A Simple Introduction to Ergodic Theory

Karma Dajani and Sjoerd Dirksin

December 18, 2008

Contents

Chapter 1

Introduction and preliminaries

1.1 What is Ergodic Theory?

It is not easy to give a simple definition of Ergodic Theory because it uses techniques and examples from many fields such as probability theory, statistical mechanics, number theory, vector fields on manifolds, group actions of homogeneous spaces and many more.

The word *ergodic* is a mixture of two Greek words: *ergon* (work) and *odos* (path). The word was introduced by Boltzmann (in statistical mechanics) regarding his hypothesis: *for large systems of interacting particles in equilibrium, the time average along a single trajectory equals the space average.* The hypothesis as it was stated was false, and the investigation for the conditions under which these two quantities are equal lead to the birth of ergodic theory as is known nowadays.

A modern description of what ergodic theory is would be: it is the study of the long term average behavior of systems evolving in time. The collection of all states of the system form a space X, and the evolution is represented by either

– a transformation $T : X \to X$, where Tx is the state of the system at time $t = 1$, when the system (i.e., at time $t = 0$) was initially in state x. (This is analogous to the setup of discrete time stochastic processes).

– if the evolution is continuous or if the configurations have spacial structure, then we describe the evolution by looking at a group of transformations G (like $\mathbb{Z}^2$, $\mathbb{R}$, $\mathbb{R}^2$) acting on X, i.e., every $g \in G$ is identified with a transformation $T_g : X \to X$, and $T_{gg'} = T_g \circ T_{g'}$.

The space X usually has a special structure, and we want T to preserve the basic structure on X. For example
–if X is a measure space, then T must be measurable.
–if X is a topological space, then T must be continuous.
–if X has a differentiable structure, then T is a diffeomorphism.
In this course our space is a probability space $(X, \mathcal{B}, \mu)$, and our time is discrete. So the evolution is described by a measurable map $T : X \to X$, so that $T^{-1}A \in \mathcal{B}$ for all $A \in \mathcal{B}$. For each $x \in X$, the orbit of x is the sequence

$$x, Tx, T^2x, \ldots .$$

If T is invertible, then one speaks of the two sided orbit

$$\ldots, T^{-1}x, x, Tx, \ldots .$$

We want also that the evolution is in steady state i.e. stationary. In the language of ergodic theory, we want T to be *measure preserving.*

1.2 Measure Preserving Transformations

Definition 1.2.1 *Let $(X, \mathcal{B}, \mu)$ be a probability space, and $T : X \to X$ measurable. The map T is said to be measure preserving with respect to μ if $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{B}$.*

This definition implies that for any measurable function $f : X \to \mathbb{R}$, the process

$$f, f \circ T, f \circ T^2, \ldots$$

is stationary. This means that for all Borel sets $B_1, \ldots, B_n$, and all integers $r_1 < r_2 < \ldots < r_n$, one has for any $k \geq 1$,

$$\mu\left(\{x : f(T^{r_1}x) \in B_1, \ldots f(T^{r_n}x) \in B_n\}\right) =$$
$$\mu\left(\{x : f(T^{r_1+k}x) \in B_1, \ldots f(T^{r_n+k}x) \in B_n\}\right).$$

In case T is invertible, then T is measure preserving if and only if $\mu(TA) = \mu(A)$ for all $A \in \mathcal{B}$. We can generalize the definition of measure preserving to the following case. Let $T : (X_1, \mathcal{B}_1, \mu_1) \to (X_2, \mathcal{B}_2, \mu_2)$ be measurable, then T is measure preserving if $\mu_1(T^{-1}A) = \mu_2(A)$ for all $A \in \mathcal{B}_2$. The following gives a useful tool for verifying that a transformation is measure preserving. For this we need the notions of algebra and semi-algebra.

Recall that a collection $\mathcal{S}$ of subsets of X is said to be a *semi-algebra* if (i) $\emptyset \in \mathcal{S}$, (ii) $A \cap B \in \mathcal{S}$ whenever $A, B \in \mathcal{S}$, and (iii) if $A \in \mathcal{S}$, then $X \setminus A = \cup_{i=1}^{n} E_i$ is a disjoint union of elements of $\mathcal{S}$. For example if $X = [0,1)$, and $\mathcal{S}$ is the collection of all subintervals, then $\mathcal{S}$ is a semi-algebra. Or if $X = \{0,1\}^{\mathbb{Z}}$, then the collection of all cylinder sets $\{x : x_i = a_i, \ldots, x_j = a_j\}$ is a semi-algebra.

An *algebra* $\mathcal{A}$ is a collection of subsets of X satisfying:(i) $\emptyset \in \mathcal{A}$, (ii) if $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$, and finally (iii) if $A \in \mathcal{A}$, then $X \setminus A \in \mathcal{A}$. Clearly an algebra is a semi-algebra. Furthermore, given a semi-algebra $\mathcal{S}$ one can form an algebra by taking all finite disjoint unions of elements of $\mathcal{S}$. We denote this algebra by $\mathcal{A}(\mathcal{S})$, and we call it the *algebra generated* by $\mathcal{S}$. It is in fact the smallest algebra containing $\mathcal{S}$. Likewise, given a semi-algebra $\mathcal{S}$ (or an algebra $\mathcal{A}$), the σ-algebra generated by $\mathcal{S}$ ($\mathcal{A}$) is denoted by $\mathcal{B}(\mathcal{S})$ ($\mathcal{B}(\mathcal{A})$), and is the smallest σ-algebra containing $\mathcal{S}$ (or $\mathcal{A}$).

A *monotone class* $\mathcal{C}$ is a collection of subsets of X with the following two properties

- if $E_1 \subseteq E_2 \subseteq \ldots$ are elements of $\mathcal{C}$, then $\cup_{i=1}^{\infty} E_i \in \mathcal{C}$,
- if $F_1 \supseteq F_2 \supseteq \ldots$ are elements of $\mathcal{C}$, then $\cap_{i=1}^{\infty} F_i \in \mathcal{C}$.

The *monotone class generated* by a collection $\mathcal{S}$ of subsets of X is the smallest monotone class containing $\mathcal{S}$.

Theorem 1.2.1 *Let $\mathcal{A}$ be an algebra of X, then the σ-algebra $\mathcal{B}(\mathcal{A})$ generated by $\mathcal{A}$ equals the monotone class generated by $\mathcal{A}$.*

Using the above Theorem, one can get an easier criterion for checking that a transformation is measure preserving.

Theorem 1.2.2 *Let $(X_i, \mathcal{B}_i, \mu_i)$ be probability spaces, $i = 1, 2$, and $T : X_1 \to X_2$ a transformation. Suppose $\mathcal{S}_2$ is a generating semi-algebra of $\mathcal{B}_2$. Then, T is measurable and measure preserving if and only if for each $A \in \mathcal{S}_2$, we have $T^{-1}A \in \mathcal{B}_1$ and $\mu_1(T^{-1}A) = \mu_2(A)$.*

Proof. Let

$$\mathcal{C} = \{B \in \mathcal{B}_2 : T^{-1}B \in \mathcal{B}_1, \text{ and } \mu_1(T^{-1}B) = \mu_2(B)\}.$$

Then, $\mathcal{S}_2 \subseteq \mathcal{C} \subseteq \mathcal{B}_2$, and hence $\mathcal{A}(\mathcal{S}_2) \subset \mathcal{C}$. We show that $\mathcal{C}$ is a monotone class. Let $E_1 \subseteq E_2 \subseteq \ldots$ be elements of $\mathcal{C}$, and let $E = \cup_{i=1}^{\infty} E_i$. Then, $T^{-1}E = \cup_{i=1}^{\infty} T^{-1}E_i \in \mathcal{B}_1$.

$$\begin{aligned} \mu_1(T^{-1}E) = \mu_1(\cup_{n=1}^{\infty} T^{-1}E_n) &= \lim_{n\to\infty} \mu_1(T^{-1}E_n) \\ &= \lim_{n\to\infty} \mu_2(E_n) \\ &= \mu_2(\cup_{n=1}^{\infty} E_n) \\ &= \mu_2(E). \end{aligned}$$

Thus, $E \in \mathcal{C}$. A similar proof shows that if $F_1 \supseteq F_2 \supseteq \ldots$ are elements of $\mathcal{C}$, then $\cap_{i=1}^{\infty} F_i \in \mathcal{C}$. Hence, $\mathcal{C}$ is a monotone class containing the algebra $\mathcal{A}(\mathcal{S}_2)$. By the monotone class theorem, $\mathcal{B}_2$ is the smallest monotone class containing $\mathcal{A}(\mathcal{S}_2)$, hence $\mathcal{B}_2 \subseteq \mathcal{C}$. This shows that $\mathcal{B}_2 = \mathcal{C}$, therefore T is measurable and measure preserving. □

For example if
– $X = [0, 1)$ with the Borel σ-algebra $\mathcal{B}$, and μ a probability measure on $\mathcal{B}$. Then a transformation $T : X \to X$ is measurable and measure preserving if and only if $T^{-1}[a, b) \in \mathcal{B}$ and $\mu\left(T^{-1}[a, b)\right) = \mu\left([a, b)\right)$ for any interval $[a, b)$.
– $X = \{0, 1\}^{\mathbb{N}}$ with product σ-algebra and product measure μ. A transformation $T : X \to X$ is measurable and measure preserving if and only if

$$T^{-1}\left(\{x : x_0 = a_0, \ldots, x_n = a_n\}\right) \in \mathcal{B},$$

and

$$\mu\left(T^{-1}\{x : x_0 = a_0, \ldots, x_n = a_n\}\right) = \mu\left(\{x : x_0 = a_0, \ldots, x_n = a_n\}\right)$$

for any cylinder set.

Exercise 1.2.1 Recall that if A and B are measurable sets, then

$$A\Delta B = (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A).$$

Show that for any measurable sets A, B, C one has

$$\mu(A\Delta B) \leq \mu(A\Delta C) + \mu(C\Delta B).$$

Another useful lemma is the following (see also ([KT]).

Lemma 1.2.1 *Let $(X, \mathcal{B}, \mu)$ be a probability space, and $\mathcal{A}$ an algebra generating $\mathcal{B}$. Then, for any $A \in \mathcal{B}$ and any $\epsilon > 0$, there exists $C \in \mathcal{A}$ such that $\mu(A\Delta C) < \epsilon$.*

Proof. Let

$$\mathcal{D} = \{A \in \mathcal{B} : \text{ for any } \epsilon > 0, \text{ there exists } C \in \mathcal{A} \text{ such that } \mu(A\Delta C) < \epsilon\}.$$

Clearly, $\mathcal{A} \subseteq \mathcal{D} \subseteq \mathcal{B}$. By the Monotone Class Theorem (Theorem (1.2.1)), we need to show that $\mathcal{D}$ is a monotone class. To this end, let $A_1 \subseteq A_2 \subseteq \cdots$ be a sequence in $\mathcal{D}$, and let $A = \bigcup_{n=1}^{\infty} A_n$, notice that $\mu(A) = \lim_{n\to\infty} \mu(A_n)$. Let $\epsilon > 0$, there exists an N such that $\mu(A\Delta A_N) = |\mu(A) - \mu(A_N)| < \epsilon/2$. Since $A_N \in \mathcal{D}$, then there exists $C \in \mathcal{A}$ such that $\mu(A_N\Delta C) < \epsilon/2$. Then,

$$\mu(A\Delta C) \leq \mu(A\Delta A_N) + \mu(A_N\Delta C) < \epsilon.$$

Hence, $A \in \mathcal{D}$. Similarly, one can show that $\mathcal{D}$ is closed under decreasing intersections so that $\mathcal{D}$ is a monotone class containg $\mathcal{A}$, hence by the Monotone class Theorem $\mathcal{B} \subseteq \mathcal{D}$. Therefore, $\mathcal{B} = \mathcal{D}$, and the theorem is proved.

□

1.3 Basic Examples

(a) *Translations* – Let $X = [0, 1)$ with the Lebesgue σ-algebra $\mathcal{B}$, and Lebesgue measure λ. Let $0 < \theta < 1$, define $T : X \to X$ by

$$Tx = x + \theta \bmod 1 = x + \theta - \lfloor x + \theta \rfloor.$$

Then, by considering intervals it is easy to see that T is measurable and measure preserving.

(b) *Multiplication by 2 modulo 1* – Let $(X, \mathcal{B}, \lambda)$ be as in example (a), and let $T : X \to X$ be given by

$$Tx = 2x \bmod 1 = \begin{cases} 2x & 0 \leq x < 1/2 \\ 2x - 1 & 1/2 \leq x < 1. \end{cases}$$

For any interval $[a, b)$,

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

and

$$\lambda\left(T^{-1}[a,b)\right) = b-a = \lambda\left([a,b)\right).$$

Although this map is very simple, it has in fact many facets. For example, iterations of this map yield the binary expansion of points in $[0,1)$ i.e., using T one can associate with each point in $[0,1)$ an infinite sequence of 0's and 1's. To do so, we define the function a_1 by

$$a_1(x) = \begin{cases} 0 & \text{if } 0 \le x < 1/2 \\ 1 & \text{if } 1/2 \le x < 1, \end{cases}$$

then $Tx = 2x - a_1(x)$. Now, for $n \ge 1$ set $a_n(x) = a_1(T^{n-1}x)$. Fix $x \in X$, for simplicity, we write a_n instead of $a_n(x)$, then $Tx = 2x - a_1$. Rewriting we get $x = \frac{a_1}{2} + \frac{Tx}{2}$. Similarly, $Tx = \frac{a_2}{2} + \frac{T^2x}{2}$. Continuing in this manner, we see that for each $n \ge 1$,

$$x = \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} + \frac{T^n x}{2^n}.$$

Since $0 < T^n x < 1$, we get

$$x - \sum_{i=1}^{n} \frac{a_i}{2^i} = \frac{T^n x}{2^n} \to 0 \text{ as } n \to \infty.$$

Thus, $x = \sum_{i=1}^{\infty} \frac{a_i}{2^i}$. We shall later see that the sequence of digits $a_1, a_2, \ldots$ forms an i.i.d. sequence of Bernoulli random variables.

(c) *Baker's Transformation* – This example is the two-dimensional version of example (b). The underlying probability space is $[0,1)^2$ with product Lebesgue σ-algebra $\mathcal{B} \times \mathcal{B}$ and product Lebesgue measure $\lambda \times \lambda$. Define $T : [0,1)^2 \to [0,1)^2$ by

$$T(x,y) = \begin{cases} (2x, \frac{y}{2}) & 0 \le x < 1/2 \\ (2x-1, \frac{y+1}{2}) & 1/2 \le x < 1. \end{cases}$$

Exercise 1.3.1 Verify that T is invertible, measurable and measure preserving.

(d) *β-transformations* – Let $X = [0,1)$ with the Lebesgue σ-algebra $\mathcal{B}$. Let $\beta = \frac{1+\sqrt{5}}{2}$, *the golden mean.* Notice that $\beta^2 = \beta + 1$. Define a transformation

$T : X \to X$ by

$$Tx = \beta x \bmod 1 = \begin{cases} \beta x & 0 \le x < 1/\beta \\ \beta x - 1 & 1/\beta \le x < 1. \end{cases}$$

Then, T is **not** measure preserving with respect to Lebesgue measure (give a counterexample), but is measure preserving with respect to the measure μ given by

$$\mu(B) = \int_B g(x)\,\mathrm{d}x,$$

where

$$g(x) = \begin{cases} \frac{5+3\sqrt{5}}{10} & 0 \le x < 1/\beta \\ \frac{5+\sqrt{5}}{10} & 1/\beta \le x < 1. \end{cases}$$

Exercise 1.3.2 Verify that T is measure preserving with respect to μ, and show that (similar to example (b)) iterations of this map generate expansions for points $x \in [0, 1)$ (known as *β-expansions)* of the form

$$x = \sum_{i=1}^{\infty} \frac{b_i}{\beta^i},$$

where $b_i \in \{0, 1\}$ and $b_i b_{i+1} = 0$ for all $i \ge 1$.

(e) *Bernoulli Shifts* – Let $X = \{0, 1, \ldots k-1\}^{\mathbb{Z}}$ (or $X = \{0, 1, \ldots k-1\}^{\mathbb{N}}$), $\mathcal{F}$ the σ-algebra generated by the cylinders. Let $p = (p_0, p_1, \ldots, p_{k-1})$ be a positive probability vector, define a measure μ on $\mathcal{F}$ by specifying it on the cylinder sets as follows

$$\mu\left(\{x : x_{-n} = a_{-n}, \ldots, x_n = a_n\}\right) = p_{a_{-n}} \ldots p_{a_n}.$$

Let $T : X \to X$ be defined by $Tx = y$, where $y_n = x_{n+1}$. The map T, called the *left shift*, is measurable and measure preserving, since

$$T^{-1}\{x : x_{-n} = a_{-n}, \ldots x_n = a_n\} = \{x : x_{-n+1} = a_{-n}, \ldots, x_{n+1} = a_n\},$$

and

$$\mu\left(\{x : x_{-n+1} = a_{-n}, \ldots, x_{n+1} = a_n\}\right) = p_{a_{-n}} \ldots p_{a_n}.$$

Notice that in case $X = \{0, 1, \ldots k-1\}^{\mathbb{N}}$, then one should consider cylinder sets of the form $\{x : x_0 = a_0, \ldots x_n = a_n\}$. In this case

$$T^{-1}\{x : x_0 = a_0, \ldots, x_n = a_n\} = \cup_{j=0}^{k-1}\{x : x_0 = j, x_1 = a_0, \ldots, x_{n+1} = a_n\},$$

and it is easy to see that T is measurable and measure preserving.

(f) *Markov Shifts* – Let $(X, \mathcal{F}, T)$ be as in example (e). We define a measure ν on $\mathcal{F}$ as follows. Let $P = (p_{ij})$ be a stochastic $k \times k$ matrix, and $q = (q_0, q_1, \ldots, q_{k-1})$ a positive probability vector such that $qP = q$. Define ν on cylinders by

$$\nu\left(\{x : x_{-n} = a_{-n}, \ldots x_n = a_n\}\right) = q_{a_{-n}} p_{a_{-n}a_{-n+1}} \ldots p_{a_{n-1}a_n}.$$

Just as in example (e), one sees that T is measurable and measure preserving.

(g) *Stationary Stochastic Processes*– Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and

$$\ldots, Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, \ldots$$

a stationary stochastic process on Ω with values in $\mathbb{R}$. Hence, for each $k \in \mathbb{Z}$

$$\mathbb{P}\left(Y_{n_1} \in B_1, \ldots, Y_{n_r} \in B_r\right) = \mathbb{P}\left(Y_{n_1+k} \in B_1, \ldots, Y_{n_r+k} \in B_r\right)$$

for any $n_1 < n_2 < \cdots < n_r$ and any Lebesgue sets $B_1, \ldots, B_r$. We want to see this process as coming from a measure preserving transformation.
Let $X = \mathbb{R}^{\mathbb{Z}} = \{x = (\ldots, x_1, x_0, x_1, \ldots) : x_i \in \mathbb{R}\}$ with the product σ-algebra (i.e. generated by the cylinder sets). Let $T : X \to X$ be the left shift i.e. $Tx = z$ where $z_n = x_{n+1}$. Define $\phi : \Omega \to X$ by

$$\phi(\omega) = (\ldots, Y_{-2}(\omega), Y_{-1}(\omega), Y_0(\omega), Y_1(\omega), Y_2(\omega), \ldots).$$

Then, ϕ is measurable since if $B_1, \ldots, B_r$ are Lebesgue sets in $\mathbb{R}$, then

$$\phi^{-1}\left(\{x \in X : x_{n_1} \in B_1, \ldots x_{n_r} \in B_r\}\right) = Y_{n_1}^{-1}(B_1) \cap \ldots \cap Y_{n_r}^{-1}(B_r) \in \mathcal{F}.$$

Define a measure μ on X by

$$\mu(E) = \mathbb{P}\left(\phi^{-1}(E)\right).$$

On cylinder sets μ has the form,

$$\mu\left(\{x \in X : x_{n_1} \in B_1, \ldots x_{n_r} \in B_r\}\right) = \mathbb{P}\left(Y_{n_1} \in B_1, \ldots, Y_{n_r} \in B_r\right).$$

Since

$$T^{-1}\left(\{x : x_{n_1} \in B_1, \dots x_{n_r} \in B_r\}\right) = \{x : x_{n_1+1} \in B_1, \dots x_{n_r+1} \in B_r\},$$

stationarity of the process Y_n implies that T is measure preserving. Furthermore, if we let $\pi_i : X \to \mathbb{R}$ be the natural projection onto the i^{th} coordinate, then $Y_i(\omega) = \pi_i(\phi(\omega)) = \pi_0 \circ T^i(\phi(\omega))$.

