ERGODIC THEORY: INTERACTIONS WITH COMBINATORICS AND NUMBER THEORY

TOM WARD

Draft May 31, 2007

This article gives a brief overview of some of the ways in which number theory and combinatorics interacts with ergodic theory. The main themes are illustrated by examples related to recurrence, mixing, orbit counting, and Diophantine analysis.

Glossary

Almost everywhere (abbreviated a.e.)

A property that makes sense for each point x in a measure space $(X, \mathcal{B}, \mu)$ is said to hold almost everywhere (or a.e.) if the set $N \subset X$ on which it does not hold satisfies $N \in \mathcal{B}$ and $\mu(N) = 0$.

Čech–Stone compactification of $\mathbb{N}$, $\beta\mathbb{N}$

A compact Hausdorff space that containing $\mathbb{N}$ as a dense subset with the property that any map from $\mathbb{N}$ to a compact Hausdorff space K extends uniquely to a continuous map $\beta\mathbb{N} \to K$. This property and the fact that $\beta\mathbb{N}$ is a compact Hausdorff space containing $\mathbb{N}$ characterizes $\beta\mathbb{N}$ up to homeomorphism.

Curvature

An intrinsic measure of the curvature of a Riemannian manifold depending only on the Riemannian metric; in the case of a surface it determines whether the surface is locally convex (positive curvature), locally saddle-shaped (negative) or locally flat (zero).

Diophantine approximation

The author thanks Richard Sharp for useful comments on the draft version.

Theory of the approximation of real numbers by rational numbers: how small can the distance from a given irrational real number to a rational number be made in terms of the denominator of the rational?

Equidistributed

A sequence is equidistributed if the asymptotic proportion of time it spends in an interval is proportional to the length of the interval.

Ergodic

A measure-preserving transformation is ergodic if the only invariant functions are equal to a constant a.e.; equivalently if the transformation exhibits the convergence in the quasi-ergodic hypothesis.

Ergodic theory

The study of statistical properties of orbits in abstract models of dynamical systems; more generally properties of measure-preserving (semi-)group actions on measure spaces.

Geodesic (flow)

The shortest path between two points on a Riemannian manifold; such a geodesic path is uniquely determined by a starting point and the initial tangent vector to the path (that is, a point in the unit tangent bundle). The transformation on the unit tangent bundle defined by flowing along the geodesic defines the geodesic flow.

Haar measure (on a compact group)

If G is a compact topological group, the unique measure μ defined on the Borel sets of G with the property that $\mu(A+g)=\mu(A)$ for all $g \in G$ and $\mu(G)=1$.

Measure-theoretic entropy

A numerical invariant of measure-preserving systems that reflects the asymptotic growth in complexity of measurable partitions refined under iteration of the map.

Mixing

A measure-preserving system is mixing if measurable sets (events) become asymptotically independent as they are moved apart in time (under iteration).

(Quasi) Ergodic hypothesis

The assumption that, in a dynamical system evolving in time and preserving a natural measure, there are some reasonable conditions under which the 'time average' along orbits of an observable (that is, the average value of a function defined on the phase space) will converge to the 'space average' (that is, the integral of the function with respect to the preserved measure).

Recurrence

Return of an orbit in a dynamical system close to its starting point infinitely often.

S-unit theorems

A circle of results stating that linear equations in fields of zero characteristic have only finitely many solutions taken from finitely-generated multiplicative subgroups of the multiplicative group of the field (apart from infinite families of solutions arising from vanishing sub-sums).

Topological entropy

A numerical invariant of topological dynamical systems that measures the asymptotic growth in the complexity of orbits under iteration. The *variational principle* states that the topological entropy of a topological dynamical system is the supremum over all invariant measures of the measure-theoretic entropies of the dynamical systems viewed as measurable dynamical systems.

I. Definition of the Subject

Number theory is a branch of pure mathematics concerned with the properties of numbers in general, and integers in particular. The areas of most relevance to this article are *Diophantine analysis* (the study of how real numbers may be approximated by rational numbers, and the consequences for solutions of equations in integers); *analytic number theory*, and in particular asymptotic estimates for the number of primes smaller than X as a function of X; *equidistribution*, and questions about how the digits of real numbers are distributed. Combinatorics is concerned with identifying structures in discrete objects; of most interest here is that part of combinatorics connected with Ramsey theory, asserting that large subsets of highly structured objects must automatically contain large replicas of that structure. Ergodic theory is the study of asymptotic behavior of group actions preserving a probability measure; it has proved to be a powerful part of dynamical systems with wide applications.

II. Introduction

Ergodic theory, part of the mathematical study of dynamical systems, has pervasive connections with number theory and combinatorics. This article briefly surveys how these arise through a small sample of results. Unsurprisingly, many details are suppressed, and of course the selection of topics reflects the author's interests far more than it does the full extent of the flow of ideas between ergodic theory and number theory. In addition the selection of topics has been chosen in part to be complementary to those in related articles in the Encyclopedia. A particularly enormous lacuna is the theory of arithmetic dynamical systems itself — the recent monograph by Silverman [118] gives a comprehensive overview.

More sophisticated aspects of this connection – in particular the connections between ergodic theory on homogeneous spaces and Diophantine analysis – are covered in the articles "Ergodic Theory on Homogeneous Spaces and Metric Number Theory" by Kleinbock and "Ergodic Theory: Rigidity" by Niţică; more sophisticated overviews of the connections with combinatorics may be found in the article "Ergodic Theory: Recurrence" by Frantzikinakis and McCutcheon.

III. Ergodic Theory

While the early origins of ergodic theory lie in the quasi-ergodic hypothesis of classical Hamiltonian dynamics, the mathematical study of ergodic theory concerns various properties of group actions on measure spaces, including 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 actions of lattices in Lie groups.
(5) Measure-preserving actions of Lie groups.

The ideas and conditions surrounding the quasi-ergodic hypothesis were eventually placed on a firm mathematical footing by developments starting in 1890. For a single measure-preserving transformation $T : X \to X$ of a probability space $(X, \mathcal{B}, \mu)$, Poincaré [66] showed a *recurrence theorem*: if $E \in \mathcal{B}$ is any measurable set, then for a.e. $x \in E$ there is an infinite set of return times, $0 < n_1 < n_2 < \cdots$ with $T^{n_j}(x) \in E$ (of course Poincaré noted this in a specific setting, concerned with a natural invariant measure for the "three-body" problem in planetary motion).

Poincaré's qualitative result was made quantitative in the 1930s, when von Neumann [90] used the approach of Koopman [45] to show the *mean ergodic theorem*: if $f \in L^2(\mu)$ then there is some $\overline{f} \in L^2(\mu)$ for which

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n - \overline{f} \right\|_2 \longrightarrow 0 \text{ as } N \longrightarrow \infty;$$

clearly $\overline{f}$ then has the property that $\|\overline{f} - \overline{f} \circ T\|_2 = 0$ and $\int \overline{f} \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu$. Around the same time, Birkhoff [9] showed the more delicate *pointwise ergodic theorem*: for any $g \in L^1(\mu)$ there is some $\overline{g} \in L^1(\mu)$ for which

$$\frac{1}{N} \sum_{n=0}^{N-1} g(T^n x) \to \overline{g}(x) \text{ a.e.;}$$

again it is then clear that $\overline{g}(Tx) = \overline{g}(x)$ a.e. and $\int \overline{g} \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu$.

The map T is called *ergodic* if the invariance condition forces the function (f or g) to be equal to a constant a.e. Thus an ergodic map has the property that the *time* or *ergodic average* $(1/N) \sum_{n=0}^{N-1} f \circ T^j$ converges to the *space average* $\int \mathrm{d}\mu$. An overview of ergodic theorems and their many extensions may be found in the article "Ergodic Theorems" by del Junco.

Thus ergodic theory at its most basic level makes strong statements about the asymptotic behavior of orbits of a dynamical system as seen by *observables* (measurable functions on the space X). Applying the ergodic theorem to the indicator function of a measurable set A shows that ergodicity guarantees that a.e. orbit spends an asymptotic proportion of time in A equal to the volume $\mu(A)$ of that set (as measured by the invariant measure). This points to the start of the pervasive connections between ergodic theory and number theory – but as this and other articles relate, the connections extend far beyond this.

IV. Frequency of Returns

In this section we illustrate the way in which a dynamical point of view may unify, explain and extend quite disparate results from number theory.

IV.1. **Normal numbers.** Borel [11] showed (as a consequence of what became the Borel–Cantelli Lemma in probability) that a.e. real number (with respect to Lebesgue measure) is *normal* to every base: that is, has the property that any block of k digits in the base-r expansion appears with asymptotic frequency r^{-k}.

