INTRODUCTION TO ERGODIC THEORY

LECTURES BY MARYAM MIRZAKHANI
NOTES BY TONY FENG

CONTENTS

1. Disclaimer 2
2. Introduction 3
2.1. Overview 3
2.2. Spectral invariants 4
2.3. Entropy 5
2.4. Examples 5
3. Mean Ergodic Theorems 8
3.1. Preliminaries 8
3.2. Poincaré Recurrence Theorem 8
3.3. Mean ergodic theorems 9
3.4. Some remarks on the Mean Ergodic Theorem 11
3.5. A generalization 13
4. Ergodic Transformations 14
4.1. Ergodicity 14
4.2. Ergodicity via Fourier analysis 15
4.3. Toral endormophisms 16
4.4. Bernoulli Shifts 17
5. Mixing 19
5.1. Mixing transformations 19
5.2. Weakly mixing transformations 19
5.3. Spectral perspective 20
5.4. Hyperbolic toral automorphism is mixing 21
6. Pointwise Ergodic Theorems 23
6.1. The Radon-Nikodym Theorem 23
6.2. Expectation 24
6.3. Birkhoff's Ergodic Theorem 25
6.4. Some generalizations 28
6.5. Applications 29
7. Topological Dynamics 31
7.1. The space of T-invariant measures 31
7.2. The ergodic decomposition theorem 33
8. Unique ergodicity 34
8.1. Equidistribution 34

8.2. Examples 36
8.3. Minimality 38
9. Spectral Methods 40
9.1. Spectral isomorphisms 40
9.2. Ergodic spectra 40
9.3. Fourier analysis 41
10. Entropy 42
10.1. Motivation 42
10.2. Partition information 42
10.3. Definition of entropy 45
10.4. Properties of Entropy 46
10.5. Sinai's generator theorem 49
10.6. Examples 50
11. Measures of maximal entropy 51
11.1. Examples 51
12. Solutions to Selected Exercises 53

1. Disclaimer

These are notes that I "live-TeXed" during a course offered by Maryam Mirzakhani at Stanford in the fall of 2014. I have tried to edit the notes somewhat, but there are undoubtedly still errors and typos, for which I of course take full responsibility.

Only about 80% of the lectures is contained here; some of the remaining classes I missed, and some parts of the notes towards the end were too incoherent to include. It is possible (but unlikely) that I will come back and patch those parts at some point in the future.

2. Introduction

2.1. **Overview.** The overarching goal is to understand measurable transformations of a measure space $(X,\mu,\mathscr{B})$. Here μ is usually a probability measure on X and $\mathscr{B}$ is the σ-algebra of measurable subsets.

Definition 2.1. We will consider a transformation $T\colon X\to X$ *preserves* μ if for all $\alpha\in\mathscr{B}$ we have

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

In particular, we require $T^{-1}(\mathscr{B})\subset\mathscr{B}$ for this to make sense.

Remark 2.2. It is *not* necessarily the case that $\mu(\alpha)=\mu(T\alpha)$, or even that $T(\mathscr{B})\subset\mathscr{B}$.

We are interested in the following kinds of questions concerning this setup.

> Can we understand the *orbits* of T on X?

More precisely, for $x\in X$ the orbit of T on x is

$$\{x,Tx,T^2x,\dots\}.$$

Natural questions one might ask: is it *periodic*? Is it *dense*? Is it *equidistributed* (whatever that means)?

- A basic example that already leads to interesting questions is $X=S^1$ with $\mu=$ the Lebesgue measure. One measure-preserving transformation is $Tx=2x$, the "doubling map" (although it is not immediately obvious that this is measure-preserving).
- Another basic but interesting example is the "rotation operator" $R_\alpha(\theta)=\theta+\alpha$. Viewing $S^1=[0,1]$ with endpoints identified, the orbit is the "distribution" of $\{n\alpha\}=n\alpha-[n\alpha]$ for $\alpha\in\mathbb{R}$. For this simple problem it is easy to show that the qualitative behavior of the orbit depends on the rationality of α.

  There are already subtle extensions of this problem: what about the distribution of $\{n^2\alpha\},\{n^3\alpha\}$, or more generally $\{p(n)\alpha\}$ where $p(n)$ is some polynomial? We will see techniques that can resolve these problems.

> Given (X,μ) that are sufficiently nice, can we "classify" all μ-preserving transformations $T\colon X\to X$? Can we find invariants that distinguish them?

A tricky thing about this is that since we are considering measure spaces, we can throw out sets of measure zero. This means that topological intuition is not so useful here.

Remark 2.3. If μ is a *regular* measure, e.g. if $X\subset\overline{X}$ where $\overline{X}$ is metrizable, and $\mathscr{B}$ is the Borel σ-algebra, then $(X,\mu,\mathscr{B})$ turns out to be isomorphic to a "standard probability space", which is a disjoint union of intervals with Lebesgue measure and discrete spaces. In particular, we see that topological ideas like dimension, etc. are useless for distinguishing probability spaces.

So then what kinds of invariants *can* you use? We will discuss two flavors.

2.2. **Spectral invariants.** A simpler class of invariants are the "spectral invariants," which are qualitative features reflected in the "spectral theory" of T (we will explain what we mean by this later).

2.2.1. *Ergodicity.* The simplest incarnation is *irreducibility*. Morally, μ is reducible if it can be decomposed as $\mu = \mu_1 + \mu_2$ where μ_1, μ_2 are T-invariant measures that are *singular* with respect to each other (which rules out "trivial" decompositions like $\mu = \frac{1}{2}\mu + \frac{1}{2}\mu$). If μ is irreducible then it is called *ergodic*.

Remark 2.4. This is one of several possible definitions of ergodicity. A different one is that if A is T-invariant and measurable, then $\mu(A) = 0$ or 1 (here μ is a probability measure).

Theorem 2.5 (Ergodic Decomposition Theorem)**.** *If $(X, \mu, \mathscr{B})$ is a regular measure space and μ is T-invariant, then there exists $(Y, \nu, \mathscr{C})$ and a map*

$$Y \to \{\text{space of } T\text{-invariant measures on } X\}$$

denoted by $y \mapsto \mu_y$ such that

$$\mu = \int_Y \mu_y \, d\nu$$

and μ_y is ergodic for ν-almost every $y \in Y$.

Definition 2.6. We say that $(X, T) \cong (Y, S)$ if there exists a full-measure subset $X' \subset X$ which is T-invariant, and a full-measure subset $Y' \subset Y$ which is S-invariant, and a map $\phi : X' \to Y'$ such that the diagram commutes

$$\begin{tikzcd} X' \arrow[r] \arrow[d, "T"] & Y' \arrow[d, "S"] \\ X' \arrow[r] & Y' \end{tikzcd}$$

and ϕ has an inverse satisfying the obvious analogous properties.

The point is that we can throw away a set of measure 0 and get the natural notion of isomorphism. In particular, an ergodic transformation will not be isomorphic to a non-ergodic transformation.

2.2.2. *Mixing.* Another such invariant is *mixing*, which says that if A, B are measurable then

$$\lim_{n\to\infty} \mu(A \cap T^{-n} B) = \mu(A)\mu(B).$$

We don't want to dwell on the formal definitions now, but it turns out that this is stronger than ergodicity. There are variants on this: weakly mixing, strongly mixing, exponentially mixing, etc.

2.2.3. *"Spectral" explained.* Why do we call these spectral invariants? Because they are related to the action of T on $L^2_\mu(X)$. That is, we have the Hilbert space of square-integrable functions on X equipped with inner product

$$\langle f_1, f_2\rangle = \int_X f_1 \overline{f_2}\, d\mu.$$

The map $T : X \to X$ induces by pullback a (unitary) operator

$$u_T : L^2_\mu(X) \to L^2_\mu(X).$$

The condition of mixing can be interpreted in terms of the spectral theory of the operator u_T. In fact, we will see that you can distinguish between rotations R_α, R_β based on their spectral properties. However, many non-equivalent operators have the same action on $L^2_\mu(X)$, so we can't distinguish them in this way.

2.3. **Entropy.** To distinguish some operators we will require a different kind of invariant, which is more refined in the sense that it does not depend only on the spectral properties.

Example 2.7. We'll now discuss a family of examples, the *hyperbolic toral automorphisms*. Let $n = 2$ for concreteness, although you can do this for $n \geq 2$ too. Let $A \in \mathrm{SL}_2(\mathbb{Z})$ be a matrix having no eigenvalue of modulus 1 (hence the "hyperbolic"). Then we have the natural action of A on $\mathbb{R}^2$, which sends integral points to integral points, and hence induces an action on $T^2 = \mathbb{R}^2/\mathbb{Z}^2$. Now A preserves the Lebesgue measure on $\mathbb{R}^2$ ($\det = 1$) and hence T^2. How can we distinguish between (T^2, A) and (T^2, B) for $A, B \in \mathrm{SL}_2(\mathbb{Z})$?

This is *extremely* difficult to do, even though they are non-isomorphic. The spectral invariants ergodicity and mixing are not enough. The only way we know is to use a very powerful invariant called *entropy*, which quantifies "how complicated" the system is. This roughly measures the growth of the number of periodic points, although periodic points aren't useful here (there are only countably many, and we can throw away sets of measure 0).

In many examples, this is the only way we know how to show that the measure spaces are not the same. The second part of the course will deal with entropy, how to define and calculate it. In the setting where X is a compact hyperbolic space and T is continuoous, there are some corollaries on counting periodic points and behavior of "long" periodic points.

2.4. **Examples.** We'll now give some examples of measure-preserving transformations that will crop up repeatedly in the course.

(1) The *rotation* operator $R_\alpha : S^1 \to S^1$ sending $\theta \mapsto \theta + \alpha$ preserves the Lebesgue measure. Its orbits are related to the distribution of $\{n\alpha\}, \{n^2\alpha\}, \dots$ in $[0,1]$.

(2) The *doubling* map T_2 (or more generally T_3, T_m) on S^1 sending $z \mapsto z^2$ (respectively z^3, z^m). You can check that in fact $\mu(A) = \mu(T^{-1}A)$ if μ is the Lebesgue measure (if A is an interval, then $T^{-1}A$ consists of two components each having half the length of A).

So then one can ask when are (S^1, T_2, μ) and (S^1, T_m, μ) are the same? It turns out that they are different if $m \neq 2$ (which we can show using entropy), but this leads into a big open question. The Lebesgue measure is invariant under T_m. There are "many" measures invariant under T_k (the Lebesgue is the "nicest" one) for any *particular* k.

Conjecture 2.8. *If μ is a probability measure invariant under T_2 and T_3 then it is either supported on a finite set or Lebesgue.*

This is a huge, difficult open problem. In contrast:

Theorem 2.9 (Furstenberg)**.** *A closed subset of S^1 which is invariant under T_2 or T_3 is either S^1 or a finite set.*

This illustrates the contrast between topology and measure theory. Sometimes something that is hard in one world is easy in another.

It is known in some general situations that if μ has positive entropy under certain maps, like T_2 and T_3, then it is Lebesgue.

(3) The *Gauss map* $T : [0,1] \to [0,1]$ defined by

$$T(x) = \begin{cases} \frac{1}{x} \mod 1 & x \neq 0, \\ 0 & x = 0. \end{cases}$$

There is a measure μ on $[0,1]$ invariant under T (not the Lebesgue), which has the form

$$\mu(B) = \frac{1}{\log 2} \int_B \frac{dx}{1+x}.$$

It turns out that $\mu(T^{-1}B) = \mu(B)$.

Let $x \in \mathbb{R}$ have the continued fraction expansion

$$x = a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}$$

We will prove the following rather remarkable result.

Theorem 2.10. *For almost every x, the frequency of k among the continued fraction expansion of x is*

$$\frac{1}{\log k} \log\left(\frac{(k+1)^2}{k(k+2)}\right)$$

and

$$\lim_{n\to\infty} (a_1 \dots a_n)^{1/n} = \prod_{k=1}^{\infty} \left(1 + \frac{1}{k^2+2k}\right)^{\frac{\log k}{\log 2}}.$$

The key input to prove these sorts of statements is that T is measure-preserving and ergodic.

(4) *Geodesic flow on hyperbolic surfaces.* $X = \mathbb{H}^2/\Gamma$ is a hyperbolic surface, inheriting the hyperbolic structure from $(\mathbb{H}^2, ds = \frac{|dz|}{\mathrm{Im}(z)})$. The geodesics are the (semi)circles perpendicular to boundary, including the straight lines. Let T^1X denote the unit tangent bundle to X. You can consider the map $\mathcal{T}^\ell \colon T^1X \to T^1X$ (the map on the unit tangent bundle) sending $v \mapsto g_\ell v$ (taking a tangent vector to its position after flowing for time ℓ).

So for each ℓ, we have a triple $(T^1X, \mathscr{T}^\ell, \mu)$, where μ is the Lebesgue measure. It is not clear if $(T^1X, \mathscr{T}^\ell, \mu) \cong (T^1x, \mathscr{T}^{\ell'}, \mu)$ unless $\ell = \pm\ell'$. In fact the answer is that they are not equivalent, and you can prove this using entropy.

The significance of the question is that periodic points for this transformation are related to closed geodesics on X.

3. Mean Ergodic Theorems

3.1. **Preliminaries.** We are interested in understanding the geometry of (X, T, μ) where $T: X \to X$ preserves the (probability) measure μ, i.e. for all $B \in \mathscr{B}$ (the σ-algebra of measurable sets) we have $\mu(T^{-1}(B)) = \mu(B)$.

Recall how we discussed that if T is measure-preserving and $f: X \to \mathbb{R}$ (or $\mathbb{C}$) is some measurable function, we can pull back via T to get another function

$$u_T(f) := f \circ T: X \to \mathbb{R} \text{ (or } \mathbb{C}).$$

Defining the Banach spaces $L^\infty(X,\mu)$, $L^1(X,\mu)$, or $L^p(X,\mu)$ as usual, we see that $T \rightsquigarrow u_T$, an operator on the corresponding Banach space. Moreover, this is an *isometry*.

Lemma 3.1. *The measure μ on X is T-invariant if and only if for all $f \in L^1(X,\mu)$ we have:*

$$\int f\, d\mu = \int f \circ T\, d\mu \tag{1}$$

Proof. One direction is trivial: assuming (1) for all (almost everywhere) bounded test functions f, we can set $f = \chi_B$ where B is measurable and we immediately obtain that $\mu(B) = \mu(T^{-1}(B))$.

Conversely, suppose that we know (1) for all χ_B where B is measurable. A basic fact from measure theory is that there exists a sequence $f_n \uparrow f$ almost everywhere, where f_n is a simple function: a finite linear combination of indicator functions. By dominated convergence

$$\lim_{n\to\infty} \int f_n \to \int f.$$

For each f_n, we have

$$\int f_n \circ T = \int f_n$$

by assumption, so we obtain the result in the limit. $\square$

3.2. **Poincaré Recurrence Theorem.** We now study the Poincaré Recurrence Theorem, which is a kind of "pigeonhole principle" for measure-preserving transformations. The idea is that if we consider the sequence of points $x, Tx, T^2x, \ldots$ then it should "return close to x" infinitely many times (hence "recurrence").

Definition 3.2. We say that (X, T, μ) is a *measure-preserving system* if (X, μ) is a measure space and $T: X \to X$ preserves μ.

Let (X, T, μ) be a measure-preserving system, where μ is actually a probability measure. For a point $x \in X$, we consider the orbit $x, T(x), T^2(x)$, etc. We want to show that most points come back very close to themselves many times.

Theorem 3.3 (Poincaré)**.** *For any measurable subset $E \subset X$, for almost every $x \in E$ there exist $n_1 < n_2 < \ldots$ such that $T^{n_1}(x), T^{n_2}(x), \ldots \subset E$.*

An interesting question is what can we say about the sequence $n_1(x) < n_2(x) < \ldots$. The theorem says that the sequence is infinite, but we might want to quantify whether or not the recurrence happens "often." In fact, it does: for "nice" maps T, $n_i(x) \sim \alpha i$. Essentially, there is a finite expected time for recurrence to occur.

Proof. The idea is to try to bound the measure of the set of points that don't come back to E. Let

$$B = \{x \in E : T^n(x) \notin E \forall n \geq 1\}.$$

First one has to check that this is measurable:

$$B = E \cap T^{-1}(X - E) \cap T^{-2}(X - E) \cap \ldots \cap T^{-k}(X - E) \cap \ldots$$

This is an (admittedly infinite) intersection of measurable sets, hence measurable.

We claim that $B, T^{-1}(B), \ldots, T^{-k}(B)$ are disjoint. Indeed, any $y \in T^{-1}(B)$ satisfies $T(y) \in E$ so $y \notin B$. Since $\mu(B) = \mu(T^{-1}(B)) = \ldots$, and we are dealing with a probability measure, we immediately see that $\mu(B) = 0$. If $x \in E$ is *not* recurrent, then $x \in T^{-N}(B)$ for some N, so we are done.