(h) *Random Shifts* – Let $(X, \mathcal{B}, \mu)$ be a probability space, and $T : X \to X$ an invertible measure preserving transformation. Then, T^{-1} is measurable and measure preserving with respect to μ. Suppose now that at each moment instead of moving forward by T ($x \to Tx$), we first flip a fair coin to decide whether we will use T or T^{-1}. We can describe this random system by means of a measure preserving transformation in the following way.
Let $\Omega = \{-1, 1\}^{\mathbb{Z}}$ with product σ-algebra $\mathcal{F}$ (i.e. the σ-algebra generated by the cylinder sets), and the uniform product measure $\mathbb{P}$ (see example (e)), and let $\sigma : \Omega \to \Omega$ be the left shift. As in example (e), the map σ is measure preserving. Now, let $Y = \Omega \times X$ with the product σ-algebra, and product measure $\mathbb{P} \times \mu$. Define $S : Y \to Y$ by

$$S(\omega, x) = (\sigma\omega, T^{\omega_0}x).$$

Then S is invertible (why?), and measure preserving with respect to $\mathbb{P} \times \mu$. To see the latter, for any set $C \in \mathcal{F}$, and any $A \in \mathcal{B}$, we have

$$\begin{aligned}
(\mathbb{P} \times \mu)\left(S^{-1}(C \times A)\right) &= (\mathbb{P} \times \mu)\left(\{(\omega, x) : S(\omega, x) \in (C \times A)\right) \\
&= (\mathbb{P} \times \mu)\left(\{(\omega, x) : \omega_0 = 1, \sigma\omega \in C, Tx \in A\right) \\
&+ (\mathbb{P} \times \mu)\left(\{(\omega, x) : \omega_0 = -1, \sigma\omega \in C, T^{-1}x \in A\right) \\
&= (\mathbb{P} \times \mu)\left(\{\omega_0 = 1\} \cap \sigma^{-1}C \times T^{-1}A\right) \\
&+ (\mathbb{P} \times \mu)\left(\{\omega_0 = -1\} \cap \sigma^{-1}C \times TA\right) \\
&= \mathbb{P}\left(\{\omega_0 = 1\} \cap \sigma^{-1}C\right)\mu\left(T^{-1}A\right) \\
&+ \mathbb{P}\left(\{\omega_0 = -1\} \cap \sigma^{-1}C\right)\mu\left(TA\right) \\
&= \mathbb{P}\left(\{\omega_0 = 1\} \cap \sigma^{-1}C\right)\mu(A) \\
&+ \mathbb{P}\left(\{\omega_0 = -1\} \cap \sigma^{-1}C\right)\mu(A) \\
&= \mathbb{P}(\sigma^{-1}C)\mu(A) = \mathbb{P}(C)\mu(A) = (\mathbb{P} \times \mu)(C \times A).
\end{aligned}$$

(h) *continued fractions* – Consider $([0, 1), \mathcal{B})$, where $\mathcal{B}$ is the Lebesgue σ-algebra. Define a transformation $T : [0, 1) \to [0, 1)$ by $T0 = 0$ and for $x \neq 0$

$$Tx = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor.$$

Exercise 1.3.3 *Show that T is* **not** *measure preserving with respect to Lebesgue measure, but is measure preserving with respect to the so called Gauss probability measure μ given by*

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

An interesting feature of this map is that its iterations generate the continued fraction expansion for points in $(0,1)$. For if we define

$$a_1 = a_1(x) = \begin{cases} 1 & \text{if } x \in (\frac{1}{2}, 1) \\ n & \text{if } x \in (\frac{1}{n+1}, \frac{1}{n}], n \geq 2, \end{cases}$$

then, $Tx = \frac{1}{x} - a_1$ and hence $x = \dfrac{1}{a_1 + Tx}$. For $n \geq 1$, let $a_n = a_n(x) = a_1(T^{n-1}x)$. Then, after n iterations we see that

$$x = \frac{1}{a_1 + Tx} = \ldots = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n + T^n x}}}.$$

In fact, if $\dfrac{p_n}{q_n} = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n}}}$, then one can show that $\{q_n\}$ are monotonically increasing, and

$$|x - \frac{p_n}{q_n}| < \frac{1}{q_n^2} \to 0 \text{ as } n \to \infty.$$

The last statement implies that

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}}.$$

1.4 Recurrence

Let T be a measure preserving transformation on a probability space $(X, \mathcal{F}, \mu)$, and let $B \in \mathcal{F}$. A point $x \in B$ is said to be *B-recurrent* if there exists $k \geq 1$ such that $T^k x \in B$.

Theorem 1.4.1 (Poincaré Recurrence Theorem) *If $\mu(B) > 0$, then a.e. $x \in B$ is B-recurrent.*

Proof Let F be the subset of B consisting of all elements that are not B-recurrent. Then,

$$F = \{x \in B : T^k x \notin B \text{ for all } k \geq 1\}.$$

We want to show that $\mu(F) = 0$. First notice that $F \cap T^{-k}F = \emptyset$ for all $k \geq 1$, hence $T^{-l}F \cap T^{-m}F = \emptyset$ for all $l \neq m$. Thus, the sets $F, T^{-1}F, \ldots$ are pairwise disjoint, and $\mu(T^{-n}F) = \mu(F)$ for all $n \geq 1$ (T is measure preserving). If $\mu(F) > 0$, then

$$1 = \mu(X) \geq \mu\left(\cup_{k\geq 0} T^{-k}F\right) = \sum_{k\geq 0} \mu(F) = \infty,$$

a contradiction. □

The proof of the above theorem implies that almost every $x \in B$ returns to B infinitely often. In other words, there exist infinitely many integers $n_1 < n_2 < \ldots$ such that $T^{n_i}x \in B$. To see this, let

$$D = \{x \in B : T^k x \in B \text{ for finitely many } k \geq 1\}.$$

Then,

$$D = \{x \in B : T^k x \in F \text{ for some } k \geq 0\} \subseteq \cup_{k=0}^{\infty} T^{-k}F.$$

Thus, $\mu(D) = 0$ since $\mu(F) = 0$ and T is measure preserving.

1.5 Induced and Integral Transformations

1.5.1 Induced Transformations

Let T be a measure preserving transformation on the probability space $(X, \mathcal{F}, \mu)$. Let $A \subset X$ with $\mu(A) > 0$. By Poincaré's Recurrence Theorem almost every $x \in A$ returns to A infinitely often under the action of T.

For $x \in A$, let $n(x) := \inf\{n \geq 1 : T^n x \in A\}$. We call $n(x)$ the *first return time* of x to A.

Exercise 1.5.1 Show that n is measurable with respect to the σ-algebra $\mathcal{F} \cap A$ on A.

By Poincaré Theorem, $n(x)$ is finite a.e. on A. In the sequel we remove from A the set of measure zero on which $n(x) = \infty$, and we denote the new set again by A. Consider the σ-algebra $\mathcal{F} \cap A$ on A, which is the restriction of $\mathcal{F}$ to A. Furthermore, let μ_A be the probability measure on A, defined by

$$\mu_A(B) = \frac{\mu(B)}{\mu(A)}, \quad \text{for } B \in \mathcal{F} \cap A,$$

so that $(A, \mathcal{F} \cap A, \mu_A)$ is a probability space. Finally, define the induced map $T_A : A \to A$ by

$$T_A x = T^{n(x)} x, \quad \text{for } x \in A.$$

From the above we see that T_A is defined on A. What kind of a transformation is T_A?

Exercise 1.5.2 Show that T_A is measurable with respect to the σ-algebra $\mathcal{F} \cap A$.

Proposition 1.5.1 *T_A is measure preserving with respect to μ_A.*

Proof For $k \geq 1$, let

$$A_k = \{x \in A : n(x) = k\}$$

$$B_k = \{x \in X \setminus A : Tx, \ldots, T^{k-1}x \notin A, T^k x \in A\}.$$

Notice that $A = \bigcup_{k=1}^{\infty} A_k$, and

$$T^{-1}A = A_1 \cup B_1 \quad \text{and} \quad T^{-1}B_n = A_{n+1} \cup B_{n+1}. \tag{1.1}$$

Let $C \in \mathcal{F} \cap A$, since T is measure preserving it follows that $\mu(C) = \mu(T^{-1}C)$. To show that $\mu_A(C) = \mu_A(T_A^{-1}C)$, we show that

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

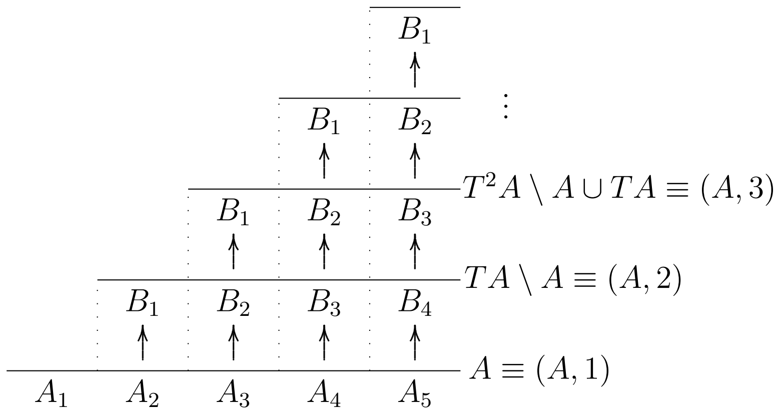

Figure 1.1: A tower.

Now,

$$T_A^{-1}(C) \;=\; \bigcup_{k=1}^{\infty} A_k \cap T_A^{-1}C \;=\; \bigcup_{k=1}^{\infty} A_k \cap T^{-k}C,$$

hence

$$\mu\left(T_A^{-1}(C)\right) \;=\; \sum_{k=1}^{\infty} \mu\left(A_k \cap T^{-k}C\right).$$

On the other hand, using repeatedly (1.1) , one gets for any $n \geq 1$,

$$\begin{aligned}
\mu\left(T^{-1}(C)\right) &= \mu(A_1 \cap T^{-1}C) + \mu(B_1 \cap T^{-1}C) \\
&= \mu(A_1 \cap T^{-1}C) + \mu(T^{-1}(B_1 \cap T^{-1}C)) \\
&= \mu(A_1 \cap T^{-1}C) + \mu(A_2 \cap T^{-2}C) + \mu(B_2 \cap T^{-2}C) \\
&\vdots \\
&= \sum_{k=1}^{n} \mu(A_k \cap T^{-k}C) + \mu(B_n \cap T^{-n}C).
\end{aligned}$$

Since

$$1 \;\geq\; \mu\left(\bigcup_{n=1}^{\infty} B_n \cap T^{-n}C\right) \;=\; \sum_{n=1}^{\infty} \mu(B_n \cap T^{-n}C),$$

it follows that

$$\lim_{n\to\infty} \mu(B_n \cap T^{-n}C) \;=\; 0.$$

Thus,

$$\mu(C) = \mu(T^{-1}C) = \sum_{k=1}^{\infty} \mu\left(A_k \cap T^{-k}C\right) = \mu(T_A^{-1}C).$$

This shows that $\mu_A(C) = \mu_A(T_A^{-1}C)$, which implies that T_A is measure preserving with respect to μ_A. $\square$

Exercise 1.5.3 Assume T is invertible. Without using Proposition 1.5.1 show that for all $C \in \mathcal{F} \cap A$,

$$\mu_A(C) = \mu_A(T_A C).$$

Exercise 1.5.4 Let $G = \dfrac{1+\sqrt{5}}{2}$, so that $G^2 = G + 1$. Consider the set

$$X = [0, \frac{1}{G}) \times [0,1) \bigcup [\frac{1}{G}, 1) \times [0, \frac{1}{G}),$$

endowed with the product Borel σ-algebra, and the normalized Lebesgue measure $\lambda \times \lambda$. Define the transformation

$$\mathcal{T}(x,y) = \begin{cases} (Gx, \dfrac{y}{G}), & (x,y) \in [0, \frac{1}{G}) \times [0,1] \\ \\ (Gx-1, \dfrac{1+y}{G}), & (x,y) \in [\frac{1}{G}, 1) \times [0, \frac{1}{G}). \end{cases}$$

(a) Show that $\mathcal{T}$ is measure preserving with respect to $\lambda \times \lambda$.

(b) Determine explicitely the induced transformation of $\mathcal{T}$ on the set $[0,1) \times [0, \frac{1}{G})$.

1.5.2 Integral Transformations

Let S be a measure preserving transformation on a probability space $(A, \mathcal{F}, \nu)$, and let $f \in L^1(A, \nu)$ be positive and integer valued. We now construct a measure preserving transformation T on a probability space $(X, \mathcal{C}, \mu)$, such that the original transformation S can be seen as the induced transformation on X with return time f.

(1) $X = \{(y,i) : y \in A \text{ and } 1 \leq i \leq f(y),\ i \in \mathbb{N}\}$,

(2) $\mathcal{C}$ is generated by sets of the form

$$(B,i) = \{(y,i) : y \in B \text{ and } f(y) \geq i\} ,$$

where $B \subset A$, $B \in \mathcal{F}$ and $i \in \mathbb{N}$.

(3) $\mu(B,i) = \dfrac{\nu(B)}{\displaystyle\int_A f(y)\,\mathrm{d}\nu(y)}$ and then extend μ to all of X.

(4) Define $T : X \to X$ as follows:

$$T(y,i) = \begin{cases} (y,i+1), & \text{if } i+1 \leq f(y), \\ (Sy,1), & \text{if } i+1 > f(y). \end{cases}$$

Now $(X, \mathcal{C}, \mu, T)$ is called an *integral system* of $(A, \mathcal{F}, \nu, S)$ under f. We now show that T is μ-measure preserving. In fact, it suffices to check this on the generators.

Let $B \subset A$ be $\mathcal{F}$-measurable, and let $i \geq 1$. We have to discern the following two cases:

(1) If $i > 1$, then $T^{-1}(B,i) = (B,i-1)$ and clearly

$$\mu(T^{-1}(B,i)) = \mu(B,i-1) = \mu(B,i) = \frac{\nu(B)}{\int_A f(y)\,\mathrm{d}\nu(y)} .$$

(2) If $i = 1$, we write $A_n = \{y \in A : f(y) = n\}$, and we have

$$T^{-1}(B,1) = \bigcup_{n=1}^{\infty} (A_n \cap S^{-1}B,\, n) \quad \text{(disjoint union).}$$

Since $\bigcup_{n=1}^{\infty} A_n = A$ we therefore find that

$$\begin{aligned} \mu(T^{-1}(B,1)) &= \sum_{n=1}^{\infty} \frac{\nu(A_n \cap S^{-1}B)}{\displaystyle\int_A f(y)\,\mathrm{d}\nu(y)} = \frac{\nu(S^{-1}B)}{\displaystyle\int_A f(y)\,\mathrm{d}\nu(y)} \\ &= \frac{\nu(B)}{\displaystyle\int_A f(y)\,\mathrm{d}\nu(y)} = \mu(B,1) . \end{aligned}$$

This shows that T is measure preserving. Moreover, if we consider the induced transformation of T on the set $(A, 1)$, then the first return time $n(x, 1) = \inf\{k \geq 1 : T^k(x, 1) \in (A, 1)\}$ is given by $n(x, 1) = f(x)$, and $T_{(A,1)}(x, 1) = (Sx, 1)$.

1.6 Ergodicity

Definition 1.6.1 *Let T be a measure preserving transformation on a probability space $(X, \mathcal{F}, \mu)$. The map T is said to be ergodic if for every measurable set A satisfying $T^{-1}A = A$, we have $\mu(A) = 0$ or 1.*

Theorem 1.6.1 *Let $(X, \mathcal{F}, \mu)$ be a probability space and $T : X \to X$ measure preserving. The following are equivalent:*

(i) T is ergodic.

(ii) If $B \in \mathcal{F}$ with $\mu(T^{-1}B \Delta B) = 0$, then $\mu(B) = 0$ or 1.

(iii) If $A \in \mathcal{F}$ with $\mu(A) > 0$, then $\mu\left(\cup_{n=1}^{\infty} T^{-n}A\right) = 1$.

(iv) If $A, B \in \mathcal{F}$ with $\mu(A) > 0$ and $\mu(B) > 0$, then there exists $n > 0$ such that $\mu(T^{-n}A \cap B) > 0$.

Remark 1.6.1

1. In case T is invertible, then in the above characterization one can replace T^{-n} by T^n.

2. Note that if $\mu(B \triangle T^{-1}B) = 0$, then $\mu(B \setminus T^{-1}B) = \mu(T^{-1}B \setminus B) = 0$. Since

$$B = \left(B \setminus T^{-1}B\right) \cup \left(B \cap T^{-1}B\right),$$

and

$$T^{-1}B = \left(T^{-1}B \setminus B\right) \cup \left(B \cap T^{-1}B\right),$$

we see that after removing a set of measure 0 from B and a set of measure 0 from $T^{-1}B$, the remaining parts are equal. In this case we say that B equals $T^{-1}B$ modulo sets of measure 0.
3. In words, (iii) says that if A is a set of positive measure, almost every $x \in X$ eventually (in fact infinitely often) will visit A.
4. (iv) says that elements of B will eventually enter A.

Proof of Theorem 1.6.1

(i)⇒(ii) Let $B \in \mathcal{F}$ be such that $\mu(B \Delta T^{-1}B) = 0$. We shall define a measurable set C with $C = T^{-1}C$ and $\mu(C\Delta B) = 0$. Let

$$C = \{x \in X : T^n x \in B \text{ i.o. }\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} T^{-k}B.$$

Then, $T^{-1}C = C$, hence by (i) $\mu(C) = 0$ or 1. Furthermore,

$$\begin{aligned}
\mu(C\Delta B) &= \mu\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} T^{-k}B \cap B^c\right) + \mu\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} T^{-k}B^c \cap B\right) \\
&\le \mu\left(\bigcup_{k=1}^{\infty} T^{-k}B \cap B^c\right) + \mu\left(\bigcup_{k=1}^{\infty} T^{-k}B^c \cap B\right) \\
&\le \sum_{k=1}^{\infty} \mu\left(T^{-k}B\Delta B\right).
\end{aligned}$$

Using induction (and the fact that $\mu(E\Delta F) \le \mu(E\Delta G) + \mu(G\Delta F)$), one can show that for each $k \ge 1$ one has $\mu\left(T^{-k}B\Delta B\right) = 0$. Hence, $\mu(C\Delta B) = 0$ which implies that $\mu(C) = \mu(B)$. Therefore, $\mu(B) = 0$ or 1.

(ii)⇒(iii) Let $\mu(A) > 0$ and let $B = \bigcup_{n=1}^{\infty} T^{-n}A$. Then $T^{-1}B \subset B$. Since T is measure preserving, then $\mu(B) > 0$ and

$$\mu(T^{-1}B\Delta B) = \mu(B \setminus T^{-1}B) = \mu(B) - \mu(T^{-1}B) = 0.$$

Thus, by (ii) $\mu(B) = 1$.

(iii)⇒(iv) Suppose $\mu(A)\mu(B) > 0$. By (iii)

$$\mu(B) = \mu\left(B \cap \bigcup_{n=1}^{\infty} T^{-n}A\right) = \mu\left(\bigcup_{n=1}^{\infty}(B \cap T^{-n}A)\right) > 0.$$

Hence, there exists $k \ge 1$ such that $\mu(B \cap T^{-k}A) > 0$.

(iv)⇒(i) Suppose $T^{-1}A = A$ with $\mu(A) > 0$. If $\mu(A^c) > 0$, then by (iv) there exists $k \ge 1$ such that $\mu(A^c \cap T^{-k}A) > 0$. Since $T^{-k}A = A$, it follows that $\mu(A^c \cap A) > 0$, a contradiction. Hence, $\mu(A) = 1$ and T is ergodic. $\square$

1.7 Other Characterizations of Ergodicity

We denote by $L^0(X, \mathcal{F}, \mu)$ the space of all complex valued measurable functions on the probability space $(X, \mathcal{F}, \mu)$. Let

$$L^p(X, \mathcal{F}, \mu) = \{f \in L^0(X, \mathcal{F}, \mu) : \int_X |f|^p \, \mathrm{d}\mu(x) < \infty\}.$$

We use the subscript $\mathbb{R}$ whenever we are dealing only with real-valued functions.
Let $(X_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2$ be two probability spaces, and $T : X_1 \to X_2$ a measure preserving transformation i.e., $\mu_2(A) = \mu_1(T^{-1}A)$. Define the *induced* operator $U_T : L^0(X_2, \mathcal{F}_2, \mu_2) \to L^0(X_1, \mathcal{F}_1, \mu_1)$ by

$$U_T f = f \circ T.$$

The following properties of U_T are easy to prove.

Proposition 1.7.1 *The operator U_T has the following properties:*

(i) *U_T is linear*

(ii) *$U_T(fg) = U_T(f)U_T(g)$*

(iii) *$U_T c = c$ for any constant c.*

(iv) *U_T is a positive linear operator*

(v) *$U_T 1_B = 1_B \circ T = 1_{T^{-1}B}$ for all $B \in \mathcal{F}_2$.*

(vi) *$\int_{X_1} U_T f \, \mathrm{d}\mu_1 = \int_{X_2} f \, \mathrm{d}\mu_2$ for all $f \in L^0(X_2, \mathcal{F}_2, \mu_2)$, (where if one side doesn't exist or is infinite, then the other side has the same property).*

(vii) *Let $p \geq 1$. Then, $U_T L^p(X_2, \mathcal{F}_2, \mu_2) \subset L^p(X_1, \mathcal{F}_1, \mu_1)$, and $||U_T f||_p = ||f||_p$ for all $f \in L^p(X_2, \mathcal{F}_2, \mu_2)$.*

Exercise 1.7.1 Prove Proposition 1.7.1

Using the above properties, we can give the following characterization of ergodicity

Theorem 1.7.1 *Let $(X, \mathcal{F}, \mu)$ be a probability space, and $T : X \to X$ measure preserving. The following are equivalent:*

(i) T is ergodic.

