Preface

Many mathematicians are aware of some of the dramatic interactions between ergodic theory and other parts of the subject, notably Ramsey theory, infinite combinatorics, and Diophantine number theory. These notes are intended to provide a gentle route to a tiny sample of these results. The intended readership is expected to be mathematically sophisticated, with some background in measure theory and functional analysis, or to have the resilience to learn some of this material along the way from other sources.

In this volume we develop the beginnings of ergodic theory and dynamical systems. While the selection of topics has been made with the applications to number theory in mind, we also develop other material to aid motivation and to give a more rounded impression of ergodic theory. Different points of view on ergodic theory, with different kinds of examples, may be found in the monographs of Cornfeld, Fomin and Sinaĭ [60], Petersen [282], or Walters [373]. Ergodic theory is one facet of dynamical systems; for a broad perspective on dynamical systems see the books of Katok and Hasselblatt [182] or Brin and Stuck [44]. An overview of some of the more advanced topics we hope to pursue in a subsequent volume may be found in the lecture notes of Einsiedler and Lindenstrauss [80] in the Clay proceedings of the Pisa Summer school.

Fourier analysis of square-integrable functions on the circle is used extensively. The more general theory of Fourier analysis on compact groups is not essential, but is used in some examples and results. The ergodic theory of commuting automorphisms of compact groups is touched on using a few examples, but is not treated systematically. It is highly developed elsewhere: an extensive treatment may be found in the monograph by Schmidt [332]. Standard background material on measure theory, functional analysis and topological groups is collected in the appendices for convenience.

Among the many *lacunae*, some stand out: Entropy theory; the isomorphism theory of Ornstein, a convenient source being Rudolph [324]; the more advanced spectral theory of measure-preserving systems, a convenient source being Nadkarni [264]; finally Pesin theory and smooth ergodic theory, a source

being Barreira and Pesin [19]. Of these omissions, entropy theory is perhaps the most fundamental for applications in number theory, and this was the reason for not including it here. There is simply too much to say about entropy to fit into this volume, so we will treat this important topic, both in general terms and in more detail in the algebraic context needed for number theory, in a subsequent volume. The notion is mentioned in one or two places in this volume, but is never used directly.

No Lie theory is assumed, and for that reason some arguments here may seem laborious in character and limited in scope. Our hope is that seeing the language of Lie theory emerge from explicit matrix manipulations allows a relatively painless route into the ergodic theory of homogeneous spaces. This will be carried further in a subsequent volume, where some of the deeper applications will be given.

Notation and Conventions

The symbols $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and $\mathbb{Z}$ denote the natural numbers, non-negative integers and integers; $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ denote the rational numbers, real numbers and complex numbers; $\mathbb{S}^1$, $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ denote the multiplicative and additive circle respectively. The elements of $\mathbb{T}$ are thought of as the elements of $[0, 1)$ under addition modulo 1. The real and imaginary parts of a complex number are denoted $x = \Re(x + \mathrm{i}y)$ and $y = \Im(x + \mathrm{i}y)$. The order of growth of real- or complex-valued functions f, g defined on $\mathbb{N}$ or $\mathbb{R}$ with $g(x) \neq 0$ for large x is compared using Landau's notation:

$$f \sim g \text{ if } \left|\frac{f(x)}{g(x)}\right| \longrightarrow 1 \text{ as } x \to \infty;$$
$$f = \mathrm{o}(g) \text{ if } \left|\frac{f(x)}{g(x)}\right| \longrightarrow 0 \text{ as } x \to \infty.$$

For functions f, g defined on $\mathbb{N}$ or $\mathbb{R}$, and taking values in a normed space, we write $f = \mathrm{O}(g)$ if there is a constant $A > 0$ with $\|f(x)\| \leqslant A\|g(x)\|$ for all x. In particular, $f = \mathrm{O}(1)$ means that f is bounded. Where the dependence of the implied constant A on some set of parameters $\mathscr{A}$ is important, we write $f = \mathrm{O}_{\mathscr{A}}(g)$. The relation $f = \mathrm{O}(g)$ will also be written $f \ll g$, particularly when it is being used to express the fact that two functions are commensurate, $f \ll g \ll f$. A sequence $a_1, a_2, \dots$ will be denoted (a_n). Unadorned norms $\|x\|$ will only be used when x lives in a Hilbert space (usually L^2) and always refer to the Hilbert space norm. For a topological space X, $C(X)$, $C_{\mathbb{C}}(X)$, $C_c(X)$ denote the space of real-valued, complex-valued, compactly supported continuous functions on X respectively, with the supremum norm. For sets A, B, denote the set difference by

$$A \smallsetminus B = \{x \mid x \in A, x \notin B\}.$$

Additional specific notation is collected in an index of notation on page 471.

Statements and equations are numbered consecutively within chapters, and exercises are numbered in sections. Theorems without numbers in the main body of the text will not be proved; appendices contain background material in the form of numbered theorems that will not be proved here.

Several of the issues addressed in this book revolve around *measure rigidity*, in which there is a natural measure that other measures are compared with. These natural measures will usually be Haar measure on a compact or locally compact group, or measures constructed from Haar measures, and these will usually be denoted m.