IV.2. **Continued fraction digits.** Analogs of normality results for the continued fraction expansion of real numbers were found by Khinchin, Kuz′min, Lévy and others. Any irrational $x \in [0,1]$ has a unique expansion as a continued fraction

$$x = \cfrac{1}{a_1(x) + \cfrac{1}{a_2(x) + \cfrac{1}{a_3(x) + \cdots}}}$$

and, just as in the case of the familiar base-r expansion, it turns out that the digits $(a_n(x))$ obey precise statistical rules for a.e. x. Gauss conjectured that the appearance of individual digits would obey the law

$$\frac{1}{N}|\{k\colon 1 \leqslant k \leqslant N, a_k(x) = j\}| \longrightarrow \frac{2\log(1+j) - \log j - \log(2+j)}{\log 2}. \qquad (1)$$

This was eventually proved by Kuz′min [46] and Lévy [51], and the probability distribution of the digits is the *Gauss–Kuz′min law*. Khinchin [43] developed this further, showing for example that

$$\lim_{n\to\infty} (a_1(x)a_2(x)\dots a_n(x))^{1/n} = \prod_{n=1}^{\infty} \left(\frac{(n+1)^2}{n(n+2)}\right)^{\log n/\log 2} = 2.68545\cdots \text{ for a.e. } x.$$

Lévy [52] showed that the denominator $q_n(x)$ of the nth convergent $\frac{p_n(x)}{q_n(x)}$ (the rational obtained by truncating the continued fraction expansion of x at the nth term) grows at a specific exponential rate,

$$\lim_{n\to\infty} \frac{1}{n} \log q_n(x) = \frac{\pi^2}{12\log 2} \text{ for a.e. } x.$$

IV.3. **First digits.** The astronomer Newcomb [59] noted that the first digits of large collections of numerical data that are not dimensionless have a specific and non-uniform distribution:

> *"The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable."*

This is now known as Benford's Law, following his popularization and possible rediscovery of the phenomenon [5]. In both cases, this was an empirical observation eventually made rigorous by Hill [35]. Arnold [96, App. 12] pointed out the dynamics behind this phenomena in certain cases, best illustrated by the statistical behavior of the sequence $1, 2, 4, 8, 1, 3, 6, 1, \dots$ of first digits of powers of 2. Empirically, the digit 1 appears about 30% of the time, while the digit 9 appears about 5% of the time.

IV.4. **Equidistribution.** Weyl [93] (and, separately, Bohl [10] and Sierpiński [82]) found an important instance of *equidistribution*. Writing $\{\cdot\}$ for the fractional part, a sequence (a_n) of real numbers is said to be equidistributed modulo 1 if, for any interval $[a,b] \subset [0,1)$,

$$\frac{1}{N}|\{k\colon 1 \leqslant k \leqslant N, \{a_k\} \in [a,b]\}| \to (b-a) \text{ as } N \to \infty;$$

equivalently if

$$\frac{1}{N}\sum_{k=1}^{N} f(a_k) \to \int_0^1 f(t)\,\mathrm{d}t \text{ as } N \to \infty$$

for all continuous functions f. Weyl showed that the sequence $\{n\alpha\}$ is equidistributed if and only if α is irrational. This result was refined and extended in many directions; for example, Hlawka [37] and others found rates for the convergence in terms of the discrepancy of the sequence, Weyl [94] proved equidistribution for $\{n^2\alpha\}$, and Vinogradov for $\{p_n\alpha\}$ where p_n is the nth prime.

IV.5. **The ergodic context.** All the results of this section are manifestations of various kinds of convergence of ergodic averages. Borel's theorem on normal numbers is an immediate consequence of the fact that Lebesgue measure on $[0,1)$ is invariant and ergodic for the map $x \mapsto bx$ modulo 1 with $b \geqslant 2$. The asymptotic properties of continued fraction digits are all a consequence of the fact that the *Gauss measure* defined by

$$\mu(A) = \frac{1}{\log 2} \int_A \frac{\mathrm{d}x}{1+x} \text{ for } A \subset [0,1]$$

is invariant and ergodic for the Gauss map $x \mapsto \left\{\frac{1}{x}\right\}$, and the orbit of an irrational number under the Gauss map determine the digits appearing in the continued fraction expansion much as the orbit under the map $x \mapsto bx \pmod 1$ determines the digits in the base b expansion.

The results on equidistribution and the frequency of first digits are related to ergodic averaging of a different sort. For example, writing $R_\alpha(t) = t + \alpha$ modulo 1 for the circle rotation by α, the first digit of 2^n is the digit j if and only if

$$\log_{10} j \leqslant R_{\log_{10}(2)}(0) < \log_{10}(j+1).$$

Thus the asymptotic frequency of appearance concerns the orbit of a *specific point*. In order to see what this means, consider a continuous map $T : X \to X$ of a compact metric space (X, d). The space $\mathcal{M}(T)$ of Borel probability measures on the Borel σ-algebra of (X, d) is a non-empty compact convex set in the weak*-topology, each extreme point is an ergodic measure for T, and these ergodic measures are mutually singular. If $\mathcal{M}(T)$ is not a singleton and $\mu_1, \mu_2 \in \mathcal{M}(T)$ are distinct ergodic measures, then for a continuous function f with $\int_X f \,\mathrm{d}\mu_1 \neq \int_X f \,\mathrm{d}\mu_2$ it is clear that the ergodic averages $\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x)$ must converge to $\int_X f \,\mathrm{d}\mu_1$ a.e. with respect to μ_1 and to $\int_X f \,\mathrm{d}\mu_2$ a.e. with respect to μ_2. Thus the presence of many invariant measures for a continuous map means that ergodic averages along the orbits of specific points need not converge to the space average with respect to a chosen invariant measure.

In the extreme situation of *unique ergodicity* (a single invariant measure, which is necessarily an extreme point of $\mathcal{M}(T)$ and hence ergodic) the convergence of ergodic averages is much more uniform. Indeed, if T is uniquely ergodic with $\mathcal{M}(T) = \{\mu\}$ then, for any continuous function $f : X \to \mathbb{R}$,

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

(see Oxtoby [61]). The circle rotation R_α is uniquely ergodic for irrational α, leading to the equidistribution results.

The ergodic viewpoint on equidistribution also places equidistribution results in a wider context. Weyl's result that $\{n^2\alpha\}$ (indeed, the fractional part of any polynomial with at least one irrational coefficient) is equidistributed for irrational α was given another proof by Furstenberg [24] using the notion of unique ergodicity.

These methods were then used in the study of nilsystems (translations on quotients of nilpotent Lie groups) by Auslander, Green and Hahn [3] and Parry [62], and these nilsystems play an essential role for polynomial (and other non-conventional) ergodic averaging (see Host and Kra [38] and Leibman [50] in the polynomial case; Host and Kra [39] in the multiple linear case). Remarkably, nilsystems are starting to play a role within combinatorics – an example is the work on the asymptotic number of 4-step arithmetic progressions in the primes by Green and Tao [34]. Pointwise ergodic theorems have also been found along sequences other than $\mathbb{N}$; notably for integer-valued polynomials and along the primes for L^2 functions by Bourgain [13], [12]. For more details, see the survey paper of del Junco [102] on ergodic theorems.

V. Ergodic Ramsey Theory and Recurrence

In 1927 van der Waerden proved a conjecture attributed to Baudet: if the natural numbers are written as a disjoint union of finitely many sets,

$$\mathbb{N} = C_1 \sqcup C_2 \sqcup \cdots \sqcup C_r, \tag{2}$$

then there must be one set C_j that contains arbitrarily long arithmetic progressions. That is, there is some $j \in \{1, \dots, r\}$ such that for any $k \geqslant 1$ there are $a \geqslant 1$ and $n \geqslant 1$ with

$$a, a+n, a+2n, \dots, a+(k-1)n \in C_j.$$

The original proof appears in van der Waerden's paper [89], and there is a discussion of how he found the proof in [123].

Work of Furstenberg and Weiss [28] and others placed the theorem of van der Waerden in the context of topological dynamics, giving alternative proofs. Specifically, van der Waerden's theorem is a consequence of *topological multiple recurrence*: the return of points under iteration in a topological dynamical system close to their starting point along finite sequences of times. The same approach readily gives dynamical proofs of Rado's extension [69] of van der Waerden's theorem, and of Hindman's theorem [36]. The theorems of Rado and Hindman introduce a new theme: given a set $A = \{n_1, n_2, \dots\}$ of natural numbers, write $FS(A)$ for the set of numbers obtained as finite sums $n_{i_1} + \cdots + n_{i_j}$ with $i_1 < i_2 < \cdots < i_j$. Rado showed that for any large n there is some C_s containing some $FS(A)$ for a set A of cardinality n. Hindman showed that there is some C_s containing some $FS(A)$ for an *infinite* set A.

In the theorem of van der Waerden, it is clear that for any reasonable notion of "proportion" or "density" one of the sets C_j must occupy a positive proportion of $\mathbb{N}$. A set $A \subset \mathbb{N}$ is said to have *positive upper density* if there are sequences (M_i) and (N_i) with $N_i - M_i \to \infty$ as $i \to \infty$ such that

$$\lim_{i\to\infty} \frac{1}{N_i - M_i} \left|\{a \in A \colon M_i < a < N_i\}\right| > 0.$$

Erdős and Turán [21] conjectured the stronger statement that any subset of $\mathbb{N}$ with positive upper density must contain arbitrary long arithmetic progressions. This statement was shown for arithmetic progressions of length 3 by Roth [72] in 1952, then for length 4 by Szemerédi [86] in 1969. The general result was eventually proved by Szemerédi [87] in 1975 in a lengthy and extremely difficult argument.