$\square$

Question: What can we say about $\mu(E \cap T^{-n}E)$ if $\mu(E) > 0$? This is a measure of how "evenly" T propagates E around.

More generally one might ask about $\mu(E_1 \cap T^{-n}E_2)$ for *distinct* sets E_1 and E_2. However, note that $\mu(E_1 \cap T^{-n}E_2)$ could be zero for all n, e.g. if X is a union of two T-invariant pieces, so this does not admit an interesting answer without further refinements.

Exercise 3.4. Show that

$$\limsup_{n>0} \mu(E \cap T^{-n}E) \geq \mu(E)^2.$$

To put this in context, one can prove that for some general classes of T (irreducible, ergodic) one has that this is the average behavior in the sense that

$$\lim_{n\to\infty} \frac{1}{n} \sum \mu(E \cap T^{-1}E) = \mu(E)^2.$$

3.3. **Mean ergodic theorems.** We know move on to the *ergodic theorems.* If (X, T, μ) is a measure-preserving tuple, we can consider for any $f \in L^1(X, \mu)$ the sequence of functions

$$f(x), f(T(x)), \ldots, f(T^n(x)), \ldots.$$

In the special case where $f = \chi_E$, this describes the recurrence of x with respect to E. One might like to ask about the limit of this sequence as $n \to \infty$, but that is too ill-behaved. However, it is better behaved after averaging.

Theorem 3.5 (Pointwise Ergodic Theorem)**.** *With notation as above,*

$$\lim_{n\to\infty} \frac{f(x) + f(T(x)) + \ldots + f(T^n(x))}{n} =: f^*(x) \text{ exists for a.e. } x \in X.$$

Furthermore f^* is measurable and T-invariant, and

$$\int f^* \, d\mu = \int f \, d\mu.$$

If $f = \chi_E$, then this describes the asymptotics of recurrence of x with respect to E.

Remark 3.6. If T is ergodic, then f^* is constant almost everywhere and thus equal to $\int f\,d\mu$. If you think of T as describing the evolution of the system in time, then this means that the for *ergodic* transformations“the space average is equal to the time average.”

Theorem 3.7 (Mean Ergodic Theorem, von Neumann)**.** *If (X,T,μ) is a measure-preserving system, let $u_T : L^2(X,\mu) \to L^2(X,\mu)$ denote the induced map. Then*

$$\lim_{n\to\infty} \frac{f + u_T(f) + \ldots + u_{T^n}(f)}{n} =: P_T(f) \in L^2(X,\mu)$$

where $P_T(f)$ is the projection of f onto the subspace

$$I = \{g \in L^2(X,\mu) \colon u_T g = g\}.$$

Proof. The proof is straightforward up to some technical machinery. The key is to explicitly describe the orthogonal complement to I, so let

$$B = \langle u_T g - g : g \in L^2(X,\mu) \rangle.$$

We claim that $B^\perp = I$. Indeed, if $u_T f = f$ then

$$\langle f, u_T g - g\rangle = \langle f, u_T g\rangle - \langle f, g\rangle = \langle u_T f, u_T g\rangle - \langle f, g\rangle = 0.$$

This shows that $I \subset B^\perp$.

We then have to show that $B^\perp \subset I$. If $f \in B^\perp$ then by definition $\langle u_T g, f\rangle = \langle g, f\rangle$ for all g. Therefore,

$$\|u_T f - f\|_2 = \langle u_T f - f, u_T f - f\rangle = 2\|u_T f\|^2 - \langle f, u_T f\rangle - \langle u_T f, f\rangle = 0.$$

So we have established that $L^2(X,\mu) = I \oplus \overline{B}$. Recall that we want to show

$$\lim_{n\to\infty} \frac{f + u_T(f) + \ldots + u_{T^n}(f)}{n} =: P_T(f) \in L^2(X,\mu).$$

To do this, we proceed as follows.

(1) Check the result for $f \in I$ (which is obvious).
(2) Check it for $f = u_T g - g$ (also obvious, since it telescopes to $\frac{1}{N}\|u_T^N g - g\|_2$).
(3) The result follows for the whole space if we can show that the left hand side is “continuous in f, so that it vanishes on all of $\overline{B}$. Well, given $\epsilon > 0$ and $h \in \overline{B}$, we can find $h_i \in B$ such that $\|h - h'\|_2 < \epsilon$. Then for all sufficiently large N we have

$$\left\| \frac{1}{N} \sum_{n=1}^{N} u_T^n h' \right\|_2 < \epsilon$$

Therefore,

$$\begin{aligned}\left\| \frac{1}{N} \sum u_T^n h \right\| &\le \left\| \frac{1}{N} \sum u_T^n (h - h') \right\|_2 + \frac{1}{N} \left\| \sum u_T^n h' \right\| \\ &< 2\epsilon.\end{aligned}$$

□

Von Neumann's Mean Ergodic Theorem deals with convergence of operators in L^2. We would actually like to have a *pointwise* result, which unfortunately doesn't follow from the L^2 convergence. However, one can obtain L^1 or pointwise convergence results:

Denote by $A_n(f) := \frac{f+u_T(f)+\ldots+u_{T^n}(f)}{n}$ the nth partial sum.

Proposition 3.8 (L^1-convergence)**.** *If $f \in L^1(X,\mu)$ then*

$$\lim_{n\to\infty} A_n(f) = \widetilde{f} \text{ in } L^1(X,\mu).$$

What is this functon $\widetilde{f}$? In the L^2 case it was projection onto a certain subspace, but since L^1 is not a Hilbert space, we can't make sense of "projection operators" as we did before. It turns out that if $\widetilde{\mathscr{B}}$ denotes the σ-algebra of T-invariant measurable sets, then $\widetilde{f}$ is $E(f \mid \widetilde{\mathscr{B}})$. We will elaborate on this later.

Remark 3.9. The same argument implies that

$$\lim_{M-N\to\infty} \frac{1}{M-N} \sum_{n=N}^{M} u_T^n(f) \to P_T(f) \text{ in } L^2(X,\mu).$$

From this we deduce the following corollary.

Corollary 3.10. *Assuming that $\mu(X) < \infty$, show that if $\mu(B) > 0$ then the set $\{n \in \mathbb{N} : \mu(B \cap T^{-n}B) > 0\}$ (which is infinite by Poincaré's recurrence theorem) has the property that the set of gaps between recurrence are bounded.*

Proof. See the solution to Exercise 3.13. □

If T is invertible, then T^{-1} is measurable, and

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

This is because if T is invertible and g is T-invariant, then g is T^{-1}-invariant, so the projection operator is the same.

3.4. **Some remarks on the Mean Ergodic Theorem.** We established the Mean Ergodic Theorem for a measure-preserving system (X,μ,T):

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} u_T^n f = P_T f$$

where P_T is the projection onto the subspace of T-invariant functions in $L^2(X,\mu)$. This holds in general, even if $\mu(X) = \infty$, but one can encounter problems such as $P_T f$ vanishing almost everywhere, even if $\int_X f\,d\mu > 0$. As a simple example, suppose f is the indicator function of $[0,1]$ and T is translation by 1 on $\mathbb{R}$.

We would like to have

$$\int_X f\,d\mu = \int_X P_T f\,d\mu.$$

When restricting to a probability space, one has $\|\cdot\|_1 \le \|\cdot\|_2$ by Cauchy-Schwarz. Therefore, if $f_n \to f$ in L^2 then one has

$$\lim \int f_n \to \int f.$$

Since

$$\int_X u_T^n f \, d\mu = \int_X f \, d\mu$$

in a probability space we are indeed guaranteed that

$$\int_X P_T f = \int_X f.$$

Suppose $f_n \to f$ in L^2 and $g \in L^2(X,\mu)$. Then

$$\langle f_n, g\rangle \to \langle f, g\rangle.$$

For measurable sets A, B of (X,μ,T) we apply this with $f = \chi_A$ and $g = \chi_B$, and $f_n = A_n f$. By the Mean Ergodic Theorem,

$$\frac{1}{N}\sum_{n=1}^{N} \mu(T^{-n}A \cap B) \to \int_B P_T(\chi_A)\, d\mu.$$

One would like to use this to show that the orbits of A intersect B, but the right hand side could be 0.

However, if T is *ergodic* then by definition, the dimension of the space of T-invariant functions is 1 (i.e. just the constants), so the right hand side is some constant times $\mu(B)$. Now, in a probability space one has

$$\int f_n \, d\mu = \int_X f \, d\mu = \mu(A).$$

We have shown:

Theorem 3.11. *If T is ergodic, then*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mu(T^{-n}A \cap B) = \mu(A)\mu(B).$$

If T is not ergodic then one can still use the same idea to try and get something (the result won't be as strong, of course).

Exercise 3.12. Let (X,μ) be a probability space and $E \subset X$ a subset of positive measure. Assume $T\colon X \to X$ is an invertible transformation preserving μ. Show that there exists $x \in X$ such that $\{n \in \mathbb{Z} \mid T^n(x) \in E\}$ has positive upper density.

Exercise 3.13. Suppose (X,μ) is a probability space. For any measurable set B and $\epsilon > 0$, show that the set

$$\{k \in \mathbb{N} \mid \mu(T^{-k}B \cap B) \ge \mu(B)^2 - \epsilon\}$$

has bounded gaps.

3.5. **A generalization.** The key ingredient to this discussion is the mean ergodic theorem, whose proof is very easy: it's just basic functional analysis. What if we want to to study more complicated things like

$$\mu(A \cap T^{-n}A \cap T^{-2n}A \cap \ldots \cap T^{-kn}A)$$

if $\mu(A) > 0$? More generally, what about $\mu(A \cap T^{-p(n)}A)$ for some polynomial $p(t) \in \mathbb{Z}[t]$? More generally still, suppose you have commuting operators $T_1, \ldots, T_k$ and want to study

$$\lim_{N\to\infty} \frac{1}{N^k} \sum_{n_1,\ldots,n_k \leq N} u_{T_1}^{n_1}(f) \ldots u_{T_k}^{n_k}(f).$$

In fact, Host-Kra showed that this kind of limit does converge in $L^2(X,\mu)$. Recurrence statements for this setting were proved by Furstenberg and Katznedson, etc. They are significantly more challenging. We remark that these do *not* involve an assumption of ergodicity.

4. Ergodic Transformations

4.1. **Ergodicity.**

Definition 4.1. Suppose T is a measure-preserving map on $(X,\mu,\mathscr{B})$. Then T is *ergodic* if $B=T^{-1}B$ for $B\in\mathscr{B}$ implies $\mu(B)=0$ or $\mu(X-B)=0$.

Remark 4.2. This makes sense even when X has infinite measure.

This definition is supposed to capture the notion of irreducibility. Given any T-invariant measure μ, it is not clear how to obtain a measure μ' that is T-invariant and ergodic with respect to T. However, such measures do exist.

Proposition 4.3. *The following are equivalent:*

(1) T *is ergodic.*
(2) $\mu(T^{-1}B\Delta B)=0 \implies \mu(B)=0$ *or* $\mu(X-B)=0$.
(3) *(Assuming* $\mu(X)=1$*) For any* $A\in\mathscr{B}$, *if* $\mu(A)>0$ *then* $\mu(\bigcup T^{-n}A)=1$.
(4) *For any* $A,B\in\mathscr{B}$ *such that* $\mu(A)\mu(B)>0$ *there exists* n *such that* $\mu(T^{-n}A\cap B)>0$.
(5) *If* $f:X\to\mathbb{C}$ *is measurable, then* $f\circ T=f$ *almost everywhere implies that* f *is equal to a constant almost everywhere.*

Remark 4.4. Condition (3) generalizes the earlier remark that $\mu(T^{-N}A\cap A)>0$ for all T-invariant measures. Recall that we said the result could fail if X were a union of two disjoint T-invariant spaces. We will later prove that if T is ergodic then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mu(T^{-n}A\cap B)=\mu(A)\mu(B).$$

Remark 4.5. The definition makes sense for any group G acting on X.

Proof. Obviously (2) $\implies$ (1). For (1) $\implies$ (2), start with some B such that $\mu(B\Delta T^{-1}B)=0$. We want to make B into a T-invariant set somehow, so the most naïve thing to do is to throw in $T^{-1}(B)$. Of course, we then have to keep going, so we set

$$C=\bigcap_{N=0}^{\infty}\bigcup_{n=N}^{\infty}T^{-n}B.$$

Then evidently $T^{-1}(C)=C$, and

$$\mu(C)=\lim_{N\to\infty}\mu\left(\bigcup_{n=N}^{\infty}T^{-n}B\right)=\mu(B).$$

Next we show that (1) $\iff$ (3). For (1) $\implies$ (3) observe that $\bigcup_n T^{-n}(A)$ is T-invariant and has positive measure, so must be full measure. Conversely, if A is T-invariant with positive measure, then $\bigcup T^{-n}(A)=A$ has full measure.

To see that (4) $\implies$ (1), let $B\subset X$ be a T-invariant set. Then taking $A=X\setminus B$, we see that A is also T-invariant. If $\mu(B)\neq 0$ and $\mu(A)\neq 0$, then there exists n such that $\mu(T^{-n}B\cap A)=\mu(B\cap A)\neq 0$, clearly a contradiction. The other direction follows from the version of the Mean Ergodic Theorem in Theorem 3.11.

Finally, we establish that (1) $\iff$ (5). By taking f to be the characteristic function of an invariant set, we see that (5) $\implies$ (1). For (1) $\implies$ (5), let f be a function such that

$f \circ T = f$ then set $A_n^k = \{x : f(x) \in [\frac{k}{n}, \frac{k+1}{n}]\}$. Then $T^{-1}A_n^k = \{x \in X : f(Tx) \in [\frac{k}{n}, \frac{k+1}{n}\}$, but since $f(T(x)) = f(x)$ this is the same set as $\{x \in X : f(x) \in [\frac{k}{n}, \frac{k+1}{n}]\}$. Therefore,

$$\mu(T^{-1}A_n^k \Delta A_n^k) = 0.$$

This implies that A_n^k has full measure or zero measure for each n, k, and it follows that f is constant almost everywhere.

$\square$

Example 4.6. Here are some examples of ergodic and non-ergodic transformations.

(1) $R_\alpha : S^1 \to S^1$ is ergodic with respect to the Lebesgue measure if α is irrational, and not ergodic if α is rational.
(2) $T_2 : S^1 \to S^1$ with respect to the Lebesgue measure is ergodic.
(3) The map $T : S^1 \times S^1 \to S^1 \times S^1$ sending $(x, y) \mapsto (x + \alpha, y + \alpha)$ is not ergodic. For instance, the function $f(x, y) = e^{2\pi i(x-y)}$ is T-invariant but not constant.
(4) The map $S : T^k \to T^k$ sending

$$(x_1, \ldots, x_k) \mapsto (x_1 + \alpha, x_2 + x_1, x_3 + x_2, \ldots, x_k + x_{k-1})$$

is ergodic if α is irrational. That is not obvious, although it's easy to see that this is measure-preserving for the Lebesgue measure.

There is a nice trick due to Furstenberg to use this to show that $\{n^2\alpha\}, \{n^3\alpha\}, \ldots, \{n^k\alpha\}$ are dense in S^1 if α is irrational.

4.2. **Ergodicity via Fourier analysis.** One approach to ergodicity on S^1 is to use Fourier analysis on $L^2(X, \mu)$, and study the action of T on the Fourier coefficients. This leads to perhaps the simplest proofs, but unfortunately they do not generalize too well.

Example 4.7. Let's try applying this idea to the rotation operator R_α. For $f \in L^2(S^1)$ we write

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

What does it mean that $f(R_\alpha(t)) = f(t)$? The rotation sends $t \mapsto t + \alpha$, so by comparing Fourier coefficients we see

$$c_n = c_n e^{2\pi i n \alpha}.$$

If α is irrational then the factor $e^{2\pi i n \alpha}$ is never 1 unless $n = 0$, so all the c_n are 0 except the constant term, i.e. f is constant almost everywhere.

Example 4.8. Next let's see what happens with the doubling map. For $f \in L^2(S^1)$ we again write

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

If $f(t) = f(2t)$ then by comparison Fourier coefficients we have $c_{2n} = c_n$. This forces $c_k = 0$ if $k \neq 0$, since $c_k = c_{2k} = c_{4k} = \ldots \to 0$, a consequence of

$$||f||_{L^2} = \sum |c_n|^2 < \infty$$

Now that we are warmed up, let's prove that (4) from Example 4.6 is ergodic. For $f \in L^2(T^k)$, we have a Fourier expansion

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

Suppose $f(\vec{x}) = f(S(\vec{x}))$.