(ii) If $f \in L^0(X, \mathcal{F}, \mu)$, with $f(Tx) = f(x)$ for all x, then f is a constant a.e.

(iii) If $f \in L^0(X, \mathcal{F}, \mu)$, with $f(Tx) = f(x)$ for a.e. x, then f is a constant a.e.

(iv) If $f \in L^2(X, \mathcal{F}, \mu)$, with $f(Tx) = f(x)$ for all x, then f is a constant a.e.

(v) If $f \in L^2(X, \mathcal{F}, \mu)$, with $f(Tx) = f(x)$ for a.e. x, then f is a constant a.e.

Proof

The implications (iii)⇒(ii), (ii)⇒(iv), (v)⇒(iv), and (iii)⇒(v) are all clear. It remains to show (i)⇒(iii) and (iv)⇒(i).

(i)⇒(iii) Suppose $f(Tx) = f(x)$ a.e. and assume without any loss of generality that f is real (otherwise we consider separately the real and imaginary parts of f). For each $n \geq 1$ and $k \in \mathbb{Z}$, let

$$X_{(k,n)} = \{x \in X : \frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}\}.$$

Then, $T^{-1}X_{(k,n)} \Delta X_{(k,n)} \subseteq \{x : f(Tx) \neq f(x)\}$ which implies that

$$\mu\left(T^{-1}X_{(k,n)} \Delta X_{(k,n)}\right) = 0.$$

By ergodicity of T, $\mu(X_{(k,n)}) = 0$ or 1, for each $k \in \mathbb{Z}$. On the other hand, for each $n \geq 1$, we have

$$X = \bigcup_{k \in \mathbb{Z}} X_{(k,n)} \text{ (disjoint union).}$$

Hence, for each $n \geq 1$, there exists a unique integer k_n such that $\mu\left(X_{(k_n,n)}\right) = 1$. In fact, $X_{(k_1,1)} \supseteq X_{(k_2,2)} \supseteq \ldots$, and $\{\frac{k_n}{2^n}\}$ is a bounded increasing sequence, hence $\lim_{n\to\infty} \frac{k_n}{2^n}$ exists. Let $Y = \bigcap_{n\geq 1} X_{(k_n,n)}$, then $\mu(Y) = 1$. Now, if $x \in Y$, then $0 \leq |f(x) - k_n/2^n| < 1/2^n$ for all n. Hence, $f(x) = \lim_{n\to\infty} \frac{k_n}{2^n}$, and f is a constant on Y.

(iv)⇒(i) Suppose $T^{-1}A = A$ and $\mu(A) > 0$. We want to show that $\mu(A) = 1$. Consider 1_A, the indicator function of A. We have $1_A \in L^2(X, \mathcal{F}, \mu)$, and $1_A \circ T = 1_{T^{-1}A} = 1_A$. Hence, by (iv), 1_A is a constant a.e., hence $1_A = 1$ a.e. and therefore $\mu(A) = 1$. □

1.8 Examples of Ergodic Transformations

Example 1–Irrational Rotations. Consider $([0,1), \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Lebesgue σ-algebra, and λ Lebesgue measure. For $\theta \in (0,1)$, consider the transformation $T_\theta : [0,1) \to [0,1)$ defined by $T_\theta x = x + \theta \pmod 1$. We have seen in example (a) that T_θ is measure preserving with respect λ. When is T_θ ergodic?

If θ is rational, then T_θ is not ergodic. Consider for example $\theta = 1/4$, then the set

$$A = [0, 1/8) \cup [1/4, 3/8) \cup [1/2, 5/8) \cup [3/4, 7/8)$$

is T_θ-invariant but $\mu(A) = 1/2$.

Exercise 1.8.1 Suppose $\theta = \frac{p}{q}$ with $gcd(p,q) = 1$. Find a non-trivial T_θ-*invariant* set. Conclude that T_θ is not ergodic if θ is a rational.

Claim. T_θ is ergodic if and only if θ is irrational.

Proof of Claim.

(⇒) The contrapositive statement is given in Exercise 1.8.1 i.e. if θ is rational, then T_θ is not ergodic.

(⇐) Suppose θ is irrational, and let $f \in L^2(X, \mathcal{B}, \lambda)$ be T_θ-invariant. Write f in its Fourier series

$$f(x) = \sum_{n\in Z} a_n e^{2\pi i n x}.$$

Since $f(T_\theta x) = f(x)$, then

$$\begin{aligned} f(T_\theta x) &= \sum_{n\in Z} a_n e^{2\pi i n(x+\theta)} = \sum_{n\in Z} a_n e^{2\pi i n\theta} e^{2\pi i n x} \\ &= f(x) = \sum_{n\in Z} a_n e^{2\pi i n x}. \end{aligned}$$

Hence, $\sum_{n\in Z} a_n(1-e^{2\pi i n\theta})e^{2\pi i n x} = 0$. By the uniqueness of the Fourier coefficients, we have $a_n(1-e^{2\pi i n\theta}) = 0$ for all $n \in \mathbb{Z}$. If $n \neq 0$, since θ is irrational we have $1 - e^{2\pi i n\theta} \neq 0$. Thus, $a_n = 0$ for all $n \neq 0$, and therefore $f(x) = a_0$ is a constant. By Theorem 1.7.1, T_θ is ergodic.

Exercise 1.8.2 Consider the probability space $([0,1), \mathcal{B}\times\mathcal{B}, \lambda\times\lambda)$, where as above $\mathcal{B}$ is the Lebesgue σ-algebra on $[0,1)$, and λ normalized Lebesgue measure. Suppose $\theta \in (0,1)$ is irrational, and define $T_\theta \times T_\theta : [0,1)\times[0,1) \to [0,1)\times[0,1)$ by

$$T_\theta \times T_\theta(x,y) = (x+\theta \ \mathit{mod}\ (1)\ , y+\theta\ \mathit{mod}\ (1)\ ).$$

Show that $T_\theta \times T_\theta$ is measure preserving, but is **not** ergodic.

Example 2–One (or Two) sided shift. Let $X = \{0,1,\ldots k-1\}^{\mathbb{N}}$, $\mathcal{F}$ the σ-algebra generated by the cylinders, and μ the product measure defined on cylinder sets by

$$\mu\left(\{x : x_0 = a_0, \ldots x_n = a_n\}\right) = p_{a_0}\ldots p_{a_n},$$

where $p = (p_0, p_1, \ldots, p_{k-1})$ is a positive probability vector. Consider the left shift T defined on X by $Tx = y$, where $y_n = x_{n+1}$ (See Example (e) in Subsection 1.3). We show that T is ergodic. Let E be a measurable subset of X which is T-invariant i.e., $T^{-1}E = E$. For any $\epsilon > 0$, by Lemma 1.2.1 (see subsection 1.2), there exists $A \in \mathcal{F}$ which is a finite disjoint union of cylinders such that $\mu(E\Delta A) < \epsilon$. Then

$$\begin{aligned} |\mu(E) - \mu(A)| &= |\mu(E\setminus A) - \mu(A\setminus E)| \\ &\leq \mu(E\setminus A) + \mu(A\setminus E) = \mu(E\Delta A) < \epsilon. \end{aligned}$$

Since A depends on finitely many coordinates only, there exists $n_0 > 0$ such that $T^{-n_0}A$ depends on different coordinates than A. Since μ is a product measure, we have

$$\mu(A\cap T^{-n_0}A) = \mu(A)\mu(T^{-n_0}A) = \mu(A)^2.$$

Further,

$$\mu(E\Delta T^{-n_0}A) = \mu(T^{-n_0}E\Delta T^{-n_0}A) = \mu(E\Delta A) < \epsilon,$$

and

$$\mu\left(E\Delta(A\cap T^{-n_0}A)\right) \le \mu(E\Delta A) + \mu(E\Delta T^{-n_0}A) < 2\epsilon.$$

Hence,

$$|\mu(E) - \mu((A\cap T^{-n_0}A))| \le \mu\left(E\Delta(A\cap T^{-n_0}A)\right) < 2\epsilon.$$

Thus,

$$\begin{aligned} |\mu(E)-\mu(E)^2| &\le |\mu(E)-\mu(A)^2| + |\mu(A)^2-\mu(E)^2| \\ &= |\mu(E)-\mu((A\cap T^{-n_0}A))| + (\mu(A)+\mu(E))|\mu(A)-\mu(E)| \\ &< 4\epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, it follows that $\mu(E) = \mu(E)^2$, hence $\mu(E) = 0$ or 1. Therefore, T is ergodic.

The following lemma provides, in some cases, a useful tool to verify that a measure preserving transformation defined on $([0,1), \mathcal{B}, \mu)$ is ergodic, where $\mathcal{B}$ is the Lebesgue σ-algebra, and μ is a probability measure equivalent to Lebesgue measure λ (i.e., $\mu(A) = 0$ if and only if $\lambda(A) = 0$).

Lemma 1.8.1 *(Knopp's Lemma) . If B is a Lebesgue set and $\mathcal{C}$ is a class of subintervals of $[0,1)$ satisfying*

(a) *every open subinterval of $[0,1)$ is at most a countable union of disjoint elements from $\mathcal{C}$,*

(b) *$\forall A \in \mathcal{C}$, $\lambda(A\cap B) \ge \gamma\lambda(A)$, where $\gamma > 0$ is independent of A,*

then $\lambda(B) = 1$.

Proof The proof is done by contradiction. Suppose $\lambda(B^c) > 0$. Given $\varepsilon > 0$ there exists by Lemma 1.2.1 a set E_ε that is a finite disjoint union of open intervals such that $\lambda(B^c \triangle E_\varepsilon) < \varepsilon$. Now by conditions (a) and (b) (that is, writing E_ε as a countable union of disjoint elements of $\mathcal{C}$) one gets that $\lambda(B\cap E_\varepsilon) \ge \gamma\lambda(E_\varepsilon)$.

Also from our choice of E_ε and the fact that

$$\lambda(B^c \triangle E_\varepsilon) \geq \lambda(B \cap E_\varepsilon) \geq \gamma\lambda(E_\varepsilon) \geq \gamma\lambda(B^c \cap E_\varepsilon) > \gamma(\lambda(B^c) - \varepsilon),$$

we have that

$$\gamma(\lambda(B^c) - \varepsilon) \; < \; \lambda(B^c \triangle E_\varepsilon) \; < \; \varepsilon \; .$$

Hence $\gamma\lambda(B^c) < \varepsilon + \gamma\varepsilon$, and since $\varepsilon > 0$ is arbitrary, we get a contradiction. □

Example 3–Multiplication by 2 modulo 1–Consider $([0,1), \mathcal{B}, \lambda)$ be as in Example (1) above, and let $T : X \to X$ be given by

$$Tx = 2x \bmod 1 = \begin{cases} 2x & 0 \leq x < 1/2 \\ 2x - 1 & 1/2 \leq x < 1, \end{cases}$$

(see Example (b), subsection 1.3). We have seen that T is measure preserving. We will use Lemma 1.8.1 to show that T is ergodic. Let $\mathcal{C}$ be the collection of all intervals of the form $[k/2^n, (k+1)/2^n)$ with $n \geq 1$ and $0 \leq k \leq 2^n - 1$. Notice that the the set $\{k/2^n : n \geq 1, \, 0 \leq k < 2^n - 1\}$ of dyadic rationals is dense in $[0,1)$, hence each open interval is at most a countable union of disjoint elements of $\mathcal{C}$. Hence, $\mathcal{C}$ satisfies the first hypothesis of Knopp's Lemma. Now, T^n maps each dyadic interval of the form $[k/2^n, (k+1)/2^n)$ linearly onto $[0,1)$, (we call such an interval dyadic of order n); in fact, $T^n x = 2^n x \bmod(1)$. Let $B \in \mathcal{B}$ be T-invariant, and assume $\lambda(B) > 0$. Let $A \in \mathcal{C}$, and assume that A is dyadic of order n. Then, $T^n A = [0,1)$ and

$$\begin{aligned} \lambda(A \cap B) &= \lambda(A \cap T^{-n}B) = \frac{1}{\lambda(A)}\lambda(T^n A \cap B) \\ &= \frac{1}{2^n}\lambda(B) = \lambda(A)\lambda(B). \end{aligned}$$

Thus, the second hypothesis of Knopp's Lemma is satisfied with $\gamma = \lambda(B) > 0$. Hence, $\lambda(B) = 1$. Therefore T is ergodic.

Exercise 1.8.3 Let $\beta > 1$ be a non-integer, and consider the transformation $T_\beta : [0,1) \to [0,1)$ given by $T_\beta x = \beta x \bmod(1) \; = \beta x - \lfloor \beta x \rfloor$. Use Lemma 1.8.1 to show that T_β is ergodic with respect to Lebesgue measure λ, i.e. if $T_\beta^{-1} A = A$, then $\lambda(A) = 0$ or 1.

Example 4–Induced transformations of ergodic transformations– Let T be an ergodic measure preserving transformation on the probability space $(X, \mathcal{F}, \mu)$, and $A \in \mathcal{F}$ with $\mu(A) > 0$. Consider the induced transformation T_A on $(A, \mathcal{F} \cap A, \mu_A)$ of T (see subsection 1.5). Recall that $T_A x = T^{n(x)} x$, where $n(x) := \inf\{n \geq 1 : T^n x \in A\}$. Let (as before)

$$\begin{aligned} A_k &= \{x \in A : n(x) = k\} \\ \\ B_k &= \{x \in X \setminus A : Tx, \ldots, T^{k-1}x \notin A, T^k x \in A\}. \end{aligned}$$

Proposition 1.8.1 *If T is ergodic on $(X, \mathcal{F}, \mu)$, then T_A is ergodic on $(A, \mathcal{F} \cap A, \mu_A)$.*

Proof Let $C \in \mathcal{F} \cap A$ be such that $T_A^{-1} C = C$. We want to show that $\mu_A(C) = 0$ or 1; equivalently, $\mu(C) = 0$ or $\mu(C) = \mu(A)$. Since $A = \bigcup_{k \geq 1} A_k$, we have $C = T_A^{-1} C = \bigcup_{k \geq 1} A_k \cap T^{-k} C$. Let $E = \bigcup_{k \geq 1} B_k \cap T^{-k} C$, and $F = E \cup C$ (disjoint union). Recall that (see subsection 1.5) $T^{-1} A = A_1 \cup B_1$, and $T^{-1} B_k = A_{k+1} \cup B_{k+1}$. Hence,

$$\begin{aligned} T^{-1} F &= T^{-1} E \cup T^{-1} C \\ &= \bigcup_{k \geq 1} \left[(A_{k+1} \cup B_{k+1}) \cap T^{-(k+1)} C\right] \cup \left[(A_1 \cup B_1) \cap T^{-1} C\right] \\ &= \bigcup_{k \geq 1} (A_k \cap T^{-k} C) \cup \bigcup_{k \geq 1} (B_k \cap T^{-k} C) \\ &= C \cup E = F. \end{aligned}$$

Hence, F is T-invariant, and by ergodicity of T we have $\mu(F) = 0$ or 1.
–If $\mu(F) = 0$, then $\mu(C) = 0$, and hence $\mu_A(C) = 0$.
–If $\mu(F) = 1$, then $\mu(X \setminus F) = 0$. Since

$$X \setminus F = (A \setminus C) \cup ((X \setminus A) \setminus E) \supseteq A \setminus C,$$

it follows that

$$\mu(A \setminus C) \leq \mu(X \setminus F) = 0.$$

Since $\mu(A \setminus C) = \mu(A) - \mu(C)$, we have $\mu(A) = \mu(C)$, i.e., $\mu_A(C) = 1$. $\square$

Exercise 1.8.4 Show that if T_A is ergodic and $\mu\left(\bigcup_{k \geq 1} T^{-k} A\right) = 1$, then, T is ergodic.

Chapter 2

The Ergodic Theorem

2.1 The Ergodic Theorem and its consequences

The Ergodic Theorem is also known as Birkhoff's Ergodic Theorem or the Individual Ergodic Theorem (1931). This theorem is in fact a generalization of the Strong Law of Large Numbers (SLLN) which states that for a sequence $Y_1, Y_2, \ldots$ of i.i.d. random variables on a probability space $(X, \mathcal{F}, \mu)$, with $\mathrm{E}|Y_i| < \infty$; one has

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i \;=\; \mathrm{E}Y_1 \text{ (a.e.)}.$$

For example consider $X = \{0,1\}^{\mathbb{N}}$, $\mathcal{F}$ the σ-algebra generated by the cylinder sets, and μ the uniform product measure, i.e.,

$$\mu\left(\{x : x_1 = a_1, x_2 = a_2, \ldots, x_n = a_n\}\right) = 1/2^n.$$

Suppose one is interested in finding the frequency of the digit 1. More precisely, for a.e. x we would like to find

$$\lim_{n\to\infty} \frac{1}{n} \#\{1 \le i \le n : x_i = 1\}.$$

Using the Strong Law of Large Numbers one can answer this question easily. Define

$$Y_i(x) \;:=\; \begin{cases} 1, & \text{if } x_i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Since μ is product measure, it is easy to see that $Y_1, Y_2, \ldots$ form an i.i.d. Bernoulli process, and $EY_i = E|Y_i| = 1/2$. Further, $\#\{1 \le i \le n : x_i = 1\} = \sum_{i=1}^{n} Y_i(x)$. Hence, by SLLN one has

$$\lim_{n\to\infty} \frac{1}{n}\#\{1 \le i \le n : x_i = 1\} = \frac{1}{2}.$$

Suppose now we are interested in the frequency of the block 011, i.e., we would like to find

$$\lim_{n\to\infty} \frac{1}{n}\#\{1 \le i \le n : x_i = 0, x_{i+1} = 1, x_{i+2} = 1\}.$$

We can start as above by defining random variables

$$Z_i(x) := \begin{cases} 1, & \text{if } x_i = 0, x_{i+1} = 1, x_{i+2} = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\frac{1}{n}\#\{1 \le i \le n : x_i = 0, x_{i+1} = 1, x_{i+2} = 1\} = \frac{1}{n}\sum_{i=1}^{n} Z_i(x).$$

It is not hard to see that this sequence is stationary but not independent. So one cannot directly apply the strong law of large numbers. Notice that if T is the left shift on X, then $Y_n = Y_1 \circ T^{n-1}$ and $Z_n = Z_1 \circ T^{n-1}$.
In general, suppose $(X, \mathcal{F}, \mu)$ is a probability space and $T : X \to X$ a measure preserving transformation. For $f \in L^1(X, \mathcal{F}, \mu)$, we would like to know under what conditions does the limit $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x)$ exist a.e. If it does exist what is its value? This is answered by the Ergodic Theorem which was originally proved by G.D. Birkhoff in 1931. Since then, several proofs of this important theorem have been obtained; here we present a recent proof given by T. Kamae and M.S. Keane in [KK].

Theorem 2.1.1 (The Ergodic Theorem) *Let $(X, \mathcal{F}, \mu)$ be a probability space and $T : X \to X$ a measure preserving transformation. Then, for any f in $L^1(\mu)$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i(x)) = f^*(x)$$

exists a.e., is T-invariant and $\int_X f\, d\mu = \int_X f^*\, d\mu$. If moreover T is ergodic, then f^* is a constant a.e. and $f^* = \int_X f\, d\mu$.

For the proof of the above theorem, we need the following simple lemma.