Furstenberg saw that Szemerédi's Theorem would follow from a deep extension of the Poincaré recurrence phenomena described in Section III and proved that extension [25] (see also the survey article by Furstenberg, Katznelson and Ornstein [106]). The *multiple recurrence* result of Furstenberg says that for any measure-preserving system $(X, \mathcal{B}, \mu, T)$ and set $A \in \mathcal{B}$ with $\mu(A) > 0$, and for any $k \in \mathbb{N}$,

$$\liminf_{N-M\to\infty} \frac{1}{N-M+1} \sum_{n=M}^{N} \mu\left(A \cap T^{-n}A \cap T^{-2n}A \cap \cdots \cap T^{-kn}A\right) > 0.$$

An immediate consequence is that in the same setting there must be some $n \geqslant 1$ for which

$$\mu\left(A \cap T^{-n}A \cap T^{-2n}A \cap \cdots \cap T^{-kn}A\right) > 0. \tag{3}$$

A general *correspondence principle*, due to Furstenberg, shows that statements in combinatorics like Szemerédi's Theorem are equivalent to statements in ergodic theory like (3).

This opened up a significant new field of *ergodic Ramsey theory*, in which methods from dynamical systems and ergodic theory are used to produce new results in infinite combinatorics. For an overview, see the articles "Ergodic Theory on Homogeneous Spaces and Metric Number Theory" by Kleinbock, "Ergodic Theory: Rigidity" by Niţică, "Ergodic Theory: Recurrence" by Frantzikinakis and McCutcheon and the survey articles of Bergelson [97], [98], [99]. The field is too large to give an overview here, but a few examples will give a flavor of some of the themes.

Call a set $R \subset \mathbb{Z}$ a *set of recurrence* if, for any finite measure-preserving invertible transformation T of a finite measure space $(X, \mathcal{B}, \mu)$ and any set $A \in \mathcal{B}$ with $\mu(A) > 0$, there are infinitely many $n \in R$ for which $\mu(A \cap T^{-n}A) > 0$. Thus Poincaré recurrence is the statement that $\mathbb{N}$ is a set of recurrence. Furstenberg and Katznelson [26] showed that if $T_1, \cdots, T_k$ form a family of commuting measure-preserving transformations and A is a set of positive measure, then

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(T_1^{-n}A \cap \cdots \cap T_k^{-n}A) > 0.$$

This remarkable multiple recurrence implies a multi-dimensional form of Szemerédi's theorem. Recently, Gowers has found a non-ergodic proof of this [32].

Furstenberg also gave an ergodic proof of Sárközy's theorem [73]: if $p \in \mathbb{Q}[t]$ is a polynomial with $p(\mathbb{Z}) \subset \mathbb{Z}$ and $p(0) = 0$, then $\{p(n)\}_{n>0}$ is a set of recurrence. This was extended to multiple polynomial recurrence by Bergelson and Leibman [7].

V.1. **Topology and coloring theorems.** The existence of idempotent ultrafilters in the Čech–Stone compactification $\beta\mathbb{N}$ gives rise to an algebraic approach to many questions in topological dynamics (this notion has its origins in the work of Ellis [104]). Using these methods, results like Hindman's finite sums theorem find elegant proofs, and many new results in combinatorics have been found. For example, in the partition (2) there must be one set C_j containing a triple x, y, z solving $x - y = z^2$.

A deeper application is to improve a strengthening of Kronecker's theorem. To explain this, recall that a set S is called *IP* if there is a sequence (n_i) of natural numbers (which do not need to be distinct) with the property that S contains all the terms of the sequence and all finite sums of terms of the sequence with distinct

indices. A set S is called IP^* if it has non-empty intersection with *every* IP set, and a set S is called IP^*_+ if there is some $t \in \mathbb{Z}$ for which $S - t$ is IP^*. Thus being IP^* (or IP^*_+) is an extreme form of 'fatness' for a set. Now let $1, \alpha_1, \ldots, \alpha_k$ be numbers that are linearly independent over the rationals, and for any $d \in \mathbb{N}$ and kd non-empty intervals $I_{ij} \subset [0,1]$ $(1 \leqslant i \leqslant d, 1 \leqslant j \leqslant k)$, let

$$D = \{n \in \mathbb{N} \colon \{n^i \alpha_j\} \in I_{ij} \text{ for all } i, j\}.$$

Kronecker showed that if $d = 1$ then D is non-empty; Hardy and Littlewood showed that D is infinite, and Weyl showed that D has positive density. Bergelson [6] uses these algebraic methods to improve the result by showing that D is an IP^*_+ set.

V.2. **Polynomialization and *IP*-sets.** As mentioned above, Bergelson and Leibman [7] extended multiple recurrence to a polynomial setting. For example, let

$$\{p_{i,j} \colon 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant t\}$$

be a collection of polynomials with rational coefficients and $p_{i,j}(\mathbb{Z}) \subset \mathbb{Z}$, $p_{i,j}(0) = 0$. Then if $\mu(A) > 0$, we have

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mu\Big(\bigcap_{i=1}^{k} \Big(\prod_{j=1}^{t} T_j^{p_{i,j}(n)} \Big)^{-1} A \Big) > 0.$$

Using the Furstenberg correspondence principle, this gives a multi-dimensional polynomial Szemerédi theorem: If $P \colon \mathbb{Z}^r \to \mathbb{Z}^\ell$ is a polynomial mapping with the property that $P(0) = 0$, and $F \subset \mathbb{Z}^r$ is a finite configuration, then any set $S \subset \mathbb{Z}^\ell$ of positive upper Banach density contains a set of the form $u + P(nF)$ for some $u \in \mathbb{Z}^\ell$ and $n \in \mathbb{N}$.

In a different direction, motivated in part by Hindman's theorem, the multiple recurrence results generalize to *IP*-sets. Furstenberg and Katznelson [27] proved a linear *IP*-multiple recurrence theorem in which the recurrence is guaranteed to occur along an *IP*-set. A combinatorial proof of this result has been found by Nagle, Rödl and Schacht [58]. Bergelson and McCutcheon [8] extended these results by proving a polynomial *IP*-multiple recurrence theorem. To formulate this, make the following definitions. Write $\mathcal{F}$ for the family of non-empty finite subsets of $\mathbb{N}$, so that a sequence indexed by $\mathcal{F}$ is an *IP*-set. More generally, an $\mathcal{F}$-sequence $(n_\alpha)_{\alpha\in\mathcal{F}}$ taking values in an abelian group is called an *IP*-sequence if $n_{\alpha\cup\beta} = n_\alpha + n_\beta$ whenever $\alpha \cap \beta = \varnothing$. An *IP*-ring is a set of the form $\mathcal{F}^{(1)} = \{\bigcup_{i\in\beta} \alpha_i \colon \beta \in \mathcal{F}\}$ where $\alpha_1 < \alpha_2 < \cdots$ is a sequence in $\mathcal{F}$, and $\alpha < \beta$ means $a < b$ for all $a \in \alpha$, $b \in \beta$; write $\mathcal{F}^m_<$ for the set of m-tuples $(\alpha_1, \cdots, \alpha_m)$ from $\mathcal{F}^m$ with $\alpha_i < \alpha_j$ for $i < j$. Write $PE(m,d)$ for the collection of all expressions of the form $T(\alpha_1, \cdots, \alpha_m) = \prod_{i=1}^{r} T_i^{p_i((n^{(b)}_{\alpha_j})_{1\leqslant b\leqslant k,\, 1\leqslant j\leqslant m})}$, $(\alpha_1, \cdots, \alpha_m) \in (F \cup \varnothing)^m_<$, where each p_i is a polynomial in a $k \times m$ matrix of variables with integer coefficients and zero constant term with degree $\leqslant d$. Then for every $m, t \in \mathbb{N}$, there is an *IP*-ring $\mathcal{F}^{(1)}$, and an $a = a(A, m, t, d) > 0$, such that, for every set of polynomial expressions $\{S_0, \cdots, S_t\} \subset PE(m,d)$,

$$\underset{(\alpha_1,\cdots,\alpha_m)\in(\mathcal{F}^{(1)})^m_<}{\to} IP - \lim \mu\Big(\bigcap_{i=0}^{t} S_i(\alpha_1, \cdots, \alpha_m)^{-1} A \Big) > 0.$$

There are a large number of deep combinatorial consequences of this result, not all of which seem accessible by other means.