The trick is that we can write $\vec{n}\cdot S(\vec{x}) = n_1\alpha\vec{e_1} + S'(n)\cdot\vec{x}$ where $S'(\vec{n}) = (n_1+n_2, n_2+n_3, \ldots, n_{k-1}+n_k, n_k)$. The nice thing about S' is that it induces an *automorphism* of $\mathbb{Z}^k$, so

$$f(S(\vec{x})) = \sum c_{\vec{n}} \cdot e^{2\pi i n_1 \alpha} e^{2\pi i S'(\vec{n})x}.$$

We conclude that $c_{S'(\vec{n})} = e^{2\pi i \alpha n_1} c_{\vec{n}}$. In particular,

$$|c_{S'(\vec{n})}| = |c_{\vec{n}}|.$$

Now we claim that the sequence of vectors $\vec{n}, S'(\vec{n}), (S')^{\circ k}(\vec{n})$ cannot be all distinct unless $c_{\vec{n}} = 0$. This is for the same reason as before:

$$||f||_2 = \sum_{\vec{n}\in\mathbb{Z}^k} |c_{\vec{n}}|^2.$$

We conclude that if $c_{\vec{n}} \neq 0$ then there exist p, q such that $(S')^{\circ p}(\vec{n}) = (S')^{\circ q}(\vec{n}) = 0$. An easy analysis shows that this implies $n_k = \ldots = n_2 = 0$. Then comparing this with the earlier equation $c_{S'(\vec{n})} = e^{2\pi i \alpha n_1} c_{\vec{n}}$ shows that $n_1 = 0$ as well. $\square$

4.3. **Toral endormophisms.** If $A \in \mathrm{GL}_n(\mathbb{Z})$, then it induces a map $\mathscr{T}_A \colon T^n \to T^n$ preserving the Lebesgue measure induced on $T^n = \mathbb{R}^n/\mathbb{Z}^n$. These are the "toral endomorphisms," which we have already encountered.

Theorem 4.9. *$\mathscr{T}_A$ is ergodic if and only if no eigenvalue of A is a root of unity.*

Since the eigenvalues of A are algebraic, this is the same as no eigenvalue having magnitude 1. For such A, we called $\mathscr{T}_A$ *hyperbolic.*

Proof. (Sketch) We use Fourier analysis again. If $f \in L^2(X,\mu)$ then we write

$$f = \sum_{\vec{n}\in\mathbb{Z}^k} c_n e^{2\pi i \langle \vec{n}, x\rangle}$$

and $f \circ T$ has expansion

$$\sum_{n\in\mathbb{Z}^k} c_n e^{2\pi i \langle \vec{n}, Ax\rangle} = \sum_{n\in\mathbb{Z}^k} c_n e^{2\pi i \langle \vec{n}A, x\rangle}$$

so

$$c_{\vec{n}} = c_{\vec{n}A} = \ldots.$$

Applying Parseval's formula as usual, we conclude that either $c_{\vec{n}} = 0$ or $\{\vec{n}, \vec{n}A, \ldots\}$ is really only a finite set. Then $\vec{n}A^k = \vec{n}$. That implies that A has an eigenvalue which is a kth root of unity.

Conversely, if $A^k\vec{n} = \vec{n}$ for some $\vec{n}$ then

$$f(x) = \sum_{j=0}^{k-1} e^{2\pi i \langle \vec{n}, A^j x\rangle}$$

is invariant under T and non-constant. $\square$

Example 4.10. If $T: X \to X$ is ergodic, $T \times T: X \times X \to X \times X$ may not necessarily be ergodic. Indeed, let $X = S^1$ and T be irrational rotation. Then $T \times T$ preserves the function $(x, y) \mapsto x - y$.

You might wonder if $T \times T$ could ever be ergodic. If T is the doubling map on S^1, then $T \times T$ is indeed ergodic.

4.4. **Bernoulli Shifts.** We can give other proofs that the map $T_d: S^1 \to S^1$ is ergodic, without referencing Fourier analysis, but putting this map into a general context called *Bernoulli shifts*.

Here is a general setting that captures all of these ergodic transformations. We have a finite alphabet $S = \{s_1, \ldots, s_k\}$ and real numbers $\{p_{s_1}, \ldots, p_{s_k}\}$ such that each $p_s \geq 0$ for all $s \in S$ and $\sum_{i=1}^{k} p_{s_i} = 1$. We define the *two-sided Bernoulli space*

$$\Sigma = \{(\ldots x_{-1}, x_0, x_1, \ldots): x_i \in S)\}$$

and the *Bernoulli shift* σ by $(\sigma(x)_i) = (x_{i+1})$. Note that here σ is a bijection.

We also define the *one-sided Bernoulli space*

$$\Sigma_+ = \{(x_0, x_1, \ldots): x_i \in S)\}$$

and the left shift operator σ_L on Σ_+ by

$$(\sigma_L(x)_i) = x_{i+1}.$$

Notice that here σ_L is surjective but not injective.

We equip Σ, Σ_+ with the σ-algebra generated by the fundamental "cylinders"

$$[(i_1, s_0), \ldots, (i_\ell, s_\ell)] = \{x = (x_i)_{i=0}^{\infty} \mid x_{i_0} = s_0, \ldots, x_{i_\ell} = s_\ell\}.$$

These play the role of intervals (rectangles) for the construction of the Lebesgue measure on $\mathbb{R}$ ($\mathbb{R}^n$). We then define measures μ on Σ by

$$\mu([(i_0, s_0), \ldots, (i_\ell, s_\ell)]) = p_{s_0} \cdot \ldots \cdot p_{s_\ell}$$

and similarly for μ_+ on Σ_+. This measure is evidently preserved by σ and σ_L, respectively. So (Σ, σ, μ) and $(\Sigma_+, \sigma_L, \mu_+)$ are measure-preserving systems. This turns out to be a robust framework capturing many measure-preserving systems that we have already encountered.

Example 4.11. The doubling map can be realized as a Bernoulli shift with $S = \{0, 1\}$ and $p_0 = p_1 = 1/2$, we have $(\Sigma_+, \sigma_L, \mu_+) \cong (S^1, T_2, \mu)$.

The tricky thing about Bernoulli shifts is that they are very difficult to distinguish. Even for $k = 2$, we are only choosing p_0 and p_1 such that $p_0 + p_1 = 1$ and it is already impossible to distinguish the different spaces by spectral properties. To do this one needs to introduce the notion of *entropy*.

Σ_+ can be metrized into a compact topological space, with $d(x, x') = \frac{1}{k}$ if $x_i = x'_i$ for $i = 1, \ldots, k$.

Theorem 4.12. (Σ, σ, μ_p) *is ergodic.*

Proof. We begin with a key observation that leverages the specific structure of cylinders. If $E \subset \Sigma$ is a finite union of cylinders and $F = \sigma^{-N} E$, then

$$\mu_p(E \cap \sigma^{-N} E) = \mu_p(E)^2 \text{ for large } N.$$

To see this, think of E as a set where you have restricted the values in a certain (finite) set of indices. Then σ^{-N} is a "right shift" (technically multivalued), so $\sigma^{-N} E$ is a set where you have restricted the values in another finite set of indices shifted to the right from the origina. If you shift by a large enough amount then eventually the places where you have restricted the values of E and $\sigma^{-N} E$ are *disjoint*.

Let B be a measurable set. We want to show that $T^{-1} B = B \implies \mu(B) = 0$ or 1. There exists a finite union of cylinders $E = \bigcup_{j=1}^{N} C_j$ (where each C_j is a cylinder) such that $\mu(E \Delta B) < \epsilon$, so in particular $|\mu(B) - \mu(E)| < \epsilon$. Since $\mu(B) = \mu(\sigma^{-1} B) = \ldots$,

$$\mu(B \Delta \sigma^{-N} E) = \mu(\sigma^{-N} B \Delta \sigma^{-N} E) = \mu(\sigma^{-N}(B \Delta E)) < \epsilon.$$

This holds for *all* N. Now the point is that B is commensurate with both E and $\sigma^{-N} E$, but these two sets are not commensurate with each other by the discussion of the first paragraph *unless* $\mu(E) = 0$ or 1.

More precisely, we have $\mu(B \Delta E) < \epsilon$ and $\mu(B \Delta \sigma^{-N} E) < \epsilon$. Also,

$$B \Delta (E \cap \sigma^{-N} E) \subset (B \Delta E) \cup (B \Delta \sigma^{-N} E)$$

so $\mu(B \Delta (E \cap \sigma^{-N} E)) < 2\epsilon$. In particular, $|\mu(B) - \mu(E \cap \sigma^{-N} E)| < \epsilon$. Taking $\epsilon \to 0$, we conclude that $\mu(E) = \mu(E)^2$.

$\square$

5. Mixing

5.1. **Mixing transformations.** Recall that we proved that a Bernoulli shift system (Σ, σ, μ_p) is ergodic if $\sum p_i = 1$ by using the structure of "cylinders," specifically the fact that $\mu(\sigma^{-N} A \cap B) = \mu(A)\mu(B)$ for all sufficiently large N.

By an approximation argument, this shows in fact that for any two measurable sets $\widetilde{A}$ and $\widetilde{B}$ we have

$$\lim_{N\to\infty} \mu(\sigma^{-N}\widetilde{A} \cap \widetilde{B}) = \mu(\widetilde{A})\mu(\widetilde{B}).$$

This is the prototype of a stronger property of transformations called mixing.

Definition 5.1. Let (X, T, μ) be a measure-preserving system. We say that T on X is *mixing* if for all measurable sets $\widetilde{A}$ and $\widetilde{B}$ one has

$$\lim_{n\to\infty} \mu(T^{-n}\widetilde{A} \cap \widetilde{B}) = \mu(\widetilde{A})\mu(\widetilde{B}).$$

Example 5.2. The proof of Theorem 4.12 shows that (Σ, σ, μ_p) is mixing.

Mixing implies ergodic, but not conversely. Indeed, one of our equivalent characterizations of ergodicity in Proposition 4.3 was that for all $\widetilde{A}, \widetilde{B}$ there exists n such that $\mu(T^{-n}\widetilde{A} \cap \widetilde{B}) > 0$.

Example 5.3. R_α is ergodic but not mixing. If $\widetilde{A}$ and $\widetilde{B}$ are small intervals, then it is clear that the limit $\lim_{n\to\infty} \mu(T^{-n}\widetilde{A} \cap \widetilde{B})$ will not exist (it will be zero much of the time), but jump up occasionally.

5.2. **Weakly mixing transformations.** Recall that we used the mean ergodic theorem to show that ergodicity implies if $\widetilde{A}, \widetilde{B}$ are measurable then

$$\frac{1}{n}\sum_{i=1}^{n} \mu(T^{-n}\widetilde{A} \cap \widetilde{B}) \to \mu(\widetilde{A})\mu(\widetilde{B}).$$

In other words, the "average" of the quantities approaches some expected value. Mixing says that *the quantities themselves approach this value.*

Definition 5.4. We say that T is *weakly mixing* if

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} |\mu(T^{-n}\widetilde{A} \cap \widetilde{B}) - \mu(\widetilde{A})\mu(\widetilde{B})| \to 0.$$

Example 5.5. In fact, you can easily see that R_α is not even weakly mixing, since a positive proportion of terms is positive.

Proposition 5.6. *If $\mathscr{P}$ is a semi-algebra (finite unions and intersection) generating $\mathscr{B}$, then*

- *Ergodicity $\iff$ for all $A, B \in \mathscr{P}$ then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu(T^{-i}A \cap B) = \mu(A)\mu(B),$$

- *weakly mixing* $\iff$ *for all* $A,B\in\mathscr{P}$ *then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|\mu(T^{-i}A\cap B)-\mu(A)\mu(B)|=0.$$

- *mixing* $\iff$ *for all* $A,B\in\mathscr{P}$ *then*

$$\lim_{n\to\infty}\mu(T^{-n}A\cap B)=\mu(A)\mu(B).$$

Exercise 5.7. Prove this. [Hint: there is basically nothing to do.]

Exercise 5.8. Show that T is *weakly mixing* if and only if $T\times T$ (on $X\times X$ with the product measure) is ergodic.

Example 5.9. Recall that $(x,y)\mapsto(x+\alpha,y+\alpha)$ is not ergodic, which reflects the fact that R_α is not weakly mixing.

Exercise 5.10. Showing weakly mixing $\iff$ given A,B there exists $J\subset\{1,\dots,n\dots\}$ of zero density (i.e. $\lim J\cap\{1,\dots,k\}/k\to 0$) such that

$$\lim_{\substack{n\to\infty\\ n\notin J}}\mu(T^{-n}A\cap B)=\mu(A)\mu(B).$$

5.3. **Spectral perspective.** Ergodicity, weakly mixing, and mixing are "spectral properties" of the operator u_T on $L^2(X,\mu)$. For instance, ergodic says that for $f,g\in L^2(X,\mu)$ we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(u_T^i f,g)\to(f,1)(1,g)$$

and mixing says that for all $f,g\in L^2(X,\mu)$

$$\lim_{n\to\infty}(u_T^n f,g)=(f,1)(1,g).$$

Exercise 5.11. T is mixing if and only if for all measurable sets $A\subset X$,

$$\lim_{n\to\infty}\mu(T^{-n}A\cap A)=\mu(A)^2$$

Recall that if $\mu(X)<\infty$ then $\limsup\mu(A\cap T^{-n}A)\geq\mu(A)^2$.

The exercise says that it suffices to check this in the special case $f=g$.

Here's another way to think about things. Recall that one formulation of ergodicity was that $u_T(f)=f \implies f$ is constant. Weakly mixing says if $u_T(f)=\lambda f$ (necessarily $|\lambda|=1$ because u_T is unitary), then $\lambda=1$, i.e. f is constant. In other words, weakly mixing implies that there can be no other interesting eigenvalues other than 1.

Indeed, if T is weakly mixing and $\mu_T f=\lambda f$, then we may assume that $\int f\,d\mu=0$ because f must be orthogonal to the space of constant functions (those being the eigenspace with eigenvalue 1). Then

$$\frac{1}{n}\sum_{i=1}^{n}|(u_T^n f,f)|\to 0.$$

So

$$\frac{1}{n}\sum|\lambda^i|(f,f)\to 0\implies(f,f)=0.$$

5.4. **Hyperbolic toral automorphism is mixing.** We revisit the hyperbolic toral automorphism T induced by $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ on $T = \mathbb{R}^2/\mathbb{Z}^2$. The goal is to prove that the action of T is mixing. (We already saw a Fourier-analytic proof of ergodicity, but of course this is stronger.)

Consider the eigenvectors for A, spanned by $v_1 = (\frac{1+\sqrt{5}}{2}, 1)$ and $v_2 = (\frac{1-\sqrt{5}}{2}, 1)$. If $x = y + \alpha v_1 + \beta v_2$, then

$$A^k(x-y) = \alpha A^k v_1 + \beta A^k v_2 = \alpha \lambda_1^k v_1 + \beta \lambda_2^k v_2.$$

So if $x - y$ is in the direction of v_2 then $d(T^n x, T^n y) \to 0$. So we have a foliation of T by lines parallel to v_2. If U is a little rectangle with edges parallel to v_1 and v_2, then $T^{-n}(U)$ is a rectangle stretched along the v_2 direction and squished along the v_1 direction.

These eigenvectors define foliation parallel to v_1 and v_2. Let $h_i^s(x) = x + s v_i$ be the flow along the ith foliation. This flow is ergodic in the sense that any measurable function invariant under it must be a.e. constant. Indeed, the flow defines a "first return map" $S^1 \to S^1$ which is which is rotation by an irrational angle, and we know that this is an ergodic transformation.

Let $h^s = h_1^s$ be the flow along the expanding foliation and $\lambda = \lambda_1$. Then $T^n \circ h^s(x) = h^{\lambda_1^n s} \circ T^n(x)$. If f, g are continuous then we want to prove that

$$\lim_{n\to\infty} \int_X f(x) g(T^n x)\, d\mu(x) \to \left(\int f\, d\mu \right) \left(\int g\, d\mu \right).$$

Let $I_n = \int_X f(x) g(T^n x)\, d\mu(x)$. Since h_s is measure-preserving, we can replace x by $h^s x$ for small s without affecting the integral by very much:

$$\begin{aligned} I_n &= \frac{1}{s} \int_X \left(\int_0^s f(h^{s'} x) g(T^n h^{s'} x)\, ds' \right) d\mu(x) \\ (f \text{ continuous}) &\approx \int_X f(x) \left(\frac{1}{s} \int_0^s g(T^n h^{s'} x)\, ds' \right) d\mu(x) \\ &= \int_X f(x) \left(\frac{1}{s} \int_0^s g(h^{\lambda^n s'} T^n x)\, ds' \right) d\mu(x) \\ &= \int_X f(T^{-n} x) \left(\frac{1}{\lambda^n s} \int_0^{\lambda^n s} g(h^{s'} x)\, ds' \right) d\mu(x) \end{aligned}$$

By the ergodicity of h^s ("time average is space average"), we see that

$$\frac{1}{\lambda^n s} \int_0^{\lambda^n s} g(h^{s'} x)\, ds' \to \int_X g(x) \mu(x)$$

for *any* x. Therefore,

$$\int_X f(T^{-n} x) \left(\frac{1}{\lambda^n s} \int_0^{\lambda^n s} g(h^{s'} x)\, ds' \right) d\mu(x) \to \int_X f(T^{-n} x)\, d\mu(x) \int_X g(x)\, d\mu(x).$$

In summary, the important ingredients were ergodicity of expanding foliations, and the existence of expanding/contracting directions. These ideas are the basis of the notion of *entropy*. It turns out that the Lebesgue measure has the maximum possible entropy for T, and this gives information about periodic points, etc.

