fix $\varepsilon > 0$ and define

$$g = f - f^* - \varepsilon,$$

and

$$S_n(g) = \sum_{k=0}^{n-1} g \circ T^k.$$

We will show that

$$\limsup_{n \to \infty} \frac{1}{n} S_n(g) \leqslant 0 \tag{1}$$

[4] As the reader will have noticed, 2013 was the year in which the summer school took place and these notes were written, and is being used simply to represent an arbitrary finite sequence of digits that does not start with 0.

[5] You will need to consider a rotation on $\mathbb{T}^2$.

for μ-almost every $x \in X$. Using the fact that

$$S_n(g) = S_n(f) - nf^* - n\varepsilon$$

and the same statement for $-f$ then gives the theorem. In turn (1) will follow from the following lemma.

Lemma 18. *Let*

$$A = \{x \in X \mid \sup_{k \geqslant 1} S_k(g) = \infty\}.$$

Then $\mu(A) = 0$.

Proof. We need to assemble some facts about S_k, A, and the function

$$M_n = \max\{S_k g \mid k = 0, \ldots, n\}.$$

- First note that $S_k(g)(Tx) = S_{k+1}(g)(x) - g(x)$ so that $T^{-1}A = A$, i.e. A is a T-invariant set.
- As is customary, we set $S_0(g) = 0$ so that $M_n \geqslant 0$ for all $n \geqslant 0$.
- Using the definition of M_n we see that

$$M_{n+1}(x) = \max\{0, g(x) + M_n(Tx)\}.$$

- It follows that

$$M_{n+1}(x) - M_n(Tx) = \max\{-M_n(Tx), g(x)\}. \tag{2}$$

- Using the definition of A and of M_n, this now shows that

$$\lim_{n\to\infty} (M_{n+1}(x) - M_n(Tx)) = g(x)$$

  for all $x \in A$.
- We also have

$$g(x) \leqslant M_{n+1}(x) - M_n(Tx) \leqslant \max\{0, g(x)\}$$

  by (2) and the inequality $M_n \geqslant 0$.
- The last fact shows that

$$M_{n+1}(x) - M_n(Tx)$$

  is dominated by an $L^1(X, \mu)$ function independently of n. Hence we may apply the dominated convergence theorem to obtain

$$\lim_{n\to\infty} \int_A (M_{n+1} - M_n \circ T)\, \mathrm{d}\mu = \int_A g\, \mathrm{d}\mu.$$

- We calculate

$$\int_A g\, \mathrm{d}\mu = \int_A f\, \mathrm{d}\mu - \int_A f^*\, \mathrm{d}\mu - \varepsilon\mu(A) = -\varepsilon\mu(A) \tag{3}$$

  by definition of f^*.

- Moreover,

$$\begin{aligned}\int_A (M_{n+1} - M_n \circ T)\,\mathrm{d}\mu &= \int_A M_{n+1}\,\mathrm{d}\mu - \int_A M_n \circ T\,\mathrm{d}\mu \\ &= \int_A M_{n+1}\,\mathrm{d}\mu - \int (\mathbb{1}_A M_n) \circ T\,\mathrm{d}\mu \quad (\text{since } T^{-1}A = A) \\ &= \int_A M_{n+1}\,\mathrm{d}\mu - \int_A M_n\,\mathrm{d}\mu \quad (\text{since } \mu \text{ is } T-\text{invariant}) \\ &= \int_A \underbrace{(M_{n+1} - M_n}_{\geqslant 0}\,\mathrm{d}\mu. \quad \text{by definition of } M_n\end{aligned}$$

Combining the last three facts we obtain

$$-\varepsilon\mu(A) \geqslant 0,$$

and hence $\mu(A) = 0$ as claimed. □

This completes the proof of Theorem 7. □

Example 19. We now present another measure-preserving transformation on (almost all of) $[0,1]$, which will turn out to be ergodic. At first sight this example may appear less natural than the multiplication by p map and the rotation by α map on the circle above, but it also is of algebraic origin–though the algebra in question is rather hidden (see [3, Ch. 9]). The *Gauss map* is defined by

$$\begin{aligned} T : (0,1) &\longrightarrow (0,1) \\ x &\longmapsto \left\{\frac{1}{x}\right\}, \end{aligned}$$

where $\{\cdot\}$ denotes the fractional part of a real number. Notice that strictly speaking this map does not define a dynamical system, as the orbit of any rational point eventually reaches 0, where the map is not defined. Restricting T to the set $[0,1]\smallsetminus\mathbb{Q}$ does however give a dynamical system, and this is the map we will think about. The map T preserves the *Gauss measure* μ defined by

$$\mu([a,b]) = \int_a^b \frac{1}{\log 2(1+x)}\,\mathrm{d}x,$$

and therefore defines a measure-preserving dynamical system.

We cannot resist mentioning one striking consequence of the study of the Gauss map, and in particular of the ergodicity of the Gauss measure with respect to T. For every irrational $x \in (0,1)$ there exists a sequence of best rational approximations $(\frac{p_n(x)}{q_n(x)})$ (defined using the continued fraction expansion of x), and μ-almost every $x \in (0,1)$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \left| x - \frac{p_n(x)}{q_n(x)} \right| = -\frac{\pi^2}{6\log 2}.$$

We refer to [3, Chap. 3] for the details and the history of this kind of result.

2 What is a nilflow?

Example 3 may be described in the following way. We start with a locally compact group $\mathbb{R}$, with a discrete subgroup $\mathbb{Z}$ of finite covolume (this means that the Lebesgue measure on $\mathbb{R}$ gives finite measure to any measurable fundamental domain for $\mathbb{Z} < \mathbb{R}$), and an element $\alpha \in \mathbb{R}$. This defines a measure-preserving map on the quotient space $\mathbb{R}/\mathbb{Z} = \mathbb{T}$ by rotation under α, and this is the map R_α. Much of our purpose here is concerned with understanding what phenomena arise if $\mathbb{R}$ is replaced by a non-commutative locally compact group, $\mathbb{Z}$ by a lattice in it, and Lebesgue measure by the Haar measure.

As a first step towards the study of algebraic actions on homogeneous spaces defined by non-commutative groups, we now discuss briefly a special case of *flows on nilmanifolds*.

Let[6]

$$H = \left\{ \begin{pmatrix} 1 & a & c \\ & 1 & b \\ & & 1 \end{pmatrix} \mid a, b, c \in \mathbb{R} \right\}$$

be the Heisenberg group (which is an example of a *two-step nilpotent Lie group* or a *unipotent algebraic group* depending on the language one is used to speaking). Similarly, let

$$\Gamma = \left\{ \begin{pmatrix} 1 & \ell & n \\ & 1 & m \\ & & 1 \end{pmatrix} \mid \ell, m, n \in \mathbb{Z} \right\}$$

be the group of integer points in H. Then the space

$$X = \Gamma\backslash H = \{\Gamma h \mid h \in H\}$$

is (by definition) a *nilmanifold*. This space is a compact manifold which can, for example, be obtained by gluing opposite faces of the cube $[0,1)^3$ together in the appropriate way. However, in some of the gluing of faces a twist is applied (without the twist, gluing the faces of the cube together would produce the 3-torus $\mathbb{T}^3$). In fact we use elements $\gamma \in \Gamma$ to identify elements $g \in H$ using the equivalence relation

$$\gamma g = \begin{pmatrix} 1 & \ell & n \\ & 1 & m \\ & & 1 \end{pmatrix} \begin{pmatrix} 1 & a & c \\ & 1 & b \\ & & 1 \end{pmatrix} = \begin{pmatrix} 1 & a+\ell & c+n+\ell b \\ & 1 & b+m \\ & & 1 \end{pmatrix} \sim g. \tag{4}$$

Given

$$\alpha = \begin{pmatrix} 0 & \alpha_1 & \beta \\ & 0 & \alpha_2 \\ & & 0 \end{pmatrix}$$

in the Lie algebra of H, we define a flow

$$R_t : X \longrightarrow X$$

$$x = \Gamma g \longmapsto x \exp(t\alpha) = \Gamma g \left(I + t + \frac{1}{2} t^2 \alpha^2 + \cdots \right)$$

[6]We will sometimes forget to write entries in a matrix if these are zero.

for all $t \in \mathbb{R}$. (As H is two-step nilpotent we could actually simplify the expression for the exponential map and simply write $\exp(t\alpha) = I + t + \frac{1}{2}t^2$.) Here we call the dynamical system a *flow* because we have an action of $\mathbb{R}$ instead of the $\mathbb{Z}$-action induced by a single transformation.