V.3. **Sets of Primes.** In a remarkable development, Szemerédi's theorem and some of the ideas behind ergodic Ramsey theory joined results of Goldston and Yıldırım [31] in playing a part in Green and Tao's proof [33] that the set of primes contains arbitrarily long arithmetic progressions. This profound result is surveyed from an ergodic point of view in the article of Kra [111]. As with Szemerédi's theorem itself, this result has been extended to a polynomial setting by Tao and Ziegler [88]. Given integer-valued polynomials $f_1, \ldots, f_k \in \mathbb{Z}[t]$ with

$$f_1(0) = \cdots = f_k(0) = 0$$

and any $\varepsilon > 0$, Tao and Ziegler proved that there are infinitely many integers x, m with $1 \leqslant m \leqslant x^{\varepsilon}$ for which $x + f_1(m), \ldots, x + f_k(m)$ are primes.

VI. Orbit-counting as an Analogous Development

Some of the connections between number theory and ergodic theory arise through developments that are analogous but not directly related. A remarkable instance of this concerns the long history of attempts to count prime numbers laid alongside the problem of counting closed orbits in dynamical systems.

VI.1. **Counting orbits and geodesics.** Consider first the fundamental arithmetic function $\pi(X) = |\{p \leqslant X : p \text{ is prime}\}|$. Tables of primes prepared by Felkel and Vega in the 18th century led Legendre to suggest that $\pi(X)$ is approximately $x/(\log(X) - 1.08)$. Gauss, using both computational evidence and a heuristic argument, suggested that $\pi(X)$ is approximated by

$$\mathrm{li}(X) = \int_2^X \frac{\mathrm{d}t}{\log t}.$$

Both of these suggestions imply the well-known asymptotic formula

$$\pi(X) \sim \frac{X}{\log X}. \tag{4}$$

Riemann brought the analytic ideas of Dirichlet and Chebyshev (who used the zeta function to find a weaker version of (4) with upper and lower bounds for the quantity $\frac{\pi(X)\log(X)}{X}$) to bear by proposing that the zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \left(1 - p^{-s}\right)^{-1}, \tag{5}$$

already studied by Euler, would connect properties of the primes to analytic methods. An essential step in these developments, due to Riemann, is the meromorphic extension of ζ from the region $\Re(s) > 1$ in (5) to the whole complex plane and a functional equation relating the value of the extension at s to the value at $1 - s$. Moreover, Riemann showed that the extension has readily understood real zeros, and that all the other zeros he could find were symmetric about $\Re(s) = \frac{1}{2}$. The *Riemann hypothesis* asserts that zeros in the region $0 < \Re(s) < 1$ all lie on the line $\Re(s) = \frac{1}{2}$, and this remains open.

Analytic properties of the Riemann zeta function were used by Hadamard and de la Vallée Poussin to prove (4), the *Prime Number Theorem*, in 1896. Tauberian methods developed by Wiener and Ikehara [95] later gave different approaches to the Prime Number Theorem.

These ideas initiated the widespread use of zeta functions in several parts of mathematics, but it was not until the middle of the 20th century that Selberg [79] introduced a zeta function dealing directly with quantities arising in dynamical systems: the lengths of closed geodesics on surfaces of constant curvature -1. The geodesic flow acts on the unit tangent bundle to the surface by moving a point and unit tangent vector at that point along the unique geodesic they define at unit speed. Closed geodesics are then in one-to-one correspondence with periodic orbits of the associated geodesic flow on the unit tangent bundle, and it is in this sense that the quantities are dynamical. The function defined by Selberg takes the form

$$Z(s) = \prod_{\tau} \prod_{k=0}^{\infty} \left(1 - e^{-(s+k)|\tau|}\right),$$

in which τ runs over all the closed geodesics, and $|\tau|$ denotes the length of the geodesic. In a direct echo of the Riemann zeta function, Selberg found an analytic continuation to the complex plane, and showed that the zeros of Z lie on the real axis or on the line $\Re(s) = \frac{1}{2}$ (the analogue of the Riemann hypothesis for Z; see also the paper of Hejhal [108]). The zeros of Z are closely connected to the eigenvalues for the Laplace–Beltrami operator, and thus give information about the lengths of closed geodesics via Selberg's trace formula in the same paper. Huber [40] and others used this approach to give an analogue of the prime number theorem for closed geodesics – a *prime orbit theorem*.

Sinai [83] considered closed geodesics on a manifold M with negative curvature bounded between $-R^2$ and $-r^2$, and found the bounds

$$(\dim(M) - 1)r \leqslant \liminf_{T\to\infty} \frac{\log \pi(T)}{T} \leqslant \limsup_{T\to\infty} \frac{\log \pi(T)}{T} \leqslant (\dim(M) - 1)R$$

for the number $\pi(T)$ of closed geodesics of multiplicity one with length less than T, analogous to Chebyshev's result.

The essential dynamical feature behind the geodesic flow on a manifold of negative curvature is that it is an example of an *Anosov flow* [2]. These are smooth dynamical $\mathbb{R}$-actions (equivalently, first-order differential equations on Riemannian manifolds) with the property that the tangent bundle has a continuously varying splitting into a direct sum $E^u \oplus E^s \oplus E^o$ and the action of the differential of the flow acts on E^u as an exponential expansion, on E^s as an exponential contraction, E^o is the one-dimensional bundle of vectors that are tangent to orbits, and the expansion and contraction factors are bounded. In the setting of Anosov flows, the natural orbit counting function is $\pi(X) = |\{\tau \colon \tau \text{ a closed orbit of length } |\tau| \leqslant X\}|$. Margulis [54], [55] generalized the picture to weak-mixing Anosov flows by showing a prime orbit theorem of the form

$$\pi(X) \sim \frac{e^{h_{\text{top}} X}}{h_{\text{top}} X} \tag{6}$$

for the counting function $\pi(X) = |\{\tau \colon \tau \text{ a closed orbit of length } |\tau| \leqslant X\}|$ where as before h_{top} denotes the topological entropy of the flow. Integral to Margulis' work is a result on the spatial distribution of the closed geodesics reflected in a flow-invariant probability measure, now called the Margulis measure.

Anosov also studied discrete dynamical systems with similar properties: diffeomorphisms of compact manifolds with a similar splitting of the tangent space (though in this setting E^o disappears). The archetypal Anosov diffeomorphism is

a hyperbolic toral automorphism of the sort considered in Section VII.1; for such automorphisms of the 2-torus Adler and Weiss [1] constructed Markov partitions, allowing the dynamics of the toral automorphism to be modeled by a topological Markov shift, and used this to determine when two such automorphisms are measurably isomorphic. Sinai [84], Ratner [70], Bowen [14], [16] and others developed the construction of Markov partitions in general for Anosov diffeomorphisms and flows.

Around the same time, Smale [85] introduced a more permissive hyperbolicity axiom for diffeomorphisms, Axiom A. Maps satisfying Axiom A are diffeomorphisms satisfying the same hypothesis as that of Anosov diffeomorphisms, but only on the set of points that return arbitrarily close under the action of the flow (or iteration of the map).

Thus Markov partitions, and with them associated transfer operators became a substitute for the geometrical Laplace–Beltrami operators of the setting considered by Selberg. Bowen [15] extended the uniform distribution result of Margulis to this setting and found an analogue of Chebychev's theorem for closed orbits. Parry [63] (in a restricted case) and Parry and Pollicott [65] went on to prove the prime orbit theorem in this more general setting. The methods are an adaptation of the Ikehara–Wiener Tauberian approach to the prime number theorem.

Thus many facets of the prime number theorem story find their echoes in the study of closed orbits for hyperbolic flows: the role played by meromorphic extensions of suitable zeta functions, Tauberian methods, and so on. Moreover, related results from number theory have analogues in dynamics, for example Mertens' theorem [57] in the work of Sharp [80] and Noorani [60] and Dirichlet's theorem in work of Parry [64].

The "elementary" proof (not using analytic methods) of the prime number theorem by Erdös [20] and Selberg [78] (see the survey by Goldfeld [30] for the background to the results and the unfortunate priority dispute) has an echo in some approaches to orbit-counting problems from an elementary (non-Tauberian) perspective, including work of Lalley [48] on special flows and Everest, Miles, Stevens and the author [22] in the algebraic setting.

In a different direction Lalley [47] found orbit asymptotics for closed orbits satisfying constraints in the Axiom A setting without using Tauberian theorems. His more direct approach is still analytic, using complex transfer operators (the same objects used to by Parry and Pollicott to study the dynamical zeta function at complex values) and indeed somewhat parallels a Tauberian argument.

Further resonances with number theory arise here. For example, there are results on the distribution of closed orbits for group extensions (analogous to Chebotarev's theorem) and for orbits with homological constraints (see Sharp [81], Katsuda and Sunada [42]).

Of course the great diversity of dynamical systems subsumed in the phrase "prime orbit theorem" creates new problems and challenges, and in particular if there is not much geometry to work with then the reliance on Markov partitions and transfer operators makes it difficult to find higher-order asymptotics.