6. Pointwise Ergodic Theorems

We now work towards an L^1-version of the mean ergodic theorem. Let (X,μ,T) be a measure-preserving system with $\mu(X)<\infty$ and T an ergodic μ-invariant measure. If $f\in L^1(X,\mu)$ then this will say that for almost every $x\in X$

$$\lim_{n\to\infty}\frac{f(x)+f(Tx)+\ldots+f(T^nx)}{n}\to\int_X f(x)\,d\mu.$$

One can weaken this in several ways: if T is not ergodic and $\mu(X)$ is not necessarily finite, then the limit exists as some T-invariant f^* whic is $E(f\mid\mathscr{B}^T)$, $\mathscr{B}^T$ being the σ-algebra of T-invariant subsets, or equivalently the almost-invariant subsets B_0 satisfying $\mu(T^{-1}B_0\Delta B_0)=0$. If $\mu(X)<\infty$ then

$$\int_X f\,d\mu=\int_X f^*\,d\mu.$$

6.1. The Radon-Nikodym Theorem.

Definition 6.1. We say that ν is *absolutely continuous* with respect to ν, and write $\nu\ll\mu$, if $\mu(B)=0\implies\nu(B)=0$.

Example 6.2. Two measures on $[0,1]$ with disjoint support are singular with respect to each other, hence not absolutely continuous.

Example 6.3. If μ is the Lebesgue measure on $[0,1]$, then one can define an absolutely continuous ν measure with respect to μ by picking a positive function f and setting

$$\nu(B)=\int_B f\,d\mu$$

The Radon-Nikodym theorem is a converse to this construction.

Theorem 6.4 (Radon-Nikodym). *Let (X,B,μ) be a probability space. Let ν be a measure defined on $\mathscr{B}$ such that $\nu\ll\mu$. Then there exists a non-negative measurable function f such that*

$$\nu(B)=\int_B f\,d\mu.$$

Furthermore, if

$$\nu(B)=\int_B g\,d\mu$$

then $f=g$ almost everywhere.

Remark 6.5. The (almost everywhere) uniqueness justifies the notation $f=\frac{d\nu}{d\mu}$. So

$$\nu(B)=\int_B\frac{d\nu}{d\mu}\,d\mu.$$

Other basic properties are justified: if $\nu_1,\nu_2\ll\mu$ then

$$\frac{d(\nu_1+\nu_2)}{d\mu}=\frac{d\nu_1}{d\mu}+\frac{d\nu_2}{d\mu}$$

and if $\lambda \ll \nu \ll \mu$ then
$$\frac{d\lambda}{d\mu} = \frac{d\lambda}{d\nu}\frac{d\nu}{d\mu}.$$

Example 6.6. An example to keep in mind why the finiteness hypothesis is necessary: compare the counting measure
$$\nu(A) = \begin{cases} |A| & |A| < \infty \\ \infty & \text{otherwise.} \end{cases}$$

Indeed, the Lebesgue measure is absolutely continuous with respect to this counting measure, but $\frac{d\mu}{d\nu} = 0$ almost everywhere.

6.2. **Expectation.** Let $\mathscr{A} \subset \mathscr{B}$ be a sub σ-algebra and μ a measure on $\mathscr{B}$. If $f \in L^1(X, \mathscr{B}, \mu)$, f might not be $\mathscr{A}$-measurable. We want to define some function "expectation function" $E(f \mid \mathscr{A}) \in L^1(X, \mathscr{A}, \mu)$ which captures the idea of projecting f to $\mathscr{A}$.

How do we construct this operator $E(\cdot \mid \mathscr{A})$? If we were working with L^2 then we could define a projection map, but we cannot do that here. If f is non-negative, then we define
$$\nu(A) = \int_A f\, d\mu.$$

Then $\nu \ll \mu|_{\mathscr{A}}$. By the Radon-Nikodym theorem, there exists $E(f \mid A)$ such that
$$\nu(A) = \int_A E(f \mid \mathscr{A})\, d\mu.$$

If f is actually measurable for $\mathscr{A}$, then $E(f \mid \mathscr{A}) = f$.

By construction, for all $A \in \mathscr{A}$ we have
$$\int_A f\, d\mu = \int_A E(f \mid \mathscr{A})\, d\mu \text{ for all } A \in \mathscr{A}.$$

Example 6.7. If $\mathscr{A}$ consists of sets of measure 0 or full measure, then any $\mathscr{A}$-measurable function is constant almost everywhere, and
$$E(f \mid \mathscr{A}) = \int_X f\, d\mu.$$

Example 6.8. If $\mathscr{A}$ is generated by a finite partition $A_1, \dots, A_n$ of X, then $\mathscr{A}$-measurable functions are constant on each A_i, so
$$E(f \mid \mathscr{A})(x) = \frac{1}{\mu(A_i)} \int_{A_i} f\, d\mu \text{ if } x \in A_i.$$

So far we have restricted our discussion to non-negative functions f, but we can extend the definition in the usual way: write $f = f_+ - f_-$ where f_+ and f_- are the positive and negative parts.

Properties. It is easy to check the following properties of the expectation.

- $E(f_1 + f_2 \mid \mathscr{A}) = E(f_1 \mid \mathscr{A}) + E(f_2 \mid \mathscr{A})$,

- $E(f \mid \mathscr{B}) = f$,
- $E(f \mid \mathscr{A}) \circ T = E(f \circ T \mid T^{-1}\mathscr{A})$.

6.3. **Birkhoff's Ergodic Theorem.**

Theorem 6.9 (Birkhoff's Ergodic Theorem)**.** *Let (X, T, μ) be a system and let $\mathscr{A}$ be the σ-algebra generated by T-invariant measurable sets, i.e. A such that $\mu(T^{-1}A\Delta A) = 0$. (So if T is ergodic, then $\mathscr{A}$ is the trivial σ-algebra.) If $f \in L^1(X, \mathscr{B})$ then for almost all x*

$$\lim_{n\to\infty} \frac{f(x) + f(Tx) + \ldots + f(T^n x)}{n+1} = E(f \mid \mathscr{A})(x) =: f^*(x).$$

The limit $f^*(x)$ is $\mathscr{A}$-measurable (i.e. T-invariant) and for any T-invariant subset A (i.e. $\mu(T^{-1}A\Delta A) = 0$),

$$\int_A f\, d\mu = \int_A f^*\, d\mu.$$

Proof. Let $f \in L^1(X, \mu, \mathscr{B})$ and E be the set of x such that $f(x) + \ldots + f(T^n x) \geq 0$ *for at least one n*.

Claim. We claim that

$$\int_E f\, d\mu \geq 0.$$

This is the main part of the argument. It does *not* use the fact that $\mu(X) < \infty$.

Lemma 6.10 (Maximal inequality)**.** *Let $f \in L^1(X, \mu, \mathscr{B})$ and define $f_0 = 0$,*

$$f_n(x) = f + f \circ T + \ldots + f \circ T^{n-1},$$

and

$$F_n(x) = \max_{0 \leq j \leq n} f_j(x).$$

Then

$$\int_{x\colon F_n(x) > 0} f\, d\mu \geq 0.$$

Let $E_n = \{x \colon F_n(x) > 0\}$. The difference between the claim and the lemma is that in the claim, we are integrating over $E = \bigcup_n E_n$.

Proof. We claim that if $F_n(x) > 0$ then $f(x) \geq F_n(x) - F_n \circ T(x)$. To see this, observe that $F_n \geq f_j$ for all $0 \leq j \leq n$, hence $F_n \circ T \geq f_j \circ T$, so

$$F_n \circ T(x) + f(x) \geq f_j \circ T(x) + f(x) = f_{j+1}(x).$$

Therefore, $F_n \circ T(x) + f(x) \geq \max_{1 \leq j \leq n+1} f_j(x)$. Since $F_m(x) \geq 0$, this the same as the maximum including $j = 0$.

Now,

$$\int_{E_n} f(x)\, d\mu \geq \int_{E_n} F_n(x)\, d\mu - \int_{E_n} F_n \circ T(x)\, d\mu.$$

We claim that we can replace E_n in the second interval with X, because outside E_n the function F_n is 0. Since F_n is non-negative, we can also extend the integral to X in the third integral. So

$$\int_{E_n} f(x)\,d\mu \geq \int_X F_n(x)\,d\mu - \int_X F_n \circ T(x)\,d\mu = 0.$$

Here we are using that F_n is measurable and the T-invariance of the measure. □

Now for the claim, we write $E = \bigcup_n E_n$. Then $f\chi_{E_n} \to f\chi_E$ and $f \in L^1$, so we can use the dominated convergence theorem to conclude that

$$\lim_{n\to\infty} \int f\chi_{E_n}\,d\mu \to \int f\chi_E\,d\mu.$$

Having established the claim, let's turn out attention to the ergodic theorem. We want to analyze

$$\lim_{n\to\infty} \frac{f(x)+f(Tx)+\ldots+f(T^n x)}{n+1}$$

but we don't know that the limit exists. So instead, we study

$$f^*(x) = \limsup_{n\to\infty} \frac{f(x)+f(Tx)+\ldots+f(T^n x)}{n+1}$$
$$f_*(x) = \liminf_{n\to\infty} \frac{f(x)+f(Tx)+\ldots+f(T^n x)}{n+1}$$

We want to prove that given a, b the set

$$E_{a,b} := \{x \colon f_*(x) < a < b < f^*(x)\}$$

has measure 0 in X. Since $\mathbb{R}$ is separable, we can let a, b range over $\mathbb{Q}$ to deduce the result.

A useful observation about this is that f^*, f_* are both T-invariant, hence $E_{a,b}$ is T-invariant. This is shown by analyzing the identity

$$A_n f(Tx) = \frac{n+1}{n} A_{n+1}(x) + \frac{f(x)}{n}.$$

Remark 6.11. If T were ergodic then we would automatically know that $\mu(E_{a,b}) = 0$ or 1.

A corollary of the claim is that if $g \in L^1(X,\mu)$ and

$$B_\alpha = \{x \colon \sup_{n\geq 1} \frac{1}{n} \sum_{j=0}^{n-1} g(T^j(x)) > \alpha\}$$

then for any set A which is T-invariant (up to measure 0),

$$\int_{B_\alpha \cap A} g\,d\mu \geq \alpha\mu(B_\alpha \cap A).$$

Indeed, this follows immediately from applying the claim with $f := g - \alpha$.

So our goal is to show that $\mu(E_{a,b})$ has measure 0. By the corollary applied to the observation that $f^*(x) > b$ on $E_{a,b}$,

$$\int_{E_{a,b}} f\, d\mu \geq b\mu(E_{a,b})$$

On the other hand, since $f_*(x) < a$ on $E_{a,b}$

$$a\mu(E_{a,b}) > \int_{E_{a,b}} f\, d\mu.$$

But $b > a$, so this is only possible $\mu(E_{a,b}) = 0$. Thus $f_* = f^*(x)$ for almost all x, so the limit exists and is T-invariant.

Remark 6.12. We see that the proof so far doesn't use $\mu(X) < \infty$, but without that assumption then the limit could be 0 for instance. We need it to show that the limit satisfies

$$\int_A \widetilde{f} = \int_A f$$

for any A satisfying $\mu(T^{-1}A\Delta A) = 0$.

It is easy to see that the limit $\widetilde{f}$ is in $L^1(X,\mu)$. Indeed, each $A_n(f)$ is in L^1, so

$$|A_n f(x)| = \left| \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) \right|$$

hence

$$\int_X |A_n f(x)|\, d\mu \leq \int |f(x)|\, d\mu$$

since μ is T-invariant. That implies that $f^* \in L^1(X,\mu)$.

We now want to show that $f^* = E(f \mid \mathscr{B}_T)$, i.e. the integrals over any T-invariant B of f and f^* are equal. We can reduce to showing that

$$\int_X f = \int_X f^*.$$

To do this, fix a very large n and set

$$D_k^n = \{x : \frac{k}{n} \leq f^*(x) \leq \frac{k+1}{n}\}.$$

Then obviously

$$\frac{k}{n}\mu(D_k^n) \leq \int_{D_k^n} f^*\, d\mu \leq \frac{k+1}{n}\mu(D_k^n).$$

We claim that in fact

$$\frac{k}{n}\mu(D_k^n) \leq \int_{D_k^n} f\, d\mu \leq \frac{k+1}{n}\mu(D_k^n).$$

Why? Let's focus on proving the lower bound. Fix $\epsilon > 0$. Then the maximal inequality implies that

$$\left(\frac{k}{n} - \epsilon\right)\mu(D_k^n) \le \int_{D_k^n} f\, d\mu$$

In fact we can replace D_k^n with $D_k^n \cap B$ for any T-invariant subset B, by restricting all the results to B. Anyway, this shows that

$$\left|\int_{D_k^n} f - \int_{D_k^n} f^*\right| \le \frac{1}{n}\mu(D_k^n).$$

Summing over k we find that

$$\left|\int_B f - \int_B f^*\right| \le \frac{1}{n}$$

and then letting $n \to \infty$ finishes off the proof. □

6.4. **Some generalizations.**

Remark 6.13. If T is ergodic but $\mu(X) = \infty$ and $f \in L^1(X, \mu)$, then

$$\lim_{n\to\infty} \frac{f(x) + f(T(x)) + \ldots + f(T^{n-1}(x))}{n} \xrightarrow{\text{a.e.}} 0.$$

Even though Birkhoff's Theorem applies and tells us that the limit exists, we unfortunately don't (necessarily) have nice properties of the limit such as $\int f = \int \lim A_n(f)$. There is a way of "fixing" this result, due to Hopf.

Theorem 6.14 (Hopf)**.** *Let T be ergodic on (X, μ). If $f_1, f_2 \in L^1(X, \mu)$ and $\int f_2\, d\mu \neq 0$ then*

$$\lim_{n\to\infty} \frac{f_1(x) + f_1(Tx) + \ldots + f_1(T^n x)}{f_2(x) + f_2(Tx) + \ldots + f_2(T^n x))} = \frac{\int_X f_1}{\int_X f_2}.$$

Another "fix" is the following.

Theorem 6.15 (Hopf)**.** *Assume that there exists $g \in L^1(X, \mu)$ such that $g(x) > 0$ almost everywhere, and that for almost every x,*

$$g(x) + g(T(x)) + \ldots + g(T^n(x)) \to \infty.$$

Then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n f(T^i x)}{\sum_{i=1}^n g(T^i x)} =: \phi(x) \in L^1(X, \mu)$$

and

$$\int_X f\, d\mu = \int_X g\phi\, d\mu.$$

The proof uses "only" the maximal inequality, proceeding along the following lines.

(1) First prove that the $\limsup = \liminf$ almost everywhere.

(2) Partition the space into chunks

$$\frac{k}{n} \leq \phi(x) \leq \frac{k+1}{n}$$

Exercise 6.16. Write out a detailed proof.

Before, we considered sets of the form

$$B_\alpha := \{\sum f(T^i x) > n\alpha\}$$

and deduced that

$$\int_{B_\alpha} f\, d\mu \geq \alpha\mu(B_\alpha).$$

Here we consider a set of the form

$$B_\alpha := \{\sum f(T^i x) > \alpha \sum g(T^i x)\}$$

and show that

$$\int_{B_\alpha} f \geq \alpha \int_{B_\alpha} g.$$

Remark 6.17. Recall that T is mixing if

$$\lim_{n\to\infty} \mu(T^{-n}A \cap B) = \frac{\mu(A)\mu(B)}{\mu(X)}.$$

If $\mu(X) = \infty$, then instead the definition should be

$$\lim_{n\to\infty} \frac{\mu(T^{-n}A \cap B)}{\mu(T^{-n}A' \cap B)} \to \frac{\mu(A)}{\mu(A')}.$$

The proof gives no information about "for which points x does the limit exist."