Theorem 20 (A uniquely ergodic nilflow)**.** *If* $\alpha_2 \neq 0$ *and* $\frac{\alpha_1}{\alpha_2} \notin \mathbb{Q}$ *then*

$$\{R_t \mid t \in \mathbb{R}\}$$

is a uniquely ergodic flow on X*. This means that*

$$\frac{1}{T}\int_0^T f\left(R_t(x)\right)\,\mathrm{d}t \longrightarrow \int_X f\,\mathrm{d}\lambda_X$$

as $T \to \infty$ *for any* $f \in C_c(X)$*, where* λ_X *is the image of the Lebesgue measure*[7] *on* $[0,1)^3$ *in* X*.*

Proof. Define a factor map

$$\pi : X \longrightarrow \mathbb{T}^2$$

by

$$\pi : \Gamma \begin{pmatrix} 1 & a & c \\ & 1 & b \\ & & 1 \end{pmatrix} \longmapsto \begin{pmatrix} a \\ b \end{pmatrix} + \mathbb{Z}^2 \in \mathbb{T}^2.$$

The calculation in (4) shows that π is well-defined. This map is often called the *maximal abelian factor*. By essentially the same argument as that used in Corollary 13, the maximal abelian factor is ergodic with respect to the induced flow

$$\widehat{R_t} : \mathbb{T}^2 \ni \begin{pmatrix} a \\ b \end{pmatrix} \longmapsto \begin{pmatrix} a + \alpha_1 t \\ b + \alpha_2 \end{pmatrix} \in \mathbb{T}^2,$$

and in fact is uniquely ergodic (that is, there is no other invariant measure for the flow). Our main task therefore is to extend this unique ergodicity result from $\mathbb{T}^2$ to X. This cannot be done using the standard theory of Fourier series as harmonic analysis works quite differently on the space X.

As a first step, we show that the flow R_t is ergodic. Suppose therefore that $f \in L^2(X, \lambda_X)$ is an invariant function for R_t. Let

$$g_s = \begin{pmatrix} 1 & s & \\ & 1 & \\ & & 1 \end{pmatrix},$$

and notice that[8]

$$\|f(x) - f(xg_s)\|_2 \longrightarrow 0$$

[7] The reader should check that λ_X is preserved by R_t for all $t \in \mathbb{R}$.

[8] In these expressions we are (inaccurately but efficiently) using $f(x)$ as shorthand for the function $x \mapsto f(x)$ rather than the value of the function f at x.

as $s \to 0$. This requires a proof, but we will leave this step to the reader and only note that elements of $L^2(X, \lambda_X)$ can be approximated by continuous functions (see also [3, Sect. 11.3.2]). Using the invariance property

$$f(x) = f(x \exp(t\alpha))$$

of f gives

$$\begin{aligned} \|f(x) - f(xg_s)\|_2 &= \|f(x \exp(t\alpha)) - f(xg_s \exp(t\alpha))\|_2 \\ &= \|f(y) - f(y \exp(-t\alpha) g_s \exp(t\alpha))\|_2 \end{aligned}$$

where we first used the invariance and then the substitution

$$y = x \exp(t\alpha),$$

which amounts to a unitary transformation on $L^2(X, \lambda_X)$. We now calculate

$$\exp(-t\alpha) g_s \exp(t\alpha) = \begin{pmatrix} 1 & s & \alpha_2 ts \\ & 1 & 0 \\ & & 1 \end{pmatrix}. \tag{5}$$

We set $s = \frac{1}{n}$ and notice that $t = t_n$ can be chosen arbitrarily. We choose t_n so that

$$\frac{\alpha_2 t_n}{n} = r$$

for some arbitrary $r \in \mathbb{R}$. Therefore the matrix in (5) approaches

$$\begin{pmatrix} 1 & 0 & r \\ & 1 & 0 \\ & & 1 \end{pmatrix}$$

as $n \to \infty$, and we get

$$\left\| f(x) - f\left(x \begin{pmatrix} 1 & 0 & r \\ & 1 & 0 \\ & & 1 \end{pmatrix} \right) \right\|_2 = 0,$$

or equivalently that f is invariant under

$$\begin{pmatrix} 1 & 0 & r \\ & 1 & 0 \\ & & 1 \end{pmatrix}$$

for all $r \in \mathbb{R}$. This step is an instance of the *Mautner phenomenon* (which we will discuss again in Section 4).

This shows that

$$f(x) = f\left(\Gamma \begin{pmatrix} 1 & a & c \\ & 1 & b \\ & & 1 \end{pmatrix} \right)$$

almost surely does not depend on c, so f is a function on the maximal abelian factor $\mathbb{T}^2$. However, as we already know that the induced flow is ergodic on $\mathbb{T}^2$ this shows that f must be constant almost everywhere, and the flow R_t is thus seen to be ergodic.

Once more we know ergodicity, and so almost every point is generic. We wish to upgrade this ergodicity to unique ergodicity. This is not always possible (see, for example, the map from Exercise 2 which clearly has many different invariant measures supported on periodic orbits). Moreover, the simple commutative or isometric argument from Corollary 13 also does not work in our current situation. Despite this, it is possible in this situation to upgrade the argument to show unique ergodicity using an observation of Furstenberg. We give a slightly modified version of this argument.

Notice first that for each $y \in \mathbb{T}^2$ the fibre $\pi^{-1}y$ is isomorphic to $\mathbb{T}$, and write $\lambda_{\pi^{-1}y}$ to denote the Lebesgue measure on $\pi^{-1}y$. We note that

$$\frac{1}{T}\int_0^T (R_t)_*\,\lambda_{\pi^{-1}y}\,\mathrm{d}t \longrightarrow \lambda_X \tag{6}$$

in the weak*-topology on the space of probability measures $\mathcal{M}(X)$ on X, which via Riesz representation is identified with a subset of the dual of the Banach space $C(X)$. In fact we already know (6), but not quite in the language used above. The statement can be translated as follows. Given a continuous function $f \in C(X)$, the measure on the left-hand side of (6) is defined to be the measure that integrates f to

$$\frac{1}{T}\int_0^T \int_{\pi^{-1}y} f\,\mathrm{d}\lambda_{\pi^{-1}(\widehat{R}_t y)}\,\mathrm{d}t.$$

Here we can identify the inner integral with the function F defined by

$$F(y) = \int_{\pi^{-1}y} f\,\mathrm{d}\lambda_{\pi^{-1}y} \in C(\mathbb{T}^2)$$

evaluated at $\widehat{R}_t(y)$. However, on $\mathbb{T}^2$ we already have unique ergodicity and so

$$\frac{1}{T}\int_0^T F\left(\widehat{R}_t y\right)\,\mathrm{d}t \longrightarrow \int_{\mathbb{T}^2} F = \int_X f\,\mathrm{d}\lambda_X$$

by Fubini's theorem. Combining these we see the weak*-convergence in (6).

