

LECTURES ON ERGODIC THEORY

KARL PETERSEN

1. MEASURE-PRESERVING DYNAMICAL SYSTEMS AND CONSTRUCTIONS

1.1. Sources of the Subject.

1.1.1. *Physics.* Ideal gas. The *state* of a system of N particles is specified completely by a point in the phase space $X = \mathbb{R}^{6N} = \{(q_i, p_i) : i = 1, \dots, N\}$. The system evolves (in continuous time or from one discrete time instant to the next) according to deterministic physical laws (Hamilton's equations). We are interested in *average values* of observables (like kinetic energy), both for a single particle over the long term (*time average*) or over all particles at a particular instant (*ensemble average* or *space average*). Note that the "long time" might actually be a fraction of a second, so existence of these averages is connected with the existence (or observability) of physical quantities. And the space average looks like an average over a set of points, but it is in fact the average over all possible states of the system, hence over an ensemble of many "ideal" systems. Such models of ideal gases lead to apparent paradoxes. Consider a frictionless billiard table with the balls set up in their usual initial clump, except for the cue ball, which is traveling towards this clump at high speed. Of course the balls are scattered and careen around the table, colliding (perfectly elastically, we assume) with each other and the cushions. But with probability 1 the balls will all return at some moment arbitrarily close to their initial positions *and initial velocities*—in fact they will do so infinitely many times!

Often a physical system is described by a system of differential equations. Sometimes many initial points have trajectories (solutions of the system) that approach a fixed point or limit cycle as $t \rightarrow \infty$. Some nonlinear systems, including those that describe fluid flow or certain biological or chemical oscillations, have many trajectories approaching "strange attractors"—closed invariant limit sets supporting more complicated dynamics and more interesting invariant measures than are possible on a single periodic trajectory. In such a case the complicated long-term equilibrium behavior can be fruitfully studied by means of the concepts of ergodic theory.

1.1.2. *Statistics.* Let X_1, X_2, \dots be a stationary stochastic process, i.e. a sequence of measurable functions on a probability space for which $P\{X_{i+n} \in A_i : i =$

$1, \dots, k\}$ (for Borel sets $A_i \subset \mathbb{R}$) is always independent of n . According to the Strong Law of Large Numbers, under some conditions

$$\frac{1}{n} \sum_{k=1}^n X_k \rightarrow \int_{\Omega} X_1 dP \text{ a.e. ,}$$

that is, the averages of the terms in the sequence tend to the mean, or expectation, of any one of them. Considerations of this kind have been proposed as a basis for the determination of probabilities in terms of frequencies of events—indeed, sometimes even as a basis for the theory of probability.

1.1.3. *Mathematics.* E. Borel showed that a.e. number is normal to every base. Champernowne showed that .0123456789101112131415161718192021... is normal base 10. Van der Waerden et al. showed that if the integers are colored by finitely many colors, then there are arbitrarily long monochromatic arithmetic progressions. Szemerédi showed that every subset of \mathbb{N} that has positive upper density contains arbitrarily long arithmetic progressions. Does the sequence of primes contain arbitrarily long arithmetic progressions? How can subsets of Euclidean space be generated which have fractional dimensions, and how can their dimensions be calculated? What is the asymptotic behavior of an operator on a Banach space? These are examples of questions that can be handled by ergodic-theoretic methods.

1.1.4. *Information Theory.* A message is a sequence of symbols, usually coming from a finite alphabet. To think of it as a *stream* of symbols, we introduce the shift transformation, which makes time go by (by making the sequence go by our viewing window). We may want to *recode* the sequence in order to achieve *reliability* (protect against errors or noise in the transmission)—this can be done by adding some *redundancy*—or to achieve *efficiency*, which can be done by *removing* redundancy. What are the theoretical possibilities, and what practical algorithms can be found to implement them? This depends heavily on the structural and statistical properties of the signal streams.

Of course there are overlaps in the above, and gains in insight to be made in combining the areas, for example coding the successive states of a statistical-mechanical system by a finite alphabet and applying the viewpoint of information theory.

1.2. Models.

- (1) Differentiable dynamics, topological dynamics, ergodic theory.
- (2) mpt=automorphism of \mathcal{B} >nonsingular map>operator on L^p
- (3) X =Lebesgue space=set of all possible *states* of the system
- (4) \mathcal{B} = σ -algebra of subsets of X =*observable events* (reflecting impossibility of perfect measurements on the system)
- (5) μ =measure on X ; gives the *probabilities* of the observable events
- (6) $f : X \rightarrow \mathbb{R}$, measurable function, is a *measurement* on the system or a *random variable*

- (7) $T : X \rightarrow X$, measure-preserving transformation or *m.p.t.*: one-to-one onto (up to sets of measure 0), $T^{-1}\mathcal{B} = \mathcal{B}$, $\mu T^{-1} = \mu$. T makes the system *develop in time*. The invariance of μ means that we are in an *equilibrium* situation, but not necessarily a *static* one!
- (8) *orbit* or *trajectory* of a point $x \in X$ is $\mathcal{O} = \{T^n x : n \in \mathbb{Z}\}$. This represents one entire *history* of the system. Think of x as the initial point, at time 0, for a system of differential equations; then at time n the solution which passes through x at time 0 is at the point $T^n x$.
- (9) A *coding* is accomplished by a (finite measurable) *partition = experiment = simple function*. It converts the orbit of a point into an *itinerary*. It is also an example of a kind of *coarse-graining*, the production of *factors* of the given system.

1.3. Fundamental questions.

- (1) Existence of long-term averages—ergodic theorems.
- (2) Ergodic hypothesis and its variants: recurrence, ergodicity, various kinds of mixing.
- (3) Classification. Isomorphism. Spectral invariants. Entropy.
- (4) Realization (as diffeomorphism of a manifold, say), systematic construction (Jewett-Krieger, Vershik).
- (5) Genericity, stability of various properties.
- (6) The study of particular examples.
- (7) Applications in other subjects, such as information theory, number theory, geometry, mechanics, statistical mechanics, population dynamics, etc.

1.4. Constructions.

- (1) Factors and products.
- (2) Flow under a function, Poincaré map.
- (3) Induced transformations, towers, Rokhlin lemma, rank, Kakutani equivalence, loosely Bernoulli
 - (T has *rank* $\leq r$ if for every measurable set A and every $\epsilon > 0$ there are r disjoint columns of sets, each level in a column being the image under T of the level below it, such that a union of certain levels of the columns equals A up to a set of measure $< \epsilon$.)
 - (Two transformations are *Kakutani equivalent* if they are both (isomorphic to) first-return maps, presumably to different subsets, in the same dynamical system—or, equivalently, if the flows under certain ceiling functions built over them are isomorphic—or, still equivalently, if one can be constructed from the other by a sequence of tower building and inducing. A transformation is *loosely Bernoulli* if it is Kakutani equivalent either to an irrational rotation, a finite-entropy Bernoulli system, or an infinite-entropy Bernoulli system.

The horocycle flow on a compact surface of constant negative curvature is loosely Bernoulli, but its Cartesian square is not (Ratner). (So all irrational rotations are “the same”—*and* the same as a horocycle map!) (See [Ferenczi].)

The T, T^{-1} map (skew product of a Bernoulli shift with itself) is K but not loosely Bernoulli (Kalikow).

- (4) Cutting and stacking
- (5) Skew products
- (6) Inverse limits, natural extensions
- (7) Joinings and disjointness

A *joining* of two systems is an invariant probability measure on their cross-product that projects to the given measures on the two factors. Two systems are *disjoint* if their only joining is product measure.

$\mathcal{E} = \{I\}^\perp$, proof via example of the relatively independent joining

$\mathcal{W} = \{\text{rotations}\}^\perp$

$\mathcal{K} = \{h = 0\}^\perp$

Minimal self-joinings, prime systems:

In $X \times X$ there is always the *diagonal* self-joining μ_Δ determined by

$$\int_{X \times X} f(x_1, x_2) d\mu_\Delta = \int_X f(x, x) d\mu,$$

and the *off-diagonal joinings*

$$\int_{X \times X} f(x_1, x_2) d\mu_j = \int_X f(x, T^j x) d\mu.$$

If these are the only ergodic ones, besides possibly the product measure, we say that T has *2-fold minimal self-joinings*.

Theorem 1.1 (Rudolph). *If T has 2-fold MSJ and each T^k is ergodic, then T commutes only with its powers and has no nontrivial factor algebras.*

On X^k there are many *off-diagonal measures* like

$$\int_{X^k} f(x_1, \dots, x_k) d\lambda = \int_X f(T^{j_1}, \dots, T^{j_k}) d\mu.$$

More generally, the set of indices $\{1, \dots, k\}$ can be partitioned into disjoint subsets, each of which determines an off-diagonal measure as above, and then we could take the direct product of the resulting measures; such a measure is called a *product of off-diagonals*. We say that (X, T) has *k-fold MSJ* if every k -fold self-joining of (X, T) is in the convex hull of products of off-diagonals.

Theorem 1.2 (del Junco-Rahe-Swanson). *The Chacon system has MSJ of all orders.*

Some other examples coming from MSJ:

- (a) a transformation without roots

- (b) transformation with two non-isomorphic square roots
- (c) two systems with no common factors that are not disjoint

Theorem 1.3 (J. King). *If T has 4-fold MSJ, then it has MSJ of all orders.*

Theorem 1.4 (Glasner, Host, Rudolph). *If T has 3-fold MSJ, then it has MSJ of all orders.*

(See JPT].)