Lemma 2.1.1 *Let $M > 0$ be an integer, and suppose $\{a_n\}_{n\geq 0}$, $\{b_n\}_{n\geq 0}$ are sequences of non-negative real numbers such that for each $n = 0, 1, 2, \ldots$ there exists an integer $1 \leq m \leq M$ with*

$$a_n + \cdots + a_{n+m-1} \geq b_n + \cdots + b_{n+m-1}.$$

Then, for each positive integer $N > M$, one has

$$a_0 + \cdots + a_{N-1} \geq b_0 + \cdots + b_{N-M-1}.$$

Proof of Lemma 2.1.1 Using the hypothesis we recursively find integers $m_0 < m_1 < \ldots < m_k < N$ with the following properties

$$\begin{gathered} m_0 \leq M,\ m_{i+1} - m_i \leq M \text{ for } i = 0, \ldots, k-1, \text{ and } N - m_k < M, \\ a_0 + \ldots + a_{m_0-1} \geq b_0 + \ldots + b_{m_0-1}, \\ a_{m_0} + \ldots + a_{m_1-1} \geq b_{m_0} + \ldots + b_{m_1-1}, \\ \vdots \\ a_{m_{k-1}} + \ldots + a_{m_k-1} \geq b_{m_{k-1}} + \ldots + b_{m_k-1}. \end{gathered}$$

Then,

$$\begin{aligned} a_0 + \ldots + a_{N-1} &\geq a_0 + \ldots + a_{m_k-1} \\ &\geq b_0 + \ldots + b_{m_k-1} \geq b_0 + \ldots b_{N-M-1}. \end{aligned}$$

$\square$

Proof of Theorem 2.1.1 Assume with no loss of generality that $f \geq 0$ (otherwise we write $f = f^+ - f^-$, and we consider each part separately). Let $f_n(x) = f(x) + \ldots + f(T^{n-1}x)$, $\overline{f}(x) = \limsup_{n\to\infty} \dfrac{f_n(x)}{n}$, and $\underline{f}(x) = \liminf_{n\to\infty} \dfrac{f_n(x)}{n}$. Then $\overline{f}$ and $\underline{f}$ are T-invariant. This follows from

$$\begin{aligned} \overline{f}(Tx) &= \limsup_{n\to\infty} \frac{f_n(Tx)}{n} \\ &= \limsup_{n\to\infty} \left[\frac{f_{n+1}(x)}{n+1} \cdot \frac{n+1}{n} - \frac{f(x)}{n} \right] \\ &= \limsup_{n\to\infty} \frac{f_{n+1}(x)}{n+1} = \overline{f}(x). \end{aligned}$$

(Similarly $\underline{f}$ is T-invariant). Now, to prove that f^* exists, is integrable and T-invariant, it is enough to show that

$$\int_X \underline{f}\,\mathrm{d}\mu \geq \int_X f\,\mathrm{d}\mu \geq \int_X \overline{f}\,\mathrm{d}\mu.$$

For since $\overline{f} - \underline{f} \geq 0$, this would imply that $\overline{f} = \underline{f} = f^*$. a.e.
We first prove that $\int_X \overline{f} d\mu \leq \int_X f\,\mathrm{d}\mu$. Fix any $0 < \epsilon < 1$, and let $L > 0$ be any real number. By definition of $\overline{f}$, for any $x \in X$, there exists an integer $m > 0$ such that

$$\frac{f_m(x)}{m} \geq \min(\overline{f}(x), L)(1-\epsilon).$$

Now, for any $\delta > 0$ there exists an integer $M > 0$ such that the set

$$X_0 = \{x \in X : \exists\ 1 \leq m \leq M \text{ with } f_m(x) \geq m\ \min(\overline{f}(x), L)(1-\epsilon)\}$$

has measure at least $1-\delta$. Define F on X by

$$F(x) = \begin{cases} f(x) & x \in X_0 \\ L & x \notin X_0. \end{cases}$$

Notice that $f \leq F$ (why?). For any $x \in X$, let $a_n = a_n(x) = F(T^n x)$, and $b_n = b_n(x) = \min(\overline{f}(x), L)(1-\epsilon)$ (so b_n is independent of n).We now show that $\{a_n\}$ and $\{b_n\}$ satisfy the hypothesis of Lemma 2.1.1 with $M > 0$ as above. For any $n = 0, 1, 2, \ldots$

–if $T^n x \in X_0$, then there exists $1 \leq m \leq M$ such that

$$\begin{aligned} f_m(T^n x) &\geq m\ \min(\overline{f}(T^n x), L)(1-\epsilon) \\ &= m\ \min(\overline{f}(x), L)(1-\epsilon) \\ &= b_n + \ldots + b_{n+m-1}. \end{aligned}$$

Hence,

$$\begin{aligned} a_n + \ldots + a_{n+m-1} &= F(T^n x) + \ldots + F(T^{n+m-1}x) \\ &\geq f(T^n x) + \ldots + f(T^{n+m-1}x) = f_m(T^n x) \\ &\geq b_n + \ldots + b_{n+m-1}. \end{aligned}$$

–If $T^n x \notin X_0$, then take $m = 1$ since

$$a_n = F(T^n x) = L \geq \min(\overline{f}(x), L)(1-\epsilon) = b_n.$$

Hence by Lemma 2.1.1 for all integers $N > M$ one has

$$F(x) + \ldots + F(T^{N-1}x) \geq (N - M) \min(\overline{f}(x), L)(1 - \epsilon).$$

Integrating both sides, and using the fact that T is measure preserving one gets

$$N \int_X F(x) \, d\mu(x) \geq (N - M) \int_X \min(\overline{f}(x), L)(1 - \epsilon) \, d\mu(x).$$

Since

$$\int_X F(x) \, d\mu(x) = \int_{X_0} f(x) \, d\mu(x) + L\mu(X \setminus X_0),$$

one has

$$\begin{aligned} \int_X f(x) \, d\mu(x) &\geq \int_{X_0} f(x) \, d\mu(x) \\ &= \int_X F(x) \, d\mu(x) - L\mu(X \setminus X_0) \\ &\geq \frac{(N - M)}{N} \int_X \min(\overline{f}(x), L)(1 - \epsilon) \, d\mu(x) - L\delta. \end{aligned}$$

Now letting first $N \to \infty$, then $\delta \to 0$, then $\epsilon \to 0$, and lastly $L \to \infty$ one gets together with the monotone convergence theorem that $\overline{f}$ is integrable, and

$$\int_X f(x) \, d\mu(x) \geq \int_X \overline{f}(x) \, d\mu(x).$$

We now prove that

$$\int_X f(x) \, d\mu(x) \leq \int_X \underline{f}(x) \, d\mu(x).$$

Fix $\epsilon > 0$, for any $x \in X$ there exists an integer m such that

$$\frac{f_m(x)}{m} \leq (\underline{f}(x) + \epsilon).$$

For any $\delta > 0$ there exists an integer $M > 0$ such that the set

$$Y_0 = \{x \in X : \exists \, 1 \leq m \leq M \text{ with } f_m(x) \leq m \, (\underline{f}(x) + \epsilon)\}$$

has measure at least $1-\delta$. Define G on X by

$$G(x) = \begin{cases} f(x) & x \in Y_0 \\ 0 & x \notin Y_0. \end{cases}$$

Notice that $G \leq f$. Let $b_n = G(T^n x)$, and $a_n = \underline{f}(x)+\epsilon$ (so a_n is independent of n). One can easily check that the sequences $\{a_n\}$ and $\{b_n\}$ satisfy the hypothesis of Lemma 2.1.1 with $M > 0$ as above. Hence for any $M > N$, one has

$$G(x) + \ldots + G(T^{N-M-1}x) \leq N(\underline{f}(x) + \epsilon).$$

Integrating both sides yields

$$(N-M)\int_X G(x)d\mu(x) \leq N(\int_X \underline{f}(x)d\mu(x) + \epsilon).$$

Since $f \geq 0$, the measure ν defined by $\nu(A) = \int_A f(x)\,\mathrm{d}\mu(x)$ is absolutely continuous with respect to the measure μ. Hence, there exists $\delta_0 > 0$ such that if $\mu(A) < \delta$, then $\nu(A) < \delta_0$. Since $\mu(X \setminus Y_0) < \delta$, then $\nu(X \setminus Y_0) = \int_{X\setminus Y_0} f(x)d\mu(x) < \delta_0$. Hence,

$$\begin{aligned} \int_X f(x)\,\mathrm{d}\mu(x) &= \int_X G(x)\,\mathrm{d}\mu(x) + \int_{X\setminus Y_0} f(x)\,\mathrm{d}\mu(x) \\ &\leq \frac{N}{N-M}\int_X (\underline{f}(x)+\epsilon)\,\mathrm{d}\mu(x) + \delta_0. \end{aligned}$$

Now, let first $N \to \infty$, then $\delta \to 0$ (and hence $\delta_0 \to 0$), and finally $\epsilon \to 0$, one gets

$$\int_X f(x)\,\mathrm{d}\mu(x) \leq \int_X \underline{f}(x)\,\mathrm{d}\mu(x).$$

This shows that

$$\int_X \underline{f}\,\mathrm{d}\mu \geq \int_X f\,\mathrm{d}\mu \geq \int_X \overline{f}\,\mathrm{d}\mu,$$

hence, $\overline{f} = \underline{f} = f^*$ a.e., and f^* is T-invariant. In case T is ergodic, then the T-invariance of f^* implies that f^* is a constant a.e. Therefore,

$$f^*(x) = \int_X f^*(y)d\mu(y) = \int_X f(y)\,\mathrm{d}\mu(y).$$

□

Remarks
(1) Let us study further the limit f^* in the case that T is not ergodic. Let $\mathcal{I}$ be the sub-σ-algebra of $\mathcal{F}$ consisting of all T-invariant subsets $A \in \mathcal{F}$. Notice that if $f \in L^1(\mu)$, then the *conditional expectation* of f given $\mathcal{I}$ (denoted by $E_\mu(f|\mathcal{I})$), is the unique a.e. $\mathcal{I}$-measurable $L^1(\mu)$ function with the property that
$$\int_A f(x)\,\mathrm{d}\mu(x) = \int_A E_\mu(f|\mathcal{I})(x)\,\mathrm{d}\mu(x)$$
for all $A \in \mathcal{I}$ i.e., $T^{-1}A = A$. We claim that $f^* = E_\mu(f|\mathcal{I})$. Since the limit function f^* is T-invariant, it follows that f^* is $\mathcal{I}$-measurable. Furthermore, for any $A \in \mathcal{I}$, by the ergodic theorem and the T-invariance of 1_A,
$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}(f1_A)(T^ix) = 1_A(x)\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}f(T^ix) = 1_A(x)f^*(x) \ a.e.$$
and
$$\int_X f1_A(x)\,\mathrm{d}\mu(x) = \int_X f^*1_A(x)\,\mathrm{d}\mu(x).$$
This shows that $f^* = E_\mu(f|\mathcal{I})$.

(2) Suppose T is ergodic and measure preserving with respect to μ, and let ν be a probability measure which is equivalent to μ (i.e. μ and ν have the same sets of measure zero so $\mu(A) = 0$ if and only if $\nu(A) = 0$), then for every $f \in L^1(\mu)$ one has
$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}f(T^i(x)) = \int_X f\,\mathrm{d}\mu$$
ν a.e.

Exercise 2.1.1 (*Kac's Lemma*) Let T be a measure preserving and ergodic transformation on a probability space $(X, \mathcal{F}, \mu)$. Let A be a measurable subset of X of positive μ measure, and denote by n the first return time map and let T_A be the induced transformation of T on A (see section 1.5). Prove that
$$\int_A n(x)\,\mathrm{d}\mu = 1.$$

Conclude that $n(x) \in L^1(A, \mu_A)$, and that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} n(T_A^i(x)) = \frac{1}{\mu(A)},$$

almost everywhere on A.

Exercise 2.1.2 Let $\beta = \dfrac{1+\sqrt{5}}{2}$, and consider the transformation $T_\beta : [0,1) \to [0,1)$ given by $T_\beta x = \beta x \bmod(1) = \beta x - \lfloor \beta x \rfloor$. Define b_1 on $[0,1)$ by

$$b_1(x) = \begin{cases} 0 & \text{if } 0 \le x < 1/\beta \\ 1 & \text{if } 1/\beta \le x < 1, \end{cases}$$

Fix $k \ge 0$. Find the a.e. value (with respect to Lebesgue measure) of the following limit

$$\lim_{n\to\infty} \frac{1}{n} \#\{1 \le i \le n : b_i = 0, b_{i+1} = 0, \ldots, b_{i+k} = 0\}.$$

Using the Ergodic Theorem, one can give yet another characterization of ergodicity.

Corollary 2.1.1 *Let $(X, \mathcal{F}, \mu)$ be a probability space, and $T : X \to X$ a measure preserving transformation. Then, T is ergodic if and only if for all $A, B \in \mathcal{F}$, one has*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \mu(T^{-i}A \cap B) = \mu(A)\mu(B). \tag{2.1}$$

Proof Suppose T is ergodic, and let $A, B \in \mathcal{F}$. Since the indicator function $1_A \in L^1(X, \mathcal{F}, \mu)$, by the ergodic theorem one has

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} 1_A(T^i x) = \int_X 1_A(x)\, \mathrm{d}\mu(x) = \mu(A) \text{ a.e.}$$

Then,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_{T^{-i}A\cap B}(x) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_{T^{-i}A}(x)1_B(x)\\
&= 1_B(x)\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_A(T^ix)\\
&= 1_B(x)\mu(A) \text{ a.e.}
\end{aligned}$$

Since for each n, the function $\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_{T^{-i}A\cap B}$ is dominated by the constant function 1, it follows by the dominated convergence theorem that

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n}\mu(T^{-i}A\cap B) &= \int_X \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} 1_{T^{-i}A\cap B}(x)\,\mathrm{d}\mu(x)\\
&= \int_X 1_B\mu(A)\,\mathrm{d}\mu(x) = \mu(A)\mu(B).
\end{aligned}$$

Conversely, suppose (2.1) holds for every $A, B \in \mathcal{F}$. Let $E \in \mathcal{F}$ be such that $T^{-1}E = E$ and $\mu(E) > 0$. By invariance of E, we have $\mu(T^{-i}E\cap E) = \mu(E)$, hence

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

On the other hand, by (2.1)

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mu(T^{-i}E\cap E) = \mu(E)^2.$$

Hence, $\mu(E) = \mu(E)^2$. Since $\mu(E) > 0$, this implies $\mu(E) = 1$. Therefore, T is ergodic. □

To show ergodicity one needs to verify equation (2.1) for sets A and B belonging to a generating semi-algebra only as the next proposition shows.

Proposition 2.1.1 *Let $(X, \mathcal{F}, \mu)$ be a probability space, and $\mathcal{S}$ a generating semi-algebra of $\mathcal{F}$. Let $T : X \to X$ be a measure preserving transformation. Then, T is ergodic if and only if for all $A, B \in \mathcal{S}$, one has*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\mu(T^{-i}A\cap B) = \mu(A)\mu(B). \tag{2.2}$$

Proof We only need to show that if (2.2) holds for all $A, B \in \mathcal{S}$, then it holds for all $A, B \in \mathcal{F}$. Let $\epsilon > 0$, and $A, B \in \mathcal{F}$. Then, by Lemma 1.2.1 (subsection 1.2) there exist sets A_0, B_0 each of which is a finite disjoint union of elements of $\mathcal{S}$ such that

$$\mu(A\Delta A_0) < \epsilon, \text{ and } \mu(B\Delta B_0) < \epsilon.$$

Since,

$$(T^{-i}A \cap B)\Delta(T^{-i}A_0 \cap B_0) \subseteq (T^{-i}A\Delta T^{-i}A_0) \cup (B\Delta B_0),$$

it follows that

$$\begin{aligned} |\mu(T^{-i}A \cap B) - \mu(T^{-i}A_0 \cap B_0)| &\leq \mu\left[(T^{-i}A \cap B)\Delta(T^{-i}A_0 \cap B_0)\right] \\ &\leq \mu(T^{-i}A\Delta T^{-i}A_0) + \mu(B\Delta B_0) \\ &< 2\epsilon. \end{aligned}$$

Further,

$$\begin{aligned} |\mu(A)\mu(B) - \mu(A_0)\mu(B_0)| &\leq \mu(A)|\mu(B) - \mu(B_0)| + \mu(B_0)|\mu(A) - \mu(A_0)| \\ &\leq |\mu(B) - \mu(B_0)| + |\mu(A) - \mu(A_0)| \\ &\leq \mu(B\Delta B_0) + \mu(A\Delta A_0) \\ &< 2\epsilon. \end{aligned}$$

Hence,

$$\begin{aligned} &\left|\left(\frac{1}{n}\sum_{i=0}^{n-1}\mu(T^{-i}A \cap B) - \mu(A)\mu(B)\right) - \left(\frac{1}{n}\sum_{i=0}^{n-1}\mu(T^{-i}A_0 \cap B_0) - \mu(A_0)\mu(B_0)\right)\right| \\ \leq\ & \frac{1}{n}\sum_{i=0}^{n-1}\left|\mu(T^{-i}A \cap B) + \mu(T^{-i}A_0 \cap B_0)\right| - |\mu(A)\mu(B) - \mu(A_0)\mu(B_0)| \\ <\ & 4\epsilon. \end{aligned}$$

Therefore,

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

$\square$

Theorem 2.1.2 *Suppose μ_1 and μ_2 are probability measures on $(X, \mathcal{F})$, and $T : X \to X$ is measurable and measure preserving with respect to μ_1 and μ_2. Then,*

(i) if T is ergodic with respect to μ_1, and μ_2 is absolutely continuous with respect to μ_1, then $\mu_1 = \mu_2$,

(ii) if T is ergodic with respect to μ_1 and μ_2, then either $\mu_1 = \mu_2$ or μ_1 and μ_2 are singular with respect to each other.

Proof (i) Suppose T is ergodic with respect to μ_1 and μ_2 is absolutely continuous with respect to μ_1. For any $A \in \mathcal{F}$, by the ergodic theorem for a.e. x one has

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i x) = \mu_1(A).$$

Let

$$C_A = \{x \in X : \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i x) = \mu_1(A)\},$$

then $\mu_1(C_A) = 1$, and by absolute continuity of μ_2 one has $\mu_2(C_A) = 1$. Since T is measure preserving with respect to μ_2, for each $n \geq 1$ one has

$$\frac{1}{n} \sum_{i=0}^{n-1} \int_X 1_A(T^i x)\, \mathrm{d}\mu_2(x) = \mu_2(A).$$

On the other hand, by the dominated convergence theorem one has

$$\lim_{n\to\infty} \int_X \frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i x) d\mu_2(x) = \int_X \mu_1(A)\, \mathrm{d}\mu_2(x).$$

This implies that $\mu_1(A) = \mu_2(A)$. Since $A \in \mathcal{F}$ is arbitrary, we have $\mu_1 = \mu_2$.

(ii) Suppose T is ergodic with respect to μ_1 and μ_2. Assume that $\mu_1 \neq \mu_2$. Then, there exists a set $A \in \mathcal{F}$ such that $\mu_1(A) \neq \mu_2(A)$. For $i = 1, 2$ let

$$C_i = \{x \in X : \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} 1_A(T^j x) = \mu_i(A)\}.$$

By the ergodic theorem $\mu_i(C_i) = 1$ for $i = 1, 2$. Since $\mu_1(A) \neq \mu_2(A)$, then $C_1 \cap C_2 = \emptyset$. Thus μ_1 and μ_2 are supported on disjoint sets, and hence μ_1 and μ_2 are mutually singular. □

We end this subsection with a short discussion that the assumption of ergodicity is not very restrictive. Let T be a transformation on the probability space $(X, \mathcal{F}, \mu)$, and suppose T is measure preserving but not necessarily ergodic. We assume that X is a complete separable metric space, and $\mathcal{F}$ the corresponding Borel σ-algebra (in order to make sure that the conditional expectation is well-defined a.e.). Let $\mathcal{I}$ be the sub-σ-algebra of T-invariant measurable sets. We can decompose μ into T-invariant ergodic components in the following way. For $x \in X$, define a measure μ_x on $\mathcal{F}$ by

$$\mu_x(A) = E_\mu(1_A|\mathcal{I})(x).$$

Then, for any $f \in L^1(X, \mathcal{F}, \mu)$,

$$\int_X f(y)\, \mathrm{d}\mu_x(y) = E_\mu(f|\mathcal{I})(x).$$

Note that

$$\mu(A) = \int_X E_\mu(1_A|\mathcal{I})(x)\, \mathrm{d}\mu(x) = \int_X \mu_x(A)\, \mathrm{d}\mu(x),$$

and that $E_\mu(1_A|\mathcal{I})(x)$ is T-invariant. We show that μ_x is T-invariant and ergodic for a.e. $x \in X$. So let $A \in \mathcal{F}$, then for a.e. x

$$\mu_x(T^{-1}A) = E_\mu(1_A \circ T|\mathcal{I})(x) = E_\mu(I_A|\mathcal{I})(Tx) = E_\mu(I_A|\mathcal{I})(x) = \mu_x(A).$$

Now, let $A \in \mathcal{F}$ be such that $T^{-1}A = A$. Then, 1_A is T-invariant, and hence $\mathcal{I}$-measurable. Then,

$$\mu_x(A) = E_\mu(1_A|\mathcal{I})(x) = 1_A(x) \;\; a.e.$$

Hence, for a.e. x and for any $B \in \mathcal{F}$,

$$\mu_x(A \cap B) = E_\mu(1_A 1_B|\mathcal{I})(x) = 1_A(x) E_\mu(1_B|\mathcal{I})(x) = \mu_x(A)\mu_x(B).$$

In particular, if $A = B$, then the latter equality yields $\mu_x(A) = \mu_x(A)^2$ which implies that for a.e. x, $\mu_x(A) = 0$ or 1. Therefore, μ_x is ergodic. (One in fact needs to work a little harder to show that one can find a set N of μ-measure zero, such that for any $x \in X \setminus N$, and any T-invariant set A, one has $\mu_x(A) = 0$ or 1. In the above analysis the a.e. set depended on the choice of A. Hence, the above analysis is just a rough sketch of the proof of what is called *the ergodic decomposition* of measure preserving transformations.)