We have not tried to be exhaustive in tracing the history of the ideas used here, but have tried to indicate some of the rich history of mathematical developments that have contributed to ergodic theory. Certain references to earlier and to related material is generally collected in endnotes at the end of each chapter; the presence of these references should not be viewed in any way as authoritative. Statements in these notes are informed throughout by a desire to remain rooted in the familiar territory of ergodic theory. The standing assumption is that, unless explicitly noted otherwise, metric spaces are complete and separable, compact groups are metrizable, discrete groups are countable, countable groups are discrete, and measure spaces are assumed to be Borel probability spaces (this assumption is only relevant starting with Section 5.3; see Definition 5.13 for the details). A convenient summary of the measure-theoretic background may be found in the work of Royden [320] or of Parthasarathy [280].

Acknowledgements

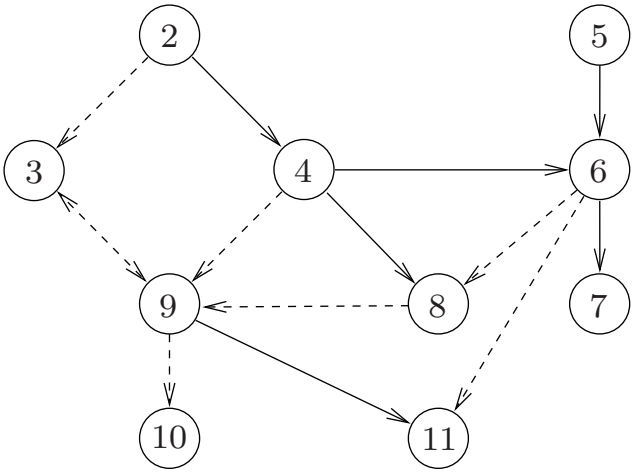
It is inevitable that we have borrowed ideas and used them inadvertently without citation, and certain that we have misunderstood, misrepresented or misattributed some historical developments; we apologize for any egregious instances of this. We are grateful to several people for their comments on drafts of sections, including Alex Abercrombie, Menny Aka, Sarah Bailey-Frick, Tania Barnett, Vitaly Bergelson, Michael Björklund, Florin Boca, Will Cavendish, Tushar Das, Jerry Day, Jingsong Chai, Alexander Fish, Anthony Flatters, Nikos Frantzikinakis, Jenny George, John Griesmer, Shirali Kadyrov, Cor Kraaikamp, Beverly Lytle, Fabrizio Polo, Christian Röttger, Nimish Shah, Ronggang Shi, Christoph Übersohn, Alex Ustian, Peter Varju and Barak Weiss; the second named author also thanks John and Sandy Phillips for sustaining him with coffee at Le Pas Opton in Summer 2006 and 2009.

We both thank our previous and current home institutions Princeton University, the Clay Mathematics Institute, The Ohio State University, Eidgenössische Technische Hochschule Zürich, and the University of East Anglia, for support, including support for several visits, and for providing the rich mathematical environments that made this project possible. We also thank the National Science Foundation for support under NSF grant DMS-0554373.

February 28, 2010 Manfred Einsiedler, Zürich
Thomas Ward, Norwich

Leitfaden

The dependencies between the chapters is illustrated below, with solid lines indicating logical dependency and dotted lines indicating partial or motivational links.



Some possible shorter courses could be made up as follows.

- Chapters 2 & 4: A gentle introduction to ergodic theory and topological dynamics.
- Chapters 2 & 3: A gentle introduction to ergodic theory and the continued fraction map (the dotted line indicates that only parts of Chapter 2 are needed for Chapter 3).
- Chapters 2, 3, & 9: As above, with the connection between the Gauss map and hyperbolic surfaces, and ergodicity of the geodesic flow.
- Chapters 2, 4, & 8: An introduction to ergodic theory for group actions.

The highlights of this book are Chapters 7 and 11. Some more ambitious courses could be made up as follows.

- To Chapter 6: Ergodic theory up to conditional measures and the ergodic decomposition.
- To Chapter 7: Ergodic theory including the Furstenberg–Katznelson–Ornstein proof of Szemerédi's theorem.
- To Chapter 11: Ergodic theory and an introduction to dynamics on homogeneous spaces, including equidistribution of horocycle orbits. A minimal path to equidistribution of horocycle orbits on $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ would include the discussions of ergodicity from Chapter 2, genericity from Chapter 4, Haar measure from Chapter 8, the hyperbolic plane from Chapter 9, and ergodicity and mixing from Chapter 11.

Contents

Chapter 1
Motivation

Our main motivation throughout the book will be to understand the applications of ergodic theory to certain problems outside of ergodic theory, in particular to problems in number theory. As we will see, this requires a good understanding of particular examples, which will often be of an algebraic nature. Therefore, we will start with a few concrete examples, and state a few theorems arising from ergodic theory, some of which we will prove within this volume. In Section 1.8 we will discuss ergodic theory as a subject in more general terms[(1)].