Dolgopyat [18] has nonetheless managed to push the Markov methods to obtain uniform bounds on iterates of the associated transfer operators to the region $\Re(s) > \sigma_0$ with $\sigma_0 < 1$. This result has wide implications; an example most relevant to the analogy with number theory is the work of Pollicott and Sharp [67] in which

Dolgopyat's result is used to show that for certain geodesic flows there is a two-term prime orbit theorem of the form

$$\pi(X) = \operatorname{li}\left(e^{h_{\text{top}}X}\right) + \mathrm{O}\left(e^{cX}\right)$$

for some $c < h_{\text{top}}$.

For non-positive curvature manifolds less is known: Knieper [44] finds upper and lower bounds for the function counting closed geodesics on rank-1 manifolds of non-positive curvature of the form

$$A\frac{e^{hX}}{X} \leqslant \pi(X) \leqslant Be^{hX}$$

for constants $A, B > 0$.

VI.2. **Counting orbits for group endomorphisms.** A prism through which to view some of the deeper issues that arise in Section VI.1 is provided by group endomorphisms. The price paid for having simple closed formulas for all the quantities involved is of course a severe loss of generality, but the diversity of examples illustrates many of the phenomena that may be expected in more general settings when hyperbolicity is lost.

Consider an endomorphism $T : X \to X$ of a compact group with the property that $\mathrm{F}_n(T) < \infty$ for all $n \geqslant 1$. The number of closed orbits of length n under T is then

$$\mathrm{O}_n(T) = \frac{1}{n}\sum_{d|n} \mu(n/d)\,\mathrm{F}_d(T). \tag{7}$$

In simple situations (hyperbolic toral automorphisms for example) it is straightforward to show that

$$\pi_T(X) = |\{\tau\colon \tau \text{ a closed orbit under } T \text{ of length} \leqslant X\}| \sim \frac{e^{(X+1)h_{\text{top}}(T)}}{X}. \tag{8}$$

Waddington [91] considered quasihyperbolic toral automorphisms, showing that the asymptotic (8) in this case is multiplied by an explicit almost-periodic function bounded away from zero and infinity.

This result has been extended further into non-hyperbolic territory, which is most easily seen via the so-called connected S-integer dynamical systems introduced by Chothi, Everest and the author [17]. Fix an algebraic number field $\mathbb{K}$ with set of places $P(\mathbb{K})$ and set of infinite places $P_\infty(\mathbb{K})$, an element of infinite multiplicative order $\xi \in \mathbb{K}^*$, and a finite set $S \subset P(\mathbb{K}) \setminus P_\infty(\mathbb{K})$ with the property that $|\xi|_w \leqslant 1$ for all $w \notin S \cup P_\infty(\mathbb{K})$. The associated ring of S-integers is

$$R_S = \{x \in \mathbb{K}\colon |x|_w \leqslant 1 \text{ for all } w \notin S \cup P_\infty(\mathbb{K})\}.$$

Let X be the compact character group of R_S, and define the endomorphism $T : X \to X$ to be the dual of the map $x \mapsto \xi x$ on R_S. Following Weil [125], write $\mathbb{K}_w$ for the completion at w, and for w finite, write r_w for the maximal compact subring of $\mathbb{K}_w$. Notice that if $S = P$ then $R_S = \mathbb{K}$ and $\mathrm{F}_n(T) = 1$ for all $n \geqslant 1$ by the product formula for $\mathbb{A}$-fields. As the set S shrinks, more and more periodic orbits come into being, and if S is as small as possible (given ξ) then the resulting system is (more or less) hyperbolic or quasi-hyperbolic.

For S finite, it turns out that there are still sufficiently many periodic orbits to have the growth rate result (10), but the asymptotic (8) is modified in much the

same way as Waddington observed for quasi-hyperbolic toral automorphisms:

$$\liminf_{X\to\infty} \frac{X\pi_T(X)}{e^{(X+1)h_{\mathrm{top}}(T)}} > 0 \tag{9}$$

and there is an associated pair (X^*, a_T), where X^* is a compact group and $a_T \in X^*$, with the property that if $a_T^{N_j}$ converges in X^* as $j \to \infty$, then there is convergence in (9).

A simple special case will illustrate this. Taking $\mathbb{K} = \mathbb{Q}$, $\xi = 2$ and $S = \{3\}$ gives a compact group endomorphism T with

$$\mathrm{F}_n(T) = (2^n - 1)|2^n - 1|_3.$$

For this example the results of [17] are sharper: The expression in (9) converges along (X_j) if and only if 2^{X_j} converges in the ring of 3-adic integers $\mathbb{Z}_3$, the expression has uncountably many limit points, and the upper and lower limits are transcendental.

Similarly, the dynamical analogue of Mertens' theorem found by Sharp may be found for S-integer systems with S finite. Writing

$$\mathcal{M}_T(N) = \sum_{|\tau|\leqslant N} \frac{1}{e^{h(T)|\tau|}},$$

it is shown in [17] that for an ergodic S-integer map T with $\mathbb{K} = \mathbb{Q}$ and S finite, there are constants $k_T \in \mathbb{Q}$ and C_T such that

$$\mathcal{M}_T(N) = k_T \log N + C_T + \mathrm{O}\left(1/N\right).$$

Without the restriction that $\mathbb{K} = \mathbb{Q}$, it is shown that there are constants $k_T \in \mathbb{Q}$, C_T and $\delta > 0$ with

$$\mathcal{M}_T(N) = k_T \log N + C_T + \mathrm{O}(N^{-\delta}).$$

VII. Diophantine Analysis as a Toolbox

Many problems in ergodic theory and dynamical system exploit ideas and results from number theory in a direct way; we illustrate this by describing a selection of dynamical problems that call on particular parts of number theory in an essential way. The example of mixing in Section VII.2 is particularly striking for two reasons: the results needed from number theory are relatively recent, and the ergodic application directly motivated a further development in number theory.

VII.1. **Orbit growth and convergence.** The analysis of periodic orbits – how their number grows as the length grows and how they spread out through space – is of central importance in dynamics (see Katok [41] for example). An instance of this is that for many simple kinds of dynamical systems $T : X \to X$ (where T is a continuous map of a compact metric space (X, d)) the logarithmic growth rate of the number of periodic points exists and coincides with the topological entropy $h(T)$ (an invariant giving a quantitative measure of the average rate of growth in orbit complexity under T). That is, writing

$$\mathrm{F}_n(T) = |\{x \in X : T^n x = x\}|,$$

we find

$$\frac{1}{n} \log \mathrm{F}_n(T) \longrightarrow h_{\mathrm{top}}(T) \tag{10}$$

for many of the simplest dynamical systems. For example, if $X = \mathbb{T}^r$ is the r-torus and $T = T_A$ is the automorphism of the torus corresponding to a matrix A in $GL_r(\mathbb{Z})$, then T_A is ergodic with respect to Lebesgue measure if and only if no eigenvalue of A is a root of unity. Under this assumption, we have

$$\mathrm{F}_n(T_A) = \prod_{i=1}^{r} |\lambda_i^n - 1|$$

and

$$h_{\mathrm{top}}(T_A) = \sum_{i=1}^{r} \log \max\{1, |\lambda_i|\} \tag{11}$$

where $\lambda_1, \ldots, \lambda_r$ are the eigenvalues of A. It follows that the convergence in (10) is clear under the assumption that T_A is *hyperbolic* (that is, no eigenvalue has modulus one). Without this assumption the convergence is less clear: for $r \geqslant 4$ the automorphism T_A may be ergodic without being hyperbolic. That is, while no eigenvalues are unit roots some may have unit modulus. As pointed out by Lind [53] in his study of these *quasihyperbolic* automorphisms, the convergence (10) does still hold for these systems, but this requires a significant Diophantine result (the theorem of Gel′fond [29] suffices; one may also use Baker's theorem [4]). Even further from hyperbolicity lie the family of S-integer systems [17], [92]; their orbit-growth properties are intimately tied up with Artin's conjecture on primitive roots and prime divisors of linear recurrence sequences.

VII.2. **Mixing and additive relations in fields.** The problem of higher-order mixing for commuting group automorphisms provides a striking example of the dialogue between ergodic theory and number theory, in which deep results from number theory have been used to solve problems in ergodic theory, and questions arising in ergodic theory have motivated further developments in number theory.

An action T of a countable group Γ on a probability space $(X, \mathcal{B}, \mu)$ is called *k-fold mixing* or *mixing on $(k+1)$ sets* if

$$\mu\left(A_0 \cap T^{-g_1} A_1 \cap \cdots \cap T^{-g_k} A_k\right) \longrightarrow \mu(A_0) \cdots \mu(A_k) \tag{12}$$

as

$$g_i g_j^{-1} \longrightarrow \infty \text{ for } i \neq j$$

with the convention that $g_0 = 1_\Gamma$, for any sets $A_0, \ldots, A_k \in \mathcal{B}$; $g_n \to \infty$ in Γ means that for any finite set $F \subset \Gamma$ there is an N with $n > N \implies g_n \notin F$. For $k = 1$ the property is called simply *mixing*. This notion for single transformations goes back to the foundational work of Rohlin [71], where he showed that ergodic group endomorphisms are mixing of all orders (and so the notion is not useful for distinguishing between group endomorphisms as measurable dynamical systems). He raised the (still open) question of whether any measure-preserving transformation can be mixing without being mixing of all orders.