- (8) Lebesgue spaces, Rokhlin partitions, ergodic decomposition

2. ERGODIC THEOREMS

Theorem 2.1 (Mean Ergodic Theorem (von Neumann, 1932)). *Let $T : \mathcal{H} \rightarrow \mathcal{H}$ be a contraction on a Hilbert space \mathcal{H} . Then for each $v \in \mathcal{H}$, the averages*

$$A_n v = \frac{1}{n} \sum_{k=0}^{n-1} T^k v$$

converge to the projection of v onto the closed subspace $\{u \in \mathcal{H} : Tu = u\}$.

Theorem 2.2 (Pointwise Ergodic Theorem (Birkhoff, 1931)). *Let $T : X \rightarrow X$ be a m.p.t. on a measure space (X, \mathcal{B}, μ) and $f \in L^1(X, \mathcal{B}, \mu)$. Then the averages*

$$A_n f(x) = \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x)$$

converge a.e. on X to an (a.e.) invariant function. They also converge in the mean of order 1 on every invariant set of finite measure.

There were some historical shenanigans about which of these theorems came first. Also papers have been published arguing which one deserves to be regarded as the “real Ergodic Theorem”. I could say quite a lot about this philosophical question, arguing all sides of it from various viewpoints, especially in connection with some of the following theorems, such as Wiener-Wintner, return-times, and random sampling.

Theorem 2.3 (Maximal Ergodic Theorem (Wiener 1939, Yosida-Kakutani 1939)). *For a m.p.t. on X and measurable function f on X , define the maximal function f^* of f by $f^* = \sup_n A_n f$. If $f \in L^1$, then*

$$\int_{\{f^* \geq 0\}} f d\mu \geq 0.$$

This is used for proving that the set of functions for which a.e. convergence of the averages occurs is *closed* in L^1 . For these averages, it is not hard to find a dense set of functions for which a.e. convergence takes place: invariant functions plus coboundaries $g - Tg$. For some of the theorems given below, *neither* the dense set of functions nor the maximal inequality is easy to come by. Bourgain used harmonic-analytic methods to produce *quadratic-variation estimates* which are stronger than maximal inequalities and imply a.e. convergence directly.

Proofs: By towers (Kolmogorov and later), positivity (Hopf and Garsia), transference (Cotlar, Calderón), ideas from nonstandard analysis (Kamae, Katznelson-Weiss).

Theorem 2.4 (Subadditive Ergodic Theorem (Kingman, 1968)). *If $T : X \rightarrow X$ is a m.p.t. on a probability space and $\{F_n\}$ is a subadditive sequence of integrable functions on X in that*

$$F_{n+m} \leq F_n + F_m \circ T^n \text{ for all } m, n \in \mathbb{N},$$

and if

$$\gamma = \inf \frac{1}{n} \int_X F_n d\mu > -\infty,$$

then F_n/n converges a.e. and in L^1 to an invariant function.

Theorem 2.5 (Wiener-Wintner (1941)). *If $T : X \rightarrow X$ is a m.p.t. and $f \in L^1$, then for a.e. $x \in X$*

$$\frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) e^{ik\theta}$$

converges for all θ .

(Bourgain showed that if f is in the orthocomplement of the eigenfunctions, then the convergence (to 0) is uniform in θ .)

Theorem 2.6 (Ergodic Hilbert Transform (Cotlar, 1955)). *If $T : X \rightarrow X$ is a m.p.t. and $f \in L^1$, then*

$$H_n f(x) = \sum_{k=-n}^n \frac{f(T^k x)}{k}$$

converges for a.e. x .

The following is a double maximal inequality for the helical transform.

Theorem 2.7 (Campbell-Petersen, 1989). *If $T : X \rightarrow X$ is a m.p.t. and for $f \in L^2$*

$$H^{**} f(x) = \sup_{n, \theta} \left| \sum_{k=-n}^n \frac{f(T^k x) e^{ik\theta}}{k} \right|,$$

then

$$\mu\{x : H^{**} f(x) > \lambda\} \leq \frac{C \|f\|_2^2}{\lambda^2} \text{ for each } \lambda > 0.$$

The following theorem is due to Bourgain, with other proofs by Bourgain-Furstenberg-Katznelson-Ornstein and Rudolph.

Theorem 2.8 (Return Times Theorem). *Let $T : X \rightarrow X$ be a m.p.t. and $f \in L^\infty(X)$. Then a.e. $x \in X$ has the following property: given a measure-preserving system (Y, \mathcal{C}, ν, S) and $g \in L^1(Y)$,*

$$\frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) g(S^k y)$$

converges a.e. $d\nu(y)$.

Theorem 2.9 (Nonlinear averages). *For polynomials $q_1(n), \dots, q_r(n) \in \mathbb{Z}[n]$ and (usually bounded) measurable f_1, \dots, f_r on X , let*

$$A_n = \frac{1}{n} \sum_{k=0}^{n-1} f_1(T^{q_1(k)}x) \dots f_r(T^{q_r(k)}x).$$

- (1) (Furstenberg) *If each f_i is the characteristic function of a measurable set A of positive measure, and if $q_i(k) = ik$, then the \liminf of the integrals of the A_n is positive. This implies Szemerédi's Theorem.*
- (2) (Furstenberg) *For $r = 1$, we have L^2 convergence.*
- (3) (Conze-Lesigne) *For $r \leq 3$ and all $\deg q_i = 1$, we have L^1 convergence.*
- (4) (Furstenberg-Weiss) *For $r = 2$, $q_1(n) = n, q_2(n) = n^2$, we have weak convergence to the product of the integrals of f_1 and f_2 .*
- (5) (Bergelson) *For T weakly mixing and the q_i and $q_i - q_j$ ($i \neq j$) not constant, we have L^2 convergence to the product of the integrals of the f_i . (Also have L^p convergence for $f_i \in L^{p_i}, \sum \frac{1}{p_i} \leq \frac{1}{p}$, as was pointed out by Derrien and Lesigne.)*
- (6) (Bourgain) *For $r = 1$, we have a.e. convergence for $f \in L^p, p > 1$.*
- (7) (Bourgain) *For $r = 2$ and all $\deg q_i = 1$, we have a.e. convergence ($f_1, f_2 \in L^2$).*
- (8) (Derrien-Lesigne; presented at Alexandria 1993): *If T is a K -automorphism then we have a.e. convergence for $f_i \in L^{p_i}, \sum 1/p_i < 1$.*

There are partial results when we consider not just powers of T but several commuting m.p.t.'s. There are also some similar results and many open questions when other subsequences (such as the primes) replace polynomially-generated subsequences. For the squares or primes, a.e. convergence of the averages is *not known* for $f \in L^1$. Further, averages of this kind can be combined with questions of the Wiener-Wintner type or with "singular averages" as in the Hilbert or helical transforms. Michael Lacey, for one, is attempting a unified approach to such questions by proving certain quadratic-variation statements for singular integrals and transferring them to the ergodic-theoretic context. There is also great interest in trying to identify the limits for the averages above, even of trying to find the factor algebras generated by the set of all limits (as for the usual ergodic theorem it is the σ -algebra of invariant sets). Curiously, even though the acting group in these questions is at most \mathbb{Z}^d , the answers seem to bring in noncommutative harmonic analysis on certain nilpotent Lie groups.

The following theorem is a sort of random ergodic theorem, in a strong form with universally representative sampled sequences.

Theorem 2.10 (Lacey-Petersen-Rudolph-Wierdl). *Let $T : \Omega \rightarrow \Omega$ be an ergodic m.p.t., $\delta : \Omega \rightarrow \mathbb{Z}$ an integrable function, $\tau_k(\omega) = \delta(\omega) + \delta(T\omega) + \dots + \delta(T^{k-1}\omega)$ for $k \geq 1$, and, for any measure-preserving system (Y, \mathcal{C}, ν, U) and integrable function g on Y ,*

$$A_n g(y) = \frac{1}{n} \sum_{k=1}^n g(U^{\tau_k(\omega)}y).$$

If δ has nonzero mean, then for a.e. $\omega \in \Omega$ we always have a.e. convergence $d\nu(y)$ of the $A_n g(y)$ for every $(Y, \mathcal{C}, \nu, U, g)$. On the other hand, if δ has mean 0, then if,

for example, (Ω, T) is a Bernoulli (i.i.d.) sequence of ± 1 's, each appearing with probability $1/2$, in any aperiodic system Y we will be able to find a counterexample, even a characteristic function g , even with "strong sweeping out".

(Compare to original Ulam-von Neumann Random Ergodic Theorem, which is easily proved by taking a skew product.)

On the other hand, we always have convergence of the fixed subsequence n^n of the averages. Similarly for higher-dimensional actions (several commuting m.p.t.'s). The general question of just which (Ω, T, δ) a.s. always produce a.e. convergence for every (Y, U, g) is very much open.

There are many other ergodic theorems, including for actions of $\mathbb{Z}^d, \mathbb{R}^d$, amenable groups, free groups, etc..

3. SPECTRAL PROPERTIES AND MIXING

3.1. The unitary operator and spectral measure of a m.p.t. Given a m.p.t. $T : X \rightarrow X$, it now seems extremely natural to consider T also as a unitary operator on $L^2(X)$ acting according to $Tf(x) = f(Tx)$, but when Koopman first had this insight it was greeted with great *éclat* by von Neumann and the rest. Of course T also acts by composition on the other L^p spaces too. Properties of T that are preserved under unitary equivalence of the associated unitary operators are called *spectral properties*. T_1 and T_2 are called *spectrally isomorphic* if their associated unitary operators are unitarily equivalent. This is a coarser equivalence relation than (point) isomorphism.

