Lectures on dynamical systems and entropy

Michael Hochman[1]

June 27, 2014

[1]Please report any errors to mhochman@math.huji.ac.il

Contents

Preface

These are notes from an introductory course on dynamical systems and entropy given at the Hebrew University of Jerusalem in the spring semester of 2014. The course covers the basic theorems of ergodic theory and topological dynamics with an emphasis on entropy theory. These notes are evolving...

Chapter 1

Introduction

At its most basic level, dynamical systems theory is about understanding the long-term behavior of a map $T : X \to X$ under iteration. X is called the *phase space* and the points $x \in X$ may be imagined to represent the possible states of the "system". The map T determines how the system evolves with time: time is discrete, and from state x it transitions to state Tx in one unit of time. Thus if at time 0 the system is in state x, then the state at all future times $t = 1, 2, 3, \ldots$ are determined: at time $t = 1$ it will be in state Tx, at time $t = 2$ in state $T(Tx) = T^2x$, and so on; in general we define

$$T^n x = \underbrace{T \circ T \circ \ldots \circ T}_{n}(x)$$

so $T^n x$ is the state of the system at time n, assuming that at time zero it is in state x. The "future" trajectory of an initial point x is called the (forward) orbit, denoted

$$O_T(x) = \{x, Tx, T^2x, \ldots\}$$

When T is invertible, $y = T^{-1}x$ satisfies $Ty = x$, so it represents the state of the world at time $t = -1$, and we write $T^{-n} = (T^{-1})^n = (T^n)^{-1}$. The one can also consider the full or two-sided orbit

$$O_T^{\pm}(x) = \{T^n x \,:\, n \in \mathbb{Z}\}$$

There are many questions one can ask. Does a point $x \in X$ necessarily return close to itself at some future time, and how often this happens? If we fix another set A, how often does x visit A? If we cannot answer this for all points, we would like to know the answer at least for typical points. What is the behavior of pairs of points $x, y \in X$: do they come close to each other? given another pair x', y', is there some future time when x is close to x' and y is close to y'? If $f : X \to \mathbb{R}$, how well does the value of f at time 0 predict its value at future times? How does randomness arise from deterministic evolution of time? And so on.

The set-theoretic framework developed so far there is relatively little that can be said besides trivialities, but things become more interesting when more structure is given to X and T. For example, X may be a topological space, and T continuous; or X may be a compact manifold and T a differentiable map (or k-times differentiable for some k); or there may be a measure on X and T may preserve it (we will come give a precise definition shortly). The first of these settings is called *topological dynamics,* the second *smooth dynamics*, and the last is *ergodic theory.* Our main focus in this course is ergodic theory, though we will also touch on some subjects in topological dynamics.

One might ask why these various assumptions are natural ones to make. First, in many cases, all these structures are present. In particular a theorem of Liouville from celestial mechanics states that for Hamiltonian systems, e.g. systems governed by Newton's laws, all these assumptions are satisfied. Another example comes from the algebraic setting of flows on homogeneous spaces. At the same time, in some situations only some of these structures is available; an example is can be found in the applications of ergodic theory to combinatorics, where there is no smooth structure in sight. Thus the study of these assumptions individually is motivated by more than mathematical curiosity.

In these notes we focus primarily on ergodic theory, which is in a sense the most general of these theories. It is also the one with the most analytical flavor, and a surprisingly rich theory emerges from fairly modest axioms. The purpose of this course is to develop some of these fundamental results. We will also touch upon some applications and connections with dynamics on compact metric spaces.

Chapter 2

Measure preserving transformations

2.1 Measure preserving transformations

Our main object of study is the following.

Definition 2.1.1. A measure preserving system is a quadruple $\mathcal{X} = (X, \mathcal{B}, \mu, T)$ where $(X, \mathcal{B}, \mu)$ is a probability space, and $T : X \to X$ is a measurable, measure-preserving map: that is

$$T^{-1}A \in \mathcal{B} \qquad \text{and} \qquad \mu(T^{-1}A) = \mu(A) \qquad \text{for all } A \in \mathcal{B}$$

If T is invertible and T^{-1} is measurable then it satisfies the same conditions, and the system is called invertible.

Example 2.1.2. Let X be a finite set with the σ-algebra of all subsets and normalized counting measure μ, and $T : X \to X$ a bijection. This is a measure preserving system, since measurability is trivial and

$$\mu(T^{-1}A) = \frac{1}{|X|}|T^{-1}A| = \frac{1}{|X|}|A| = \mu(A)$$

This example is very trivial but many of the phenomena we will encounter can already be observed (and usually are easy to prove) for finite systems. It is worth keeping this example in mind.

Example 2.1.3. The identity map on any measure space is measure preserving.

Example 2.1.4 (Circle rotation)**.** Let $X = S^1$ with the Borel sets $\mathcal{B}$ and normalized length measure μ. Let $\alpha \in \mathbb{R}$ and let $R_\alpha : S^1 \to S^1$ denote the rotation by angle α, that is, $z \mapsto e^{2\pi i\alpha}z$ (if $\alpha \notin 2\pi\mathbb{Z}$ then this map is not the identity). Then R_α preserves μ; indeed, it transforms intervals to intervals of equal length. If we consider the algebra of half-open intervals with endpoints

in $\mathbb{Q}[\alpha]$, then T preserves this algebra and preserves the measure on it, hence it preserves the extension of the measure to $\mathcal{B}$, which is μ.

This example is sometimes described as $X = \mathbb{R}/\mathbb{Z}$, then the map is written additively, $x \mapsto x + \alpha$.

This example has the following generalization: let G be a compact group with normalized Haar measure μ, fix $g \in G$, and consider $R_g : G \to G$ given by $x \to gx$. To see that $\mu(T^{-1}A) = \mu(A)$, let $\nu(A) = \mu(g^{-1}A)$, and note that ν is a Borel probability measure that is right invariant: for any $h \in H$, $\nu(Bh) = \mu(g^{-1}Bh) = \mu(g^{-1}B) = \nu(B)$. This $\nu = \mu$.

Example 2.1.5 (Doubling map)**.** Let $X = [0, 1]$ with the Borel sets and Lebesgue measure, and let $Tx = 2x \bmod 1$. This map is onto is ,not 1-1, in fact every point has two pre-images which differ by $\frac{1}{2}$, except for 1, which is not in the image. To see that T_2 preserves μ, note that for any interval $I = [a, a + r) \subseteq [0, 1)$,

$$T_2^{-1}[a, a + r) = [\frac{a}{2}, \frac{a + r}{2}) \cup [\frac{a}{2} + \frac{1}{2}, \frac{a + r}{2} + \frac{1}{2})$$

which is the union of two intervals of length half the length; the total length is unchanged.

Note that TI is generally of larger length than I; the property of measure preservation is defined by $\mu(T^{-1}A) = \mu(A)$.

This example generalizes easily to $T_a x = ax \bmod 1$ for any $1 < a \in \mathbb{N}$. For non-integer $a > 1$ Lebesgue measure is not preserved.

If we identify $[0, 1)$ with $\mathbb{R}/\mathbb{Z}$ then the example above coincides with the endomorphism $x \mapsto 2x$ of the compact group $\mathbb{R}/\mathbb{Z}$. More generally one can consider a compact group G with Haar measure μ and an endomorphism $T : G \to G$. Then from uniqueness of Haar measure one again can show that T preserves μ.

Example 2.1.6. (Symbolic spaces and product measures) Let A be a finite set, $|A| \geq 2$, which we think of as a discrete topological space. Let $X^+ = A^{\mathbb{N}}$ and $X = A^{\mathbb{Z}}$ with the product σ-algebras. In both cases there is a map which shifts "to the right",

$$(\sigma x)_n = x_{n+1}$$

In the case of X this is an invertible map (the inverse is $(\sigma x)_n = x_{n-1}$). In the one-sided case X^+, the shift is not 1-1 since for every sequence $x = x_1 x_2 \ldots \in A^{\mathbb{N}}$ we have $\sigma^{-1}(x) = \{x_0 x_1 x_2 \ldots : x_0 \in A\}$.

Let p be a probability measure on A and $\mu = p^{\mathbb{Z}}$, $\mu^+ = p^{\mathbb{N}}$ the product measures on X, X^+, respectively. By considering the algebra of cylinder sets $[a] = \{x : x_i = a_i\}$, where a is a finite sequence of symbols, one may verify that σ preserves the measure.

Example 2.1.7. (Stationary processes) In probability theory, a sequence $\{\xi_n\}_{n=1}^{\infty}$ of random variables is called *stationary* if the distribution of a consecutive n-tuple $(\xi_k, \ldots, \xi_{k+n-1})$ does not depend on where it behind; i.e. $(\xi_1, \ldots, \xi_n) =$

$(\xi_k, \ldots, \xi_{k+n-1})$ in distribution for every k and n. Intuitively this means that if we observe a finite sample from the process, the values that we see give no information about when the sample was taken.

From a probabilistic point of view it rarely matters what the sample space is and one may as well choose it to be $(X, \mathcal{B}) = (Y^{\mathbb{N}}, \mathcal{C}^{\mathbb{N}})$, where $(Y, \mathcal{C})$ is the range of the variables. On this space there is again defined the shift map $\sigma : X \to X$ given by $\sigma((y_n)_{n=1}^{\infty}) = (y_{n+1})_{n=1}^{\infty}$. For any $A_1, \ldots, A_n \in \mathcal{C}$ and k let

$$A^i = \underbrace{Y \times \ldots \times Y}_{k} \times A_1 \times \ldots \times A_n \times Y \times Y \times Y \times \ldots$$

Note that $\mathcal{B}$ is generated by the family of such sets. If P is the underlying probability measure, then stationarity means that for any $A_1, \ldots, A_n$ and k,

$$P(A^0) = P(A^k)$$

Since $A^k = \sigma^{-k} A^0$ this shows that the family of sets B such that $P(\sigma^{-1} B) = P(B)$ contains all the sets of the form above. Since this family is a σ-algebra and the sets above generate $\mathcal{B}$, we see that σ preserves P.

There is a converse to this: suppose that P is a σ-invariant measure on $X = Y^{\mathbb{N}}$. Define $\xi_n(y) = y_n$. Then (ξ_n) is a stationary process.

Example 2.1.8. (Hamiltonian systems) The notion of a measure-preserving system emerged from the following class of examples. Let $\Omega = \mathbb{R}^{2n}$; we denote $\omega \in \Omega$ by $\omega = (p, q)$ where $p, q \in \mathbb{R}^n$. Classically, p describes the positions of particles and q their momenta. Let $H : \Omega \to \mathbb{R}$ be a smooth map and consider the differential equation

$$\begin{aligned} \frac{d}{dt} p_i &= -\frac{\partial H}{\partial q_i} \\ \frac{d}{dt} \dot{q}_i &= \frac{\partial H}{\partial p_i} \end{aligned}$$

Under suitable assumptions, for every initial state $\omega = (p_0, q_0) \in \Omega$ and $t \in \mathbb{R}$ there is determines a unique solution $\gamma_\omega(t) = (p(t), q(t))$, and $\omega_t = \gamma_\omega(t)$ is the state of the world after evolving for a period of t started from ω.

Thinking of t as fixed, we have defined a map $T_t : \Omega \to \Omega$ by $T_t \omega = \gamma_\omega(t)$. Clearly

$$T_0(\omega) = \gamma_\omega(0) = \omega$$

We claim that this is an action of $\mathbb{R}$. Indeed, notice that $\sigma(s) = \gamma_\omega(t + s)$ satisfies $\sigma(0) = \gamma_\omega(t) = \omega_t$ and $\dot{\sigma}(s) = \dot{\gamma_{\omega_t}}(t + s)$, and so $A(\sigma, \dot{\sigma}) = A(\gamma_\omega(t + s), \dot{\gamma}_\omega(t + s)) = 0$. Thus by uniqueness of the solution, $\gamma_{\omega_t}(s) = \gamma_\omega(t + s)$. This translates to

$$T_{t+s}(\omega) = \gamma_\omega(t + s) = \gamma_{\omega_t}(s) = T_s \omega_t = T_s(T_t \omega)$$

and of course also $T_{t+s} = T_{s+t} = T_t T_s \omega$. Thus $(T_t)_{t \in \mathbb{R}}$ is action of $\mathbb{R}$ on Ω.

It often happens that Ω contains compact subsets which are invariant under the action. For example there may be a notion of energy $E : \Omega \to \mathbb{R}$ that is preserved, i.e. $E(T_t\omega) = E(\omega)$, and then the level sets $M = E^{-1}(e_0)$ are invariant under the action. E is nice enough, M will be a smooth and often compact manifold. Furthermore, by a remarkable theorem of Liouville, if the equation governing the evolution is a Hamiltonian equation (as is the case in classical mechanics) then the flow preserves volume, i.e. $\mathrm{vol}(T_tU) = \mathrm{vol}(U)$ for every t and open (or Borel) set U. The same is true for the volume form on M.

2.2 Recurrence

One of deep and basic properties of measure preserving systems is that they display "recurrence", meaning, roughly, that for typical x, anything that happens along its orbit happens infinitely often. This phenomenon was first discovered by Poincaré and bears his name.

Given a set A and $x \in A$ it will be convenient to say that x returns to A if $T^n x \in A$ for some $n > 0$; this is the same as $x \in A \cap T^{-n}A$. We say that x returns for A infinitely often if there are infinitely many such n.

The following proposition is, essentially, the pigeon-hole principle.

Proposition 2.2.1. *Let A be a measurable set, $\mu(A) > 0$. Then there is an n such that $\mu(A \cap T^{-n}A) > 0$.*

Proof. Consider the sets $A, T^{-1}A, T^{-2}A, \ldots, T^{-k}A$. Since T is measure preserving, all the sets $T^{-i}A$ have measure $\mu(A)$, so for $k > 1/\mu(A)$ they cannot be pairwise disjoint mod μ (if they were then $1 \geq \mu(X) \geq \sum_{i=1}^{k} \mu(T^{-i}A) > 1$, which is impossible). Therefore there are indices $0 \leq i < j \leq k$ such that $\mu(T^{-i}A \cap T^{-j}A) > 0$. Now,

$$T^{-i}A \cap T^{-j}A = T^{-i}(A \cap T^{-(j-i)}A)$$

so $\mu(A \cap T^{-(j-i)}A) > 0$, as desired. □

Theorem 2.2.2 (Poincare recurrence theorem)**.** *If $\mu(A) > 0$ then μ-a.e. $x \in A$ returns to A.*

Proof. Let

$$E = \{x \in A : T^n x \notin A \text{ for } n > 0\} = A \setminus \bigcup_{n=1}^{\infty} T^{-n}A$$

Thus $E \subseteq A$ and $T^{-n}E \cap E \subseteq T^{-n}E \cap A = \emptyset$ for $n \geq 1$ by definition. Therefore by the previous corollary, $\mu(E) = 0$. □

Corollary 2.2.3. *If $\mu(A) > 0$ then μ-a.e. $x \in A$ returns to A infinitely often.*

Proof. Let E be as in the previous proof. For any k-tuple $n_1 < n_2 < \ldots < n_k$, the set of points $x \in A$ which return to A only at times $n_1, \ldots, n_k$ satisfy $T^{n_k} x \in E$. Therefore,

$$\{x \in A : x \text{ returns to } A \text{ finitely often}\} = \bigcup_k \bigcup_{n_1 < \ldots < n_k} T^{-n_k} E$$

Hence the set on the left is the countable union of set of measure 0. □

In order to discuss of recurrence for individual points we suppose now assume that X is a metric space.

Definition 2.2.4. Let X be a metric space and $T : X \to X$. Then $x \in X$ is called *forward recurrent* if there is a sequence $n_k \to \infty$ such that $T^{n_k} x \to x$.

Proposition 2.2.5. *Let $(X, \mathcal{B}, \mu, T)$ by a measure-preserving system where X is a separable metric space and the open sets are measurable. Thenμ-a.e. x is forward recurrent.*

Proof. Let $A_i = B_{r_i}(x_i)$ be a countable sequence of balls that generate the topology. By Theorem 2.2.2, there are sets $A_i' \subseteq A_i$ of full measure such that every $x \in A_i'$ returns to A_i. Let $X_0 = X \setminus \bigcup (A_i \setminus A_i')$, which is of full μ-measure. For $x \in X_0$ if $x \in A_i$ then x returns to A_i infinitely often, so it returns to within $|\operatorname{diam} A_n|$ of itself infinitely often. Since x belongs to sets A_n of arbitrarily small diameter, x is recurrent. □

When the phenomenon of recurrence was discovered it created quite a stir. Indeed, by Liouville's theorem it applies to Hamiltonian systems, such as planetary systems and the motion of molecules in a gas. In these settings, Poincaré recurrence seems to imply that the system is stable in the strong sense that it nearly returns to the same configuration infinitely often. This question arose original in the context of stability of the solar system in a weaker sense, i.e., will it persist indefinitely or will the planets eventually collide with the sun, or fly off into deep space. Stability in the strong sense above contradicts our experience. One thing to note, however, is the time frame for this recurrence is enormous, and in the celestial-mechanical or thermodynamics context it does not say anything about the short-term stability of the systems.

Chapter 3

Ergodicity

3.1 Ergodicity

In this section and the following ones we will study how it may be decomposed into simpler systems.

Definition 3.1.1. Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system. A measurable set $A \subseteq X$ is invariant if $T^{-1}A = A$. The system is *ergodic* if there are no non-trivial invariant sets; i.e. every invariant set has measure 0 or 1.

If A is invariant then so is $X \setminus A$. Indeed,

$$T^{-1}(X \setminus A) = T^{-1}X \setminus T^{-1}A = X \setminus A$$

Thus, ergodicity is an irreducibility condition: a non-ergodic system the dynamics splits into two (nontrivial) parts which do not "interact", in the sense that an orbit in one of them never enters the other.

Example 3.1.2. Let X be a finite set with normalized counting measure, and $T : X \to X$ a 1-1 map. If X consists of a single orbit then the system is ergodic, since any invariant set that is not empty contains the whole orbit. In general, X splits into the disjoint (finite) union of orbits, and each of these orbits is invariant and of positive measure. Thus the system is ergodic if and only if it consists of a single orbit.

Note that every (invertible) system splits into the disjoint union of orbits. However, these typically have measure zero, so do not in themselves prevent ergodicity.

Example 3.1.3. By taking disjoint unions of measure preserving systems with the normalized sum of the measures, one gets many examples of non-ergodic systems.

Definition 3.1.4. A function $f : X \to Y$ for some set Y is invariant if $f(Tx) = f(x)$ for all $x \in X$.

The primary example is 1_A when A is invariant.

Lemma 3.1.5. *The following are equivalent:*

1. $(X, \mathcal{B}, \mu, T)$ *is ergodic.*
2. *If* $T^{-1}A \subseteq A$ *then* $\mu(A) = 0$ *or* 1.
3. *If* $\mu(A) \geq 0$ *then* $B = \bigcup_{n=0}^{\infty} T^{-n}A$ *has measure* 0 *or* 1.
4. *If* $\mu(A) > 0$ *and* $\mu(B) > 0$ *then there is an* n *with* $\mu(A \cap T^{-n}B) > 0$.
5. *If* $T^{-1}A = A \bmod \mu$ *then* $\mu(A) = 0$ *or* 1.

Proof. (1) implies (2): If $T^{-1}A \subseteq A$ let $B = \bigcap_{n=0}^{\infty} T^{-n}A$. Then

$$T^{-n}B = \bigcap_{n=0}^{\infty} T^{-(n+1)}A = \bigcap_{n=1}^{\infty} T^{-n}A = B$$

where the last equality is because $T^{-n}A \subseteq A$ for all n. Thus by ergodicity, $\mu(B) = 0$ or 1.

(2) implies (3): One notes that $T^{-1}B \subseteq B$.

(3) implies (4): by (3), $\bigcup_{n=0}^{\infty} T^{-n}B$ has full measure, and so there must be an n as desired.

(4) implies (5): If $\mu(A) > 0$ then $B = \bigcup_{n=0}^{\infty} T^{-n}A$ must have measure 1, since otherwise its complement $A' = X \setminus B$ and the set B contradicts (4). But up to measure 0 all the sets $T^{-n}A$ are the same, so $\mu(A) = \mu(B) = 1$.

(5) implies (1): trivial, since every invariant set is invariant up to measure 0. □

Example 3.1.6 (Irrational circle rotation)**.** Let $R_\alpha(x) = e^{2\pi i\alpha}x$ be an irrational circle rotation ($\alpha \notin \mathbb{Q}$) on S^1 with Lebesgue measure μ. We claim that this system is ergodic.

Let $\mu(A) > 0$. Let $z \in A$ be a density point for Lebesgue measure, so $\frac{1}{2r}\mu(B_r(z)) \to 1$ as $r \to 0$. Here $B_r(z)$ is the ball of radius r when the circle is parameterized as $\mathbb{R}/2\pi\mathbb{Z}$, which is an arc of comparable length.

Choose r small enough that $\mu(B_r(z)) > 0.99 \cdot 2 \cdot r$.

Choose n so that $d(R_\alpha^{-n}z, z) < 0.01$. This can be done because $\alpha \notin \mathbb{Q}$ and hence all orbits are dense under $R_\alpha^{-1} = R_{-\alpha}$.

Now, $B_r(A \cap R_\alpha^{-n}(z))$ and $A \cap B_r(z)$ are subsets of $B_{1.01r}(z)$, each with measure $0.99 \cdot 2 \cdot r$, while $B_{1.01r}(z)$ has measure $2.02r$. Hence they must intersect in a set of positive measure. In particular $\mu(R_\alpha^{-n}A \cap A) > 0$. By (4) of the the previous lemma R_α is ergodic.

3.2 Induced action on functions

Given a map $T : X \to Y$ there is an induced map $\widehat{T}$ on functions with domain Y, given by

$$\widehat{T}f(x) = f(Tx)$$

On the space $f : Y \to \mathbb{R}$ or $f : Y \to \mathbb{C}$ the operator $\widehat{T}$ has some obvious properties: it is linear, positive ($f \geq 0$ implies $\widehat{T}f \geq 0$), multiplicative ($\widehat{T}(fg) = \widehat{T}f \cdot \widehat{T}g$). Also $|\widehat{T}f| = \widehat{T}|f|$ and $\widehat{T}(f^c) = (\widehat{T}f)^c$.

When $(X, \mathcal{B})$ and $(Y, \mathcal{C})$ are measurable spaces and T is measurable, the induced map $\widehat{T}$ acts on the space of measurable functions on Y.

Lemma 3.2.1. *If $(X, \mathcal{B}, \mu, T)$ is a measure preserving system then for every measurable function $f \geq 0$ on X, $\int f\, d\nu = \int \widehat{T}f\, d\mu$. Furthermore for $p \geq 1$ and $f \in L^p$, Tf is well defined and $\|f\|_p = \|Tf\|_p$.*

Proof. For $A \in \mathcal{C}$ note that $\widehat{T}1_A(x) = 1_A(Tx) = 1_{T^{-1}A}(x)$, hence

$$\int \widehat{T}1_A\, d\mu = \mu(T^{-1}A) = (\widehat{T}\mu)(A) = \int 1_A\, d\widehat{T}\mu$$

This proves $\int f\, d\nu = \int \widehat{T}f\, d\mu$ for indicator functions. Every non-negative function is the increasing pointwise limit of simple functions so the same follows for them by monotone convergence. For L^1 functions the same holds by writing f as a difference of integrable non-negative functions.

Let $f = g$ almost everywhere. Then

$$\begin{aligned} \{x : Tf(x) \neq Tg(x)\} &= \{x : f(Tx)) \neq g(Tx)\} \\ &= T^{-1}\{x : f(x) \neq g(x)\} \end{aligned}$$

$\mu(\{x : f(x) \neq g(x)\}) = 0$ and T preserves μ, also $Tf = Tg$ a.e., so T is well defined on equivalence class in L^p. By the first part,

$$\int |Tf|^p d\mu = \int T(|f|^p) d\mu = \int |f|^p d\mu < \infty$$

this shows that $Tf \in L^p$ if f is and that in this case the norms are the same.

The case $p = \infty$ is obtained by taking $p \to \infty$ (or directly). □

Corollary 3.2.2. *In a measure preserving system $\widehat{T}$ is a norm-preserving self-map of L^p, and if T is invertible then $\widehat{T}$ is an isometry of L^p.*

The operator $\widehat{T}$ on L^2 is sometimes called the *Koopman operator*. When T is invertible it is a unitary operator and opens up the door for using spectral techniques to study the underlying system. We will return to this idea later.

We will follow the usual convention and write T instead of $\widehat{T}$. This introduces slight ambiguity but the meaning should usually be clear from he context.

A function $f : X \to Y$ is called invariant if $Tf = f$.

Lemma 3.2.3. *The following are equivalent in a measure preserving system* $(X, \mathcal{B}, \mu, T)$*:*

1. *The system is ergodic.*
2. *Every measurable invariant function is constant a.e.*
3. *If* $f \in L^1$ *and* $Tf = f$ *a.e. then* f *is a.e. constant.*

Proof. Observe that f is invariant (a.e. invariant) then $\{f < c\}$ is invariant (resp. a.e. invariant) for every constant c; and that f is constant (a.e. constant) if and only if $\{f < c\}$ is $\emptyset$ or X (respectively measure 0 or 1) for every c.

The lemma now follows from Lemma 3.1.5. □

3.3 Mixing

Although a wide variety of ergodic systems can be constructed or shown abstractly to exist, it is surprisingly difficult to verify ergodicity of naturally arising systems. In fact, in most cases where ergodicity can be proved because the system satisfies a stronger "mixing" property.

Definition 3.3.1. $(X, \mathcal{B}, \mu, T)$ is called *mixing* if for every pair A, B of measurable sets,

$$\mu(A \cap T^{-n}B) \to \mu(A)\mu(B) \qquad \text{as } n \to \infty$$

It is immediate from the definition that mixing systems are ergodic. The advantage of mixing over ergodicity is that it is enough to verify it for a "dense" family of sets A, B. It is better to formulate this in a functional way.

Lemma 3.3.2. *For fixed* $f \in L^2$ *and* n*, the map* $(f, g) \mapsto \int f \cdot T^n g \, d\mu$ *is multilinear and* $\left\| \int f \cdot T^n g \, d\mu \right\|_2 \leq \|f\|_2 \|g\|_2$.

Proof. Using Cauchy-Schwartz and the previous lemma,

$$\int f \cdot T^n g \, d\mu \leq \|f\|_2 \|T^n g\|_2 = \|f\|_2 \|g\|_2$$ □

Proposition 3.3.3. $(X, \mathcal{B}, \mu, T)$ *is mixing if and only if for every* $f, g \in L^2$*,*

$$\int f \cdot T^n g \, d\mu \to \int f \, d\mu \cdot \int g \, d\mu \qquad \text{as } n \to \infty$$

Furthermore this limit holds for ail $f, g \in L^2$ *if and only if it holds for* f, g *in a dense subset of* L^2*.*

Proof. We prove the second statement first. Suppose the limit holds for $f, g \in V$ with $V \subseteq L^2$ dense. Now let $f, g \in L^2$ and for $\varepsilon > 0$ let $f', g' \in V$ with

$\|f - f'\| < \varepsilon$ and $\|g - g'\| < \varepsilon$. Then

$$\begin{aligned}\left\|\int f \cdot T^n g \, d\mu\right\| &\leq \left\|\int (f - f' + f') \cdot T^n (g - g' + g') \, d\mu\right\| \\ &\leq \left\|\int (f - f') \cdot T^n g \, d\mu\right\| + \left\|\int f \cdot T^n (g - g') \, d\mu\right\| + \\ &\quad + \left\|\int (f - f') \cdot T^n (g - g') \, d\mu\right\| + \left\|\int f' \cdot T^n g' \, d\mu\right\| \\ &\leq \varepsilon \|g\| + \|f\| \varepsilon + \varepsilon^2 + \left\|\int f' \cdot T^n g' \, d\mu\right\|\end{aligned}$$

Since $\left\|\int f' \cdot T^n g' \, d\mu\right\| \to 0$ and ε was arbitrary this shows that $\left\|\int f \cdot T^n g \, d\mu\right\| \to 0$, as desired.

For the first part, using the identities $\int 1_A \, d\mu = \mu(A)$, $T^n 1_A = 1_{T^{-n}A}$ and $1_A 1_B = 1_{A \cap B}$, we see that mixing is equivalent to the limit above for indicator functions, and since the integral is multilinear in f, g it holds for linear combinations of indicator functions and these combinations are dense in L^2, we are done by what we proved above. □

Example 3.3.4. Let $X = A^{\mathbb{Z}}$ for a finite set A, take the product σ-algebra, and μ a product measure with marginal given by a probability vector $p = (p_a)_{a \in A}$. Let $\sigma : X \to X$ be the shift map $(\sigma x)_n = x_{n+1}$. We claim that this map is mixing and hence ergodic.

To prove this note that if $f(x) = \widetilde{f}(x_1, \ldots, x_k)$ depends on the first k coordinates of the input, then $\sigma^n f(x) = \widetilde{f}(x_{k+1}, \ldots, x_{k+n})$. If f, g are two such functions then for n large enough, $\sigma^n g$ and f depend on different coordinates, and hence, because μ is a product measure, they are independent in the sense of probability theory:

$$\int f \cdot \sigma^n g \, d\mu = \int f \, d\mu \cdot \int \sigma^n g \, d\mu = \int f \, d\mu \cdot \int g \, d\mu$$

so the same is true when taking $n \to \infty$. Mixing follows from the previous proposition.

Chapter 4

The ergodic theorem

4.1 Preliminaries

We have seen that in a measure preserving system, a.e. $x \in A$ returns to A infinitely often. Now we will see that more is true: these returns occur with a definite frequency which, in the ergodic case, is just $\mu(A)$; in the non-ergodic case the limit is $\mu_x(A)$, where μ_x is the ergodic component to which x belongs.

This phenomenon is better formulated at an analytic level in terms of averages of functions along an orbit. To this end let us introduce some notation. Let $T : V \to V$ be a linear operator of a normed space V, and suppose T is a contraction, i.e. $\|Tf\| \leq \|f\|$. This is the case when T is induced from a measure-preserving transformation (in fact we have equality). For $v \in V$ define

$$S_N v = \frac{1}{N} \sum_{n=0}^{N-1} T^n v$$

Note that in the dynamical setting, the frequency of visits x to A up to time N is $S_N 1_A(x) = \frac{1}{N} \sum_{n=0}^{N-1} 1_A(T^n x)$. Clearly S_N is linear, and since T is a contraction $\|T^n v\| \leq \|v\|$ for $n \geq 1$, so by the triangle inequality, $\|S_N v\| \leq \frac{1}{N} \sum_{n=0}^{N-1} \|T^n v\| \leq \|v\|$. Thus S_N are also contractions. This has the following useful consequence.

Lemma 4.1.1. *Let $T : V \to V$ as above and let $S : V \to V$ be another bounded linear operator. Suppose that $V_0 \subseteq V$ is a dense subset and that $S_N v \to Sv$ as $N \to \infty$ for all $v \in V_0$. Then the same is true for all $v \in V$.*

Proof. Let $v \in V$ and $w \in V_0$. Then

$$\limsup_{N\to\infty} \|S_N v - Sv\| \leq \limsup_{N\to\infty} \|S_N v - S_N w\| + \limsup_{N\to\infty} \|S_N w - Sv\|$$

Since $\|S_N v - S_N w\| = \|S_N(v - w)\| \leq \|v - w\|$ and $S_N w \to Sw$ (because $w \in V_0$), we have

$$\limsup_{N\to\infty} \|S_N v - Sv\| \leq \|v - w\| + \|Sw - Sv\| \leq (1 + \|S\|) \cdot \|v - w\|$$

Since $\|v - w\|$ can be made arbitrarily small, the lemma follows. □

4.2 Mean ergodic theorem

Historically, the first ergodic theorem is von-Neuman's "mean" ergodic theorem, which can be formulated in a purely Hilbert-space setting (and it is not hard to adapt it to L^P). Recall that if $T : V \to V$ is a bounded linear operator of a Hilbert space then $T^* : V \to V$ is the adjoint operator, characterized by $\langle v, Tw\rangle = \langle T^*v, w\rangle$ for $v, w \in V$, and satisfies $\|T^*\| = \|T\|$.

Lemma 4.2.1. *Let $T : V \to V$ be a contracting linear operator of a Hilbert space. Then $v \in V$ is T-invariant if and only if it is T^*-invariant.*

Remark 4.2.2. When T is unitary (which is one of the main cases of interest to us) this lemma is trivial. Note however that without the contraction assumption this is false even in $\mathbb{R}^d$.

Proof. Since $(T^*)^* = T$ it suffices to prove that $T^*v = v$ implies $Tv = v$.

$$\begin{aligned}
\|v - Tv\|^2 &= \langle v - Tv, v - Tv\rangle \\
&= \|v\|^2 + \|Tv\|^2 - \langle Tv, v\rangle - \langle v, Tv\rangle \\
&= \|v\|^2 + \|Tv\|^2 - \langle v, T^*v\rangle - \langle T^*v, v\rangle \\
&= \|v\|^2 + \|Tv\|^2 - \langle v, v\rangle - \langle v, v\rangle \\
&= \|Tv\|^2 - \|v\|^2 \\
&\leq 0
\end{aligned}$$

where the last inequality is because T is a contraction. □

Let use write U for the space of T-invariant vectors:

$$U = \{v \in V \,:\, Tv = v\}$$

This is a closed linear subspace of V.

Lemma 4.2.3. $U^\perp = \overline{\{w = Tw \,:\, w \in V\}}$.