A class of group actions that are particularly easy to understand are the *algebraic dynamical systems* studied systematically by Schmidt [115]: here X is a compact abelian group, each T^g is a continuous automorphism of X, and μ is the Haar measure on X. Schmidt [75] related mixing properties of algebraic dynamical systems with $\Gamma = \mathbb{Z}^d$ to statements in arithmetic, and showed that a mixing action on a connected group could only fail to mix in a certain way. Later Schmidt and the author [76] showed that for X connected, mixing implies mixing of all orders.

The proof proceeds by showing that the result is exactly equivalent to the following statement: if $\mathbb{K}$ is a field of characteristic zero, and G is a finitely generated subgroup of the multiplicative group $\mathbb{K}^\times$, then the equation

$$a_1x_1 + \cdots + a_nx_n = 1 \tag{13}$$

for fixed $a_1, \ldots, a_n \in \mathbb{K}^\times$ has a finite number of solutions $x_1, \ldots, x_n \in G$ for which no subsum $\sum_{i\in I} a_ix_i$ with $I \subsetneq \{1, \ldots, n\}$ vanishes. The bound on the number of solutions to (13) follows from the profound extensions to W. Schmidt's subspace theorem in Diophantine geometry [77] by Evertse and Schlickewei (see [23], [68], [74] for the details).

The argument in [76] may be cast as follows: failure of k-fold mixing in a connected algebraic dynamical system implies (via duality) an infinite set of solutions to an equation of the shape (13) in some field of characteristic zero. The S-unit theorem means that this can only happen if there is some proper subsum that vanishes infinitely often. This infinite family of solutions to a homogeneous form of (13) with fewer terms can then be translated back via duality to show that the system fails to mix for some strictly lower order, proving that mixing implies mixing of all orders by induction.

Mixing properties for algebraic dynamical systems without the assumption of connectedness are quite different, and in particular it is possible to have mixing actions that are not mixing of all orders. This is a simple consequence of the fact that the constituents of a disconnected algebraic dynamical system are associated with fields of positive characteristic, where the presence of the Frobenius automorphism can prevent higher-order mixing. Ledrappier [49] pointed this out via examples of the following shape. Let

$$X = \left\{ x \in \mathbb{F}_2^{\mathbb{Z}^2} : x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod 2 \right\}$$

and define the $\mathbb{Z}^2$-action T to be the natural shift action,

$$(T^{(n,m)}x)_{(a,b)} = x_{(a+n,b+m)}.$$

It is readily seen that this action is mixing with respect to the Haar measure. The condition $x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod 2$ implies that, for any $k \geqslant 1$,

$$x_{(0,2^k)} = \sum_{j=0}^{2^k} \binom{2^k}{j} x_{(j,0)} = x_{(0,0)} + x_{(2^k,0)} \pmod 2 \tag{14}$$

since every entry in the 2^kth row of Pascal's triangle is even apart from the first and the last. Now let $A = \{x \in X : x_{(0,0)} = 0\}$ and let $x_* \in X$ be any element with $x_{(0,0)} = 1$. Then X is the disjoint union of A and $A + x_*$, so

$$\mu(A) = \mu(A + x_*) = \tfrac{1}{2}.$$

However, (14) shows that

$$x \in A \cap T_{-(2^k,0)}A \implies x \in T_{-(0,2^k)}A,$$

so

$$A \cap T_{-(2^k,0)}A \cap T_{-(0,2^k)}(A + x_*) = \varnothing$$

for all $k \geqslant 1$, which shows that T cannot be mixing on three sets.

The full picture of higher-order mixing properties on disconnected groups is rather involved; see Schmidt's monograph [115]. A simple illustration is the construction by Einsiedler and the author [19] of systems with any prescribed order of mixing. When such systems fail to be mixing of all orders, they fail in a very specific way – along dilates of a specific *shape* (a finite subset of $\mathbb{Z}^d$). In the example above, the shape that fails to mix is $\{(0,0),(1,0),(0,1)\}$. This gives an order of mixing as detected by shapes; computing this is in principle an algebraic problem. On the other hand, there is a more natural definition of the order of mixing, namely the largest k for which (12) holds; computing this is in principle a Diophantine problem. A conjecture emerged (formulated explicitly by Schmidt [116]) that for any algebraic dynamical system, if every set of cardinality $r \geqslant 2$ is a mixing shape, then the system is mixing on r sets.

This question motivated Masser [56] to prove an appropriate analogue of the S-unit theorem on the number of solutions to (13) in positive characteristic as follows. Let H be a multiplicative group and fix $n \in \mathbb{N}$. An infinite subset $A \subset H^n$ is called *broad* if it has both of the following properties:

- if $h \in H$ and $1 \leqslant j \leqslant n$, then there are at most finitely many $(a_1, \dots, a_n)$ in A with $a_j = h$;
- if $n \geqslant 2$, $h \in H$ and $1 \leqslant i < j \leqslant n$ then there are at most finitely many $(a_1, \dots, a_n) \in H$ with $a_i a_j^{-1} = h$.

Then Masser's theorem says the following. Let $\mathbb{K}$ be a field of characteristic $p > 0$, let G be a finitely-generated subgroup of $\mathbb{K}^\times$ and suppose that the equation

$$a_1 x_1 + \cdots + a_n x_n = 1$$

has a broad set of solutions $(x_1, \dots, x_n) \in G^n$ for some constants $a_1, \dots, a_n \in \mathbb{K}^\times$. Then there is an $m \leqslant n$, constants $b_1, \dots, b_m \in \mathbb{K}^\times$ and some $(g_1, \dots, g_m) \in G^m$ with the following properties:

- $g_j \neq 1$ for $1 \leqslant j \leqslant m$;
- $g_i g_j^{-1} \neq 1$ for $1 \leqslant i < j \leqslant m$;
- there are infinitely many k for which

$$b_1 g_1^k + b_2 g_2^k + \cdots + b_m g_m^k = 1.$$

The proof that shapes detect the order of mixing in algebraic dynamics then proceeds much as in the connected case.

VIII. Future Directions

The interaction between ergodic theory, number theory and combinatorics continues to expand rapidly, and many future directions of research are discussed in the articles "Ergodic Theory on Homogeneous Spaces and Metric Number Theory" by Kleinbock, "Ergodic Theory: Rigidity" by Niţică and "Ergodic Theory: Recurrence" by Frantzikinakis and McCutcheon. Some of the directions most relevant to the examples discussed in this article include the following.

The recent developments mentioned in Section V.3 clearly open many exciting prospects involving finding new structures in arithmetically significant sets (like the primes). The original conjecture of Erdős and Turán [21] asked if $\sum_{a \in A \subset \mathbb{N}} \frac{1}{a} = \infty$ is sufficient to force the set A to contain arbitrary long arithmetic progressions, and remains open. This would of course imply both Szemerédi's theorem [87] and the result of Green and Tao [33] on arithmetic progressions in the primes. More

generally, it is clear that there is still much to come from the dialogue subsuming the four parallel proofs of Szemerédis: one by purely combinatorial methods, one by ergodic theory, one by hypergraph theory, and one by Fourier analysis and additive combinatorics. For an overview, see the survey papers of Tao [119], [120], [121].

In the context of the orbit-counting results in Section VI, a natural problem is to on the one hand obtain finer asymptotics with better control of the error terms, and on the other to extend the situations that can be handled. In particular, relaxing the hypotheses related to hyperbolicity (or negative curvature) is a constant challenge.