1.1 Examples of Ergodic Behavior

The *orbit* of a point $x \in X$ under a transformation $T : X \to X$ is the set $\{T^n(x) \mid n \in \mathbb{N}\}$. The structure of the orbit can say a great deal about the original point x. In particular, the behavior of the orbit will sometimes detect special properties of the point. A particularly simple instance of this appears in the next example.

Example 1.1. Write $\mathbb{T}$ for the quotient group $\mathbb{R}/\mathbb{Z} = \{x + \mathbb{Z} \mid x \in \mathbb{R}\}$, which can be identified with a circle (as a topological space, this can also be obtained as a quotient space of $[0, 1]$ by identifying 0 with 1); there is a natural bijection between $\mathbb{T}$ and the half-open interval $[0, 1)$ obtained by sending the coset $x+\mathbb{Z}$ to the fractional part of x. Let $T : \mathbb{T} \to \mathbb{T}$ be defined by $T(x) = 10x \pmod 1$. Then $x \in \mathbb{T}$ is rational if and only if the orbit of x under T is finite. To see this, assume first that $x = \frac{p}{q}$ is rational. In this case the orbit of x is some subset of $\{0, \frac{1}{q}, \ldots, \frac{q-1}{q}\}$. Conversely, if the orbit is finite then there must be integers m, n with $1 \leqslant n < m$ for which $T^m(x) = T^n(x)$. It follows that $10^m x = 10^n x + k$ for some $k \in \mathbb{N}$, so x is rational.

Detecting the behavior of the orbit of a given point is usually not so straightforward. Ergodic theory generally has more to say about the orbit of

"typical" points, as illustrated in the next example. Write χ_A for the indicator function of a set,

$$\chi_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A. \end{cases}$$

Example 1.2. This example recovers a result due to Borel [40]. We shall see later that the map $T : \mathbb{T} \to \mathbb{T}$ defined by $T(x) = 10x \pmod 1$ preserves Lebesgue measure m on $[0, 1)$ (see Definition 2.1), and is *ergodic* with respect to m (see Definition 2.13). A consequence of the pointwise ergodic theorem (Theorem 2.30) is that for any interval

$$A(j, k) = [\tfrac{j}{10^k}, \tfrac{j+1}{10^k}),$$

we have

$$\frac{1}{N} \sum_{i=0}^{N-1} \chi_{A(j,k)}(T^i x) \longrightarrow \int_0^1 \chi_{A(j,k)}(x)\, \mathrm{d}m(x) = \frac{1}{10^k} \tag{1.1}$$

as $N \to \infty$, for almost every x (that is, for all x in the complement of a set of zero measure, which will be denoted a.e.). For any block $j_1 \dots j_k$ of k decimal digits, the convergence in equation (1.1) with $j = 10^{k-1}j_1 + 10^{k-2}j_2 + \cdots + j_k$ shows that the block $j_1 \dots j_k$ appears with asymptotic frequency $\frac{1}{10^k}$ in the decimal expansion of almost every real number in $[0, 1]$.

Even though the ergodic theorem only concerns the orbital behavior of typical points, there are situations where one is able to describe the orbits for *all* starting points.

Example 1.3. We show later that the circle rotation $R_\alpha : \mathbb{T} \to \mathbb{T}$ defined by $R_\alpha(t) = t + \alpha \pmod 1$ is *uniquely ergodic* if α is irrational (see Definition 4.9 and Example 4.11). A consequence of this is that for any interval $[a, b) \subseteq [0, 1) = \mathbb{T}$,

$$\frac{1}{N} \sum_{n=0}^{N-1} \chi_{[a,b)}(R_\alpha^n(t)) \longrightarrow b - a \tag{1.2}$$

as $N \to \infty$ for *every* $t \in \mathbb{T}$ (see Theorem 4.10 and Lemma 4.17). As pointed out by Arnol′d and Avez [7] this equidistribution result may be used to find the density of appearance of the digits[(2)] in the sequence $1, 2, 4, 8, 1, 3, 6, 1, \dots$ of first digits of the powers of 2:

$$\mathbf{1}, \mathbf{2}, \mathbf{4}, \mathbf{8}, \mathbf{1}6, \mathbf{3}2, \mathbf{6}4, \mathbf{1}28, \mathbf{2}56, \mathbf{5}12, \mathbf{1}024, \dots .$$

A set $A \subseteq \mathbb{N}$ is said to have *density* $\mathbf{d}(A)$ if

$$\mathbf{d}(A) = \lim_{k \to \infty} \frac{1}{k} \left| A \cap [1, k] \right|$$

exists. Notice that 2^n has first digit k for some $k \in \{1, 2, \ldots, 9\}$ if and only if

$$\log_{10} k \leqslant \{n \log_{10} 2\} < \log_{10}(k+1),$$

where we write $\{t\}$ for the fractional part of the real number t.

Since $\alpha = \log_{10} 2$ is irrational, we may apply equation (1.2) to deduce that

$$\begin{aligned} \frac{\left|\{n \,|\, 0 \leqslant n \leqslant N-1,\, 1^{\text{st}} \text{ digit of } 2^n \text{ is } k\}\right|}{N} &= \frac{1}{N} \sum_{n=0}^{N-1} \chi_{[\log_{10} k, \log_{10}(k+1))}(R_\alpha^n(0)) \\ &\to \log_{10}\left(\frac{k+1}{k}\right) \end{aligned}$$

as $N \to \infty$.