Proof. Write $U' = \overline{\{w = Tw \,:\, w \in V\}}$. Then It is enough to show that $(U')^\perp = U$. Now,

$$\begin{aligned}
w \perp U' &\iff \forall v \in V \quad \langle w, v - Tv\rangle = 0 \\
&\iff \forall v \in V \quad \langle w, v\rangle - \langle w, Tv\rangle = 0 \\
&\iff \forall v \in V \quad \langle w, v\rangle - \langle T^*w, v\rangle = 0 \\
&\qquad \forall v \in V \quad \langle w - T^*w, v\rangle = 0 \\
&\iff w = T^*w \\
&\iff w = Tw
\end{aligned}$$

where the last line was from the previous lemma. This proves the claim. □

Theorem 4.2.4 (Hilbert-space mean ergodic theorem)**.** *Let T be a linear contraction of a Hilbert space V, i.e. $\|Tv\| \leq \|v\|$. Let $V_0 \leq V$ denote the closed subspace of T-invariant vectors (i.e. $V_0 = \ker(T - I)$) and π the orthogonal projection to V_0. Then*

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n v \to \pi v \qquad \text{for all } v \in V$$

Proof. If $v \in V_0$ then $S_N v = v$ and so $S_N v \to v = \pi v$ trivially. Since $V = V_0 \oplus V_0^\perp$ and S_N is linear, it suffices for us to show that $S_N v \to 0$ for $v \in V_0^\perp$. By the previous lemma,

$$V_0^\perp = \overline{\{v - Tv \,:\, v \in V\}} \tag{4.1}$$

Thus, by Lemma 4.1.1 we must only show that $S_N(v - Tv) \to 0$ for $v \in V$, and this follows from

$$\begin{aligned} S_N(v - Tv) &= \frac{1}{N}\sum_{n=0}^{N-1} T^n(v - Tv) \\ &= \frac{1}{N}(w - T^{N+1}w) \\ &\to 0 \end{aligned}$$

where in the last step we used $\left\|w - T^{N+1}w\right\| \leq \|w\| + \left\|T^{N+1}w\right\| \leq 2\|w\|$. □

Now let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let T denote also the Koopman operator induced on L^2 by T. Then the space V_0 of T-invariant vectors is just $L^2(X, \mathcal{I}, \mu)$, where $\mathcal{I} \subseteq \mathcal{B}$ is the σ-algebra of invariant sets, and the orthogonal projection π to V_0 is just the conditional expectation operator, $\pi f = \mathbb{E}(f|\mathcal{I})$ (see the Appendix). We derive the following:

Corollary 4.2.5 (Dynamical mean ergodic theorem)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system, let $\mathcal{I}$ denote the σ-algebra of invariant sets, and let π denote the orthogonal projection from $L(X, \mathcal{B}, \mu)$ to the closed subspace $L^2(X, \mathcal{I}, \mu)$. Then for every $f \in L^2$,*

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \mathbb{E}(f|\mathcal{I}) \qquad \text{in } L^2$$

In particular, if the system is ergodic then the limit is constant:

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \int f \, d\mu \qquad \text{in } L^2$$

Specializing to $f = 1_A$, and noting that L^2-convergence implies, for example, convergence in probability, the last result says that on an arbitrarily large part of the space, the frequency of visits of an orbit to A up to time N is arbitrarily close to $\mu(A)$, if N is large enough.

4.3 The pointwise ergodic theorem

Very shortly after von-Neuman's mean ergodic theorem (and appearing in print before it), Birkhoff proved a stronger version in which convergence takes place a.e. and in L^1.

Theorem 4.3.1 (Pointwise ergodic theorem)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system, let $\mathcal{I}$ denote the σ-algebra of invariant sets. Then for any $f \in L^1(\mu)$,*

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \mathbb{E}(f|\mathcal{I}) \qquad \text{a.e. and in } L^1$$

In particular, if the system is ergodic then the limit is constant:

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \int f\, d\mu \qquad \text{a.e. and in } L^1$$

We shall see several proofs of this result. The first and most "standard" proof follows the same scheme as the mean ergodic theorem: one first establishes the statement for a dense subspace $V \subseteq L^1$, and then uses some continuity property to extend to all of L^1. The first step is nearly identical to the proof of the mean ergodic theorem.

Proposition 4.3.2. *There is a dense subspace $V \subseteq L^1$ such that the conclusion of the theorem holds for every $f \in V$.*

Proof. We temporarily work in L^2. Let V_1 denote the set of invariant $f \in L^2$, for which the theorem holds trivially because $S_N f = f$ for all N. Let $V_2 \subseteq L^2$ denote the linear span of functions of the form $f = g - Tg$ for $g \in L^\infty$. The theorem also holds for these, since

$$\left\|g + T^{N+1}g\right\|_\infty \leq \|g\|_\infty + \left\|T^{N+1}g\right\|_\infty = 2\|g\|_\infty$$

and therefore

$$\frac{1}{N}\sum_{n=0}^{N-1} T^n(g - Tg) = \frac{1}{N}(g - T^{N+1}g) \to 0 \qquad \text{a.e. and in } L^1$$

Set $V = V_1 + V_2$. By linearity of S_N, the theorem holds for $f \in V_1 + V_2$. Now, L^∞ is dense in L^2 and T is continuous on L^2, so $\overline{V}_2 = \overline{\{g - Tg : g \in L^2\}}$. In the proof of the mean ergodic theorem we saw that $L^2 = V_1 \oplus \overline{V}_2$, so $V = V_1 \oplus V_2$ is dense in L^2, and hence in L^1, as required. □

By Lemma 4.1.1, this proves the ergodic theorem in the sense of L^1-convergence for all $f \in L^1$. In order to similarly extend the pointwise version to all of L^1 we need a little bit of "continuity", which is provided by the following.

Theorem 4.3.3 (Maximal inequality)**.** *Let $f \in L^1$ with $f \geq 0$ and $S_N f = \frac{1}{N}\sum_{n=0}^{N-1} T^n f$. Then for every t,*

$$\mu\left(x \,:\, \sup_N S_N f(x) > t\right) \leq \frac{1}{t}\int f\, d\mu$$

Before giving the proof let us show how this finishes the proof of the ergodic theorem. Write $S = \mathbb{E}(\cdot|\mathcal{I})$, which is a bounded linear operator on L^1, let $f \in L^1$ and $g \in V$. Then

$$\begin{aligned} |S_N f - Sf| &\leq |S_N f - S_N g| + |S_N g - Sg| \\ &\leq S_N|f-g| + |S_N g - Sf| \end{aligned}$$

Now, $S_N g \to Sg$ a.e., hence $|S_N g - Sf| \to |S(g-f)| \leq S|f-g|$ a.e. Thus,

$$\limsup_{N\to\infty} |S_N f - Sf| \leq \limsup_{N\to\infty} S_N|f-g| + S|g-f|$$

If the left hand side is $> \varepsilon$ then at least one of the terms on the right is $> \varepsilon/2$. Therefore,

$$\mu\left(\limsup_{N\to\infty} |S_N f - Sf| > \varepsilon\right) \leq \mu\left(\limsup_{N\to\infty} S_N|f-g| > \varepsilon/2\right) + \mu\left(S|g-f| > \varepsilon/2\right)$$

Now, by the maximal inequality, the first term on the right side is bounded by $\frac{1}{\varepsilon/2}\|f-g\|$, and by Markov's inequality and the identity $\int Sh\, d\mu = \int h\, d\mu$, the second term is bounded by $\frac{1}{\varepsilon/2}\|g-f\|$ as well. Thus for any $\varepsilon > 0$ and $g \in V$ we have found that

$$\mu\left(\limsup_{N\to\infty} |S_N f - Sf| > \varepsilon\right) \leq \frac{4}{\varepsilon}\|f-g\|$$

For each fixed $\varepsilon > 0$, the right hand side can be made arbitrarily close to 0, hence $\limsup |S_N f - Sf| = 0$ a.e. which is just $S_N f \to Sf = \mathbb{E}(f|\mathcal{I})$, as claimed.

We now return to the maximal inequality which will be proved by reducing it to a purely combinatorial statement about functions on the integers. Given a function $\widehat{f} : \mathbb{N} \to [0,\infty)$ and a set $\emptyset \neq I \subseteq \mathbb{N}$, the average of $\widehat{f}$ over I is denoted

$$S_I \widehat{f} = \frac{1}{|I|}\sum_{i\in I} \widehat{f}(i)$$

In the following discussion we write $[i,j]$ also for integer segments, i.e. $[i,j] \cap \mathbb{Z}$.

Proposition 4.3.4 (Discrete maximal inequality)**.** *Let $\widehat{f} : \mathbb{N} \to [0,\infty)$. Let $J \subseteq I \subseteq \mathbb{N}$ be finite intervals, and for each $j \in J$ let $I_j \subseteq I$ be a sub-interval of I whose left endpoint is j. Suppose that $S_{I_j}\widehat{f} > t$ for all $j \in J$. Then*

$$S_I \widehat{f} > t \cdot \frac{|J|}{|I|}$$

Proof. Suppose first that the intervals $\{I_j\}$ are disjoint. Then together with $U = I \setminus \bigcup I_j$ they form a partition of I, and by splitting the average $S_I \widehat{f}$ according to this partition, we have the identity

$$S_I \widehat{f} = \frac{|U|}{|I|} S_U \widehat{f} + \sum \frac{|I_j|}{|I|} S_{I_j} \widehat{f}$$

Since $\widehat{f} \geq 0$ also $S_U \widehat{f} \geq 0$, and so

$$S_I \widehat{f} \geq \sum \frac{|I_j|}{|I|} S_{I_j} \widehat{f} \geq \frac{1}{|I|} \sum t|I_j| \geq t \frac{|\bigcup I_j|}{|I|}$$

Now, $\{I_j\}_{j \in J}$ is not a disjoint family, but the above applies to every disjoint sub-collection of it. Therefor we will be done if we can extract from $\{I_j\}_{j \in J}$ a disjoint sub-collection whose union is of size at least $|J|$. This is the content of the next lemma. □

Lemma 4.3.5 (Covering lemma)**.** *Let $I, J, \{I_j\}_{j \in J}$ be intervals as above. Then there is a subset $J_0 \subseteq J$ such that (a) $J \subseteq \bigcup_{i \in J_0} I_j$ and (b) the collection of intervals $\{J_i\}_{i \in J_0}$ is pairwise disjoint.*

Proof. Let $I_j = [j, j+N(j)-1]$. We define $J_0 = \{j_k\}$ by induction using a greedy procedure. Let $j_1 = \min J$ be the leftmost point. Assuming we have defined $j_1 < \ldots < j_k$ such that $I_{j_1}, \ldots, I_{j_k}$ are pairwise disjoint and cover $J \cap [0, j_k + N(j_k) - 1]$. As long as this is not all of J, define

$$j_{k+1} = \min\{I \setminus [0, j_k + N(j_k) - 1]\}$$

It is clear that the extended collection satisfies the same conditions, so we can continue until we have covered all of J. □

We return now to the dynamical setting. Each $x \in X$ defines a function $\widehat{f} = \widehat{f}_x : \mathbb{N} \to [0, \infty)$ by evaluating f along the orbit:

$$\widehat{f}(i) = f(T^i x)$$

Let

$$A = \{\sup_N S_N f > t\}$$

and note that if $T^j x \in A$ then there is an $N = N(j)$ such that $S_N f(T^j x) > t$. Writing

$$I_j = [j, j + N(j) - 1]$$

this is the same as

$$S_{I_j} \widehat{f} > t$$

Fixing a large M (we eventually take $M \to \infty$), consider the interval $I = [0, M-1]$ and the collection $\{I_j\}_{j \in J}$, where

$$J = J_x = \{0 \leq j \leq M - 1 \,:\, T^j x \in A \text{ and } I_j \subseteq [0, M-1]\}$$

The proposition then gives

$$S_{[0,M-1]}\widehat{f} > t \cdot \frac{|J|}{M}$$

In order to estimate the size of J we will restrict to intervals of some bounded length $R > 0$ (which we eventually will send to infinity). Let

$$A_R = \{ \sup_{0 \le N \le R} S_N f > t\}$$

Then

$$J \supseteq \{0 \le j \le M - R - 1 \,:\, T^j x \in A_R\}$$

and if we write $h = 1_{A_R}$, then we have

$$\begin{aligned} |J| &\geq \sum_{j=0}^{M-R-1} \widehat{h}(j) \\ &= (M-R-1) S_{[0,M-R-1]}\widehat{h} \end{aligned}$$

With this notation now in place,the above becomes

$$S_{[0,M-1]}\widehat{f}_x > t \cdot \frac{M-R-1}{M} \cdot S_{[0,M-R-1]}\widehat{h}_x \tag{4.2}$$

and notice that the average on the right-hand side is just frequency of visits to A_R up to time M.

We now apply a general principle called the *transference principle*, which relates the integral $\int g\, d\mu$ of a function $g : X \to \mathbb{R}$ its discrete averages $S_I\widehat{g}$ along orbits: using $\int g = \int T^n g$, we have

$$\begin{aligned} \int g\, d\mu &= \frac{1}{M} \sum_{m=0}^{M-1} \int T^m g\, d\mu \\ &= \int \left(\frac{1}{M} \sum_{m=0}^{M-1} T^m g \right) d\mu \\ &= \int S_{[0,M-1]}\widehat{g}_x \, d\mu(x) \end{aligned}$$

Applying this to f and using 4.2, we obtain

$$\begin{aligned} \int f\, d\mu &= S_{[0,M-1]}\widehat{f}_x \\ &> t \cdot \frac{M-R-1}{M} \cdot \int h\, d\mu \\ &= t \cdot (1 - \frac{R-1}{M}) \cdot \int 1_{A_R}\, d\mu \\ &= t \cdot (1 - \frac{R-1}{M}) \cdot \mu(A_R) \end{aligned}$$

Letting $M \to \infty$, this is

$$\int f \, d\mu > t \cdot \mu(A_R)$$

Finally, letting $R \to \infty$ and noting that $\mu(A_R) \to \mu(A)$, we conclude that $\int f \, d\mu > t \cdot \mu(A)$, which is what was claimed.

Example 4.3.6. Let $(\xi_n)_{n=1}^{\infty}$ be an independent identically distributed sequence of random variables represented by a product measure on $(X, \mathcal{B}, \mu) = (\Omega, \mathcal{F}, P)^{\mathbb{N}}$, with $\xi_n(\omega) = \xi(\omega_n)$ for some $\xi \in L^1(\Omega, \mathcal{F}, P)$. Let $\sigma : X \to X$ be the shift, which preserves μ and is ergodic, and $\xi_n = \xi_0(\sigma^n)$. Since the shift acts ergodically on product measures, the ergodic theorem implies

$$\frac{1}{N} \sum_{n=0}^{N-1} \xi_n = \frac{1}{N} \sum_{n=0}^{N-1} \sigma^n \xi_0 \to \mathbb{E}(\xi_0 | \mathcal{I}) = \mathbb{E}\xi_0 \qquad \text{a.e.}$$

Thus the ergodic theorem generalizes the law of large numbers. However it is a very broad generalization: it holds for any stationary process $(\xi_n)_{n=1}^{\infty}$ without any independence assumption, as long as the process is ergodic.

When T is invertible it is also natural to consider the two-sided averages $\overline{S}_N = \frac{1}{2N+1} \sum_{n=-N}^{N} T^n f$. Up to an extra term $\frac{1}{2N+1} f$, this is just $\frac{1}{2} S_N(T, f) + \frac{1}{2} S_N T^{-1}, f)$, where we write $S_N(T, f)$ to emphasize which map is being used. Since both of these converge in L^1 and a.e. to the same function $\mathbb{E}(f|\mathcal{I})$, the same is true for $\overline{S}_N f$.

4.4 Interpretation in the non-ergodic case

Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system, and $X = X_1 \cup X_2$ where X_1, X_2 are complementary invariant sets. Suppose that, up to measure 0, these are the only nontrivial invariant sets, so that $\mathcal{I} = \{\emptyset, X_1, X_2, X\}$ (up to measure 0). Write

$$\mu_i = \frac{1}{\mu(X_i)} \mu|_{X_i}$$

for the conditional measures on X_i. It follows that

$$\mathbb{E}(f|\mathcal{I})(x) = \int f d\mu_1 \cdot 1_{X_1}(x) + \int f d\mu_2 \cdot 1_{X_2}(x) = \begin{cases} \int f d\mu_1 & x \in X_1 \\ \int f d\mu_2 & x \in X_2 \end{cases}$$

(see Example in the appendix). Therefore, for μ-a.e. $x \in X_1$ (i.e. μ_1-a.e. x),

$$S_N f(x) \to \int f d\mu_1$$

and for μ-a.e. $x \in X_2$ (i.e. μ_2-a.e. x),

$$S_N f(x) \to \int f d\mu_2$$

In particular, for $f = 1_A$, note that the frequency of visits of the orbit of a typical point x to A varies depending on whether $x \in X_1$ or $x \in X_2$. In particular for $x \in A$, this shows that the rate of recurrence in the Poincaré recurrence theorem may depend on the point (again, on whether $x \in X_1$ or X_2).

4.5 Ergodic decomposition

Let $(X, \mathcal{B}, \mu_i, T)$ be a measure preserving system, X_1 an invariant set with $0 < \mu(X_1) < 1$, and $X_2 = X \setminus X_1$ also invariant. Let $\mu_i = \frac{1}{\mu(X_i)}\mu|_{X_i}$ are invariant measures, and

$$\mu = \mu(X_1)\mu_1 + \mu(X_2)\mu_2$$

presents μ as a convex combination of T-invariant measures. If, as in the example in the last section, there are no other invariant sets besides X_1, X_2 (up to μ-measure 0), then μ_1, μ_2 are ergodic, but if they are not ergodic one can continue to decompose them and obtain a refined convex presentation of μ. This process may or may not stop after a finite number of steps with a presentation of μ as a convex combination of ergodic measures. However, even when it does not terminate, such a presentation exists.

Let $(Y, \mathcal{C})$ and $(Z, \mathcal{D})$ be measurable spaces. A measure-values map from Y to Z is a map $\nu : Y \to \mathcal{P}(Z)$, written as $y \mapsto \nu_y$. Such a map is a measurable map if $y \mapsto \nu_y(D)$ is measurable for all $D \in \mathcal{D}$.

Given a measurable map $\nu : Y \to \mathcal{P}(Z)$ and a measure $\mu \in \mathcal{P}(Y)$, let

$$\mu = \int \nu_y d\mu(y)$$

denote the measure on $(Z, \mathcal{D})$ given by $D \mapsto \int \nu_y(D) d\mu(y)$. This is well defined by measurability and is easily checked to be a probability measure.

Definition 4.5.1. A standard probability space is a probability space $(X, \mathcal{B}, \mu)$ such that there is a complete separable metric d on X for which $\mathcal{B}$ is the Borel σ-algebra.

Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system on a probability space. Let $\mathcal{I} \subseteq \mathcal{B}$ denote the family of T-invariant measurable sets. It is easy to check that $\mathcal{I}$ is a σ-algebra.

The σ-algebra $\mathcal{I}$ in general is not countably generated. Consider for example the case of an invertible ergodic transformation on a Borel space, such as an irrational circle rotation or two-sided Bernoulli shift. Then $\mathcal{I}$ consists only of sets of measure 0 and 1. If $\mathcal{I}$ were countably generated by $\{I_n\}_{n=1}^{\infty}$, say, then for each n either $\mu(I_n) = 1$ or $\mu(X \setminus I_n) = 1$. Set $F_n = I_n$ or $F_n = X \setminus I_n$ according to these possibilities. Then $F = \bigcap F_n$ is an invariant set of measure 1 and is an atom of $\mathcal{I}$. But the atoms of $\mathcal{I}$ are the orbits, since each point in X is measurable and hence every countable set is. But this would imply that μ is

supported on a single countable orbit, contradicting the assumption that it is non-atomic.

We shall work instead with a fixed countably generated μ-dense sub-σ-algebra $\mathcal{I}_0$ of $\mathcal{I}$. Then $L^1(X, \mathcal{I}, \mu)$ is a closed subspace of $L^1(X, \mathcal{B}, \mu)$, and since the latter is separable (due to standardness of the probability space), so is the former. Choose a dense countable sequence $f_n \in L^1(X, \mathcal{I}, \mu)$, choosing representatives of the functions that are genuinely $\mathcal{I}$ measurable, not just modulo a $\mathcal{B}$-measurable nullset. Now consider the countable family of sets $A_{n,p,q} = \{p < f_n < q\}$, where $p, q \in \mathbb{Q}$, and let $\mathcal{I}_0$ be the σ-algebra that they generate. Clearly $\mathcal{I}_0 \subseteq \mathcal{I}$ and all of the f_n are $\mathcal{I}_0$-measurable, so $L^1(X, \mathcal{I}_0, \mu) = L^1(X, \mathcal{I}, \mu)$. In particular, $\mathcal{I}$ is contained in the μ-completion of $\mathcal{I}_0$.

Let $\mathcal{I}_0(y)$ denote the atom of $\mathcal{I}_0$ to which y belongs.

Theorem 4.5.2 (Ergodic decomposition theorem)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system on a standard probability space, and let $\mathcal{I}, \mathcal{I}_0$ be as above. Then there is an $\mathcal{I}_0$-measurable map $X \to \mathcal{P}(X)$, $x \mapsto \mu_x \mathcal{P}$ such that*

1. *$\mu = \int \mu_x \, d\mu(x)$*
2. *μ_y is T-invariant, ergodic, and supported on $\mathcal{I}_0(y)$ for μ-a.e. y.*
3. *For every $f \in L^1(\mu)$ we have $\mathbb{E}(f|\mathcal{I})(y) = \int f \, d\mu_y$ for μ-a.e. y.*

Furthermore the representation is unique in the sense that if $\{\mu'_y\}$ is any other family with the same properties then $\mu_y = \mu'_y$ for μ-a.e. y.

We will not prove this theorem here.

The measure μ_y is called the "ergodic component" of y (it is defined only μ-a.s.). Sometimes $\mathcal{I}_0(y)$ is also called the ergodic component, although this depends on the choice of $\mathcal{I}_0$.

Corollary 4.5.3. *With the notation of the previous theorem, for any $f \in L^1(\mu)$,*

$$S_N f \to \int f d\mu_y$$

for μ-a.e. y.

Chapter 5

Isomorphism

5.1 Isomorphism of probability spaces

Definition 5.1.1. Probability spaces $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{C}, \nu)$ are isomorphic if there are measurable sets $X_0 \subseteq X$ and $Y_0 \subseteq Y$ of full measure, and a map $f : X_0 \to Y_0$ which is 1-1 and both f and f^{-1} are measurable and $\mu(f^{-1}(C)) = \nu(C)$ for all $C \in \mathcal{C}$, $C \subseteq Y_0$. Such a map f is called an isomorphism between X and Y.

Proposition 5.1.2. *Let $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{C}, \nu)$ be standard probability spaces and $f : X \to Y$ a 1-1 measurable and measure-preserving map. Then f is an isomorphism between X and Y.*

Proof. It suffices to show that for every $A \in \mathcal{B}$ with $\mu(A) > 0$ there is a set $A' \subseteq A$ with $\mu(A') > \frac{1}{2}\mu(A)$, $f(A') \in \mathcal{C}$, and $(f|_{A'})^{-1} : f(A') \to A'$ is measurable. Indeed, given this we start with $A_1 = X$ and construct A_1', and define $A_2 = X \setminus A_1'$. Obtain A_2' and set $A_3 = X \setminus (A_1' \cup A_2')$. At the n-th step, $\mu(X \setminus (A_1' \cup \ldots \cup A_n')) < 2^{-n}$ so $X_0 = \bigcup A_n'$ has full measure, and has the required property (note that its image is measurable and since f is measure-preserving, also $f(X_0)$ has full measure).

To prove the claim, let $A \in \mathcal{B}$ and $\mu(A) > 0$. Let d_X, d_Y be complete separable metrics on X, Y respectively under which the σ-algebras are Borel. By Egorov's theorem, we can find $A'' \subseteq A$ with $\mu(A'') > \frac{1}{2}\mu(A)$ and $f|_{A''}$ is continuous in these metrics. Now recall that a Borel probability measure on a separable complete metric space is inner regular (also outer regular), so we can find $A' \subseteq A''$ with $\mu(A') > \frac{1}{2}\mu(A)$ and A' compact. Now $f|_{A'} : A' \to f(A')$ is a continuous injection whose domain is a compact set so the image is a compact set and the inverse is continuous, hence measurable. □

Proposition 5.1.3. *Any two standard probability spaces whose measures are non-atomic are isomorphic.*

We do not prove this here.

5.2 Isomorphism of measure preserving systems

Proposition 5.2.1. *Measure-preserving systems $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ are isomorphic if there are sets of full measure $X_0 \subseteq X$, $Y_0 \subseteq Y$, invariant in the sense that $T^{-1}X_0 = X_0, S^{-1}Y_0 = Y_0$, and a bijection map $f : X_0 \to Y_0$ such that f, f^{-1} are measurable and $S \circ f = f \circ T$, i.e. $Sf(x) = f(Tx)$ for all $x \in X_0$. In this case f is called an isomorphism.*

Clearly isomorphic systems have isomorphic underlying probability spaces.

Proposition 5.2.2. *Let $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ be measure-preserving systems on standard Borel spaces, and let $f : X \to Y$ be an isomorphism satisfying $Sf = fT$. Then the systems are isomorphic.*

Proof. By the previous proposition there are $X_0 \subseteq X$, $Y_0 \subseteq Y$ of full measure so that $f(X_0) = Y_0$ and $f|_{X_0}, (f|_{X_0})^{-1}$ are measurable. Replacing X_0 by $\bigcup_N \bigcap_{n>N} T^{-n}X_0$ and Y_0 by $\bigcup_N \bigcap_{n>N} S^{-n}Y_0$ ensures that they are invariant. □

We shall generally not distinguish between isomorphic systems.

5.3 Spectral isomorphism

If $\pi : X \to Y$ is an isomorphism between measure-preserving systems $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$, then there is an induced map $\pi : L^p(\nu) \to L^p(\mu)$ for every $p \geq 1$ given by $\pi f(x) = f(\pi x)$. The fact that this map is well defined, preserves integrals and norms is proved in the same way as the case$Y = X$ which we have already seen.

In particular follows for $p = 2$ that π is a unitary equivalence between the operators $T : L^2(\mu) \to L^2(\mu)$ and $S : L^2(\nu) \to L^2(\nu)$, that is, π is a linear isometry satisfying $\pi S = T\pi$. This relationship means that the properties of the operators S, T on L^2 are the same: they have the same eigenvalues, equal eigenvalues have eigenspaces of equal dimensions, and more generally for $v \in L^2(\nu)$ the spectral measures of v and πv under S, T, respectively, are the same.

This shows ergodicity is an isomorphism invariant, since it is equivalent to the eigenvalue 1 having a 1-dimensional eigenspace. Similarly, mixing is an isomorphism invariant, since it can be characterized by the property $\langle f, T^n g\rangle \to \langle f, 1\rangle \langle g, 1\rangle$, where in L^2 language (the function 1 is characterized by the property that $\|1\|^2$ and $T1 = 1$; also -1 has this property by using -1 instead of 1 leads to the same condition).

Example 5.3.1. Let $\mathcal{B}$ be the Borel σ-algebra of S^1 and μ normalized length measure. For $\theta \in S^1$ let $R_\theta : S^1 \to S^1$ be given by $R_\theta z = \theta z$. Fix $\alpha_1, \alpha_2 \in \mathbb{R}$ and let $\theta_i = 2^{2\pi i \alpha_i} \in S^1$. Then $(S^1, \mathcal{B}, \mu, R_{\theta_1}) \cong (X, \mathcal{B}, \mu, R_{\theta_1})$ if and only if either α_1, α_2 are roots of unity of the same order, or $\alpha_1 = \pm\alpha_2$.

Proof. Let $\chi_n(z) = z^n$. Then $\{\chi_n)_{n\in\mathbb{Z}}$ form an orthonormal basis of $L^2(\mu)$ (for instance, they are uniformly dense in $C(S^1)$ by Stone-Weierstrass, and $C(S^1)$ is dense in L^2). Now

$$R_\theta \chi_n(z) = \chi_n(\theta z) = (\theta z)^n = \theta^n \chi_n(z)$$

so χ_n is an eigenfunction of R_θ with eigenvalue θ^n. It follows that the eigenvalues of R_θ are $\{\theta^n\}_{n\in\mathbb{Z}}$ (there are no other eigenvalues because, since $R_\theta : L^2 \to L^2$ is unitary, eigenvectors for different eigenvalues are orthogonal, but $\{\chi_n\}_{n\in\mathbb{Z}}$ spans a dense subspace of L^2, so its orthogonal complement is trivial).

Now, if the systems were isomorphic that $\theta_1 = (\theta_2)^n$ and $\theta_2 = (\theta_1)^m$ for some $m, n \in \mathbb{Z} \setminus \{0\}$, which means that $\alpha_1 = n\alpha_2 \bmod 1$ and $\alpha_2 = m\alpha_1 \bmod 1$, hence $\alpha_1 = mn\alpha_1 \bmod 1$. For this to occur, either $m = n = 1$ or $m = n = -1$, or else $\alpha_1 = k/(mn-1)$ for some $k \in \mathbb{Z}$. In the latter case also $\alpha_2 = mk/(mn-1)$, and evidently θ_1, θ_2 are roots of unity of the same order (since m is relatively prime to $mn - 1$). □

5.4 Spectral isomorphism of Bernoulli measures

A unitary operator $T : H \to H$ has countable Lebesgue spectrum if there are unit vectors $v_0, v_1, v_2, \ldots \in H$ that $\{v_0\} \cup \{T^n v_i\}_{n\in\mathbb{Z}, i\in\mathbb{N}}$ is an orthonormal basis of H. Any two such operators are unitarily equivalent, since if $T' : H' \to H'$ and $\{u_i'\}$ are another such system then we can define $\pi(T^n u_i) = (T')^n u_i'$ and this map extends to a unitary map $H \to H'$ with $T'\pi = \pi T$.

Let $(X_0, \mathcal{B}_0, \mu_0)$ be a nontrivial separable probability space (μ_0 is not a delta-measure). Let $X = X^{\mathbb{Z}}$, $\mathcal{B} = \mathcal{B}_0^{\mathbb{Z}}$ and $\mu = \mu_0^{\mathbb{Z}}$ be the product space and σ denote the shift. Let $H = L_0^2(\mu)$, i.e. the space of functions with integral 0. Note that $\sigma : H \to H$ is a unitary operator.

Claim 5.4.1. $\sigma|_H$ has countable Lebesgue spectrum.

Proof. Let $\{f_i\}_{i=0}^\infty$ be an orthonormal basis of $L^2(\mu_0)$ with $f_0 = 1$ (if $L^2(\mu_0)$ is finite dimensional there are finitely many functions).

Let I denote the set of maps $i : \mathbb{Z} \to \{1, 2, 3, \ldots\}$ such that $i(n) = 1$ for all but finitely many $n \in \mathbb{Z}$. For $i \in I$ let

$$f_i(x) = \prod_{n\in\mathbb{Z}} f_{i(n)}(x_n) = \prod_{n\in\mathbb{Z}} \sigma^n \widetilde{f}_{i(n)}$$

Note that the sum is only formally infinite, since all but finitely many of the terms are 1.

The family $\{f_i\}_{i\in I}$ is an orthogonal basis of $L_0^2(\mu)$. Indeed, $\|f_i\| = 1$, and

given $i, j \in I$, $|f_i| \equiv 1$ with $i \neq j$,

$$\begin{aligned}\langle f_i, f_j \rangle &= \int \prod_{n\in\mathbb{Z}} f_{i(n)}(x_n) f_{j(n)}(x_n) d\mu(x) \\ &= \prod_{n\in\mathbb{Z}} \int f_{i(n)}(x_n) f_{j(n)}(x_n) d\mu(x) \\ &= \prod_{n\in\mathbb{Z}} \int f_{i(n)}(\xi) f_{j(n)}(\xi) d\mu_0(\xi) \\ &= 0\end{aligned}$$

because $i(n) \neq j(n)$ for some n and $f_{i(n)} \perp f_{j(n)}$ in $L^2(\mu_0)$. To see that $\{f_i\}_{i\in I}$ span a dense subspace of L^2 note that it is immediate that the span is dense in the set of $L^1(\mu)$ functions depending on coordinates $-N, \ldots, N$, and the increasing union of these subspaces is evidently dense in $L^2(\mu)$.

Now, if we let σ denote the shift on $\mathbb{N}^{\mathbb{Z}}$, clearly

$$\sigma f_i = f_{\sigma i}$$

Let $\overline{1} = (\ldots 1, 1, 1, \ldots) \in I$, and $I' = I \setminus \{\overline{1}\}$. Then σ acts on I' without finite orbits, so we can find $I'' \subseteq I'$ such that $\{\sigma^n f_i\}_{n\in\mathbb{Z}, i\in I''} = \{f_i\}_{i\in I'}$, and this shows that σ has countable Lebesgue spectrum. □

In particular, one the spectral level, all shift maps with a product measures are isomorphic.

An early challenge in ergodic theory was:

Problem 5.4.2. Are the product measures $\{\frac{1}{2}, \frac{1}{2}\}^{\mathbb{Z}}$ and $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}^{\mathbb{Z}}$ with the shift map isomorphic?

As probability spaces the two are isomorphic, and spectrally they are isomorphic. It turns out they are not isomorphic; the tool that distinguishes them is entropy.

Chapter 6

Entropy

6.1 Shannon entropy

Let $(X, \mathcal{B}, \mu)$ be a probability space. A partition of X is a countable collection $\mathcal{A}$ of disjoint measurable sets whose union has full measure. We will focus almost exclusively on finite partitions $\mathcal{A} = \{A_1, \ldots, A_n\}$.

How can one quantify how "large" a partition is (relative to a measure)? A crude measure is $|\mathcal{A}|$, the number of elements in the partition. Only slightly more refined would be to count the sets $A \in \mathcal{A}$ of positive mass. Both these options ignore how mass is distributed. For example, a measure on two points may give them both mass $1/2$, or give one mass 0.9999 and mass 0.0001 to the other; certainly the first is more uniform than the second.

Definition 6.1.1. The Shannon entropy of μ with respect to $\mathcal{A}$ is the non-negative number

$$H_\mu(A) = -\sum_{A \in \mathcal{A}} \mu(A) \log \mu(A)$$

By convention the logarithm is taken in base 2 and $0 \log 0 = 0$. For infinite partitions $H_\mu(\mathcal{A})$ may be infinite.