Let $x \in X$ be arbitrary, let $y = \pi(x)$ and fix $\delta > 0$. Then we can split $\lambda_{\pi^{-1}y}$ into a convex combination

$$\lambda_{\pi^{-1}y} = (2\delta)\lambda_{I_\delta(x)} + (1-2\delta)\lambda_{\pi^{-1}y \smallsetminus I_\delta(x)}, \tag{7}$$

where $I_\delta(x)$ denotes the δ-neighbourhood of x in $\pi^{-1}y$, $\lambda_{I_\delta(x)}$ is the Lebesgue measure on $I_\delta(x)$ normalized to be a probability measure, and $\lambda_{\pi^{-1}y \smallsetminus I_\delta(x)}$ is defined similarly.

Now consider the analogue of the left-hand side of (6) for $\lambda_{I_\delta(x)}$, giving

$$\frac{1}{T}\int_0^T (R_t)_*\,\lambda_{I_\delta(x)}\,\mathrm{d}t \in \mathcal{M}(X). \tag{8}$$

By the Tychonoff–Alaoglu theorem on $C(X)^*$ we can choose a sequence (T_n) with $T_n \to \infty$ as $n \to \infty$ such that (8) converges to come probability measure $\nu_{x,\delta}$. From the construction it is easy to see that $\nu_{x,\delta}$ is R_t-invariant (because the interval $[0,T]$ is almost translation-invariant). Furthermore, we can choose another subsequence so that the same expression for $\nu_{\pi^{-1}y \smallsetminus I_\delta}$ also converges to an invariant probability measure $\widetilde{\nu_{x,\delta}}$. Comparing this with (6) and (7) we get

$$\lambda_X = 2\delta\nu_{x,\delta} + (1-2\delta)\widetilde{\nu_{x,\delta}}.$$

We did not mention this before, but an ergodic measure is an extreme point of the convex set of invariant probability measures (see [3, Th. 4.4] for the details). It follows that

$$\lambda_X = \nu_{x,\delta},$$

and what is also curious is that the limit measure obtained is independent of the chosen subsequence (T_n). Now if a sequence in a compact topological space has the property that the limit of convergent subsequence is independent of the subsequence, then the original sequence (or, in our case, map from $(0,\infty)$ to $\mathcal{M}(X)$) must itself converge.

Finally, fix some $f \in C(X)$ and $\varepsilon > 0$. Choose $\delta > 0$ as in the condition for uniform continuity of f, and notice that $z \in I_\delta$ has the property that

$$\mathsf{d}\left(R_t(x), R_t(z)\right) < \delta$$

for all $t \in \mathbb{R}$. Now we can conclude the argument as in the proof of Corollary 13. □

We note that nilflows (or rather their discrete analogues, nilrotations) have recently become extremely important in the area of interaction between ergodic theory, Ramsey theory, and number theory (see the surveys of Bergelson [1], [2] for an overview of Ramsey theory in this context).

Exercise 21. Let

$$N = \left\{ \begin{pmatrix} 1 & a_{12} & a_{13} & a_{14} \\ & 1 & a_{23} & a_{24} \\ & & 1 & a_{34} \\ & & & 1 \end{pmatrix} \mid a_{ij} \in \mathbb{R} \right\},$$

$$\Gamma = \left\{ \begin{pmatrix} 1 & a_{12} & a_{13} & a_{14} \\ & 1 & a_{23} & a_{24} \\ & & 1 & a_{34} \\ & & & 1 \end{pmatrix} \mid a_{ij} \in \mathbb{Z} \right\},$$

and define $X = \Gamma\backslash N$. Characterize those flows $R_t(x) = x\exp(t\alpha)$ for $t \in \mathbb{R}$ which are ergodic by studying the maximal abelian quotient (and the appropriate Mautner phenomenon). Can you generalize Theorem 20 to these nilflows?

3 A brief review of dynamics on the modular surface

Starting in this section we will consider more general group actions. Let G be a group and let X be a set (called the space). Then a *G-action* is simply a homomorphism Φ from G to the group of bijections of X. We will also write $g{\cdot}x = \Phi(g)(x)$ for the action of $g \in G$ on $x \in X$. Usually the space has some structure and we require the action to preserve this structure. If G is a *topological group* (where the group operations are continuous) and X is a topological space, then we call the G-action *continuous* if

$$(g, x) \in G \times X \to g{\cdot}x \in X$$

is continuous. If X has a measure, then we call the G-action *measure-preserving* if $\Phi(g)$ is measure-preserving for all $g \in G$.

3.1 The space

We recall (see, for example, [3, Ch. 9]) that the upper half-plane

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

equipped with the Riemannian metric

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

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

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \longmapsto g{\cdot}z = \frac{az + b}{cz + d}. \tag{9}$$

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

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

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

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

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

where

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

This action gives rise to the simply transitive action of

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

on the unit tangent bundle

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

so that

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

by sending g to $\mathrm{D}\, g(\mathrm{i}, \mathrm{i})$.

The region E illustrated by shading in Figure 1. is a fundamental region for the action of the discrete subgroup $\mathrm{PSL}_2(\mathbb{Z})$ on $\mathbb{H}$ (strictly speaking we should describe carefully which parts of the boundary of the hyperbolic triangle shaded belong to the domain but as the boundary is a nullset one usually ignores that issue — we will comply with this tradition), see Exercise 22.

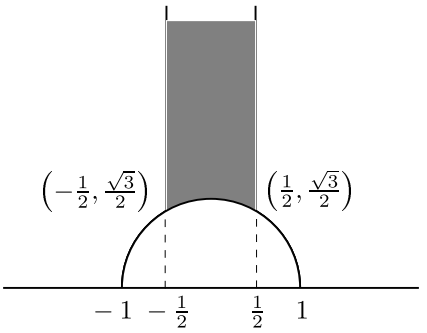

Figure 1. A fundamental domain $E \subset \mathbb{H}$ for the action of $\mathrm{SL}_2(\mathbb{Z})$.

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

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

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

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

(Strictly speaking we should describe more carefully which vectors attached to points $z \in \partial E$ are allowed in F.) Furthermore, we can lift the set $F \subset \mathrm{PSL}_2(\mathbb{R})$ to a surjective[9] set $F \subset \mathrm{SL}_2(\mathbb{R})$ for $\mathrm{SL}_2(\mathbb{Z})$. We claim that this argument shows that

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

has finite volume. In order to see this, we recall some basic facts from [3, Ch. 9]:

[9]That is, F has the property that the natural quotient map from $\mathrm{SL}_2(\mathbb{R})$ to $\mathrm{SL}_2(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{R})$ is surjective when restricted to F.