The unitary operator T on $L^2(X)$ has a *spectral measure*, a countably additive set function E defined on the Borel subsets of $[-\pi, \pi]$ with $E(B)$ being a projection on L^2 for each Borel $B \subset [-\pi, \pi]$ and such that $T^k = \int_{-\pi}^{\pi} e^{ik\theta} dE(\theta)$ for all $k \in \mathbb{Z}$. Moreover, for each $f, g \in L^2$, $(E(\cdot)f, g)$ is a complex-valued Borel measure and $(T^k f, g) = \int_{-\pi}^{\pi} e^{ik\theta} d(E(\theta)f, g)$ for all $k \in \mathbb{Z}$. In the sense of absolute continuity, the minimal positive measure type that dominates all these measures $(E(\cdot)f, g)$ is called the *maximal spectral type* of T .

3.2. Recurrence.

Theorem 3.1 (Poincaré Recurrence Theorem (1899)). *Let $T : X \rightarrow X$ be a m.p.t. on a space X of finite measure.*

- (1) *If $A \subset X$ and $\mu(A) > 0$, then there is $n \geq 1$ with $\mu(T^n A \cap A) > 0$.*
- (2) *If $A \subset X$ and $\mu(A) > 0$, then for almost every $x \in A$ there is a positive integer $n(x)$ such that $T^{n(x)}x \in A$.*

Theorem 3.2 (Khinchine Recurrence Theorem (1934)). *If $A \subset X$ and $\mu(A) > 0$, then for each $\epsilon > 0$, the set of integers n for which $\mu(T^n A \cap A) \geq \mu(A)^2 - \epsilon$ is relatively dense (i.e., has bounded gaps).*

The following result provides an ergodic-theoretic proof of Szemerédi's Theorem, according to which every subset of the positive integers with positive upper (even Banach) density contains arbitrarily long arithmetic progressions.

Theorem 3.3 (Furstenberg Multiple Recurrence Theorem). *If $A \subset X$ and $\mu(A) > 0$, then given any positive integer k there is a positive integer n such that $\mu(A \cap T^n A \cap T^{2n} A \cap \dots \cap T^{kn} A) > 0$.*

3.3. Equivalent conditions for ergodicity.

- (1) Every invariant ($\mu(T^{-1}A \triangle A) = 0$) set has measure 0 or 1.
- (2) Every invariant ($f \circ T = f$ a.e.) measurable function is constant a.e..
- (3) 1 is a simple eigenvalue of the unitary operator associated to T .
- (4) Equality of time means and space means: For each $f \in L^1(X)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int_X f d\mu.$$

- (5) For every measurable set $A \subset X$, a.e. point $x \in X$ has *mean sojourn time*

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

equal to the measure of A .

- (6) For each $f, g \in L^2(X)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} (T^k f, g) = (f, 1) \overline{(g, 1)}.$$

- (7) For each pair of measurable sets $A, B \subset X$,

$$\frac{1}{n} \sum_{k=0}^{n-1} \mu(T^k A \cap B) \rightarrow \mu(A)\mu(B).$$

- (8) $(X, T) \perp ([0, 1], I)$.

3.4. Equivalent conditions for strong mixing.

- (1) For each pair of measurable sets $A, B \subset X$, $\mu(T^n A \cap B) \rightarrow \mu(A)\mu(B)$ as $n \rightarrow \infty$.
- (2) For each $f, g \in L^2(X)$, $(T^n f, g) \rightarrow (f, 1) \overline{(g, 1)}$ as $n \rightarrow \infty$.
- (3) (Rényi, 1958) For each measurable $A \subset X$, $\mu(T^n A \cap A) \rightarrow \mu(A)^2$ as $n \rightarrow \infty$.
- (4) For each $f \in L^2(X)$, $T^n f \rightarrow \int f d\mu$ *weakly*.
- (5) (Blum-Hanson, 1960) For any increasing sequence $\{k_j\}$ of positive integers and $f \in L^2(X)$,

$$\left\| \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=0}^{n-1} T^{k_j} f - \int f d\mu \right\|_2 = 0.$$

- (6) (Ornstein, 1972) T^n is ergodic for all n , and there is c such that for each pair of measurable sets $A, B \subset X$, $\limsup \mu(T^n A \cap B) \leq c\mu(A)\mu(B)$.

3.5. Equivalent conditions for weak mixing.

- (1) For each pair of measurable sets $A, B \subset X$,

$$\frac{1}{n} \sum_{k=0}^{n-1} |\mu(T^k A \cap B) - \mu(A)\mu(B)| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

- (2) For each $f, g \in L^2(X)$,

$$\frac{1}{n} \sum_{k=0}^{n-1} |(T^k f, g) - (f, 1)\overline{(g, 1)}| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

- (3) For each pair of measurable sets $A, B \subset X$, there is a set of density zero $J \subset \mathbb{N}$ such that $\mu(T^n A \cap B) \rightarrow \mu(A)\mu(B)$ as $n \rightarrow \infty$, $n \notin J$. (Actually, it's possible to select a single set J of density 0 that works simultaneously for all A, B .)
- (4) $T \times T$ is ergodic on $X \times X$.
- (5) $T \times S$ is ergodic on $X \times Y$ for each ergodic $S : Y \rightarrow Y$.
- (6) T has no measurable eigenfunctions ($Tf = \lambda f$ a.e.) besides the constants. Equivalently, the unitary operator on L^2 associated to T has (only) *continuous spectrum* (i.e., no discrete—point—spectrum other than the eigenvalue 1 with associated eigenspace the constants).
- (7) (X, T) is disjoint from every ergodic group rotation.
- (8) (Krengel) The set of *weakly wandering functions* $f \in L^2$ —i.e., those f for which there is an increasing sequence of integers n_k such that the functions $T^{n_k} f$ are pairwise orthogonal—is dense in $L^2(X)$.

(Recently this was strengthened by Bergelson, Kornfeld, and Mityagin, also by del Junco and Reinhold-Larsson, to show that the sequence n_k can be taken to be not too sparse, in that it can be an *IP-set*, the set of all finite sums of distinct elements of some infinite subset $G \subset \mathbb{N}$.)

- (9) (Krengel) The set of *weakly wandering partitions*—i.e., those P for which there is an increasing sequence of integers n_k such that the partitions $T^{n_k} P$ are mutually independent (any finite subfamily is independent)—is dense in the space of finite partitions of X .

Usually it is enough to verify any of the above conditions for ergodicity, strong mixing, or weak mixing involving sets only for all the members in some *generating semialgebra* (closed under finite intersections, complement of any member is a finite disjoint union of members) for the σ -*algebra* \mathcal{B} of all measurable subsets of X .

3.6. Other kinds of mixing.

- (1) *Mild mixing*: There are no nonconstant *rigid functions* for T , i.e. if $f \in L^2(X)$ and there are $k_j \rightarrow \infty$ with $T^{k_j} f \rightarrow f$ in L^2 , then f is constant.
- (2) *Light mixing*: For each pair of measurable sets $A, B \subset X$ with positive measure, $\liminf \mu(T^n A \cap B) > 0$.
- (3) *Partial mixing*: For some $0 < \alpha < 1$, for each pair of measurable sets $A, B \subset X$, $\liminf \mu(T^n A \cap B) \geq \alpha \mu(A)\mu(B)$.

- (4) *Mixing of order n* : For any measurable sets $A_1, \dots, A_n \subset X$, $\mu(T^{m_1}A_1 \cap \dots \cap T^{m_n}A_n) \rightarrow \mu(A_1) \dots \mu(A_n)$ as the m_i and $|m_i - m_j|, i \neq j$, tend to ∞ .

Theorem 3.4 (Kalikow). *Every mixing rank 1 transformation is mixing of all orders (and MSJ of all orders).*

This has been extended to finite-rank mixing systems by V.V. Ryzhikov.

Theorem 3.5 (Host). *If the maximal spectral type of T is singular with respect to Lebesgue measure and T is 2-mixing, then T is mixing of all orders.*

- (5) *Lebesgue spectrum of multiplicity $\leq M$* : There are a set Λ of cardinality M and functions $f_{\lambda,j}$, $\lambda \in \Lambda$ and $j \in \mathbb{Z}$, such that $Tf_{\lambda,j} = f_{\lambda,j+1}$ and $\{1\} \cup \{f_{\lambda,j} : \lambda \in \Lambda, j \in \mathbb{Z}\}$ forms an orthonormal basis for $L^2(X)$.

Question 3.1. Is there a m.p.t. with *simple* Lebesgue spectrum ($M = 1$)?

Question 3.2. Does there exist a system with maximal spectral type absolutely continuous with respect to Lebesgue measure, but not equivalent to it?

- (6) *K-automorphisms*: The following are equivalent definitions. Note that being K or Bernoulli are *not* spectral properties.

- (a) There is a subalgebra $\mathcal{A} \subset \mathcal{B}$ such that $T^{-1}\mathcal{A} \subset \mathcal{A}$, $\cup_{-\infty}^{\infty} T^n \mathcal{A}$ generates \mathcal{B} , and the *tail* $\cap_{-\infty}^0 T^n \mathcal{A}$ is trivial (i.e., consists only of sets of measure 0 or 1).

- (b) For every (finite measurable) partition P of X , for every measurable set $E \subset X$, $\sup\{|\mu(A \cap E) - \mu(A)\mu(E)| : A \in P_{-\infty}^{-n}\} \rightarrow 0$.

Let us say that two partitions α and β are ϵ -*independent*, and write $\alpha \perp^\epsilon \beta$, if there is a family \mathfrak{G} of atoms of β such that $\mu(\cup \mathfrak{G}) > 1 - \epsilon$ and for each $B \in \mathfrak{G}$,

$$|\text{dist}(\alpha|\beta) - \text{dist}(\alpha)| = \sum_{A \in \alpha} |\mu(A|B) - \mu(A)| < \epsilon.$$

(Equivalently, there is $f(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$ such that

$$\sum_{A,B} |\mu(A \cap B) - \mu(A)\mu(B)| < f(\epsilon).)$$

- (c) For each partition P of X there are $\epsilon_{n,m}$ with $\epsilon_{n,m} \rightarrow 0$ as $n \rightarrow \infty$ for each m such that $P_{-m}^0 \perp^{\epsilon_{n,m}} P_{-n-m}^{-n}$ for all m, n .