Observe that $H_\mu(\mathcal{A})$ depends only on the probability vector $(\mu(A))_{A \in \mathcal{A}}$. For a probability vector $\underline{p} = (p_i)$ it is convenient to introduce the notation

$$H(\underline{p}) = H(p_1, p_2, \ldots) = -\sum_i p_i \log p_i$$

Example 6.1.2. For $\underline{p} = (p, 1-p)$ the entropy $H(p) = -p \log p - (1-p) \log(1-p)$ depends on the single variable p. It is an exercise in calculus to verify that $h(\cdot)$ is strictly concave on $[0, 1]$, increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$, with a unique maximum value $h(1/2) = 1$ and minimal values $h(0) = h(1) = 0$. Thus, the entropy is minimal when all the mass is on one atom of $\mathcal{A}$, and maximal when it is uniformly distributed.

Properties of entropy

(E1) $0 \le H(\mu, \mathcal{A}) \le \log |\mathcal{A}|$:

(a) $H(\mu, \mathcal{A}) = 0$ if and only if $\mu(A) = 1$ for some $A \in \mathcal{A}$.

(b) $H(\mu, \mathcal{A}) = \log |\mathcal{A}|$ if and only if μ is uniform on $\mathcal{A}$, that is, $\mu(A) = 1/|\mathcal{A}|$ for $A \in \mathcal{A}$.

(E2) $H(\cdot, \mathcal{A})$ is concave: for probability measures μ, ν on and $0 < \alpha < 1$,

$$H(\alpha\mu + (1-\alpha)\nu, \mathcal{A}) \ge \alpha H(\mu, \mathcal{A}) + (1-\alpha)H(\nu, \mathcal{A})$$

with equality if and only if $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$.

Proof. We first prove (E2). Since $f(t) = -t \log t$ is strictly concave, by Jensen's inequality,

$$\begin{aligned} H(\alpha\mu + (1-\alpha)\nu, \mathcal{A}) &= \sum_{A \in \mathcal{A}} f(\alpha\mu(A) + (1-\alpha)\nu(A)) \\ &\ge \sum_{A \in \mathcal{A}} (\alpha f(\mu(A)) + (1-\alpha) f(\nu(A))) \\ &= \alpha H(\mu, \mathcal{A}) + (1-\alpha) H(\nu, \mathcal{A}) \end{aligned}$$

with equality if and only if $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$.

The left inequality of (E1) is trivial. For the right one consider the function $F(\underline{p}) = -\sum_{A \in \mathcal{A}} p_A \log p_A$ on the simplex Δ of probability vectors $\underline{p} = (p_A)_{A \in \mathcal{A}}$. It suffices to show that the unique maximum is attained at $\underline{p}^* = (1/|\mathcal{A}|, \dots, 1/|\mathcal{A}|)$, since $F(\underline{p}^*) = \log |\mathcal{A}|$. The simplex Δ is compact and convex and by (E2), $H(\cdot)$ is strictly concave, so there is a unique maximizing point $\underline{p}^*$. Since $F(\cdot)$ is invariant under permutation of its variables, the maximizing point $\underline{p}^*$ must be similarly invariant, and hence all its coordinates are equal. Since it is a probability vector they are are equal to $1/|\mathcal{A}|$. □

For a set B of positive measure, let μ_B denote the conditional probability measure $\mu_B(C) = \mu(B \cap C)/\mu(B)$. Note that for a partition $\mathcal{B}$ we have the identity

$$\mu = \sum_{B \in \mathcal{B}} \mu(B) \cdot \mu_B \tag{6.1}$$

The *conditional entropy* of μ and $\mathcal{A}$ given another partition $\mathcal{B} = \{B_i\}$ is defined by

$$H(\mu, \mathcal{A}|\mathcal{B}) = \sum_{B \in \mathcal{B}} \mu(B) H(\mu_B, \mathcal{A})$$

This is just the average over $B \in \mathcal{B}$ of the entropy of $\mathcal{A}$ with respect to the conditional measure on B.

Definition 6.1.3. Let $\mathcal{A}, \mathcal{B}$ be partitions of the same space.

1. The *join* of $\mathcal{A}, \mathcal{B}$ is the partition

$$\mathcal{A} \vee \mathcal{B} = \{A \cap B \,:\, A \in \mathcal{A}\,,\, B \in \mathcal{B}\}$$

2. $\mathcal{A}$ *refines* $\mathcal{B}$ (up to measure 0) if every $A \in \mathcal{A}$ is contained in some $B \in \mathcal{B}$ (up to measure 0).

3. $\mathcal{A}, \mathcal{B}$ are *independent* if $\mu(A \cap B) = \mu(A)\mu(B)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$.

Properties of entropy (continued)

(E2') $H(\cdot, \mathcal{A}|\mathcal{B})$ is concave:

(E3) $H(\mu, \mathcal{A} \vee \mathcal{B}) = H(\mu, \mathcal{A}) + H(\mu, \mathcal{B}|\mathcal{A})$

(E4) $H(\mu, \mathcal{A} \vee \mathcal{B}) \geq H(\mu, \mathcal{A})$ with equality if and only if $\mathcal{A}$ refines $\mathcal{B}$ up to μ-measure 0.

(E5) $H(\mu, \mathcal{A} \vee \mathcal{B}) \leq H(\mu, \mathcal{A}) + H(\mu, \mathcal{B})$ with equality if and only if $\mathcal{A}, \mathcal{B}$ are independent. Equivalently, $H_\mu(\mathcal{B}|\mathcal{A}) \leq H(\mathcal{B})$ with equality if and only if $\mathcal{A}, \mathcal{B}$ are independent.

Proof. For (E3), by algebraic manipulation,

$$\begin{aligned}
H(\mu, \mathcal{A} \vee \mathcal{B}) \quad &= \\
&= \quad - \sum_{A \in \mathcal{A}, B \in \mathcal{B}} \mu(A \cap B) \log \mu(A \cap B) \\
&= \quad \sum_{A \in \mathcal{A}} \mu(A) \sum_{B \in \mathcal{B}} \frac{\mu(A \cap B)}{\mu(A)} \left(- \log \frac{\mu(A \cap B)}{\mu(A)} - \log \mu(A) \right) \\
&= \quad - \sum_{A \in \mathcal{A}} \mu(A) \log \mu(A) \sum_{B \in \mathcal{B}} \mu_A(B) - \sum_{A \in \mathcal{A}} \mu(A) \sum_{B \in \mathcal{B}} \mu_A(B) \log \mu_A(B) \\
&= \quad H(\mu, \mathcal{A}) + H(\mu, \mathcal{B}|\mathcal{A})
\end{aligned}$$

The inequality in (E4) follows from (E3) since $H(\mu, \mathcal{B}|\mathcal{A}) \geq 0$; there is equality if and only if $H(\mu_A, \mathcal{B}) = 0$ for all $A \in \mathcal{A}$ with $\mu(A) > 0$. By (E1), this occurs precisely when, on each $A \in \mathcal{A}$ with $\mu(A) \neq 0$, the measure μ_A is supported on a single atom of $\mathcal{B}$, which means that $\mathcal{A}$ refines $\mathcal{B}$ up to measure 0.

For (E2'), let $\mu = \alpha\eta + (1 - \alpha)\theta$. For $B \in \mathcal{B}$ let $\beta_B = \frac{\alpha\eta(B)}{\mu(B)}$. Then $(1 - \beta_B) = \frac{(1-\alpha)\theta(B)}{\mu(B)}$ and

$$\mu_B = \beta_B \eta_B + (1 - \beta_B)\theta_B$$

hence

$$\begin{aligned}
H(\mu, \mathcal{A}|\mathcal{B}) \quad &= \\
&= \quad \sum_{B \in \mathcal{B}} \mu(B) H(\mu_B, \mathcal{B}) && \text{by definition} \\
&\geq \quad \sum_{B \in \mathcal{B}} \mu(B) \left(\beta_B H(\eta_B, \mathcal{A}) + (1 - \beta_B) H(\theta_B, \mathcal{A}) \right) && \text{by concavity (E2)} \\
&= \quad \sum_{B \in \mathcal{B}} \left(\alpha\eta(B) \cdot H(\eta_B, \mathcal{A}) + (1 - \alpha)\theta(B) \cdot H(\theta_B, \mathcal{A}) \right) \\
&= \quad \alpha H(\eta, \mathcal{A}|\mathcal{B}) + (1 - \alpha) H(\theta, \mathcal{A}|\mathcal{B})
\end{aligned}$$

Finally, (E5) follows from (E1) an (E2). First,

$$H(\mu, \mathcal{B}|\mathcal{A}) = \sum_{B\in\mathcal{B}} \mu(B) H(\mu_B, \mathcal{A}) \leq H(\sum_{B\in\mathcal{B}} \mu(B)\mu_B, \mathcal{A}) = H(\mu, \mathcal{A})$$

It is clear that if $\mathcal{A}, \mathcal{B}$ are independent there is equality. To see this is the only way it occurs, one again uses strict convexity of $H(\underline{p})$, which shows that the independent case is the unique maximizer. □

There are a few generalizations of these properties which are useful:

Properties of entropy (continued):

1. $H(\mathcal{A}, \mathcal{B}|\mathcal{C}) = H(\mathcal{B}|\mathcal{C}) + H(\mathcal{A}|\mathcal{B} \vee \mathcal{C})$.
2. If $\mathcal{C}$ refines $\mathcal{B}$ then $H(\mathcal{A}|\mathcal{C}) \leq H(\mathcal{A}|\mathcal{B})$, with equality if and only if $\mathcal{B} = \mathcal{C}$.

Proof. For (1) expand both sides using (E3). For (2) use (1), noting that $\mathcal{C} = \mathcal{C} \vee \mathcal{B}$ since $\mathcal{C}$ refines $\mathcal{B}$. □

Corollary 6.1.4. *If $\mathcal{C}$ refined $\mathcal{B}$ then*

Remark. The definition of entropy may seem somewhat arbitrary. However, up to normalization, it is essentially the only possible definition if we wish (E1)–(E6) to hold. A proof of this can be found in Shannon's original paper on information theory and entropy, [**?**].

6.2 Entropy conditioned on a sigma-algebra

Let $(X, \mathcal{F}, \mu)$ be a probability space and $\mathcal{E} \subseteq \mathcal{F}$ a sub-σ-algebra. For a set $A \in \mathcal{F}$ write

$$\mu_x(A|\mathcal{E}) = \mathbb{E}(1_A|\mathcal{E})(x)$$

This is well defined for μ-a.e.x and is called the conditional probability of A given $\mathcal{E}$ (at x). Note that if $\mathcal{E}$ is generated by a finite partition $\mathcal{E}_0 = \{E_1, \ldots, E_N\}$ then, as we have seen, $\mathbb{E}(1_A|\mathcal{E})(x) = \mu(A \cap E_i)/\mu(E_i) = \mu_{E_i}(A)$ where $E_i = \mathcal{E}_0(x)$ is the element containing x, so $\mu_x(A|\mathcal{E}) = \mu_{\mathcal{E}_0(x)}(A)$.

Note that if $A_1, \ldots, A_n \in \mathcal{F}$ are disjoint then

$$\begin{aligned}
\mu_x(\bigcup_{i=1}^{n} A_i) &= \mathbb{E}(1_{\bigcup A_i}|\mathcal{E})(x) \\
&= \mathbb{E}(\sum_{i=1}^{n} 1_{A_i}|\mathcal{E})(x) \\
&= \sum_{i=1}^{n} \mathbb{E}(1_{A_i}|\mathcal{E})(x) \\
&= \sum_{i=1}^{n} \mu_x(A_i|\mathcal{E})
\end{aligned}$$

These equalities hold a.s. In particular if $\mathcal{A}$ is a finite sub-algebra of $\mathcal{F}$ then the above holds a.s. for all sequences of disjoint sets in $\mathcal{A}$ (since there are finitely many such sequences and for each it holds a.s.) and also a.s. $\mu_x(A|\mathcal{E}) \geq 0$, by positivity of the conditional expectation operator. Hence $A \mapsto \mu_x(A|\mathcal{E})$ is a probability measure on $\sigma(\mathcal{A})$ for a.e. x.

Remark 6.2.1. It is possible to show that $\mu_x(\cdot|\mathcal{E})$ extends to a probability measure on the full σ-algebra $\mathcal{F}$ in a way that satisfies many good properties, this is the disintegration of μ over $\mathcal{E}$. We will not use this.

Definition 6.2.2. Let $\mathcal{A}$ be a finite partition and and $\mathcal{E} \subseteq \mathcal{F}$ a sub-σ-algebra, Then the conditional entropy of $\mathcal{A}$ on $\mathcal{E}$ is

$$H_\mu(\mathcal{A}|\mathcal{E}) = \int H_{\mu_x(\cdot|\mathcal{E})}(\mathcal{A})\, d\mu(x)$$

Note that if $\mathcal{E}$ is generated by the partition $\{E_1, \ldots, E_N\}$ the formula above reduces to $\sum \mu(E_i) H_{\mu_{E_i}}(\mathcal{A})$ which is the usual definition of conditional entropy.

Lemma 6.2.3. *Let $\mathcal{A}$ be a finite partition and $\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq \mathcal{F}$ sub-σ-algebras. Then $H_\mu(\mathcal{A}|\mathcal{E}_1) \geq H_\mu(\mathcal{A}|\mathcal{E}_2)$.*

Proof. Since $L^2(\mathcal{E}_1) \subseteq L^2(\mathcal{E}_2)$, The composition of projections $L^2(\mathcal{F}) \to L^2(\mathcal{E}_2) \to L^2(\mathcal{E}_1)$ is just the projection $L^2(\mathcal{F}) \to L^2(\mathcal{E}_1)$. Since projection and conditional expectation agree on L^2, for $A \in \mathcal{A}$ we have

$$\mathbb{E}(\mu_x(A|\mathcal{E}_2)|\mathcal{E}_1) = \mathbb{E}(\mathbb{E}(1_A|\mathcal{E}_2)|\mathcal{E}_1) = \mathbb{E}(1_A|\mathcal{E}_1) = \mu_x(A|\mathcal{E}_1)$$

Let $\underline{p}(x) = (\mu_x(A|\mathcal{E}_1))_{A\in\mathcal{A}}$ and $\underline{q}(x) = (\mu_x(A|\mathcal{E}_2))_{A\in\mathcal{A}}$, so that $H_\mu(\mathcal{A}|\mathcal{E}_1) = \int H(\underline{p}(x))d\mu(x)$, $H_\mu(\mathcal{A}|\mathcal{E}_2) = \int H(\underline{q}(x))d\mu(x)$, and by the above $\underline{p}(x) = \mathbb{E}(\underline{q}(x)|\mathcal{E}_1)$. Thus, by concavity of the entropy function,

$$\begin{aligned} H_\mu(\mathcal{A}|\mathcal{E}_1) &= \int H\left(\mathbb{E}(\underline{q}(x)|\mathcal{E}_1)\right)\, d\mu(x) \\ &\geq \int \mathbb{E}\left(H(\underline{q}(x))|\mathcal{E}_1\right)\, d\mu(x) \\ &= \int H(\underline{q}(x))\, d\mu(x) \\ &= H_\mu(\mathcal{A}, \mathcal{E}_2) \qquad \square \end{aligned}$$

Lemma 6.2.4. *Let $\mathcal{E} \subseteq \mathcal{F}$ be a sub-σ-algebra and $\mathcal{A}$ a partition. Then $H(\mu, \mathcal{A}|\mathcal{E}) = 0$ if and only if $\mathcal{A}$ is $\mathcal{E}$-measurable up to measure 0.*

Proof. The claim follows from the following chain of equivalent statements:

1. $\mathcal{A}$ is $\mathcal{E}$-measurable up to measure 0.
2. $\mathbb{E}(1_A|\mathcal{E}) = 1_A$ a.e. for all $A \in \mathcal{A}$ (e.g. because in L^2 conditional expectation is the projection to $L^2(\mathcal{E})$ and its fixed-point set is its range).

3. $\mu_x(\cdot|\mathcal{E})$ is supported on a single $\mathcal{A}$-atom for a.e. x (Since $\mu_x(A|\mathcal{E}) = \mathbb{E}(1_A|\mathcal{E})(x)$, if $\mathbb{E}(1_A|\mathcal{E})(x) = 1_A$ a.e. for $A \in \mathcal{A}$ then a.s. μ_x takes only values $0, 1$. Conversely suppose it is atomic meaning $\mathbb{E}(1_A|\mathcal{E})$ is $0, 1$-valued for each $A \in \mathcal{A}$. Then $\mathbb{E}(1_A|\mathcal{E})(x) = 1_E$ for some set $E \in \mathcal{E}$. We have

$$\mathbb{E}(1_{X\setminus E}1_A|\mathcal{E}) = 1_{X\setminus E}\mathbb{E}(1_A|\mathcal{E}) = 1_{X\setminus E}1_E = 0$$

Since $1_{X\setminus E}1_A \geq 0$ and conditional expectation is a positive integral-preserving operator, this implies that $1_{X\setminus E}1_A = 0$ a.e., or $A \subseteq E$. If $A \neq E$ mod μ then $\mu(A) < \mu(E)$ and we would have $\int 1_A < \int 1_E = \int \mathbb{E}(1_A|\mathcal{E})$ which is impossible).

4. $H_{\mu_x}(\mathcal{A}|\mathcal{E}) = 0$ a.s.

5. $H_\mu(\mathcal{A}|\mathcal{E}) = 0$ (because $H_\mu(\mathcal{A}|\mathcal{E}) = \int H_{\mu_x}(\mathcal{A})d\mu(x)$ and $H_{\mu_x}(\mathcal{A}) \geq 0$).

□

Recall that to families of sets $\mathcal{A}$ and $\mathcal{E}$ are independent if $\mu(A \cap E) = \mu(A)\mu(E)$ for $A \in \mathcal{A}$ and $E \in \mathcal{E}$.

Lemma 6.2.5. *Let $\mathcal{E} \subseteq \mathcal{F}$ be a sub-σ-algebra and $\mathcal{A}$ a partition. Then $H_\mu(\mathcal{A}|\mathcal{E}) = H_\mu(\mathcal{A})$ if and only if $\mathcal{A}$ is independent of $\mathcal{E}$.*

Proof. If $\mathcal{A}$ is independent of $\mathcal{E}$ then $1_A - \mu(A) \perp 1_E$ for all $E \in \mathcal{E}$. By approximation we find that $1_A - \mu(A) \perp L^2(\mathcal{E})$. Since conditional expectation is projection this shows that $\mu_x(A|\mathcal{E}) = \mu(A)$ a.s., so $H_{\mu_x(\cdot|\mathcal{E})}(\mathcal{A}) = H_\mu(\mathcal{A})$ a.s., so $H_\mu(\mathcal{A}|\mathcal{E}) = \int H_{\mu_x(\cdot|\mathcal{E})}(\mathcal{A})d(x) = H_\mu(\mathcal{A})$.

Conversely suppose $H_\mu(\mathcal{A}|\mathcal{E}) = H_\mu(\mathcal{A})$. Let $E \in \mathcal{E}$ and $\mathcal{E}_0 = \{E, X \setminus E\}$. Then

$$H_\mu(\mathcal{A}) \geq H_\mu(\mathcal{A}|\mathcal{E}_0) \geq H_\mu(\mathcal{A}|\mathcal{E})$$

All are equalities, hence $\mathcal{A}$ is independent of E and $x\setminus E$, since $E \in \mathcal{E}$ is arbitrary $\mathcal{A}$ is independent of $\mathcal{E}$. □

Proposition 6.2.6. *Let $\mathcal{E}_1 \subseteq \mathcal{E}_2 \subseteq \ldots \subseteq \mathcal{F}$ be σ-algebras and $\mathcal{E} = \sigma(\mathcal{E}_1, \mathcal{E}_2, \ldots)$. Then for any finite partition $\mathcal{A}$,*

$$\lim_{n\to\infty} H_\mu(\mathcal{A}|\mathcal{E}_n) = H_\mu(\mathcal{A}|\mathcal{E})$$

Proof. By the martingale convergence theorem, $\mu_x(A|\mathcal{E}_n) = \mathbb{E}(1_A|\mathcal{E}_n) \to \mathcal{E}(1_A|\mathcal{E}) = \mu_x(A)$ a.e. for every $A \in \mathcal{A}$, and hence a.e. simultaneously for $A \in \mathcal{A}$. Thus by continuity of $\underline{p} \mapsto H(\underline{p})$, and identifying $\mu_x(\cdot|\mathcal{E})$ with the probability vector indexed by $\mathcal{A}$, we have

$$H_{\mu_x(\cdot|\mathcal{E}_n)}(\mathcal{A}) = H(\mu_x(\cdot|\mathcal{E}_n)) \to H(\mu_x(\cdot|\mathcal{E})) = H_{\mu_x(\cdot|\mathcal{E})}(\mathcal{A})$$

a.e. as $n \to \infty$. Also by (E1) we have the uniform bound

$$H_{\mu_x(\cdot|\mathcal{E}_n)}(\mathcal{A}) \leq \log|\mathcal{A}|$$

Therefore by bounded convergence

$$H_\mu(\mathcal{A}|\mathcal{E}_n) = \int H_{\mu_x(\cdot|\mathcal{E}_n)}(\mathcal{A})\, d\mu(x) \to \int H_{\mu_x(\cdot|\mathcal{E})}(\mathcal{A})\, d\mu(x) = H_\mu(\mathcal{A}|\mathcal{E}_n)$$

as $n \to \infty$. □

Remark 6.2.7. Mere generally one can show that

$$H_\mu(\mathcal{A}|\mathcal{E}) = \sup\{H_\mu(\mathcal{A}|\mathcal{E}_0) \,:\, \mathcal{E}_0 \subseteq \mathcal{E} \text{ a finite sub-}\sigma\text{-algebra}\}$$

When $\mathcal{B}_n$ are finite partitions, $\bigvee_{k=1}^n \mathcal{B}_k$ is a finite partition which we often identify with the algebra it generates. However$\bigvee_{k=1}^\infty \mathcal{B}_k$ is generally not a finite partition and we define it instead to be the σ-algebra $\sigma(\mathcal{B}_1, \mathcal{B}_2, \ldots)$. With this convention and the theorem above, we have

$$H_\mu(\mathcal{A}| \bigvee_{k=1}^{n} \mathcal{B}_k) \to H_\mu(\mathcal{A}| \bigvee_{k=1}^{\infty} \mathcal{B}_k)$$

6.3 Entropy of discrete random variables

Every partition $\mathcal{A} = \{A_i\}_{i\in I}$ of a probability space $(X, \mathcal{F}, \mu)$ defines an I-valued random variable $x \mapsto \mathcal{A}(x)$, where $\mathcal{A}(x) = i$ if and only if $x \in \mathcal{A}_i$.

Conversely, if $\xi : X \to I$ is a random variable and I is countable then the sets $A_i = \xi^{-1}(i)$ form a partition $\mathcal{A}$ of X, and evidently $\mathcal{A}(x) = \xi(x)$.

If $\xi : X \to I$ and $\zeta : Y \to J$ are random variables corresponding partitions $\mathcal{A}, \mathcal{B}$ respectively, then the pair (ξ, ζ) is an $I \times J$-valued random variable corresponding to the partition $\mathcal{A} \vee \mathcal{B}$. The random variables are independent if and only the corresponding partitions are.

Definition 6.3.1. Let $\mathcal{A}, \mathcal{B}$ be the partitions corresponding to random variables ξ, ζ defined on a common probability space. Then we denote

$$H(\xi) = H(\mathcal{A}) \qquad , \qquad H(\xi, \zeta) = H(\mathcal{A} \vee \mathcal{B}) \qquad , \qquad H(\xi|\zeta) = H(\mathcal{A}|\mathcal{B})$$

etc.

One interprets $H(\xi)$ as a measure of the randomness of ξ: If it takes on 1 value e.s. then $H(\xi) = 0$, if it takes on n values then $H(\xi) \leq \log n$ with equality if and only if $\xi(a) = \frac{1}{n}$ for each of these values; etc.

6.4 Entropy of a partition in a measure-preserving system

For a map $f : X \to Y$ and a partition $\mathcal{A}$ of Y write $f^{-1}\mathcal{A} = \{f^{-1}A \,:\, A \in \mathcal{A}\}$. This is a partition of X. Assuming f is measurable, $f^{-1}\mathcal{A}$ consists of measurable

sets, and given measures μ on X and ν on Y and assuming that f preserves the measure, the probability vector $(\mu(f^{-1}A)_{A\in A})$ is equal to $(\nu(A))_{A\in\mathcal{A}}$, hence

$$H(\nu,\mathcal{B}) = H_\mu(f^{-1}\mathcal{A})$$

Definition 6.4.1. Let $(X,\mathcal{F},\mu,T)$ be a measure-preserving system and $\mathcal{A}$ a partition of X. For $n\geq 0$ we write

$$\mathcal{A}^n = \bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}$$

Note that if ξ is the random variable corresponding to $\mathcal{A}$ then $\mathcal{A}^n$ is the random variable corresponding to the vector $(\xi, T\xi, \ldots, T^{n-1}\xi)$, which is the initial n variables of the stationary process (ξ_n) where $\xi_n = T^n\xi$.

The partition $\mathcal{A}^n$ is obtained by iteratively splitting each atom of $\mathcal{A}^{n-1}$ into at most $|\mathcal{A}|$ parts (according to $T^{-n}\mathcal{A}$). These parts may have unequal (relative) masses. In particular the masses of the atoms of $\mathcal{A}^n$ may be unequal. Entropy measures the average mass of these atoms in logarithmic scale.

Definition 6.4.2. The entropy $h_\mu(T,\mathcal{A})$ of a partition $\mathcal{A}$ of a measure preserving system $(X,\mathcal{F},\mu,T)$ is the limit

$$\begin{aligned} h_\mu(T,\mathcal{A}) &= \lim_{n\to\infty}\frac{1}{n}H_\mu(\mathcal{A}^n) \\ &= \lim_{n\to\infty}\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}) \end{aligned}$$

Lemma 6.4.3. *The limit in the definition exists.*

Proof. Let

$$a_n = H_\mu(\mathcal{A}^n)$$

Then the existence of the limit will follow if we show that a_n is sub-additive, i.e. $a_{m+n}\leq a_m + a_n$. Indeed,

$$\begin{aligned} a_{m+n} &= H_\mu(\bigvee_{k=0}^{m+n-1} T^{-k}\mathcal{A}) \\ &= H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{A} \vee \bigvee_{k=m}^{m+n-1} T^{-k}) \\ &\leq H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{A}) + H_\mu(\bigvee_{k=m}^{m+n-1} T^{-k}) \qquad \text{by (E3)} \\ &= a_m + H_\mu(T^{-m}\mathcal{A}^n) \\ &= a_m + H_\mu(\mathcal{A}^n) \\ &= a_m + a_n \end{aligned}$$

□

Example 6.4.4. Let A be a finite set, $\mu_0 \in \mathcal{P}(A)$, let $X = A^{\mathbb{Z}}$ and $\mu = \mu_0^{\mathbb{Z}}$ the product measure. Let σ denote the shift map. Let $\mathcal{A}$ be the partition of X according tot he 0-th coordinate, i.e. for $a \in A$ let

$$[a] = \{x \in A^{\mathbb{Z}} : a_0 = a\}$$

and$\mathcal{A} = \{[a]\}_{a\in A}$. Then

$$T^{-k}[a] = \{x \in A^{\mathbb{Z}} : x_k = a\}$$

and because μ is a product measure, it follows that the partitions $T^{-k}\mathcal{A}$, $k = 0, 1, \ldots$, form an independent family. Therefore by an iterated application of (E5),

$$H_\mu(\mathcal{A}^n) = nH_\mu(\mathcal{A}) = nH(\mu_0)$$

where we identify μ_0 with the probability vector $(\mu_0(a))_{a\in A}$. Thus

$$h_\mu(T, \mathcal{A}) = nH(\mu_0)$$

Example 6.4.5. Let $X = S^1$, μ length measure and R_θ a rotation. Let $\mathcal{A}$ denote the partition of X into northern and southern hemispheres (with some convention for the endpoints). Then $R_\rho^{-n}\mathcal{A}$ is also a partition into two intervals. The partition $\mathcal{A}^n$ is then also a partition into intervals, and these are determined by the endpoints of the intervals $T^{-k}\mathcal{A}$, $k = 0, \ldots, n-1$. There are at most $2n$ such endpoints (exactly $2n$ if θ is irrational) and so $\mathcal{A}^n$ consists of at most $2n$ intervals. Hence $H_\mu(\mathcal{A}^n) \le \log 2n$ by (E1) and

$$0 \le h_\mu(T, \mathcal{A}) \le \lim_{n\to\infty} \frac{\log 2n}{n} = 0$$

so $h_\mu(\mathcal{A}^n) = 0$.

Lemma 6.4.6 (Elemntary properties)**.** *1.* $0 \le h_\mu(T, \mathcal{A}) \le \log |\mathcal{A}|$

2. $h_\mu(T, \mathcal{A}) \le h_\mu(T, \mathcal{A} \vee \mathcal{B}) \le h_\mu(T, \mathcal{A}) + h_\mu(T, \mathcal{B})$

3. $h_\mu(T, \mathcal{A}) = h_\mu(T, \mathcal{A}^k)$ *for all* $k \ge 1$.

4. $h_\mu(T^m, \mathcal{A}^m) = mh_\mu(T, \mathcal{A})$ *for all* $k \in \mathbb{N}$.

5. If T *is invertible then* $h_\mu(T, \mathcal{A}) = h_\mu(T^{-1}, \mathcal{A})$.

Proof. These are all easy consequences of the properties of Shannon entropy. For example,. to prove (3) note that

$$\begin{aligned}
(\mathcal{A}^m)^n &= \bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}^m \\
&= \bigvee_{k=0}^{n-1} T^{-k}\Big(\bigvee_{j=0}^{m-1} T^{-j}\mathcal{A}\Big) \\
&= \bigvee_{k=0}^{n+m-1} T^{-k}\mathcal{A} \\
&= \mathcal{A}^{n+m-1}
\end{aligned}$$

so

$$\lim_{n\to\infty}\frac{1}{n}H_\mu((\mathcal{A}^m)^n)=\lim_{n\to\infty}\frac{n+m-1}{n}\cdot\frac{1}{n+m-1}H_\mu(\mathcal{A}^{n+m-1})=h_\mu(T,\mathcal{A})\ \square$$

Proposition 6.4.7. *Let $(X,\mathcal{F},\mu,T)$ be a measure-preserving system and $\mathcal{A}$ a finite partition. Then*

$$h_\mu(T,\mathcal{A})=H_\mu(\mathcal{A}|\bigvee_{k=1}^{\infty}T^{-k}\mathcal{A})$$

Proof. Using (E3) we have

$$\begin{aligned}H_\mu(\bigvee_{k=0}^{n-1}T^{-k}\mathcal{A}) &= H_\mu((\bigvee_{k=0}^{n-2}T^{-k}\mathcal{A})\vee T^{-(n-1)}\mathcal{A})\\ &= H_\mu(T^{-(n-1)}\mathcal{A})+H_\mu(\bigvee_{k=0}^{n-2}T^{-k}\mathcal{A}|T^{-(n-1)}\mathcal{A})\end{aligned}$$

Iterating this and using the previous lemma, we have

$$H_\mu(\bigvee_{k=0}^{n-1}T^{-k}\mathcal{A})=\sum_{k=0}^{n-1}H_\mu(T^{-k}\mathcal{A}|\bigvee_{m=k+1}^{n-1}T^{-m}\mathcal{A})$$

Now by the measure preserving property, $H_\mu(T^{-i}\mathcal{B}|T^{-i}\mathcal{C})=H_\mu(\mathcal{B}|\mathcal{C})$. Therefore the above is

$$\begin{aligned}H_\mu(\bigvee_{k=0}^{n-1}T^{-k}\mathcal{A}) &= \sum_{k=0}^{n-1}H_\mu(\mathcal{A}|\bigvee_{m=1}^{n-k-1}T^{-m}\mathcal{A})\\ &=\end{aligned}$$

Since $H_\mu(\mathcal{A}|\bigvee_{m=1}^{k}T^{-m}\mathcal{A})\to H_\mu(\mathcal{A}|\bigvee_{m=1}^{\infty}T^{-m}\mathcal{A})$ as $m\to\infty$, by (E5), hence by Cesaro's theorem,

$$\begin{aligned}h_\mu(T,\mathcal{A}) &= \lim_{n\to\infty}\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1}T^{-k}\mathcal{A})\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}H_\mu(\mathcal{A}|\bigvee_{m=1}^{k}T^{-m}\mathcal{A})\\ &= H_\mu(\mathcal{A}|\bigvee_{m=1}^{\infty}T^{-m}\mathcal{A})\end{aligned}$$

as claimed. $\square$

Remark 6.4.8. From the proof and the fact that the sequence $H_\mu(\mathcal{A}|\bigvee_{m=1}^{k}T^{-m}\mathcal{A})$ is non-increasing in k we find that in fact $\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1}T^{-k}\mathcal{A})$ is decreasing as well.

Example 6.4.9. Here is a surprising application to stochastic processes. Let $(\xi_n)_{n=-\infty}^{\infty}$ be an ergodic stationary finite-valued process and write $\mathcal{A}_{\xi_n}$ for the partition associated to ξ_n. Assuming as we may that the process arises from a m.p.s. then $\xi_n = T^n\xi_0$ and $\mathcal{A}_{\xi_n} = T^{-n}\mathcal{A}_{\xi_0}$.