Primary Literature

1. R. L. Adler and B. Weiss. *Similarity of automorphisms of the torus*. Memoirs of the American Mathematical Society, No. 98. American Mathematical Society, Providence, R.I., 1970.
2. D. V. Anosov. Geodesic flows on closed Riemannian manifolds of negative curvature. *Trudy Mat. Inst. Steklov.*, 90:209, 1967.
3. L. Auslander, L. Green, and F. Hahn. *Flows on homogeneous spaces*. With the assistance of L. Markus and W. Massey, and an appendix by L. Greenberg. Annals of Mathematics Studies, No. 53. Princeton University Press, Princeton, N.J., 1963.
4. A. Baker. *Transcendental number theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, second edition, 1990.
5. F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78:551–572, 1938.
6. V. Bergelson. Minimal idempotents and ergodic Ramsey theory. In *Topics in dynamics and ergodic theory*, volume 310 of *London Math. Soc. Lecture Note Ser.*, pages 8–39. Cambridge Univ. Press, Cambridge, 2003.
7. V. Bergelson and A. Leibman. Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.*, 9(3):725–753, 1996.
8. V. Bergelson and R. McCutcheon. An ergodic IP polynomial Szemerédi theorem. *Mem. Amer. Math. Soc.*, 146(695):viii+106, 2000.
9. G. D. Birkhoff. Proof of the ergodic theorem. *Proc. Nat. Acad. Sci. U.S.A.*, 17:656–660, 1931.
10. P. Bohl. Über ein in der Theorie der säkularen Störungen vorkommendes Problem. *J. für Math.*, 135:189–283, 1909.
11. E. Borel. Les probabilités denombrables et leurs applications arithmetiques. *Rend. Circ. Math. Palermo*, 27:247–271, 1909.
12. J. Bourgain. An approach to pointwise ergodic theorems. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 204–223. Springer, Berlin, 1988.
13. J. Bourgain. On the maximal ergodic theorem for certain subsets of the integers. *Israel J. Math.*, 61(1):39–72, 1988.
14. R. Bowen. Markov partitions for Axiom A diffeomorphisms. *Amer. J. Math.*, 92:725–747, 1970.
15. R. Bowen. The equidistribution of closed geodesics. *Amer. J. Math.*, 94:413–423, 1972.
16. R. Bowen. Symbolic dynamics for hyperbolic flows. *Amer. J. Math.*, 95:429–460, 1973.
17. V. Chothi, G. Everest, and T. Ward. S-integer dynamical systems: periodic points. *J. Reine Angew. Math.*, 489:99–132, 1997.
18. D. Dolgopyat. On decay of correlations in Anosov flows. *Ann. of Math. (2)*, 147(2):357–390, 1998.
19. M. Einsiedler and T. Ward. Asymptotic geometry of non-mixing sequences. *Ergodic Theory Dynam. Systems*, 23(1):75–85, 2003.
20. P. Erdös. On a new method in elementary number theory which leads to an elementary proof of the prime number theorem. *Proc. Nat. Acad. Sci. U. S. A.*, 35:374–384, 1949.
21. P. Erdős and P. Turán. On some sequences of integers. *J. London Math. Soc.*, 11:261–264, 1936.
22. G. Everest, R. Miles, S. Stevens, and T. Ward. Orbit-counting in non-hyperbolic dynamical systems. *J. Reine Angew. Math.*, 2007.

23. J.-H. Evertse and H. P. Schlickewei. The absolute subspace theorem and linear equations with unknowns from a multiplicative group. In *Number theory in progress, Vol. 1 (Zakopane-Kościelisko, 1997)*, pages 121–142. de Gruyter, Berlin, 1999.
24. H. Furstenberg. Strict ergodicity and transformation of the torus. *Amer. J. Math.*, 83:573–601, 1961.
25. H. Furstenberg. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Analyse Math.*, 31:204–256, 1977.
26. H. Furstenberg and Y. Katznelson. An ergodic Szemerédi theorem for commuting transformations. *J. Analyse Math.*, 34:275–291 (1979), 1978.
27. H. Furstenberg and Y. Katznelson. An ergodic Szemerédi theorem for IP-systems and combinatorial theory. *J. Analyse Math.*, 45:117–168, 1985.
28. H. Furstenberg and B. Weiss. Topological dynamics and combinatorial number theory. *J. Analyse Math.*, 34:61–85 (1979), 1978.
29. A. O. Gel′fond. *Transcendental and algebraic numbers*. Translated from the first Russian edition by Leo F. Boron. Dover Publications Inc., New York, 1960.
30. D. Goldfeld. The elementary proof of the prime number theorem: an historical perspective. In *Number theory (New York, 2003)*, pages 179–192. Springer, New York, 2004.
31. D. A. Goldston and C. Y. Yıldırım. Small gaps between primes I. 2005. `arXiv:math.NT/0504336`.
32. T. Gowers. Hypergraph regularity and the multidimensional Szemerédi Theorem. 2006. `www.dpmms.cam.ac.uk/∼wtg10`.
33. B. Green and T. Tao. The primes contain arbitrarily long arithmetic progressions. 2004. `arXiv:math.NT/0404188`.
34. B. Green and T. Tao. Linear equations in primes. 2006. `arXiv:math.NT/0606088`.
35. T. P. Hill. Base-invariance implies Benford's law. *Proc. Amer. Math. Soc.*, 123(3):887–895, 1995.
36. N. Hindman. Finite sums from sequences within cells of a partition of $\mathbb{N}$. *J. Combinatorial Theory Ser. A*, 17:1–11, 1974.
37. E. Hlawka. Discrepancy and uniform distribution of sequences. *Compositio Math.*, 16:83–91 (1964), 1964.
38. B. Host and B. Kra. Convergence of polynomial ergodic averages. *Israel J. Math.*, 149:1–19, 2005. Probability in mathematics.
39. B. Host and B. Kra. Nonconventional ergodic averages and nilmanifolds. *Ann. of Math. (2)*, 161(1):397–488, 2005.
40. H. Huber. Zur analytischen Theorie hyperbolischen Raumformen und Bewegungsgruppen. *Math. Ann.*, 138:1–26, 1959.
41. A. Katok. Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst. Hautes Études Sci. Publ. Math.*, (51):137–173, 1980.
42. A. Katsuda and T. Sunada. Closed orbits in homology classes. *Inst. Hautes Études Sci. Publ. Math.*, (71):5–32, 1990.
43. A. I. Khinchin. *Continued fractions*. The University of Chicago Press, Chicago, Ill.-London, 1964.
44. G. Knieper. On the asymptotic geometry of nonpositively curved manifolds. *Geom. Funct. Anal.*, 7(4):755–782, 1997.
45. B. Koopman. Hamiltonian systems and transformations in Hilbert spaces. *Proc. Nat. Acad. Sci. U.S.A.*, 17:315–318, 1931.
46. R. O. Kuz′min. A problem of Gauss. *Dokl. Akad. Nauk*, pages 375–380, 1928.
47. S. P. Lalley. Distribution of periodic orbits of symbolic and Axiom A flows. *Adv. in Appl. Math.*, 8(2):154–193, 1987.
48. S. P. Lalley. The "prime number theorem" for the periodic orbits of a Bernoulli flow. *Amer. Math. Monthly*, 95(5):385–398, 1988.
49. F. Ledrappier. Un champ markovien peut être d'entropie nulle et mélangeant. *C. R. Acad. Sci. Paris Sér. A-B*, 287(7):A561–A563, 1978.
50. A. Leibman. Convergence of multiple ergodic averages along polynomials of several variables. *Israel J. Math.*, 146:303–315, 2005.
51. P. Levy. Sur les lois de probabilité dont dependent les quotients complets et incomplets d'une fraction continue. *Bull. Soc. Math. France*, 57:178–194, 1929.

52. P. Lévy. Sur quelques points de la théorie des probabilités dénombrables. *Ann. Inst. H. Poincaré*, 6(2):153–184, 1936.
53. D. A. Lind. Dynamical properties of quasihyperbolic toral automorphisms. *Ergodic Theory Dynamical Systems*, 2(1):49–68, 1982.
54. G. A. Margulis. Certain applications of ergodic theory to the investigation of manifolds of negative curvature. *Funkcional. Anal. i Priložen.*, 3(4):89–90, 1969.
55. G. A. Margulis. *On some aspects of the theory of Anosov systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004. With a survey by Richard Sharp: Periodic orbits of hyperbolic flows, Translated from the Russian by Valentina Vladimirovna Szulikowska.
56. D. W. Masser. Mixing and linear equations over groups in positive characteristic. *Israel J. Math.*, 142:189–204, 2004.
57. F. Mertens. Ein Beitrag zur analytischen Zahlentheorie. *J. Reine Angew. Math.*, 78:46–62, 1874.
58. B. Nagle, V. Rödl, and M. Schacht. The counting lemma for regular k-uniform hypergraphs. *Random Structures Algorithms*, 28(2):113–179, 2006.
59. S. Newcomb. Note on the frequency of the use of digits in natural numbers. *Amer. J. Math.*, 4(1):39–40, 1881.
60. M. S. Md. Noorani. Mertens' theorem and closed orbits of ergodic toral automorphisms. *Bull. Malaysian Math. Soc. (2)*, 22(2):127–133, 1999.
61. J. C. Oxtoby. Ergodic sets. *Bull. Amer. Math. Soc.*, 58:116–136, 1952.
62. W. Parry. Ergodic properties of affine transformations and flows on nilmanifolds. *Amer. J. Math.*, 91:757–771, 1969.
63. W. Parry. An analogue of the prime number theorem for closed orbits of shifts of finite type and their suspensions. *Israel J. Math.*, 45(1):41–52, 1983.
64. W. Parry. Bowen's equidistribution theory and the Dirichlet density theorem. *Ergodic Theory Dynam. Systems*, 4(1):117–134, 1984.
65. W. Parry and M. Pollicott. An analogue of the prime number theorem for closed orbits of Axiom A flows. *Ann. of Math. (2)*, 118(3):573–591, 1983.
66. H. Poincaré. Sur le problème des trois corps et les équations de la Dynamique. *Acta Math.*, 13:1–270, 1890.
67. M. Pollicott and R. Sharp. Exponential error terms for growth functions on negatively curved surfaces. *Amer. J. Math.*, 120(5):1019–1042, 1998.
68. A. J. van der Poorten and H. P. Schlickewei. Additive relations in fields. *J. Austral. Math. Soc. Ser. A*, 51(1):154–170, 1991.
69. R. Rado. Studien zur Kombinatorik. *Math. Z.*, 36(1):424–470, 1933.
70. M. Ratner. Markov partitions for Anosov flows on n-dimensional manifolds. *Israel J. Math.*, 15:92–114, 1973.
71. V. A. Rohlin. On endomorphisms of compact commutative groups. *Izvestiya Akad. Nauk SSSR. Ser. Mat.*, 13:329–340, 1949.
72. K. Roth. Sur quelques ensembles d'entiers. *C. R. Acad. Sci. Paris*, 234:388–390, 1952.
73. A. Sárközy. On difference sets of sequences of integers. III. *Acta Math. Acad. Sci. Hungar.*, 31(3-4):355–386, 1978.
74. H. P. Schlickewei. S-unit equations over number fields. *Invent. Math.*, 102(1):95–107, 1990.
75. K. Schmidt. Mixing automorphisms of compact groups and a theorem by Kurt Mahler. *Pacific J. Math.*, 137(2):371–385, 1989.
76. K. Schmidt and T. Ward. Mixing automorphisms of compact groups and a theorem of Schlickewei. *Invent. Math.*, 111(1):69–76, 1993.
77. W. M. Schmidt. Norm form equations. *Ann. of Math. (2)*, 96:526–551, 1972.
78. A. Selberg. An elementary proof of the prime-number theorem. *Ann. of Math. (2)*, 50:305–313, 1949.
79. A. Selberg. Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc. (N.S.)*, 20:47–87, 1956.
80. R. Sharp. An analogue of Mertens' theorem for closed orbits of Axiom A flows. *Bol. Soc. Brasil. Mat. (N.S.)*, 21(2):205–229, 1991.
81. R. Sharp. Closed orbits in homology classes for Anosov flows. *Ergodic Theory Dynam. Systems*, 13(2):387–408, 1993.
82. W. Sierpiński. Sur la valeur asymptotique d'une certaine somme. *Bull Intl. Acad. Polonmaise des Sci. et des Lettres (Cracovie)*, pages 9–11, 1910.