Thus the first digit $k \in \{1, \ldots, 9\}$ appears with density $\log_{10}\left(\frac{k+1}{k}\right)$, and it follows in particular that the digit 1 is the most common leading digit in the sequence of powers of 2.

Exercises for Section 1.1

Exercise 1.1.1. A point $x \in X$ is said to be *periodic* for the map $T : X \to X$ if there is some $k \geqslant 1$ with $T^k(x) = x$, and *pre-periodic* if the orbit of x under T is finite. Describe the periodic points and the pre-periodic points for the map $x \mapsto 10x \pmod 1$ from Example 1.1.

Exercise 1.1.2. Prove that the orbit of any point $x \in \mathbb{T}$ under the map R_α on $\mathbb{T}$ for α irrational is dense (that is, for any $\varepsilon > 0$ and $t \in \mathbb{T}$ there is some $k \in \mathbb{N}$ for which $T^k x$ lies within ε of t). Deduce that for any finite block of decimal digits, there is some power of 2 that begins with that block of digits.

1.2 Equidistribution for Polynomials

A sequence $(a_n)_{n \in \mathbb{N}}$ of numbers in $[0, 1)$ is said to be equidistributed if

$$\mathbf{d}(\{n \in \mathbb{N} \mid a \leqslant a_n < b\}) = b - a$$

for all a, b with $0 \leqslant a < b \leqslant 1$. A classical result of Weyl [380] extends the equidistribution of the numbers $(n\alpha)_{n \in \mathbb{N}}$ modulo 1 for irrational α to the values of any polynomial with an irrational coefficient*.

* Numbered theorems like Theorem 1.4 in the main text are proved in this volume, but not necessarily in the chapter in which they first appear.

Theorem 1.4 (Weyl). *Let $p(n) = a_k n^k + \cdots + a_0$ be a real polynomial with at least one coefficient among $a_1, \ldots, a_k$ irrational. Then the sequence $(p(n))$ is equidistributed modulo* 1.

Furstenberg extended unique ergodicity to a dynamically defined extension of the irrational circle rotation described in Example 1.3, giving an elegant ergodic-theoretic proof of Theorem 1.4. This approach will be discussed in Section 4.4.

Exercises for Section 1.2

Exercise 1.2.1. Describe what Theorem 1.4 can tell us about the leading digits of the powers of 2.

1.3 Szemerédi's Theorem

Szemerédi, in an intricate and difficult combinatorial argument, proved a long-standing conjecture of Erdős and Turán [85] in his paper [357]. A set S of integers is said to have *positive upper Banach density* if there are sequences (m_j) and (n_j) with $n_j - m_j \to \infty$ as $j \to \infty$ with the property that

$$\lim_{j\to\infty} \frac{|S \cap [m_j, n_j]|}{n_j - m_j} > 0.$$

Theorem 1.5 (Szemerédi). *Any subset of the integers with positive upper Banach density contains arbitrarily long arithmetic progressions.*

Furstenberg [102] (see also his book [103] and the article of Furstenberg, Katznelson and Ornstein [107]) showed that Szemerédi's theorem would follow from a generalization of Poincaré's recurrence theorem, and proved that generalization. The connection between recurrence and Szemerédi's theorem will be explained in Section 7.3, and Furstenberg's proof of the generalization of Poincaré recurrence needed will be presented in Chapter 7. There are a great many more theorems in this direction which we cannot cover, but it is worth noting that many of these further theorems to date only have proofs using ergodic theory.

More recently, Gowers [122] has given a different proof of Szemerédi's theorem, and in particular has found the following effective form of it*.

Theorem (Gowers). For every integer $s \geqslant 1$ and sufficiently large integer N, every subset of $\{1, 2, \ldots, N\}$ with at least

* Theorems and other results that are not numbered will not be proved in this volume, but will also not be used in the main body of the text.

$$N(\log\log N)^{-2^{-2^{s+9}}}$$

elements contains an arithmetic progression of length s.

Typically proofs using ergodic theory are not effective: Theorem 1.5 easily implies a finitistic version of Szemerédi's theorem, which states that for every s and constant $c > 0$ and all sufficiently large $N = N(s, c)$, any subset of $\{1, \ldots, N\}$ with at least cN elements contains an arithmetic progression of length s. However, the dependence of N on c is not known by this means, nor is it easily deduced from the proof of Theorem 1.5. Gowers' Theorem, proved by different methods, does give an explicit dependence.

We mention Gowers' Theorem to indicate some of the limitations of ergodic theory. While ergodic methods have many advantages, proving quite general theorems which often have no other proofs, they also have disadvantages, one of them being that they tend to be non-effective.

Subsequent development of the combinatorial and arithmetic ideas by Goldston, Pintz and Yıldırım [118]$^{(3)}$ and Gowers, and of the ergodic method by Host and Kra [159] and Ziegler [392], has influenced some arguments of Green and Tao [127] in their proof of the following long-conjectured result. This is a good example of how asking for effective or quantitative versions of existing results can lead to new qualitative theorems.