The process is said to be *deterministic* if ξ_0 is determined a.s. by $(\xi_n)_{n<0}$, or in other words, $\mathcal{A}_{\xi_0}$ is measurable (mod μ) with respect to $\bigvee_{n=-\infty}^{-1}\mathcal{A}_{\xi_n} = \bigvee_{n=-\infty}^{-1} T^{-n}\mathcal{A}_{\xi_0}$. This is the same as $h_\mu(T, \mathcal{A}_{\xi_0}) = H(\mathcal{A}_\xi | \bigvee_{n=-\infty}^{-1} T^{-n}\mathcal{A}_{\xi_0}) = 0$. But this is also the same as $H(\mathcal{A}_\xi | \bigvee_{n=1}^{\infty} T^{-n}\mathcal{A}_{\xi_0}) = 0$, so ξ_0 is measurable with respect to $\xi_1, \xi_2, \ldots$.

Thus if the past determines the future, the future determines the past! There is no known proof of this fact that does not use entropy.

6.5 Entropy of a measure preserving system

We have defined the entropy of a partition in a m.p.s. However, different partitions can give different entropies. For example, in any system the trivial partition into one set has entropy zero. To obtain a number associated to the system alone we have the following.

Definition 6.5.1. The Alamogordo-Sinai entropy (or just entropy) of a measure preserving system $(X, \mathcal{F}, \mu, T)$ is

$$h_\mu(T) = \sup\{h_\mu(T, \mathcal{A}) \,:\, \mathcal{A} \text{ a finite patition of } X\}$$

It is possible to have $h_\mu(T) = \infty$. Indeed the entropy $h_\mu(T, \mathcal{A})$ is finite when $\mathcal{A}$ is finite but the upper bound $\log|\mathcal{A}|$ tends to infinity when the size of the partition does.

Proposition 6.5.2. *Entropy is an isomophism invariant, i.e. isomorphic systems have the same entropy.*

Proof. Suppose $(X_i, \mathcal{F}_i, \mu_i, T_i)$, $i = 1, 2$, are m.p.s.s and f an isomorphism between them. For any sets $B_0, \ldots, B_k \in \mathcal{F}_2$ and $B = \bigcap_{i=0}^{k} T^{-1}B_k$, we have

$$f^{-1}(B) = \bigcap_{i=0}^{k} T^{-i}(f^{-1}A_i)$$

It follows that for any partition $\mathcal{B}$ of $\mathcal{F}_2$ and $\mathcal{A} = f^{-1}\mathcal{B}$, there is a measure-preserving identification between $\mathcal{B}^k$ and $\mathcal{A}^k$ given by f^{-1}, hence $H_\mu(\mathcal{A}^k) = H_\mu(\mathcal{B}^k)$ and so $h_\mu(T_1, \mathcal{A}) = h_\mu(T_2, \mathcal{B})$. Thus

$$\begin{aligned} h_\mu(T_!) &= \sup\{h_\mu(T_1, \mathcal{A}) \,:\, \mathcal{A} \text{ a finite patition of } X_1\} \\ &\geq \sup\{h_\mu(T_1, f^{-1}(\mathcal{B})) \,:\, \mathcal{B} \text{ a finite patition of } X_2\} \\ &= \sup\{h_\mu(T_2, \mathcal{B}) \,:\, \mathcal{B} \text{ a finite patition of } X_2\} \\ &= h_\mu(T_2) \end{aligned}$$

The reverse inequality follows by symmetry, proving the claim. □

Lemma 6.5.3. $h_\mu(T^k) = |k| h_\mu(T)$.

Proof. Let $k \in \mathbb{N}_+$. For any finite partition $\mathcal{A}$ we saw that $h_\mu(T, \mathcal{A}) = k h_\mu(T^k, \mathcal{A}^k)$, hence $h_\mu(T) \leq k h_\mu(T^k)$. On the other hand since $\bigvee_{j=0}^{k-1-j} \mathcal{A}$ refines $\mathcal{A}$,

$$\begin{aligned} h_\mu(T^k, \mathcal{A}) &\leq h_\mu(T^k, \bigvee_{j=0}^{k-1} T^{-j}\mathcal{A}) \\ &= \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} \bigvee_{j=0}^{k-1} (T^{-ki-j}\mathcal{A})) \\ &= \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{kn-1} T^{-i}\mathcal{A}) \\ &= k h_\mu(T, \mathcal{A}) \end{aligned}$$

For $k = -1$, we already saw that $h_\mu(T, \mathcal{A}) = h_\mu(T^{-1}, \mathcal{A})$ for all $\mathcal{A}$ and the claim follows. For general $k < 0$ it follows by applying the first part to T^{-1}. □

Calculating entropy is potentially difficult, since one must take into account all partitions. In practice, it is enough to consider a dense family of partitions, and sometimes even a single one.

Definition 6.5.4. A partition $\mathcal{A}$ in an invertible measure preserving system $(X, \mathcal{F}, \mu, T)$ is a generating partition if $\bigvee_{n=-\infty}^{\infty} T^{-n}\mathcal{A} = \mathcal{F}$ up to μ-measure 0 (that is $\mathcal{F} = \sigma(\mathcal{A}_n : n \in \mathbb{Z})$). If $\bigvee_{n=0}^{\infty} T^{-n}\mathcal{A} = \mathcal{F}$ we say that $\mathcal{A}$ is a one-sided generator (this definition makes sense also when T is not invertible).

Proposition 6.5.5. *Let $\mathcal{A}, \mathcal{B}$ be partitions in an invertible measure preserving system $(X, \mathcal{F}, \mu, T)$. Then*

$$h_\mu(T, \mathcal{A} \vee \mathcal{B}) = h_\mu(\mathcal{B}) + H_\mu(\mathcal{A} | \bigvee_{n=1}^{\infty} T^{-n}\mathcal{A} \vee \bigvee_{n=-\infty}^{\infty} T^{-n}\mathcal{B})$$

and in any system (even not invertible),

$$h_\mu(T, \mathcal{A} \vee \mathcal{B}) \leq h_\mu(\mathcal{B}) + H_\mu(\mathcal{A} | \bigvee_{n=1}^{\infty} T^{-n}\mathcal{A} \vee \bigvee_{n=0}^{\infty} T^{-n}\mathcal{B})$$

Proof. For each n, using (E3) once and then again inductively as in Proposition

??,

$$\begin{aligned} H_\mu(\bigvee_{k=0}^{n-1} T^{-k}(\mathcal{A}\vee\mathcal{B})) &= H_\mu((\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A})\vee(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B})) \\ &= H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}) + H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}|\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}) \\ &= H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}) + \sum_{m=0}^{n-1} H_\mu(T^{-m}\mathcal{A}|\bigvee_{k=m+1}^{n-1} T^{-k}\mathcal{A}\vee\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}) \\ &= H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}) + \sum_{m=0}^{n-1} H_\mu(\mathcal{A}|\bigvee_{k=1}^{n-m-1} T^{-k}\mathcal{A}\vee\bigvee_{k=-m}^{n-m-1} T^{-k}\mathcal{B}) \end{aligned}$$

Dividing by n and taking $n\to\infty$ the left hand side and the first term on the right tend to $h_\mu(T,\mathcal{A}\vee\mathcal{B})$ and $h_\mu(T,\mathcal{B})$ respectively. To evaluate the limit

$$\lim_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1} H_\mu(\mathcal{A}|\bigvee_{k=1}^{n-m-1} T^{-k}\mathcal{A}\vee\bigvee_{k=-m}^{n-m-1} T^{-k}\mathcal{B})$$

note that for every m we have

$$H_\mu(\mathcal{A}|\bigvee_{k=1}^{n-m-1} T^{-k}\mathcal{A}\vee\bigvee_{k=-m}^{n-m-1} T^{-k}\mathcal{B}) \geq H_\mu(\mathcal{A}|\bigvee_{k=1}^{\infty} T^{-k}\mathcal{A}\vee\bigvee_{k=-\infty}^{\infty} T^{-k}\mathcal{B})$$

hence the right hand side is a lower bound for the limit. On the other hand, for $m>\sqrt{n}$ we have

$$H_\mu(\mathcal{A}|\bigvee_{k=1}^{n-m-1} T^{-k}\mathcal{A}\vee\bigvee_{k=-m}^{n-m-1} T^{-k}\mathcal{B}) \leq H_\mu(\mathcal{A}|\bigvee_{k=1}^{\sqrt{n}} T^{-k}\mathcal{A}\vee\bigvee_{k=-\sqrt{n}}^{\sqrt{n}} T^{-k}\mathcal{B})$$

and for every m the terms are bounded by $\log|\mathcal{A}|$, hence

$$\sum_{m=0}^{n-1} H_\mu(\mathcal{A}|\bigvee_{k=1}^{n-m-1} T^{-k}\mathcal{A}\vee\bigvee_{k=-m}^{n-m-1} T^{-k}\mathcal{B}) \leq \frac{\sqrt{n}}{n}\cdot|\mathcal{A}|+\frac{n-\sqrt{n}}{n}\cdot H_\mu(\mathcal{A}|\bigvee_{k=1}^{\sqrt{n}} T^{-k}\mathcal{A}\vee\bigvee_{k=-\sqrt{n}}^{\sqrt{n}} T^{-k}\mathcal{B})$$

The left hand side tends to $H_\mu(\mathcal{A}|\bigvee_{k=1}^{\infty} T^{-k}\mathcal{A}\vee\bigvee_{k=-\infty}^{\infty} T^{-k}\mathcal{B})$, completing the proof.

In the non-invertible case we start with the same identity and note that conditioning only on $T^{-k}\mathcal{B}$ for non-negative k only can only increase the entropy. The rest is the same. □

Theorem 6.5.6. *Let $\mathcal{B}$ be a generating (or one-sided generating) partition in a measure preserving system $(X,\mathcal{F},\mu,T)$. Then $h_\mu(T)=h_\mu(T,\mathcal{B})$.*

Proof. We prove the case of an invertible system, the other is similar. We must show that $h_\mu(T,\mathcal{A}) \le h_\mu(T,\mathcal{B})$ for any finite partition $\mathcal{A}$. Indeed, fixing $\mathcal{A}$,

$$\begin{aligned} h_\mu(T,\mathcal{A}) &\le h_\mu(T,\mathcal{A}\vee\mathcal{B}) \\ &= h_\mu(T,\mathcal{B}) + h_\mu(T,\mathcal{A}|\bigvee_{k=1}^{\infty} T^{-k}\mathcal{A}\vee\bigvee_{k=-\infty}^{\infty} T^{-k}\mathcal{B}) \\ &\le h_\mu(T,\mathcal{B}) + h_\mu(T,\mathcal{A}|\bigvee_{k=-\infty}^{\infty} T^{-k}\mathcal{B}) \\ &= h_\mu(T,\mathcal{B}) \end{aligned}$$

by Lemma **??**. □

Corollary 6.5.7. *Let μ_0 be a measure on a finite set A. Then the entropy of the product system $\mu_0^{\mathbb{Z}}$ with the shift is $H(\mu_0)$. In particular the product measures $\{\frac{1}{2},\frac{1}{2}\}^{\mathbb{Z}}$ and $\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}^{\mathbb{Z}}$ with the shift maps are not isomorphic.*

Proof. For a finite set A, the partition according to the 0-coordinate generates in the system $(A^{\mathbb{Z}},\sigma)$, so the entropy is the entropies of this partition, which is just $H(\mu_0)$. □

In the absence of a generating partition, entropy can also be computed as follows.

Theorem 6.5.8. *Let $\mathcal{E}$ be an algebra of sets which generate together the σ-algebra $\mathcal{F}$ of a measure preserving system $(X,\mathcal{F},\mu,T)$. Then $h_\mu(T) = \sup h_\mu(T,\mathcal{B})$, where the supremum is over $\mathcal{E}$-measurable partitions $\mathcal{B}$.*

Proof. Write $\alpha = \sup H_\mu(T,\mathcal{B})$ for $\mathcal{B}$ as above. Evidently $H_\mu(T) \ge \alpha$ so we only need to prove the opposite inequality. For any finite partition $\mathcal{A}$, it is possible to find a refining sequence $\mathcal{C}_n$, $n = 1,2,\dots$, of $\mathcal{E}$-measurable partitions such that $\mathcal{A} \in \bigvee_{n=1}^{\infty}\mathcal{C}_n$. Then the argument in the proof of the previous theorem shows that $H_\mu(T,\mathcal{A}) \le \lim H_\mu(T,\mathcal{C}_n) \le \alpha$. □

Chapter 7

The Shannon-McMillan-Breiman Theorem

7.1 Example: Bernoulli measures

The partition $\mathcal{A}^n = \bigvee_{i=0}^{n-1} T^{-i}\mathcal{A}$ is obtained by iteratively splitting each atom of $\mathcal{A}^{n-1}$ into at most $|\mathcal{A}|$ parts (the atoms of $T^{-n}\mathcal{A}$). These parts may have unequal (relative) masses, and thus the masses of the atoms of $\mathcal{A}^n$ may be unequal. The entropy $h_\mu(T, \mathcal{A})$ measures the *average* mass of these atoms in logarithmic scale. It turns out that the typical atoms does not get very far from the mean (and even more is true).

Example 7.1.1. Let $(\xi_n)_{n=0}^\infty$ be a $\{0,1\}$-valued n i.i.d. process with $\mathbb{P}(\xi_n = 0) = p$ and $\mathbb{P}(\xi_n = 1) = 1 - p$ for some $0 < p \leq \frac{1}{2}$. If $p = \frac{1}{2}$ then for every sequence $a \in \{0,1\}^n$, $\mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_n) = 2^{-n}$, independent of choice of a. But if $p < \frac{1}{2}$ then different sequences may yield different probabilities, the minimal one being $a = 00\dots0$ with probability p^n and the largest being $a = 11\dots1$ with probability $(1-p)^n$. In general, writing $p_0 = p$ and $p_1 = 1-p$, we have

$$\begin{aligned} \mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_2) &= \prod_{i=1}^n p_{a_i} \\ &= p^{\#\{1 \leq i \leq n \,:\, a_i = 0\}} \cdot p^{\#\{1 \leq i \leq n \,:\, a_i = 1\}} \end{aligned}$$

Now, for an infinite realization $a \in \{0,1\}^\infty$ of the process, by the ergodic theorem (or law of large numbers),

$$\#\{1 \leq i \leq n \,:\, a_i = 0\} = n(p + o(1))$$

and

$$\#\{1 \leq i \leq n \,:\, a_i = 1\} = n(1 - p + o(1))$$

Therefore with probability one over the choice of a,

$$\begin{aligned} \mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_2) &= p^{n(p+o(1))}(1-p)^{n(1-p+o(1))} \\ &= 2^{n(-p\log p-(1-p)\log(1-p)+o(1))} \\ &= 2^{nH(p)+o(n)} \end{aligned}$$

In other words, with probability one over the choice of a,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_2) = H(p)$$

So, although different realizations have initial segments with different probabilities, asymptotically the probabilities are a.s. the same (when measured in this way).

In particular, for any $\varepsilon > 0$, the set of sequences

$$\Sigma_n = \{a \in \{0,1\}^n \,:\, 2^{-(H(p)+\varepsilon)n} \leq \mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_n) \leq 2^{-(H(p)-\varepsilon)n}\}$$

satisfies that a.s., $x_1 \dots x_n \in \Sigma_n$ for all large enough n; hence $P(\xi_1 \dots \xi_n \in \Sigma_n) \to 1$ as $n \to \infty$. This tells us that most realizations of the first n variables occur with "comparable" probabilities.

We will see shortly that this phenomenon is very general indeed.

7.2 Maker's theorem

Theorem 7.2.1. *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system. Let $f_n \in L^1$ and $f_n \to f$ a.e. Suppose that $\sup_n |f_n| \in L^1$. Then*

$$\frac{1}{N} \sum_{n=0}^{N-1} T^n f_{N-n} \to \mathbb{E}(f|\mathcal{I})$$

a.e. and in L^1, where $\mathcal{I} \subseteq \mathcal{F}$ is the σ-algebra of T-invariant sets. Also,

$$\frac{1}{N} \sum_{n=0}^{N-1} T^n f_n \to \mathbb{E}(f|\mathcal{I})$$

Proof. We prove the first statement, and begin under the assumption that T is ergodic, so $\mathcal{I}$ is trivial.

We first claim that we may assume that $f \equiv 0$. By the ergodic theorem $\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \mathbb{E}(f|\mathcal{I})$ a.e. and in L^1, so in order to prove the theorem it is enough to show that $\frac{1}{N}\sum_{n=0}^{N-1} T^n(f_{N-n} - f) \to 0$ a.e. and in L^1. Since $\sup_n |f_n - f| \in L^1$, we have reduced to the case $f \equiv 0$.

Assume now $f \equiv 0$. Let $\varepsilon > 0$ and let

$$g = \sup_n |f_n|$$

By assumption $g \in L^1$, so we can fix $\delta > 0$ such that for any set E with $\mu(E) < \delta$ we have $\int_E g d\mu < \varepsilon$.

Since $f_n \to 0$ a.e., there is an n_0 and a set A with $\mu(A) > 1 - \delta$ such that $|f_n(x)| < \varepsilon$ for $x \in X$ and all for $n > n_0$.

Now consider $f'_n = 1_A f_n$ and $f''_n = 1_{X \setminus A} f_n$, so $f_n = f'_n + f''_n$. Since $|f'_n| < \varepsilon$ for $n > n_0$ and $|f'_n| \leq g$, we have

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} |T^n f'_{N-n}| &< \frac{1}{N} \sum_{n=0}^{N-n_0-1} \varepsilon + \frac{1}{N} \sum_{n=N-n_0-1}^{N-1} T^n g \\ &< \varepsilon + \frac{1}{N} \left(\sum_{n=0}^{N-1} T^n g - \sum_{n=0}^{N-n_0-1} T^n g \right) \end{aligned} \tag{7.1}$$

The last term on the right tends to 0 a.e. and in L^1 as $N \to \infty$. On the other hand

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} T^n |f''_{N-n}| &\leq \frac{1}{N} \sum_{n=0}^{N-1} T^n |1_{X \setminus A} g| \\ &\to \int_{X \setminus A} g \, d\mu \\ &< \varepsilon \end{aligned} \tag{7.2}$$

a.e. and in L^1, because $\mu(X \setminus A) < \delta$. Combining the two inequalities we conclude that

$$\begin{aligned} \limsup_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n |f_{N-n}| &\leq \limsup_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n |f'_{N-n}| + \limsup_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n |f''_{N-n}| \\ &\leq 2\varepsilon \end{aligned}$$

so $\frac{1}{N} \sum_{n=0}^{N-1} T^n |f_{N-n}| \to 0$ a.e., and similarly, taking the L^1-norm, in L^1.

In the case that $\mathcal{I}$ is non-trivial we proceed in the same manner, but in (7.2) the conclusion becomes

$$\frac{1}{N} \sum_{n=0}^{N-1} T^n |f''_{N-n}| \to \mathbb{E}(1_{X \setminus A} g | \mathcal{I})$$

Now, $\int \mathbb{E}(1_{X \setminus A} g | \mathcal{I}) d\mu = \int 1_{X \setminus A} g d\mu < \varepsilon$, and since $1_{X \setminus A} g \geq 0$ and conditional expectation is a positive operator, $\mathbb{E}(1_{X \setminus A} g | \mathcal{I}) \geq 0$ a.s. Thus by Markov's inequality

$$\mu(\mathbb{E}(1_{X \setminus A} g | \mathcal{I}) \geq \sqrt{\varepsilon}) \leq \frac{\int \mathbb{E}(1_{X \setminus A} g | \mathcal{I}) d\mu}{\sqrt{\varepsilon}} < \sqrt{\varepsilon}$$

We find that

$$\mu\left(x : \limsup_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} T^n|f''_{N-n}|(x) > \sqrt{\varepsilon}\right) < \sqrt{\varepsilon}$$

and so as before, combining the above with (7.1),

$$\mu\left(x : \limsup_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} T^n|f_{N-n}|(x) > \varepsilon + \sqrt{\varepsilon}\right) < \sqrt{\varepsilon}$$

Taking $\varepsilon_k = 2^{-k}$, we have $\sum \sqrt{\varepsilon_k} < \infty$. Applying Borel-Cantelli, we find that a.e. x is in finitely many of the events above and hence a.s. $\frac{1}{N}\sum_{n=0}^{N-1} T^n|f_{N-n}|(x) \to 0$ as desired. □

7.3 The Shannon-McMillan-Breiman theorem

Theorem 7.3.1. *Let $(\xi_n)_{n=0}^{\infty}$ be an ergodic stationary process with values in a finite set I. Let $p(a_1 \dots a_n) = \mathbb{P}(\xi_1 \dots \xi_n = a_1 \dots a_n)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log p(\xi_1 \dots \xi_n)$$

exists a.s. and is a.s. constant.

Equivalently, let $(X, \mathcal{F}, \mu, T)$ be an ergodic measure preserving system and $\mathcal{A}$ a finite partition. Then for a.e. x,

$$\lim_{n\to\infty} \frac{1}{n} \log \mu(\mathcal{A}^n(x)) = h_\mu(T, \mathcal{A})$$

Proof. The two versions are related by passing to the dynamical realization of the random variables, and setting $\mathcal{A} = \{\xi_0^{-1}(i)\}_{i\in I}$; conversely, by defining $\xi_n(x) = \mathcal{A}(T^n x)$. We prove the second version. We give the proof first for the case that T is invertible.

Fix x. Defining $\mathcal{A}^0 = \{X\}$ to be the trivial partition, we have

$$\begin{aligned} \mu(\mathcal{A}^n(x)) &= \mu(\bigcap_{k=1}^{n-1} \mathcal{A}^k(x)) \\ &= \prod_{k=0}^{n-1} \frac{\mu(\mathcal{A}^k(x))}{\mu(\mathcal{A}^{k-1}(x))} \end{aligned}$$

Hence

$$\log \mu(\mathcal{A}^n(x)) = \sum_{k=1}^{n} \log \frac{\mu(\mathcal{A}^k(x))}{\mu(\mathcal{A}^{k-1}(x))}$$

Observe that

$$\left(\bigvee_{i=m+k}^{n+k} T^{-i}\mathcal{A}\right)(x) = \left(T^{-k}\bigvee_{i=m}^{n} T^{-i}\mathcal{A}\right)(x) = \left(\bigvee_{i=m}^{n} T^{-i}\mathcal{A}\right)(T^k x)$$

Therefore, if we define

$$f_k = -\log \frac{\mu((\bigvee_{i=-k}^{0} T^{-i}\mathcal{A})(x))}{\mu((\bigvee_{i=-k}^{-1} T^{-i}\mathcal{A})(x))}$$

Hence

$$-\log \frac{\mu(\mathcal{A}^{k+1}(x))}{\mu(\mathcal{A}^k(x))} = f_k(T^k x)$$

Combining all of the above,

$$-\frac{1}{n}\log \mu(\mathcal{A}^n(x)) = \frac{1}{n}\sum_{k=0}^{n-1} f_k(T^k x)$$

We shall complete the proof by showing that the f_k satisfy the hypothesis of maker's theorem, and then identify the limit. Let

$$f(x) = -\log \mu_x(\mathcal{A}(x)|\bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A})$$

Claim 7.3.2. $f_k \to f$ a.e.

Proof. By the martingale theorem,

$$\begin{aligned}
\mu(\mathcal{A}(x)|\bigvee_{i=-k}^{-1} T^{-i}\mathcal{A}) &= \mathbb{E}(1_{\mathcal{A}(x)}|\bigvee_{i=-k}^{-1} T^{-i}\mathcal{A})(x) \\
&\xrightarrow[n\to\infty]{} \mathbb{E}(1_{\mathcal{A}(x)}|\bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A})(x) \\
&= f(x)
\end{aligned}$$

□

Claim 7.3.3. $\sup_k |f_k| \in L^1$

Proof. Let

$$E_t = \{x : \sup_k f_k(x) > t\}$$

It suffices for us to show that $\mu(E_t) < C \cdot 2^{-t}$ where C is independent of t, since then

$$0 \le \sup_k |f_k| \le \sum_{n=0}^{\infty} 1_{E_n}$$

and the right hand side is integrable.

For each $A \in \mathcal{A}$ consider the family $\mathcal{U}_A$ of sequences $(A_i)_{i=1}^k$ of any length for which $A_i \in T^{-i}\mathcal{A}$, and such that

$$-\log \frac{\mu(A \cap \bigcap_{i=1}^k A_i)}{\mu(\bigcap_{i=1}^k A_i)} > t$$

but $(A_i)_{i=1}^\ell$ does not satisfy this for any $1 \leq \ell < k$. Evidently the sets $\bigcap_{i=1}^k A_i$ are pairwise disjoint as (A_i) ranges over $\mathcal{U}_A$, and every $x \in A \cap E_t$ belongs to such an intersection. Therefore it suffices for us to show that

$$\mu(A \cap \bigcup_{(A_i)\in\mathcal{U}_A} \bigcap A_i) < 2^{-t}$$

since then, summing over $A \in \mathcal{A}$, we have $\mu(E_t) < |\mathcal{A}| \cdot 2^{-t}$. To show the inequality above, observe that for each $(A_i) \in \mathcal{U}_A$ we have

$$\mu(\bigcap_{i=1}^k A_i \cap A) = \mu(\bigcap_{i=1}^k A_i) \cdot \frac{\mu(A \cap \bigcap_{i=1}^k A_i)}{\mu(\bigcap_{i=1}^k A_i)} < 2^{-t} \cdot \mu(\bigcap_{i=1}^k A_i)$$

Therefore, using the fact that the sets $\bigcap_{i=1}^k A_i$ are pairwise disjoint for $(A_i) \in \mathcal{U}_A$,

$$\begin{aligned}
\mu(A \cap \bigcup_{(A_i)\in\mathcal{U}_A} \bigcap A_i) &= \sum_{(A_i)\in\mathcal{U}_A} \mu(A \cap \bigcap A_i) \\
&< 2^{-t} \sum_{(A_i)\in\mathcal{U}_A} \mu(\bigcap A_i) \\
&\leq 2^{-t}\mu(\bigcup_{(A_i)\in\mathcal{U}_A} \bigcap A_i) \\
&\leq 2^{-t}
\end{aligned}$$

as desired. □

□

We can now apply Makers theorem and deduce that $-\frac{1}{n}\log\mu(\mathcal{A}^n(x)) \to \mathbb{E}(f|\mathcal{I})$ a.s. as $n \to \infty$, where $\mathcal{I}$ is the σ-algebra of T-invariant sets. Since our system is ergodic this is simply $\int f d\mu$, and we have already seen that this is the entropy of the system.

Remark 7.3.4. The proof shows that convergence holds also in the non-ergodic case, and the limit is $\mathbb{E}(f|\mathcal{I})$. If $\mu = \int \nu_x d\mu(x)$ is the ergodic decomposition of μ, then $\mathbb{E}(f|\mathcal{I})(x) = \int f\, d\nu_x$. It is also not too hard to show that $\int f d\nu_x = h_{\nu_x}(T, \mathcal{A})$ a.s. Therefore $\frac{1}{n}\log\mu(\mathcal{A}^n(x)) \to h_{\nu_x}(T)$ a.s.

7.4 Entropy-typical sequences

Let $(X, \mathcal{F}, \mu, T)$ be an ergodic measure preserving system and $\mathcal{A} = \{A_1, \ldots, A_r\}$ a finite partition. This gives a map $\xi : X \to \{1, \ldots, r\}$ given by $\xi(x) = i$ if and only if $x \in A_i$. For every n, we obtain a map $\xi_{\mathcal{A},n} : X \to \{1, \ldots, r\}^n$ given by

$$\xi_{\mathcal{A},n}(x) = (\xi(x), \xi(Tx), \ldots, \xi(T^n x))$$

The vector above is called the $(\mathcal{A}, n)$-itinerary of x with respect to $\mathcal{A}$. We also have $\xi_{\mathcal{A},\infty} : X \to \{1, \ldots, r\}^{\mathbb{N}}$ given by

$$\xi_{\mathcal{A},\infty} = (\xi(x), \xi(Tx), \xi(T^2 x), \ldots)$$

This is the full itinerary of x with respect to $\mathcal{A}$.

For every n this gives us a measure $\mu_{\mathcal{A},n} \in \mathcal{P}(\{1, \ldots, r\}^n)$, given by the push-forward of μ by $\xi_{\mathcal{A},n}$:

$$\mu_{\mathcal{A},n}(E) = \mu(x \,:\, \xi_{\mathcal{A},n}(x) \in E)$$

and similarly $\mu_{\mathcal{A},\infty} \in \mathcal{P}(\{1, \ldots, r\}^{\mathbb{N}})$.

Let $h = h_\mu(T, \mathcal{A})$. Given $\varepsilon > 0$ define

$$\Sigma_{n,\varepsilon} = \{a \in \{1, \ldots, r\}^n \,:\, 2^{-n(h+\varepsilon)} \le \mu_{\mathcal{A},n}(a) \le 2^{-n(h-\varepsilon)}\}$$

Proposition 7.4.1. *For every $\varepsilon > 0$, for μ-a.e. x, we have $\xi_{\mathcal{A},n}(x) \in \Sigma_{n,\varepsilon}$ for all large enough n.*

Proof. By the Shannon-McMillan-Breiman theorem, for μ-a.e. x,

$$\frac{1}{n} \log \mu(\mathcal{A}^n(x)) \to h \qquad \text{as } n \to \infty$$

Since $\mu(\mathcal{A}^n(x)) = \mu_{\mathcal{A},n}(\xi_{\mathcal{A},n}(x))$ the claim follows. □

Corollary 7.4.2 (Shannon-McMillan theorem)**.** *For every $\varepsilon > 0$,*

$$\mu(\xi_{\mathcal{A},n}^{-1}(\Sigma_{n,\varepsilon})) \to 1$$

Furthermore, for every set $A \subseteq X$ of positive measure, for all large enough n,

$$\mu(A \cap \xi_{\mathcal{A},n}^{-1}(\Sigma_{n,\varepsilon})) \to \mu(A)$$

Proof. By SMB, $\xi_{\mathcal{A},n}(x) \in (\Sigma_{n,\varepsilon})$ for all large enough n, for a.e. x. Thus a.s., $1_{\xi_{\mathcal{A},n}^{-1}(\Sigma_{n,\varepsilon})} \to 1$ a.e. as $n \to \infty$, so a.s. $1_{A \cap \xi_{\mathcal{A},n}^{-1}(\Sigma_{n,\varepsilon})} \to 1_A$ as $n \to \infty$. Integrating this gives the claim. □

Corollary 7.4.3. *$|\Sigma_{n,\varepsilon}| \le 2^{(h+\varepsilon)n}$, and for every large enough n, $|\Sigma_{n,\varepsilon}| \ge \frac{1}{2} 2^{n(h-\varepsilon)}$.*

Proof. The first inequality follows from

$$1 \geq \mu_{\mathcal{A},n}(\Sigma_{n,\varepsilon}) = \sum_{a \in \Sigma_{n,\varepsilon}} \mu_{\mathcal{A},n}(a) \geq |\Sigma_{n,\varepsilon}| 2^{-n(h+\varepsilon)}$$

The measure $\mu_{\mathcal{A},n}$ gives every $a \in \Sigma_{n,\varepsilon}$ mas at least $2^{-n(h-\varepsilon)}$, and these masses must sum to at most 1, giving the first bound. For the second, note that for large enough n, $\mu_{\mathcal{A},n}(\Sigma_{n,\varepsilon}) > \frac{1}{2}$ (since it $\to 1$), so

$$\frac{1}{2} < \mu_{\mathcal{A},n}(\Sigma_{n,\varepsilon}) = \sum_{a \in \Sigma_{n,\varepsilon}} \mu_{\mathcal{A},n}(a) \leq |\Sigma_{n,\varepsilon}| 2^{-n(h-\varepsilon)}$$

this is the second inequality. □

Combining the last two corollaries gives the so-called *Asymptotic Equipartition Property*, which is a central tool in information theory:

> If (ξ_n) is a stationary ergodic finite-valued process with entropy h, then for large n, the random word $w = \xi_1, \ldots, \xi_n$ is essentially chosen uniformly from a set of 2^{hn} words. More precisely, with probability $1 - o(1)$, the word w is drawn from a set of size $2^{n(h+o(1))}$ and has probability $2^{-n(h+o(1))}$.

We will see applications of this in a later section.

Chapter 8

A combinatorial approach to entropy

Our definition of entropy was based on the entropy of partitions. We now start over and discuss a purely combinatorial definition for ergodic systems.

8.1 Two combinatorial lemmas

Let $\binom{n}{t}$ denote the number of subsets of $\{1,\dots,n\}$ of size $\le t$ (allowing non-integer t).

Lemma 8.1.1. *For every $0 \le \alpha \le \frac{1}{2}$, the number of subsets of $\{1,\dots,n\}$ of size $< \alpha n$ is at most $2^{H(\alpha)n}$.*

Proof. Since $H(\cdot)$ in increasing for $0 \le \alpha < \frac{1}{2}$ there is no loss of generality in assuming $m = \alpha n$ is an integer. By the binomial theorem,

$$\begin{aligned} 1 &= \sum_{k=0}^{n} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \\ &\ge \sum_{k \le m} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \end{aligned}$$

Since $\alpha \le \frac{1}{2}$, the quantity $\delta^k(1-\delta)^k$ decreases as k increases, so for $0 \le k \le m$ it is minimal when $k = m$, giving

$$\begin{aligned} 1 &\ge \sum_{k \le m} \binom{n}{k} \alpha^m (1-\alpha)^{n-m} \\ &= \sum_{k \le m} \binom{n}{k} 2^{n\left(\frac{m}{n}\log\alpha + (1-\frac{m}{n})\log(1-\alpha)\right)} \\ &= \sum_{k \le m} \binom{n}{k} 2^{-nH(\alpha)} \end{aligned}$$

since $m/n = \alpha$. Dividing by $2^{-nH(\alpha)}$ proves the lemma. □

Our next goal gives bounds on the number of words w that can be obtained by concatenating words from given sets of given sizes, while allowing for some extra "errors".