6.5. **Applications.**

Example 6.18. Recall the "times b" map $T_b\colon S^1 \to S^1$ sending $z \mapsto z^b$. Write $x \in [0,1]$ in terms of a "base b expansion" $0.x_0x_1x_2x_3\ldots$. Then $T_b(x) = 0.x_1x_2x_3\ldots$. We proved that T_b is ergodic. This corresponds to the Bernoulli shift with $p_0 = \frac{1}{b}, p_1 = \frac{1}{b}, \ldots, p_{b-1} = \frac{1}{b}$.

Given $x \in [0,1]$, we can write $x = x_0x_1x_2\ldots x_j$. Then $x_j = k \iff T_b^j(x) \in [\frac{k}{b}, \frac{k+1}{b})$. By Birkhoff's ergodic theorem for $\chi_{[k/b,k+1/b)}$ we have that for almost all x,

$$\lim_{n\to\infty} \frac{\#\{x_i : i \leq n, x_i = k\}}{n} = \frac{1}{b}.$$

This can be generalized to strings of digits: a particular string $(k_1 \ldots k_\ell)$ appears a proportion of $\frac{1}{b^\ell}$ of the time.

Definition 6.19. A point x is *normal* if for all b, in the base b expansion $0.x_0x_1\ldots$ we have

$$\lim_{n\to\infty} \frac{\#\{i : x_i = k, i \leq n\}}{n} = \frac{1}{b}.$$

Birkhoff's ergodic theorem implies that almost all x are normal. However, it is an open question to produce any provably normal point x.

Example 6.20. Consider the Gauss map $T(x) = \frac{1}{x} \mod 1$. If the continued fraction expansion of x is

$$x = x_0 + \cfrac{1}{x_1 + \frac{1}{x_2 + \frac{1}{x_3 + \dots}}}$$

then $x_1 = [1/T(x)], \dots x_n = [1/T^n(x)]$.

The ergodicity of T is then tied with the distribution of $(x_0, x_1, x_2, \dots)$.

For instance, consider the interval $I_k = (\frac{1}{k+1}, \frac{1}{k})$. Then $T^n(x) \in I_k \implies x_n = k$. An invariant measure for T is

$$\mu(B) = \int_B \frac{1}{x+1}\, dx.$$

You can check this on intervals $[a,b]$, so $T^{-1}[a,b] = \bigcup_{k=1}^{\infty} [\frac{1}{b+n}, \frac{1}{a+n}]$.

♠♠♠ TONY: [question: how do you motivate this measure?]

One can prove that T is in fact ergodic with respect to this measure. Then Birkhoff's Ergodic Theorem implies that for almost every x, the frequency of k is the measure of $(\frac{1}{k+1}, \frac{1}{k})$ under μ, and the result turns out to be

$$\frac{1}{\log 2} \log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

One can also use the theorem to do "weighted averages."

Example 6.21. Let $(2^n) = [2, 4, 8, 16, 32, 64, \dots]$. We ask: what is the frequency of ℓ as the first digit in in $x_n \in 2^n$ as $n \to \infty$? We claim that the frequence of ℓ is $\log_{10}(1 + 1/\ell)$.

The number 2^m has d as first digit if it lies in

$$d10^n \le 2^m \le (d+1)10^n$$

for some n. Equivalently,

$$n + \log_{10} d \le m \log 2 \le n + \log_{10}(d+1).$$

Thus $\{m \log 2\} \in [\log d, \log(d+1)] \mod 1$. By Birkhoff's Ergodic Theorem,

$$\{m\alpha\} \in (\log d, \log(d+1))$$

with proportion $\log(1 + 1/d)$ for almost every x. However, right now we are interested in the particular value $x = 0$, so the result does not quite follow from Birkhoff's Ergodic Theorem. Therefore, we need a stronger result.

7. Topological Dynamics

7.1. **The space of T-invariant measures.** Suppose you have a measure-preserving system (X,μ,T) such that X is "compact" and metrizable (these are not essential assumptions, but will be very helpful). The measure is assumed to be Borel (i.e. Borel sets are measurable). Sometimes we will want to assume that T is continuous.

Example 7.1. The rotation map $R_\alpha\colon x\mapsto x+\alpha$ on S^1 and the times d map $T_d\colon z\mapsto z^d$ on S^1 satisfy these conditions.

In general, we can consider the "space of finite T-invariant measures on X, which we denote by $\mathcal{M}^T(X)$. Unfortunately, this "does not help" for understanding measurable dynamics, in the sense that if (X_1,T_1,μ_1) and (X_2,T_2,μ_2) are two systems, then $\mathcal{M}^{T_1}(X_1)$ and $\mathcal{M}^{T_2}(X_2)$ are not related, since in the measurable setting you can throw away sets of measure zero, but there could be lots of interesting T-invariant measures supported on such a set.

There are interesting "classification" theorems that illustrate the disparity between topological and measure-theoretic results. Any regular, ergodic, measure-preserving system (X,T,μ) is isomorphic to a measure-preserving system (X',T',μ') such that μ' is the *only* ergodic measure invariant under T'. Also, it can be shown that the system is isomorphic to a "nice" measure-preserving system on T^2. The moral is that topological dynamics and measure-preserving dynamics very different.

Let $\mathcal{M}_1(X)$ be the space of finite measures on X. This is equipped with the weak* topology. If $C(X)$ is the space of continuous maps $X\to\mathbb{R}$ (or $\mathbb{C}$), then the Riesz Representation Theorem says that $C^*(X)$ is basically the same as the space of "signed measures" on X. Furthermore, $C(X)$ is separable. Then $\mu_k\to\mu$ in the weak* topology if and only if for all $f\in C(X)$,

$$\int f\,d\mu_k\to\int f\,d\mu.$$

It is a fact that $\mathcal{M}_1(X)$ is compact and convex (closed) with respect to the weak* topology. If T is continuous, then there is a map $T_*\colon\mathcal{M}_1(X)\to\mathcal{M}_1(X)$ sending $\mu\mapsto T_*\mu$, i.e.

$$\int f\,d\,T_*\mu=\int f\circ T\,d\mu$$

and moreover this map is continuous with respect to the weak* topology.

Proposition 7.2. *Let X be a compact metrizable space and $T\colon X\to X$ a continuous map. Then $\mathcal{M}_1^T(X)$ is non-empty.*

The content of the proposition is that there always exist non-trivial invariant (probability) measures on a compact metrizable space. How might one construct such an invariant measure? For any $x\in X$, we can consider the sequence $x,Tx,\dots T^nx,\dots$ and define

$$\mu_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{T^ix}\in\mathcal{M}_1(x).$$

Since X is compact, there is a convergent subsequence (in the weak* topology), and we will show shortly that it is T-invariant.

In fact, there is a more general construction. If $\{x_n \in X\}_{n\in\mathbb{N}}$, then one can consider

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{T^i x_n} \in \mathscr{M}_1(x).$$

and again extract a convergent subsequence.

Proof. Let (ν_n) be a sequence in $\mathscr{M}_1(X)$ and let

$$\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} T_*^j(\nu_n).$$

We claim that any weak limit is T-invariant. For any continuous function f on X, we have

$$\begin{aligned}\left|\int f\circ T\, d\mu_{n_j} - \int f\, d\mu_{n_j}\right| &= \frac{1}{n_j}\left|\int \left(\sum f\circ T^{i+1} - f\circ T^i\right) d\nu_{n_j}\right| \\ &= \frac{1}{n_j}\left|\int f\circ T^{n_j+1} - f\, d\nu_{n_j}\right| \\ &\le \frac{2}{n_j}\|f\|_\infty \to 0.\end{aligned}$$

$\square$

Proposition 7.3. *Let X be compact and T measurable. The extreme ponts in $\mathscr{M}_1^T(X)$ are in bijection with ergodic measures for T.*

An extreme point is a measure μ such that if $\mu = \mu_1 + \mu_2$ then $\mu_1 = t\mu$ and $\mu_2 = (1-t)\mu$. (Recall that $\mathscr{M}_1^T(X)$ is convex.) These intuitively correspond to extreme points in the hull of a convex body.

Proof. If μ is not ergodic, then there exists E such that $\mu(E\Delta T^{-1}E) = 0$ and $0 < \mu(E) < 1$, then we can write $\mu = \mu(E)\frac{1}{\mu(E)}\mu|_E + (1-\mu(E))\frac{1}{\mu(X\backslash E)}\mu_{X\backslash E}$, a convex combination of two probability measures which are singular with respect to each other, hence μ is not extremal.

If μ is not extremal, then $\mu = t\mu_1 + (1-t)\mu_2$ and it is easy to see that μ can't be ergodic. The reason is that $\mu_1(A) \le \frac{1}{t}\mu(A)$, so μ_1 is absolutely continuous with respect to μ and by the Radon-Nikodym theorem there exists some ϕ such that

$$\mu_1(A) = \int_A \phi\, d\mu$$

and since μ_1, μ are both T-invariant, ϕ must be. Since T is ergodic, ϕ must be constant almost everywhere, hence $\mu_1 = \mu$. $\square$

7.2. **The ergodic decomposition theorem.** We now want to establish a kind of converse result, asserting that every T-invariant measure is a "linear combination" of the extremal ones. This is *only true* if X is compact.

Theorem 7.4. *Let $\mu \in \mathscr{M}_1^T(X)$. Then there exists a measure λ on $\mathscr{M}_1(X)$ such that $\lambda(E^T) = 1$, where $\mathscr{E}^T$ is the set of extremal measures, such that for all $f \in C(X)$,*

$$\int f\, d\mu = \int_{\mathscr{E}^T(X)} \left(\int_X f\, d\nu \right) d\lambda(\nu).$$

As a consequence, if $|\mathscr{M}_1^T(X)| > 1$ i.e. there is more than one invariant measure, then there exists more than one ergodic invariant measure.

Remark 7.5. This is the only thing reasonable to hope for, because $\mathscr{M}_1^T(X)$ could be a really large space. There are (X, T) where $\mathscr{M}_1^T(X)$ is "finite-dimensional" (only finitely many ergodic measures invariant under T), but in general things are much more complicated, and the set of extremal points may not even be closed. It can even be dense.

Example 7.6. For the time 1 geodesic flow on the unit tangent bundle on a hyperbolic surface X, one can construct ergodic measures of the following form. One ergodic measure α is supported on a closed geodesic, and another ergodic measure β is supported on another closed geodesic. One can then take measures supported on some intertwining of these geodesics, wrapping around n times and renormalizing. In the limit this becomes just $\alpha + \beta$. The set of ergodic measures is dense in the space of all invariant measures for geodesic flow on $T^1(X)$.

8. Unique ergodicity

8.1. Equidistribution.

Definition 8.1. A sequence $x_n \in X$ becomes *equidistributed* with respect to μ if $\mu \in \mathscr{M}_1(X)$ (X compact) if for all $f \in C(X)$,

$$\frac{1}{n}\sum_{j=1}^{n} f(x_j) \to \int f(x)\, d\mu.$$

If (X, T, μ) is a system then we say that $x \in X$ is *generic* if $(x, Tx, T^2x, \ldots)$ becomes equidistributed on X with respect to μ.

Proposition 8.2. *If T is ergodic, then almost every $x \in X$ is generic.*

Proof. By the Pointwise Ergodic Theorem, if $f \in C(X)$ then

$$\frac{1}{n}\sum_{i=1}^{n} f(T^i x) \to \int f\, d\mu$$

for almost every x. However, we are not quite doen yet because the "bad set" depends on f, and there are uncountably many possibilities for f.

What saves us is that in fact $C(X)$ is separable, so we can restriction our attention to the functions in a separable basis $\{f_n\}$ for $C(X)$. Then there is full measure subset $X' \subset X$ such that

$$\frac{1}{N}\sum_{i=1}^{N} f_n(T^i x) \to \int_X f_n\, d\mu$$

for all $x \in X'$ and n. Then any f, you can choose some f_n such that $\|f - f_n\| < \epsilon$, and

$$\int f - 2\epsilon \le \liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f_n(T^i x) \le \limsup_{N\to\infty} \frac{1}{N}\sum f_n(T^i x) \le \int f + 2\epsilon.$$

□

What if we really want to know that *every* point is generic (not just almost every point)? This is tied in with the notion of *unique ergodicity.*

Theorem 8.3. *Let $T: X \to X$ be continuous on X a compact metrizable space. Then the following are equivalent.*

(1) $\#\mathscr{M}_1^T(X) = 1$.

(2) *For every $f \in C(X)$ and every $x \in X$, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(T^i x) = C(f).$$

(3) *For every $f \in C(X)$ and every $x \in X$, we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(T^i x) = \int f\, d\mu$$

uniformly, where μ is the unique T-invariant measure.

(4) *For f in a dense subset of $C(X)$,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(T^i x) = \int f\,d\mu$$

where μ is the unique T-invariant measure.

Definition 8.4. If (X,T) is a system in which the above conditions are satisfied, then we say that it is *uniquely ergodic.*

Proof. (1) $\implies$ (2). Assume that μ is the unique ergodic measure. For $x \in X$, we can consider

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\delta_{T^k x}$$

which is a T-invariant probability measure, necessarily equal to μ (in the weak* topology). That means that for any $f \in C(X)$,

$$\frac{1}{N}\sum_{n=1}^{N} f(T^n x) \to \int f\,d\mu.$$

(2) $\implies$ (3). Letting $\mu_n = \frac{1}{n}\sum_{k=1}^{n}\delta_{T^k x}$ denote the nth measure in the sequence, we have

$$\int f\,d\mu_n = \frac{1}{n}\sum_{k=1}^{n} f(T^k x).$$

Supposing that the convergence is not uniform, then we may choose $g \in C(X)$ such that for all N_0, there exists $N > N_0$ and $x_j \in X$ such that

$$\left|\frac{1}{N}\sum_{n=1}^{N} g(T^n x_j) - C(g)\right| > \epsilon$$

but among such N there exists weak*-convergent subsequence $\mu_{N_i} \to \nu$, so

$$\left|\int_X g\,d\mu - C(g)\right| \geq \epsilon$$

a contradiction.

The equivalence of (3) and (4) follows from general approximation arguments.

Let's show (3) $\implies$ (1). If $A_N f(x) \to C(f)$ which is constant and independent of x, then we want to show that there is only one ergodic measure. Indeed, for every T-invariant measure μ we have

$$\int A_N f(x)\,d\mu \to \int_X C(f)\,d\mu = C(f)$$

and on the other hand

$$\int A_N f(x)\,d\mu = \int f\,d\mu$$

so for any two T-invariant measures μ, ν we have

$$\int f\,d\mu = C(f) = \int f\,d\nu$$

for all $f \in C(X)$, hence $\mu = \nu$.

$\square$

Remark 8.5. The uniformity of convergence doesn't follow from generalities: it is *not true* that convergence to a continuous function on a compact space is automatically uniform. For example, take a sequence of functions on $[0,1]$ where the nth element "spikes" on $[0,1/n]$.

8.2. **Examples.** On S^1, $\{x_n\}_{n=1}^{\infty}$ become equidistributed with respect to the Lebesgue measure m if for any $f \in C(X)$,

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to \int f\,dm.$$

This is equivalent to: for any interval (a,b),

$$\frac{1}{n}\#\{j \le n : x_j \in (a,b)\} \to |b-a|$$

and it's in fact enough to show that if $k \neq 0$, then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} e^{2\pi i k x_j} = 0$$

because the trigonometric polynomials are dense in $C(X)$. This isn't necessarily easy to check: for instance the question of whether or not $(3/2)^n$ is equidistributed is still open.

Theorem 8.6. *If $R_\alpha(x) := x + \alpha$ on $\mathbb{R}/\mathbb{Z} = S^1$, where α is irrational, then for every $x \in S^1$ the sequence $x, R_\alpha x, R_\alpha^2 x, \dots$ becomes equidistributed with respect to the Lebesgue measure. In particular, (S^1, R_α) is uniquely ergodic.*

Proof. We have to check that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} e^{2\pi i k(x+n\alpha)} = 0.$$

But this can be written as

$$e^{2\pi i k x}\sum_{n=1}^{N} e^{2\pi i k n \alpha}.$$

letting $z = e^{2\pi i k \alpha}$, the sum is

$$\frac{1}{N}\sum_{n=1}^{N} z^n = \frac{1}{N}\frac{1-z^N}{1-z} \to 0.$$

$\square$

Let $T: X \to X$ be a continuous map of a compact and metrizable space. Let $\phi: X \to S^1$ be a continuous function. Then we can construct a new pair $(\widehat{X}, \widehat{T})$ where $\widehat{X} = X \times S^1$ and $\widehat{T}: (x,s) \mapsto (T(x), s + \varphi(x))$.

Example 8.7. We've seen a special case of this before: $S^1 \times S^1 \to S^1 \times S^1$ sending $(x,y) \mapsto (x+\alpha, x+y)$ is the construction for $X = S^1$, $T = R_\alpha$, $\phi = \mathrm{Id}$. We proved that this is ergodic using Fourier analysis.