83. Ja. G. Sinaĭ. Asymptotic behavior of closed geodesics on compact manifolds with negative curvature. *Izv. Akad. Nauk SSSR Ser. Mat.*, 30:1275–1296, 1966.
84. Ja. G. Sinaĭ. Construction of Markov partitionings. *Funkcional. Anal. i Priložen.*, 2(3):70–80, 1968.
85. S. Smale. Differentiable dynamical systems. *Bull. Amer. Math. Soc.*, 73:747–817, 1967.
86. E. Szemerédi. On sets of integers containing no four elements in arithmetic progression. *Acta Math. Acad. Sci. Hungar.*, 20:89–104, 1969.
87. E. Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, 27:199–245, 1975.
88. T. Tao and T. Ziegler. The primes contain arbitrarily long polynomial progressions. 2006. `arXiv:math.NT/0610050`.
89. B. L. van der Waerden. Beweis einer Baudet'schen Vermutung. *Nieuw. Arch. Wisk.*, 15:212–216, 1927.
90. J. von Neumann. Proof of the quasi-ergodic hypothesis. *Proc. Nat. Acad. Sci. U.S.A.*, 18:70–82, 1932.
91. S. Waddington. The prime orbit theorem for quasihyperbolic toral automorphisms. *Monatsh. Math.*, 112(3):235–248, 1991.
92. T. Ward. Almost all S-integer dynamical systems have many periodic points. *Ergodic Theory Dynam. Systems*, 18(2):471–486, 1998.
93. H. Weyl. Über die *Gibbs*sche Erscheinung und verwandte Konvergenzphänomene. *Rendiconti del Circolo Matematico di Palermo*, 30:377–407, 1910.
94. H. Weyl. Uber die Gleichverteilung von Zahlen mod Eins. *Math. Ann.*, 77:313–352, 1916.
95. N. Wiener. Tauberian theorems. *Ann. of Math. (2)*, 33(1):1–100, 1932.

Books and Reviews

96. V. I. Arnol′d and A. Avez. *Ergodic problems of classical mechanics*. Translated from the French by A. Avez. W. A. Benjamin, Inc., New York-Amsterdam, 1968.
97. V. Bergelson. Ergodic Ramsey theory—an update. In *Ergodic theory of $\mathbf{Z}^d$ actions (Warwick, 1993–1994)*, volume 228 of *London Math. Soc. Lecture Note Ser.*, pages 1–61. Cambridge Univ. Press, Cambridge, 1996.
98. V. Bergelson. Ergodic theory and Diophantine problems. In *Topics in symbolic dynamics and applications (Temuco, 1997)*, volume 279 of *London Math. Soc. Lecture Note Ser.*, pages 167–205. Cambridge Univ. Press, Cambridge, 2000.
99. V. Bergelson. Combinatorial and Diophantine applications of ergodic theory. In *Handbook of dynamical systems. Vol. 1B*, pages 745–869. Elsevier B. V., Amsterdam, 2006. Appendix A by A. Leibman and Appendix B by Anthony Quas and Máté Wierdl.
100. I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinaĭ. *Ergodic theory*. Springer-Verlag, New York, 1982.
101. K. Dajani and C. Kraaikamp. *Ergodic theory of numbers*, volume 29 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, 2002.
102. A. del Junco. Ergodic theorems. In *Encyclopedia of Complexity and Systems Science*. Springer, 2008.
103. M. Denker, C. Grillenberger, and K. Sigmund. *Ergodic theory on compact spaces*. Springer-Verlag, Berlin, 1976. Lecture Notes in Mathematics, Vol. 527.
104. R. Ellis. *Lectures on topological dynamics*. W. A. Benjamin, Inc., New York, 1969.
105. H. Furstenberg. *Recurrence in ergodic theory and combinatorial number theory*. Princeton University Press, Princeton, N.J., 1981. M. B. Porter Lectures.
106. H. Furstenberg, Y. Katznelson, and D. Ornstein. The ergodic theoretical proof of Szemerédi's theorem. *Bull. Amer. Math. Soc. (N.S.)*, 7(3):527–552, 1982.
107. E. Glasner. *Ergodic theory via joinings*, volume 101 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2003.
108. D. A. Hejhal. The Selberg trace formula and the Riemann zeta function. *Duke Math. J.*, 43(3):441–482, 1976.
109. M. Iosifescu and C. Kraaikamp. *Metrical theory of continued fractions*, volume 547 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.
110. A. Katok and B. Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1995. With a supplementary chapter by Anatole Katok and Leonardo Mendoza.

111. B. Kra. The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view. *Bull. Amer. Math. Soc. (N.S.)*, 43(1):3–23 (electronic), 2006.
112. U. Krengel. *Ergodic theorems*, volume 6 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1985. With a supplement by Antoine Brunel.
113. R. McCutcheon. *Elemental methods in ergodic Ramsey theory*, volume 1722 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1999.
114. K. Petersen. *Ergodic theory*, volume 2 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1989. Corrected reprint of the 1983 original.
115. K. Schmidt. *Dynamical systems of algebraic origin*, volume 128 of *Progress in Mathematics*. Birkhäuser Verlag, Basel, 1995.
116. K. Schmidt. The dynamics of algebraic $\mathbb{Z}^d$-actions. In *European Congress of Mathematics, Vol. I (Barcelona, 2000)*, volume 201 of *Progr. Math.*, pages 543–553. Birkhäuser, Basel, 2001.
117. F. Schweiger. *Ergodic theory of fibred systems and metric number theory*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1995.
118. J. H. Silverman. *The Arithmetic of Dynamical Systems*, volume 241 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2007.
119. T. Tao. Arithmetic progressions and the primes. *Collect. Math.* 2006, Vol. Extra, 37–88.
120. T. Tao. The dichotomy between structure and randomness, arithmetic progressions, and the primes. 2005. `arXiv:math/0512114v2`.
121. T. Tao. What is good mathematics?. 2007. `arXiv:math/0702396v1`.
122. H. Totoki. *Ergodic theory*. Lecture Notes Series, No. 14. Matematisk Institut, Aarhus Universitet, Aarhus, 1969.
123. B. L. van der Waerden. How the proof of Baudet's conjecture was found. In *Studies in Pure Mathematics (Presented to Richard Rado)*, pages 251–260. Academic Press, London, 1971.
124. P. Walters. *An introduction to ergodic theory*, volume 79 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1982.
125. A. Weil. *Basic number theory*. Die Grundlehren der mathematischen Wissenschaften, Band 144. Springer-Verlag New York, Inc., New York, 1967.

School of Mathematics, University of East Anglia, Norwich NR4 7TJ, United Kingdom

E-mail address: `t.ward@uea.ac.uk`