- (d) *K-mixing*: For every finite partition α and every measurable set B ,

$$\sup_{A \in \alpha_{-\infty}^{-n}} |\mu(A \cap B) - \mu(A)\mu(B)| \rightarrow 0.$$

This shows that for (α, T) to be K is between weak Bernoulli and α -mixing (see below).

- (e) For any finite partition P of X and $\epsilon > 0$ there is an N such that for any positive integer n and integers $m_1, \dots, m_n > N$ with $m_j - m_{j-1} > N$, $j = 2, \dots, n$,

$$\|E(B) \bigvee_j T^{m_j} P - \mu(B)\|_1 < \epsilon \text{ for all } B \in \mathcal{B}.$$

- (f) For every nontrivial finite partition P of X , $h(T, P) > 0$; i.e., T has *completely positive entropy*: there are no nontrivial 0-entropy factors.
 (g) For every partition $P = \{P_i\}$, $h(T^k, P) \rightarrow H(P) = -\sum_i \mu(P_i) \log_2 \mu(P_i)$.

For *noninvertible* measure-preserving T , the property corresponding to K is *exactness*: $\bigcap_{n=1}^{\infty} T^{-n} \mathcal{B} = \{\emptyset, X\}$ up to sets of measure 0.

- (7) *Weakly Bernoulli*: There is a generating partition P of X such that for every $\epsilon > 0$ there is an N such that $P_0^m \perp^\epsilon P_{-N-m}^{-N}$ for all m .
 (8) (*Finite-state*) *Bernoulli*: There is a generating partition P of X such that the $T^i P$ are independent (any finite number taken together are independent).

General Bernoulli: Shift on $(\Omega, \mathcal{F}, P)^{\mathbb{Z}}$.

Theorem 3.6. *Bernoulli* \Leftrightarrow *weakly Bernoulli* $\Rightarrow K \Rightarrow$ *countable Lebesgue spectrum and n -mixing for all n* .

Theorem 3.7. *Lebesgue spectrum* \Rightarrow *2-mixing* \Rightarrow *partial mixing* \Rightarrow *light mixing* \Rightarrow *mild mixing* \Rightarrow *weak mixing*.

- (9) *Some statistical kinds of mixing*: Taking the sup over $A \in \alpha_{-\infty}^0, B \in \alpha_n^\infty$, (or f, g measurable with respect to these σ -algebras for ρ , and over all finite partitions measurable with respect to these algebras for β and I), define

$$\begin{aligned} \alpha(n) &= \sup |\mu(A \cap B) - \mu(A)\mu(B)|, \\ \phi(n) &= \sup |\mu(A|B) - \mu(B)|, \\ \psi(n) &= \sup \left| \frac{\mu(A \cap B)}{\mu(A)\mu(B)} - 1 \right|, \\ \rho(n) &= \sup |\text{corr}(f, g)|, \\ \beta(n) &= \sup \sum_{A, B} |\mu(A \cap B) - \mu(A)\mu(B)|, \\ I(n) &= \sup - \sum_{A, B} \mu(A \cap B) \log \frac{\mu(A \cap B)}{\mu(A)\mu(B)}. \end{aligned}$$

Then

$\alpha(n) \rightarrow 0$ defines *α -mixing*, or *strong mixing in the sense of Rosenblatt*;

$\phi, \psi, \rho \rightarrow 0$ define *ϕ, ψ, ρ -mixing*, respectively;

$\beta(n) \rightarrow 0$ defines *absolute regularity*, or *weakly Bernoulli* (see above);

and $I(n) \rightarrow 0$ defines *information regularity*.

The following implications hold:

$$\begin{array}{ccccccc} \psi(n) \rightarrow 0 & \Rightarrow & \phi(n) \rightarrow 0 & \Rightarrow & \rho(n) \rightarrow 0 \\ \downarrow & & \downarrow & & \downarrow \\ I(n) \rightarrow 0 & \Rightarrow & \beta(n) \rightarrow 0 & \Rightarrow & \alpha(n) \rightarrow 0. \end{array}$$

(See R.C. Bradley, On a very weak Bernoulli condition, *Stochastics* 13 (1984), 61–81.)

4. SPECTRAL-THEORETIC AND HARMONIC-ANALYTIC APPROACHES TO A.E. CONVERGENCE

Theorem 4.1 (Gaposhkin,1981). *A normal operator $T : X \rightarrow X$ with spectral measure E and a function $f \in L^2(X)$ satisfy the pointwise ergodic theorem if and only if $E(-2^{-n}, 0)f(x) \rightarrow 0$ for a.e. x .*

Theorem 4.2 (Campbell-Petersen,1989). *For the normal operator T associated with a m.p.t. $T : X \rightarrow X$ with spectral measure E , for any $f \in L^2(X)$ and any sequence of positive numbers $\epsilon_n \rightarrow 0$, $E(-\epsilon_n, 0)f(x) \rightarrow 0$ for a.e. x .*

Theorem 4.3 (White, 1989, inspired by Bourgain). *For a function ϕ on \mathbb{Z} , let $A_n\phi(j) = \frac{1}{2n+1} \sum_{m=-n}^n \phi(p)$. There is a constant C such that whenever $\{n_k\}$ is a sequence of positive integers satisfying $n_{k+1} > n_k^8$ for all k , then for all $K > 0$ and all $\phi \in l^2$*

$$\sum_{k=1}^K \left\| \max_{n_{k-1} \leq n < n_k} |A_n\phi - A_{n_{k-1}}\phi| \right\|_{l^2}^2 \leq C\|\phi\|_{l^2}^2.$$

Consequently, for the ergodic averages

$$B_n f(x) = \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x)$$

for a m.p.t. $T : X \rightarrow X$ and measurable function f on X , there is a constant C such that whenever $\{n_k\}$ are as above,

$$\sum_{k=1}^K \left\| \max_{n_{k-1} \leq n < n_k} |B_n f - B_{n_{k-1}} f| \right\|_{L^2}^2 \leq C\|f\|_{L^2}^2.$$

This is a strengthening of Birkhoff's Ergodic Theorem. Bourgain's methods, and an array of related techniques which apply harmonic analysis to obtain a.e. convergence results in ergodic theory, are being applied and developed further to obtain many new theorems like those mentioned in the preceding lecture. (See [Rosenblatt-Wierdl].)

5. EXAMPLES

5.1. Group rotations. A typical example is $x \rightarrow x + \alpha \pmod{1}$ on $[0, 1]$, where α is irrational. More generally we may consider a compact abelian group, which we might as well take to be monothetic, with translation by a generator (element generating a dense subgroup) g_0 . Haar measure is invariant. These systems have *discrete spectrum*, and *every discrete spectrum system arises in this way*. The eigenfunctions are the characters of the group, and the eigenvalues are the values of the characters at the generating element. Two such systems are isomorphic iff they are spectrally isomorphic. They all have entropy 0. Many of these facts can be proved by means of harmonic analysis on the underlying group. There are applications in number theory, for example to uniform distribution mod 1.

5.2. Bernoulli shifts. On a one- or two-sided countably infinite product of copies of a probability space (usually a finite discrete space), we consider the product measure and the shift transformation. This is the ergodic-theoretic way to treat independent identically-distributed stochastic processes. These systems are mixing of all orders and have countable Lebesgue spectrum (hence are all spectrally isomorphic). Kolmogorov and Sinai showed that two of them cannot be isomorphic unless they have the same entropy; Ornstein showed the converse. $\mathcal{B}(1/2, 1/2)$ is isomorphic to the Lebesgue-measure-preserving transformation $x \rightarrow 2x \bmod 1$ on $[0, 1]$; similarly, $\mathcal{B}(1/3, 1/3, 1/3)$ is isomorphic to $x \rightarrow 3x \bmod 1$. Furstenberg asked whether the only nonatomic measure invariant for both $x \rightarrow 2x \bmod 1$ and $x \rightarrow 3x \bmod 1$ on $[0, 1]$ is Lebesgue measure. Lyons showed that if one of the actions is K , then the measure must be Lebesgue, and Rudolph showed the same thing under the weaker hypothesis that one of the actions has positive entropy. For recent work on this question, see [Host, Parry].

5.3. Markov shifts. On the space of sequences on an alphabet with n symbols, put a measure determined on cylinder sets by an n -dimensional probability vector p (the initial distribution) and a stochastic transition matrix P such that $pP = p$. This is the ergodic-theoretic formulation of a homogeneous stationary Markov chain (the transition probabilities do not change with time, and we have reached the equilibrium state). Aperiodic Markov chains are strongly mixing, in fact are isomorphic to Bernoulli shifts (probably using a different partition). If some transition probabilities are allowed to be 0, so that the graph showing the transitions could be a proper subgraph of the full graph on n symbols, then the underlying topological system is called a *subshift of finite type*. In the irreducible case, it has a unique measure of maximal entropy, which is a Gibbs measure.

5.4. Some symbolic systems. We consider some systems that are given by the shift transformation on a subset of the set of (usually doubly-infinite) sequences on a finite alphabet, usually $\{0, 1\}$. These subsets are *subshifts*—closed shift-invariant subsets. Associated with each subshift is its *language*, the set of all finite blocks seen in all sequences in the subshift. These languages are *factorial* (every subword of a word in the language is also in the language) and *extendable* (every word in the language extends on both sides to longer words in the language). In fact these two properties characterize the languages (subsets of the set of finite-length words on an alphabet) associated with subshifts.