Let A be a finite set and $\alpha > 0$. For each n let $\Sigma_n \subseteq A^n$ be a set with $|\Sigma_n| \leq |A|^{\alpha n}$. For $n = n_1 + \ldots + n_k$ 0bserve that $\Sigma_{n_1}\Sigma_{n_2}\ldots\Sigma_{n_k} \subseteq A^{n_1+\ldots+n_k} = A$ is a collection of words of size

$$|\Sigma_{n_1}\Sigma_{n_2}\ldots\Sigma_{n_k}| = |\Sigma_{n_1}| \cdot |\Sigma_{n_2}| \cdot \ldots \cdot |\Sigma_{n_k}| \leq |A|^{\alpha n_1} \cdot \ldots \cdot |A|^{\alpha n_k} = |A|^{\alpha n}$$

so concatenation of sets Σ_n of this type results in a set with a similar cardinality bound. We want to generalize this to allow more freedom in concatenating elements of Σ_i, both in terms of the lengths and also allowing a small fraction of "errors".

Definition 8.1.2. We say that a word $w \in A^n$ is ε-covered by the sets Σ_i if one can write

$$w = u_1 w_1 u_2 w_2 \ldots u_k w_k u_{k+1} \tag{8.1}$$

where $w_i \in \Sigma_{|w_i|}$ and $|w_i| > 1/\varepsilon$, and the u_i are (possibly empty) words satisfying $\sum |u_i| < \varepsilon n$ (thus $\sum |w_j| \geq (1-\varepsilon)n$).

Note that the assumption $|w_i| > 1/\varepsilon$ implies that the number k of words w_j satisfies $k < \varepsilon n$.

Lemma 8.1.3 (Covering lemma)**.** *Let $2 \leq |A| < \infty$, $\Sigma_n \subseteq A^n$ and $|\Sigma_n| < |A|^{\alpha n}$. Then for each n, the number of words $w \in A^n$ that can be ε-covered by the Σ_i is at most $|A|^{n(\alpha+\varepsilon+H(2\varepsilon))}$.*

Proof. Let $w = u_1 w_1 u_2 w_2 \ldots u_k w_k u_{k+1}$ be a word of length n with $w_i \in \Sigma_{|w_i|}$, $u_i \in A^*$ and $\sum |u_i| < \varepsilon n$. Every such word can be constructed by the following three-step procedure:

1. Choose the positions of the first and last letter of the w_i. This amounts to choosing a set $I = \{a_1 \leq b_1 < a_2 \leq b_2 <\} \subseteq \{1, \ldots, n\}$ of size $< 2\varepsilon n$.

2. For each index in the set $U = \{1, \ldots, n\} \setminus \bigcup [a_i, b_i]$, choose a symbol from A (this specifies the words u_i).

3. For each interval $J_i = [a_i, b_i]$ specify a word from $\Sigma_{|J_i|}$ (these are the w_j).

By the previous lemma, the number of ways to choose I as in (1) is $\leq 2^{nH(2\varepsilon)}$. Since by assumption U in (2) satisfies $|U| < \varepsilon n$, the number of choices of the symbols in (2) is $|A|^{|U|} < |A|^{\varepsilon n}$. Finally, for each J_i as in (3) the number of choices is $\leq |\Sigma|_{|J_i|} \leq |A|^{\alpha |J_i|}$, so, numbering $J_1, \ldots, J_k$ the intervals in (3), the total number of choices is

$$\leq \prod |A|^{\alpha |J_i|} \leq |A|^{\alpha \sum |J_i|} = |A|^{\alpha n}$$

Altogether the number of words w is bounded by

(choices at step (1))·(choices at step (2))·(choices at step (3)) $\leq 2^{nH(2\varepsilon)}|A|^{\varepsilon n}|A|^{\alpha n}$

as claimed. □

The proof gives the following version as well, which is in "base 2":

Lemma 8.1.4. *Let $2 \leq |A| < \infty$, $\Sigma_n \subseteq A^n$ and $|\Sigma_n| < 2^{\alpha n}$. Then for each n, the number of words $w \in A^n$ that can be ε-covered by the Σ_i is at most $2^{n(\alpha+\varepsilon \log |A|+H(2\varepsilon))}$.*

8.2 Alternative definition of entropy

Let $(X, \mathcal{B}, \mu, T)$ be an ergodic process and $\mathcal{A}$ a finite partition. Define $\xi_{\mathcal{A},n}$ and $\mu_{\mathcal{A},n}$ as before. For every $\varepsilon > 0$ let

$$N_n(T, \mathcal{A}, \varepsilon) = \min\{|\Sigma| : \Sigma \subseteq \mathcal{A}^n \text{ and } \mu_{\mathcal{A},n}(\Sigma) > 1 - \varepsilon\}$$

Definition 8.2.1. Let

$$s(T, \mathcal{A}) = \lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log N_n(T, \mathcal{A}, \varepsilon)$$

The outer limit exists because clearly $\varepsilon_1 < \varepsilon_2$ implies $N_n(T, \mathcal{A}, \varepsilon_1) \geq N_n(T, \mathcal{A}, \varepsilon_2)$, hence the inner limit is increasing as a function of ε.

Lemma 8.2.2. *For every $\varepsilon > 0$, for all sufficiently small $\delta > 0$ the following holds: For $k > 1/\varepsilon$ and $\frac{1}{k} \log N_k(T, \mathcal{A}, \delta) < s$, then for all large enough n,*

$$\frac{1}{n} \log N_n(T, \mathcal{A}, \varepsilon) < s + \varepsilon$$

Proof. Let δ be small enough that $2\delta \log |A| + H(4\delta) < \varepsilon$. Assume that $k > 1/\varepsilon$ and $\frac{1}{k} \log N_k(T, \mathcal{A}, \delta) < s$. Then we can choose $\Sigma \subseteq A^k$ such that $|\Sigma| \leq 2^{sk} =$ and $\mu(\Sigma) > 1 - \delta$. Let

$$E = \{x \in X : \xi_{\mathcal{A},k}(x) \in \Sigma\}$$

so $\mu(E) > 1 - \delta$. By the ergodic theorem, for all large enough n, there is a set $X_n \subseteq X$ with $\mu(X_n) > 1 - \delta$ and such that, for $x \in X_n$,

$$\frac{1}{n} \sum_{i=0}^{n} 1_E(T^i x) > 1 - \delta$$

For $x \in X_n$ let

$$I = I(x) = \{0 \leq i \leq n - k : T^i x \in E\}$$

and note that, assuming $n > k/\delta = 1/\varepsilon\delta$,

$$|I| > (1 - \delta)n - k > (1 - 2\delta)n$$

Consider the collection of intervals $[i, i+k-1]$ for $i \in I$. By Lemma 4.3.5 we can choose a sub-collection $I' \subseteq I$ such that the intervals $[i, i+k-1]$, $i \in I'$, are pairwise disjoint, and their total length is at least $|I| > (1-2\delta)n$.

Now, for $i \in I'$ we have $T^i x \in E$, hence

$$\xi_{\mathcal{A},n}(x)|_{[i,i+k-1]} = \xi_{\mathcal{A},k}(T^i x) \in \Sigma$$

It follows that $\xi_{\mathcal{A},n}(x)$ is $(1-2\delta)$-covered by words from Σ. By Lemma **??**, the collection Λ_n of words with this property has size at most $2^{sn+2\delta \log |A| + H(4\delta)} < 2^{(s+\varepsilon)n}$, and $\xi_{\mathcal{A},n}(x) \in \Lambda_n$ for all $x \in X_n$, and $\mu(X_n) > 1-\delta > 1-\varepsilon$. Thus

$$\frac{1}{n} N_n(T, \mathcal{A}, \varepsilon) \leq s + \varepsilon$$

as claimed. □

Corollary 8.2.3. $s(T, \mathcal{A}) = \lim_{(\varepsilon,n)\to(0,\infty)} \frac{1}{n} N_n(T, \mathcal{A}, \varepsilon)$.

8.3 An alternative proof of the Shannon-McMillan-Breiman theorem

Lemma 8.3.1. *Let $(X, \mathcal{B}, \mu, T)$ be a measure-preserving system and $f : X \to \mathbb{R}$ measurable satisfying $f(Tx) \geq f(x)$ a.e. Then f is a.e. invariant ($Tf = f$ a.e.) and if T is ergodic it is a.e. constant. The holds assuming $f(Tx) \leq f(x)$ a.e.*

Proof. First suppose that f is bounded. Suppose that $f(Tx) > f(x)$ on a set of positive measure. Since $f(Tx) \geq f(x)$ everywhere else, $\int f(Tx) d\mu(x) > \int f(x) d\mu(x)$, and this contradicts measure-preservation.

In general, for $M > 0$ define

$$f_M(x) = \begin{cases} -M & f(x) \leq -M \\ f(x) & -M < f(x) < M \\ M & f(x) > M \end{cases}$$

Then F_M is bounded and satisfies the same hypothesis, so by the first case f_M is a.e. invariant for each M, and since $f = \lim_{M\to\infty} f_M$ also f is a.e. invariant. □

Theorem 8.3.2. *For an ergodic measure preserving system $(X < \mathcal{B}, \mu, T)$ and finite partition $\mathcal{A}$, $\lim_{n\to\infty} -\frac{1}{n} \log \mu(\mathcal{A}^n(x))$ exists a.e.*

Proof. Let

$$\begin{aligned} s^+ &= \limsup_{n\to\infty} -\frac{1}{n} \log \mu(\mathcal{A}^n(x)) \\ s^- &= \liminf_{n\to\infty} -\frac{1}{n} \log \mu(\mathcal{A}^n(x)) \end{aligned}$$

so $s^- \leq s^+$. We first claim that both functions are invariant. Indeed, $\mathcal{A}^{n-1}(Tx) \supseteq \mathcal{A}^n(x)$ so

$$\log \mu(\mathcal{A}^{n-1}(Tx)) \geq \log \mu(\mathcal{A}^n(x))$$

Hence

$$\begin{aligned} s^+(x) &= \limsup_{n\to\infty} -\frac{1}{n}\log\mu(\mathcal{A}^n(x)) \\ &\geq \limsup_{n\to\infty} -\frac{1}{n}\log\mu(\mathcal{A}^{n-1}(Tx)) \\ &= \limsup_{n\to\infty} -\frac{1}{n-1}\log\mu(\mathcal{A}^{n-1}(Tx)) \\ &= s^+(Tx) \end{aligned}$$

By the previous lemma, s^+ is a.s. constant. The same works for s^-.

We denote the a.s. values of s^-, s^+ by $\alpha \leq \beta$ respectively.

Suppose for the sake of contradiction that $\alpha < \beta$. Fix $\varepsilon > 0$ and $\alpha < \alpha' < \beta' < \beta$. Write

$$\Sigma_n^- = \{w \in \mathcal{A}^n \,:\, \mu_{\mathcal{A},n}(w) > 2^{-\alpha' n}\}$$

By assumption $s^- = \alpha$ a.e., which implies that for a.e. x there are infinitely many n such that $\xi_{\mathcal{A},n}(x) \in \Sigma_n^-$, and we can choose one such $n = n(x)$ satisfying $n > 1/\varepsilon$.

Choose a set $E \subseteq X$ of measure $\mu(E) > 1-\varepsilon$ and an n_0 such that for $x \in E$ we have $1/\varepsilon < n(x) < n_0$.

By the pointwise ergodic theorem, there is a set $X_0 \subseteq X$ with $\mu(X_0) > 0$, and an N_0, such that for every $x \in X_0$ and $N > N_0$,

$$\frac{1}{N}\sum_{k=0}^{N-1} 1_E(x) > 1-\varepsilon$$

Whenever $x \in X_0$, $N > N_0$, $i < N-k$ and $T^i x \in E$ we have $\xi_{\mathcal{A},N_0}(x)|_{[i,i+n(T^ix)-1]} = \xi_{\mathcal{A},n(T^i(x))}(T^ix) \in \Sigma_{n(T^ix)}$. It follows as is the proof of the previous proposition that for $N > N_0$ the collection of words

$$\Lambda_N = \{\xi_{\mathcal{A},N}(x) \,:\, x \in X\}$$

can be ε-covered by the Σ_j^-, hence satisfies

$$|\Lambda_N| \leq |A|^{N(\alpha'+2\varepsilon+H(2\varepsilon)}$$

and taking ε small we may assume that $\alpha' + 2\varepsilon + H(2\varepsilon) < \beta'$.

Now observe that the set

$$\begin{aligned} \mu\left(x \in X_0 \,:\, \mu(\mathcal{A}^N(x)) < 2^{-N\beta'}) \text{ and } \xi_{\mathcal{A},N}(x) \in \Lambda_n\right) &\leq |\Lambda_n|\cdot 2^{-N\beta'} \\ &< \Lambda_n^{\delta n} \end{aligned}$$

where $\delta = \beta' - \alpha' - 2\varepsilon - H(2\varepsilon) > 0$. Therefore by Borel-Cantelli, a.e. x belongs to only finitely many of the sets above. But by definition of s^+, a.e. $x \in X_0$ satisfies $s^+(x) = \beta > \beta'$, which means that $\mu(\mathcal{A}^N(x)) < 2^{-N\beta'}$ infinitely many N. These last two statements together contradict $\mu(X_0) > 0$, completing the proof. □

Chapter 9

Applications of entropy

9.1 Shannon coding

Let A be a finite set and

$$A^* = \bigcup_{n=0}^{\infty} A^n$$

the set of finite sequences over A. We call such sequences *words* or *blocks*. For $a = a_1 \dots a_n \in A^n$ we write $|a| = n$ for its length.

A *code* is a function $c : A^* \to B^*$, where B is another finite set. Throughout this section we will take $B = \{0, 1\}$.

Given a probability measure μ on A^* and a code c, the *average coding length* of c (with respect to μ) is

$$|c|_\mu = \int |c(a)| \, d\mu(a)$$

This is the average number of bits needed to represent a word a (chosen according to μ) in the coding c.

Given a measure $\mu \in \mathcal{P}(A^*)$ one would like to find a code with minimal average coding length. Of course, we would also like the coding to be *μ-a.s. faithful*, meaning that μ-a.s. $c(a)$ determines a. In other words c should be an injection on $U = \operatorname{supp} \mu$. In fact, we will restrict to codes satisfying a stronger property: that the map $c(w_1, \dots, w_n) = c(w_1) \dots c(w_n)$, $w_i \in A^*$, obtained by applying c to each $w_i \in U$ and concatenating, is injective as a map $U^n \to \{0, 1\}^n$ for all n. If this holds we say that c is *uniquely* decodable.

Lemma 9.1.1. *If c is uniquely decodable then $\sum_{u \in U} 2^{-|c(u)|} \leq 1$.*

Proof. Suppose c is uniquely decodable. Let $U = \operatorname{dom} c$ and assume first that $|c(w)| \leq L$ for all $w \in U$. For each m,

$$(\sum_{u \in U} 2^{-|c(u)|})^m = \sum_{(u_1 \dots u_m) \in U^m} 2^{-\sum_{j=1}^m |c(u_j)|} = \sum_{(u_1 \dots u_m) \in U^m} 2^{-|c(u_1, \dots, u_m)|}$$

Divide the codewords according to length:

$$= \sum_{\ell=1}^{Lm} \sum_{\underline{u} \in U^{\leq m} \,:\, c(\underline{u})=\ell} 2^{-\ell} \leq \sum_{\ell=1}^{Lm} 2^{-\ell} 2^{\ell} = Lm$$

taking m-th roots and $m \to \infty$, this gives $\sum_{u \in U} 2^{-|c(u)|} \leq 1$ as desired. In the general case, apply the bounded-length result to restrictions $c|_{U_M}$ where $U_L = \{w \in U \,:\, |c(w)| \leq L\}$ and take $L \to \infty$. □

Proposition 9.1.2 (Essentially Shannon)**.** *If c is a uniquely decodable then $|c|_\mu \geq H(\mu) = -\sum_{w \in A^*} \mu(w) \log \mu(w)$.*

Proof. Again let $U = \operatorname{dom} c$. It suffices to show that if $\ell : U \to \mathbb{N}$ satisfies $\sum_{w \in L} 2^{-\ell(w)} \leq 1$, then $\sum_{w \in U} \mu(w)\ell(w) \geq H(\mu)$. Consider

$$\begin{aligned} \Delta &= H(p) - \sum_{w \in U} \mu(w)\ell(w) \\ &= -\sum \mu(w) \left(\log \mu(w) + \ell(w)\right) \end{aligned}$$

We want to show that $\Delta \leq 0$. Let $\nu(w) = 2^{-\ell(w)} / \sum_{w \in U} 2^{-\ell(w)}$, so ν is a probability measure on U and $\ell(w) \geq -\log \nu(w)$ (because $\sum 2^{-\ell(w)} \leq 1$). Therefore

$$\begin{aligned} \Delta &\leq -\sum \mu(w) \left(\log \mu(w) - \log \nu(w)\right) \\ &= -\sum \mu(w) \left(\log \frac{\mu(w)}{\nu(w)}\right) \\ &= \sum \mu(w) \left(\log \frac{\nu(w)}{\mu(w)}\right) \\ &\leq \log \sum (\mu(w) \cdot \frac{\nu(w)}{\mu(w)}) \\ &= \log 1 \\ &= 0 \end{aligned}$$

where in the second inequality we used concavity of the logarithm function. □

Corollary 9.1.3. *Let (ξ_n) be a stationary ergodic process with values in a finite set A and h is its entropy. Write μ_n for the distribution of $(\xi_1, \ldots, \xi_n)$. Then for any decodable code $c : A^n \to \{0,1\}^*$, $\frac{1}{n}|c|_{\mu_n} \geq h$.*

Proof. $h = \inf_k \frac{1}{k} H(\xi_1 \ldots \xi_k) \leq \frac{1}{n} H(\xi_1 \ldots \xi_n) \leq |c|_{\mu_n}$ by the last proposition. □

The remarkable fact is the this theoretical lower bound on decodable coding can be achieved:

Proposition 9.1.4. *Let (ξ_n) be a stationary ergodic process with values in a finite set A and h is its entropy. Then for every n there is a code $c_n : A^n \to \{0,1\}^*$ such that $\lim_{n\to\infty} \frac{1}{n}|c|_{\mu_n} = h$.*

Proof. By the Shannon-McMillan theorem, for every $\varepsilon > 0$, for n large enough,

$$\mathbb{P}(\xi_1 \dots \xi_n \in \Sigma_{n,\varepsilon}) > 1 - \varepsilon$$

Therefore we can choose $\varepsilon(n) \to 0$ such that

$$\mathbb{P}(\xi_1 \dots \xi_n \in \Sigma_{n,\varepsilon(n)}) \to 1$$

Enumerate

$$\Sigma_{n,\varepsilon(n)} = \{w_{n,1}, \dots, w_{n,N(n)}\}$$

and recall that $|N(n)| \le 2^{n(h+\varepsilon(n))}$. Therefore for each $w_{n,j}$ the binary representation $[j]$ of j contains at most $\lceil \log N(n) \rceil \le \lceil n(h+\varepsilon) \rceil$ bits. Also, enumerate as $A^n = (u_{n,k})_{k=0}^{|A|^n-1}$.

$$c_n(w) = \begin{cases} 0[j] & \text{if } w = w_{n,j} \in \Sigma_{n,\varepsilon(n)} \\ 1[k] & \text{if } w \notin \Sigma_{n,\varepsilon(n)} \text{ and } w = u_{n,k} \end{cases}$$

Then

$$|c_n(w)| = \begin{cases} 1 + \lceil n(h+\varepsilon) \rceil & w \in \Sigma_{n,\varepsilon(n)} \\ 1 + \lceil n \log |A| \rceil & \text{otherwise} \end{cases}$$

so

$$|c_n|_{\mu_n} = (1 + \lceil n(h+\varepsilon) \rceil)\mu_n(\Sigma_{n,\varepsilon(n)}) + (1 + \lceil n \log |A| \rceil)(1 - \mu_n(\Sigma_{n,\varepsilon(n)}))$$

Dividing by n and using $\mu_n(\Sigma_{n,\varepsilon(n)}) \to 1$,

$$\frac{1}{n}|c_n|_{\mu_n} \to h$$

as claimed. □

9.2 Return times

Let $(\xi_n)_{n=0}^{\infty}$ be an ergodic stationary process with values in a finite alphabet A. Assume without loss of generality that $\xi_n = T^n \xi_0$ for T an ergodic transformation of the probability space $(X, \mathcal{B}, \mu)$. Let $h = h_\mu(T)$.

For $n \in \mathbb{N}$ let r_n denote the first index at which the initial n symbols repeat, that is,

$$r_n = \min\{k \ge 1 : \xi_0, \dots, \xi_{n-1} = \xi_k \dots \xi_{k+n-1}\}$$

so r_n is an integer-valued random variable. Note that by Poincaré recurrence, $r_n < \infty$ a.e.

Theorem 9.2.1 (Wyner-Ziv , Ornstein-Weiss)**.** $\frac{1}{n} \log r_n \to h$ *a.s.*

We prove this in two stages, first the upper bound, then the lower bound. For $w \in A^n$ write

$$[w] = \{\xi_0 \dots \xi_{n-1} = w\}$$

and for $\Delta < A^n$ write

$$[\Delta] = \bigcup_{w \in \Delta} [w] = \{\xi_0 \dots \xi_{n-1} \in \Delta_n\}$$

Proof that $\limsup \frac{1}{n} r_n \leq h$ *a.s.* Let $\varepsilon > 0$, it suffices to show that a.s. $r_n \leq 2^{(h+\varepsilon)n}$ for all but finitely many n, i.e. that a.e. $x \in X$ belongs to

$$E_n \quad = \quad \{r_n > 2^{(h+\varepsilon)n}\}$$

finitely often.

Let $\Sigma_{n,\varepsilon/2} \subseteq A^n$ denote the sets of "good" words defined in the Section **??**, so $|\Sigma_{n,\varepsilon/3}| \leq 2^{(h+\varepsilon/2)n}$ and almost surely $\xi_1 \dots \xi_n \in \Sigma_{n,\varepsilon/23}$ for all large enough n. In other words, a.e. $x \in X$ belongs to

$$F_n = \{\xi_{01} \dots \xi_{n-1} \in \Sigma_{n,\varepsilon/2}\}$$

for all but finitely many F_ns.

Thus, it is enough for us to show that a.e. every x belongs to $F_n \cap E_n$ only finitely many times. This will follow from Borel-Cantelly once we show that $\sum \mu(E_n \cap F_n) < \infty$.

Suppose that $w \in \Sigma_{n,\varepsilon/2}$. If $x \in [w] \cap E_n$ then $r_n(x) > 2^{(h+\varepsilon)n}$, so $T^i x \notin [w]$ for $1 \leq i \leq 2^{(h+\varepsilon)n}$. Thus the sets

$$T^{-i}([w] \cap E_n]) \qquad 1 \leq i \leq 2^{(h+\varepsilon)n}$$

are disjoint and of eq1ual measure, so

$$\mu([w] \cap E_n) < 2^{-(h+\varepsilon)n}$$

Since $F_n = \bigcup_{w \in \Sigma_{n,\varepsilon/2}} [w]$,

$$\mu(F_n \cap E_n) = \sum_{w \in \Sigma_{n,\varepsilon/2}} \mu([w] \cap W_n) < \sum_{w \in \Sigma_{n,\varepsilon/2}} 2^{-(h+\varepsilon)n} \leq 2^{(h+\varepsilon/2)n} \cdot 2^{-(h+\varepsilon)n} = 2^{-\frac{1}{2}\varepsilon n}$$

and the claim follows. □

Lemma 9.2.2. *Let $\varepsilon > 0$ and let $\Delta_n \subseteq A^n$ be sets such that $|\Delta_n| < 2^{(h-\varepsilon)n}$. Then a.s. $\xi_0 \dots \xi_{n-1} \in \Delta_n$ only x finitely often (i.e. $\mu(\limsup[\Delta_n]) = 0$).*

Proof. Let $\Sigma_{n,\varepsilon/2}$ be as usual. Then a.e. x is in $[\Sigma_{n,\varepsilon/2}]$ for all but finitely many n, so it suffices to show that a.s. $\xi_0 \dots \xi_{n-1} \in \Sigma_{n,\varepsilon/2} \Delta \Delta_n$ only finitely often. This follows from Borel Cantelli, since every $w \in \Sigma_{n,\varepsilon/2}$ satisfies $\mu([w]) < 2^{-(h-\varepsilon/2)n}$ and so

$$\mu([\Delta_n] \cap [\Sigma_{n,\varepsilon/2}]) \leq \sum_{w \in \Delta_n} \mu([w]) < \sum_{w \in \Delta_n} 2^{-(h-\varepsilon/2)n} \leq |\Delta_n| 2^{-(h-\varepsilon/2)n} < 2^{-\varepsilon n/2}$$

so $\sum \mu([\Delta_n] \cap [\Sigma_{n,\varepsilon/2}]) < \infty$. □

Lemma 9.2.3. *Let $t, \rho > 0$ and $N \in \mathbb{N}$. Let $W \subseteq A^N$ denote the set of words $w \in A^N$ can be written as $u_0 w_1 u_1 w_2 u_2 \dots w_k u_k$, where*

1. $1/\rho < |w_i|$.

2. $\sum |w_i| > (1-\rho)Nn$
3. *If w_i begins at index $m(i)$ then it repeats in w at an index $m(i) < m'(i) < m(i) + 2^{t|w_i|}$.*

Then $|W| \le |A|^{(\rho + H(2\rho) + t(1-\rho))n}$.

Proof. We claim that a word $w = u_1 w_1 \dots w_{k-1} w_k$ as above is specified completely by the following information:

a the set of intervals $[m(i), m(i) + |w_i| - 1]$ that describe the positions of the w_i.

1. The words u_k.
2. The distances $m'(i) - m(i)$ at which w_i repeats.

Indeed, given this we can reconstruct w as follows: the positions of the w_i, u_i and the symbols of the u_i are given explicitly. Now reconstruct the symbols of w_i from right to left: if we have constructed the symbols at positions $j+1 \dots N$, and j is in w_i, then the symbol at j is the same as the one at $j + m'(i) - m(i)$, which is already known.

Finally we count how many choices we have: Since $|w_i| > 1/\rho$ we have $k < \rho N$, so the set of indices $m(i), m(i) + |w_i| - 1$ has size $\le 2\rho N$. Thus there are $\le 2^{H(2\rho)}$ choices for these indices. Since $\sum |u_i| < \rho N$, the number of choices of symbols is $< 2^{\rho N}$. And finally, $m'(i) - m(i) < 2^{t|w_i|}$, so the number of choices of these distances is $\le \prod 2^{t|w_i|} = 2^{t \sum |w_i|} = 2^{t(1-\rho)N}$. Thus the total number of choices satisfies the stated bound. □

Proof that $\limsup \frac{1}{n} r_n \ge h$ *a.s.* We may assume $h > 0$. Let $0 < \varepsilon < h$, we must show that a.s. $r_n \ge 2^{(h-\varepsilon)n}$ for all but finitely many n. In other words, writing

$$r_- = \liminf_{n \to \infty} \frac{1}{n} \log r_n$$

we must show that $r_- \ge h - \varepsilon$ a.s.

Note that if $r_n(x) = k$ then $\xi_0 \dots \xi_{n-1} = \xi_k \dots \xi_{k+n-1}$, hence $\xi_1 \dots \xi_n = \xi_{k+1} \dots \xi_{k+1+(n-1)-1}$, so

$$r_{n-1}(Tx) \le r_n(x)$$

It follows that $r_-(Tx) \le r_-(x)$, so by ergodicity r_- is a.s. constant.

Thus in order to show $r_- \ge h - \varepsilon$ a.s. it suffices to show that $r_- < h - \varepsilon$ a.s. is impossible.

Thus, suppose that $r_- < h - \varepsilon$ a.s. Fix another parameter $\delta > 0$. Thus a.e. x satisfies $r_n < 2^{(h-\varepsilon)n}$ for arbitrarily large n and in particular for $n > 1/\delta$ so there is an L such that the set

$$E = \left\{ x : r_n(x) < 2^{(h-\varepsilon)n} \text{ for some } \frac{1}{\varepsilon} < n < L \right\}$$

satisfies

$$\mu(E) > 1 - \delta$$

For large N, let

$$F_N = \left\{ x \,:\, \frac{1}{N} \sum_{n=0}^{N-1} 1_E(T^n x) > 1 - \delta \right\}$$

and let

$$\Delta_N = (\xi_0(x) \dots \xi_{N-1}(x) \,:\, x \in F_N\}$$

By the ergodic theorem a.e. x belongs to F_N for all but finitely many N. We will arrive at a contradiction by showing that $|\Delta_n| < 2^{(h-\varepsilon/2)n}$.

Indeed, for $x \in F_N$ and $w = \xi_0(x) \dots \xi_{N-1}(x)$ let

$$I = \{0 \le i \le N - L - 2^{(h-\varepsilon)L} \,:\, T^i x \in E\}$$

and note that assuming N is large enough,

$$\frac{1}{N}|I| \ge 1 - \delta - \frac{L + 2^{(h-\varepsilon)L}}{N} < 1 - 2\delta$$

For each $i \in I$ there is a $1 \le n = n(i) \le L$ such that $r_{n(i)}(T^i x) < 2^{(h-\varepsilon)n}$. Let $j(i) = i + n(i) - 1$. This means that the subword of length $n(i)$ starting at i in w repeats in w at an index $i + 1 \le i' < i + 2^{(h-\varepsilon)n}$. Apply the covering Lemma 4.3.5 to the collection $\{[i, i + n(i) = 1]\}_{i \in I}$ and obtain a sub-collection $I_0 \subseteq I$ such that $\{[i, i + n(i) = 1]\}_{i \in I_0}$ are pairwise disjoint and their union is of size at least $|I| > (1 - 2\delta)N$. Now w satisfies the hypothesis of Lemma **??** with $t = h - \varepsilon$ and $\rho = 2\delta$, so

$$|\Delta_n| \le 2^{(2\delta + H(4\delta) + (h-\varepsilon)(1-2\delta))N} < 2^{(h-\varepsilon/2)n}$$

assuming δ is small enough. □

9.3 The Lempel-Ziv algorithm

In this section we sketch without proofs the Lempel-Ziv compression algorithm, which as input accepts a sequence $x = x_1, x_2, \dots, x_n$ from a given finite alphabet A, and outputs a sequence $c(x) = y_1, y_2, \dots, y_n$ of bits $0, 1$ such that the map $x \to c(x)$ is 1-1 and constitutes a faithful code (when restricted to A^n). The important property of c is that it compresses optimally universally: for any ergodic process (ξ_n) of entropy h, we have $\mathbb{E}(\frac{1}{n}|c_n(\xi_1 \dots \xi_n)|) \to h$; and in fact a.s. $\frac{1}{n}|c_n(\xi_1 \dots \xi_n)| \to h$.

The algorithm is as follows: on input $x_1, \dots, x_n$,

1. Let $i = 1$.

2. Let $j \in (i, n]$ be the largest index such that the word $x_i \dots x_{j-1}$ appears at an index $i' < i$.

3. Output $j-1-i$, $i-i'$.

4. Output x_{j+1}.

5. If $j=n$ halt, otherwise set $i=j+1$ and go to 2.

From the output one easily reconstructs x inductively: if we have reconstructed $x_1,\dots,x_k$ and $i=k+1$ and we read $(m_1,m_2,a)\in\times\mathbb{N}\times\mathbb{AN}$ from the input, then we know that the next word has length m_1, we can find the first m_1-1 symbols by going back m_2 steps in the word we have already constructed, and the last symbol is a.

The main idea is that the distance m_2 is of order $2^{h(m_1-1)}$. This is similar to the return times theorem. Thus m_2 can be coded using $h(m_1-1)$ bits. In addition we must record the length m_1, which requites $\log m_1 = o(m_1)$ bits, and the symbol a which requires $O(1)=o(m_1)$ bits if $m_1\to\infty$. Note that the sequence of i's chosen by the algorithm parses x into distinct blocks, so for each L, the difference $j-i$ can be less than L only a bounded number of times (independent of n).

Chapter 10

The Pinsker algebra and CPE-systems

10.1 Factors and relative entropy

Definition 10.1.1. Factor

Remark 10.1.2. Identification of factors with sub-σ-algebras

Example 10.1.3. Trivial factors, factor generated by a partition/family of sets, product systems and marginal projections.

Definition 10.1.4. Entropy of a partition and system relative to a factor.

Proposition 10.1.5. $h_\mu(T, \mathcal{A}|\mathcal{B}) = H(\mathcal{A}|\bigvee_{i=1}^{\infty} T^i\mathcal{A} \vee \mathcal{B})$.

Remark 10.1.6. The usual definition is relative to the trivial factor.

Recall that

$$h_\mu(T, \mathcal{A} \vee \mathcal{B}) = H(\mathcal{A}|\bigvee_{i=1}^{\infty} T^i\mathcal{A} \vee \bigvee_{i=-\infty}^{\infty} T^i\mathcal{B}) + h_\mu(T, \mathcal{B})$$

Proposition 10.1.7. $h_\mu(T|\mathcal{E}) = h_\mu(T) - h_\mu(T|_\mathcal{E})$, *assuming that* $h_\mu(T|\mathcal{E}) < \infty$.