2.2 Characterization of Irreducible Markov Chains

Consider the Markov Chain in Example(f) subsection 1.3. That is $X = \{0, 1, \ldots N-1\}^{\mathbb{Z}}$, $\mathcal{F}$ the σ-algebra generated by the cylinders, $T : X \to X$ the left shift, and μ the Markov measure defined by the stochastic $N \times N$ matrix $P = (p_{ij})$, and the positive probability vector $\pi = (\pi_0, \pi_1, \ldots, \pi_{N-1})$ satisfying $\pi P = \pi$. That is

$$\mu(\{x : x_0 = i_0, x_1 = i_1, \ldots x_n = i_n\}) = \pi_{i_0} p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n}.$$

We want to find necessary and sufficient conditions for T to be ergodic. To achieve this, we first set

$$Q = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k,$$

where $P^k = (p_{ij}^{(k)})$ is the k^{th} power of the matrix P, and P^0 is the $k \times k$ identity matrix. More precisely, $Q = (q_{ij})$, where

$$q_{ij} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)}.$$

Lemma 2.2.1 *For each $i, j \in \{0, 1, \ldots N-1\}$, the limit $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)}$ exists, i.e., q_{ij} is well-defined.*

Proof For each n,

$$\frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)} = \frac{1}{\pi_i} \frac{1}{n} \sum_{k=0}^{n-1} \mu(\{x \in X : x_0 = i, x_k = j\}).$$

Since T is measure preserving, by the ergodic theorem,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_k=j\}}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_0=j\}}(T^k x) = f^*(x),$$

where f^* is T-invariant and integrable. Then,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_0=i,x_k=j\}}(x) = 1_{\{x:x_0=i\}}(x) \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_0=j\}}(T^k x) = f^*(x) 1_{\{x:x_0=i\}}(x).$$

Since $\frac{1}{n}\sum_{k=0}^{n-1} 1_{\{x:x_0=i,x_k=j\}}(x) \leq 1$ for all n, by the dominated convergence theorem,

$$\begin{aligned}
q_{ij} &= \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(\{x \in X : x_0 = i, x_k = j\}) \\
&= \frac{1}{\pi_i} \int_X \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_0=i,x_k=j\}}(x)\, \mathrm{d}\mu(x) \\
&= \frac{1}{\pi_i} \int_X f^*(x) 1_{\{x:x_0=i\}}(x)\, \mathrm{d}\mu(x) \\
&= \frac{1}{\pi_i} \int_{\{x:x_0=i\}} f^*(x)\, \mathrm{d}\mu(x)
\end{aligned}$$

which is finite since f^* is integrable. Hence q_{ij} exists. □

Exercise 2.2.1 Show that the matrix Q has the following properties:
(a) Q is stochastic.
(b) $Q = QP = PQ = Q^2$.
(c) $\pi Q = \pi$.

We now give a characterization for the ergodicity of T. Recall that the matrix P is said to be irreducible if for every $i, j \in \{0, 1, \ldots N-1\}$, there exists $n \geq 1$ such that $p_{ij}^{(n)} > 0$.

Theorem 2.2.1 *The following are equivalent,*

(i) T is ergodic.

(ii) All rows of Q are identical.

(iii) $q_{ij} > 0$ for all i, j.

(iv) P is irreducible.

(v) 1 is a simple eigenvalue of P.

Proof
(i) ⇒ (ii) By the ergodic theorem for each i, j,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{x:x_0=i,x_k=j\}}(x) = 1_{\{x:x_0=i\}}(x)\pi_j.$$

By the dominated convergence theorem,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(\{x \in X : x_0 = i, x_k = j\}) = \pi_i \pi_j.$$

Hence,

$$q_{ij} = \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(\{x \in X : x_0 = i, x_k = j\}) = \pi_j,$$

i.e., q_{ij} is independent of i. Therefore, all rows of Q are identical.

(ii) $\Rightarrow$ (iii) If all the rows of Q are identical, then all the columns of Q are constants. Thus, for each j there exists a constant c_j such that $q_{ij} = c_j$ for all i. Since $\pi Q = \pi$, it follows that $q_{ij} = c_j = \pi_j > 0$ for all i, j.

(iii) $\Rightarrow$ (iv) For any i, j

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)} = q_{ij} > 0.$$

Hence, there exists n such that $p_{ij}^{(n)} > 0$, therefore P is irreducible.

(iv) $\Rightarrow$ (iii) Suppose P is irreducible. For any state $i \in \{0, 1, \ldots, N-1\}$, let $S_i = \{j : q_{ij} > 0\}$. Since Q is a stochastic matrix, it follows that $S_i \neq \emptyset$. Let $l \in S_i$, then $q_{il} > 0$. Since $Q = QP = QP^n$ for all n, then for any state j

$$q_{ij} = \sum_{m=0}^{N-1} q_{im} p_{mj}^{(n)} \geq q_{il} p_{lj}^{(n)}$$

for any n. Since P is irreducible, there exists n such that $p_{lj}^{(n)} > 0$. Hence, $q_{ij} > 0$ for all i, j.

(iii) $\Rightarrow$ (ii) Suppose $q_{ij} > 0$ for all $j = 0, 1, \ldots, N-1$. Fix any state j, and let $q_j = \max_{0 \leq i \leq N-1} q_{ij}$. Suppose that not all the q_{ij}'s are the same. Then there exists $k \in \{0, 1, \ldots, N-1\}$ such that $q_{kj} < q_j$. Since Q is stochastic and $Q^2 = Q$, then for any $i \in \{0, 1, \ldots, N-1\}$ we have,

$$q_{ij} = \sum_{l=0}^{N-1} q_{il} q_{lj} < \sum_{l=0}^{N-1} q_{il} q_j = q_j.$$

This implies that $q_j = \max_{0\leq i\leq N-1} q_{ij} < q_j$, a contradiction. Hence, the columns of Q are constants, or all the rows are identical.

(ii) $\Rightarrow$ (i) Suppose all the rows of Q are identical. We have shown above that this implies $q_{ij} = \pi_j$ for all $i,j \in \{0,1,\ldots,N-1\}$. Hence $\pi_j = \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} p_{ij}^{(k)}$.

Let

$$A = \{x : x_r = i_0, \ldots, x_{r+l} = i_l\}, \text{ and } B = \{x : x_s = j_0, \ldots, x_{s+m} = j_m\}$$

be any two cylinder sets of X. By Proposition 2.1.1 in Section 2, we must show that

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

Since T is the left shift, for all n sufficiently large, the cylinders $T^{-n}A$ and B depend on different coordinates. Hence, for n sufficiently large,

$$\mu(T^{-n}A\cap B) = \pi_{j_0}p_{j_0j_1}\ldots p_{j_{m-1}j_m}p_{j_mi_0}^{(n+r-s-m)}p_{i_0i_1}\ldots p_{i_{l-1}i_l}.$$

Thus,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \mu(T^{-k}A\cap B)\\
= \;& \pi_{j_0}p_{j_0j_1}\ldots p_{j_{m-1}j_m}p_{i_0i_1}\ldots p_{i_{l-1}i_l} \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} p_{j_mi_0}^{(k)}\\
= \;& (\pi_{j_0}p_{j_0j_1}\ldots p_{j_{m-1}j_m})(\pi_{i_0}p_{i_0i_1}\ldots p_{i_{l-1}i_l})\\
= \;& \mu(B)\mu(A).
\end{aligned}$$

Therefore, T is ergodic.

(ii) $\Rightarrow$ (v) If all the rows of Q are identical, then $q_{ij} = \pi_j$ for all i,j. If $vP = v$, then $vQ = v$. This implies that for all j, $v_j = (\sum_{i=0}^{N-1} v_i)\pi_j$. Thus, v is a multiple of π. Therefore, 1 is a simple eigenvalue.

(v) $\Rightarrow$ (ii) Suppose 1 is a simple eigenvalue. For any i, let $q_i^* = (q_{i0},\ldots,q_{i(N-1)})$ denote the i^{th} row of Q then, q_i^* is a probability vector. From $Q = QP$, we get $q_i^* = q_i^*P$. By hypothesis π is the only probability vector satisfying $\pi P = P$, hence $\pi = q_i^*$, and all the rows of Q are identical. $\square$

2.3 Mixing

As a corollary to the ergodic theorem we found a new definition of ergodicity; namely, asymptotic average independence. Based on the same idea, we now define other notions of *weak* independence that are stronger than ergodicity.

Definition 2.3.1 *Let $(X, \mathcal{F}, \mu)$ be a probability space, and $T : X \to X$ a measure preserving transformation. Then,*

(i) T is weakly mixing if for all $A, B \in \mathcal{F}$, one has

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left|\mu(T^{-i}A \cap B) - \mu(A)\mu(B)\right| = 0. \tag{2.3}$$

(ii) T is strongly mixing if for all $A, B \in \mathcal{F}$, one has

$$\lim_{n\to\infty} \mu(T^{-i}A \cap B) = \mu(A)\mu(B). \tag{2.4}$$

Notice that strongly mixing implies weakly mixing, and weakly mixing implies ergodicity. This follows from the simple fact that if $\{a_n\}$ is a sequence of real numbers such that $\lim_{n\to\infty} a_n = 0$, then $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} |a_i| = 0$, and hence $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} a_i = 0$. Furthermore, if $\{a_n\}$ is a bounded sequence, then the following are equivalent:

(i) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} |a_i| = 0$

(ii) $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} |a_i|^2 = 0$

(iii) there exists a subset J of the integers of density zero, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \# \left(\{0, 1, \ldots, n-1\} \cap J\right) = 0,$$

such that $\lim_{n\to\infty, n\notin J} a_n = 0$.

Using this one can give three equivalent characterizations of weakly mixing transformations, can you state them?

Exercise 2.3.1 Let $(X, \mathcal{F}, \mu)$ be a probability space, and $T : X \to X$ a measure preserving transformation. Let $\mathcal{S}$ be a generating semi-algebra of $\mathcal{F}$.

(a) Show that if equation (2.3) holds for all $A, B \in \mathcal{S}$, then T is weakly mixing.

(b) Show that if equation (2.4) holds for all $A, B \in \mathcal{S}$, then T is strongly mixing.

Exercise 2.3.2 Consider the one or two-sided Bernoulli shift T as given in Example (e) in subsection 1.3, and Example (2) in subsection 1.8. Show that T is strongly mixing.

Exercise 2.3.3 Let $(X, \mathcal{F}, \mu)$ be a probability space, and $T : X \to X$ a measure preserving transformation. Consider the transformation $T \times T$ defined on $(X \times X, \mathcal{F} \times \mathcal{F}, \mu \times \mu)$ by $T \times T(x, y) = (Tx, Ty)$.

(a) Show that $T \times T$ is measure preserving with respect to $\mu \times \mu$.

(b) Show that $T \times T$ is ergodic, if and only if T *is* weakly mixing.

Chapter 3

Measure Preserving Isomorphisms and Factor Maps

3.1 Measure Preserving Isomorphisms

Given a measure preserving transformation T on a probability space $(X, \mathcal{F}, \mu)$, we call the quadruple $(X, \mathcal{F}, \mu, T)$ *a dynamical system* . Now, given two dynamical systems $(X, \mathcal{F}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$, what should we mean by: *these systems are the same*? On each space there are two important structures:

(1) The measure structure given by the σ-algebra and the probability measure. Note, that in this context, sets of measure zero can be ignored.

(2) The dynamical structure, given by a measure preserving transformation.

So our notion of *being the same* must mean that we have a map

$$\psi : (X, \mathcal{F}, \mu, T) \to (Y, \mathcal{C}, \nu, S)$$

satisfying

(i) ψ is one-to-one and onto a.e. By this we mean, that if we remove a (suitable) set N_X of measure 0 in X, and a (suitable) set N_Y of measure 0 in Y, the map $\psi : X \setminus N_X \to Y \setminus N_Y$ is a bijection.

(ii) ψ is measurable, i.e., $\psi^{-1}(C) \in \mathcal{F}$, for all $C \in \mathcal{C}$.

(iii) ψ preserves the measures: $\nu = \mu \circ \psi^{-1}$, i.e., $\nu(C) = \mu\left(\psi^{-1}(C)\right)$ for all $C \in \mathcal{C}$.

Finally, we should have that

(iv) ψ preserves the dynamics of T and S, i.e., $\psi \circ T = S \circ \psi$, which is the same as saying that the following diagram commutes.

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

This means that T-orbits are mapped to S-orbits:

$$\begin{array}{ccccc} x \rightarrow & Tx \rightarrow & T^2x \rightarrow & \cdots \rightarrow & T^n x \rightarrow \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \psi(x) \rightarrow & S\left(\psi(x)\right) \rightarrow & S^2\left(\psi(x)\right) \rightarrow & \cdots \rightarrow & S^n\left(\psi(x)\right) \rightarrow \end{array}$$

Definition 3.1.1 *Two dynamical systems* $(X, \mathcal{F}, \mu, T)$ *and* $(Y, \mathcal{C}, \nu, S)$ *are isomorphic if there exist measurable sets* $N \subset X$ *and* $M \subset Y$ *with* $\mu(X \backslash N) = \nu(Y \setminus M) = 0$ *and* $T(N) \subset N$, $S(M) \subset M$, *and finally if there exists a measurable map* $\psi : N \rightarrow M$ *such that* (i)–(iv) *are satisfied.*

Exercise 3.1.1 Suppose $(X, \mathcal{F}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ are two isomorphic dynamical systems. Show that

(a) T is ergodic if and only if S is ergodic.

(b) T is weakly mixing if and only if S is weakly mixing.

(c) T is strongly mixing if and only if S is strongly mixing.

Examples

(1) Let $K = \{z \in \mathbb{C} : |z| = 1\}$ be equipped with the Borel σ-algebra $\mathcal{B}$ on K, and Haar measure (i.e., normalized Lebeque measure on the unit circle).

Define $S : K \to K$ by $Sz = z^2$; equivalently $Se^{2\pi i\theta} = e^{2\pi i(2\theta)}$.. One can easily check that S is measure preserving. In fact, the map S is isomorphic to the map T on $([0,1), \mathcal{B}, \lambda)$ given by $Tx = 2x \pmod 1$ (see Example (b) in subsection 1.3, and Example (3) in subsection 1.8). Define a map $\phi : [0,1) \to K$ by $\phi(x) = e^{2\pi ix}$. We leave it to the reader to check that ϕ is a measurable isomorphism, i.e., ϕ is a measurable and measure preserving bijection such that $S\phi(x) = \phi(Tx)$ for all $x \in [0,1)$.

(2) Consider $([0,1), \mathcal{B}, \lambda)$, the unit interval with the Lebesgue σ-algebra, and Lebesgue measure. Let $T : [0,1) \to [0,1)$ be given by $Tx = Nx - \lfloor Nx \rfloor$. Iterations of T generate the N-adic expansion of points in the unit interval. Let $Y := \{0, 1, \ldots, N-1\}^{\mathbb{N}}$, the set of all sequences $(y_n)_{n\geq 1}$, with $y_n \in \{0, 1, \ldots, N-1\}$ for $n \geq 1$. We now construct an isomorphism between $([0,1), \mathcal{B}, \lambda, T)$ and $(Y, \mathcal{F}, \mu, S)$, where $\mathcal{F}$ is the σ-algebra generated by the cylinders, and μ the uniform product measure defined on cylinders by

$$\mu(\{(y_i)_{i\geq 1} \in Y : y_1 = a_1, y_2 = a_2, \ldots, y_n = a_n\}) = \frac{1}{N^n},$$

for any $(a_1, a_2, a_3, \ldots) \in Y$, and where S is the left shift.

Define $\psi : [0,1) \to Y = \{0, 1, \ldots, N-1\}^{\mathbb{N}}$ by

$$\psi : x = \sum_{k=1}^{\infty} \frac{a_k}{N^k} \mapsto (a_k)_{k\geq 1},$$

where $\sum_{k=1}^{\infty} a_k/N^k$ is the N-adic expansion of x (for example if $N = 2$ we get the binary expansion, and if $N = 10$ we get the decimal expansion). Let

$$C(i_1, \ldots, i_n) = \{(y_i)_{i\geq 1} \in Y : y_1 = i_1, \ldots, y_n = i_n\}.$$

In order to see that ψ is an isomorphism one needs to verify measurability and measure preservingness on cylinders:

$$\psi^{-1}(C(i_1, \ldots, i_n)) = [\frac{i_1}{N} + \frac{i_2}{N^2} + \cdots + \frac{i_n}{N^n}, \frac{i_1}{N} + \frac{i_2}{N^2} + \cdots + \frac{i_n + 1}{N^n})$$

and

$$\lambda\left(\psi^{-1}(C(i_1, \ldots, i_n))\right) = \frac{1}{N^n} = \mu(C(i_1, \ldots, i_n))).$$

Note that

$$\mathcal{N} = \{(y_i)_{i\geq 1} \in Y : \text{ there exists a } k \geq 1 \text{ such that } y_i = N-1 \text{ for all } i \geq k\}$$

is a subset of Y of measure 0. Setting $\tilde{Y} = Y \setminus \mathcal{N}$, then $\psi : [0,1) \to \tilde{Y}$ is a bijection, since every $x \in [0,1)$ has a unique N-adic expansion **generated** by T. Finally, it is easy to see that $\psi \circ T = S \circ \psi$.

Exercise 3.1.2 Consider $([0,1)^2, \mathcal{B} \times \mathcal{B}, \lambda \times \lambda)$, where $\mathcal{B} \times \mathcal{B}$ is the product Lebesgue σ-algebra, and $\lambda \times \lambda$ is the product Lebesgue measure Let $T : [0,1)^2 \to [0,1)^2$ be given by

$$T(x,y) = \begin{cases} (2x, \dfrac{1}{2}y), & 0 \le x < \frac{1}{2} \\ \\ (2x-1, \dfrac{1}{2}(y+1)), & \frac{1}{2} \le x < 1. \end{cases}$$

Show that T is isomorphic to the two-sided Bernoulli shift S on $\big(\{0,1\}^{\mathbb{Z}}, \mathcal{F}, \mu\big)$, where $\mathcal{F}$ is the σ-algebra generated by cylinders of the form

$$\Delta = \{x_{-k} = a_{-k}, \ldots, x_\ell = a_\ell \,:\, a_i \in \{0,1\},\ i = -k, \ldots, \ell\}, \quad k, \ell \ge 0,$$

and μ the product measure with weights $(\frac{1}{2}, \frac{1}{2})$ (so $\mu(\Delta) = (\frac{1}{2})^{k+\ell+1}$).

Exercise 3.1.3 Let $G = \dfrac{1+\sqrt{5}}{2}$, so that $G^2 = G + 1$. Consider the set

$$X = [0, \frac{1}{G}) \times [0,1) \bigcup [\frac{1}{G}, 1) \times [0, \frac{1}{G}),$$

endowed with the product Borel σ-algebra. Define the transformation

$$\mathcal{T}(x,y) = \begin{cases} (Gx, \dfrac{y}{G}), & (x,y) \in [0, \frac{1}{G}) \times [0,1] \\ \\ (Gx - 1, \dfrac{1+y}{G}), & (x,y) \in [\frac{1}{G}, 1) \times [0, \frac{1}{G}). \end{cases}$$

(a) Show that $\mathcal{T}$ is measure preserving with respect to normalized Lebesgue measure on X.

(b) Now let $\mathcal{S} : [0,1) \times [0,1) \to [0,1) \times [0,1)$ be given by

$$\mathcal{S}(x,y) = \begin{cases} (Gx, \dfrac{y}{G}), & (x,y) \in [0, \frac{1}{G}) \times [0,1] \\ \\ (G^2x - G, \dfrac{G+y}{G^2}), & (x,y) \in [\frac{1}{G}, 1) \times [0,1). \end{cases}$$

Show that $\mathcal{S}$ is measure preserving with respect to normalized Lebesgue measure on $[0,1)\times[0,1)$.

(c) Let $Y=[0,1)\times[0,\frac{1}{G})$, and let U be the induced transformation of $\mathcal{T}$ on Y, i.e., for $(x,y)\in Y$, $U(x,y)=\mathcal{T}^{n(x,y)}$, where $n(x,y)=\inf\{n\geq 1:\mathcal{T}^n(x,y)\in Y\}$. Show that the map $\phi:[0,1)\times[0,1)\to Y$ given by

$$\phi(x,y)=(x,\frac{y}{G})$$

defines an isomorphism from $\mathcal{S}$ to U, where Y has the induced measure structure (see Section 1.5).

3.2 Factor Maps

In the above section, we discussed the notion of isomorphism which describes when two dynamical systems are considered the same. Now, we give a precise definition of what it means for a dynamical system to be a subsystem of another one.