5.4.1. Prouhet-Thue-Morse. 0

0 1
 0 1 10
 0 1 10 0110
 ⋮

At each stage write down the opposite ($0' = 1, 1' = 0$) or mirror image of what is available so far. Or, repeatedly apply the *substitution* $0 \rightarrow 01, 1 \rightarrow 10$. The n 'th entry is the sum, mod 2, of the digits in the dyadic expansion of n . Using Keane's *block multiplication* according to which if B is a block, $B \times 0 = B, B \times 1 = B'$,

and $B \times (\omega_1 \dots \omega_n) = (B \times \omega_1) \dots (B \times \omega_n)$, we may also obtain this sequence as

$$0 \times 01 \times 01 \times 01 \times \dots$$

The orbit closure of this sequence is uniquely ergodic (there is a unique shift-invariant Borel probability measure, which is then necessarily ergodic). It is isomorphic to a skew product over the von Neumann-Kakutani adding machine, or odometer (see below). Generalized Morse systems, that is, orbit closures of sequences like $0 \times 001 \times 001 \times 001 \times \dots$, are also isomorphic to skew products over compact group rotations.

5.4.2. *Chacon system.* This is the orbit closure of the sequence generated by the substitution $0 \rightarrow 0010, 1 \rightarrow 1$. It is uniquely ergodic and is one of the first systems shown to be weakly mixing but not strongly mixing. It is *prime* (has no nontrivial factors) (del Junco 1978), and in fact has *minimal self joinings* (del Junco-Rahe-Swanson 1980). It also has a nice description by means of cutting up the unit interval and stacking the pieces, using spacers (see below). This system has singular spectrum. It is not known whether or not its Cartesian square is loosely Bernoulli.

5.4.3. *Sturmian systems.* Take the orbit closure of the sequence $\omega_n = \chi_{[0,\beta)}(n\alpha)$, where α is irrational. This is a uniquely ergodic system that is isomorphic to rotation by α on the unit interval. Some of these systems ($\beta = 1 - \alpha$) have *minimal complexity* in the sense that the number of n -blocks grows as slowly as possible ($n + 1$).

5.4.4. *Toeplitz systems.* Fill in the blanks alternately with 0's and 1's. If done correctly, you get a uniquely ergodic system which is isomorphic to a rotation on a compact abelian group. (See 1998 notes on symbolic dynamics.)

5.4.5. *Sofic systems.* These are images of SFT's under continuous factor maps (finite codes, or block maps). They correspond to *regular languages*—languages whose words are recognizable by finite automata. These are the same as the languages defined by *regular expressions*—finite expressions built up from \emptyset (empty set), ϵ (empty word), $+$ (union of two languages=sets of finite words), \cdot (all concatenations of words from two languages), and $*$ (all finite concatenations of elements). They also have the characteristic property that the family of all *follower sets* of all blocks seen in the system is a finite family; similarly for *predecessor sets*. These are also generated by *phase-structure grammars* which are *linear*, in the sense that every production is either of the form $A \rightarrow Bw$ or $A \rightarrow w$, where A and B are variables and w is a string of terminals (symbols in the alphabet of the language).

[A *phase-structure grammar* consists of alphabets V of *variables* and A of *terminals*, a set of *productions*, which is finite set of pairs of words (α, w) , usually written $\alpha \rightarrow w$, of words on $V \cup A$, and a *start symbol* S . The associated language consists of all words on the alphabet A of terminals which can be made by starting with S and applying a finite sequence of productions.]

5.4.6. *Context-free systems.* These are generated by phase-structure grammars in which all productions are of the form $A \rightarrow w$, where A is a variable and w is a string of variables and terminals.

5.4.7. *Coded systems.* These are systems all of whose blocks are concatenations of some (finite or infinite) list of blocks. These are the same as the closures of increasing sequences of SFT's (Krieger). Alternatively, they are the closures of the images under finite edge-labelings of irreducible countable-state topological Markov chains. They need not be context-free. Squarefree languages are not coded, in fact do not contain any coded systems of positive entropy. (See Blanchard and Hansel.)

5.5. **Adic transformations.** A.M. Vershik has introduced a family of models, called adic transformations, into ergodic theory and dynamical systems. One begins with a graph which is arranged in levels, finitely many vertices on each level, with connections only from each level to the adjacent ones. The space X consists of the set of all infinite paths in this graph; it is a compact metric space in a natural way. Each level is supposed to be ordered from left to right (or alternatively we are given an order on the *edges* into each vertex), and then X is partially ordered as follows: x and y are comparable if they agree from some point on, in which case we say that $x < y$ if at the last level n where they are different, the vertex x_n of x is to the left of the vertex y_n of y . A map T is defined by letting Tx be the smallest y that is larger than x , if there is one. In nice situations, T is a homeomorphism after the deletion of perhaps countably many maximal and minimal elements and their orbits. This is a nice combinatorial way to present the *cutting and stacking* method of constructing m.p.t.'s. The adic viewpoint also leads to a nice version of the Jewett-Krieger Theorem:

Theorem 5.1 (Vershik). *Every ergodic measure-preserving transformation on a Lebesgue space is isomorphic to a uniquely ergodic adic transformation. Moreover, the isomorphism can be chosen so that a given countable dense invariant subalgebra of the measurable sets goes over into the algebra of cylinder sets.*

A vertex in the graph corresponds to a stack or column in the cutting and stacking picture.

A finite path in the graph corresponds to a level of a tower.

The order of the paths corresponds to the order up the stack.

Connections to the next level in the graph show how the stack is cut up and restacked.

Question 5.1. Is there a *relative version* of Vershik's Theorem? I.e., can factor maps be realized as continuous homomorphisms of uniquely ergodic systems, perhaps as natural maps between adic systems? (See Furstenberg's Math. Systems Th. 1967

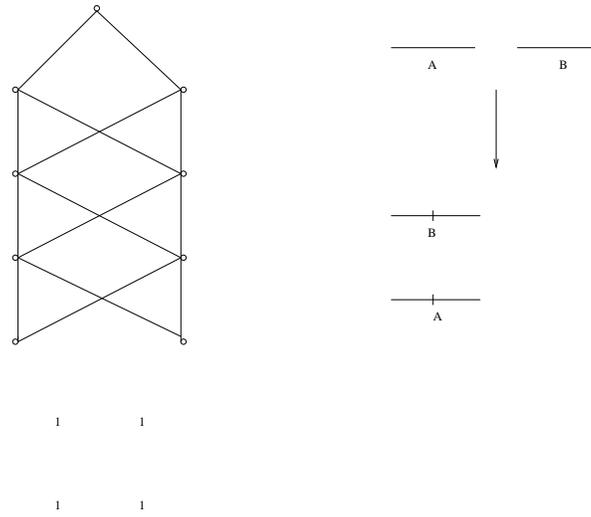


FIGURE 1. Odometer

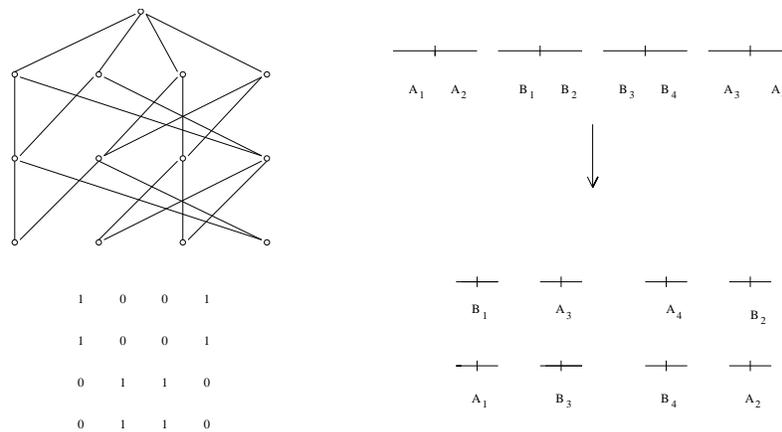


FIGURE 2. Prouhet-Thue-Morse

paper and 1981 book for some function-analytic approaches. Also Weiss's paper on realizability of factor diagrams by uniquely ergodic systems: Strictly ergodic models for dynamical systems, Bull. Amer. Math. Soc. 13 (1985), 143–146.)

5.6. Horocycle and geodesic flows.

5.7. Billiards.

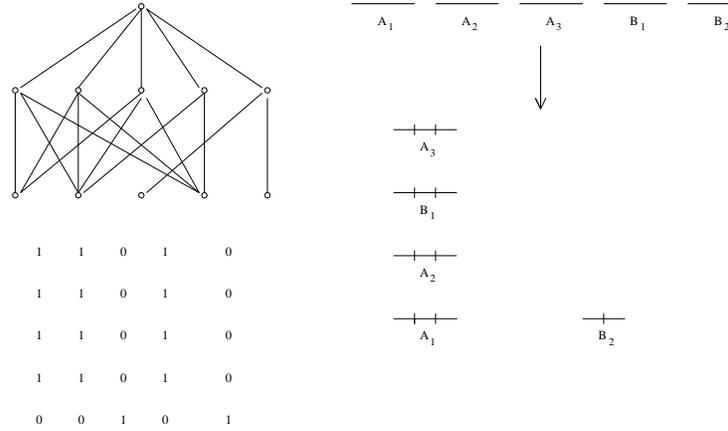


FIGURE 3. Chacon

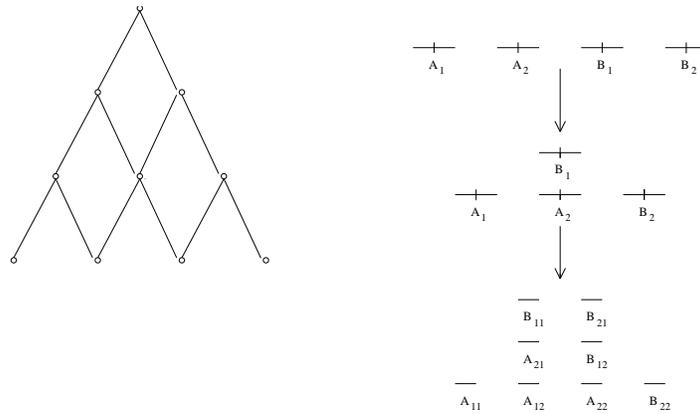


FIGURE 4. Pascal: $1^q 0^p 01 \dots \rightarrow 0^p 1^q 10 \dots, p, q \geq 0$.