Proof. For any partitions $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{B} \subseteq \mathcal{E}$ we have

$$\begin{aligned} h_\mu(T) &\geq h_\mu(\mathcal{A} \vee \mathcal{B}) \\ &= H(\mathcal{A}|\bigvee_{i=1}^{\infty} T^i\mathcal{A} \vee \bigvee_{i=-\infty}^{\infty} T^i\mathcal{B}) + h_\mu(T, \mathcal{B}) \\ &\geq H(\mathcal{A}|\bigvee_{i=1}^{\infty} T^i\mathcal{A} \vee \mathcal{E}) + h_\mu(T, \mathcal{B}) \\ &= h_\mu(T, \mathcal{A}|\mathcal{E}) + h_\mu(T, \mathcal{B}) \end{aligned}$$

This shows that $h_\mu(T) \geq h_\mu(T|\mathcal{E}) + h_\mu(T|_\mathcal{E})$. On the other hand, by choosing $\mathcal{A}$ fine enough we can ensure that $|h_\mu(T) - h_\mu(\mathcal{A} \vee \mathcal{B})|$ is arbitrarily small and likewise $|h_\mu(T, \mathcal{A}|\mathcal{E}) - h_\mu(T|\mathcal{E})|$. Also choosing $\mathcal{B}$ fine enough we can ensure that $|h_\mu(T, \mathcal{B}) - h_\mu(T|_\mathcal{E})|$ is arbitrarily small. Finally, we can choose $\mathcal{B}$ so that $|H(\mathcal{A}|\bigvee_{i=1}^\infty T^i\mathcal{A} \vee \mathcal{E}) - H(\mathcal{A}|\bigvee_{i=1}^\infty T^i\mathcal{A} \vee \bigvee_{i=-\infty}^\infty T^i\mathcal{B})|$ is arbitrarily small. This controls all the inequalities above and allows us to reverse them with an arbitrarily small error. This proves the claim. □

Corollary 10.1.8. *If $\mathcal{A}$ generates the system and $\mathcal{B}$ generates a factor then the relative entropy is $H(\mathcal{A}|\bigvee_{i=1}^\infty T^i\mathcal{A} \vee \bigvee_{i=-\infty}^\infty T^i\mathcal{B})$.*

Example 10.1.9. Continuity of entropy in the space of k-partitions.

10.2 The Pinsker algebra

Let $\mathcal{P}_A = \{A, X \setminus A\}$

Definition 10.2.1. The Pinsker algebra is $\Pi = \{A \in \mathcal{F} : h_\mu(T, \mathcal{P}_A) = 0\}$.

Clearly Π is T-invariant.

Proposition 10.2.2. Π *is countably generated (mod μ) .*

Proof. Immediate from separability of the space of 2-partitions in L^1 and continuity of entropy. □

Proposition 10.2.3. Π *is a σ-algebra.*

Proof. Let $A \in \sigma(\Pi)$. Thus there are $A_n \in \Pi$ with $A \in \sigma(A_1, A_2, \ldots)$ up to measure 0. Letting $\mathcal{B}_n = \bigvee_{i=1}^n \mathcal{P}_{A_i}$, we have

$$h_\mu(T, \mathcal{B}_n) \leq \sum_{i=1}^n h_\mu(T, \mathcal{P}_{A_i}) = 0$$

so $h_\mu(T, \mathcal{B}_n) = 0$. Also

$$h_\mu(T, \mathcal{A}|\mathcal{B}_n) \to h_\mu(T, \mathcal{A}|\bigvee_{i=1}^\infty \mathcal{P}_{A_i}) = h_\mu(T, \mathcal{A}|\Pi) = 0$$

Hence

$$0 \leq h_\mu(T, \mathcal{P}_A) \leq h_\mu(T, \mathcal{P}_A \vee \mathcal{B}_n) = h_\mu(T, \mathcal{B}_n) + h_\mu(T, \mathcal{A}|\mathcal{B}_n) \to 0$$

so $A \in \Pi$. □

10.3 The tail algebra and Pinsker's theorem

Definition 10.3.1. $\mathcal{T}^{\pm}$of a process.

Definition 10.3.2. Let $\mathcal{A}$ be a partition. Then $\mathcal{T}^{-}(\mathcal{A}) = \bigcap_{n\in\mathbb{N}} \bigvee_{i=-\infty}^{-n} T^{-i}\mathcal{A}$ and $\mathcal{T}^{+}(\mathcal{A}) = \bigcap_{n\in\mathbb{N}} \bigvee_{i=n}^{\infty} T^{-i}\mathcal{A}$.

Theorem 10.3.3. *If $\mathcal{A}$ generates then $\Pi = \mathcal{T}^{\pm}(\mathcal{A})$.*

Proof. Let $\mathcal{B} \in \mathcal{T}^{-}(\mathcal{A})$. Since $\mathcal{A}$ generates we have

$$h_\mu(T) = h_\mu(T, \mathcal{A}) \leq h_\mu(T, \mathcal{A} \vee \mathcal{B}) \leq h_\mu(T)$$

and so

$$\begin{aligned} h_\mu(T) &= h_\mu(T, \mathcal{A} \vee \mathcal{B}) \\ &= h_\mu(T, \mathcal{B}) + h_\mu(T, \mathcal{A}|\mathcal{B}) \\ &= h_\mu(T, \mathcal{B}) + H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A} \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{B}) \\ &= h_\mu(T, \mathcal{B}) + H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) \\ &= h_\mu(T, \mathcal{B}) + h_\mu(\mathcal{A}) \end{aligned}$$

where in the last transition we used that $\mathcal{T}^{-}(\mathcal{A}) \subseteq \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}$ and $T^j\mathcal{B} \in \mathcal{T}^{-}(\mathcal{A})$ for all j, hence $\bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{B} \subseteq \mathcal{T}^{-}(\mathcal{A})$. Subtracting $h_\mu(T, \mathcal{A})$ from both sides gives $h_\mu(T, \mathcal{B}) = 0$.

Now suppose that $\mathcal{B} \in \Pi$. Then we again have, for every k,

$$\begin{aligned} H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) &= h_\mu(T) \\ &= h_\mu(T, \mathcal{B}) + H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A} \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{B}) \\ &= H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A} \vee \bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{B}) \\ &\leq H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A} \vee T^{-k}\mathcal{B}) \\ &\leq H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) \end{aligned}$$

so we have for all k,

$$H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) = H_\mu(\mathcal{A}| \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A} \vee T^{-k}\mathcal{B})$$

An elementary calculation using the conditional entropy formula shows that this implies for all k that

$$H_\mu(T^{-k}\mathcal{B}|\bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) = H_\mu(T^{-k}\mathcal{B}|\bigvee_{i=-\infty}^{0} T^{-i}\mathcal{A})$$

or equivalently, for all k,

$$H_\mu(\mathcal{B}|\bigvee_{i=-\infty}^{k} T^{-i}\mathcal{A}) = H_\mu(\mathcal{B}|\bigvee_{i=-\infty}^{k+1} T^{-i}\mathcal{A})$$

Now, since $\mathcal{A}$ generates, we know that

$$\lim_{n\to\infty} H(\mathcal{B}|\bigvee_{i=-\infty}^{n} T^{-i}\mathcal{A}) = H(\mathcal{B}|\bigvee_{i=-\infty}^{\infty} T^{-i}\mathcal{A}) = 0$$

but since

$$H(\mathcal{B}|\bigvee_{i=-\infty}^{n} T^{-i}\mathcal{A}) = H(\mathcal{B}|\bigvee_{i=-\infty}^{n-1} T^{-i}\mathcal{A}) = \ldots = H(\mathcal{B}|\bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A})$$

we find that

$$\lim_{n\to\infty} H(\mathcal{B}|\bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}) = 0$$

so $\mathcal{B} \in \bigvee_{i=-\infty}^{-1} T^{-i}\mathcal{A}$. The same argument shows that $\mathcal{B} \in \bigvee_{i=-\infty}^{-k} T^{-i}\mathcal{A}$ for all k, so $\mathcal{B} \in \mathcal{T}^-(\mathcal{A})$. □

Corollary 10.3.4. $\mathcal{T}^+ = \mathcal{T}^-$.

10.4 Systems with completely positive entropy

Definition 10.4.1. CPE (K) systems

Definition 10.4.2. A system has uniform mixing if for every partition $\mathcal{P}$, $h_\mu(T^n, \mathcal{P}) \to H_\mu(\mathcal{P})$ as $n \to \infty$. In other words,

$$\sup_N \left(\frac{1}{N} H_\mu(\bigvee_{i=1}^{N} T^{-nN}\mathcal{P}) - H_\mu(\mathcal{P}) \right) = o(1) \qquad \text{as } n \to \infty$$

Theorem 10.4.3. *A system is CPE if and only if it has uniform mixing.*

Proof. If $h_\mu(T, Q) = 0$ then $h_\mu(T^n, Q) = 0$ for all n so there is no uniform mixing.

In the other direction if the system is CPE, then $\mathcal{T}^-(\mathcal{P}) \subseteq \Pi$ is trivial, so from the martingale theorem,

$$H_\mu(\mathcal{P}| \bigvee_{i=-\infty}^{-n} T^{-i}\mathcal{P}) \to H_\mu(P) \qquad \text{as } n \to \infty$$

since $\bigvee_{i=-\infty}^{-1} T^{-ni}\mathcal{P} \subseteq \bigvee_{i=-\infty}^{-n} T^{-i}\mathcal{P}$ we have

$$H_\mu(\mathcal{P}| \bigvee_{i=-\infty}^{-n} T^{-i}\mathcal{P}) \leq H_\mu(\mathcal{P}| \bigvee_{i=-\infty}^{-1} T^{-ni}\mathcal{P}) \leq H_\mu(P)$$

hence

$$h_\mu(T^n, \mathcal{P}) = H(\mathcal{P}| \bigvee_{i=-\infty}^{-1} T^{-ni}\mathcal{P}) \to H_\mu(P) \qquad \text{as } n \to \infty \qquad \square$$

Proposition 10.4.4. *If T is uniformly mixing (equivalently CPE) then for any partition $\mathcal{P}$ and any k,*

$$H(\bigvee_{i=0}^{k-1} T^{-in}\mathcal{P}) \to kH(\mathcal{P})$$

In particular, for any functions $f_0, \ldots, f_{k-1} \in L^\infty(\mu)$,

$$\int f_0(x) \cdot f_1(T^n x) \cdot f_3(T^{-2n}x) \cdot \ldots \cdot f_{k-1}(T^{(k-1)n}x) d\mu(x) \to \prod \int f_i d\mu$$

and in particular T is strongly mixing.

Proof sketch. First one shows that it is enough to prove this for simple functions, hence for indicator functions. Let $f_i = 1_{A_i}$ and let $\mathcal{P}$ the partition determines by $A_1, \ldots, A_{k-1}$. Now,

$$\begin{aligned} \int f_0(x) \cdot \ldots \cdot f_{k-1}(T^{(k-1)n}x) d\mu(x) &= \mathbb{E}_\mu(f_0|T^n f_1, \ldots, T^{(k-1)n} f_{k-1}) \mathbb{E}_\mu(T^n f_1 \cdot \ldots \cdot T^{(k-1)n} f_{k-1}) \\ &= \mathbb{E}_\mu(f_0|T^n f_1, \ldots, T^{(k-1)n} f_{k-1}) \mathbb{E}_\mu(f_1 \cdot T^n f_2 \cdot \ldots \cdot T^{(k-2)n} f_{k-1}) \end{aligned}$$

For n large we can make $H_\mu(\mathcal{P}| \bigvee_{i=1}^{k-1} T^{-in}\mathcal{P})$ as close to $H_\mu(\mathcal{P})$ as we like. This implies that $\mathbb{E}_\mu(f_0|T^n f_1, \ldots, T^{(k-1)n} f_{k-1})$ will be arbitrarily close to $\int f_0 d\mu$ when n is large. This takes care of the first term on the right-hand side, for the second we induct. $\square$

Chapter 11

Topological dynamics

11.1 Topological dynamical systems

Definition 11.1.1. A topological dynamical system is a pair (X, T) where X is a compact metric space and $T : X \to X$ a continuous map.

One can drop the assumption of metrizability with few changes to the resulting theory. The assumption of compactness is more essential. Some authors also assume that T is onto, and/or invertible, but we do not make these assumptions.

We define the orbit (forward and two-sided) of a point in the same manner as in the measurable case.

Definition 11.1.2. Two topological systems (X, T) and (Y, S) are isomorphic if there is a homeomorphism $\pi : X \to Y$ such that $\pi T = S\pi$.

Note that homeomorphism of the phase space is a pre-requisite for isomorphism. such an obstruction rarely exists in the measurable category since most "natural" measure spaces are isomorphic as measure spaces.

Definition 11.1.3. A topological system (Y, S) is a factor of a topological system (X, T) if there is an onto continuous map $\pi : X \to Y$ such that $\pi T = S\pi$.

Again, there may be topological obstructions to the existence of a factor map.

Example 11.1.4. Full shift: for a finite alphabet A let $X = A^{\mathbb{N}}$ or $X = A^{\mathbb{Z}}$ with the product topology (A is discrete). The shift map is then continuous.

Example 11.1.5. Circle rotation

Definition 11.1.6. A subsystem of a topological system (X, T) is a subset $Y \subseteq X$ which is closed, non-empty, and invariant ($TY \subseteq Y$).

Example 11.1.7. The orbit closure of a point is a subsystem.

Symbolic example: $X \subseteq \{0, 1\}^{\mathbb{Z}}$, $x \in X$ if and only if no two consecutive 1s.

11.2 Transitivity

Definition 11.2.1. $x \in X$ is (forward) transitive if the forward orbit $\{T^n x\}_{n\geq 0}$ is dense in X. It is a bi-transitive point if the two-sided orbit $\{T^n x\}_{n\in\mathbb{Z}}$ is dense. A system is transitive if it contains a transitive point.

Example 11.2.2. In $\{0,1\}^{\mathbb{Z}}$ take the point $0000\ldots0abcd\ldots$ where $a, b, c, d, \ldots$ is an enumeration of $\{0,1\}^*$.

Proposition 11.2.3. *If (X,T) supports an invariant ergodic Borel probability measure μ and $\mu(U) > 0$ for every open set $U \neq \emptyset$ then μ-a.e point is transitive.*

Proof. Fix a countable basis $\{U_i\}$ for the topology on X. By the ergodic theorem, for each i, a.e. x satisfies $\frac{1}{N}\sum_{n=0}^{N-1} 1_{U_i}(T^n x) \to \int 1_{U_i} d\mu = \mu(U_i) > 0$. In particular there is an n such that $1_{U_i}(T^n x) = 1$, that is, $T^n x \in U_i$. Since there are countably many sets U_i, a.e. x satisfies this for all i simultaneously. Such an x has a dense forward orbit in X. □

Proposition 11.2.4. *Suppose T is invertible. Then the following are equivalent:*

1. *(X,T) is bi-transitive.*
2. *For every pair of open sets $U, V \neq \emptyset$ there is an $n \in \mathbb{Z}$ with $T^{-n}U \cap V \neq \emptyset$.*
3. *The set of bi-transitive points in X is a dense G_δ subset of X.*

Proof. (1) implies (2): given U, V, let x be a transitive point. Then there is an n such that $T^n x \in U$ and an m such that $T^m x \in V$. Thus $x \in T^{-n}U \cap T^{-m}V$, so $T^{-(n-m)}U \cap V \neq \emptyset$.

(2) implies (3): let U_i be a basis for the topology if X. By (2), for each i the set $\bigcup_{n\in\mathbb{Z}} T^{-n}U_i$ is dense in X and of course it is open. Thus $X_0 = \bigcap_i \bigcup_{n\in\mathbb{Z}} T^{-n}U_i$ is a dense G_δ set. If $x \in X_0$ then for each i there is an n such that $x \in T^{-n}U_i$. This implies that $\{T^n x\}$ intersects every open set, so x is bi-transitive.

(3) implies (1) trivially. □

For non-invertible systems we have the following:

Proposition 11.2.5. *Assume X has no isolated points. Then the following are equivalent:*

1. *(X,T) is transitive.*
2. *For every pair of open sets $U, V \neq \emptyset$ there is an $n \in \mathbb{N}$ with $T^{-n}U \cap V \neq \emptyset$.*
3. *The set of transitive points in X is a dense G_δ subset of X.*

Proof. (2) implies (3) is proved as in the previous proposition, using the sets $\bigcup_{n\in\mathbb{N}} T^{-n}U_i$ instead of $\bigcup_{n\in\mathbb{Z}} T^{-n}U_i$. Again, (3) implies (1) trivially, so we only need to prove that (1) implies (2). For this let $\emptyset \neq U, V$ be open and suppose $T^n x \in U$ and $T^m x \in V$. As before we have $T^{-(n-m)}U \cap V \neq \emptyset$ but we would like $n - m \in \mathbb{N}$ and this might fail. To correct this, note first that U contains infinitely many points (since it is non-empty and X has no isolated points). Therefore there is a $y \in U$ with $y \neq T^i x$ for all $0 \leq i \leq m$. Let $U' \subseteq U$ be an open neighborhood of y' disjoint from $x, Tx, \ldots, T^m x$. Then there is an $n \in \mathbb{N}$ with $T^n x \in U' \subseteq U$. Clearly $T^n x \neq x, Tx, \ldots, T^m x$ so $n > m$. Now proceed as before. □

Remark 11.2.6. In a transitive system satisfying one of the assumptions above the cases (T invertible or X without isolated points), for every open sets $U, V \neq \emptyset$, the set $\{n \in \mathbb{N} : T^{-n}U \cap V \neq \emptyset\}$ is infinite.

Example 11.2.7. Consider the orbit closure X of $x = (0, 1, 1, 1, 1, \ldots) \in \{0, 1\}^{\mathbb{N}}$. $X = \{x, y\}$ where $y = (1, 1, 1 \ldots)$. Then x is transitive but y is not, and $\{y\}$ is open in X.

Remark 11.2.8. The propositions above can be viewed as a topological version of the ergodic theorem: in a transitive system a "typical" point (in the Baire category sense) is transitive. Compare: In an ergodic system a typical point visits every set in a dense countable algebra of sets).

However, there is no analog of the ergodic decomposition theorem: a non-transitive system does not decompose into disjoint transitive systems. For example, the orbit closure X of $(\ldots 000111 \ldots)$ and Y of $(\ldots 000222 \ldots)$ in $\{0, 1, 2\}^{\mathbb{Z}}$ both contain 0, so their union $Z = X \cup Y$ cannot be partitioned into two transitive (closed)) systems.

Definition 11.2.9. (X, T) is topologically mixing if for every pair of open sets $U, V \neq \emptyset$, $\{n : T^{-n}U \cap V \neq \emptyset\}$ is co-finite in $\mathbb{N}$.

Lemma 11.2.10. *(obvious) Topological mixing implies transitivity.*

Example 11.2.11. $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ are mixing, hence transitive, for every compact A.

Indeed, it is enough to check that the condition is satisfied for U, V in a basis of the topology. The basis consisting of cylinder sets clearly satisfies it.

11.3 Minimality

Definition 11.3.1. (X, T) is minimal if it has no non-trivial subsystems, i.e. if $Y \subseteq X$ is a subsystem then $Y = X$ or $Y = \emptyset$.

Remark 11.3.2. If (X, T) us minimal then T is surjective, since $T(X)$ is a subsystem.

Lemma 11.3.3. *A system is minimal if and only if every point is transitive.*

Proof. If x is a non-transitive point, then its orbit closure is a non-trivial subsystem. Conversely, if $Y \subseteq X$ is a non-trivial subsystem and $x \in Y$ then the orbit of x is contained in Y and hence not dense. □

Example 11.3.4. Circle rotation.

Proposition 11.3.5. *Every topological dynamical system has a minimal subsystem.*

Proof. Let $\mathcal{S}$ be the class of subsystems of X ordered by inclusion. Since the intersection of a decreasing family of subsystems is a subsystem (non-empty because X is compact), by Zorn's lemma there is a minimal element for this order. By definition this is a subsystem which does not have proper subsystems, so it is minimal. □

We next develop a more usable characterization of minimality in terms of the visit times of points to sets.

Definition 11.3.6. A set $I \subseteq \mathbb{N}$ (or $I \subseteq \mathbb{Z}$) is syndetic if there is a constant L (the syndeticity constant) such that every interval $[a, a+L] \subseteq \mathbb{N}$ (resp. $\mathbb{Z}$) contains a point from I. In other words, the gaps in I are of length at most L.

Proposition 11.3.7. *The following are equivalent:*

1. *(X, T) is minimal.*
2. *For every open set $U \neq \emptyset$ there is an N such that $X = \bigcup_{n=0}^{N-1} T^{-n}$.*
3. *For every $x \in X$ and open set $U \neq \emptyset$, the set $\{n \in \mathbb{N} : T^n x \in U\}$ is syndetic.*

Proof. Assume (1). The given $U \neq \emptyset$ and $x \in X$, the point x is transitive so $T^n x \in U$ for some $n \in \mathbb{N}$, hence $x \in T^{-n}U$. Therefore $X = \bigcup_{n=0}^{\infty} T^{-n}U$. All the sets int he union are open so by compactness there is a finite sub-cover, giving (2).

Assume (2). fix x and U and let N be as in (2) for U. Since $X = \bigcup_{n=0}^{N-1} T^{-n}U$, for every k we have

$$X = T^{-kN}X = \bigcup_{n=0}^{N-1} T^{-kN-n}U = \bigcup_{n=kN}^{(k+1)N-1} T^{-n}U$$

so for each k there is an $kN \leq n_k < (k+1)N$ with $x \in T^{-n_k}U$, equivalently $T^{n_k}x \in U$. Since $n_{k+1} - n_k \leq 2N$, the sequence $\{n_k\}$ is syndetic and so is $\{n : T^n x \in U\}$.

Finally (3) implies that every x is transitive, so (X, T) is minimal. □

Corollary 11.3.8. *If A is finite and $X \subseteq A^{\mathbb{Z}}$ is minimal for the shift σ, then for every word $a \in A^*$ such that a appears in some point of X, there is a constant $L = L(a)$ such that for every $x \in X$ the set $\{i : x_i \dots x_{i+|a|-1} = a\}$ of appearances of a in x is syndetic with constant L.*

Example 11.3.9. Morse minimal system.

Remark 11.3.10. Minimality is a "indecomposability" condition and also implies that every point has a "well distributed" orbit. However this analog of ergodicity still fails to have an "ergodic decomposition": most systems are not unions of their minimal subsystems. For example, in $\{0,1\}^{\mathbb{Z}}$ there are many points in which some word appears non-syndetically, and hence the point does not belong to a minimal system. For example, $x = \ldots 0001111 \ldots$.

Definition 11.3.11. A point $x \in X$ is recurrent is if there is a sequence $n_k \to \infty$ such that $T^{n_k} x \to x$.

Lemma 11.3.12. *In a minimal system every point is recurrent.*

Proof. Let $x \in X$. If $Tx = x$ then the conclusion is obvious. Otherwise, there is some $y \in T^{-1}x$ with $y \neq x$. Since x is transitive, there is a sequence $n_k \to \infty$ such that $T^{n_k} x \to y$. Then $T^{n_k+1} x \to Ty = x$. □

Remark 11.3.13. The converse is false, e.g. in the disjoint union of two minimal systems every point is recurrent, but the system is not minimal.

Corollary 11.3.14. *Every system has a recurrent point.*

Proof. Choose a minimal subsystem and take any point in it. □

Remark 11.3.15. This proof of the existence of recurrent points is due to Birkhoff. It is purely topological. An alternative proof can be used by introducing invariant measures, and using Poincaré's theorem.

11.4 Invariant measures and unique ergodicity

Definition 11.4.1. Let $\mathcal{P}(X)$ denote the space of Borel probability measures on X and $\mathcal{P}_T(X)$ the subset of invariant ones

$$\mathcal{P}_T(X) = \{\mu \in \mathcal{P}(X) : \mu(T^{-1}A) = \mu(A) \text{ for all Borel sets } A \subseteq X\}$$

Lemma 11.4.2. *$\mu \in \mathcal{P}(X)$ is invariant if and only if $\int f d\mu = \int f \circ T d\mu$ for all $f \in C(X)$.*

Proof. Exercise. □

Definition 11.4.3. The weak-* topology on $C(X)$ is the weakest topology such that $\mu \mapsto \int f d\mu$ is continuous for each $f \in C(X)$.

Lemma 11.4.4. *$\mathcal{P}_T(X)$ is closed (compact) in the weak-* topology.*

For more on the weak-* topology see the Appendix.

Proposition 11.4.5. *Every topological dynamical system has invariant (and ergodic) Borel probability measures.*

Proof. Let $x \in X$ be any point and let

$$\mu_N = \frac{1}{N} \sum_{n=0}^{N-1} \delta_{T^n x}$$

Thus

$$\int f d\mu_N = \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x)$$

By compactness there is a subsequence $N_k \to \infty$ and $\mu \in \mathcal{P}(X)$ such that $\mu_{N_k} \to \mu$. We claim that μ is invariant. Indeed,

$$\begin{aligned}
\left| \int f d\mu - \int f \circ T d\mu \right| &= \lim_{k\to\infty} \left| \int f d\mu_{N_k} - \int f \circ T d\mu_{N_k} \right| \\
&= \lim_{k\to\infty} \left| \frac{1}{N_k} \sum_{n=0}^{N_k-1} f(T^n x) - \frac{1}{N_k} \sum_{n=0}^{N_k-1} f(T^{n+1} x) \right| \\
&= \lim_{k\to\infty} \frac{1}{N_k} |f(x) - f(T^{N_k+1} x)| \\
&= \lim_{k\to\infty} \frac{2 \|f\|_\infty}{N_k} \\
&= 0
\end{aligned}$$

Thus μ is invariant. □

Remark 11.4.6. The set of invariant measures is weak-* closed. In fact the ergodic measures are precisely the extreme points.

(If μ is not ergodic then there is a non-trivial invariant set A. Then $\mu = \mu(A)\mu_A + (1 - \mu(A))\mu_{X\setminus A}$ presents μ as a non-trivial convex combination of invariant measures, so μ is not an extreme point. Conversely, in the case that T is invertible, if $\mu = t + (1-t)\eta$ then $\nu \ll \mu$ and the Radon-Nykodim derivative $d\nu/d\mu$ can be shown to be invariant and is non-trivial if $\nu \neq \eta$. Thus μ is not ergodic. In the non-invertible case invariance of $d\nu/d\mu$ is not as trivial to show, but can be done, see Walters or my notes from last year).

Remark 11.4.7. The full shift has many ergodic measures, in fact they are dense in the set of invariant measures.

Definition 11.4.8. $x \in X$ is generic for a measure μ if $\frac{1}{N} \sum_{n=0}^{N-1} \delta_{T^n x} \to \mu$, i.e.

$$\frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) \to \int f d\mu \qquad \text{for all } f \in C(X) \tag{11.1}$$

The argument in the previous proposition shows that if x is generic for μ then μ is invariant.

Proposition 11.4.9. *The generic points of an ergodic measure have full measure.*

Proof. Let μ be an ergodic measure. It is enough to show that (11.1) holds for f in a countable dense algebra $\mathcal{F} \subseteq C(X)$. Such an algebra exists by stone-Weierstrass. Now, for every $f \in X$ the ergodic theorem implies that (11.1) holds for a.e. x, hence it a.s. holds for all $f \in \mathcal{F}$, as desired. □

Corollary 11.4.10. *If μ, ν are distinct ergodic measures then they are mutually singular.*

Proof. Let X_μ denote the set of generic points for μ and X_ν the set of generic points for ν. It is clear that these are Borel sets. Since $\mu \neq \nu$ no point can be generic for both so $X_\mu \cap X_\nu = \emptyset$. But $\mu(X_\mu) = \nu(X_\nu) = 0$ by the proposition, so $\mu \perp \nu$. □

Example 11.4.11. A point may be generic for a non-singular measure. For example let $x = 0^1 1^2 0^3 1^4 0^5 1^6 \dots$ where $a^n = aaa \dots a$ n-times. We leave it as an exercise to show that $x \in \{0,1\}^{\mathbb{N}}$ is generic for $\frac{1}{2}\delta_{0^\infty} + \frac{1}{2}\delta_{1^\infty}$, which is non ergodic since e.g. 0^∞ and 1^∞ are invariant sets of positive measure for it.

Definition 11.4.12. (X, T) is uniquely ergodic if it has a unique invariant probability measure.

Remark 11.4.13. If the measure is unique, it must be ergodic (otherwise its ergodic components would give additional ergodic measures, contradicting uniqueness).

Remark 11.4.14. A uniquely ergodic system with a fully supported invariant measure is minimal.

Proof. Similar to the proof that for a fully supported ergodic measure a.e. point is transitive; here we use the fact that every point is generic for μ to deduce that every point is transitive. The details are left as an exercise. □

Lemma 11.4.15. *The following are equivalent:*

1. *(X, T) is uniquely ergodic.*
2. *There is a measure $\mu \in \mathcal{P}(X)$ such that every point in X is generic for μ.*
3. *For every $f \in C(X)$, $\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to c(f)$ converges uniformly to a constant $c(f)$.*

Proof. (3) implies (2) since $f \mapsto c(f)$ defines a positive linear functional, hence by Riesz's theorem $c(f) = \int f d\mu$ for some $\mu \in \mathcal{P}(X)$, and then by definition every x is generic for μ.

(2) implies (1) because if there were two distinct invariant measures there would be distinct ergodic ones, and each would have generic points, contradicting (2).

(1) implies (3): Let μ be the unique invariant measure, so $\frac{1}{N}\sum_{n=0}^{N-1} T^n f \to \int f d\mu$ pointwise for every $f \in C(X)$. If convergence is not uniform then we can find a sequence of points x_i and $N_i \to \infty$ such that $\lim_{i\to\infty} \frac{1}{N_i}\sum_{n=0}^{N_i-1} f(T^n x_i) \neq$

$\int f d\mu$. Passing to a further subsequence we find an accumulation point μ' of $\frac{1}{N_i}\sum_{n=0}^{N_i-1}\delta_{T^n x_i}$ such that $\int f d\mu' \neq \int f d\mu$, so $\mu' \neq \mu$. But μ' is invariant (the same calculation as in Lemma **??** holds). This contradicts unique ergodicity. □

Exercise 11.4.16. The Morse minimal system is uniquely ergodic.

11.5 Isometries (we skip this in class)

Proposition 11.5.1. *Let (Y,d) be a compact metric space and $S: Y \to Y$ an isometry with a dense orbit. Then there is a compact metric group G and $g \in G$ and a homeomorphism $\pi: Y \to G$ such that $L_g\pi = \pi S$. Furthermore if ν is an invariant measure on Y then it is ergodic and $\pi\nu$ is Haar measure on G.*

Proof. Consider the group Γ of isometries of Y with the sup metric,

$$d(\gamma,\gamma') = \sup_{y\in Y} d(\gamma(y),\gamma'(y))$$

Then (Γ,d) is a complete metric space, and note that it is right invariant: $d(\gamma\circ\delta,\gamma\circ\delta) = d(\gamma,\gamma')$.

Let $y_0 \in Y$ have dense orbit and set $Y_0 = \{S^n y_0\}_{n\in\mathbb{Z}}$. If the orbit is finite, $Y = Y_0$ is a finite set permuted cyclically by S, so the statement is trivial. Otherwise $y \in Y_0$ uniquely determines n such that $S^n y_0 = y$ and we can define $\pi: Y_0 \to \Gamma$ by $y \mapsto S^n \in \Gamma$ for this n.

We claim that π is an isometry. Fix $y, y' \in Y_0$, so $y = S^n y_0$ and $y' = S^{n'} y_0$, so

$$d(\pi y, \pi y') = \sup_{z\in Y} d(S^n z, S^{n'} z)$$

Given $z \in Y$ there is a sequence $n_k \to \infty$ such that $S^{n_k} y_0 \to z$. But then

$$\begin{aligned} d(S^n z, S^{n'} z) &= d(S^n(\lim S^{n_k} y_0), S^{n'}(\lim S^{n_k} y_0)) \\ &= \lim d(S^n S^{n_k} y_0, S^{n'} S^{n_k} y_0) \\ &= \lim d(S^{n_k}(S^n y_0), S^{n_k}(S^{n'} y_0)) \\ &= \lim d(S^n y_0, S^{n'} y_0) \\ &= d(S^n y_0, S^{n'} y_0) \\ &= d(y, y') \end{aligned}$$

Thus $d(\pi y, \pi y') = d(y, y')$ and π is an isometry $Y_0 \hookrightarrow \Gamma$. Furthermore, for $y = S^n y_0 \in Y_0$,

$$\pi(Sy) == \pi(SS^n y) = S^{n+1} = L_S S^n = L_S \pi(y)$$

It follows that π extends uniquely to an isometry with $Y \hookrightarrow \Gamma$ also satisfying $\pi(Sy) = S(\pi y)$. The image $\pi(Y_0)$ is compact, being the continuous image of the compact set Y. Since $\pi(Y_0) = \{S^n\}_{n\in\mathbb{Z}}$ and this is a group its closure is also a group G.

Finally, suppose ν is an invariant measure on Y. Then $m = \pi\nu$ is L_S invariant on G. Since it is invariant under L_S it is invariant under $\{L_S^n\}_{n\in\mathbb{Z}}$, and this is a dense set of elements in G. Thus m it is invariant under every translation in G, and there is only one such measure up to normalization: Haar measure. The same argument applies to every ergodic components of m (w.r.t. L_S) and shows that the ergodic components are also Haar measure. Thus m is L_S-ergodic and since π is an isomorphism, (Y, ν, S) is ergodic. □

Corollary 11.5.2. *Irrational circle rotations are minimal and uniquely ergodic.*

Chapter 12

Topological Entropy via Covers

12.1 Definition

Let (X, T) be a topological dynamical system.

Definition 12.1.1. .

1. An *open cover* of X is a collection of open sets whose union is X.
2. If $\mathcal{U}, \mathcal{V}$ are open covers of X their *join* is $\mathcal{U} \vee \mathcal{V} = \{U \cap V \,:\, U \in \mathcal{U}\,,\, V \in \mathcal{V}\}$, it is also an open cover of X.
3. An open cover $\mathcal{U}$ refines an open cover $\mathcal{V}$ if every $U \in \mathcal{U}$ is a subset of some $V \in \mathcal{V}$
4. If $T : X \to X$ is a continuous map then $T^{-1}\mathcal{U} = \{T^{-1}U \,:\, U \in \mathcal{U}\}$ is an open cover.

Lemma 12.1.2. .

1. $T^{-1}(\mathcal{U} \vee \mathcal{V}) = T^{-1}(\mathcal{U}) \vee T^{-1}(\mathcal{V})$.
2. *If* $\mathcal{U}$ *refines* $\mathcal{V}$ *then* $T^{-1}(\mathcal{U})$ refines $T^{-1}(\mathcal{V})$.