- $\mathrm{SL}_2(\mathbb{R})$ is unimodular (see Exercise 23).
- $\mathrm{SL}_2(\mathbb{R}) = NAK$ with[10]

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

  and $K = \mathrm{SO}(2)$, in the sense that every $g \in \mathrm{SL}_2(\mathbb{R})$ can be written uniquely[11] as a product $g = nak$ with $n \in N$, $a \in A$ and $k \in K$.
- Let $B = NA = AN$ be the subgroup $B = \left\{ \begin{pmatrix} a & t \\ & a^{-1} \end{pmatrix} \mid a > 0, t \in \mathbb{R} \right\}$. The Haar measure $m_{\mathrm{SL}_2(\mathbb{R})}$ decomposes in the coordinates $g = bk$, meaning that

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

  where $\propto$ denotes proportionality (with the constant of proportionality dependent only on the choices of Haar measures). Moreover, the left Haar measure m_B decomposes in the coordinate system

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

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

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

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

Combining these facts we get

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

The argument above also helps us to understand the space

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

globally: it is, apart from some difficulties arising from the distinguished points

$$\mathrm{i}, \tfrac{1}{2} + \tfrac{\sqrt{3}}{2}\mathrm{i} \in E,$$

the unit tangent bundle of the surface[12] $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. This surface may be thought of as being obtained by gluing the two vertical sides in Figure 1. together using the action of

$$\begin{pmatrix} 1 & \pm 1 \\ & 1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

[10] We sometimes indicate by $*$ any entry of a matrix which is only restricted to be a real number, and omit entries that are zero.

[11] This is a simple instance of the more general Iwasawa decomposition of a connected real semi-simple Lie group.

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

and the third side to itself using the action of

$$\begin{pmatrix} & -1 \\ 1 & \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

In particular, X_2 is non-compact.

3.2 The geodesic flow—the subgroup A

As we have seen, the unit tangent bundle of a 'hyperbolic surface' $\Gamma\backslash\mathbb{H}$ can be identified with a quotient of $\mathrm{SL}_2(\mathbb{R})$ by the discrete subgroup $\Gamma < \mathrm{SL}_2(\mathbb{R})$. More generally if G is a linear Lie group and $\Gamma < G$ is a discrete subgroup then one can define the quotient space $X = \Gamma\backslash G$. On such a space there is still a natural G-action defined by $g\cdot x = xg^{-1}$ for $x \in X$ and $g \in G$.

We recall that

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

defines the geodesic flow on X_2, whose orbits may also be described in the fundamental region as in Figure 2..

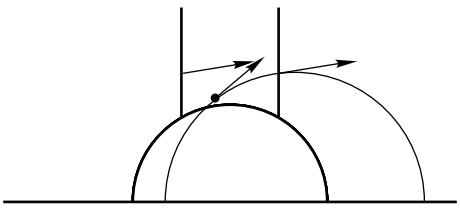

Figure 2. The geodesic flow follows the circle determined by the arrow which intersects $\mathbb{R} \cup \{\infty\} = \partial\mathbb{H}$ normally, and is moved back to F via a Möbius transformation in $\mathrm{SL}_2(\mathbb{Z})$ once the orbit leaves F.

The diagonal subgroup

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

is also called the *torus* or *Cartan subgroup*. We claim that A acts ergodically on X_2 with respect to the Haar measure m_{X_2} (see Section 4 and [3, Sec. 9.5]). Even so, there are many different types of A-orbits, which include the following:

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

matrix $g_{\text{golden}} \in K$ has the property[13] that

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

Now notice that

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

This identity shows that the orbit $\mathrm{SL}_2(\mathbb{Z}) g_{\text{golden}} A$ is compact (see also Figure 3. in which $\lambda = \frac{1+\sqrt{5}}{2}$).

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

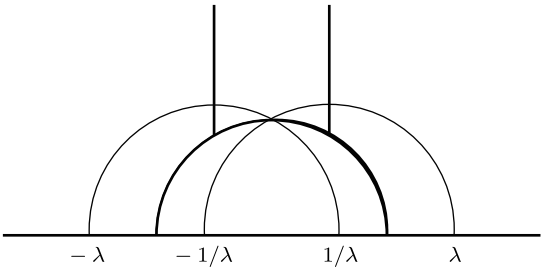

Figure 3. The union of the two geodesics considered in X_2 with both directions allowed is a periodic A-orbit, and comprises the orbit $\mathrm{SL}_2(\mathbb{Z}) g_{\text{golden}} A$.

Finally we would like to point out that there is a correspondence between rational (or arithmetic) objects and closed A-orbits as in the first two types of A-orbit considered above (see Exercise 25 and 26).

3.3 The horocycle flow–the subgroup $U^- = N$

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

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

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

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

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

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

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

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

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

This implies that

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

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

Geometrically, the horocycle orbits $U^-\cdot x = xU^-$ can be described as circles touching the real axis with the arrows (that is, the tangent space component) normal to the circle pointing inwards or as horizontal lines with the arrows pointing upwards, as in Figure 4..

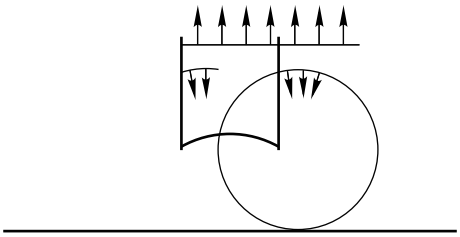

Figure 4. The picture shows the two types of horocycle orbits; the orbits in X_2 can again be understood by using the appropriate Möbius transformation whenever the orbit leaves the fundamental domain.

We recall that U^- also acts ergodically on X_2 with respect to the Haar measure m_{X_2} (see Section 4 and [3, Sec. 11.3]). However, unlike the case of A-orbits, the classification of U^--orbits on X_2 is shorter. The possibilities are as follows:

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

This gives the complete list of types of U^--orbits, and once more gives substance to the claim that there is a correspondence between rational objects and closed orbits (see Exercise 27).

3.4 Exercises for Section 3

Exercise 22. Let E be as in Figure 1.

1. Use $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ to show that $\mathrm{SL}_2(\mathbb{Z})\cdot E$ is 'uniformly open', meaning that there exists some $\delta > 0$ such that $z \in \mathrm{SL}_2(\mathbb{Z})\cdot E$ implies that
$$B_\delta(z) \subset \mathrm{SL}_2(\mathbb{Z})\cdot E.$$
Conclude that $\mathrm{SL}_2(\mathbb{Z})\cdot E = \mathbb{H}$.
2. Suppose that both z and $\gamma\cdot z$ lie in E for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. Show that either $\gamma = \pm I$ or $z \in \partial E$.
3. Conclude that E can be modified (by defining which parts of the boundary of E should be included) to become a fundamental domain.

Exercise 23. Let $d \geqslant 2$. Show that
$$m_{\mathrm{SL}_d(\mathbb{R})}(B) = m_{\mathbb{R}^{d^2}}(\{tb : t \in [0,1], b \in B\})$$
for any measurable $B \subset \mathrm{SL}_d(\mathbb{R})$ defines a (bi-invariant) Haar measure on $\mathrm{SL}_d(\mathbb{R})$, where $m_{\mathbb{R}^{d^2}}$ is the Lebesgue measure on the matrix algebra $\mathrm{Mat}_{dd}(\mathbb{R})$ viewed as the vector space $\mathbb{R}^{d^2}$.

Exercise 24. Show that $\mathrm{SL}_2(\mathbb{R})$ is generated by the unipotent subgroups
$$\begin{pmatrix} 1 & * \\ & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & \\ * & 1 \end{pmatrix}.$$

Exercise 25. Show that $\mathrm{SL}_2(\mathbb{Z})gA$ is a divergent trajectory (that is, the map sending $a \in A$ to $\mathrm{SL}_2(\mathbb{Z})ga$ is a proper map) if and only if $ga \in \mathrm{SL}_2(\mathbb{Q})$ for some $a \in A$.

Exercise 26. Show that to any compact A-orbit in $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ one can attach a real quadratic number field K such that the length of the orbit is $\log|\xi|$, where ξ in $\mathcal{O}_K^*$ is a unit in the order $\mathcal{O}_K$ of K. Prove that there are only countably many such orbits.