5.8. Hamiltonian systems and other smooth examples. A general question, still open in regard to certain of its major aspects, is the problem of realization of abstract measure-preserving systems by means of diffeomorphisms on manifolds preserving measures given by smooth volume forms. Closely related is the identification of the *attractors* in smooth dynamical systems, going back at least to Smale's finding horseshoes in systems with homoclinic points. These questions are on the way to answering which ergodic-theoretic systems actually arise in the physical world, and therefore opening the possibility of studying chaotic and turbulent natural phenomena by ergodic-theoretic techniques.

5.9. Gaussian systems. Consider a real-valued stationary process $\{f_k : -\infty < i < \infty\}$ on a probability space (Ω, \mathcal{F}, P) . The process (and the associated measure-preserving system consisting of the shift and a shift-invariant measure on $\mathbb{R}^{\mathbb{Z}}$) is called *Gaussian* if any d of the f_k form an \mathbb{R}^d -valued Gaussian random variable on

Ω : with $E(f_k) = m$ for all k and

$$A_{ij} = \int_{\Omega} (f_{k_i} - m)(f_{k_j} - m) dP = C(k_i - k_j) \text{ for } i, j = 1, \dots, d,$$

for each Borel set $B \subset \mathbb{R}$,

$$P\{\omega : (f_{k_1}(\omega), \dots, f_{k_d}(\omega)) \in B\} = \frac{1}{2\pi^{d/2}\sqrt{\det A}} \int_B \exp\left[-\frac{1}{2}(x - (m, \dots, m))^{\text{tr}} A^{-1}(x - (m, \dots, m))\right] dx_1 \dots dx_d.$$

The function $C(k)$ is positive semidefinite and hence has an associated measure σ on $[0, 2\pi)$ such that

$$C(k) = \int_0^{2\pi} e^{ikt} d\sigma(t).$$

Theorem 5.2. *The Gaussian system is ergodic if and only if the “spectral measure” σ is continuous (i.e., nonatomic), in which case it is also weakly mixing. It is mixing if and only if $C(k) \rightarrow 0$ as $|k| \rightarrow \infty$. If σ is singular with respect to Lebesgue measure, then the entropy is 0; otherwise the entropy is infinite (Thierry de la Rue, thesis).*

5.10. Continued fraction map. This is the map $T : [0, 1] \rightarrow [0, 1]$ given by $Tx = 1/x \bmod 1$, and it corresponds to the shift $[0; a_1, a_2, \dots] \rightarrow [0; a_2, a_3, \dots]$ on the continued fraction expansions of points in the unit interval (a map on $\mathbb{N}^{\mathbb{N}}$). It preserves a unique finite measure equivalent to Lebesgue measure, the *Gauss measure* $dx/(\log 2)(1+x)$. It is Bernoulli with entropy $\pi^2/6 \log 2$ —the natural partition into intervals is a weak Bernoulli generator (Adler, 1975). By using the Ergodic Theorem, Khintchine and Lévy showed that

$$(a_1 \dots a_n)^{1/n} \rightarrow \prod_{k=1}^{\infty} \left[1 + \frac{1}{k^2 + 2k}\right]^{\log k / \log 2} \quad \text{a.e. as } n \rightarrow \infty;$$

$$\text{if } [0; a_1, \dots, a_n] = \frac{p_n}{q_n}, \text{ then } \frac{1}{n} \log q_n \rightarrow \frac{\pi^2}{12 \log 2} \text{ a.e.};$$

$$\frac{1}{n} \log \left|x - \frac{p_n(x)}{q_n(x)}\right| \rightarrow \frac{\pi^2}{6 \log 2} \text{ a.e.};$$

and if m is Lebesgue measure (or any equivalent measure) and μ is Gauss measure, then for each interval I , $m(T^{-n}I) \rightarrow \mu(I)$, in fact exponentially fast, with a best constant 0.30366... (Kuzmin, Lévy, Wirsing).

5.11. The Farey map. This is the map $U : [0, 1] \rightarrow [0, 1]$ given by $Ux = x/(1-x)$ if $0 \leq x \leq 1/2$, $Ux = (1-x)/x$ if $1/2 \leq x \leq 1$. It is ergodic for the σ -finite infinite measure dx/x (Rényi/Parry). It is also ergodic for the *Minkowski measure* $d?$, which gives it maximal entropy. This map corresponds to the shift on the *Farey tree* of rational numbers which provide the *intermediate convergents* (best one-sided) as well as the continued fraction (best two-sided) rational approximations to irrational numbers.

5.12. **f -expansions.** Generalizing the continued fraction map, let $f : [0, 1] \rightarrow [0, 1]$ and let $\{I_n\}$ be a finite or infinite partition of $[0, 1]$ into subintervals. We study the map f by coding itineraries with respect to the partition $\{I_n\}$. For many examples, absolutely continuous (with respect to Lebesgue measure) invariant measures can be found and their dynamical properties determined.

5.13. **β -shifts.** This is the special case of f -expansions when $f(x) = \beta x \bmod 1$ for some fixed $\beta > 1$. This map of the interval is called the β -transformation. With a proper choice of partition, it is represented by the shift on a certain subshift of the set of all sequences on the alphabet $\{0, 1, \dots, \lfloor \beta \rfloor\}$, called the β -shift. A point x is expanded as an infinite series in negative powers of β with coefficients from this set; $d_n(x) = \lfloor \beta f^n(x) \rfloor$. These were first studied by Bissinger, Rényi and Parry; there are good summaries by Bertrand-Mathis and Blanchard.

For $\beta = \frac{1+\sqrt{5}}{2}$, $1 \sim 110000\dots$

For $\beta = \frac{3}{2}$, $1 \sim 101000001\dots$ (not eventually periodic).

Every β -shift is coded.

The topological entropy of a β -shift is $\log \beta$. There is a unique measure of maximal entropy.

A β -shift is a subshift of finite type iff the β -expansion of 1 is finite. It's sofic iff the expansion of 1 is eventually periodic. If β is a Pisot-Vijayaragavhan number (algebraic integer all of whose conjugates have modulus less than 1), then the β -shift is sofic. If the β -shift is sofic, then β is a Perron number (algebraic integer of maximum modulus among its conjugates).

Theorem 5.3 (Parry). *Every strongly transitive (for every nonempty open set U , $\cup_{n>0} T^n U = X$) piecewise monotonic map on $[0, 1]$ is topologically conjugate to a β -transformation.*

6. ENTROPY

6.1. **Definition.** A finite (or sometimes countable) measurable partition

$$\alpha = \{A_1, \dots, A_r\}$$

of X is thought of as the set of possible outcomes of an experiment (performed at time 0) or as an alphabet of symbols used to form messages (the experiment could consist of receiving and reading one symbol). The *entropy* of the partition is

$$H(\alpha) = \sum_{A \in \alpha} -\mu(A) \log \mu(A) \quad (\text{the logs can be base } e, 2, \text{ or } r);$$

it represents the amount of information gained=amount of uncertainty removed when the experiment is performed or one symbol is received (averaged over all possible states of the world—the amount of information gained if the outcome is A (i.e., we learn to which cell of α the world actually belongs) is $-\log \mu(A)$). (Note

that this is large when $\mu(A)$ is small.) Notice that the information gained when we learn that an event A occurred is additive for independent events.

The partition

$$T^{-1}\alpha = \{T^{-1}A : A \in \alpha\}$$

represents performing the experiment α (or reading a symbol) at time 1, and $\alpha \vee T^{-1}\alpha \vee \dots \vee T^{-n+1}\alpha$ represents the result of n repetitions of the experiment (or the reception of a string of n symbols). Then $H(\alpha \vee T^{-1}\alpha \vee \dots \vee T^{-n+1}\alpha)/n$ is the average information gain per repetition (or per symbol received), and

$$h(\alpha, T) = \lim_{n \rightarrow \infty} \frac{1}{n} H(\alpha \vee T^{-1}\alpha \vee \dots \vee T^{-n+1}\alpha)$$

is the long-term time average of the information gained per unit time. (This limit exists because of the *subadditivity* of H : $H(\alpha \vee \beta) \leq H(\alpha) + H(\beta)$.)

The *entropy of the system* (X, \mathcal{B}, μ, T) is defined to be

$$h_\mu(T) = \sup_{\alpha} h(\alpha, T),$$

the maximum information per unit time available from any finite- (or countable-) state stationary process generated by the system.

Theorem 6.1 (Kolmogorov-Sinai). *If T has a finite generator α —a partition α such that the smallest σ -algebra that contains all $T^j\alpha, j \in \mathbb{Z}$, is \mathcal{B} —then $h_\mu(T) = h(\alpha, T)$. (Similarly if T has a countable generator with finite entropy.)*

Theorem 6.2. *If $\{\alpha_k\}$ is an increasing sequence of finite partitions which generates \mathcal{B} up to sets of measure 0, then $h(\alpha_k, T) \rightarrow h(T)$ as $k \rightarrow \infty$.*

6.2. Conditioning. For a countable measurable partition α and sub- σ -algebra \mathcal{F} of \mathcal{B} , we define the *conditional information function* of α given \mathcal{F} by

$$I_{\alpha|\mathcal{F}}(x) = - \sum_{A \in \alpha} \log \mu(A|\mathcal{F})(x) \chi_A(x);$$

this represents the information gained by performing the experiment α (if the world is in state x) after we already know for each member of \mathcal{F} whether or not it contains the point x . The *conditional entropy* of α given \mathcal{F} is

$$H(\alpha|\mathcal{F}) = \int_X I_{\alpha|\mathcal{F}}(x) d\mu(x);$$

this is the average over all possible states x of the information gained from the experiment α . When \mathcal{F} is the σ -algebra generated by a partition β , we often just write β in place of \mathcal{F} .