Definition 12.1.3. For an open cover $\mathcal{U}$ we denote

$$N(\mathcal{U}) = \min\{|\mathcal{V}| \,:\, \mathcal{V} \subseteq \mathcal{U} \text{ is an open cover}\}$$

and

$$H(\mathcal{U}) = \log N(\mathcal{U})$$

Remark 12.1.4. By compactness every open cover has a finite sub-cover, so $N(\mathcal{U}) \in \mathbb{N}$.

Lemma 12.1.5. .

1. *$N(\mathcal{U}) \geq 1$ and $H(\mathcal{U}) \geq 0$, with equality if and only if $X \in \mathcal{U}$*
2. *$\mathcal{U}$ refines $\mathcal{V}$ implies $N(\mathcal{U}) \geq N(\mathcal{V})$ and $H(\mathcal{U}) \geq H(\mathcal{V})$.*
3. *$H(\mathcal{U} \vee \mathcal{V}) \leq H(\mathcal{U}) + H(\mathcal{V})$.*
4. *$H(T^{-1}\mathcal{U}) \leq H(\mathcal{U})$ and if T is onto then equality.*

Theorem 12.1.6. *$\lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U})$ exists for every open cover $\mathcal{U}$ of X.*

Proof. Write $a_n = H(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U})$. Then

$$\begin{aligned} a_{m+n} &= H(\bigvee_{i=0}^{(m+n)-1} T^{-i}\mathcal{U}) \\ &= H(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{U} \vee \bigvee_{i=m}^{(m+n)-1} T^{-i}\mathcal{U})) \\ &\leq H(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{U}) + H(\bigvee_{i=m}^{(m+n)-1} T^{-i}\mathcal{U})) \\ &\leq H(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{U}) + H(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U})) \\ &= a_m + a_n \end{aligned}$$

and the claim follows from sub-additivity. □

Definition 12.1.7. The *topological entropy* of (X, T) and an open cover $\mathcal{U}$ is

$$h_{top}(T,\mathcal{U}) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=-0}^{n-1} T^{-i}\mathcal{U})$$

Proposition 12.1.8. .

1. *$0 \leq h_{top}(T,\mathcal{U}) \leq H(\mathcal{U})$.*
2. *If $\mathcal{U}$ refines $\mathcal{V}$ then $h_{top}(T,\mathcal{U}) \geq h_{top}(T,\mathcal{V})$.*

Definition 12.1.9. The *topological entropy* of (X, T) is

$$h_{top}(T) = \sup_{\mathcal{U}} h_{top}(T,\mathcal{U})$$

where the supremum is over open covers $\mathcal{U}$ of X.

Remark 12.1.10. We can take the sup over finite sub-covers.

Proposition 12.1.11. .

1. $h_{top}(T) \geq 0$.
2. If $Y \subseteq X$ is a subsystem then $h_{top}(T) \geq h_{top}(T|_Y)$.
3. If T is invertible then $h_{top}(T) = h_{top}(T^{-1})$.

Definition 12.1.12. A topological system (Y, S) is a *factor* of (X, T) if there is a continuous onto map $\pi : X \to Y$ such that $\pi \circ T = S \circ \pi$.

Theorem 12.1.13. *If (Y, S) is a factor of (X, T) then $h_{top}(T) \geq h_{top}(S)$.*

Proof. Let $\pi : X \to Y$ be a factor map. If $\mathcal{U}$ is an open cover of $\mathcal{U}$ then $\pi^{-1}\mathcal{U} = \{\pi^{-1}U : U \in \mathcal{U}\}$ is an open cover of X and $N(\pi^{-1}\mathcal{U}) = N(\mathcal{U})$. Also $\pi^{-1}(\bigvee_{i=0}^{n-1} S^{-i}\mathcal{U}) = \bigvee_{i=0}^{n-1} T^{-i}\pi^{-1}\mathcal{U}$. Combining these two facts we find that $h_{top}(T, \pi^{-1}\mathcal{U}) = h_{top}(S, \mathcal{U})$. This shows that

$$h_{top}(T) = \sup_{\mathcal{V}} h_{top}(T, \mathcal{V}) \geq \sup_{\mathcal{U}} h_{top}(S, \mathcal{U}) = h_{top}(S) \qquad \square$$

12.2 Expansive systems

Definition 12.2.1. (X, T) is (forward) expansive if there is an $\varepsilon > 0$ such that for every $x, y \in X$ with $X \neq y$ there is an $n \in \mathbb{N}$ such that $d(T^n x, T^n y) > \varepsilon$. It is two-sided expansive if T is invertible and the same holds but allowing $n \in \mathbb{Z}$. The constant ε is called the expansiveness constant.

Remark 12.2.2. Although the definition uses the metric, the property of expansiveness is independent of the metric: if d' is another metric on X giving the same topology, then, since X is compact, for every $\varepsilon > 0$ there is an ε' such that if $d'(u, v) < \varepsilon'$ then $d(u, v) < \varepsilon$. It follows that if (X, T) is expansive and ε is the constant in the definition then ε' satisfies the same property for d'.

Lemma 12.2.3. *If ε is as in the definition of expansiveness, then for every $\delta > 0$ there is an $N = N(\delta)$ such that if $x, y \in X$ and $d(x, y) \geq \delta$ then there is an $n \in \{0, 1, \ldots, N-1\}$ with $d(T^n x, T^n y) > \varepsilon$.*

Proof. If not then there is some $\delta > 0$ such that for every N there is a pair $x_N, y_N \in X$ with $d(x_N, y_N) \geq \delta$ and $d(T^n x_k, T^n y_k) \leq \varepsilon$ for all $0 \leq n < N$. Passing to subsequence we can assume that $x_{N_k} \to x$ and $y_{N_k} \to y$. Evidently $d(x, y) \geq \delta$, so $x \neq y$, but for every n we have $n < N_k$ for all large k and by continuity of T, $d(T^n x, T^n y) = \lim d(T^n x_{N_k}, T^n y_{N_k}) \leq \varepsilon$. This contradicts expansiveness. $\square$

Lemma 12.2.4. *For any cover $\mathcal{U}$ and any N, $h_{top}(T, \mathcal{U}) = h_{top}(T, \bigvee_{i=0}^{N-1} T^{-i}\mathcal{U})$.*

Proof. Since $\bigvee_{i=0}^{N-1} T^{-i}\mathcal{U}$ refines $\mathcal{U}$ we certainly have $\leq$. For the other direction write $\mathcal{V} = \bigvee_{i=0}^{N-1} T^{-i}\mathcal{U}$ and notice that

$$\bigvee_{i=0}^{M-1} T^{-i}\mathcal{V} = \bigvee_{i=0}^{(N+M)-1} T^{-i}\mathcal{U}$$

hence

$$\begin{aligned} h_{top}(T,\mathcal{V}) &= \limsup \frac{1}{n} \log N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{V}) \\ &= \limsup \frac{1}{n} \log N(\bigvee_{i=0}^{n+N-1} T^{-i}\mathcal{U}) \\ &= h_{top}(T,\mathcal{U}) \qquad \square \end{aligned}$$

Proposition 12.2.5. *If (X,T) is expansive with expansive constant ε, and $\mathcal{U}$ is a cover of X by sets of diameter $\leq \varepsilon$, then $h_{top}(T) = h_{top}(T,\mathcal{U})$.*

Proof. It suffices to show that for every open cover $\mathcal{V}$ we have $h_{top}(T,\mathcal{U}) \geq h_{top}(T,\mathcal{V})$.

Let δ be a Lebesgue covering number of $\mathcal{V}$, so for every $x \in X$ we have $\overline{B_\delta}(x) \subseteq V$ for some $V \in \mathcal{V}$. Let $N = N(\delta)$ be as in the lemma and $\mathcal{U}' = N(\bigvee_{i=0}^{N-1} T^{-i}\mathcal{U})$. We claim that every element of $\mathcal{U}'$ has diameter $< \delta$. Indeed, if $d(x,y) \geq \delta$ then there is some $0 \leq n < N$ with $d(T^n x, T^n y) > \varepsilon$, and hence $T^n x, T^n y$ cannot both belong to the same element of $\mathcal{U}$, hence x, y do not belong to the same element of $T^{-n}\mathcal{U}$. This shows that every x, y which belong to the same element of $\mathcal{U}'$ satisfy $d(x,y) < \delta$ as claimed.

It follows that $\mathcal{U}'$ refines $\mathcal{V}$, hence $h_{top}(\mathcal{U}') \geq h_{top}(\mathcal{V})$. But $h_{top}(\mathcal{U}) = h_{top}(\mathcal{U}')$ be the previous lemma and the proposition follows. $\square$

Corollary 12.2.6. *An expansive map has finite topological entropy.*

Example 12.2.7. Let $X = A^{\mathbb{N}}$ for a finite set A and T the shift. Then $h_{top}(T) = \log |A|$.

Indeed, define the metric by

$$d(x,y) = 2^{-n} \quad \text{where} \quad n = \min\{i \in \mathbb{N} : x_i \neq y_i\}$$

Note that if $x_1 \neq y_1$ then $d(x,y) \geq \frac{1}{2}$. Since $x \neq y$ implies that $x_n \neq y_n$ for some n, and $(T^n x)_1 = x_n \neq y_n = (T^n y)_1$, we have $d(T^n x, T^n y) \geq \frac{1}{2}$, so T is expansive with constant $\frac{1}{2}$. Also note that if $x_1 = y_1$ then $d(x,y) \leq \frac{1}{2}$, so the cylinder sets

$$[a] = \{x \in X : x_1 = a\}$$

are open (and closed) sets of diameter $\frac{1}{2}$. By the proposition, $h_{top}(T) = h_{top}(T,\mathcal{U})$ for the partition $\mathcal{U} = \{[a] : a \in A\}$. Finally, $\bigvee_{i=1}^{n} T^{-i}\mathcal{U}$ is the partition of X according tot he initial n-segments of sequences $x \in X$ and consists of $|A|^n$ pairwise disjoint sets, so it has no proper subcovers and $N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}) = |A|^n$. Thus $h_{top}(T,\mathcal{U}) = \log |A|$, as claimed.

Corollary 12.2.8. *Let A, B be finite sets. If $|B| > |A|$ then there is no factor map from $A^{\mathbb{Z}} \to B^{\mathbb{Z}}$. In particular $A^{\mathbb{Z}}, B^{\mathbb{Z}}$ are isomorphic if and only if $|A| = |B|$.*

Example 12.2.9. Let A be finite and $X \subseteq A^{\mathbb{Z}}$a sybsystem. Let

$$L_n(X) = \#\{w \in A^n \,:\, w \text{ appears in } X\}$$

Then $h_{top}(T|_X) = \lim_{n\to\infty} \frac{1}{n} \log L_n(X)$.

Indeed, T is expansive with the same constant as before so for the partition $\mathcal{U}$ into cylinders $[a]\cap X$, $a \in A$, we have again $h_{top}(T|_X) = \lim \frac{1}{m} \log N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U})$. But $N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}) = L_n(X)$ and the claim follows.

Chapter 13

Topological Entropy via Separated Sets

13.1 Spanning and separating sets

Definition 13.1.1. Let (X, d) be a compact metric space and $\varepsilon > 0$.

1. The ε-covering number $cov(X, d, \varepsilon)$ is the minimal number of points in an ε-dense set, i.e.

$$cov(X, d, \varepsilon) = \min\{n : \exists x_1, \ldots, x_n \in X \text{ s.t. } X = \bigcup_{i=1}^{n} \overline{B_\varepsilon}(x_i)\})$$

2. The ε-separation number, $sep(X, d, \varepsilon)$, is the maximal number of ε-separated points, i.e.

$$sep(X, d, \varepsilon) = \max\{n : \exists y_1, \ldots, y_n \in X \text{ s.t. } d(y_i, y_j) > \varepsilon \text{ for all } i \neq j\}$$

Remark 13.1.2. By compactness, both numbers are finite.
If $\varepsilon' < \varepsilon$ then $cov(X, d, \varepsilon') \geq cov(X, d, \varepsilon)$ and $sep(X, d, \varepsilon') \geq sep(X, d, \varepsilon)$.

Lemma 13.1.3. $cov(X, d, \varepsilon/2) \geq sep(X, d, \varepsilon) \geq cov(X, d, \varepsilon)$

Proof. Suppose that $x_1, \ldots, x_n$ is a maximal ε-separated set, so $n = sep(X, d, \varepsilon)$. If $X \not\subseteq \bigcup \overline{B_\varepsilon}(x_i)$ there is an $x \in X$ such that $d(x, x_i) \geq> \varepsilon$ for all i and then $x_1, \ldots, x_n, x$ would also be ε-separated, contradicting maximality. Hence $X = \bigcup B_\varepsilon(x_i)$ and $cov(X, d, \varepsilon) \leq n = sep(X, d, \varepsilon)$.

On the other hand if $X = \bigcup_{i=1}^{m} \overline{B_{\varepsilon/2}}(y_i)$ then for any ε-separated set $x_1, \ldots, x_n$, no two of the points x_i are in the same ball $\overline{B_\varepsilon}(y_j)$, but each x_i is in at least one such ball, hence $n \leq m$. It follows that $cov(X, d, \varepsilon/2) \geq sep(X, d, \varepsilon)$. □

13.2 Bowen's definition of entropy

Definition 13.2.1. If (X,T) is a topological dynamical system, d a metric on X, then

$$d_n(x,y) = \max_{0\le i\le n-1} d(T^n x, T^n y)$$

Observe that the ε-ball around x in d_n is $\bigcap_{i=0}^{n-1} B_\varepsilon(T^n x)$.

Definition 13.2.2. For $\varepsilon > 0$,

$$\begin{aligned} h_{sep}(T,d,\varepsilon) &= \limsup_{n\to\infty} \frac{1}{n} \log sep(X,d_n,\varepsilon) \\ h_{cov}(T,d,\varepsilon) &= \limsup_{n\to\infty} \frac{1}{n} \log cov(X,d_n,\varepsilon) \end{aligned}$$

and

$$\begin{aligned} h_{sep}(T,d) &= \lim_{\varepsilon\to 0} h_{sep}(T,\varepsilon) \\ &= \sup_{\varepsilon\to 0} h_{sep}(T,\varepsilon) \end{aligned}$$

and

$$\begin{aligned} h_{cov}(T,d) &= \lim_{\varepsilon\to 0} h_{cov}(T,\varepsilon) \\ &= \sup_{\varepsilon\to 0} h_{cov}(T,\varepsilon) \end{aligned}$$

Remark 13.2.3. Since $cov(X,d,\varepsilon/2) \ge sep(X,d,\varepsilon) \ge cov(X,d,\varepsilon)$ we have

$$h_{cov}(T,d,\varepsilon) \le h_{sep}(T,d,\varepsilon) \le h_{cov}(T,d,\varepsilon/2)$$

so

$$h_{sep}(T,d) = h_{cov}(T,d)$$

Lemma 13.2.4. *$h_{sep}(T), h_{cov}(T)$ are independent of the metric (depend only on the topology).*

Proof. Let d, d' be two metrics compatible with the topology on X. For every $\varepsilon > 0$ there is an $\varepsilon' > 0$ such that if $d'(x,y) < \varepsilon'$ then $d(x,y) < \varepsilon$. Thus $B'_{\varepsilon'}(x) \subseteq N_\varepsilon(x)$, where B' denotes the ball with respect to d'. It follows that $cov(X,d',\varepsilon') \ge cov(X,d,\varepsilon)$ and $cov(X,d'_n,\varepsilon') \ge cov(X,d_n,\varepsilon)$. Hence $h_{cov}(T,d',\varepsilon') \ge h_{cov}(T,d,\varepsilon)$. Hence

$$h_{cov}(T,d') = \sup_{\varepsilon'} h_{cov}(T,d,\varepsilon') \ge \sup_{\varepsilon} h_{cov}(T,d,\varepsilon) = h_{cov}(T,d)$$

The other inequality is symmetric. The claim about h_{sep} follows from the fact that it is the same as h_{cov}. □

In view of the last lemma, from now on we drop the metric from the notation and write $h_{cov}(T), h_{sep}(T)$.

Example 13.2.5. If T is an isometry, then $d_n = d$. Hence $cov(X,d_n,\varepsilon) = cov(X,d,\varepsilon)$ is independent of n and $\frac{1}{n} cov(X,d_n,\varepsilon) \to 0$. Taking ε also, we have $h_{cov}(T) = 0$.

13.3 Equivalence of the definitions

For an open cover $\mathcal{U}$ write $\operatorname{diam}\mathcal{U} = \max\{\operatorname{diam} U : U \in \mathcal{U}\}$.

Proposition 13.3.1. *Let $\mathcal{U}_n$ be open covers with $\operatorname{diam}\mathcal{U}_n \to 0$. Then*

$$h_{top}(T) = \lim_{n\to\infty} h_{top}(T, \mathcal{U}_n)$$

Proof. First, for any open cover $\mathcal{V}$, let δ be a Lebesgue number for $\mathcal{V}$. Then for large enough n we have that $\operatorname{diam}\mathcal{U}_n < \delta$ so $\mathcal{U}_n$ refines $\mathcal{V}$ and $h_{top}(T, \mathcal{U}_n) \geq h_{top}(\mathcal{V})$. In particular, taking $\mathcal{V} = \mathcal{U}_{n_0}$, this shows that $\lim h_{top}(T, \mathcal{U}_n)$ exists, and that the limit is at least as large as $\sup_{\mathcal{V}} h_{top}(T, \mathcal{V})$. Since it also does not exceed this supremum and the supremum is equal by definition to $h_{top}(T)$, we are done. □

Proposition 13.3.2. *If $\mathcal{U}$ is an open cover with Lebesgue number δ then*

$$N(\bigvee_{i=1}^{n-1} T^{-i}\mathcal{U}) \leq cov(X, d_n, \delta/2) \leq sep(X, d_n, \varepsilon/2)$$

Proof. We have already seen the right inequality. For the left one, notice that in the metric d_n the open cover $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}$ has Lebesgue number δ. Therefore if $\mathcal{U}_n$ is an optimal cover of (X, d_n) by $\delta/2$ balls, then its elements have diameter δ and it refines $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}$. Thus $N(\bigvee_{i=1}^{n-1} T^{-i}\mathcal{U}) \leq N(\mathcal{U}_n) = cov(X, d_n, \delta/2)$. □

Proposition 13.3.3. *If $\mathcal{U}$ is an open cover with $\operatorname{diam}\mathcal{U} \leq \varepsilon$, then*

$$cov(X, d_n, \varepsilon) \leq sep(X, d_n, \varepsilon) \leq N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U})$$

Proof. The left inequality was already proved. For the right one, note that if $x_1, \ldots, x_m$ is ε-separated in d_n then for each x_i, x_j there is some $0 \leq k \leq n-1$ such that $d(T^k x_i, T^k x_j) > \varepsilon$. This means that $T^k x_i, T^k x_j$ do not lie in a common element of $\mathcal{U}$, equivalently x_i, x_j do not lie in a common element of $T^{-k}\mathcal{U}$, so they do not lie in a common element of $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}$. This means that a subcover of $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}$ must contain at least m sets. Taking a maximal separated set, with $m = sep(X, d_n, \varepsilon)$, we find that $N(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{U}) \geq sep(X, d_n, \varepsilon)$. □

Theorem 13.3.4. $h_{top}(T) = h_{sep}(T) = h_{cov}(T)$.

Proof. Let $\mathcal{U}_n$ be open covers with $\operatorname{diam}\mathcal{U}_n < 1/n$, so

$$h_{top}(T) = \lim_{n\to\infty} h_{top}(T, \mathcal{U}_n)$$

Now for each n, by the previous proposition with $\varepsilon = 1/n$,

$$\begin{aligned} h_{top}(T, \mathcal{U}_n) &= \limsup_{N\to\infty} \frac{1}{N} \log N(\bigvee_{i=0}^{N-1} T^{-i}\mathcal{U}_n) \\ &\geq \limsup_{N\to\infty} \frac{1}{N} \log cov(X, d_n, 1/n) \\ &= h_{cov}(T, d, 1/n) \end{aligned}$$

so taking $n \to \infty$ we conclude

$$h_{top}(T) \geq h_{cov}(T)$$

On the other hand let δ_n be the Lebesgue covering number of $\mathcal{U}_n$ and note that $\delta_n \leq \operatorname{diam} \mathcal{U}_n \to 0$. Then by the other proposition,

$$\begin{aligned} h_{top}(T, \mathcal{U}_n) &= \limsup_{N\to\infty} \frac{1}{N} \log N(\bigvee_{i=0}^{N-1} T^{-i}\mathcal{U}_n) \\ &\leq \limsup_{N\to\infty} \frac{1}{N} \log cov(X, d_n, \delta_n/2) \\ &= h_{cov}(T, d, \delta_n/2) \end{aligned}$$

again taking $n \to \infty$ we obtain

$$h_{top}(T) \leq h_{cov}(T)$$

as claimed. □

Chapter 14

Interplay Between Measurable and Topological Entropy

14.1 The Variational Principle

Theorem 14.1.1. *Let (X,T) be a topological dynamical system. Then*

$$h_{top}(T) = \sup\{h_\mu(T) \,:\, \mu \in \mathcal{P}_T(X)\}$$

The proof has two directions which we prove separately. First we have a lemma about the size of the "hamming ball" around a word. Let A be a finite set, and for $n \in \mathbb{N}$ define a metric on A^n by

$$u_n(a,b) = \frac{1}{n}\#\{0 \le i \le n-1 \,:\, a_i \neq b_i\}$$

This is called the Hamming distance.

Lemma 14.1.2. *For any $a \in A^n$ and $\varepsilon > 0$,*

$$|\{b \in A^n \,:\, u_n(a,b) < \varepsilon\}| \le 2^{n(H(\varepsilon)+\varepsilon \log |A|)}$$

Proof. In order to obtain a point b in the set above, one first chooses a set $I \subseteq \{0,\dots,n-1\}$ of size $|I| \le \varepsilon n$, and then changes the values of a_i for $i \in I$. The number of choices for I is $\binom{n}{\varepsilon n} \le 2^{H(\varepsilon)}$, and the number of ways to modify a at the indices in I is $(|A|-1)^{|I|} \le 2^{\varepsilon n \log |A|}$. Thus the number of b's in the set above is bounded by the product of these two bounds. □

Corollary 14.1.3. *Let $t > 0$ and let $U \subseteq A^n$ be a set with $|U| > 2^{tn \log |A|}$. Then for every $\varepsilon > 0$ there is a subset $U' \subseteq U$ with $|U'| > 2^{n((t-\varepsilon)\log |A| - H(\varepsilon))}$ such that $u_n(a,b) \ge \varepsilon$ for all distinct $a,b \in U'$.*

Proof. Choose U' by induction, greedily: start with an arbitrary $a \in U$ and replace U by $U \setminus \{b \in A^n : u_n(b,a) < \varepsilon\}$. We have thus removed at most $2^{n(H(\varepsilon)+\varepsilon \log |A|)}$ points from U. Choose another point in U and iterate. This can go on for k steps provided that $k \cdot 2^{n(H(\varepsilon)+\varepsilon \log |A|)} < |U|$ so certainly it can go on for $2^{nt \log |A|}/2^{n(H(\varepsilon)+\varepsilon \log |A|)}$ steps, so the set U' we end up with has the desired size and clearly also the second property. □

Proposition 14.1.4. *$h_{top}(T) \geq h_\mu(T)$ for every $\mu \in \mathcal{P}_T(X)$.*

Proof. Let $\mu \in \mathcal{P}_T(X)$ and let $\mathcal{P}$ be a finite measurable partition of X, we must show that $h_{top}(T) \geq h_\mu(T, \mathcal{P})$.

Fix $\varepsilon > 0$ and for each $A_i \in \mathcal{P}$ let $A' \subseteq A$ be a compact set such that $\mu(A') > (1 - \frac{\varepsilon}{2})\mu(A)$. Let $A'' = A \setminus A'$ and let $\mathcal{Q} = \{A', A''\}_{A \in \mathcal{P}}$. This is again a measurable partition of X and it refines $\mathcal{P}$ so $h_\mu(T, \mathcal{Q}) \geq h_\mu(T, \mathcal{P})$. Denote

$$h = h_\mu(T, \mathcal{Q})$$

Let $E = \bigcup_{A \in \mathcal{P}} A''$ and note that $\mu(E) < \frac{\varepsilon}{2}$. Denote

$$\delta = \min_{A \neq B \in \mathcal{P}} d(A', B') > 0$$

(using compactness), and observe that if $x, y \in X$ and for some $i < n$ we have $\mathcal{Q}(T^i x) \neq \mathcal{Q}(T^i y)$ and $T^i x, T^i y \notin E$, then $d(T^i x, T^i y) \geq \delta$, hence $d_n(x, y) \geq \delta$. Our goal is to produce a large set of points for which each pair satisfies the above.

By the SMB theorem and the ergodic theorem, for μ-a.e. x for all large n there is a set $F_n \subseteq X$ such that for $x \in F_n$,

$$2^{-n(h+\varepsilon)} < \mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}(x)) < 2^{-n(h-\varepsilon)}$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1} 1_E(T^i x) < \frac{\varepsilon}{2}$$

Let U_n be the set of words $a = (\mathcal{Q}(T^i x))_{i=0}^{n-1}$ for $x \in F_n$. Then $|U_n| > 2^{n(h+\varepsilon)}$, so by the corollary above, there is a subset $U_n' \subseteq U_n$ with $|U_n'| > 2^{n(h+\varepsilon-\varepsilon \log |A|-H(\varepsilon))}$ and $u_n(a,b) \geq \varepsilon$ for distinct $a, b \in U_n'$. For each $a \in U_n'$ there is an $x = x(a) \in F_n$ such that $a_i = \mathcal{Q}(T^i x)$. Now, for distinct $a, b \in U_n'$, for $x = x(a)$ and $y = y(b)$ we know that $\frac{1}{n}\sum_{i=0}^{n-1} 1_E(T^i x) < \frac{\varepsilon}{2}$ and $\frac{1}{n}\sum_{i=0}^{n-1} 1_E(T^i y) < \frac{\varepsilon}{2}$, and $u_n(a,b) \geq \varepsilon$. This implies that there must be some $0 \leq i \leq n-1$ such that $\mathcal{Q}(T^i x) \neq \mathcal{Q}(T^i y)$ and $T^i x, T^i y \notin E$. As discussed above, this implies that $d_n(x, y) > \delta$. Thus, the collection $\{x(a) : a \in U_n'\}$ is δ-separated in d_n so for all large enough n

$$sep(X, d_n, \delta) \geq 2^{n(h+\varepsilon-\varepsilon \log |A|-H(\varepsilon))}$$

This implies that

$$h_{top}(T) \geq h - (\log |A| - 1)\varepsilon + H(\varepsilon)$$

Since ε was arbitrary this shows that $h_{top}(T) \geq h$ as desired. □

For the other direction we need two easy facts topologically nice partitions.

Lemma 14.1.5. *Let μ be Borel a measure on a metric space (X, d). Then for every $\varepsilon > 0$ there exists a Borel partition $\mathcal{P}$ of X such that* $\operatorname{diam} A < \varepsilon$ *and $\mu(\partial A) = 0$ for $A \in \mathcal{P}$.*

Proof. Fix $\varepsilon > 0$. For $x \in X$ the sets $\partial B_r(x)$ are pairwise disjoint Borel sets, and there are continuum many of them for $r < \varepsilon/2$, so they cannot all have positive measure. Thus there is some $r = r(x)$ such that $\mu(\partial B_r(x)) = 0$. By compactness we can choose a finite sequence $x_1, \ldots, x_n$ such that $B_{r(x_i)}(x)$ cover X. The partition generated by these balls has the desired properties. □

Lemma 14.1.6. *Let $\mathcal{P}$ be a partition of a metric space X. Let $\mu_n, \mu \in \mathcal{P}(X)$ and suppose that $\mu_n \to \mu$ weak-* and that $\mu(\partial A) = 0$ for every $A \in \mathcal{P}$. Then $H(\mu_n, \mathcal{P}) \to H(\mu, \mathcal{P})$.*

Proof. If suffices to note that under the assumptions, $\mu_n(\partial A) \to \mu(\partial A)$ for every $A \in \mathcal{P}$. □

Lemma 14.1.7. *Let μ be a T-invariant measure on X and $\mathcal{P}$ a partition. Then for every $k < n$,*

$$\begin{aligned} \frac{1}{n} H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) &\leq \frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{k} H_\mu(\bigvee_{j=i}^{i+k-1} T^{-j}\mathcal{P}) + O(\frac{k \log |\mathcal{P}|}{n}) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{k} H_{T^i\mu}(\bigvee_{j=0}^{k-1} T^{-j}\mathcal{P}) + O(\frac{k \log |\mathcal{P}|}{n}) \end{aligned}$$

In particular, writing $\nu = \frac{1}{n}\sum_{i=0}^{n-1} T^i\mu$, we have

$$\frac{1}{n} H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) \leq \frac{1}{k} H_\nu(\bigvee_{j=0}^{k-1} T^{-j}\mathcal{P}) + O(\frac{k \log |\mathcal{P}|}{n})$$

Proof. The first two statements are identical by measure preservation, which gives the identity $H_\mu(\bigvee_{j=i}^{i+k-1} T^{-j}\mathcal{P}) = H_{T^i\mu}(\bigvee_{j=0}^{k-1} T^{-j}\mathcal{P})$. To derive the second statement from the first, use concavity of entropy in the measure argument and the definition of ν to deduce that

$$H_\nu(\bigvee_{j=0}^{k-1} T^{-j}\mathcal{P}) \geq \frac{1}{n}\sum_{i=0}^{n-1} H_{T^i\mu}(\bigvee_{j=0}^{k-1} T^{-j}\mathcal{P})$$

and apply the first part. Thus we must prove the first equality.

For this, note that for any $0 \le m < k$ there is a maximal integer n_m such that $k \cdot n_m + m \le n$. We decompose the partition $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}$ into "blocks" of length k starting at integers $j \equiv m \bmod k$. Using the inequality $H_\mu(\mathcal{U} \vee \mathcal{V}) \le H_\mu(\mathcal{U}) + H_\mu(\mathcal{V})$,

$$\begin{aligned}
H_\mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) &\le H_\mu\left(\bigvee_{i=0}^{k(n_m+1)+m-1} T^{-i}\mathcal{P}\right) \\
&= H_\mu\left(\left(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{P}\right) \vee \left(\bigvee_{i=m}^{k(n_m+1)+m-1} T^{-i}\mathcal{P}\right)\right) \\
&= H_\mu\left(\left(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{P}\right) \vee \left(\bigvee_{j=0}^{n_m}\bigvee_{i=0}^{k-1} T^{-(jk+m+i)}\mathcal{P}\right)\right) \\
&\le H_\mu\left(\bigvee_{i=0}^{m-1} T^{-i}\mathcal{P}\right) + \sum_{j=0}^{n_m} H_\mu\left(\bigvee_{i=0}^{k-1} T^{-(jk+m+i)}\mathcal{P}\right) \\
&= \sum_{j=0}^{n_m} H_\mu\left(\bigvee_{i=0}^{k-1} T^{-(jk+m+i)}\mathcal{P}\right) + O(k \log |\mathcal{P}|)
\end{aligned}$$

where in the last line we used that $\bigvee_{i=0}^{m-1} T^{-i}\mathcal{P}$ has at most $|\mathcal{P}|^m \le |\mathcal{P}|^k$ atoms and hence entropy at most $k \log |\mathcal{P}|$. Now average over $m = 0, \ldots, k-1$ and divide by n. Noticing that every expression of the form $\bigvee_{j=i}^{i+k-1} T^{-j}\mathcal{P}$ occurs once for $0 \le j < n$, we obtain the stated inequality. □

Proposition 14.1.8. *For $\varepsilon > 0$ there exists $\mu \in \mathcal{P}_T(X)$ such that $h_\mu(T) \ge h_{top}(T) - \varepsilon$.*

Proof. By Bowen's definition of topological entropy, there is a $\delta > 0$ such that $h_{top}(T, d, \delta) > h_{top}(T) - \varepsilon$ and for every large enough n there is a set $X_n \subseteq X$ with $|X_n| > 2^{n(h_{top}(T)-\varepsilon)}$ such that $d_n(x, y) \ge \delta$ for distinct $x, y \in X_n$. Let

$$\xi_n = \frac{1}{|X_n|} \sum_{x \in X_n} \delta_{x_n}$$

and set

$$\begin{aligned}
\mu_n &= \frac{1}{n} \sum_{i=0}^{n-1} T^i \xi_n \\
&= \frac{1}{|X_n|} \sum_{x \in X_n} \left(\frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i x}\right)
\end{aligned}$$

Let μ be any accumulation point of μ_n. It is easily checked that μ is invariant. We claim that $h_\mu(T) \ge h_{top}(T) - \varepsilon$.

Choose a partition $\mathcal{P}$ such that $\operatorname{diam} A < \varepsilon$ and $\mu(\partial A) = 0$ for $A \in \mathcal{P}$, as can be done by Lemma **??**. Since $\partial(T^{-1}A) = T^{-1}(\partial A)$, by measure preservation, $\mu(T^{-i}\partial A) = 0$ for every $A \in \mathcal{P}$, hence the same is true for $A \in \bigvee_{i=0}^{n-1} T^{-i}P$ (since the boundary of such A is a finite union of boundaries of atoms of $T^{-i}P$). Therefore, using the previous lemma and the definition of μ_n, for each k we have

$$\begin{aligned}\frac{1}{k}H_\mu(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) &= \lim_{n\to\infty} \frac{1}{k}H_{\mu_n}(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) \\ &\geq \lim_{n\to\infty} \frac{1}{n}H_{\xi_n}(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P})\end{aligned}$$

Now, ξ_n consists of $|X_n|$ equally weighted atoms, and for any two of these atoms x, y there is an $0 \leq i \leq n-1$ such that $d(T^i x, T^i y) \geq \delta$, so $T^i x, T^i y$ are in different atoms of $\mathcal{P}$ (since the atoms of $\mathcal{P}$ have diameter $< \delta$). Thus, each atom of μ_n lies in a different atom of $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}$, so for all large enough n,

$$H_{\xi_n}(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) = \log|X_n| > n(h_{top}(T) - \varepsilon)$$

We have concluded that for each k we have $\frac{1}{k}H_\mu(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) \geq h_{top}(T) - \varepsilon$. Taking $k \to \infty$ we conclude that $h_\mu(T, \mathcal{P}) \geq h_{top}(T) - \varepsilon$, as desired. □

14.2 The entropy function

Let T be a measurable transformation of a standard Borel space X. For $\mu \in \mathcal{P}_T(X)$ we denote $h(\mu) = h_\mu(T)$. This defines a function

$$h : \mathcal{P}_T(X) \to [0, \infty)$$

This is a measurable function since for any P clearly $\mu \mapsto h_\mu(T, \mathcal{P})$ is measurable, and h is the supremum of such maps over a dense set of partitions.