Exercise 27. Show that $\mathrm{SL}_2(\mathbb{Z})gU^-$ is compact if and only if
$$g(\infty) \in \mathbb{Q} \cup \{\infty\}.$$
Show that if $\mathrm{SL}_2(\mathbb{Z})gU^-$ is compact, then any other compact orbit is of the form $\mathrm{SL}_2(\mathbb{Z})gaU^-$ for some $a \in A$.

Exercise 28. Show that $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R}) \cong \{\mathbb{Z}^2 g \mid g \in \mathrm{SL}_2(\mathbb{R})\}$ can be identified with lattices $\mathbb{Z}^2 g \subset \mathbb{R}^2$ of co-volume $\det g = 1$. Use the isomorphism with $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{T}^1\mathbb{H}$ discussed in this section to characterize compact subsets K of $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ in terms of elements of the lattices $\mathbb{Z}^2 g$ for $\mathrm{SL}_2(\mathbb{Z})g \in K$. More precisely, calculate the relationship between the shortest vector $ng \in \mathbb{Z}^2 g$ and the imaginary part of $g\mathrm{i} \in \mathbb{H}$ under the assumption that the representative $g \in \mathrm{SL}_2(\mathbb{R})$ has been chosen with $g\mathrm{i} \in E$.

4 Mautner phenomenon

We continue our study of the special (but important) group $G = \mathrm{SL}_2(\mathbb{R})$. Any element $g \in \mathrm{SL}_2(\mathbb{R})$ is conjugate to one of the following three type of elements:

- an $\mathbb{R}$-diagonal matrix, that is one of the form $a = \begin{pmatrix} \lambda & \\ & \lambda^{-1} \end{pmatrix}$ with $\lambda \in \mathbb{R}$;
- a unipotent matrix $u = \begin{pmatrix} 1 & \pm 1 \\ & 1 \end{pmatrix}$; or
- a matrix in the compact subgroup $\mathrm{SO}(2)$, that is one of the form

$$k = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

for some $\phi \in \mathbb{R}$.

For the last case we can make no claim concerning ergodicity of the action of g (because a compact group only acts ergodically when it acts transitively). However, for the first two types we will prove the Mautner phenomenon (Proposition 29), which gives ergodicity of these elements in many cases.

Proposition 29 (Mautner for $\mathrm{SL}_2(\mathbb{R})$)**.** *Let $G = \mathrm{SL}_2(\mathbb{R})$ act unitarily on a Hilbert space $\mathcal{H}$, and suppose that $g \neq \pm I$ is unipotent or $\mathbb{R}$-diagonalizable and fixes a vector $v_0 \in \mathcal{H}$. Then all of G fixes v_0 also.*

Suppose $g \in G$ satisfies the hypotheses of Proposition 29, and $h \in G$ has the property that hgh^{-1} fixes $v_0 \in \mathcal{H}$. Then g fixes $\pi^{-1}(h)v_0$ and so $v_0 = \pi^{-1}(h)v_0$ is fixed by G as needed. Thus it is sufficient to consider one representative of each conjugacy class for the proof of Proposition 29.

The proof relies on the following key lemma.

Lemma 30 (The key lemma)**.** *Let $\mathcal{H}$ be a Hilbert space carrying a unitary representation of a topological group G. Suppose that $v_0 \in \mathcal{H}$ is fixed by some subgroup $L \leqslant G$. Then v_0 is also fixed under every other element $h \in G$ with the property that*

$$B_\delta^G(h) \cap LB_\delta^G(I)L \neq \emptyset \tag{10}$$

for every $\delta > 0$.

Proof. By assumption, there exist three sequences (g_n) in G, (ℓ_n) in L, and (ℓ'_n) in L with $g_n \to e$ and $\ell_n g_n \ell'_n \to h$ as $n \to \infty$. This implies that

$$\|\pi(\ell_n g_n \ell'_n)v_0 - v_0\| = \|\pi(\ell_n)(\pi(g_n\ell'_n)v_0 - \pi(\ell_n^{-1})v_0)\| = \|\pi(g_n)v_0 - v_0\|$$

by invariance of v_0 under all elements of L and unitarity of $\pi(\ell_n)$. However, the left hand side converges to $\|\pi(h)v_0 - v_0\|$ by continuity[14] of the representation and the right hand side converges to 0. □

[14]This continuity is a very general property. For example, it always holds for the unitary representation derived from a continuous measure-preserving actions of G, see [3, Sect. 11.3.2].

Proof of Proposition 29. For $a = \begin{pmatrix} \lambda & \\ & \lambda^{-1} \end{pmatrix}$ with $\lambda \neq \pm 1$ a direct calculation shows that we can apply Lemma 30 with $L = a^{\mathbb{Z}}$ and any element of the unipotent subgroups $\begin{pmatrix} 1 & * \\ & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & \\ * & 1 \end{pmatrix}$ in $\mathrm{SL}_2(\mathbb{R})$. For example,

$$a^n \begin{pmatrix} 1 & s \\ & 1 \end{pmatrix} a^{-n} = \begin{pmatrix} 1 & \lambda^{2n} s \\ & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}$$

if $\lambda^{2n} \to 0$ as $n \to \infty$. It follows that if a fixes some $v_0 \in \mathcal{H}$, then so do these two unipotent subgroups, and as they together generate $\mathrm{SL}_2(\mathbb{R})$ (see Exercise 24), we obtain Proposition 29 for this case.

If $u = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$ then

$$\begin{aligned} u^n \begin{pmatrix} 1+\delta & \\ & 1/(1+\delta) \end{pmatrix} u^{-n} &= \begin{pmatrix} 1 & n \\ & 1 \end{pmatrix} \begin{pmatrix} 1+\delta & \\ & 1/(1+\delta) \end{pmatrix} \begin{pmatrix} 1 & -n \\ & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1+\delta & -2\delta n \\ & 1/(1+\delta) \end{pmatrix} \end{aligned}$$

can be made (since n can be chosen arbitrary) to converge to $\begin{pmatrix} 1 & s \\ & 1 \end{pmatrix}$ for $\delta \to 0$. It follows that if v_0 is fixed by $\begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$ then it is also fixed by $\begin{pmatrix} 1 & s \\ & 1 \end{pmatrix}$ for any $s \in \mathbb{R}$ by Lemma 30 applied with

$$L = \left\{ \begin{pmatrix} 1 & n \\ & 1 \end{pmatrix} \mid n \in \mathbb{Z} \right\}.$$

Applying Lemma 30 once more with

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

to the matrix

$$\begin{pmatrix} 1 & s_1 \\ & 1 \end{pmatrix} \begin{pmatrix} 1 & \\ \delta & 1 \end{pmatrix} \begin{pmatrix} 1 & s_2 \\ & 1 \end{pmatrix} = \begin{pmatrix} 1+\delta s_1 & s_2(1+\delta s_1) + s_1 \\ \delta & 1+\delta s_2 \end{pmatrix} = g_\delta \tag{11}$$

with s_1 chosen to have

$$1 + \delta s_1 = \mathrm{e}^{\alpha},$$

and with s_2 chosen to have

$$s_2(1+\delta s_1) + s_1 = 0$$

shows that v_0 is also fixed by

$$\begin{pmatrix} \mathrm{e}^{\alpha} & \\ & \mathrm{e}^{-\alpha} \end{pmatrix} = \lim_{\delta \to 0} g_\delta.$$

Applying the previous (diagonal) case, we see once again that v_0 is fixed by all of $\mathrm{SL}_2(\mathbb{R})$. □

Suppose now $X = \Gamma\backslash \mathrm{SL}_2(\mathbb{R})$ is a compact (or, more generally, a finite volume) quotient of $\mathrm{SL}_2(\mathbb{R})$ by a discrete subgroup. Then clearly every G-invariant set on X is either empty or everything due to transitivity, and so G acts ergodically on X with respect to the Haar measure m_X (which is induced by the Haar measure on $\mathrm{SL}_2(\mathbb{R})$ normalized to be a probability measure). Applying Theorem 29 it follows[15] that any nontrivial diagonal matrix (that is, any nontrivial element of the geodesic flow) and every nontrivial unipotent matrix (that is, any nontrivial element of the horocycle flow) also acts ergodically on X with respect to the Haar measure.