Proposition 6.3. (1) $H(\alpha \vee \beta|\mathcal{F}) = H(\alpha|\mathcal{F}) + H(\beta|\mathcal{B}(\alpha) \vee \mathcal{F})$.
 (2) $H(\alpha|\mathcal{F})$ is increasing in its first variable and decreasing in its second.

Theorem 6.4. *For any finite (or countable finite-entropy) partition α ,*

$$h(\alpha, T) = H(\alpha|\mathcal{B}(T^{-1}\alpha \vee T^{-2}\alpha \vee \dots)).$$

6.3. Examples.

6.3.1. *Bernoulli shifts.* $h = -\sum p_i \log p_i$. Consequently $\mathcal{B}(1/2, 1/2)$ is not isomorphic to $\mathcal{B}(1/3, 1/3, 1/3)$.

6.3.2. *Markov shifts.* $h = -\sum p_i \sum P_{ij} \log P_{ij}$.

6.3.3. *Discrete spectrum.* $h = 0$. (Similarly for *rigid* systems—ones for which there is a sequence $n_k \rightarrow \infty$ with $T^{n_k} f \rightarrow f$ for all $f \in L^2$.) Similarly for any system with a *one-sided generator*, for then $h(\alpha, T) = H(\alpha|\alpha_1^\infty) = H(\alpha|\mathcal{B}) = 0$. It's especially easy to see for an irrational rotation of the circle, for if α is the partition into two disjoint arcs, then α_0^n only has $2(n+1)$ sets in it.

6.3.4. *Products.* $h(T_1 \times T_2) = h(T_1) + h(T_2)$.

6.3.5. *Factors.* If $\pi : T \rightarrow S$, then $h(T) \geq h(S)$.

6.3.6. *Bounded-to-One Factors.* $h(T) = h(S)$. See Parry-Tuncel, p. 56.

6.3.7. *Skew products.* $h(T \times \{S_x\}) = h(T) + h_T(S)$. Here the action is $(x, y) \rightarrow (Tx, S_x y)$, with each S_x a m.p.t. on Y , and the second term is the *fiber entropy*

$$h_T(S) = \sup \left\{ \int_X H(\beta | S_x^{-1} \beta \vee S_x^{-1} S_{T_x}^{-1} \beta \vee \dots) d\mu(x) : \beta \text{ is a finite partition of } Y \right\}.$$

6.3.8. *Automorphism of the torus.* $h = \sum_{|\lambda_i| > 1} \log |\lambda_i|$ (the λ_i are the eigenvalues of the integer matrix with determinant ± 1).

6.3.9. *Pesin's Formula.* If $\mu \ll m$ (Lebesgue measure on the manifold), then

$$h_\mu(f) = \int \sum_{\lambda_k(x) > 0} q_k(x) \lambda_k(x) d\mu(x),$$

where the $\lambda_k(x)$ are the *Lyapunov exponents* and $q_k(x) = \dim(V_k(x) \setminus V_{k-1}(x))$ are the dimensions of the corresponding subspaces.

6.3.10. *Induced transformation (first-return map).* For $A \subset X$, $h(T_A) = h(T)/\mu(A)$.

6.3.11. *Finite rank ergodic. $h = 0$.*

Proof. Suppose rank = 1, let P be a partition into two sets (labels 0 and 1), let $\epsilon > 0$. Take a tower of height L with levels approximately P -constant (possible by rank 1; we could even take them P -constant) and $\mu(\text{junk}) < \epsilon$. Suppose we follow the orbit of a point $N \gg L$ steps; how many different P, N -names can we see? Except for a set of measure $< \epsilon$, we hit the junk $n \sim \epsilon N$ times. There are L starting places (levels of the tower); $C(N, n)$ places with uncertain choices of 0, 1; and 2^n ways to choose 0 or 1 for these places. So the sum of $\mu(A) \log \mu(A)$ over A in P_0^{n-1} is \leq the log of the number of names seen in the good part minus the log of $2^N (\epsilon/2^N) \log(\epsilon/2^N)$, and dividing by N gives

$$\frac{\log L}{N} + NH(\epsilon, 1 - \epsilon) + \frac{N\epsilon}{N} + \frac{\epsilon(-\log \epsilon + N)}{N} \sim 0.$$

Similarly for any finite partition P . Also for rank r —then we have to take care (not easily) about the different possible ways to switch columns when spilling over the top. \square

6.4. Ornstein's Isomorphism Theorem. Two Bernoulli shifts are isomorphic if and only if they have the same entropy.

6.5. Shannon-McMillan-Breiman Theorem. For a finite measurable partition α , and $x \in X$, let $\alpha(x)$ denote the member of α to which x belongs, and let $\alpha_0^{n-1} = \alpha \vee T^{-1}\alpha \vee \dots \vee T^{-n+1}\alpha$. If T is an ergodic m.p.t. on X , then

$$\frac{-\log \mu(\alpha_0^{n-1}(x))}{n} \rightarrow h(\alpha, T) \text{ a.e. and in } L^1.$$

6.6. Mañé, Brin-Katok Formula. If T is ergodic and

$$B(x, \delta, n) = \{y : d(T^j x, T^j y) < \delta \text{ for all } j = 1, \dots, n\},$$

then

$$h(T) = \sup_{\delta} \limsup_n -\frac{1}{n} \log \mu B(x, \delta, n).$$

6.7. Wyner-Ziv-Ornstein-Weiss entropy calculation algorithm. For a stationary ergodic sequence $\{\omega_1, \omega_2, \dots\}$ on a finite alphabet and $n \geq 1$, let $R_n(\omega) =$ the first place to the right of 1 at which the initial n -block of ω reappears (not overlapping its first appearance). Then

$$\frac{\log R_n(\omega)}{n} \rightarrow h \text{ a.e.}$$

This is also related to the *Lempel-Ziv parsing algorithm*, in which a comma is inserted in a string ω each time a block is completed (beginning at the preceding comma) which has not yet been seen in the sequence.

6.8. Topological entropy. Let X be a compact metric space and $T : X \rightarrow X$ a homeomorphism.

First definition (Adler-Konheim-McAndrew): For an open cover \mathcal{U} of X , let $N(\mathcal{U})$ denote the minimum number of elements in a subcover of \mathcal{U} , $H(\mathcal{U}) = \log N(\mathcal{U})$,

$$h(\mathcal{U}, T) = \lim_{n \rightarrow \infty} \frac{1}{n} H(\mathcal{U} \vee T^{-1}\mathcal{U} \vee \dots \vee T^{-n+1}\mathcal{U}),$$

and

$$h(T) = \sup_{\mathcal{U}} h(\mathcal{U}, T).$$

Second definition (Bowen): For $n \in \mathbb{N}$ and $\epsilon > 0$, a subset $A \subset X$ is called n, ϵ -separated if given $a, b \in A$ with $a \neq b$, there is $k \in \{0, \dots, n-1\}$ with $d(T^k a, T^k b) \geq \epsilon$. We let $S(n, \epsilon)$ denote the *maximum* possible cardinality of an n, ϵ -separated set. Then

$$h(T) = \lim_{\epsilon \rightarrow 0^+} \limsup_{n \rightarrow \infty} \frac{1}{n} \log S(n, \epsilon).$$

Third definition (Bowen): For $n \in \mathbb{N}$ and $\epsilon > 0$, a subset $A \subset X$ is called n, ϵ -spanning if given $x \in X$ there is $a \in A$ with $d(T^k a, T^k x) \leq \epsilon$ for all $k = 0, \dots, n-1$. We let $R(n, \epsilon)$ denote the *minimum* possible cardinality of an n, ϵ -spanning set. Then

$$h(T) = \lim_{\epsilon \rightarrow 0^+} \limsup_{n \rightarrow \infty} \frac{1}{n} \log R(n, \epsilon).$$

Example 6.1. If (X, T) is a subshift ($X =$ a closed shift-invariant subset of the set of all doubly infinite sequences on a finite alphabet, $T = \sigma =$ shift transformation), then

$$h(\sigma) = \lim_{n \rightarrow \infty} \frac{\log(\text{number of } n\text{-blocks seen in sequences in } X)}{n}.$$

Theorem 6.5 (“Variational Principle”). $h(T) = \sup\{h_\mu(T) : \mu \text{ is an invariant (ergodic) Borel measure on } X\}$.

6.9. Some information theory. A *source* is a finite-state stationary stochastic process, which we model by a shift-invariant measure μ on the space $X = A^{\mathbb{Z}}$ of all sequences on a finite alphabet A .

A *channel* has, in addition to a source, an output alphabet B , the set $Y = B^{\mathbb{Z}}$ of all sequences on that alphabet, and a family of measures $\{\nu_x : x \in X\}$ on Y , with $\nu_{\sigma x} = \sigma \nu_x$, which determine an *input-output measure* λ on $X \times Y$ by

$$\lambda(E \times F) = \int_E \nu_x(F) d\mu(x).$$

This is supposed to give the probability that the output signal is in F , given that the input signal came from the set E .

The *output measure* γ on Y is given by $\gamma(F) = \lambda(\pi_Y^{-1}F)$. The idea is that if messages are input to the channel with statistics described by μ , then the output,

with the output measure, is regarded as a new source, the result of connecting the initial source to the channel and seeing what comes across.