Theorem 8.8 (Furstenberg)**.** *Suppose (X,T) is uniquely ergodic with unique T-invariant measure μ. If $\widehat{\mu} = \mu \times m$ (the Lebesgue measure) is $\widehat{T}$-ergodic, then $(\widehat{X}, \widehat{T})$ is uniquely ergodic and $\widehat{\mu}$ is the unique $\widehat{T}$-invariant measure on $\widehat{X}$.*

Proof. For $t \in S^1$, let τ_t be the map defined by $(x,s) \mapsto (x, s+t)$, which commutes with $\widehat{T}$. Then if ν_0 is $\widehat{T}$-invariant, $\nu_t := (\tau_t)_* \nu_0$ is also $\widehat{T}$-invariant. We define

$$\widehat{\nu} = \int_t \nu_t \, dt.$$

We claim that $\widehat{\nu} = \mu \times m$. If this is true, then that expresses an ergodic measure as an integral of other invariant measures, which is impossible unless almost all of them are the same.

Consider the projection map $\widehat{X} \to X$. Then the pushforward of ν_0 on X is a T-invariant probability measure, hence equal to μ. Thus

$$\begin{aligned}
\int_{\widehat{X}} f \, d\widehat{\nu} &= \int_{S^1} \int_X f \, d\nu_t \, dt \\
&= \int_{S^1} \int_X f(x, s+t) \, d\nu_0 \, dt \\
&= \int_X \left(\int_{S^1} f(x,t) \, dt \right) d\nu_0 \\
&= \int_X f \, d\mu \, dm
\end{aligned}$$

hence $\widehat{\nu} = \mu \times m$. Then there exists t_0 such that $\nu_{t_0} = \mu \times m$, hence $\nu_0 = \mu \times m$. $\square$

Example 8.9. As mentioned above, we already proved that $(x,y) \mapsto (x+\alpha, x+y)$ is ergodic, hence *uniquely ergodic* by the theorem. Let's see some interesting consequences.

For all (x,y), the orbit becomes equidistributed in $S^1 \times S^1$, and

$$T^n(x,y) = (x + n\alpha, Y + nx + \frac{n^2 - n}{2}\alpha) = (x + n\alpha, y + n(x - \alpha) + n^2\alpha).$$

Applying this to $(x,y) = (\alpha, 0)$ we see that $T^n(x,y) = (x + n\alpha, n^2\alpha)$. That means that for all $f \in C(S^1)$, applying Theorem 8.3 to $F(x,y) := f(y)$ we have

$$\frac{1}{N} \sum_{n=1}^{N-1} f(n^2\alpha) \to \int_{S^1} f(y) \, dy$$

hence $\{n^2\alpha\}$ is equidistributed in S^1.

Using this technique, Furstenberg proved that if $p(t)$ is any polynomial with at least one irrational coefficient, then $\{p(n)\alpha\}$ is equidistributed.

8.3. **Minimality.** A set equidistributed with respect to the Lebesgue measure must be dense, but a dense set need not be equidistributed with respect to the Lebesgue measure. We saw that unique ergodicity is equivalent to every point having equidistributed orbit, so a natural relaxation is to study dense orbits.

Definition 8.10. We say that (X, T) is *minimal* if every orbit is dense.

If a system is uniquely ergodic for the Lebesgue measure, then it must be minimal by the observations above. However, if the unique measure is not Borel then there is no implication in either direction.

Example 8.11. The doubling map $T_2 \colon S^1 \to S^1$ is uniquely ergodic, but not minimal. Indeed, this has a (unique) fixed point, and it turns out that the only invariant ergodic measure is a mass supported at this point. But the orbit of the fixed point is obviously not dense.

In some nice situations, the two can be proved to be equivalent.

In fact, the irrational rotation is in some sense the "only" uniquely ergodic transformation, as the following theorem describes.

Theorem 8.12. *Let $T \colon S^1 \to S^1$ be a homeomorphism with no periodic points. Then there exists an irrational rotation $S \colon S^1 \to S^1$ and map $\phi \colon S^1 \to S^1$ such that $\phi \circ T = S \circ \phi$, i.e. the diagram commutes:*

$$\begin{tikzcd} S^1 \arrow[r, "\phi"] \arrow[d, "T"'] & S^1 \arrow[d, "S"] \\ S^1 \arrow[r, "\phi"] & S^1 \end{tikzcd}$$

If T is minimal, then ϕ is a homeomorphism.

There is no analogous fact for $S^1 \times S^1$.

Example 8.13. Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ and let f be the associated map $\mathbb{R}^2/\mathbb{Z}^2 \cong S^1 \times S^1$. Then the periodic points are dense, and if $P_n(f) = \#\{x \mid f^n(x) = x\}$ then

$$P_n(f) = \left(\frac{3+\sqrt{5}}{2}\right)^n + \left(\frac{3-\sqrt{5}}{2}\right)^n - 2.$$

Note that this grows *exponentially* with n, and $\lim_{n\to\infty} \frac{\log P_n(f)}{n}$ exists. There will be many ergodic measures, supported on periodic orbits.

To see why this formula is true, note that (x, y) is periodic if $A^n(x, y) - (x, y) \in \mathbb{Z}^2$. Therefore, $(A^n - 1)(x, y) \in \mathbb{Z}^2$. This translates the question of counting periodic points to a question of counting lattice points: how many lattice points are there in $(A^n - I)([0,1] \times [0,1])$? (That number is precisely $P_n(f)$.)

Theorem 8.14 (Pick's Theorem). *The area of a lattice triangle in* $\mathbb{R}^2$ *is*

$$i + \frac{b}{2} - 1$$

where i *is the number of interior lattice points and* b *is the number of boundary lattice points.*

Remark 8.15. There is a generalization to higher dimensions.

In our case, the area of $(A^n - I)([0,1] \times [0,1])$ is precisely the determinant. To use Pick's theorem, one has to check that there are no other integral points on the boundary.

9. Spectral Methods

9.1. **Spectral isomorphisms.** Our goal is to distinguish between R_α and R_β where α, β are irrational rotations, by considering the induced actions on L^2_m (where m is the Lebesgue measure on S^1).

We have discussed how a triple (X, T, μ) gives an operator U_T on $L^2(X, \mu)$.

Definition 9.1. We say that T_1 and T_2 are *spectrally isomorphic*, and write $U_{T_1} \cong U_{T_2}$, if we can we find $W\colon L^2(X_1, \mu_1) \to L^2(X_2, \mu_2)$ such that $\langle W f_1, W f_2\rangle = \langle f_1, f_2\rangle$ and

$$U_{T_2} \circ W = W \circ U_{T_1}$$

i.e. the following diagram commutes:

$$\begin{tikzcd} L^2(X_1,\mu_1) \arrow[r, "U_{T_1}"] \arrow[d, "W"'] & L^2(X_1,\mu_1) \arrow[d, "W"] \\ L^2(X_2,\mu_2) \arrow[r, "U_{T_2}"'] & {}^2(X_2,\mu_2) \end{tikzcd}$$

9.2. **Ergodic spectra.**

Proposition 9.2. *Let (X, T, μ) be a T-invariant probability measure, where T is ergodic. Then*

(1) $U_T f = \lambda f, f \in L^2(\mu) \implies |\lambda| = 1$ and $|f|$ is constant,
(2) Eigenfunctions correspond to different eigenvalues are orthogonal,
(3) If f, g are both eigenfunctions for λ then $f = cg$ for some constant c,
(4) The eigenvalues form a subgroup of the unit circle.

Remark 9.3. It is possible that the only eigenvalue is 1 and the only eigenfunctions are the constants.

Definition 9.4. We say that (X, T, μ) has *discrete spectrum* if there exists an orthonormal basis for $L^2(X, \mu)$ consisting of eigenfunctions.

It is a fact that if T_1, T_2 have discrete spectra, then they are spectrally isomorphic if and only if they have the same eigenvalues.

Remark 9.5. (X, T, μ) is weakly mixing if and only if 1 is the only eigenfunction for U_T on $L^2(X, T, \mu)$.

Proof. (1) Recall that if T is measure-preserving then U_T is unitary, i.e.

$$\langle U_T f, U_T f\rangle = |\lambda|\langle f, f\rangle = \langle f, f\rangle \implies |\lambda| = 1.$$

Also,

$$|U_T f| = |\lambda| \cdot |f| \implies |U_T f| = |f|$$

so $|f|$ is constant almost everywhere (using ergodicity of T).

(2) We have

$$\langle U_T f, U_T g\rangle = \langle f, g\rangle = \lambda\overline{\mu}\langle f, g\rangle$$

so if $\lambda\overline{\mu} \neq 1$ then $\langle f, g\rangle = 0$.

(3) Suppose $f(T(x)) = \lambda f(x)$ and $g(T(x)) = \lambda g(x)$. If $|g| \neq 0$ almost everywhere then $h = \frac{f}{g}$ is T-invariant, hence constant almost everywhere.
(4) If $f(T(x)) = \lambda f(x)$ and $g(T(x)) = \mu g(x)$ then $\overline{g} \circ T = \overline{\mu g}$, and $f\overline{g} \circ T = \lambda\overline{\mu}(f\overline{g})$.

□

9.3. **Fourier analysis.**

Example 9.6. Consider the rotation R_α. If $f_n(e^{2\pi i x}) = e^{2\pi i n x}$ then $f_n(R_\alpha z) = e^{2\pi i n \alpha} f_n(z)$. Therefore, the maps $z \mapsto z^n$ are all eigenfunctions for R_α.

Theorem 9.7 (Fourier analysis). *The set of f_n forms a basis of $L^2(S^1, m)$.*

Therefore R_α has discrete spectrum with eigenvalues $\{e^{2\pi i n \alpha}\}$.

This is enough to distinguish two rotations R_α and R_β. If they were measure-theoretically isomorphic, then they would be spectrally isomorphic.

Remark 9.8. We can do the same argument for any compact abelian group G. Let $\widehat{G}$ denote the character group. If G is compact metrizable, then $\widehat{G}$ is countable and discrete. For each $a \in G$, there is a map $f_a : G \to G$ sending $x \mapsto ax$.

Theorem 9.9. *The characters of G give an orthonormal basis for $L^2(G, m)$ where m is the Haar measure.*

The eigenvalues are $\{\gamma(a)\}_{\gamma \in \widehat{G}}$. Then we have the following theorem, which asserts that "every" ergodic, measure-preserving map is a rotation.

Theorem 9.10. *If T is an ergodic, measure-preserving map with discrete spectrum, then (X, T, μ) is "conjugate" to a rotation on some compact abelian group. If (X, μ) is regular, then we can replace "conjugate" by "isomorphic."*

Exercise 9.11. Find of a proof of this.

Definition 9.12. We say that (X_1, T_1, μ_1) is *conjugate* to (X_2, T_2, μ_2) if there exists a map $W : L^2(X_1, \mu_1) \to L^2(X_2, \mu_2)$ such that
(1) $(Wf, Wg) = (f, g)$,
(2) $W \circ U_{T_1} = U_{T_2} \circ W$,
(3) W sends bounded functions to bounded functions,
(4) $W(fg) = W(f)W(g)$ for bounded functions f, g.

Definition 9.13. Say $T : X \to X$ is invertible. We say that (X, T, μ) has *countable Lebesgue spectrum* if there are functions $f_0 = 1, f_1, f_2, \ldots, f_n$ such that $\{U_T^i f_k\}_{i,k}$ forms an orthonormal basis for $L^2(X, \mu)$.

The point is as follows.
(1) Any two invertible, measure-preserving maps with countable Lebesgue spectrum are spectrally isomorphic. This is clear by sending the countable spectra to each other.
One can check that having a countable Lebesgue spectrum implies mixing.
(2) Two-sided Bernoulli shifts all have countable Lebesgue spectrum. Therefore, they cannot be distinguished by spectral methods.

10. Entropy

10.1. **Motivation.** We want to motivate the notion of *entropy* for measure-preserving maps (X,T,μ). Consider (S^1,R_α): we mentioned that the operator U_T on $L^2(X,\mu)$ has discrete spectrum. Conversely, any transformation with discrete spectrum looks like rotation on a compact abelian group.

On the other hand, the Bernoulli shifts, which encompass most of the examples we have seen, are all spectrally isomorphic (as they have countable spectra), but they are not measure theoretically isomorphic.

What is the difference between (S^1,R_α) and Bernoulli shifts? The rotation is an *isometry*, and in particular

$$d(x,x')<\epsilon \implies d(T^n x,T^n x')<\epsilon.$$

The Bernoulli shift is much more "violent."

Example 10.1. Baker's transformation is defined on $[0,1]$ by

$$T_B(x,y)=\begin{cases}(2x,y/2) & x\le 1/2\\ (2x-y,(y+1)/2) & x\ge \frac{1}{2}.\end{cases}$$

One can check that this is the same as the bi-infinite Bernoulli shift $((x_i))_{i=-\infty}^{\infty}$. Geometrically, this splits a rectangle down the middle (vertically), and then stacks the halfs vertically, and then crushes them down.

Now we prepare ourselves to define the entropy. The entropy of a system (X,T,μ) is a non-negative number such that:

(1) It is invariant under measurable isomorphism. Therefore, it can distinguish between the Bernoulli shifts $(1/2,1/2)$ and $(1/3,1/3,1/3)$.
(2) Given (X,T), in "many nice situations" there is a *unique* measure of maximal entropy μ for (X,T) (even though there is no way to classify all invariant measures).

However, many interesting measures can have zero entropy.

Example 10.2. For the irrational rotation R_α on S^1, it will be the case $h_\mu(R_\alpha)=0$. This reflects the fact that there are no fixed points. So sometimes we get no information from entropy. However, we'll see that the map $z\mapsto z^2$ has non-zero entropy on S^1.

It tends to be the case that if X is compact, in nice cases (e.g. hyperbolic toral automorphisms such as induced by $\begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}$), the Lebesgue measure has the maximum possible entropy.

We would also like to establish methods to compute entropy. For instance, if $\mu=\mu_1+\mu_2$ then we want to describe h_μ in terms of h_{μ_1} and h_{μ_2}. There is a relationship, but it isn't very simple.

10.2. **Partition information.** The idea of defining entropy is to ask, how much do you "gain" from applying T? Entropy should be a measure of "chaos." So we partition X into finitely many measurable sets $\{P_1,\dots,P_k\}$. That means that $\mu(P_i\cap P_j)=0$ for $i\neq j$ and $\mu(X-\bigcup_{i=1}^k P_i)=0$. The idea is that the "information" you get from $\mathscr{P}$, which depends only on the numbers $\mu(P_1),\dots,\mu(P_k)$, i.e. is a function $H(\mu(P_1),\dots,\mu(P_k))$.

From T and a partition $\mathscr{P}$, we get more partitions $T^{-1}(\mathscr{P}), T^{-2}(\mathscr{P}), \ldots$. Intuitively each of these taken individually has the "same amount of information." But in general, if $\mathscr{P}_1 = \{A_i\}_i$ and $\mathscr{P}_2 = \{B_j\}_j$ are two different partitions then we can form their *join* $\mathscr{P}_1 \vee \mathscr{P}_2 = \{A_i \cap B_j\}_{i,j}$, and this contains "more information" than either.

Now we consider the *growth* of the function H on the partitions $\mathscr{P} \vee T^{-1}\mathscr{P} \vee \ldots \vee T^{-k}(\mathscr{P})$ as $k \to \infty$. Intuitively this tells us about the amount of new information obtained by T; if $T^{-1}(\mathscr{P}) = \mathscr{P}$ then H will not grow at all.

Definition 10.3. For a partition $\mathscr{P} = \{P_1, \ldots, P_k\}$ and a measure μ, we define

$$\begin{aligned} H_\mu(\mathscr{P}) = H(\mu(P_1), \ldots, \mu(P_k)) &= \sum_{i=1}^{k} \mu(P_i) \log(1/\mu(P_i)) \\ &= -\sum \mu(P_i) \log(\mu(P_i)). \end{aligned}$$

The expression here is the same as that from information theory.