5 Mixing of the geodesic flow

The following theorem gives a significant strengthening of ergodicity.

Theorem 31 (Howe–Moore)**.** *Suppose that $G = \mathrm{SL}_2(\mathbb{R})$ acts on a probability space (X, μ), and that the action is measure-preserving and ergodic with respect to μ. Then the action of G is also* mixing*, meaning that*

$$\langle f_1 \circ g, f_2\rangle \longrightarrow \int f_1 \,\mathrm{d}\mu \int \overline{f_2}\,\mathrm{d}\mu$$

whenever $g \to \infty$ in G, for any $f_1, f_2 \in L^2(X, \mu)$.

Proof for the geodesic flow. Let $f_1 \in L^2(X, \mu)$ and let

$$a_t = \begin{pmatrix} \mathrm{e}^{-t/2} & \\ & \mathrm{e}^{t/2} \end{pmatrix}$$

be an element of the diagonal subgroup. We will consider the case $t \to \infty$; the case $t \to -\infty$ is similar. Then

$$\|f_1(xa_t)\|_2 = \|f\|_2,$$

and so we may choose (by the Tychonoff–Alaoglu theorem) a sequence (t_n) with

$$\lim_{n\to\infty} f_1(xa_{t_n}) = f_1^*(x)$$

in the weak$*$ topology. In other words, we have

$$\lim_{n\to\infty} \langle f_1(xa_{t_n}), f_2\rangle = \langle f_1^*, f_2\rangle$$

[15]Hopefully without destroying the beauty of the argument in the eye of the reader, we wish to point out a small technical detail. Clearly every G-invariant set (measurable or not) is empty or everything but for Proposition 29 another notion of invariance is important. A set is called G-invariant modulo the Haar measure m_X if it is measurable and $m_X(B \triangle g\cdot B) = 0$ for all $g \in G$, or equivalently if the characteristic function is an invariant function in $L^2(X, m_X)$. What is needed is that any G-invariant set modulo m_X has measure zero or one. This requires a small argument on the same order of complexity as the uniqueness theorem for the Haar measure on locally compact groups.

for all $f_2 \in L^2(X,\mu)$. We claim that f_1^* is invariant under the action of the unipotent subgroup

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

In fact

$$\begin{aligned} \langle f_1^*(xu_s), f_2 \rangle &= \langle f_1^*, f_2(xu_{-s}) \rangle \\ &= \lim_{n\to\infty} \langle f_1(xa_{t_n}), f_2(xu_{-s}) \rangle \\ &= \lim_{n\to\infty} \langle f_1(xu_s a_{t_n}), f_2 \rangle \\ &= \lim_{n\to\infty} \langle f_1(xa_{t_n} u_{\mathrm{e}^{-t_n} s}), f_2 \rangle . \end{aligned}$$

However,

$$\| f_1(yu_{\mathrm{e}^{-t_n} s}) - f_1(y) \|_2 \longrightarrow 0$$

as $\mathrm{e}^{-t_n} s \to 0$, and since the substitution $y = xa_{t_n}$ is unitary we obtain

$$\langle f_1^*(xu_s), f_2 \rangle = \lim_{n\to\infty} \langle f_1(xa_{t_n}), f_2 \rangle = \langle f_1^*, f_2 \rangle .$$

This shows that

$$f_1^*(xu_s) = f_1^*(x)$$

for almost every x and all $s \in \mathbb{R}$. Now the assumption that G acts ergodically implies together with Proposition 29 that f_1^* coincides with a constant almost everywhere, which quickly shows

$$f_1^* = \int f_1 \, \mathrm{d}\mu.$$

This is the desired conclusion. □

The general case of $g \in \mathrm{SL}_2(\mathbb{R})$ going to infinity (not necessarily in the diagonal subgroup) surprisingly can be derived from the above case. Here one uses the Cartan decomposition of matrices which states that any g can be written as $k_1 a k_2$ where a is diagonal and $k_1, k_2 \in \mathrm{SO}(2)$ belong to the compact group of rotations $\mathrm{SO}(2)$.

Exercise 32. Show the Cartan decomposition of elements of $\mathrm{SL}_2(\mathbb{R})$ and finish the proof of Theorem 31.

6 Unique ergodicity on compact quotients

We will now show the unique ergodicity of the horocycle flow on compact quotients of $\mathrm{SL}_2(\mathbb{R})$, but we will derive this in a much more general framework for Lie groups and their horospherical subgroups.

Suppose that G is a connected linear Lie group and let $a \in G$ be an $\mathbb{R}$-diagonalizable element that acts as a mixing transformation[16] on all the quotients of G appearing below. Let

$$G_a^- = \left\{ g \in G \mid a^n g a^{-n} \to I \text{ as } n \to \infty \right\}$$

be the stable horospherical subgroup of a.

Theorem 33 (Unique ergodicity of horospherical actions[17])**.** *Let G be a linear Lie group, $\Gamma < G$ a uniform lattice, and let $a \in G$ be $\mathbb{R}$-diagonalizable. Suppose a acts mixingly on $X = \Gamma\backslash G$. Then the action of G_a^- is uniquely ergodic.*

Proof. Let us assume compatibility of the Haar measures in the sense that

$$m_X(\pi(B)) = m_G(B)$$

for any injective[18] Borel subset $B \subset G$, and that $m_X(X) = 1$.

Since a is diagonalizable (we will even assume that a is diagonal) and G is linear, the subgroups G_a^- and

$$P_a = \left\{ g \in G \mid a^n g a^{-n} \text{ stays bounded as } n \to -\infty \right\}$$

can easily be defined in terms of the vanishing of certain matrix entries, and so are closed subgroups. Together they define a local coordinate system, in the sense that $P_a G_a^-$ contains an open neighborhood of the identity[19], and the implied representation of elements of G in that neighborhood is unique. In fact, if $u_1 p_1 = u_2 p_2$ with $u_1, u_2 \in G_a^-$ and $p_1, p_2 \in P_a$ then

$$g = u_2^{-1} u_1 = p_2 p_1^{-1}$$

has $a^n g a^{-n} \to I$ as $n \to \infty$ and stays bounded as $n \to -\infty$, which together show[20] that $g = I$. Moreover, the Haar measure of G restricts to the product of a Haar measure on G_a^- and a left Haar measure on P_a.

We let $B_0 \subset G_a^-$ be a neighborhood of the identity with compact closure such that $m_{G_a^-}(\partial B_0) = 0$ and define $B_n = a^{-n} B_0 a^n$. We claim that

$$\frac{1}{m_{G_a^-}(B_n)} \int_{B_n} f(u \cdot x) \, \mathrm{d}m_{G_a^-}(u) \longrightarrow \int_X f \, \mathrm{d}m_X \tag{12}$$

for any $f \in C(X)$ and any $x \in X$.