Definition 3.2.1 *Let $(X,\mathcal{F},\mu,T)$ and $(Y,\mathcal{C},\nu,S)$ be two dynamical systems. We say that S is a factor of T if there exist measurable sets $M_1\in\mathcal{F}$ and $M_2\in\mathcal{C}$, such that $\mu(M_1)=\nu(M_2)=1$ and $T(M_1)\subset M_1$, $S(M_2)\subset M_2$, and finally if there exists a measurable and measure preserving map $\psi:M_1\to M_2$ which is surjective, and satisfies $\psi(T(x))=S(\psi(x))$ for all $x\in M_1$. We call ψ a factor map.*

Remark Notice that if ψ is a factor map, then $\mathcal{G}=\psi^{-1}\mathcal{C}$ is a T-invariant sub-σ-algebra of $\mathcal{F}$, since

$$T^{-1}\mathcal{G}=T^{-1}\psi^{-1}\mathcal{C}=\psi^{-1}S^{-1}\mathcal{C}\subseteq\psi^{-1}\mathcal{C}=\mathcal{G}.$$

Examples Let T be the Baker's transformation on $([0,1)^2,\mathcal{B}\times\mathcal{B},\lambda\times\lambda)$, given by

$$T(x,y)=\begin{cases}(2x,\frac{1}{2}y), & 0\leq x<\frac{1}{2}\\ \\ (2x-1,\frac{1}{2}(y+1)), & \frac{1}{2}\leq x<1,\end{cases}$$

and let S be the left shift on $X = \{0,1\}^{\mathbb{N}}$ with the σ-algebra $\mathcal{F}$ generated by the cylinders, and the uniform product measure μ. Define $\psi : [0,1)\times[0,1) \to X$ by

$$\psi(x,y) = (a_1, a_2, \ldots),$$

where $x = \sum_{n=1}^{\infty} \dfrac{a_n}{2^n}$ is the binary expansion of x. It is easy to check that ψ is a factor map.

Exercise 3.2.1 Let T be the left shift on $X = \{0,1,2\}^{\mathbb{N}}$ which is endowed with the σ-algebra $\mathcal{F}$, generated by the cylinder sets, and the uniform product measure μ giving each symbol probability 1/3, i.e.,

$$\mu\left(\{x \in X : x_1 = i_1, x_2 = i_2, \ldots, x_n = i_n\}\right) = \frac{1}{3^n},$$

where $i_1, i_2, \ldots, i_n \in \{0,1,2\}$.
Let S be the left shift on $Y = \{0,1\}^{\mathbb{N}}$ which is endowed with the σ-algebra $\mathcal{G}$, generated by the cylinder sets, and the product measure ν giving the symbol 0 probability 1/3 and the symbol 1 probability 2/3, i.e.,

$$\mu\left(\{y \in Y : y_1 = j_1, y_2 = j_2, \ldots, y_n = j_n\}\right) = (\frac{2}{3})^{j_1+j_2+\ldots+j_n}(\frac{1}{3})^{n-(j_1+j_2+\ldots+j_n)},$$

where $j_1, j_2, \ldots, j_n \in \{0,1\}$. Show that S is a factor of T.

Exercise 3.2.2 Show that a factor of an ergodic (weakly mixing/strongly mixing) transformation is also ergodic (weakly mixing/strongly mixing).

3.3 Natural Extensions

Suppose $(Y, \mathcal{G}, \nu, S)$ is a *non-invertible* measure-preserving dynamical system. An invertible measure-preserving dynamical system $(X, \mathcal{F}, \mu, T)$ is called a *natural extension* of $(Y, \mathcal{G}, \nu, S)$ if S is a factor of T and the factor map ψ satisfies $\vee_{m=0}^{\infty} T^m \psi^{-1}\mathcal{G} = \mathcal{F}$, where

$$\bigvee_{k=0}^{\infty} T^k \psi^{-1}\mathcal{G}$$

is the smallest σ-algebra containing the σ-algebras $T^k\psi^{-1}\mathcal{G}$ for all $k \geq 0$.

Example Let T on $\left(\{0,1\}^{\mathbb{Z}}, \mathcal{F}, \mu\right)$ be the two-sided Bernoulli shift, and S on $\left(\{0,1\}^{\mathbb{N}\cup\{0\}}, \mathcal{G}, \nu\right)$ be the one-sided Bernoulli shift, both spaces are endowed with the uniform product measure. Notice that T is invertible, while S is not. Set $X = \{0,1\}^{\mathbb{Z}}$, $Y = \{0,1\}^{\mathbb{N}\cup\{0\}}$, and define $\psi : X \to Y$ by

$$\psi\left(\ldots, x_{-1}, x_0, x_1, \ldots\right) \;=\; \left(x_0, x_1, \ldots\right).$$

Then, ψ is a factor map. We claim that

$$\bigvee_{k=0}^{\infty} T^k \psi^{-1}\mathcal{G} \;=\; \mathcal{F}.$$

To prove this, we show that $\bigvee_{k=0}^{\infty} T^k \psi^{-1}\mathcal{G}$ contains all cylinders generating $\mathcal{F}$.

Let $\Delta = \{x \in X \,:\, x_{-k} = a_{-k}, \ldots, x_\ell = a_\ell\}$ be an arbitrary cylinder in $\mathcal{F}$, and let $D = \{y \in Y \,:\, y_0 = a_{-k}, \ldots, y_{k+\ell} = a_\ell\}$ which is a cylinder in $\mathcal{G}$. Then,

$$\psi^{-1}D = \{x \in X \,:\, x_0 = a_{-k}, \ldots, x_{k+\ell} = a_\ell\} \quad \text{and} \quad T^k\psi^{-1}D = \Delta.$$

This shows that

$$\bigvee_{k=0}^{\infty} T^k \psi^{-1}\mathcal{G} = \mathcal{F}.$$

Thus, T is the natural extension of S.

Chapter 4

Entropy

4.1 Randomness and Information

Given a measure preserving transformation T on a probability space $(X, \mathcal{F}, \mu)$, we want to define a nonnegative quantity $h(T)$ which measures the average uncertainty about where T moves the points of X. That is, the value of $h(T)$ reflects the amount of 'randomness' generated by T. We want to define $h(T)$ in such a way, that (i) the amount of information gained by an application of T is proportional to the amount of uncertainty removed, and (ii) that $h(T)$ is isomorphism invariant, so that isomorphic transformations have equal entropy.

The connection between entropy (that is randomness, uncertainty) and the transmission of information was first studied by Claude Shannon in 1948. As a motivation let us look at the following simple example. Consider a source (for example a ticker-tape) that produces a string of symbols $\cdots x_{-1}x_0x_1\cdots$ from the alphabet $\{a_1, a_2, \ldots, a_n\}$. Suppose that the probability of receiving symbol a_i at any given time is p_i, and that each symbol is transmitted independently of what has been transmitted earlier. Of course we must have here that each $p_i \geq 0$ and that $\sum_i p_i = 1$. In ergodic theory we view this process as the dynamical system $(X, \mathcal{F}, \mathcal{B}, \mu, T)$, where $X = \{a_1, a_2, \ldots, a_n\}^{\mathbb{N}}$, $\mathcal{B}$ the σ-algebra generated by cylinder sets of the form

$$\Delta_n(a_{i_1}, a_{i_2}, \ldots, a_{i_n}) := \{x \in X : x_{i_1} = a_{i_1}, \ldots, x_{i_n} = a_{i_n}\}$$

μ the product measure assigning to each coordinate probability p_i of seeing the symbol a_i, and T the left shift. We define the entropy of this system by

$$H(p_1, \ldots, p_n) = h(T) := -\sum_{i=1}^{n} p_i \log_2 p_i \,. \tag{4.1}$$

If we define $\log p_i$ as the amount of uncertainty in transmitting the symbol a_i, then H is the average amount of information gained (or uncertainty removed) per symbol (notice that H is in fact an expected value). To see why this is an appropriate definition, notice that if the source is degenerate, that is, $p_i = 1$ for some i (i.e., the source only transmits the symbol a_i), then $H = 0$. In this case we indeed have no randomness. Another reason to see why this definition is appropriate, is that H is maximal if $p_i = \frac{1}{n}$ for all i, and this agrees with the fact that the source is most random when all the symbols are equiprobable. To see this maximum, consider the function $f : [0,1] \to \mathbb{R}_+$ defined by

$$f(t) = \begin{cases} 0 & \text{if } t = 0, \\ -t \log_2 t & \text{if } 0 < t \leq 1. \end{cases}$$

Then f is continuous and concave downward, and Jensen's Inequality implies that for any $p_1, \ldots, p_n$ with $p_i \geq 0$ and $p_1 + \ldots + p_n = 1$,

$$\frac{1}{n} H(p_1, \ldots, p_n) = \frac{1}{n} \sum_{i=1}^{n} f(p_i) \leq f(\frac{1}{n} \sum_{i=1}^{n} p_i) = f(\frac{1}{n}) = \frac{1}{n} \log_2 n \,,$$

so $H(p_1, \ldots, p_n) \leq \log_2 n$ for all probability vectors $(p_1, \ldots, p_n)$. But

$$H(\frac{1}{n}, \ldots, \frac{1}{n}) = \log_2 n,$$

so the maximum value is attained at $(\frac{1}{n}, \ldots, \frac{1}{n})$.

4.2 Definitions and Properties

So far H is defined as the average information per symbol. The above definition can be extended to define the information transmitted by the occurrence of an event E as $-\log_2 P(E)$. This definition has the property that the information transmitted by $E \cap F$ for independent events E and F is the sum of the information transmitted by each one individually, i.e.,

$$-\log_2 P(E \cap F) = -\log_2 P(E) - \log_2 P(F).$$

The only function with this property is the logarithm function to any base. We choose base 2 because information is usually measured in bits.

In the above example of the ticker-tape the symbols were transmitted independently. In general, the symbol generated might depend on what has been received before. In fact these dependencies are often 'built-in' to be able to check the transmitted sequence of symbols on errors (think here of the Morse sequence, sequences on compact discs etc.). Such dependencies must be taken into consideration in the calculation of the average information per symbol. This can be achieved if one replaces the symbols a_i by blocks of symbols of particular size. More precisely, for every n, let $\mathcal{C}_n$ be the collection of all possible n-blocks (or cylinder sets) of length n, and define

$$H_n := -\sum_{C \in \mathcal{C}_n} P(C) \log P(C) \, .$$

Then $\frac{1}{n} H_n$ can be seen as the average information per symbol when a block of length n is transmitted. The entropy of the source is now defined by

$$h := \lim_{n \to \infty} \frac{H_n}{n} \, . \tag{4.2}$$

The existence of the limit in (4.2) follows from the fact that H_n is a *subadditive sequence*, i.e., $H_{n+m} \leq H_n + H_m$, and proposition (4.2.2) (see proposition (4.2.3) below).

Now replace the source by a measure preserving system $(X, \mathcal{B}, \mu, T)$. How can one define the entropy of this system similar to the case of a source? The symbols $\{a_1, a_2, \ldots, a_n\}$ can now be viewed as a partition $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ of X, so that X is the disjoint union (up to sets of measure zero) of $A_1, A_2, \ldots, A_n$. The source can be seen as follows: with each point $x \in X$, we associate an infinite sequence $\cdots x_{-1}, x_0, x_1, \cdots$, where x_i is a_j if and only if $T^i x \in A_j$. We define the *entropy of the partition* α by

$$H(\alpha) = H_\mu(\alpha) := -\sum_{i=1}^{n} \mu(A_i) \log \mu(A_i) \, .$$

Our aim is to define the entropy of the transformation T which is independent of the partition we choose. In fact $h(T)$ must be the maximal entropy over all possible finite partitions. But first we need few facts about partitions.

Exercise 4.2.1 Let $\alpha = \{A_1, \ldots, A_n\}$ and $\beta = \{B_1, \ldots, B_m\}$ be two partitions of X. Show that

$$T^{-1}\alpha := \{T^{-1}A_1, \ldots, T^{-1}A_n\}$$

and

$$\alpha \vee \beta := \{A_i \cap B_j : A_i \in \alpha, B_j \in \beta\}$$

are both partitions of X.

The members of a partition are called the *atoms* of the partition. We say that the partition $\beta = \{B_1, \ldots, B_m\}$ is a *refinement* of the partition $\alpha = \{A_1, \ldots, A_n\}$, and write $\alpha \leq \beta$, if for every $1 \leq j \leq m$ there exists an $1 \leq i \leq n$ such that $B_j \subset A_i$ (up to sets of measure zero). The partition $\alpha \vee \beta$ is called the *common refinement* of α and β.

Exercise 4.2.2 Show that if β is a refinement of α, each atom of α is a finite (disjoint) union of atoms of β.

Given two partitions $\alpha = \{A_1, \ldots A_n\}$ and $\beta = \{B_1, \ldots, B_m\}$ of X, we define the *conditional entropy of* α *given* β by

$$H(\alpha|\beta) := -\sum_{A\in\alpha}\sum_{B\in\beta} \log\left(\frac{\mu(A\cap B)}{\mu(B)}\right)\mu(A\cap B)\,.$$

(Under the convention that $0 \log 0 := 0$.)

The above quantity $H(\alpha|\beta)$ is interpreted as the average uncertainty about which element of the partition α the point x will enter (under T) if we already know which element of β the point x will enter.

Proposition 4.2.1 *Let α, β and γ be partitions of X. Then,*

(a) $H(T^{-1}\alpha) = H(\alpha)$;

(b) $H(\alpha \vee \beta) = H(\alpha) + H(\beta|\alpha)$;

(c) $H(\beta|\alpha) \leq H(\beta)$;

(d) $H(\alpha \vee \beta) \leq H(\alpha) + H(\beta)$;

(e) If $\alpha \leq \beta$, *then* $H(\alpha) \leq H(\beta)$;

(f) $H(\alpha \vee \beta | \gamma) \;=\; H(\alpha|\gamma) + H(\beta|\alpha \vee \gamma)$*;*

(g) If $\beta \leq \alpha$*, then* $H(\gamma|\alpha) \leq H(\gamma|\beta)$*;*

(h) If $\beta \leq \alpha$*, then* $H(\beta|\alpha) \;=\; 0$*.*

(i) We call two partitions α *and* β independent *if*

$$\mu(A \cap B) \;=\; \mu(A)\mu(B) \quad \textit{for all} \quad A \in \alpha,\, B \in \beta\,.$$

If α *and* β *are independent partitions, one has that*

$$H(\alpha \vee \beta) \;=\; H(\alpha) + H(\beta)\,.$$

Proof We prove properties (b) and (c), the rest are left as an exercise.

$$\begin{aligned}
H(\alpha \vee \beta) &= -\sum_{A\in\alpha}\sum_{B\in\beta} \mu(A\cap B)\log \mu(A\cap B) \\
&= -\sum_{A\in\alpha}\sum_{B\in\beta} \mu(A\cap B)\log \frac{\mu(A\cap B)}{\mu(A)} \\
&+ -\sum_{A\in\alpha}\sum_{B\in\beta} \mu(A\cap B)\log \mu(A) \\
&= H(\beta|\alpha) + H(\alpha).
\end{aligned}$$

We now show that $H(\beta|\alpha) \leq H(\beta)$. Recall that the function $f(t) = -t\log t$ for $0 < t \leq 1$ is concave down. Thus,

$$\begin{aligned}
H(\beta|\alpha) &= -\sum_{B\in\beta}\sum_{A\in\alpha} \mu(A\cap B)\log \frac{\mu(A\cap B)}{\mu(A)} \\
&= -\sum_{B\in\beta}\sum_{A\in\alpha} \mu(A)\frac{\mu(A\cap B)}{\mu(A)}\log \frac{\mu(A\cap B)}{\mu(A)} \\
&= \sum_{B\in\beta}\sum_{A\in\alpha} \mu(A) f\left(\frac{\mu(A\cap B)}{\mu(A)}\right) \\
&\leq \sum_{B\in\beta} f\left(\sum_{A\in\alpha} \mu(A)\frac{\mu(A\cap B)}{\mu(A)}\right) \\
&= \sum_{B\in\beta} f(\mu(B)) = H(\beta).
\end{aligned}$$

□

Exercise 4.2.3 Prove the rest of the properties of Proposition 4.2.1

Now consider the partition $\bigvee_{i=0}^{n-1} T^{-i}\alpha$, whose atoms are of the form $A_{i_0} \cap T^{-1}A_{i_1} \cap \ldots \cap T^{-(n-1)}A_{i_{n-1}}$, consisting of all points $x \in X$ with the property that $x \in A_{i_0}$, $Tx \in A_{i_1}$, ... , $T^{n-1}x \in A_{i_{n-1}}$.

Exercise 4.2.4 Show that if α is a finite partition of $(X, \mathcal{F}, \mu, T)$, then

$$H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) = H(\alpha) + \sum_{j=1}^{n-1} H(\alpha | \bigvee_{i=1}^{j} T^{-i}\alpha).$$

To define the notion of the entropy of a transformation with respect to a partition, we need the following two propositions.

Proposition 4.2.2 *If $\{a_n\}$ is a subadditive sequence of real numbers i.e., $a_{n+p} \leq a_n + a_p$ for all n, p, then*

$$\lim_{n\to\infty} \frac{a_n}{n}$$

exists.

Proof Fix any $m > 0$. For any $n \geq 0$ one has $n = km + i$ for some i between $0 \leq i \leq m-1$. By subadditivity it follows that

$$\frac{a_n}{n} = \frac{a_{km+i}}{km+i} \leq \frac{a_{km}}{km} + \frac{a_i}{km} \leq k\frac{a_m}{km} + \frac{a_i}{km}.$$

Note that if $n \to \infty$, $k \to \infty$ and so $\limsup_{n\to\infty} a_n/n \leq a_m/m$. Since m is arbitrary one has

$$\limsup_{n\to\infty} \frac{a_n}{n} \leq \inf \frac{a_m}{m} \leq \liminf_{n\to\infty} \frac{a_n}{n}.$$

Therefore $\lim_{n\to\infty} a_n/n$ exists, and equals $\inf a_n/n$. □

Proposition 4.2.3 *Let α be a finite partitions of $(X, \mathcal{B}, \mu, T)$, where T is a measure preserving transformation. Then, $\lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha)$ exists.*

Proof Let $a_n = H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) \geq 0$. Then, by Proposition 4.2.1, we have

$$\begin{aligned} a_{n+p} &= H(\bigvee_{i=0}^{n+p-1} T^{-i}\alpha) \\ &\leq H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) + H(\bigvee_{i=n}^{n+p-1} T^{-i}\alpha) \\ &= a_n + H(\bigvee_{i=0}^{p-1} T^{-i}\alpha) \\ &= a_n + a_p. \end{aligned}$$

Hence, by Proposition 4.2.2

$$\lim_{n\to\infty} \frac{a_n}{n} = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha)$$

exists. □

We are now in position to give the definition of the entropy of the transformation T.

Definition 4.2.1 *The entropy of the measure preserving transformation T with respect to the partition α is given by*

$$h(\alpha, T) \,=\, h_\mu(\alpha, T) := \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha)\,,$$

where

$$H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) \,=\, - \sum_{D \in \bigvee_{i=0}^{n-1} T^{-i}\alpha} \mu(D) \log(\mu(D))\,.$$

Finally, the entropy of the transformation T is given by

$$h(T) \,=\, h_\mu(T) \,:=\, \sup_\alpha h(\alpha, T)\,.$$

The following theorem gives an equivalent definition of entropy..

Theorem 4.2.1 *The entropy of the measure preserving transformation T with respect to the partition α is also given by*

$$h(\alpha, T) = \lim_{n\to\infty} H(\alpha | \bigvee_{i=1}^{n-1} T^{-i}\alpha).$$

Proof Notice that the sequence $\{H(\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha)$ is bounded from below, and is non-increasing, hence $\lim_{n\to\infty} H(\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha)$ exists. Furthermore,

$$\lim_{n\to\infty} H(\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} H(\alpha|\bigvee_{i=1}^{j} T^{-i}\alpha).$$

From exercise 4.2.4, we have

$$H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) = H(\alpha) + \sum_{j=1}^{n-1} H(\alpha|\bigvee_{i=1}^{j} T^{-i}\alpha).$$

Now, dividing by n, and taking the limit as $n \to \infty$, one gets the desired result $\square$

Theorem 4.2.2 *Entropy is an isomorphism invariant.*

Proof Let $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ be two isomorphic measure preserving systems (see Definition 1.2.3, for a definition), with $\psi : X \to Y$ the corresponding isomorphism. We need to show that $h_\mu(T) = h_\nu(S)$.