There is an elementary calculation due to Khinchin (50s, "Mathematical Foundations of Information Theory") characterizing H as the *unique* function satisfying the following properties:

(1) $H(p_1, \ldots, p_k) \geq 0$ and is 0 if and only if one $p_i = 1$,
(2) H is continuous in $p_1, \ldots, p_k$,
(3) $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$,
(4) H is maximized when $p_1 = \ldots = p_k = \frac{1}{k}$,
(5) If $\mathscr{A}$ and $\mathscr{B}$ are two partitions of X, then $H(\mathscr{A} \vee \mathscr{B}) = H(\mathscr{A}) + H(\mathscr{B} \mid \mathscr{A})$ where
$$H(\mathscr{B} \mid \mathscr{A}) = \sum_{A \in \mathscr{A}} \mu(A) \cdot H_A(\mathscr{B})$$
and
$$H_A(\mathscr{B}) = H\left(\frac{\mu(B_1 \cap A)}{\mu(A)}, \ldots, \frac{\mu(B_k \cap A)}{\mu(A)}\right).$$

Definition 10.4. We define

$$H_\mu(\mathscr{B} \mid \mathscr{A}) = -\sum_{A_i \in \mathscr{A}} \sum_{B_j \in \mathscr{B}_j} \mu(A_i \cap B_j) \log\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) \geq 0.$$

The entropy of a partition $\mathscr{P}$ will eventually be defined as essentially the *growth rate* of $H_\mu(\mathscr{P} \vee T^{-1}\mathscr{P} \vee \ldots \vee T^{-k}\mathscr{P})$ as $k \to \infty$.

Basic Properties. Let α, β, γ be partitions.

(1) $H_\mu(\alpha \vee \beta) = H_\mu(\alpha) + H_\mu(\beta \mid \alpha)$.
(2) $H_\mu(\beta \mid \alpha) \leq H_\mu(\beta)$.
(3) $H_\mu(\alpha \vee \beta \mid \gamma) = H_\mu(B \mid \gamma) + H_\mu(\alpha \mid \beta \vee \gamma)$.
(4) $H_\mu(\alpha \mid \beta \vee \gamma) \leq H(\alpha \mid \beta)$.

The key technical ingredient is the convexity of $x \log x$. Recall that if ψ is convex on (a, b) and $x_i \in (a, b)$, $t_i \in (0, 1)$ such that $\sum t_i = 1$, then

$$\psi(\sum_i t_i x_i) \leq t_i \sum_i \psi(x_i).$$

More generally, if μ is a probability measure, $f \in L^1_\mu(X)$, and ψ is convex then *Jensen's inequality* says that

$$\psi\left(\int f(x)\,d\mu(x)\right) \le \int \psi(x)(f(x))\,d\mu(x).$$

We say that ψ is *strictly convex* if the $\le$ can be replaced by $<$ unless $f(x)$ is (almost everywhere) constant.

Corollary 10.5. *If $\mathscr{P} = \{P_1, \dots, P_k\}$ then $H_\mu(\mathscr{P}) \le \log k$ and equality holds when $\mu(P_1) = \dots = \mu(P_i)$.*

Proof. Let $\phi(x) = x \log x$. If there exists P_i such that $\mu(P_i) \ne 1/k$, then

$$\phi(\sum_{i=1}^{k} \frac{1}{k}\mu(P_i)) < \sum_{i=1}^{k} \frac{1}{k}\phi(\mu(P_i)).$$

Since $\sum \mu(P_i) = 1$, this says that

$$-\frac{\log k}{k} < \frac{1}{k}\sum_{i=1}^{k} -\mu(P_i)\log\mu(P_i) \implies \sum \mu(P_i)\log\mu(P_i) < \log k.$$

Tracing through the equality condition gives the result of the conclusion. □

Now let's prove some of the basic properties.

(1) We have

$$\begin{aligned}
H_\mu(\alpha \vee \beta) &= -\sum_{i,j} \mu(A_i \cap B_j)\log\mu(A_i \cap B_j) \\
&= -\sum_{i,j} \mu(A_i \cap B_j)\log\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) - \sum_{i,j} \mu(A_i \cap B_j)\log\mu(A_i) \\
&= H_\mu(\beta \mid \alpha) + H_\mu(\alpha).
\end{aligned}$$

(2) We have

$$\begin{aligned}
H_\mu(\beta \mid \alpha) &= -\sum_{j=1}^{\ell}\sum_{i=1}^{k} \mu(A_i \cap B_j)\log\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) \\
&= -\sum_{j=1}^{\ell}\sum_{i=1}^{k} \mu(A_i)\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right)\log\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) \\
&\le -\sum_{j=1}^{\ell} \phi\left(\sum_{i=1}^{k} \mu(A_i)\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) \\
&= -\sum_{j=1}^{\ell} \phi(\mu(B_j)) \\
&= H_\mu(\beta).
\end{aligned}$$

10.3. **Definition of entropy.** By using the basic properties, we immediately obtain:

Corollary 10.6. $H_\mu(\alpha \vee \beta) \leq H_\mu(\alpha) + H_\mu(\beta)$.

Lemma 10.7 (Subadditive sequence lemma)**.** *If $(a_n)_n$ is a positive subadditive sequence (i.e. $a_{m+n} \leq a_m + a_n$) then*

$$\lim_{n\to\infty} \frac{a_n}{n} = \ell = \inf_{n\geq 1} \frac{a_n}{n}.$$

Proof. Left as exercise. □

Let $\mathscr{P}$ be a partition of X and define $\mathscr{P}^{(n)} = \bigvee_{i=0}^{n-1} T^{-i}\mathscr{P}$. When then claim that $\{H_\mu(\mathscr{P}^n)\}_n$ forms a subadditive sequence. To see this, note that

$$H_\mu(\mathscr{P}^{(m+n)}) \leq H_\mu(\mathscr{P}^{(m)}) + H_\mu(\mathscr{P}^{(n)})$$

because

$$\mathscr{P}^{(m+n)} = \bigvee_{k=0}^{m+n-1} T^{-k}\mathscr{P} = \left(\bigvee_{k=0}^{n-1} T^{-k}\mathscr{P}\right) \vee \left(T^{-n}(\bigvee_{k=0}^{m-1} T^{-k}\mathscr{P})\right)$$

and Corollary 10.6 implies that

$$\begin{aligned} H_\mu(P^{(m+n)}) &\leq H_\mu(\mathscr{P}^{(m)}) + H_\mu(T^{-n}\mathscr{P}^{(m)}) \\ &\leq H_\mu(\mathscr{P}^{(m)}) + H_\mu(\mathscr{P}^{(n)}). \end{aligned}$$

Therefore, Lemma 10.7 implies that

$$\lim_{n\to\infty} \frac{H_\mu(\mathscr{P}^{(n)})}{n} \text{ exists.}$$

We define the limiting value to be $h_\mu(T, \mathscr{P})$.

Definition 10.8. For a triple (X, μ, T) we define the *entropy* to be

$$h_\mu(T) := \sup_{\text{finite partitions } \mathscr{P}} \{h_\mu(T, \mathscr{P})\}.$$

Remark 10.9. This may seem impossible to compute because one has to check all finite partitions, but it turns out that if $\mathscr{P}$ generates the σ-algebra then $h_\mu(T, \mathscr{P}) = h_\mu(T)$. Thus in nice situations it suffices to compute the entropy of a single partition.

Example 10.10. Let $T : S^1 \to S^1$ be the squaring map and μ the Lebesgue measure. Set $\mathscr{P} = \{[0, 1/2), [1/2, 1)\}$. Then

$$\mathscr{P}^{(n)} = \mathscr{P} \vee T^{-1}\mathscr{P} \vee \ldots \vee T^{-n+1}\mathscr{P},$$

and one can check that this is $\{[\frac{i}{2^{n+1}}, \frac{i+1}{2^{n+1}}]\}$ for $i = 0, 1, \ldots, 2^{n+1} - 1$. So

$$H_\mu(\mathscr{P}^{(n)}) = -2^{n+1} \times \frac{1}{2^{n+1}} \log(1/2^{n+1}) = (n+1)\log 2$$

so $h(T, \mathscr{P}) = \log 2$.

In fact, it is true that $h_\mu(T) = \log 2$. It is not clear how to check this now, since the definition is in terms of *all* partitions, but we shall later see a criterion for checking that a given partition suffices to compute the entropy.

Theorem 10.11 (Kolmogorov-Sinai)**.** *The entropy is invariant under measurable isomorphisms, i.e. if $\pi\colon X_1 \to X_2$ is a measurable isomorphism such that the diagram*

$$\begin{tikzcd} X_1 \arrow[r, "\pi"'] \arrow[d, "T_1"'] & X_2 \arrow[d, "T_2"] \\ X_1 \arrow[r, "\pi"'] & X_2 \end{tikzcd}$$

commutes, then you can easily check that $h_{\mu_1}(T_1) = h_{\mu_2}(T_2)$.

Proof. If $\{A_1, \dots, A_n\}$ is a partition of X_1 then $\{\pi(A_1), \dots, \pi(A_n)\}$ is a partition of X_2 then

$$h_{\mu_1}(T_1, \alpha_1) = h_{\mu_2}(T_2, \pi(\alpha_1)).$$

□

10.4. **Properties of Entropy.** Last time we defined the entropy of a finite measure space. Today we will prove some basic properties about it.

Definition 10.12. We say that $\mathscr{P}$ *generates the σ-algebra of measurable sets on X* if $\bigvee_{i=0}^{\infty} T^{-i}\mathscr{P}$ generates it in the usual sense, i.e. given A measurable in X, for all $\epsilon > 0$ there exists $B \in \bigvee_{i=0}^{\infty} T^{-i}\mathscr{P}$ such that $\mu(A\Delta B) < \epsilon$.

The goal is to prove the following theorem of Sinai.

Theorem 10.13. *If $\mathscr{P}$ generates the σ-algebra of measurable sets, then $h_\mu(T) = h_\mu(T, \mathscr{P})$.*

Example 10.14. This theorem gives an effective method to compute entropy.

(1) For irrational rotation $R_\alpha\colon S^1 \to S^1$, we can check that any interval plus its complement generates the σ-algebra of Lebesgue-measurable sets. Here the number of intervals grows linearly (about $2n$), of length $1/2n$. Then
$$H_\mu(R_\alpha, \mathscr{P}) \approx \lim_{n\to\infty} \frac{\log n}{n} = 0.$$
(2) For the T_d map $S^1 \to S^1$, any interval plus its complement generates the σ-algebra of Lebesgue-measurable sets. Here the number of intervals grows exponentially, and each has length about $1/d^n$. Then
$$H_\mu(R_\alpha, \mathscr{P}) \approx \lim_{n\to\infty} \frac{n \log d}{n} = \log d.$$

Let $\mathscr{A}, \mathscr{C}$ be two partitions. We should have

$$H_\mu(A \vee \mathscr{C}) = H_\mu(A) + H_\mu(\mathscr{C} \mid \mathscr{A}).$$

If $\mathscr{A} = \{A_i\}$ and $\mathscr{C} = \{C_j\}$, recall that we defined

$$H_\mu(A \mid \mathscr{C}) = \sum_{i,j} \mu(A_i \cap C_j) \log\left(\frac{\mu(A_i \cap C_j)}{\mu(C_j)}\right).$$

Remark 10.15. Some basic remarks:

(1) $H_\mu(\mathscr{A} \mid \mathscr{C}) = 0 \iff \mathscr{A} \prec \mathscr{C}$, i.e. every $A_i \in \mathscr{A}$ is a union of elements of $\mathscr{C}$.
(2) $H_\mu(\mathscr{A} \mid \mathscr{C}) = H_\mu(A) \iff \mathscr{A}$ and $\mathscr{C}$ are independent, i.e. $\mu(A_i \cap C_j) = \mu(A_i)\mu(C_j)$ for any i, j.

Proposition 10.16. *The identity*

$$H_\mu(\mathscr{A} \mid \mathscr{C}) + H_\mu(\mathscr{C} \mid \mathscr{A}) = d_\mu(\mathscr{A}, \mathscr{C})$$

defines a metric on the space of finite partitions (up to sets of measure 0*).*

Proof. Definiteness follows from (1) above. We have to check the triangle inequality, which follows if we can establish:

$$H_\mu(\mathscr{A} \mid \mathscr{D}) \le H_\mu(\mathscr{A} \mid \mathscr{C}) + H_\mu(\mathscr{C} \mid \mathscr{D}).$$

Well,

$$\begin{aligned} H_\mu(\mathscr{A} \mid \mathscr{D}) &\le H_\mu(\mathscr{A} \vee \mathscr{C} \mid \mathscr{D}) \\ &= H_\mu(\mathscr{C} \mid \mathscr{D}) + H_\mu(\mathscr{A} \mid \mathscr{C} \vee \mathscr{D}) \\ &\le H_\mu(\mathscr{C} \mid \mathscr{D}) + H_\mu(\mathscr{A} \mid \mathscr{C}) \end{aligned}$$

□

Note that there is a partial order on partitions by $\alpha \prec \beta$ if β is finer than α.

Lemma 10.17. *Let α, β, γ be partitions.*

(1) If $\alpha \prec \beta$, then $H_\mu(\alpha \mid \gamma) \le H_\mu(\beta \mid \gamma)$.
(2) If $\alpha \prec \beta$ then $H_\mu(\gamma \mid \alpha) \ge H_\mu(\gamma \mid \beta)$.
(3) If T preserves μ, then $H_\mu(\alpha \mid \beta) = H_\mu(T^{-1}\alpha \mid T^{-1}\beta)$.

Proof. (1) Note that a special case of (1) is $H_\mu(\alpha) \le H_\mu(\beta)$ if $\alpha \prec \beta$, so let's try to see this first. Well, if $\alpha \prec \beta$ then $\alpha \vee \beta = \beta$, so $H_\mu(\alpha \vee \beta) = H_\mu(\alpha) + H_\mu(\beta \mid \alpha) \ge H_\mu(\alpha)$.

The general argument just works by putting in γ everywhere.

$$\begin{aligned} H_\mu(\beta \mid \gamma) &= H_\mu(\alpha \vee \beta \mid \gamma) \\ &= H_\mu(\alpha \mid \gamma) + H_\mu(\beta \mid \gamma \vee \alpha) \\ &\ge H_\mu(\alpha \mid \gamma). \end{aligned}$$

(2) If $\alpha = \{A_i\}$, $\beta = \{B_j\}$, and $\gamma = \{C_k\}$, then we have

$$\begin{aligned} H_\mu(\gamma \mid \beta) &= -\sum_j \sum_k \mu(C_j \cap B_k) \log\left(\frac{\mu(C_j \cap B_k)}{\mu(B_k)}\right) \\ &= -\sum_{j,k} \mu(B_k) \frac{\mu(C_j \cap B_k)}{\mu(B_k)} \log\left(\frac{\mu(C_j \cap B_k)}{\mu(B_k)}\right) \\ &= -\sum_{i,j,k} \mu(A_i \cap B_k) \frac{\mu(C_j \cap B_k)}{\mu(B_k)} \log\left(\frac{\mu(C_j \cap B_k)}{\mu(B_k)}\right) \end{aligned}$$

Now we claim that

$$\frac{\mu(A_i \cap C_j)}{\mu(A_i)} \log\left(\frac{\mu(A_i \cap C_j)}{\mu(A_i)}\right) \ge -\sum_k \frac{\mu(A_i \cap B_k)}{\mu(A_i)} \cdot \frac{\mu(C_j \cap B_k)}{\mu(B_k)} \log\left(\frac{\mu(C_j \cap B_k)}{\mu(B_k)}\right).$$

Granting the claim, we find that

$$H_\mu(\gamma \mid \beta) \le -\sum_{i,j} \mu(A_i \cap C_j) \log\left(\frac{\mu(A_i \cap C_j)}{\mu(A_i)}\right) = H_\mu(\gamma \mid \alpha).$$

Therefore we are reduced to proving the claimed identity. If $\alpha < \beta$, then

$$\sum_k \frac{\mu(A_i \cap B_k)}{\mu(A_i)} \cdot \frac{\mu(C_j \cap B_k)}{\mu(B_k)} = \frac{\mu(A_i \cap C_j)}{\mu(A_i)}.$$

Applying $\phi(x) = -x \log x$ and convexity completes the proof.

(3) is obvious. □

Corollary 10.18. *We have:*

(1) $\frac{1}{n} H_\mu(\bigvee_{i=0}^{n-1} T^{-i} \mathscr{A})$ *is a decreasing sequence (with limit* $h_\mu(T, \mathscr{A})$*).*

(2) $h_\mu(T, \mathscr{A}) = \lim_{n\to\infty} H_\mu(\mathscr{A} \mid \bigvee_{i=1}^{n-1} T^{-i} \mathscr{A})$.

Proof. (1) We want to show that

$$nH(\bigvee_{i=0}^{n} T^{-i} \mathscr{A}) \le (n+1) H(\bigvee_{i=0}^{n-1} T^{-i} \mathscr{A}).$$

Expanding both sides out, we find that this is equivalent to

$$nH_\mu(\mathscr{A} \mid \bigvee_{i=1}^{n} T^{-i} \mathscr{A}) \le \sum_{j=0}^{n-1} H_\mu(\mathscr{A} \mid \bigvee_{i=1}^{j} T^{-i} \mathscr{A})$$

which is immediate from the fact that conditioning on a larger partition decreases the entropy (Lemma 10.17).