[16] Unless a specific other probability measure is identified, a property of a transformation on a homogeneous space like ergodicity, mixing, and so on, is meant with respect to the measure induced by the Haar measure on G.

[17] This is an example of a circle of results developed among others by Dani and Veech.

[18] The set B is called injective if the quotient map $\pi : G \to X$ is injective when restricted to B.

[19] This can be quickly checked using the Lie algebras of G_a^- (and of P_a), which are simply the sum of the eigenspaces of Ad_a for all eigenvalues of absolute value less than one (respectively greater than or equal to one).

[20] This is a consequence of considering the eigenvalue decomposition of the matrix $g - I$ in $\mathrm{Mat}_d(\mathbb{R})$ for the linear map $\mathrm{Mat}_d(\mathbb{R}) \ni v \longmapsto ava^{-1}$.

Assuming this for now, it follows that $\mu = m_X$ is the only G_a^--invariant probability measure. Indeed if μ is another such measure then

$$\int_X f \,\mathrm{d}\mu = \int_X \frac{1}{m_{G_a^-}(B_n)} \int_{B_n} f(u{\cdot}x) \,\mathrm{d}m_{G_a^-}(u) \,\mathrm{d}\mu(x) \longrightarrow \int_X f \,\mathrm{d}m_X$$

by dominated convergence. As this would hold for any $f \in C(X)$ we deduce that $\mu = m_X$, as claimed.

Now fix a point $x \in X = \Gamma\backslash G$ and a function $f \in C(X)$. By compactness f is uniformly continuous, so given $\varepsilon > 0$ there is a $\delta > 0$ for which

$$\mathsf{d}(h, e) < \delta \implies |f(h{\cdot}y) - f(y)| < \varepsilon \tag{13}$$

where d is a left-invariant metric on G (giving rise to the metric on X). Now we can choose a compact neighborhood $V \subset P_a$ of the identity whose boundary has measure zero with

$$\mathsf{d}(a^{-n}ha^n, e) < \delta$$

for $h \in V$ and $n \geqslant 0$. Then

$$\frac{1}{m_{G_a^-}(B_n)} \int_{B_n} f(u{\cdot}x) \,\mathrm{d}m_{G_a^-}(u)$$

is within ε of

$$\frac{1}{m_{G_a^-}(B_n) m_{P_a}(a^{-n}Va^n)} \int_{B_n} \int_{a^{-n}Va^n} f(hu{\cdot}x) \,\mathrm{d}m_{P_a}(h) \,\mathrm{d}m_{G_a^-}(u)$$

because of (13). Using $B_n = a^{-n}B_0a^n$, the latter may in turn be written as

$$\frac{1}{m_G(VB_0)} \int_{VB_0} f(a^{-n}ga^n{\cdot}x) \,\mathrm{d}m_G(g), \tag{14}$$

since m_G is locally the product of m_{P_a} and $m_{G_a^-}$. Now notice that the map

$$G_a^- \ni u \mapsto u{\cdot}x$$

is injective for all $x \in X$, for otherwise the injectivity radius at $a^n{\cdot}x$ would shrink to zero, contradicting the compactness of X. By a simple compactness argument, we may assume that the above δ is small enough to ensure that the map

$$VB_0 \ni g \mapsto g{\cdot}x$$

is injective for all $x \in X$. Thus (14) can also be written as

$$\frac{1}{m_G(VB_0)} \int_X f(a^{-n}y) \mathbb{1}_{VB_0a^n{\cdot}x}(y) \,\mathrm{d}m_X. \tag{15}$$

In the sequence (or in any of its subsequences) $(a^n{\cdot}x)_{n\geqslant 1}$ we can find (by compactness) a subsequence $(a^{n_k}{\cdot}x)_{k\geqslant 1}$ converging to some $z \in X$. Since[21]

$$\|\mathbb{1}_{VB_0a^{n_k}{\cdot}x} - \mathbb{1}_{VB_0{\cdot}z}\|_2 \longrightarrow 0$$

[21] Here we are making use of the fact that $m_G(\partial(VB_0)) = 0$, which follows since m_G is the product measure in the local coordinate system $G_a^-P_a$ of G that we use and we already know that $m_{G_a^-}(\partial B_0) = 0$ and $m_{P_a}(\partial V) = 0$.

by dominated convergence as $k \to \infty$, we see that the expression (15) converges to

$$\frac{1}{m_G(VB_0)} \int_X f \, \mathrm{d}m_X \int \mathbb{1}_{VB_0 \cdot z} \, \mathrm{d}m_X$$

as $n \to \infty$ because a defines a mixing transformation. This proves (12) for the given function f up to an error of ε. Since f and $\varepsilon > 0$ were both arbitrary, the theorem follows. □

Notice that once unique ergodicity is proved by using the Følner sequence

$$\left(a^{-n} B_0 a^n\right)_{n \geqslant 1},$$

then the pointwise everywhere convergence of the ergodic averages also follows for other Følner sets (see Exercises 34–35).

Exercise 34. We let $B_n = a^{-n} B_0 a^n$ be as in the proof of Theorem 33, with

$$m_{G_a^-}(\partial B_0) = 0.$$

Show that this sequence is a Følner sequence in G_a^-, that is a sequence satisfying

$$\frac{m_{G_a^-}(B_n \triangle (KB_n))}{m_{G_a^-}(B_n)} \longrightarrow 0$$

as $n \to \infty$ for every compact subset $K \subset G_a^-$.

Exercise 35. Let a and X be as in Theorem 33. Let $F_n \subset G_a^-$ be any Følner sequence and show that

$$\frac{1}{m_{G_a^-}(F_n)} \int_{F_n} f(u \cdot x) \, \mathrm{d}m_{G_a^-}(u) \to \int_X f \, \mathrm{d}m_X$$

as $n \to \infty$, for any $f \in C(X)$ and any $x \in X$.

7 Ratner's theorems in unipotent dynamics

We let $X = \Gamma \backslash G$, where G is a connected Lie group and $\Gamma < G$ a lattice. Let

$$U = \{u_s \mid s \in \mathbb{R}\} < G$$

be a one-parameter unipotent subgroup of G. Then the U-invariant probability measures on X can be completely classified. This was conjectured by Dani (as an analog of Raghunathan's conjecture, which will be described below) and proved by Ratner. We only state these important results and refer to the original papers [9, 10, 11, 12, 8], the monograph [7] and the survey [6] for proofs and more details.

Theorem 36 (Dani's conjecture; Ratner's measure classification)**.** *If $X = \Gamma \backslash G$ and $U = \{u_s \mid s \in \mathbb{R}\} < G$ is a one-parameter unipotent subgroup, then every U-invariant ergodic probability measure μ on X is*[22] algebraic. *That is, there exists*

[22] An alternative term that is used is *homogeneous*.

a closed connected unimodular subgroup L with $U \leqslant L \leqslant G$ such that μ is the L-invariant normalized probability measure (that is, the normalized Haar measure) on a closed orbit $L{\cdot}x_0$ (for any $x_0 \in \operatorname{supp}\mu$).

In this result it is sufficient to assume that Γ is discrete or even just closed. Theorem 36 and Theorem 33 suggest other results which we now start to describe.