The point of the following lemma is that not only is Shannon entropy $H_\mu(\mathcal{P})$ is concave in the μ argument, but it is nearly convex: the defect is bounded independently of the partition.

Lemma 14.2.1. *Let $\mu, \nu \in \mathcal{P}(X)$ and $\mathcal{P}$ a partition of X. Then for any $0 < t < 1$,*

$$H_{t\mu+(1-t)\nu}(\mathcal{P}) \leq tH_\mu(\mathcal{P}) + (1-t)H_\nu(\mathcal{P}) + H(t)$$

Proof. Let X_μ, X_ν be two disjoint copies of the underlying probability space X and consider μ' on X_μ a copy of μ and ν' on X_ν a copy of ν. Let $\theta = t\mu' + (1-t)\nu'$ a probability measure on $Y = X_\mu \cup X_\nu$. Let $\mathcal{Q} = \{X_\mu, X_\nu\}$. For each $A \in \mathcal{P}$ let A_μ, A_ν denote the corresponding set in X_μ, X_ν respectively, and let $\widetilde{A} = A_\mu \cup A_\nu$.

Set $\widetilde{\mathcal{P}} = \{\widetilde{A} : A \in \mathcal{P}\}$. Now observe that $\theta(\widetilde{A}) = t\mu(A) + (1-t)\nu(A)$ for $\widetilde{A} \in \widetilde{\mathcal{P}}$, hence

$$H_\theta(\widetilde{\mathcal{P}}) = H_{t\mu+(1-t)\nu}(\mathcal{P})$$

On the other hand

$$\begin{aligned} H_\theta(\widetilde{\mathcal{P}}) &\leq H_\theta(\widetilde{\mathcal{P}} \vee \mathcal{Q}) \\ &= H_\theta(\mathcal{Q}) + H_\theta(\widetilde{\mathcal{P}}|\mathcal{Q}) \\ &= H(t) + \sum_{Q\in\mathcal{Q}} \theta(Q) H_{\theta_Q}(\widetilde{\mathcal{P}}) \\ &= H(t) + tH_\mu(\mathcal{P}) + (1-t)H_\nu(\mathcal{P})\square \end{aligned}$$

Proposition 14.2.2. *Let $(X, \mathcal{F})$ be a measurable space and $T : X \to X$ a measurable map. Let $\mathcal{P}_T(X)$ denote the set of T-invariant probability measures and for a partition $\mathcal{P}$ of X let $h_\mathcal{P} : \mathcal{P}_T(X) \to \mathbb{R}$ the entropy map, $\mu \mapsto h_\mu(T, \mathcal{P})$. Then $h_\mathcal{P}$ is affine: $h_\mathcal{P}(t\mu + (1-t)\nu) = th_\mathcal{P}(\mu) + (1-t)h_\mathcal{P}(\nu)$. Consequently also h is affine.*

Proof. Let $\theta = t\mu + (1-t)\mu$, we must show that $h_\mathcal{P}(\theta) = th_\mathcal{P}(\mu) + (1-t)h_\mathcal{P}(\nu)$. By concavity of entropy,

$$\begin{aligned} h_\theta(T, \mathcal{P}) &= \lim_{n\to\infty} \frac{1}{n} H_\theta(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) \\ &\geq \lim_{n\to\infty} \frac{1}{n} \left(tH_\mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) + (1-t)H_\nu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) \right) \\ &= th_\mu(T, \mathcal{P}) + (1-t)h_\nu(T, \mathcal{P}) \end{aligned}$$

On the other hand by the previous lemma

$$\begin{aligned} h_\theta(T, \mathcal{P}) &= \lim_{n\to\infty} \frac{1}{n} H_\theta(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) \\ &\leq \lim_{n\to\infty} \frac{1}{n} \left(tH_\mu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) + (1-t)H_\nu(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}) + H(t) \right) \\ &= th_\mu(T, \mathcal{P}) + (1-t)h_\nu(T, \mathcal{P}) \end{aligned}$$

This proves that $h_\mathcal{P}$ is affine. For the last statement take the supremum over finite partitions $\mathcal{P}$ in the identity $h_\mathcal{P}(\theta) = th_\mathcal{P}(\mu) + (1-t)h_\mathcal{P}(\nu)$. the left hand side is $h(\theta)$. It remains to show that

$$\sup_\mathcal{P} (th_\mathcal{P}(\mu) + (1-t)h_\mathcal{P}(\nu)) = \sup_\mathcal{P} th_\mathcal{P}(\mu) + \sup_\mathcal{P} (1-t)h_\mathcal{P}(\nu)$$

Indeed, $\leq$ is automatic, whereas for any partitions $\mathcal{P}', \mathcal{P}''$ we have

$$th_{\mathcal{P}'}(\mu) + (1-t)h_{\mathcal{P}''}(\nu) \leq th_{\mathcal{P}'\vee\mathcal{P}''}(\mu) + (1-t)h_{\mathcal{P}'\vee\mathcal{P}''}(\nu)$$

which gives the reverse inequality. $\square$

Proposition 14.2.3. *Let $(X, \mathcal{F})$ be a standard Borel space and $T : X \to X$ measurable. For any T-invariant measure θ with ergodic decomposition $\theta = \int \nu d\tau(\nu)$, and any partition $\mathcal{P}$, we have*

$$h_T(\theta, \mathcal{P}) = \int h_\nu(T, \mathcal{P})\, d\tau(\nu)$$

and

$$h_T(\theta) = \int h_\nu(T)\, d\tau(\nu)$$

Proof. X is a standard Borel space so it is isomorphic as a measurable space to a compact metric space, hence $\mathcal{P}(X)$ and also $\mathcal{P}_T(X)$ can be given the weak topology in which they are compact.

Fix a partition $\mathcal{P}$ of X. Since τ is a probability measure on $\mathcal{P}_T(X)$, the measurable function $h_{\mathcal{P}} : \mu \mapsto h_\mu(T, \mathcal{P})$ is continuous on compact subsets of $\mathcal{P}_T(X)$ of arbitrarily large τ measure. Fix $\varepsilon \geq 0$ and let $\mathcal{W}_\varepsilon \subseteq \mathcal{P}_T(X)$ be such a set with $\tau(\mathcal{W}_\varepsilon) = 1 - \delta$ where $\delta \leq \varepsilon$. For simplicity we assume that $\delta = \varepsilon$ which can be achieved if τ is non-atomic, otherwise we might have $\delta < \varepsilon$, the proof then continues in the same way but with slightly more complicated notation.

Write $\tau_\varepsilon = \frac{1}{1-\varepsilon}\tau|_{\mathcal{W}_\varepsilon}$ and $\tau'_\varepsilon = \frac{1}{\varepsilon}\tau|_{\mathcal{W}_\varepsilon^c}$, so these are probability measures and $\tau = (1-\varepsilon)\tau_\varepsilon + \varepsilon\tau'_\varepsilon$. Let $\mu_\varepsilon = \int \nu d\tau_\varepsilon(\nu)$ and $\mu'_\varepsilon = \int \nu d\tau'_\varepsilon(\nu)$ so that $\mu = (1-\varepsilon)\mu_\varepsilon + \varepsilon\mu'_\varepsilon$. Then by the previous proposition

$$h_\mu(T, \mathcal{P}) = (1-\varepsilon)h_{\mu_\varepsilon}(T, \mathcal{P}) + \varepsilon h_{\mu'_\varepsilon}(T, \mathcal{P})$$

Since $h_{\mu'_\varepsilon}(T, \mathcal{P}) \leq \log|\mathcal{P}|$, this gives

$$|h_\mu(T, \mathcal{P}) - h_{\mu_\varepsilon}(T, \mathcal{P})| = O(\varepsilon \log|\mathcal{P}|)$$

Now choose $\tau_{\varepsilon,n} \in \mathcal{P}(\mathcal{W}_n)$ a sequence of measures supported in finite subsets of $\mathcal{W}_\varepsilon$ which converge weak-* to τ_ε (this is always possible). Since $h_{\mathcal{P}}$ is continuous on $\mathcal{W}_\varepsilon$, this implies, using the previous proposition for the measures $\tau_{\varepsilon,n}$,

$$\begin{aligned}
\int h_{\mathcal{P}}(\nu) d\tau_\varepsilon(\nu) &= \lim_{n\to\infty} \int h_{\mathcal{P}}(\nu) d\tau_{\varepsilon,n}(\nu) \\
&= \lim_{n\to\infty} h_{\mathcal{P}}(\int \nu d\tau_{\varepsilon,n}(\nu) \\
&= h_{\mathcal{P}}(\int \nu d\tau_\varepsilon) \\
&= h_{\mathcal{P}}(\mu_\varepsilon) \\
&= h_{\mu_\varepsilon}(T, \mathcal{P})
\end{aligned}$$

where in the middle equality we again used continuity of $h_{\mathcal{P}}$ on $\mathcal{W}_\varepsilon$ and the fact that $\tau_{\varepsilon,n} \to \tau_\varepsilon$ implies $\int \nu d\tau_{\varepsilon,n}(\nu) \to \int \nu d\tau_\varepsilon(\nu)$ (which can be seen by integrating against continuous functions).

Finally, using again the bound on the integrand,

$$\left| \int h_\nu(T, \mathcal{P}) d\tau_\varepsilon(\nu) - \int h_\nu(T, \mathcal{P}) d\tau(\nu) \right| = O(\varepsilon \log|\mathcal{P}|)$$

Putting the last three equations together we have found that

$$\left| h_\mu(T,\mathcal{P}) - \int h_{\mathcal{P}}(\nu) d\tau(\nu) \right| = O(\varepsilon \log \mathcal{P})$$

and since ε was arbitrary this implies

$$h_\mu(T,\mathcal{P}) = \int h_\nu(T,\mathcal{P}) d\tau(\nu)$$

For the last statement take a refining sequence of partitions which together generate the Borel σ-algebra, and take a limit of the equation above, using monotone convergence on the right hand side. □

Corollary 14.2.4. *Let (X,T) be a topological dynamical system. Then*

$$h_{top}(T) = \sup\{h_\mu(T) \,:\, \mu \in \mathcal{P}_T(X) \text{ is ergodic}\}$$

Proof. We certainly have $\geq$. If $<$ held then there is a $u < h_{top}(T)$ such that $h_\mu(T) < u$ for ergodic μ. Then for any invariant measure μ, writing its ergodic decomposition as $\mu = \int \nu_\omega d\tau(\omega)$, we have

$$h_\mu(T) = \int h_{\nu_\omega}(T) d\tau(\omega) < \int u d\tau(\omega) < u < h_{top}(T)$$

contradicting the variational principle. □

Now consider a topological system (X,T). How continuous is the entropy function $h : \mathcal{P}_T(X) \to \mathbb{R}$? In general, it is not continuous.

Example 14.2.5. Let $X = \{0,1\}^{\mathbb{Z}}$. Then every $\mu \in \mathcal{P}_T(X)$ is the weak-* limit of periodic orbits. Let us prove this in the ergodic case. Choose a generic point x for μ (μ-a.e. x will d). Then for every k there is an $N = N_k$ such that

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) - \int f d\mu \right| < \frac{1}{k}$$

for all f that depend on the coordinates $-k, \ldots, k$. Let x_k denote the periodic point obtained by repeating the first N_k symbols of x. Let $\mu_k = \frac{1}{N_k} \sum_{n=0}^{N_k-1} \delta_{T^n x_k}$ denote the uniform measure on the orbit. Then it is clear that $\mu_k \to \mu$.

This example shows that every measure of positive entropy on $\{0,1\}^{\mathbb{Z}}$ is the weak-* limit of zero-entropy measures and so the entropy function is not continuous. In general not much more can be said, but for expansive systems (e.g. shift spaces on finite alphabets) there is some continuity nonetheless.

Proposition 14.2.6. *Let (X,T) be expansive. Then $h : \mathcal{P}_T(X) \to \mathbb{R}$ is upper semi-continuous, i.e. if $\mu_n \to \mu$ then $h(\mu) \geq \limsup h(\mu_n)$.*

Proof. Let $\mu_n \to \mu$. Let ε be the expansiveness constant and choose a finite partition $\mathcal{P}$ of X into sets of diameter $< \varepsilon$ such that $\mu(\partial A) = 0$ for $A \in \mathcal{P}$. Then for any k,

$$\begin{aligned} \frac{1}{k} H_\mu(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) &= \lim_{n\to\infty} \frac{1}{k} H_{\mu_n}(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) \\ &\geq \limsup_{n\to\infty} h_{\mu_n}(\mathcal{P}) \end{aligned}$$

because $\frac{1}{k} H_{\mu_n}(\bigvee_{i=0}^{k-1} T^{-i}\mathcal{P}) \searrow h_{\mu_n}(T)$. as $k \to \infty$. Taking $k \to \infty$ in the last inequality proves the claim. □

Corollary 14.2.7. *If (X, T) is expansive then there is an ergodic $\mu \in \mathcal{P}_T(X)$ with $h_\mu(T) = h_{top}(T)$.*

Proof. Let $\mu_n \in \mathcal{P}_T(X)$ be such that $h_{\mu_n}(T) \to h_{top}(T)$, as guaranteed by the variational principle. Let μ be a weak-* accumulation point of μ_n. Then $h_\mu(T) \geq \lim h_{\mu_n}(T) = h_{top}(T)$. Thus μ must have an ergodic component satisfying the same inequality, and the reverse inequality is automatic. □

Example 14.2.8. Let $X_n \subseteq \{0,1\}^{\mathbb{Z}}$ be subsystems (with respect to the shift σ) such that $h_{top}(X_n) \nearrow 1$. Define $X_\infty = \{.\}$ with the identity map, also denoted σ. Formally take X_n to be disjoint, fix a metric d_0 on $\{0,1\}^{\mathbb{Z}}$, and define a metric d on $X = \bigcup X_n \cup X_\infty$ by

$$d(x,y) = \begin{cases} \frac{1}{n} d_0(x,y) & x, y \in X_n \text{ for some } n \\ \frac{1}{n} - \frac{1}{m} & x, \in X_n \,,\, y \in X_m \; m \neq n \end{cases}$$

This makes X compact. Define $\sigma : X \to X$ by $\sigma|_{X_n} = \sigma_n$. This is a continuous map. Note that the X_n are invariant sets so every ergodic measure on X is an ergodic measure on one of the X_n, hence by the variational principle, all invariant measures on X have entropy < 1 (since this is true for the X_n's). On the other hand by the variational principle

$$h_{top}(X,\sigma) = \sup_{\mu\in\mathcal{P}_T(X)} h_\mu(\sigma) = \sup_n h_{top}(X_n,\sigma) = 1$$

Thus there is no measure on (X,σ) whose entropy realizes the topological entropy.

(We observe that σ is not expansive: X_n is an invariant set of diameter $\frac{1}{n} \operatorname{diam} d_0$).

Chapter 15

Appendix

15.1 The weak-* topology

Proposition 15.1.1. *Let X be a compact metric space. Then $\mathcal{P}(X)$ is metrizable and compact in the weak-* topology.*

Proof. Let $\{f_i\}_{i=1}^{\infty}$ be a countable dense subset of the unit ball in $C(X)$. Define a metric on $\mathcal{P}(X)$ by

$$d(\mu,\nu) = \sum_{i=1}^{\infty} 2^{-i} \left| \int f_i d\mu - \int f_i d\nu \right|$$

It is easy to check that this is a metric. We must show that the topology induced by this metric is the weak-* topology.

If $\mu_n \to \mu$ weak-* then $\int f_i d\mu_n - \int f_i d\mu \to 0$ as $n \to \infty$, hence $d(\mu_n, \mu) \to 0$.

Conversely, if $d(\mu_n, \mu) \to 0$, then $\int f_i d\mu_n \to \int f_i d\mu$ for every i and therefore for every linear combination of the f_is. Given $f \in C(X)$ and $\varepsilon > 0$ there is a linear combination g of the f_i such that $\|f - g\|_\infty < \varepsilon$. Then

$$\begin{aligned} \left| \int f d\mu_n - \int f d\mu \right| &< \left| \int f d\mu_n - \int g d\mu_n \right| + \left| \int g d\mu_n - \int g d\mu \right| + \left| \int g d\mu - \int f d\mu \right| \\ &< \varepsilon + \left| \int g d\mu_n - \int g d\mu \right| + \varepsilon \end{aligned}$$

and the right hand side is $< 3\varepsilon$ when n is large enough. Hence $\mu_n \to \mu$ weak-*.

Since the space is metrizable, to prove compactness it is enough to prove sequential compactness, i.e. that every sequence $\mu_n \in \mathcal{P}(X)$ has a convergent subsequence. Let $V = \operatorname{span}_{\mathbb{Q}}\{f_i\}$, which is a countable dense $\mathbb{Q}$-linear subspace of $C(X)$. The range of each $g \in V$ is a compact subset of $\mathbb{R}$ (since X is compact and g continuous) so for each $g \in V$ we can choose a convergent subsequence of $\int g d\mu_n$. Using a diagonal argument we may select a single subsequence $\mu_{n(j)}$ such that $\int g \mu_{n(j)} \to \Lambda(g)$ as $j \to \infty$ for every $g \in V$. Now, Λ is a $\mathbb{Q}$-linear

functional because

$$\begin{aligned}\Lambda(af_i + bf_j) &= k \lim \int (af_i + bf_j)\, d\mu_{n(k)} \\ &= \lim_{k\to\infty} a \int f_i d\mu_{n(k)} + b \int f_j d\mu_{n(k)} \\ &= a\Lambda(f_i) + b\Lambda(f_j)\end{aligned}$$

Λ is also uniformly continuous because, if $\|f_i - f_j\|_\infty < \varepsilon$ then

$$\begin{aligned}|\Lambda(f_i - f_j)| &= \left| \lim_{k\to\infty} \int (f_i - f_j)\, d\mu_{n(k)} \right| \\ &\leq \lim_{k\to\infty} \int |f_i - f_j| d\mu_{n(k)} \\ &\leq \varepsilon\end{aligned}$$

Thus Λ extends to a continuous linear functional on $C(X)$. Since Λ is positive (i.e. non-negative on non-negative functions), sos is its extension, so by the Riesz representation theorem there exists $\mu \in \mathcal{P}(X)$ with $\Lambda(f) = \int f d\mu$. By definition $\int g d\mu - \int g d\mu_{n(k)} \to 0$ as $k \to \infty$ for $g \in V$, hence this is true for the f_i, so $d(\mu_{n(k)}, \mu) \to 0$ Hence $\mu_{n(k)} \to \mu$ weak-* . □

15.2 Conditional expectation

When (X, B, μ) is a probability space, $f \in L^1$, and A a set of positive measure, then the conditional expectation of f on A is usually defined as $\frac{1}{\mu(A)} \int_A f\, d\mu$. When A has measure 0 this formula is meaningless, and it is not clear how to give an alternative definition. But if $\mathcal{A} = \{A_i\}_{i\in I}$ is a partition of X into measurable sets (possibly of measure 0), one can sometimes give a meaningful definition of the conditional expectation of f on $\mathcal{A}(x)$ for a.e. x, where $\mathcal{A}(x)$ is the element A_i containing x. Thus the conditional expectation of$f$ on $\mathcal{A}$ is a function that assigns to a.e. x the conditional expectation of f on the set $\mathcal{A}(x)$. Rather than partitions, we will work with σ-algebras; the connection is made by observing that if $\mathcal{E}$ is a countably-generated σ-algebra then the partition of X into the atoms of $\mathcal{E}$ is a measurable partition.

Theorem 15.2.1. *Let $(X, \mathcal{B}, \mu)$ be a probability space and $\mathcal{E} \subseteq \mathcal{B}$ a sub-σ algebra. Then there is a linear operator $L^1(X, \mathcal{B}, \mu) \to L^1(X, \mathcal{E}, \mu)$ satisfying*

1. *Chain rule: $\int \mathbb{E}(f|\mathcal{E})\, d\mu = \int f\, d\mu$.*
2. *Product rule: $\mathbb{E}(gf|\mathcal{E}) = g \cdot \mathbb{E}(f|\mathcal{E})$ for all $g \in L^\infty(X, \mathcal{E}, \mu)$.*

Proof. We begin with existence. Let $f \in L^1(X, \mathcal{B}, \mu)$ and let μ_f be the finite signed measure $d\mu_f = f d\mu$. Then $\mu_f \ll \mu$ in the measure space $(X, \mathcal{B}, \mu)$ and this remains true in $(X, \mathcal{E}, \mu)$. Let $\mathbb{E}(f|\mathcal{E}) = d\mu_f/d\mu \in L^1(X, \mathcal{E}, \mu)$, the Radon-Nykodim derivative of μ_f with respect to μ in $(X, \mathcal{E}, \mu)$.

The domain of this map is $L^1(X, \mathcal{B}, \mu)$ and its range is in $L^1(X, \mathcal{E}, \mu)$ by the properties of $d\mu_f/d\mu$.

Linearity follows from uniqueness of the Radon-Nykodim derivative and the definitions. The chain rule is also immediate:

$$\int \mathbb{E}(f|\mathcal{E})\, d\mu = \int \frac{d\mu_f}{d\mu}\, d\mu = \int f\, d\mu$$

For the product rule, let $g \in L^\infty(X, \mathcal{E}, \mu)$. We must show that $g \cdot \frac{d\mu_f}{d\mu} = \frac{d\mu_{gf}}{d\mu}$ in $(X, \mathcal{E}, \mu)$. Equivalently we must show that

$$\int_E g\frac{d\mu_f}{d\mu}\, d\mu = \int_E \frac{d\mu_{gf}}{d\mu}\, d\mu \qquad \text{for all } E \in \mathcal{E}$$

Now, for $A \in \mathcal{E}$ and $g = 1_A$ we have

$$\begin{aligned}
\int_E 1_A \frac{d\mu_f}{d\mu}\, d\mu &= \int_{A\cap E} \frac{d\mu_f}{d\mu}\, d\mu \\
&= \mu_f(A \cap E) \\
&= \int_{A\cap E} f\, d\mu \\
&= \int_E 1_A f\, d\mu \\
&= \int_E \frac{d\mu_{1_A f}}{d\mu}\, d\mu
\end{aligned}$$

so the identity holds. By linearity of these integrals in the g argument it holds linear combinations of indicator functions. For arbitrary $g \in L^\infty$ we can take a uniformly bounded sequence of such functions converging pointwise to g, and pass to the limit using dominated convergence. This proves the product rule.

To prove uniqueness, let $T : L^1(X, \mathcal{B}, \mu) \to L^1(X, \mathcal{E}, \mu)$ be an operator with these properties. Then for $f \in L^1(X, \mathcal{B}, \mu)$ and $E \in \mathcal{E}$,

$$\begin{aligned}
\int_E Tf\, d\mu &= \int 1_E Tf\, d\mu \\
&= \int T(1_E f)\, d\mu \\
&= \int 1_E f\, d\mu \\
&= \int_E f\, d\mu
\end{aligned}$$

where the second equality uses the product rule and the third uses the chain rule. Since this holds for all $E \in \mathcal{E}$ we must have $Tf = d\mu_f/d\mu$. □

Proposition 15.2.2. *The conditional expectation operator satisfies the following properties:*

1. *Positivity:* $f \geq 0$ *a.e. implies* $\mathbb{E}(f|\mathcal{E}) \geq 0$ *a.e.*
2. *Triangle inequality:* $|\mathbb{E}(f|\mathcal{I})| \leq \mathbb{E}(|f|\,|\mathcal{I})$.
3. *Contraction:* $\|\mathbb{E}(f|\mathcal{E})\|_1 \leq \|f\|_1$*; in particular,* $\mathbb{E}(\cdot|\mathcal{E})$ *is* L^1*-continuous.*
4. *Sup/inf property:* $\mathbb{E}(\sup f_i|\mathcal{E}) \geq \sup \mathbb{E}(f_i|\mathcal{E})$ *and* $\mathbb{E}(\inf f_i|\mathcal{E}) \leq \inf \mathbb{E}(f_i|\mathcal{E})$ *for any countable family* $\{f_i\}$.
5. *Jensen's inequality: if* g *is convex then* $g(\mathbb{E}(f|\mathcal{E})) \leq \mathbb{E}(g \circ f|\mathbb{E})$.
6. *Fatou's lemma:* $\mathbb{E}(\liminf f_n|\mathcal{E}) \leq \liminf \mathbb{E}(f_n|\mathcal{E})$.

Remark 15.2.3. Properties (2)–(6) are consequences of positivity only.

Proof. (1) Suppose $f \geq 0$ and $\mathbb{E}(f|\mathcal{E}) \not> 0$, so $\mathbb{E}(f|\mathcal{E}) < 0$ on a set $A \in \mathcal{E}$ of positive measure. Applying the product rule with $g = 1_A$, we have

$$\mathbb{E}(1_A f|\mathcal{E}) = 1_A \mathbb{E}(f|\mathcal{E})$$

hence, replacing f by 1_A, we can assume that $f \geq 0$ and $\mathbb{E}(f|\mathcal{E}) < 0$. But this contradicts the chain rule since $\int f\,d\mu \geq 0$ and $\int \mathbb{E}(f|\mathcal{E})\,d\mu < 0$.

(2) Decompose f into positive and negative parts, $f = f^+ - f^-$, so that $|f| = f^+ + f^-$. By positivity,

$$\begin{aligned} |\mathbb{E}(f|\mathcal{E})| &= |\mathbb{E}(f^+|\mathcal{E}) - \mathbb{E}(f^-|\mathcal{E})| \\ &\leq |\mathbb{E}(f^+|\mathcal{E})| + |\mathbb{E}(f^-|\mathcal{E})| \\ &= \mathbb{E}(f^+|\mathcal{E}) + \mathbb{E}(f^-|\mathcal{E}) \\ &= \mathbb{E}(f^+ + f^-|\mathcal{E}) \\ &= \mathbb{E}(|f|\,|\mathcal{E}) \end{aligned}$$

(3) We compute:

$$\begin{aligned} \|\mathbb{E}(f|\mathcal{E})\|_1 &= \int |\mathbb{E}(f|\mathcal{E})|\,d\mu \\ &\leq \int \mathbb{E}(|f|\,|\mathcal{E})|\,d\mu \\ &= \int |f|\,d\mu \\ &= \|f\|_1 \end{aligned}$$

where we have used the triangle inequality and the chain rule.

(4) We prove the sup version. By monotonicity and continuity it suffices to prove this for finite families and hence for two functions. The claim now follows from the identity $\max\{f_1, f_2\} = \frac{1}{2}(f_1 + f_2 + |f_1 - f_2|)$, linearity, and the triangle inequality.

(5) For an affine function $g(t) = at + b$,

$$\mathbb{E}(g \circ f|\mathcal{E}) = \mathbb{E}(af + b|\mathcal{E}) = a\mathbb{E}(f|\mathcal{E}) + b = g \circ \mathbb{E}(f|\mathcal{E})$$

If g is convex then $g = \sup g_i$ where $\{g_i\}_{i\in I}$ is a countable family of affine functions. Thus

$$\begin{aligned}\mathbb{E}(g\circ f|\mathcal{E}) &= \mathbb{E}(\sup_i g_i\circ f|\mathcal{E})\\ &\geq \sup_i \mathbb{E}(g_i\circ f|\mathcal{E})\\ &= \sup_i g_i\circ \mathbb{E}(f|\mathcal{E})\\ &= g\circ\mathbb{E}(f|\mathcal{E})\end{aligned}$$

(6) Since $\inf_{k>n} f_k \nearrow \liminf f_k$ as $n\to\infty$ the convergence is also in L^1, so by continuity and positivity the same holds after taking the conditional expectation. Thus, using the inf property,

$$\begin{aligned}\liminf_{n\to\infty}\mathbb{E}(f_n|\mathcal{E}) &= \lim_{n\to\infty}\inf_{k>n}\mathbb{E}(f_k|\mathcal{E})\\ &\geq \lim_{n\to\infty}\mathbb{E}(\inf_{k>n} f_k|\mathcal{E})\\ &= \mathbb{E}(\liminf_{n\to\infty} f_n|\mathcal{E})\ \square\end{aligned}$$

Corollary 15.2.4. *The restriction of the conditional expectation operator to $L^2(X,\mathcal{B},\mu)$ coincides with the orthogonal projection $\pi : L^2(X,\mathcal{B},\mu)\to L^2(X,\mathcal{E},\mu)$.*

Proof. Write $\pi = \mathbb{E}(\cdot|\mathcal{E})$. If $f\in L^2$ then by by convexity of $t\to t^2$ and Jensen's inequality (which is immediate for simple functions and hence holds for $f\in L^1$ by approximation),

$$\begin{aligned}\|\pi f\|_2 &= \int |\mathbb{E}(f|\mathcal{E})|^2\,d\mu\\ &\leq \int \mathbb{E}(|f|^2|\mathcal{E})\,d\mu\\ &= \int |f|^2\,d\mu \qquad\qquad \text{by the chain rule}\\ &= \|f\|_2\end{aligned}$$

Thus π maps L^2 into the subspace of $\mathcal{E}$-measurable L^2 functions, hence $\pi : L^2(X,\mathcal{B},m)\to L^2(X,\mathcal{E},\mu)$. We will now show that π is the identity on $L^2(X,\mathcal{E},\mu)$ and is π. Indeed, if $g\in L^2(X,\mathbb{E},\mu)$ then for every $A\in\mathcal{E}$

$$\begin{aligned}\pi g &= \mathbb{E}(g\cdot 1|\mathcal{E})\\ &= g\cdot\mathbb{E}(1|\mathcal{E})\end{aligned}$$

Since $\int\mathbb{E}(1|\mathcal{E}) = \int 1 = 1$, this shows that π is the identity on $L^2(X,\mathcal{E},)$. Next

if $f, g \in L^2$ then $fg \in L^1$, and

$$\begin{aligned}
\langle f, \pi g \rangle &= \int f \cdot \mathbb{E}(g|\mathcal{E})\, d\mu \\
&= \int \mathbb{E}\left(f \cdot \mathbb{E}(g|\mathcal{E})\right)\, d\mu && \text{by the chain rule} \\
&= \int \mathbb{E}\left(f|\mathcal{E}\right)\mathbb{E}(g|\mathcal{E})\, d\mu && \text{by the product rule} \\
&= \int \mathbb{E}\left(\mathbb{E}(f|\mathcal{E}) \cdot g\right)\, d\mu && \text{by the product rule} \\
&= \int \mathbb{E}(f|\mathcal{E}) \cdot g\, d\mu && \text{by the chain rule} \\
&= \langle \pi f, g \rangle
\end{aligned}$$

so π is self-adjoint. □

Example 15.2.5. Let $(X, \mathcal{B}, \mu)$ be a probability space, X_1, X_2 a partition of X, and $\mathcal{I} = \{\emptyset, X_1, X_2, X\}$, a σ-algebra. Then for any $f \in L^1(\mu, \mathcal{I})$,

$$\begin{aligned}
\mathbb{E}(f|\mathcal{I})(x) &= \frac{\int_{X_1} f d\mu}{\mu(X_1)} \cdot 1_{X_1}(x) + \frac{\int_{X_2} f d\mu}{\mu(X_2)} \cdot 1_{X_2}(x) \\
&= \begin{cases} \frac{\int_{X_1} f d\mu}{\mu(X_1)} & x \in X_1 \\ \frac{\int_{X_2} f d\mu}{\mu(X_2)} & x \in X_2 \end{cases}
\end{aligned}$$

To see this, note that since $\mathbb{E}(f|\mathcal{I})$ is $\mathcal{I}$-measurable, it has the form $a1_{X_1} + b1_{X_2}$. Thus

$$\begin{aligned}
a1_{X_1} &= 1_{X_1}(a1_{X_1} + b1_{X_2}) \\
&= 1_{X_1}\mathbb{E}(f|\mathcal{I}) \\
&= \mathbb{E}(1_{X_1} f|\mathcal{I})
\end{aligned}$$

Integrating we have

$$a\mu(X_1) = \int \mathbb{E}(1_{X_1} f|\mathcal{I}) d\mu = \int 1_{X_1} f d\mu = \int_{X_1} f d\mu$$

this shows that $a = \int f d\mu / \mu(X_1)$, and b is computed similarly.

15.3 Regularity

I'm not sure we use this anywhere, but for the record:

Lemma 15.3.1. *A Borel probability measure on a complete (separable) metric space is regular.*

Proof. It is easy to see that the family of sets A with the property that

$$\begin{aligned} \mu(A) &= \inf\{\mu(U) : U \supseteq A \text{ is open}\} \\ &= \sup\{\mu(C) : C \subseteq A \text{ is closed}\} \end{aligned}$$

contains all open and closed sets, and is a σ-algebra. Therefore every Borel set A has this property. We need to verify that in the second condition we can replace closed by compact. Clearly it is enough to show that for every closed set C and every $\varepsilon > 0$ there is a compact $K \subseteq C$ with $\mu(K >> \mu(C) - \varepsilon$.

Fix C and $\varepsilon > 0$. For every n we can find a finite family $B_{n,1}, \ldots, B_{n,k(n)}$ of δ-balls whose union $B_n = \bigcup B_{n,i}$ intersects A in a set of measure $> \mu(A) - \varepsilon/2^n$. Let $K_0 = C \cap \bigcap B_n$, so that $\mu(K_0) > \mu(C) - \varepsilon$. By construction K_0 is pre-compact, and $K = \overline{K_0} \subseteq C$, so K has the desired property. □