Theorem (Green and Tao). The set of primes contains arbitrarily long arithmetic progressions.

1.4 Indefinite Quadratic Forms and Oppenheim's Conjecture

Our purpose here is to provide enough background in ergodic theory to quickly reach some understanding of a few deeper results in number theory and combinatorial number theory where ergodic theory has made a contribution. Along the way we will develop a good portion of ergodic theory as well as some other background material. In the rest of this introductory chapter, we mention some more highlights of the many connections between ergodic theory and number theory. The results in this section, and in Sections 1.5 and 1.6, will not be covered in this book, but we plan to discuss them in a subsequent volume.

The next theorem was conjectured in a weaker form by Oppenheim in 1929 and eventually proved by Margulis in the stronger form stated here in 1986 [247], [250]. In order to state the result, we recall some terminology for quadratic forms.

A *quadratic form* in n variables is a homogeneous polynomial $Q(x_1, \ldots, x_n)$ of degree two. Equivalently, a quadratic form is a polynomial Q for which there is a symmetric $n \times n$ matrix A_Q with

$$Q(x_1, \dots, x_n) = (x_1, \dots, x_n) A_Q (x_1, \dots, x_n)^{\mathrm{t}}.$$

Since A_Q is symmetric, there is an orthogonal matrix P for which $P^{\mathrm{t}} A_Q P$ is diagonal. This means there is a different coordinate system $y_1, \dots, y_n$ for which

$$Q(x_1, \dots, x_n) = c_1 y_1^2 + \cdots + c_n y_n^2.$$

The quadratic form is called *non-degenerate* if all the coefficients c_i are non-zero (equivalently, if $\det A_Q \neq 0$), and is called *indefinite* if the coefficients c_i do not all have the same sign. Finally, the quadratic form is said to be *rational* if its coefficients (equivalently, if the entries of the matrix A_Q) are rational*.

Theorem (Margulis). Let Q be an indefinite non-degenerate quadratic form in $n \geqslant 3$ variables that is not a multiple of a rational form. Then $Q(\mathbb{Z}^n)$ is a dense subset of $\mathbb{R}$.

It is easy to see that two of the stated conditions are necessary for the result: if the form Q is definite then the elements of $Q(\mathbb{Z}^n)$ all have the same sign, and if Q is a multiple of a rational form, then $Q(\mathbb{Z}^n)$ lies in a discrete subgroup of $\mathbb{R}$. The assumption that Q is non-degenerate and n is at least 3 are also necessary, though this is less obvious (requiring in particular the notion of badly approximable numbers from the theory of Diophantine approximation, which will be introduced in Section 3.3). This shows that the theorem as stated above is in the strongest possible form. Weaker forms of this result have been obtained by other methods, but the full strength of Margulis' Theorem at the moment requires dynamical arguments (for example, ergodic methods).

Proving the theorem involves understanding the behavior of *orbits* for the action of the subgroup $\mathrm{SO}(2,1) \leqslant \mathrm{SL}_3(\mathbb{R})$ on points $x \in \mathrm{SL}_3(\mathbb{Z}) \backslash \mathrm{SL}_3(\mathbb{R})$ (the space of right cosets of $\mathrm{SL}_3(\mathbb{Z})$ in $\mathrm{SL}_3(\mathbb{R})$); these may be thought of as sets of the form $x\,\mathrm{SO}(2,1)$. As it turns out (a consequence of Raghunathan's conjectures, discussed briefly in Section 1.7), such orbits are either closed subsets of $\mathrm{SL}_3(\mathbb{Z}) \backslash \mathrm{SL}_3(\mathbb{R})$ or are dense in $\mathrm{SL}_3(\mathbb{Z}) \backslash \mathrm{SL}_3(\mathbb{R})$. Moreover, the former case happens if and only if the point x corresponds in an appropriate sense to a rational quadratic form.

Margulis' Theorem may be viewed as an extension of Example 1.3 to higher degree in the following sense. The statement that every orbit under the map $R_\alpha(t) = t + \alpha \pmod 1$ is dense in $\mathbb{T}$ is equivalent to the statement that if L is a linear form in two variables that is not a multiple of a rational form, then $L(\mathbb{Z}^2)$ is dense in $\mathbb{R}$.

* Note that the rationality of Q cannot be detected using the coefficients $c_1, \dots, c_n$ after the real coordinate change.

1.5 Littlewood's Conjecture

For a real number t, write $\langle t\rangle$ for the distance from t to the nearest integer,

$$\langle t\rangle = \min_{q\in\mathbb{Z}} |t-q|.$$

The theory of continued fractions (which will be described in Chapter 3) shows that for any real number u, there is a sequence (q_n) with $q_n \to \infty$ such that $q_n\langle q_n u\rangle < 1$ for all $n \geqslant 1$. Littlewood conjectured the following in the 1930s: for any real numbers u, v,

$$\liminf_{n\to\infty} n\langle nu\rangle\langle nv\rangle = 0.$$

Some progress was made on this for restricted classes of numbers u and v by Cassels and Swinnerton-Dyer [50], Pollington and Velani [290], and others, but the problem remains open. In 2003 Einsiedler, Katok and Lindenstrauss [79] used ergodic methods to prove that the set of exceptions to Littlewood's conjecture is extremely small.