Theorem 37 (Ratner's equidistribution theorem)**.** *Let $X = \Gamma\backslash G$ where Γ is a lattice, and let $U = \{u_s \mid s \in \mathbb{R}\} < G$ be a one-parameter unipotent subgroup. Then for any $x_0 \in X$ there exists some closed connected unimodular subgroup $L \leqslant G$ such that $U \leqslant L$,*

- *$L{\cdot}x_0$ is closed with finite L-invariant volume, and*
- $\dfrac{1}{T}\displaystyle\int_0^T f(u_s{\cdot}x_0)\,\mathrm{d}s \longrightarrow \frac{1}{\operatorname{vol}(L{\cdot}x_0)}\int_{L{\cdot}x_0} f\,\mathrm{d}m_{L{\cdot}x_0}$ *as $T \to \infty$.*

It is interesting to note that Theorem 37 in particular implies that any point $x \in X$ returns close to itself under a unipotent flow. That is, for any one-parameter unipotent subgroup $\{u_s \mid s \in \mathbb{R}\}$ and any $x \in X$ there is a sequence $(t_k)_{k\geqslant 1}$ for which $t_k \to \infty$ and $\mathsf{d}(x, u_{t_k}{\cdot}x) \to 0$ as $k \to \infty$. This close return statement is of course incomparably weaker than Ratner's equidistribution theorem, but even this weak statement does not seem to have an independent proof to our knowledge.

Theorem 37 also suggests that the closures of orbits under the action of a unipotent one-parameter subgroup should be algebraic. A more general version of that statement is the famous conjecture of Raghunathan that motivated all of the theorems above and was proved by Ratner.

Theorem 38 (Raghunathan's conjecture; Ratner's orbit closure theorem)**.** *Suppose that $X = \Gamma\backslash G$, with G a connected Lie group and Γ a lattice. Let $H < G$ be a closed subgroup generated by one-parameter unipotent subgroups. Then for any $x_0 \in X$ the orbit closure is*[23] algebraic*, meaning that there exists some closed connected unimodular subgroup L with $H \leqslant L \leqslant G$ such that*

$$\overline{H{\cdot}x_0} = L{\cdot}x_0$$

and $L{\cdot}x_0$ supports a finite L-invariant measure.

It is also interesting to ask what the structure of the set of all probability measures that are invariant and ergodic under some unipotent flow really is. This generalizes a theorem of Sarnak concerning periodic horocycle orbits. At first sight, one might only ask this out of curiosity or to satisfy the urge to complete our understanding of this aspect of these dynamical systems. However, this line of enquiry turns out to be useful for applications to number-theoretic problems. A satisfying answer to this question is given by Mozes and Shah.

Theorem 39 (Mozes–Shah equidistribution theorem)**.** *Let $X = \Gamma\backslash G$ with G a connected Lie group and Γ a lattice, and let $H_n < G$ be a sequence of subgroups*

[23] Again this is also called *homogeneous*.

generated by unipotent one-parameter subgroups. Let μ_n be an invariant ergodic probability measure for the action of H_n for all $n \geqslant 1$. Assume that[24] *$\mu_n \to \mu$ in the weak*-topology as $n \to \infty$. Then one of the following two possibilities holds.*

(1) *$\mu = 0$, and $\operatorname{supp}\mu_n \to \infty$ as $n \to \infty$ in the sense that for every compact set $K \subset X$ there is an N with $\operatorname{supp}\mu_n \cap K = \emptyset$ for $n \geqslant N$.*

(2) *$\mu = m_{L\cdot y}$ is the L-invariant probability measure on a closed finite volume orbit $L\cdot y$ for the closed connected group $L = \operatorname{Stab}_G(\mu)^\circ \leqslant G$. Moreover, μ is invariant and ergodic for the action of a one-parameter unipotent subgroup. Furthermore, suppose that $x_n = \varepsilon_n \cdot x \in \operatorname{supp}\mu_n$ for $n \geqslant 1$ and some $x \in X$ with $\varepsilon_n \to 1$ as $n \to \infty$, and suppose the connected subgroups (L_n) satisfy $\mu_n = m_{L_n \cdot x_n}$ for $n \geqslant 1$. Then $xL = yL = \operatorname{supp}\mu$ and there exists some N with $\varepsilon_n^{-1} L_n \varepsilon_n \subset L$.*

The additional information in each case is useful in applying this theorem. According to (1), once we know that for every measure μ_n there exists some point $x_n \in \operatorname{supp}\mu_n$ within a fixed compact set, the limit measure is a probability measure.

In (2), if we know that $H_n = H$ for all $n \geqslant 1$, then L has to contain H and the conjugates $\varepsilon_n^{-1} H \varepsilon_n$ as in (2). Together this often puts severe limitations on the possibilities that $L \leqslant G$ can take, and sometimes forces L to be G. This situation arises, for example, if we study long periodic horocycle orbits, or orbits of a maximal subgroup $H < G$. In any case, the final claim of (2) says that the convergence to the limit measure $m_{L\cdot x}$ is almost from within the orbit $L\cdot x$. In fact, after modifying the measures in the sequence only slightly by the elements ε_n we get

$$\operatorname{supp}\left((\varepsilon_n)_*^{-1}\mu_n\right) = \varepsilon_n^{-1} L_n \cdot x_n = \varepsilon_n^{-1} L_n \varepsilon_n \cdot x \subset L\cdot x = L\cdot y = \operatorname{supp}\mu$$

for $n \geqslant N$.

References

[1] V. Bergelson, Ergodic Ramsey theory—an update, in *Ergodic theory of $\mathbb{Z}^d$ actions (Warwick, 1993–1994)*, in *London Math. Soc. Lecture Note Ser.* **228**, pp. 1–61 (Cambridge Univ. Press, Cambridge, 1996).

[2] V. Bergelson, Ergodic Ramsey theory: a dynamical approach to static theorems, in *International Congress of Mathematicians. Vol. II*, pp. 1655–1678 (Eur. Math. Soc., Zürich, 2006).

[3] M. Einsiedler and T. Ward, *Ergodic theory with a view towards number theory*, in *Graduate Texts in Mathematics* **259** (Springer-Verlag London Ltd., London, 2011).

[4] M. Einsiedler and T. Ward, *Homogeneous dynamics and applications*, `http://maths.dur.ac.uk/~tpcc68/homogeneous/welcome.html`

[5] A. Katok and B. Hasselblatt, *Introduction to the modern theory of dynamical systems*, in *Encyclopedia of Mathematics and its Applications* **54** (Cambridge

[24] By Tychonoff-Alaoglu there always exists a subsequence that converges.

University Press, Cambridge, 1995). With a supplementary chapter by Anatole Katok and Leonardo Mendoza.

[6] D. Kleinbock, N. Shah, and A. Starkov, 'Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory', in *Handbook of dynamical systems, Vol. 1A*, pp. 813–930 (North-Holland, Amsterdam, 2002).

[7] D. W. Morris, *Ratner's theorems on unipotent flows*, in *Chicago Lectures in Mathematics* (University of Chicago Press, Chicago, IL, 2005).

[8] S. Mozes and N. Shah, On the space of ergodic invariant measures of unipotent flows, *Ergodic Theory Dynam. Systems* **15** (1995), no. 1, 149–159.

[9] M. Ratner, On measure rigidity of unipotent subgroups of semisimple groups, *Acta Math.* **165** (1990), no. 3-4, 229–309.

[10] M. Ratner, Strict measure rigidity for unipotent subgroups of solvable groups, *Invent. Math.* **101** (1990), no. 2, 449–482.

[11] M. Ratner, On Raghunathan's measure conjecture, *Ann. of Math. (2)* **134** (1991), no. 3, 545–607.

[12] M. Ratner, Raghunathan's topological conjecture and distributions of unipotent flows, *Duke Math. J.* **63** (1991), no. 1, 235–280.