Let $\beta = \{B_1, \ldots, B_n\}$ be any partition of Y, then $\psi^{-1}\beta = \{\psi^{-1}B_1, \ldots, \psi^{-1}B_n\}$ is a partition of X. Set $A_i = \psi^{-1}B_i$, for $1 \le i \le n$. Since $\psi : X \to Y$ is an isomorphism, we have that $\nu = \mu\psi^{-1}$ and $\psi T = S\psi$, so that for any $n \ge 0$ and $B_{i_0}, \ldots, B_{i_{n-1}} \in \beta$

$$\begin{aligned}
& \nu\left(B_{i_0} \cap S^{-1}B_{i_1} \cap \ldots \cap S^{-(n-1)}B_{i_{n-1}}\right) \\
= \;& \mu\left(\psi^{-1}B_{i_0} \cap \psi^{-1}S^{-1}B_{i_1} \cap \ldots \cap \psi^{-1}S^{-(n-1)}B_{i_{n-1}}\right) \\
= \;& \mu\left(\psi^{-1}B_{i_0} \cap T^{-1}\psi^{-1}B_{i_1} \cap \ldots \cap T^{-(n-1)}\psi^{-1}B_{i_{n-1}}\right) \\
= \;& \mu\left(A_{i_0} \cap T^{-1}A_{i_1} \cap \ldots \cap T^{-(n-1)}A_{i_{n-1}}\right) .
\end{aligned}$$

Setting

$$A(n) = A_{i_0} \cap \ldots \cap T^{-(n-1)}A_{i_{n-1}} \text{ and } B(n) = B_{i_0} \cap \ldots \cap S^{-(n-1)}B_{i_{n-1}},$$

we thus find that

$$\begin{aligned} h_\nu(S) &= \sup_\beta h_\nu(\beta, S) = \sup_\beta \lim_{n\to\infty} \frac{1}{n} H_\nu(\bigvee_{i=0}^{n-1} S^{-i}\beta) \\ &= \sup_\beta \lim_{n\to\infty} -\frac{1}{n} \sum_{B(n)\in\bigvee_{i=0}^{n-1} S^{-i}\beta} \nu(B(n)) \log \nu(B(n)) \\ &= \sup_{\psi^{-1}\beta} \lim_{n\to\infty} -\frac{1}{n} \sum_{A(n)\in\bigvee_{i=0}^{n-1} T^{-i}\psi^{-1}\beta} \mu(A(n)) \log \mu(A(n)) \\ &= \sup_{\psi^{-1}\beta} h_\mu(\psi^{-1}\beta, T) \\ &\le \sup_\alpha h_\mu(\alpha, T) = h_\mu(T)\,, \end{aligned}$$

where in the last inequality the supremum is taken over all possible finite partitions α of X. Thus $h_\nu(S) \le h_\mu(T)$. The proof of $h_\mu(T) \le h_\nu(S)$ is done similarly. Therefore $h_\nu(S) = h_\mu(T)$, and the proof is complete. □

4.3 Calculation of Entropy and Examples

Calculating the entropy of a transformation directly from the definition does not seem feasible, for one needs to take the supremum over **all** finite partitions, which is practically impossible. However, the entropy of a partition is relatively easier to calculate if one has full information about the partition under consideration. So the question is whether it is possible to find a partition α of X where $h(\alpha, T) = h(T)$. Naturally, such a partition contains all the information 'transmitted' by T. To answer this question we need some notations and definitions.

For $\alpha = \{A_1, \ldots, A_N\}$ and all m, $n \ge 0$, let

$$\sigma\left(\bigvee_{i=n}^{m} T^{-i}\alpha\right) \text{ and } \sigma\left(\bigvee_{i=-m}^{-n} T^{-i}\alpha\right)$$

be the smallest σ-algebras containing the partitions $\bigvee_{i=n}^{m} T^{-i}\alpha$ and $\bigvee_{i=-m}^{-n} T^{-i}\alpha$ respectively. Furthermore, let $\sigma\left(\bigvee_{i=-\infty}^{-\infty} T^{-i}\alpha\right)$ be the smallest σ-algebra containing all the partitions $\bigvee_{i=n}^{m} T^{-i}\alpha$ and $\bigvee_{i=-m}^{-n} T^{-i}\alpha$ for all n and m. We call a partition α a *generator* with respect to T if $\sigma\left(\bigvee_{i=-\infty}^{\infty} T^{-i}\alpha\right) = \mathcal{F}$, where $\mathcal{F}$ is the σ-algebra on X. If T is non-invertible, then α is said to be a generator if $\sigma\left(\bigvee_{i=0}^{\infty} T^{-i}\alpha\right) = \mathcal{F}$. Naturally, this equality is modulo sets of measure zero. One has also the following characterization of generators, saying basically, that each measurable set in X can be approximated by a finite disjoint union of cylinder sets. See also [W] for more details and proofs.

Proposition 4.3.1 *The partition α is a generator of $\mathcal{F}$ if for each $A \in \mathcal{F}$ and for each $\varepsilon > 0$ there exists a finite disjoint union C of elements of $\{\alpha_n^m\}$, such that $\mu(A \triangle C) < \varepsilon$.*

We now state (without proofs) two famous theorems known as *Kolmogorov-Sinai's Theorem* and *Krieger's Generator Theorem.* For the proofs, we refer the interested reader to the book of Karl Petersen or Peter Walter.

Theorem 4.3.1 (Kolmogorov and Sinai, 1958) *If α is a generator with respect to T and $H(\alpha) < \infty$, then $h(T) = h(\alpha, T)$.*

Theorem 4.3.2 (Krieger, 1970) *If T is an ergodic measure preserving transformation with $h(T) < \infty$, then T has a finite generator.*

We will use these two theorems to calculate the entropy of a Bernoulli shift.

Example (Entropy of a Bernoulli shift)–Let T be the left shift on $X = \{1, 2, \cdots, n\}^{\mathbb{Z}}$ endowed with the σ-algebra $\mathcal{F}$ generated by the cylinder sets, and product measure μ giving symbol i probability p_i, where $p_1 + p_2 + \ldots + p_n = 1$. Our aim is to calculate $h(T)$. To this end we need to find a partition α which generates the σ-algebra $\mathcal{F}$ under the action of T. The natural choice of α is what is known as the *time-zero partition* $\alpha = \{A_1, \ldots, A_n\}$, where

$$A_i := \{x \in X : x_0 = i\}, \ i = 1, \ldots, n.$$

Notice that for all $m \in \mathbb{Z}$,

$$T^{-m}A_i = \{x \in X : x_m = i\},$$

and

$$A_{i_0} \cap T^{-1}A_{i_1} \cap \cdots \cap T^{-m}A_{i_m} = \{x \in X : x_0 = i_0, \ldots, x_m = i_m\}.$$

In other words, $\bigvee_{i=0}^{m} T^{-i}\alpha$ is precisely the collection of cylinders of length m (i.e., the collection of all m-blocks), and these by definition generate $\mathcal{F}$. Hence, α is a generating partition, so that

$$h(T) \;=\; h(\alpha, T) \;=\; \lim_{m\to\infty} \frac{1}{m} H\left(\bigvee_{i=0}^{m-1} T^{-i}\alpha\right) .$$

First notice that $-$ since μ is product measure here $-$ the partitions

$$\alpha,\, T^{-1}\alpha,\, \cdots ,\, T^{-(m-1)}\alpha$$

are all independent since each specifies a different coordinate, and so

$$\begin{aligned} & H(\alpha \vee T^{-1}\alpha \vee \cdots \vee T^{-(m-1)}\alpha) \\ = \;& H(\alpha) + H(T^{-1}\alpha) + \cdots + H(T^{-(m-1)}\alpha) \\ = \;& mH(\alpha) \;=\; -m \sum_{i=1}^{n} p_i \log p_i . \end{aligned}$$

Thus,

$$h(T) \;=\; \lim_{m\to\infty} \frac{1}{m}(-m) \sum_{i=1}^{n} p_i \log p_i \;=\; -\sum_{i=1}^{n} p_i \log p_i \, .$$

Exercise 4.3.1 Let T be the left shift on $X = \{1, 2, \cdots , n\}^{\mathbb{Z}}$ endowed with the σ-algebra $\mathcal{F}$ generated by the cylinder sets, and the Markov measure μ given by the stochastic matrix $P = (p_{ij})$, and the probability vector $\pi = (\pi_1, \ldots , \pi_n)$ with $\pi P = \pi$. Show that

$$h(T) = -\sum_{j=1}^{n} \sum_{i=1}^{n} \pi_i p_{ij} \log p_{ij}$$

Exercise 4.3.2 *Suppose $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ are two dynamical systems. Show that*

$$h_{\mu_1\times\mu_2}(T_1 \times T_2) = h_{\mu_1}(T_1) + h_{\mu_2}(T_2).$$

4.4 The Shannon-McMillan-Breiman Theorem

In the previous sections we have considered only finite partitions on X, however all the definitions and results hold if we were to consider countable partitions of finite entropy. Before we state and prove the Shannon-McMillan-Breiman Theorem, we need to introduce the information function associated with a partition.

Let $(X, \mathcal{F}, \mu)$ be a probability space, and $\alpha = \{A_1, A_2, \ldots\}$ be a finite or a countable partition of X into measurable sets. For each $x \in X$, let $\alpha(x)$ be the element of α to which x belongs. Then, the *information function* associated to α is defined to be

$$I_\alpha(x) = -\log \mu(\alpha(x)) = -\sum_{A\in\alpha} 1_A(x) \log \mu(A).$$

For two finite or countable partitions α and β of X, we define the *conditional information function* of α given β by

$$I_{\alpha|\beta}(x) = -\sum_{B\in\beta}\sum_{A\in\alpha} 1_{(A\cap B)}(x) \log\left(\frac{\mu(A\cap B)}{\mu(B)}\right).$$

We claim that

$$I_{\alpha|\beta}(x) = -\log E_\mu(1_{\alpha(x)}|\sigma(\beta)) = -\sum_{A\in\alpha} 1_A(x) \log E(1_A|\sigma(\beta)), \tag{4.3}$$

where $\sigma(\beta)$ is the σ-algebra generated by the finite or countable partition β, (see the remark following the proof of Theorem (2.1.1)). This follows from the fact (which is easy to prove using the definition of conditional expectations) that if β is finite or countable, then for any $f \in L^1(\mu)$, one has

$$E_\mu(f|\sigma(\beta)) = \sum_{B\in\beta} 1_B \frac{1}{\mu(B)} \int_B f\, d\mu.$$

Clearly, $H(\alpha|\beta) = \int_X I_{\alpha|\beta}(x)\, \mathrm{d}\mu(x)$.

Exercise 4.4.1 Let α and β be finite or countable partitions of X. Show that

$$I_{\alpha\bigvee\beta} = I_\alpha + I_{\beta|\alpha}.$$

Now suppose $T : X \to X$ is a measure preserving transformation on $(X, \mathcal{F}, \mu)$, and let $\alpha = \{A_1, A_2, \ldots\}$ be any countable partition. Then $T^{-1} = \{T^{-1}A_1, T^{-1}A_2, \ldots\}$ is also a countable partition. Since T is measure preserving one has,

$$I_{T^{-1}\alpha}(x) = -\sum_{A_i \in \alpha} 1_{T^{-1}A_i}(x) \log \mu(T^{-1}A_i) = -\sum_{A_i \in \alpha} 1_{A_i}(Tx) \log \mu(A_i) = I_\alpha(Tx).$$

Furthermore,

$$\lim_{n\to\infty} \frac{1}{n+1} H(\bigvee_{i=0}^{n} T^{-i}\alpha) = \lim_{n\to\infty} \frac{1}{n+1} \int_X I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x)\, \mathrm{d}\mu(x) = h(\alpha, T).$$

The Shannon-McMillan-Breiman theorem says if T is ergodic and if α has finite entropy, then in fact the integrand $\dfrac{1}{n+1} I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x)$ converges a.e. to $h(\alpha, T)$. Notice that the integrand can be written as

$$\frac{1}{n+1} I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x) = -\frac{1}{n+1} \log \mu \left((\bigvee_{i=0}^{n} T^{-i}\alpha)(x) \right),$$

where $(\bigvee_{i=0}^{n} T^{-i}\alpha)(x)$ is the element of $\bigvee_{i=0}^{n} T^{-i}\alpha$ containing x (often referred to as the α-cylinder of order n containing x). Before we proceed we need the following proposition.

Proposition 4.4.1 *Let $\alpha = \{A_1, A_2, \ldots\}$ be a countable partition with finite entropy. For each $n = 1, 2, 3, \ldots$, let $f_n(x) = I_{\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha}(x)$, and let $f^* = \sup_{n\geq 1} f_n$. Then, for each $\lambda \geq 0$ and for each $A \in \alpha$,*

$$\mu\left(\{x \in A : f^*(x) > \lambda\}\right) \leq 2^{-\lambda}.$$

Furthermore, $f^* \in L^1(X, \mathcal{F}, \mu)$.

Proof Let $t \geq 0$ and $A \in \alpha$. For $n \geq 1$, let

$$f_n^A(x) = -\log E_\mu \left(1_A | \bigvee_{i=1}^{n} T^{-i}\alpha \right)(x),$$

and

$$B_n = \{x \in X : f_1^A(x) \leq t, \ldots, f_{n-1}^A(x) \leq t, f_n^A(x) > t\}.$$

Notice that for $x \in A$ one has $f_n(x) = f_n^A(x)$, and for $x \in B_n$ one has $E_\mu\left(1_A | \bigvee_{i=1}^n T^{-i}\alpha\right)(x) < 2^{-t}$. Since $B_n \in \sigma\left(\bigvee_{i=1}^n T^{-i}\alpha\right)$, then

$$\begin{aligned}
\mu(B_n \cap A) &= \int_{B_n} 1_A(x)\,\mathrm{d}\mu(x) \\
&= \int_{B_n} E_\mu\left(1_A | \bigvee_{i=1}^n T^{-i}\alpha\right)(x)\,\mathrm{d}\mu(x) \\
&\le \int_{B_n} 2^{-t}\,\mathrm{d}\mu(x) = 2^{-t}\mu(B_n).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mu\left(\{x \in A : f^*(x) > t\}\right) &= \mu\left(\{x \in A : f_n(x) > t,\ \text{for some } n\}\right) \\
&= \mu\left(\{x \in A : f_n^A(x) > t,\ \text{for some } n\}\right) \\
&= \mu\left(\cup_{n=1}^\infty A \cap B_n\right) \\
&= \sum_{n=1}^\infty \mu\left(A \cap B_n\right) \\
&\le 2^{-t}\sum_{n=1}^\infty \mu(B_n) \le 2^{-t}.
\end{aligned}$$

We now show that $f^* \in L^1(X, \mathcal{F}, \mu)$. First notice that

$$\mu\left(\{x \in A : f^*(x) > t\}\right) \le \mu(A),$$

hence,

$$\mu\left(\{x \in A : f^*(x) > t\}\right) \le \min(\mu(A), 2^{-t}).$$

Using Fubini's Theorem, and the fact that $f^* \geq 0$ one has

$$\begin{aligned}
\int_X f^*(x)\,\mathrm{d}\mu(x) &= \int_0^\infty \mu\left(\{x \in X : f^*(x) > t\}\right)\,\mathrm{d}t \\
&= \int_0^\infty \sum_{A\in\alpha} \mu\left(\{x \in A : f^*(x) > t\}\right)\,\mathrm{d}t \\
&= \sum_{A\in\alpha} \int_0^\infty \mu\left(\{x \in A : f^*(x) > t\}\right)\,\mathrm{d}t \\
&\leq \sum_{A\in\alpha} \int_0^\infty \min(\mu(A), 2^{-t})\,\mathrm{d}t \\
&= \sum_{A\in\alpha} \int_0^{-\log\mu(A)} \mu(A)\,\mathrm{d}t + \sum_{A\in\alpha} \int_{-\log\mu(A)}^\infty 2^{-t}\,\mathrm{d}t \\
&= -\sum_{A\in\alpha} \mu(A)\log\mu(A) + \sum_{A\in\alpha} \frac{\mu(A)}{\log_e 2} \\
&= H_\mu(\alpha) + \frac{1}{\log_e 2} < \infty.
\end{aligned}$$

□

So far we have defined the notion of the conditional entropy $I_{\alpha|\beta}$ when α and β are countable partitions. We can generalize the definition to the case α is a countable partition and $\mathcal{G}$ is a σ-algebra as follows (see equation (4.3)),

$$I_{\alpha|\mathcal{G}}(x) = -\log E_\mu(1_{\alpha(x)}|\mathcal{G}).$$

If we denote by $\bigvee_{i=1}^\infty T^{-i}\alpha = \sigma(\cup_n \bigvee_{i=1}^n T^{-i}\alpha)$, then

$$I_{\alpha|\bigvee_{i=1}^\infty T^{-i}\alpha}(x) = \lim_{n\to\infty} I_{\alpha|\bigvee_{i=1}^n T^{-i}\alpha}(x). \tag{4.4}$$

Exercise 4.4.2 Give a proof of equation (4.4) using the following important theorem, known as the Martingale Convergence Theorem (and is stated to our setting)

Theorem 4.4.1 *(Martingale Convergence Theorem) Let $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \cdots$ be a sequence of increasing σalgebras, and let $\mathcal{C} = \sigma(\cup_n \mathcal{C}_n)$. If $f \in L^1(\mu)$, then*

$$E_\mu(f|\mathcal{C}) = \lim_{n\to\infty} E_\mu(f|\mathcal{C}_n)$$

μ a.e. and in $L^1(\mu)$.

Exercise 4.4.3 Show that if T is measure preserving on the probability space $(X, \mathcal{F}, \mu)$ and $f \in L^1(\mu)$, then

$$\lim_{n\to\infty} \frac{f(T^n x)}{n} = 0, \ \ \mu \ a.e.$$

Theorem 4.4.2 *(The Shannon-McMillan-Breiman Theorem) Suppose T is an ergodic measure preserving transformation on a probability space $(X, \mathcal{F}, \mu)$, and let α be a countable partition with $H(\alpha) < \infty$. Then,*

$$\lim_{n\to\infty} \frac{1}{n+1} I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x) = h(\alpha, T) \ a.e.$$

Proof For each $n = 1, 2, 3, \ldots$, let $f_n(x) = I_{\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha}(x)$. Then,

$$\begin{aligned}
I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x) &= I_{\bigvee_{i=1}^{n} T^{-i}\alpha}(x) + I_{\alpha|\bigvee_{i=1}^{n} T^{-i}\alpha}(x) \\
&= I_{\bigvee_{i=0}^{n-1} T^{-i}\alpha}(Tx) + f_n(x) \\
&= I_{\bigvee_{i=1}^{n-1} T^{-i}\alpha}(Tx) + I_{\alpha|\bigvee_{i=1}^{n-1} T^{-i}\alpha}(Tx) + f_n(x) \\
&= I_{\bigvee_{i=0}^{n-2} T^{-i}\alpha}(T^2x) + f_{n-1}(Tx) + f_n(x) \\
&\vdots \\
&= I_\alpha(T^n x) + f_1(T^{n-1}x) + \ldots + f_{n-1}(Tx) + f_n(x).
\end{aligned}$$

Let $f(x) = I_{\alpha|\bigvee_{i=1}^{\infty} T^{-i}\alpha}(x) = \lim_{n\to\infty} f_n(x)$. Notice that $f \in L^1(X, \mathcal{F}, \mu)$ since $\int_X f(x)\, d\mu(x) = h(\alpha, T)$. Now letting $f_0 = I_\alpha$, we have

$$\begin{aligned}
\frac{1}{n+1} I_{\bigvee_{i=0}^{n} T^{-i}\alpha}(x) &= \frac{1}{n+1}\sum_{k=0}^{n} f_{n-k}(T^k x) \\
&= \frac{1}{n+1}\sum_{k=0}^{n} f(T^k x) + \frac{1}{n+1}\sum_{k=0}^{n}(f_{n-k} - f)(T^k x).
\end{aligned}$$

By the ergodic theorem,

$$\lim_{n\to\infty} \sum_{k=0}^{n} f(T^k x) = \int_X f(x)\, d\mu(x) = h(\alpha, T) \ a.e.$$

We now study the sequence $\{\frac{1}{n+1}\sum_{k=0}^{n}(f_{n-k}-f)(T^kx)\}$. Let

$$F_N = \sup_{k\geq N}|f_k - f|, \text{ and } f^* = \sup_{n\geq 1} f_n.$$

Notice that $0 \leq F_N \leq f^*+f$, hence $F_N \in L^1(X, \mathcal{F}, \mu)$ and $\lim_{N\to\infty} F_N(x) = 0$ a.e. Also for any k, $|f_{n-k} - f| \leq f^* + f$, so that $|f_{n-k} - f| \in L^1(X, \mathcal{F}, \mu)$ and $\lim n \to \infty |f_{n-k} - f| = 0$ a.e.
For any $N \leq n$,

$$\begin{aligned}
\frac{1}{n+1}\sum_{k=0}^{n}|f_{n-k}-f|(T^kx) &= \frac{1}{n+1}\sum_{k=0}^{n-N}|f_{n-k}-f|(T^kx) \\
&+ \frac{1}{n+1}\sum_{k=n-N+1}^{n}|f_{n-k}-f|(T^kx) \\
&\leq \frac{1}{n+1}\sum_{k=0}^{n-N}F_N(T^kx) \\
&+ \frac{1}{n+1}\sum_{k=0}^{N-1}|f_k-f|(T^{n-k}x).
\end{aligned}$$

If we take the limit as $n \to \infty$, then by exercise (4.4.3) the second term tends to 0 a.e., and by the ergodic theorem as well as the dominated convergence theorem, the first term tends to zero a.e. Hence,

$$\lim_{n\to\infty}\frac{1}{n+1}I_{\bigvee_{i=0}^{n}T^{-i}\alpha}(x) = h(\alpha, T) \ a.e.$$

$\square$

The above theorem can be interpreted as providing an estimate of the size of the atoms of $\bigvee_{i=0}^{n}T^{-i}\alpha$. For n sufficiently large, a typical element $A \in \bigvee_{i=0}^{n}T^{-i}\alpha$ satisfies

$$-\frac{1}{n+1}\log\mu(A) \approx h(\alpha, T)$$

or

$$\mu(A_n) \approx 2^{-(n+1)h(\alpha,T)}.$$

Furthermore, if α is a generating partition (i.e. $\bigvee_{i=0}^{\infty}T^{-i}\alpha = \mathcal{F}$, then in the conclusion of Shannon-McMillan-Breiman Theorem one can replace $h(\alpha, T)$ by $h(T)$.