(2) We have

$$H_\mu(\bigvee_{i=0}^{n-1} T^{-i} \mathscr{A}) = H(\mathscr{A}) + \sum_{j=1}^{n-1} H_\mu(\mathscr{A} \mid \bigvee_{i=1}^{j} T^{-i} \mathscr{A}).$$

We will use (1) plus the observation that if $\lim b_i$ exists then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} b_j = \lim_{j\to\infty} b_j.$$

Applying this observation to the sequence

$$b_j = H_\mu(\mathscr{A} \mid \bigvee_{i=1}^{j} T^{-i} \mathscr{A})$$

we deduce that

$$h_\mu(T, \mathscr{A}) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i} \mathscr{A}) = \lim_{n\to\infty} H_\mu(\mathscr{A} \mid \bigvee_{i=1}^{n} T^{-i} \mathscr{A}).$$

□

10.5. **Sinai's generator theorem.** We fix a measure-preserving system $(X,\mu,\mathscr{B},T)$.

Definition 10.19. We say that a finite partition ξ is a *one-sided generator* if $\bigvee_{i=0}^{\infty} T^{-i}\xi$ generates $\mathscr{B}$ (the σ-algebra of measurable subsets), i.e. if for all $B\in\mathscr{B}$ and $\delta>0$ there exists k and $A\in\bigvee_{i=0}^{k} T^{-i}\xi$ such that $\mu(A\Delta B)=0$.

Definition 10.20. If T is invertible and T^{-1} is also measure-preserving then we say that ξ is a *two-sided generator* if $\bigvee_{i=-\infty}^{\infty} T^{-i}\xi$ generates $\mathscr{B}$.

The goal of this section is to prove the following theorem giving an "effective" way to calculate the entropy of a measure-preserving transformation.

Theorem 10.21 (Sinai)**.** *If T is invertible and ξ is a one-sided generator, then $h_\mu(T)=h_\mu(T,\xi)$.*

An analogous result holds if ξ is a two-sided generator. Since the proof is similar, we just prove the result above. The idea is basically that $\bigvee_{i=0}^{n} T^{-i}\alpha$ eventually becomes (almost) finer than any finite partition.

Note that we can replace any partition but a finite expansion of it:

$$h_\mu(T,\alpha)=h_\mu(T,\bigvee_{i=0}^{n} T^{-i}\alpha). \tag{2}$$

Lemma 10.22. *If α,β are finite partitions, then*

$$h_\mu(T,\beta)\le h_\mu(T,\alpha)+H_\mu(\beta\mid\alpha).$$

Proof. We have

$$\begin{aligned}
\frac{1}{n}H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\beta)\frac{1}{n} &\le H_\mu(\bigvee_{i=0}^{n-1}(T^{-i}\beta\vee T^{-i}\alpha))\\
&=\frac{1}{n}H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\alpha)+\frac{1}{n}H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\beta\mid\bigvee_{i=0}^{n-1} T^{-i}\alpha))\\
&\le\frac{1}{n}H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\alpha)+\frac{1}{n}\sum_{j=0}^{n-1}H_\mu(T^{-j}\beta\mid T^{-j}\alpha)
\end{aligned}$$

and taking the limit as $n\to\infty$ completes the proof. $\square$

By the observation 2 that

$$h_\mu(T,\alpha)=h_\mu(T,\bigvee_{i=0}^{n} T^{-i}\alpha).$$

it suffices to establish the following lemma, which is the main technical ingredient of the proof.

Lemma 10.23. *If ξ is a one-sided generator, then*

$$\lim_{n\to\infty}H_\mu(\eta\mid\bigvee_{i=0}^{n} T^{-i}\xi)=0$$

Proof. By taking n to be sufficiently large, we may ensure that every part of η is approximated arbitrarily well by some parts of $\bigvee_{i=0}^{n} T^{-i}\xi$. Intuitively, that means that $H_\mu(\eta \mid \bigvee_{i=0}^{n} T^{-i}\xi)$ is very small since $T^{-i}\xi$ is nearly finer than η.

Exercise 10.24. Prove the result rigorously by analyzing the definition of the conditional information.

□

10.6. **Examples.**

Example 10.25. We consider two-sided Bernoulli shifts with k symbols and parameters $(p_1, \dots, p_k)$. Then we claim that

$$H_\mu(\sigma) = -\sum_{i=1}^{k} p_i \log p_i.$$

Indeed, consider the partition $\mathscr{P}$ obtained separating elements by the value of x_0:

$$\mathscr{P} = \bigcup \{x_0 = s_i\}.$$

Then it is easy to see that this ξ is a two-sided generator, and computation shows that

$$h_\mu(T, \xi) = -\sum p_i \log p_i.$$

In fact, we have the following classification theorem.

Theorem 10.26 (Ornstein). *Entropy is a complete invariant for full shifts.*

There are non-obvious numerical identifications, e.g. $(1/2, 1/2) \not\sim (1/3, 1/3, 1/3)$ but $(1/4, 1/4, 1/4, 1/4) \sim (1/2, 1/8, 1/8, 1/8, 1/8)$.

Remark 10.27. However, for one-sided shifts entropy is *not* a complete invariant. Intuitively, isomorphic one-sided shifts should have the same numbers of symbols, because if there are k symbols then the map is $k:1$.

Remark 10.28. What made the calculation of entropy for the full shift feasible was that for the full shift, $\{T^{-1}\xi, \dots, T^{-i}\xi\}$ form an *independent* partition, i.e.

$$\mu(T^{-j_1} A_{i_1} \cap \dots \cap T^{-jk} A_{ik}) = \prod_m \mu(A_{im}).$$

In this special setting, you don't have to calculate anything because the entropy of the join is automatically the sum of the entropies:

$$H_\mu(\bigvee T^{-i}\xi) = \sum H_\mu(T^{-i}\xi) = n H_\mu(\xi).$$

Therefore,

$$\lim_{n\to\infty} \frac{H_\mu(\xi \vee \dots \vee T^{-(n-1)}\xi)}{n} = H_\mu(\xi).$$

In fact, any invertible map with an "independent generator" is isomorphic to a two-sided Bernoulli shift.

11. Measures of maximal entropy

11.1. **Examples.** Let X be a compact topological space (perhaps metrizable) and T a continuous map $X \to X$. Let μ be a measure on X invariant under T. What can we say about the entropy of μ under T?

For instance, the two-sided Bernoulli shifts with k symbols have maximum entropy $\log k$, when $p_1 = \ldots = p_k = 1/k$. So sometimes there is a unique measure with maximum entropy for T on X. A key example to keep in mind is when T is a homeomorphism that is "expanding," i.e. there exists $\delta > 0$ such that

$$d(T^i x, T^i y) < \delta \forall i \implies x = y.$$

Example 11.1. The map $T_p : S^1 \to S^1$ sending $z \mapsto z^p$. The Lebesgue measure has entropy $\log p$. We claim that any other measure has entropy strictly less than $\log p$ (so the Lebesgue measure is the unique measure of maximal entropy).

Indeed, let ξ be the partition $\{[0,1/p),[1/p,2/p),\ldots,[p-1/p,1)\}$. Note that $\bigvee_{i=0}^{n-1} T^{-i}\xi$ is precisely the partition consisting of (the p^n) intervals of the form $[\frac{j}{p^n}, \frac{j+1}{p^n})$, so its entropy is $\log(p^n) = n\log p$. It is clear that ξ is a generator for the Lebesgue measure, so

$$h_\mu(T) = \lim_{n\to\infty} \frac{H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\xi)}{n} = \log p.$$

In fact, ξ generates any T-invariant, Radon measure. That's because any measurable set must be approximable by intervals. Therefore, $h_\mu(T) \le H_\mu(\xi) \le \log p$ for *any* such μ, and by definition

$$h_\mu(T,\xi) = \inf_{n\ge 1} \frac{H_\mu(\xi \vee T^{-1}\xi \vee \ldots \vee T^{-(n-1)}\xi)}{n} \le \frac{n\log p}{n}.$$

The quality case $H_\mu(\xi \vee \ldots \vee T^{-(n-1)}\xi) = \log(p^n)$ implies $\mu([j/p^n,(j+1)/p^n]) = 1/p^n$ and this implies that μ is the Lebesgue measure.

Example 11.2. Let's consider the hyperbolic toral automorphisms, given by $A \in \mathrm{SL}_2(\mathbb{Z})$ with eigenvalues having absolute value different from 1. One can check that T is a homeomorphism of the torus that is expansive, so that is morally why it works in this case. (Remark: if you have such a homeomorphism, then it's easy to find a generating partition. Indeed, take a partition with diameter less than δ, and it will be a generator).

Theorem 11.3. *If m is the Lebesgue measure, then T has entropy $h_m(T) = \log\rho$ where ρ is the eigenvalue of A greater than* 1.

In fact, we will show that for any μ one has $h_\mu(T) \le \log\rho$, with equality holding for the Lebesgue measure, so again the Lebesgue measure has maximal entropy.

Proof. The goal is to find a particularly nice partition ξ, from which we can calculate the entropy. In this case we can choose a partition consisting of rectangles with edges parallel to the eigenvectors v^+ and v^- with eigenvalues ρ and $1/\rho$.

Then $T^{-1}\xi$ consists of rectangles with edges parallel to v^+ and v^-, but contracted by ρ along the v^+ direction and expanded by ρ along the v^- direction. $T\xi$ has the opposite effect of contracting along v^- and expanding along v^+. Thus $\bigvee_{i=-n}^{n} T^{-i}\xi$ consists of a mesh of rectangles with length and width $\asymp \rho^{-n}$. In particular, we see that ξ is a two-sided generator. Therefore,

$$h_m(T) = h_m(T,\xi) = \lim_{n\to\infty} \frac{H_m(\bigvee_{i=-n}^{n} T^{-i}\xi)}{n}.$$

Now, the lengths of the rectangles in $T^{-i}\xi$ are in $[c_1\rho^{-n}, c_2\rho^{-n}]$ for some constants c_1, c_2 independent of n, so

$$-2\log c_2 + n\log\rho \le H_m(\bigvee_{i=-n}^{n} T^{-1}\xi) \le -2\log c_1 + n\log\rho.$$

Dividing by n and taking the limit, we see that necessarily $h_m(T) = \rho$.

Again, for equality to hold we need that all these rectangles have essentially the same measure, which recovers the Lebesgue measure. □

12. SOLUTIONS TO SELECTED EXERCISES

Exercise 3.4. It suffices to show that for any $\epsilon > 0$ and $M > 0$, we can find $m > M$ such that $\mu(T^{-m}E \cap E) \ge \mu(E)^2 - \epsilon$. Replacing T with T^k, which still preserves μ, it suffices to show that there exists any $m > 0$ such that $\mu(T^{-m}E \cap E) > \mu(E)^2 - \epsilon$.

Since T is measure-preserving,

$$\int_X \sum_{n=1}^{N} \chi_{T^{-n}E} = N\mu(E).$$

Squaring and using Cauchy-Schwarz, we find that

$$\int_X \left(\sum \chi_{T^{-n}E}\right)^2 d\mu \ge \left(\int_X \sum \chi_{T^{-n}E}\right)^2 \ge (N\mu(E))^2.$$

Expanding out the left hand side gives

$$\begin{aligned}\int_X \left(\sum \chi_{T^{-n}E}\right)^2 d\mu &= \int_X \sum_{1\le a\le b\le N} \chi_{T^{-a}E}\chi_{T^{-b}E}\, d\mu \\ &= N + \sum_{1\le a<b\le N} \mu(T^{-a}E \cap T^{-b}E) \\ &= N + \sum_{1\le a<b\le N} \mu(T^{b-a}E \cap E).\end{aligned}$$

Therefore,

$$\sup_{1\le a<b\le N} \mu(T^{b-a}E \cap E) \ge \frac{N^2\mu(E)^2 - N}{N(N-1)} = \frac{N}{N-1}\mu(E)^2 - \frac{1}{N-1}.$$

Letting $N \to \infty$ gives the desired result.

□

Solution to Exercise 5.11. Let $A_n = T^{-n}(A)$. We regard $u_T^n \chi_A = \chi_{A_n} \in L^2(X,\mu)$. By assumption,

$$\int_X \chi_{A_n}\, d\mu = \mu(A) =: \alpha$$

for each n, and

$$\lim_{n\to\infty} \langle \chi_{A_n}, \chi_{A_m}\rangle = \lim_{n\to\infty} \mu(T^{-n}A \cap T^{-m}A) = \alpha^2.$$

Therefore, if we set $f_n = \chi_{A_n} - \alpha$ then we have

$$\lim_{n\to\infty} \langle f_n, f_m\rangle = \lim_{n\to\infty} \langle \chi_{A_n}, \chi_{A_m}\rangle - \alpha^2 = 0.$$

We claim that this implies that $\lim_{n\to\infty}\langle f_n, g\rangle = 0$ for all $g \in L^2(X,\mu)$. Indeed, this is true on the closure of the subspace generated by the f_k, and also on its orthogonal complement by definition. Then taking $g = \chi_B$, we find that

$$0 = \lim_{n\to\infty} \langle f_n, g\rangle = \int_B \chi_{A_n} - \alpha\mu(B) = \mu(T^{-n}A \cap B) - \mu(A)\mu(B).$$

□

Solution to Exercise 3.13. Let $f = \chi_B$. Define

$$A_N(f) = \frac{1}{N}\sum_{n=0}^{N-1} u_T^i(f)$$

and also

$$A_{M,N}(f) = \frac{1}{N}\sum_{n=M}^{M+N-1} u_T^i(f).$$

We know that $A_N(f) \to P_T(f)$. Actually, observe that

$$\begin{aligned}||A_{M,N}(f) - P_T(f)|| &= ||u_T^M(A_N(f)) - P_T(f)|| \\ &= ||u_T^M(A_N(f)) - u_T^M P_T(f) \\ &= ||A_N(f) - P_T(f)||.\end{aligned}$$

It now suffices to show that

$$\int_B P_T(f) \geq \mu(B)^2.$$

Indeed, suppose this to be the case. Then for N large enough, we have

$$\int_B A_{M,N}(f) \geq \mu(B)^2 - \epsilon$$

since convergence in L^2 implies converge in L^1 on a probability space (here is where we use the finite measure assumption!). But

$$\int_B A_{M,N}(f) = \frac{\mu(B \cap T^{-M}(B)) + \ldots + \mu(B \cap T^{-M-N+1}(B))}{N}$$

is the average measure of $T^{-k}(B) \cap B$ for $k \in [M, M+N-1]$. Therefore, $\mu(T^{-k}(B) \cap B) > \mu(B)^2 - \epsilon$ for at least one such k.

Now let's establish the claim. As before, we use the identity

$$\int \sum_{n=1}^{N} u_T^n(f) = N\mu(B).$$

Therefore,

$$\left(\int \sum_{n=1}^{N} u_T^n(f)\right)^2 = N^2\mu(B)^2.$$

Expanding out the left hand side we find

$$N\mu(B) + 2\sum_{k=1}^{N-1}(N-k)\mu(T^{-k}(B) \cap B) = N^2\mu(B)^2.$$

On the other hand,

$$\sum_{n=0}^{N}\int_B A_n(f) = \sum_{k=0}^{N-1}(N-k)\mu(T^{-k}(B) \cap B).$$

Therefore,

$$\sum_{n=0}^{N} n \int_B A_n(f) = \frac{N^2\mu(B)^2 + N\mu(B)}{2}.$$

Now we know that $A_n(f) \to P_T(f)$, so $\int_B A_n(f)$ converges to a limit whose value must then be, by the above equation, $\mu(B)^2$. $\square$

Solution to Exercise 5.7. (1) We have to show that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu(T^{-i}A \cap B) \to \mu(A)\mu(B)$$

given that it holds for all sets in $\mathscr{P}$. For any $\epsilon > 0$, we can choose A', B' such that $\mu(A'\Delta A) < \epsilon$ and $\mu(B'\Delta B) < \epsilon$. Then

$$(T^{-i}A \cap B)\Delta(T^{-i}A' \cap B') \subset (T^{-i}A \Delta T^{-i}A') \cup (B\Delta B')$$

so

$$\left|\mu(T^{-i}A \cap B) - \mu(T^{-i}A' \cap B')\right| < 2\epsilon.$$

Therefore,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mu(T^{-i}A \cap B) - \frac{1}{n}\sum_{i=1}^{n} \mu(T^{-i}A' \cap B')\right| < 2\epsilon.$$

So both the left and right hand sides of the purported identity behave well under approximation with elements of $\mathscr{P}$.

(2) By the same argument as above, the summand $\mu(T^{-i}A \cap B) - \mu(A)\mu(B)$ behaves well under approximation by elements of $\mathscr{P}$, and we know that for elements of $\mathscr{P}$ the limit tends to 0.

(3) Follows from the same argument. $\square$