Theorem (Einsiedler, Katok & Lindenstrauss). Let

$$\Theta = \left\{(u,v)\in\mathbb{R}^2 \mid \liminf_{n\to\infty} n\langle nu\rangle\langle nv\rangle > 0\right\}.$$

Then the Hausdorff dimension of Θ is zero.

In fact the result in [79] is a little stronger, showing that Θ satisfies a stronger property that implies it has Hausdorff dimension zero. The proof relies on a partial classification of certain invariant measures on $\mathrm{SL}_3(\mathbb{Z})\backslash \mathrm{SL}_3(\mathbb{R})$. This is part of the theory of *measure rigidity*, and the particular type of phenomenon seen has its origins in work of Furstenberg [100], who showed that the natural action $t \mapsto at \pmod 1$ of the semi-group generated by two multiplicatively independent natural numbers a_1 and a_2 on $\mathbb{T}$ has, apart from finite sets, no non-trivial closed invariant sets. He asked if this system could have any non-atomic ergodic invariant measures other than Lebesgue measure. Partial results on this and related generalizations led to the formulation of far-reaching conjectures by Margulis [251], by Furstenberg, and by Katok and Spatzier [183], [184]. A special case of these conjectures concerns actions of the group A of positive diagonal matrices in $\mathrm{SL}_k(\mathbb{R})$ for $k \geqslant 3$ on the space $\mathrm{SL}_k(\mathbb{Z})\backslash \mathrm{SL}_k(\mathbb{R})$: if μ is an A-invariant ergodic probability measure on this space, is there a closed connected group $L \geqslant A$ for which μ is the unique L-invariant measure on a single closed L-orbit (that is, is μ *homogeneous*)?

In the work of Einsiedler, Katok and Lindenstrauss the conjecture stated above is proved under the additional hypothesis that the measure μ gives positive entropy to some one-parameter subgroup of A, which leads to the

theorem concerning Θ. A complete classification of these measures without entropy hypotheses would imply the full conjecture of Littlewood.

In this volume we will develop the minimal background needed for the ergodic approach to continued fractions (see Chapter 3) as well as the basic theorems concerning the action of the diagonal subgroup A on the quotient space $\mathrm{SL}_2(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{R})$ (see Chapter 9). We will also describe the connection between these two topics, which will help us to prove results about the continued fraction expansion and about the action of A.

1.6 Integral Quadratic Forms

An important topic in number theory, both classical and modern, is that of integral quadratic forms. A quadratic form $Q(x_1, \dots, x_n)$ is said to be *integral* if its coefficients are integers.

A natural problem[(4)] is to describe the range $Q(\mathbb{Z}^n)$ of an integral quadratic form evaluated on the integers. A classical theorem of Lagrange[(5)] on the sum of four squares says that $Q_0(\mathbb{Z}^4) = \mathbb{N}_0$ if

$$Q_0(x_1, x_2, x_3, x_4) = x_1^2 + x_2^2 + x_3^2 + x_4^2,$$

solving the problem for a particular form.

More generally, Kloosterman, in his dissertation of 1924, found an asymptotic formula for the number of expressions for an integer in terms of a positive definite quadratic form Q in five or more variables and deduced that any large integer lies in $Q(\mathbb{Z}^n)$ if it satisfies certain congruence conditions. The case of four variables is much deeper, and required him to make new deep developments in analytic number theory; special cases appeared in [201] and the full solution in [202], where he proved that an integral definite quadratic form Q in four variables represents all large enough integers a for which there is no *congruence obstruction*. Here we say that $a \in \mathbb{N}$ has a congruence obstruction for the quadratic form $Q(x_1, \dots, x_n)$ if a modulo d is not a value of $Q(x_1, \dots, x_n)$ modulo d for some $d \in \mathbb{N}$.

The methods that are usually applied to prove these theorems are purely number-theoretic. Ellenberg and Venkatesh [83] have introduced a method that combines number theory, algebraic group theory, and ergodic theory to prove results in this field, leading to a different proof of the following special case of Kloosterman's Theorem.

Theorem (Kloosterman). Let Q be a positive definite quadratic form with integer coefficients in at least 6 variables. Then all large enough integers that do not fail the congruence conditions can be represented by the form Q.

That is, if $a \in \mathbb{N}$ is larger than some constant that depends on Q and for every $d > 0$ there exists some $x_d \in \mathbb{Z}^n$ with $Q(x_d) = a$ modulo d, then there

exists some $x \in \mathbb{Z}^n$ with $Q(x) = a$. This theorem has purely number-theoretic proofs (see the survey by Schulze-Pillot [335]).