The *mutual information* of two partitions α and β is defined to be

$$H(\alpha; \beta) = H(\alpha) - H(\alpha|\beta) = H(\beta) - H(\beta|\alpha).$$

We will apply this to partitions of $X \times Y$, α being the lift of the *time-0 partition* of X , β a lift of the time-0 partition of Y . Then the *information transmission rate* of the channel, for a choice μ of the input measure, is

$$\begin{aligned} (1) \quad R(\mu) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(\alpha_0^n; \beta_0^n) \\ (2) \quad &= \lim_{n \rightarrow \infty} \frac{1}{n} [H(\alpha_0^n) - H(\alpha_0^n | \beta_0^n)] \\ (3) \quad &= \lim_{n \rightarrow \infty} \frac{1}{n} [H(\beta_0^n) - H(\beta_0^n | \alpha_0^n)] \\ (4) \quad &= h_\mu(\sigma_X) + h_\nu(\sigma_Y) - h_\lambda(\sigma_{X \times Y}), \end{aligned}$$

where the subscripts on σ indicate shifts acting on different spaces, X and Y being thought of as $X \times Y$ with smaller σ -algebras. The transmission rate is thus the information in the input (per unit time) minus the uncertainty about what the input was supposed to be when we have received the output and are looking at it; it is also the information per symbol coming out of the channel, minus the information (or uncertainty) that is extra beyond the information that was put in, for example the extra uncertainty due to the presence of noise.

The *capacity* of the channel is defined to be

$$C = \sup\{R(\mu)\},$$

the supremum being taken over all stationary ergodic sources that can be attached to the channel.

Shannon's theorems give *operational meaning* to capacity—they tell what you can do by means of coding to get information to go nicely across a channel. The strategy is first to *compress* information by recoding to remove redundancy. Then we *add redundancy* in a controlled way to protect against noise and errors.

Theorem 6.6 (Source Coding Theorem). *The minimum mean number of bits per symbol required to encode an ergodic source is the entropy of the source.*

Explanation: The entropy of the source is by definition (if logarithms are taken base 2) the mean amount of information that it emits, in bits per symbol. If we construct a code by sending blocks on A to blocks on $\{0, 1\}$ in a one-to-one way (presumably sending more frequent blocks to short blocks), then the expected value of the ratio of block lengths is $\geq h(\mu)$ (if we have an isomorphism).

Conversely, by the Shannon-McMillan-Breiman Theorem, given $\epsilon > 0$ we can find $n/k > h(\mu) - \epsilon$ and a coding of k -blocks on A to n -blocks on $\{0, 1\}$ which is one-to-one on a set of input sequences of measure $\geq 1 - \epsilon$.

Another precise and a.e. formulation of such a statement has been given by H. White. A *coding scheme* is a map Φ defined on a subset of $A^* =$ the set of all finite-length words on the alphabet A taking values in $\{0, 1\}^*$ such that $\Phi|A^n$ is one-to-one for each n . The idea is that we fix an n and code n -blocks on A to blocks of 0's and 1's, possibly of varying lengths, in a decodable (invertible) way. Then we vary n and see whether we can decrease the number of bits per symbol required (taking the ratio of the length of the image of an n -block to n).

Proposition 6.7 (White). *For any coding scheme Φ as above and any ergodic shift-invariant measure μ on $X = A^{\mathbb{Z}}$,*

$$\liminf_{n \rightarrow \infty} \frac{l(\Phi(x_1^n))}{n} \geq h_\mu(\sigma) \text{ a.e. (with } l(\Phi(w)) = \infty \text{ if } \Phi \text{ is not defined at } w).$$

Theorem 6.8 (Feinstein's Lemma). *For a "good" channel, given $\epsilon > 0$ there is n_0 such that for $n \geq n_0$ we can find $2^{n(C-\epsilon)}$ distinguishable code words $\{u_i\}$ in A^n : there are disjoint sets V_i of n - blocks in B^n such that for each i , $\nu_{u_i}(V_i) = \lambda\{u_i \text{ in, given } V_i \text{ out}\} > 1 - \epsilon$.*

This is proved by *random coding*: the bad blocks are counted, and by comparing exponential rates of growth it is shown that their relative number tends to 0 as $n \rightarrow \infty$.

Theorem 6.9 (First Shannon Theorem). *For a "good" channel, if $h(\mu) < C$, then for every $\epsilon > 0$ there is n_0 such that if $n \geq n_0$ there is a block code $A^n \rightarrow B^{n+m}$ and a decoder such that when the source is sent across the channel and decoded, the average probability of symbol error is $< \epsilon$.*

This is proved by using SMB and Feinstein. First we recode the source into $\sim 2^{nh}$ n -blocks, then we assign to each of these a distinguishable block. Thus this combines the ideas of Huffman (compression) and Hamming (error-correcting) coding.

Theorem 6.10 (Second Shannon Theorem). *For a "good" channel, if $h(\mu) < C$, then given $\epsilon > 0$ there is n_0 such that if $n \geq n_0$ then there are a block code and decoder as above with $R(\tilde{\mu}) > h(\tilde{\mu}) - \epsilon$ (where $\tilde{\mu}$ is the input measure to the channel determined by the original input measure μ after the encoding).*

These two conclusions can be achieved simultaneously, with a single code; this assertion might be called the *Channel Coding Theorem*.

So what is a "good" channel? It is sufficient that it be stationary, *non-anticipating* ($\nu_x\{y_0 = i\}$ depends only on $\dots x_{-2}x_{-1}x_0$), *finite-memory* (in fact $\nu_x\{y_0 = i\}$ depends only on $x_{-k} \dots x_{-2}x_{-1}x_0$), and *Nakamura ergodic*: the compound source (product source on $X \times Y$) it forms with each ergodic source should be ergodic; equivalently,

$$\frac{1}{n} \sum_{k=0}^{n-1} \int_{\sigma^k U \cap V} |\nu_x(\sigma^k W \cap Z) - \nu_x(\sigma^k W)\nu_x(Z)| d\mu(x) \rightarrow 0$$

for all cylinder sets $U, V \subset X$ and $W, Z \subset Y$. There are of course other conditions under which these theorems or variations hold.

6.10. Complexity. The (*Kolmogorov*) *complexity* $K(w)$ of a finite sequence w on a finite alphabet is defined to be the length of the shortest program that when input to a fixed universal Turing machine produces output w (or at least a coding of w by a block of 0's and 1's). For a topological dynamical system (X, T) and open cover $\mathcal{U} = \{U_0, \dots, U_{r-1}\}$ of X , for $x \in X$ and $n \geq 1$ let $\mathcal{C}(x, n) =$ the set of n -blocks w on $\{0, \dots, r-1\}$ such that $T^j x \in U_{w_j}$, $j = 1, \dots, n$. Then we define the upper and lower *complexity of the orbit* of a point $x \in X$ to be

$$\sup K(x, T) = \sup_{\mathcal{U}} \limsup_{n \rightarrow \infty} \min \left\{ \frac{K(w)}{n} : w \in \mathcal{C}(x, n) \right\}$$

and

$$\inf K(x, T) = \sup_{\mathcal{U}} \liminf_{n \rightarrow \infty} \min \left\{ \frac{K(w)}{n} : w \in \mathcal{C}(x, n) \right\}$$

Theorem 6.11 (Brudno, White). *If μ is an ergodic invariant measure on (X, T) , then*

$$\sup K(x, T) = \inf K(x, T) = h_\mu(X, T) \text{ a.e. } d\mu(x).$$

6.11. Lyapunov exponents, Hausdorff dimension, volume growth—and entropy. Let X be a compact manifold, $T : X \rightarrow X$ a \mathcal{C}^2 diffeomorphism, and μ an ergodic T -invariant Borel probability measure on X .

The *Hausdorff dimension* of μ is $HD(\mu) = \inf\{HD(A) : A \subset X, \mu(A) = 1\}$. Recall that $HD(A)$, the Hausdorff dimension of a set A , is the value of t at which the t -dimensional Hausdorff measure H^t jumps from ∞ to 0, where

$$H^t(A) = \lim_{\delta \rightarrow 0} \inf \left\{ \sum (\text{diam}(U_j))^t : U_j \text{ closed balls covering } A, \text{diam}(U_j) < \delta \right\}.$$

The *Lyapunov exponents* of T are real numbers $\lambda_1 < \lambda_2 < \dots < \lambda_r$ for which for each $x \in X$ there is a DT -invariant splitting of the tangent space at x , $\mathcal{T}_x X = E_1^x \oplus \dots \oplus E_s^x$, depending measurably on x , such that for a.e. x

$$\frac{1}{n} \log \|D_x(T^n)v\| \rightarrow \lambda_i \text{ for all } v \in E_i^x.$$

Theorem 6.12 (Pesin's Formula). *If T is ergodic, $\mu \ll m$, and $\dim(E_i) = n_i$, then $h_\mu(T) = \sum_{\lambda_i > 0} n_i \lambda_i$.*

Theorem 6.13 (Young). *If X is a compact surface and T is a \mathcal{C}^2 diffeomorphism of X with $\lambda_1 < 0 < \lambda_2$, then*

$$HD(\mu) = h_\mu(T) \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right).$$

This is proved by showing that in this case the *fractal dimension* of μ ,

$$\lim_{\delta \rightarrow 0^+} \frac{\log \mu(B_\delta(x))}{\log \delta},$$

exists for a.e. x and hence equals $HD(\mu)$, and then applying the *Mañé-Brin-Katok* result that (in a slight variation of the version stated above) for a.e. x ,

$$\lim_{\epsilon \rightarrow 0^+} \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mu \{y : d(T^k x, T^k y) \leq \epsilon \text{ for all } k = 0, 1, \dots, n\} = h_\mu(T).$$

There are many further developments along this line due to Ledrappier-Young, Thieullen, Pesin, Newhouse, Yomdin, et al.