In fact Ellenberg and Venkatesh proved in [83] a different theorem that currently does not have a purely number-theoretic proof. They considered the problem of representing a quadratic form by another quadratic form: If Q is an integral positive definite[6] quadratic form in n variables and Q' is another such form in $m < n$ variables, then one can ask whether there is a subgroup $\Lambda \leqslant \mathbb{Z}^n$ generated by m elements such that when Q is restricted to Λ the resulting form is isomorphic to Q'. This question has, for instance, been studied by Gauss in the case of $m = 2$ and $n = 3$ in the *Disquisitiones Arithmeticae* [111]. As before, there can be congruence obstructions to this problem, which are best phrased in terms of p-adic numbers. Roughly speaking, Ellenberg and Venkatesh show that for a given integral definite quadratic form Q in n variables, every integral definite quadratic form Q' in $m \leqslant n-5$ variables[7] that does not have small image values can be represented by Q, unless there is a congruence obstruction. The assumption that the quadratic form Q' does not have small image means that $\min_{x \in \mathbb{Z}^m \smallsetminus \{0\}} Q'(x)$ should be bigger than some constant that depends on Q.

The ergodic theory used in [83] is related to Raghunathan's conjecture mentioned in Section 1.4 and discussed again in Section 1.7 below, and is the result of work by many people, including Margulis, Mozes, Ratner, Shah, and Tomanov.

1.7 Dynamics on Homogeneous Spaces

Let $G \leqslant \mathrm{SL}_n(\mathbb{R})$ be a closed linear group over the reals (or over a local field; see Section 9.3 for a precise definition), let $\Gamma < G$ be a discrete subgroup[8], and let $H < G$ be a closed subgroup. For example, the case $G = \mathrm{SL}_3(\mathbb{R})$ and $\Gamma = \mathrm{SL}_3(\mathbb{Z})$ arises in Section 1.4 with $H = \mathrm{SO}(2,1)$, and arises in Section 1.5 with $H = A$. Dynamical properties of the action of right multiplication by elements of H on the homogeneous space $X = \Gamma\backslash G$ is important for numerous problems[9]. Indeed, all the results in Sections 1.4–1.6 may be proved by studying concrete instances of such systems. We do not want to go into the details here, but simply mention a few highlights of the theory.

There are many important and general results on the ergodicity and mixing behavior of natural measures on such quotients (see Chapter 2 for the definitions). These results (introduced in Chapters 9 and 11) are interesting in their own right, but have also found applications to the problem of counting integer (and, more recently, rational) points on groups (or certain other varieties). The first instance of this can be found in Margulis's thesis [252], where this approach is used to find the asymptotics for the number of closed geodesics on compact manifolds of negative curvature. Independently, Eskin and McMullen [86] found the same method and applied it to a counting prob-

lem in certain varieties, which re-proved certain cases of the theorems in the work of Duke, Rudnick and Sarnak [76] in a simpler manner. However, as discussed in Section 1.1, the most difficult – and sometimes most interesting – problem is to understand the orbit of a given point rather than the orbit of almost every point. Indeed, the solution of Oppenheim's conjecture in Section 1.4 by Margulis involved understanding the $\mathrm{SO}(2,1)$-orbit of a point in $\mathrm{SL}_3(\mathbb{Z})\backslash \mathrm{SL}_3(\mathbb{R})$ corresponding to the given quadratic form.

We need one more definition before we can state a general theorem in this direction. A subgroup $U < \mathrm{SL}_n(\mathbb{R})$ is called a *one-parameter unipotent subgroup* if U is the image of $\mathbb{R}w$ under the exponential map, for some matrix $w \in \mathrm{Mat}_{nn}$ satisfying $w^n = 0$ (that is, w is nilpotent and $\exp(tw)$ has only 1 as an eigenvalue, hence the name). For example, there is an index two subgroup $H \leqslant \mathrm{SO}(2,1)$ which is generated by one-parameter unipotent subgroups. However, notice that the diagonal subgroup A is not generated by one-parameter unipotent subgroups.

Raghunathan conjectured that if the subgroup H is generated by one-parameter unipotent subgroups, then the closures of orbits xH are always of the form xL for some closed connected subgroup L of G that contains H. This reduces the properties of orbit closures (a dynamical problem) to the algebraic problem of deciding for which closed connected subgroups L the orbit xL is closed.

Ratner [305] proved this important result using methods from ergodic theory. In fact, she deduced Raghunathan's conjecture from Dani's conjecture[(10)] regarding H-invariant measures, which she proved first in the series of papers [302], [303] and [304].

To date there have been numerous applications of the above theorem, and certain extensions of it. To name a few more seemingly unrelated applications, Elkies and McMullen [82] have applied these theorems to obtain the distribution of the gaps in the sequence of fractional parts of $\sqrt{n}$, and Vatsal [366] has studied values of certain L-functions using the p-adic version of the theorems. There are further applications of the theory too numerous to describe here, but the examples above show again the variety of fields that have connections to ergodic theory.

We will discuss a few special cases of the conjectures of Raghunathan and Dani. Example 1.3, Section 4.4, Chapter 10, Section 11.5, and Section 11.7 treat special cases, some of which were known before the conjectures were formulated.

1.8 An Overview of Ergodic Theory

Having seen some statements that qualify as being ergodic in nature, and some of the many important applications of ergodic theory to number theory, in this short section we give a brief overview of ergodic theory. If this is

not already clear, notice that it is a rather diffuse subject with ill-defined boundaries[(11)].

Ergodic theory is the study of long-term behavior in dynamical systems from a statistical point of view. Its origins therefore are intimately connected with the time evolution of systems modeled by measure-preserving actions of the reals or the integers, with the action representing the passage of time. Related approaches, using probabilistic methods to study the evolution of systems, also arose in statistical physics, where other natural symmetries – typically reflected by the presence of a $\mathbb{Z}^d$-action – arise. The rich interaction between arithmetic and geometry present in measure-preserving actions of (lattices in) Lie groups quickly emerged, and it is now natural to view ergodic theory as the study of measure-preserving group actions, containing but not limited to several special branches:

(1) The classical study of single measure-preserving transformations.
(2) Measure-preserving actions of $\mathbb{Z}^d$; more generally of countable amenable groups.
(3) Measure-preserving actions of $\mathbb{R}^d$ and more general amenable groups, called flows.
(4) Measure-preserving and more general actions of groups, in particular of Lie groups and of lattices in Lie groups.

Some of the illuminating results in ergodic theory come from the existence of (counter-)examples. Nonetheless, there are many substantial theorems. In addition to fundamental results (the pointwise and mean ergodic theorems themselves, for example) and structural results (the isomorphism theorem of Ornstein, Krieger's theorem on the existence of generators, the isomorphism invariance of entropy), ergodic theory and its way of thinking have made dramatic contributions to many other fields.

Notes to Chapter 1

[(1)](Page 1) The origins of the word 'ergodic' are not entirely clear. Boltzmann coined the word *monode* (unique μὸνος, nature εἶδος) for a set of probability distributions on the phase space that are invariant under the time evolution of a Hamiltonian system, and *ergode* for a monode given by uniform distribution on a surface of constant energy. Ehrenfest and Ehrenfest (in an influential encyclopedia article of 1912, translated as [78]) called a system *ergodic* if each surface of constant energy comprised a single time orbit — a notion called *isodic* by Boltzmann (same ισος, path ὸδός) — and *quasi-ergodic* if each surface has dense orbits. The Ehrenfests themselves suggested that the etymology of the word *ergodic* lies in a different direction (work ἔργον, path ὸδός). This work stimulated interest in the mathematical foundations of statistical mechanics, leading eventually to Birkhoff's formulation of the *ergodic hypothesis* and the notion of systems for which almost every orbit in the sense of measure spends a proportion of time in a given set in the phase space in proportion to the measure of the set.

[(2)](Page 2) Questions of this sort were raised by Gel′fand; he considered the vector of first digits of the numbers $(2^n, 3^n, 4^n, 5^n, 6^n, 7^n, 8^n, 9^n)$ and asked if (for example) there

is a value of $n > 1$ for which this vector is $(2, 3, 4, 5, 6, 7, 8, 9)$. This circle of problems is related to the classical Poncelet's porism, as explained in an article by King [194]. The influence of Poncelet's book [292] is discussed by Gray [126, Chap. 27].

(3) (Page 5) See also the account with some simplifications by Goldston, Motohashi, Pintz, and Yıldırım [117] and the survey by Goldston, Pintz and Yıldırım [119].

(4) (Page 8) In a more general form, this is the 11th of Hilbert's famous set of problems formulated for the 1900 International Congress of Mathematics.

(5) (Page 8) Bachet conjectured the result, and Diophantus stated it; there are suggestions that Fermat may have known it. The first published proof is that of Lagrange in 1770; a standard proof may be found in [87, Sect. 2.3.1] for example.

(6) (Page 9) For *indefinite* quadratic forms there is a very successful algebraic technique, namely strong approximation for algebraic groups (an account may be found in the monograph [286] of Platonov and Rapinchuk), so ergodic theory does not enter into the discussion.

(7) (Page 9) Under an additional congruence condition on Q' the method also works for $m \leqslant n - 3$.

(8) For some of the statements made here one actually has to assume that Γ is a *lattice*; see Section 9.4.3.

(9) (Page 9) Further readings from various perspectives on the ergodic theory of homogeneous spaces may be found in the books of Bekka and Mayer [21], Feres [90], Starkov [350], Witte Morris [384], [386] and Zimmer [393].

(10) (Page 10) For linear groups over local fields, and products of such groups, the conjectures of Dani (resp. Raghunathan) have been proved by Margulis and Tomanov [253] and independently by Ratner [306].

(11) (Page 11) Some of the many areas of ergodic theory that we do not treat in a substantial way, and other general sources on ergodic theory, may be found in the following books: the connection with information theory in the work of Billingsley [31] and Shields [342]; a wide-ranging overview of ergodic theory in that of Cornfeld, Fomin and Sinaĭ [60]; ergodic theory developed in the language of joinings in the work of Glasner [116]; more on the theory of entropy and generators in books by Parry [277], [279]; a thorough development of the fundamentals of the measurable theory, including the isomorphism and generator theory, in the book of Rudolph [324].