4.5 Lochs' Theorem

In 1964, G. Lochs compared the decimal and the continued fraction expansions. Let $x \in [0,1)$ be an irrational number, and suppose $x = .d_1 d_2 \cdots$ is the decimal expansion of x (which is generated by iterating the map $Sx = 10x \pmod 1$). Suppose further that

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}} = [0; a_1, a_2, \cdots] \tag{4.5}$$

is its regular continued fraction (RCF) expansion (generated by the map $Tx = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor$). Let $y = .d_1 d_2 \cdots d_n$ be the rational number determined by the first n decimal digits of x, and let $z = y + 10^{-n}$. Then, $[y, z)$ is the decimal cylinder of order n containing x, which we also denote by $B_n(x)$. Now let

$$y = \cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots + \cfrac{1}{b_l}}}$$

and

$$z = \cfrac{1}{c_1 + \cfrac{1}{c_2 + \ddots + \cfrac{1}{c_k}}}$$

be the continued fraction expansion of y and z. Let

$$m(n,x) = \max\{i \le \min(l,k) : \text{ for all } j \le i, \ b_j = c_j\}. \tag{4.6}$$

In other words, if $B_n(x)$ denotes the decimal cylinder consisting of all points y in $[0,1)$ such that the first n decimal digits of y agree with those of x, and if $C_j(x)$ denotes the continued fraction cylinder of order j containing x, i.e., $C_j(x)$ is the set of all points in $[0,1)$ such that the first j digits in their continued fraction expansion is the same as that of x, then $m(n,x)$ is the largest integer such that $B_n(x) \subset C_{m(n,x)}(x)$. Lochs proved the following theorem:

Theorem 4.5.1 *Let λ denote Lebesgue measure on $[0,1)$. Then for a.e. $x \in [0,1)$*

$$\lim_{n\to\infty} \frac{m(n,x)}{n} = \frac{6\log 2 \log 10}{\pi^2}.$$

In this section, we will prove a generalization of Lochs' theorem that allows one to compare any two known expansions of numbers. We show that Lochs' theorem is true for any two sequences of interval partitions on $[0,1)$ satisfying the conclusion of Shannon-McMillan-Breiman theorem. The content of this section as well as the proofs can be found in [DF]. We begin with few definitions that will be used in the arguments to follow.

Definition 4.5.1 *By an interval partition we mean a finite or countable partition of $[0,1)$ into subintervals. If P is an interval partition and $x \in [0,1)$, we let $P\,(x)$ denote the interval of P containing x.*

Let $\mathcal{P} = \{P_n\}_{n=1}^{\infty}$ be a sequence of interval partitions. Let λ denote Lebesgue probability measure on $[0,1)$.

Definition 4.5.2 *Let $c \geq 0$. We say that $\mathcal{P}$ has entropy c a.e. with respect to λ if*

$$-\frac{\log \lambda\,(P_n\,(x))}{n} \to c \text{ a.e.}$$

Note that we do not assume that each P_n is refined by P_{n+1}.

Suppose that $\mathcal{P} = \{P_n\}_{n=1}^{\infty}$ and $\mathcal{Q} = \{Q_n\}_{n=1}^{\infty}$ are sequences of interval partitions. For each $n \in \mathbb{N}$ and $x \in [0,1)$, define

$$m_{\mathcal{P},\mathcal{Q}}\,(n,x) = \sup\{m \mid P_n\,(x) \subset Q_m\,(x)\}\,.$$

Theorem 4.5.2 *Let $\mathcal{P} = \{P_n\}_{n=1}^{\infty}$ and $\mathcal{Q} = \{Q_n\}_{n=1}^{\infty}$ be sequences of interval partitions and λ Lebesgue probability measure on $[0,1)$. Suppose that for some constants $c > 0$ and $d > 0$, $\mathcal{P}$ has entropy c a.e with respect to λ and $\mathcal{Q}$ has entropy d a.e. with respect to λ. Then*

$$\frac{m_{\mathcal{P},\mathcal{Q}}\,(n,x)}{n} \to \frac{c}{d} \text{ a.e.}$$

Proof First we show that

$$\limsup_{n\to\infty} \frac{m_{\mathcal{P},\mathcal{Q}}(n,x)}{n} \le \frac{c}{d} \text{ a.e.}$$

Fix $\varepsilon > 0$. Let $x \in [0,1)$ be a point at which the convergence conditions of the hypotheses are met. Fix $\eta > 0$ so that $\dfrac{c+\eta}{c-\frac{c}{d}\eta} < 1+\varepsilon$. Choose N so that for all $n \ge N$

$$\lambda(P_n(x)) > 2^{-n(c+\eta)}$$

and

$$\lambda(Q_n(x)) < 2^{-n(d-\eta)}.$$

Fix n so that $\min\left\{n, \frac{c}{d}n\right\} \ge N$, and let m' denote any integer greater than $(1+\varepsilon)\frac{c}{d}n$. By the choice of η,

$$\lambda(P_n(x)) > \lambda(Q_{m'}(x))$$

so that $P_n(x)$ is not contained in $Q_{m'}(x)$. Therefore

$$m_{\mathcal{P},\mathcal{Q}}(n,x) \le (1+\varepsilon)\frac{c}{d}n$$

and so

$$\limsup_{n\to\infty} \frac{m_{\mathcal{P},\mathcal{Q}}(n,x)}{n} \le (1+\varepsilon)\frac{c}{d} \text{ a.e.}$$

Since $\varepsilon > 0$ was arbitrary, we have the desired result.

Now we show that

$$\liminf_{n\to\infty} \frac{m_{\mathcal{P},\mathcal{Q}}(n,x)}{n} \ge \frac{c}{d} \text{ a.e.}$$

Fix $\varepsilon \in (0,1)$. Choose $\eta > 0$ so that $\zeta =: \varepsilon c - \eta\left(1 + (1-\varepsilon)\frac{c}{d}\right) > 0$. For each $n \in \mathbb{N}$ let $\bar{m}(n) = \left\lfloor (1-\varepsilon)\frac{c}{d}n \right\rfloor$. For brevity, for each $n \in \mathbb{N}$ we call an element of P_n (respectively Q_n) (n,η) −good if

$$\lambda(P_n(x)) < 2^{-n(c-\eta)}$$

(respectively

$$\lambda(Q_n(x)) > 2^{-n(d+\eta)}.)$$

For each $n \in \mathbb{N}$, let

$$D_n(\eta) = \left\{ x : \begin{array}{c} P_n(x) \text{ is } (n,\eta) - \text{good and } Q_{\bar{m}(n)}(x) \text{ is } (\bar{m}(n), \eta) - \text{ good} \\ \text{and } P_n(x) \nsubseteq Q_{\bar{m}(n)}(x) \end{array} \right\}.$$

If $x \in D_n(\eta)$, then $P_n(x)$ contains an endpoint of the $(\bar{m}(n), \eta)$ −good interval $Q_{\bar{m}(n)}(x)$. By the definition of $D_n(\eta)$ and $\bar{m}(n)$,

$$\frac{\lambda(P_n(x))}{\lambda\left(Q_{\bar{m}(n)}(x)\right)} < 2^{-n\zeta}.$$

Since no more than one atom of P_n can contain a particular endpoint of an atom of $Q_{\bar{m}(n)}$, we see that $\lambda(D_n(\eta)) < 2 \cdot 2^{-n\zeta}$ and so

$$\sum_{n=1}^{\infty} \lambda(D_n(\eta)) < \infty,$$

which implies that

$$\lambda\{x \mid x \in D_n(\eta) \text{ i.o.}\} = 0.$$

Since $\bar{m}(n)$ goes to infinity as n does, we have shown that for almost every $x \in [0,1)$, there exists $N \in \mathbb{N}$, so that for all $n \geq N$, $P_n(x)$ is (n,η) −good and $Q_{\bar{m}(n)}(x)$ is $(\bar{m}(n), \eta)$ −good and $x \notin D_n(\eta)$. In other words, for almost every $x \in [0,1)$, there exists $N \in \mathbb{N}$, so that for all $n \geq N$, $P_n(x)$ is (n,η) −good and $Q_{\bar{m}(n)}(x)$ is $(\bar{m}(n), \eta)$ −good and $P_n(x) \subset Q_{\bar{m}(n)}(x)$. Thus, for almost every $x \in [0,1)$, there exists $N \in \mathbb{N}$, so that for all $n \geq N$, $m_{\mathcal{P},\mathcal{Q}}(n,x) \geq \bar{m}(n)$, so that

$$\frac{m_{\mathcal{P},\mathcal{Q}}(n,x)}{n} \geq \lfloor (1-\varepsilon)\frac{c}{d} \rfloor.$$

This proves that

$$\liminf_{n\to\infty} \frac{m_{\mathcal{P},\mathcal{Q}}(n,x)}{n} \geq (1-\varepsilon)\frac{c}{d} \text{ a.e.}$$

Since $\varepsilon > 0$ was arbitrary, we have established the theorem. □

The above result allows us to compare any two well-known expansions of numbers. Since the *commonly used* expansions are usually performed for points in the unit interval, our underlying space will be $([0,1), \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Lebesgue σ-algebra, and λ the Lebesgue measure. The expansions we have in mind share the following two properties.

Definition 4.5.3 *A surjective map $T : [0,1) \to [0,1)$ is called a number theoretic fibered map (NTFM) if it satisfies the following conditions:*

(a) there exists a finite or countable partition of intervals $\alpha = \{A_i;\, i \in D\}$ such that T restricted to each atom of α (cylinder set of order 0) is monotone, continuous and injective. Furthermore, α is a generating partition.

(b) T is ergodic with respect to Lebesgue measure λ, and there exists a T invariant probability measure μ equivalent to λ with bounded density. (Both $\frac{d\mu}{d\lambda}$ and $\frac{d\lambda}{d\mu}$ are bounded, and $\mu(A) = 0$ if and only if $\lambda(A) = 0$ for all Lebesgue sets A.).

Let T be an NTFM with corresponding partition α, and T-invariant measure μ equivalent to λ. Let $L, M > 0$ be such that

$$L\lambda(A) \le \mu(A) < M\lambda(A)$$

for all Lebesgue sets A (property (b)). For $n \ge 1$, let $P_n = \bigvee_{i=0}^{n-1} T^{-i}\alpha$, then by property (a), P_n is an interval partition. If $H_\mu(\alpha) < \infty$, then Shannon-McMillan-Breiman Theorem gives

$$\lim_{n\to\infty} -\frac{\log \mu(P_n(x))}{n} = h_\mu(T) \text{ a.e. with respect to } \mu.$$

Exercise 4.5.1 Show that the conclusion of the Shannon-McMillan-Breiman Theorem holds if we replace μ by λ, i.e.

$$\lim_{n\to\infty} -\frac{\log \lambda(P_n(x))}{n} = h_\mu(T) \text{ a.e. with respect to } \lambda.$$

Iterations of T generate expansions of points $x \in [0,1)$ with digits in D. We refer to the resulting expansion as the T-*expansion of* x.

Almost all known expansions on $[0,1)$ are generated by a NTFM. Among them are the n-adic expansions ($Tx = nx \pmod 1$, where n is a positive integer), β expansions ($Tx = \beta x \pmod 1$, where $\beta > 1$ is a real number), continued fraction expansions ($Tx = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor$), and many others (see the book Ergodic Theory of Numbers).

Exercise 4.5.2 Prove Theorem (4.5.1) using Theorem (4.5.2). Use the fact that the continued fraction map T is ergodic with respect to Gauss measure μ, given by

$$\mu(B) = \int_B \frac{1}{\log 2} \frac{1}{1+x} \, \mathrm{d}x,$$

and has entropy equal to $h_\mu(T) = \dfrac{\pi^2}{6 \log 2}$.

Exercise 4.5.3 Reformulate and prove Lochs' Theorem for any two NTFM maps S and T on $[0,1)$.

Chapter 5

Hurewicz Ergodic Theorem

In this chapter we consider a class of non-measure preserving transformations. In particular, we study invertible, non-singular and conservative transformations on a probability space. We first start with a quick review of equivalent measures, we then define non-singular and conservative transformations, and state some of their properties. We end this section by giving a new proof of Hurewicz Ergodic Theorem, which is a generalization of Birkhoff Ergodic Theorem to non-singular conservative transformations.

5.1 Equivalent measures

Recall that two measures μ and ν on a measure space $(Y, \mathcal{F})$ are equivalent if μ and ν have the same null-sets, i.e.,

$$\mu(A) = 0 \quad \Leftrightarrow \quad \nu(A) = 0, \qquad A \in \mathcal{F}.$$

The theorem of Radon-Nikodym says that if μ, ν are σ-finite and equivalent, then there exist measurable functions $f, g \geq 0$, such that

$$\mu(A) = \int_A f \, \mathrm{d}\nu \quad \text{and} \quad \nu(A) = \int_A g \, \mathrm{d}\mu.$$

Furthermore, for all $h \in L^1(\mu)$ (or $L^1(\nu)$),

$$\int h \, \mathrm{d}\mu = \int hf \, \mathrm{d}\nu \text{ and } \int h \, \mathrm{d}\nu = \int hg \, \mathrm{d}\mu.$$

Usually the function f is denoted by $\dfrac{d\mu}{d\nu}$ and the function g by $\dfrac{d\nu}{d\mu}$.

Now suppose that $(X, \mathcal{B}, \mu)$ is a probability space, and $T : X \to X$ a measurable transformation. One can define a new measure $\mu \circ T^{-1}$ on $(X, \mathcal{B})$ by $\mu \circ T^{-1}(A) = \mu(T^{-1}A)$ for $A \in \mathcal{B}$. It is not hard to prove that for $f \in L^1(\mu)$,

$$\int f \, \mathrm{d}(\mu \circ T^{-1}) = \int f \circ T \, \mathrm{d}\mu \tag{5.1}$$

Exercise 5.1.1 Starting with indicator functions, give a proof of (5.1).

Note that if T is invertible, then one has that

$$\int f \, \mathrm{d}(\mu \circ T) = \int f \circ T^{-1} \, \mathrm{d}\mu \tag{5.2}$$

5.2 Non-singular and conservative transformations

Definition 5.2.1 *Let $(X, \mathcal{B}, \mu)$ be a probability space and $T : X \to X$ an invertible measurable function. T is said to be non-singular if for any $A \in \mathcal{B}$,*

$$\mu(A) = 0 \ \textit{ if and only if } \ \mu(T^{-1}A) = 0.$$

Since T is invertible, non-singularity implies that

$$\mu(A) = 0 \ \text{ if and only if } \ \mu(T^n A) = 0, \, n \neq 0.$$

This implies that the measures $\mu \circ T^n$ defined by $\mu \circ T^n(A) = \mu(T^n A)$ is equivalent to μ (and hence equivalent to each other). By the theorem of Radon-Nikodym, there exists for each $n \neq 0$, a non-negative measurable function $\omega_n(x) = \dfrac{d\mu \circ T^n}{d\mu}(x)$ such that

$$\mu(T^n A) = \int_A \omega_n(x) \, \mathrm{d}\mu(x).$$

We have the following propositions.

Proposition 5.2.1 *Suppose $(X, \mathcal{B}, \mu)$ is a probability space, and $T : X \to X$ is invertible and non-singular. Then for every $f \in L^1(\mu)$,*

$$\int_X f(x)\,\mathrm{d}\mu(x) = \int_X f(Tx)\omega_1(x)\,\mathrm{d}\mu(x) = \int_X f(T^n x)\omega_n(x)\,\mathrm{d}\mu(x).$$

Proof We show the result for indicator functions only, the rest of the proof is left to the reader.

$$\begin{aligned}\int_X 1_A(x)\,\mathrm{d}\mu(x) &= \mu(A) = \mu(T(T^{-1}A)) \\ &= \int_{T^{-1}A} \omega_1(x)\,\mathrm{d}\mu(x) \\ &= \int_X 1_A(Tx)\omega_1(x)\,\mathrm{d}\mu(x).\end{aligned}$$

$\square$

Proposition 5.2.2 *Under the assumptions of Proposition 5.2.1, one has for all $n, m \geq 1$, that*

$$\omega_{n+m}(x) = \omega_n(x)\omega_m(T^n x), \quad \mu\,a.e.$$

Proof For any $A \in \mathcal{B}$,

$$\begin{aligned}\int_A \omega_n(x)\omega_m(T^n x)\mathrm{d}\mu(x) &= \int_X 1_A(x)\omega_m(T^n x)\mathrm{d}(\mu \circ T^n)(x) \\ &= \int_X 1_A(T^{-n}x)\omega_m(x)\mathrm{d}\mu(x) \\ &= \int_X 1_{T^n A}(x)\mathrm{d}(\mu \circ T^m)(x) \\ &= \mu \circ T^m(T^n A) = \mu(T^{m+n}A) = \int_A \omega_{n+m}(x)\mathrm{d}\mu(x).\end{aligned}$$

Hence, $\omega_{n+m}(x) = \omega_n(x)\omega_m(T^n x), \quad \mu$ a.e.

Exercise 5.2.1 Let $(X, \mathcal{B}, \mu)$ be a probability space, and $T : X \to X$ an invertible non-singular transformation. For any measurable function f, set $f_n(x) = \sum_{i=0}^{n-1} f(T^i x)\omega_i(x)$, $n \geq 1$, where $\omega_0(x) = 1$. Show that for all $n, m \geq 1$,

$$f_{n+m}(x) = f_n(x) + \omega_n(x) f_m(T^n x).$$

Definition 5.2.2 *Let $(X, \mathcal{B}, \mu)$ be a probability space, and $T : X \to X$ a measurable transformation. We say that T is conservative if for any $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists an $n \geq 1$ such that $\mu(A \cap T^{-n}A) > 0$.*

Note that if T is invertible, non-singular and conservative , then T^{-1} is also non-singular and conservative. In this case, for any $A \in \mathcal{B}$ with $\mu(A) > 0$, there exists an $n \neq 0$ such that $\mu(A \cap T^n A) > 0$.

Proposition 5.2.3 *Suppose T is invertible, non-singular and conservative on the probability space $(X, \mathcal{B}, \mu)$, and let $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for μ a.e. $x \in A$ there exist infinitely many positive and negative integers n, such that $T^n x \in A$.*

Proof Let $B = \{x \in A : T^n x \notin A \text{ for all } n \geq 1\}$. Note that for any $n \geq 1$, $B \cap T^{-n}B = \emptyset$. If $\mu(B) > 0$, then by conservativity there exists an $n \geq 1$, such that $\mu(B \cap T^{-n}B)$ is positive, which is a contradiction. Hence, $\mu(B) = 0$, and by non-singularity we have $\mu(T^{-n}B) = 0$ for all $n \geq 1$.

Now, let $C = \{x \in A; T^n x \in A \text{ for only finitely many } n \geq 1\}$, then $C \subset \bigcup_{n=1}^{\infty} T^{-n}B$, implying that

$$\mu(C) \leq \sum_{n=1}^{\infty} \mu(T^{-n}B) = 0.$$

Therefore, for almost every $x \in A$ there exist infinitely many $n \geq 1$ such that $T^n x \in A$. Replacing T by T^{-1} yields the result for $n \leq -1$. □

Proposition 5.2.4 *Suppose T is invertible, non-singular and conservative, then*

$$\sum_{n=1}^{\infty} \omega_n(x) = \infty, \quad \mu \, a.e.$$

Proof Let $A = \{x \in X : \sum_{n=1}^{\infty} \omega_n(x) < \infty\}$. Note that

$$A = \bigcup_{M=1}^{\infty} \{x \in X : \sum_{n=1}^{\infty} \omega_n(x) < M\}.$$

If $\mu(A) > 0$, then there exists an $M \geq 1$ such that the set

$$B = \{x \in X : \sum_{n=1}^{\infty} \omega_n(x) < M\}$$

has positive measure. Then, $\int_B \sum_{n=1}^{\infty} \omega_n(x)\, \mathrm{d}\mu(x) < M\mu(B) < \infty$. However,

$$\begin{aligned}\int_B \sum_{n=1}^{\infty} \omega_n(x)\, \mathrm{d}\mu(x) &= \sum_{n=1}^{\infty} \int_B \omega_n(x)\, \mathrm{d}\mu(x) \\ &= \sum_{n=1}^{\infty} \mu(T^n B) \\ &= \sum_{n=1}^{\infty} \int_X 1_{T^n B}(x)\, \mathrm{d}\mu(x) \\ &= \int_X \sum_{n=1}^{\infty} 1_B(T^{-n}x)\, \mathrm{d}\mu(x).\end{aligned}$$

Hence, $\int_X \sum_{n=1}^{\infty} 1_B(T^{-n}x)\, \mathrm{d}\mu(x) < \infty$, which implies that

$$\sum_{n=1}^{\infty} 1_B(T^{-n}x) < \infty \quad \mu\, a.e.$$

Therefore, for μ a.e. x one has $T^{-n}x \in B$ for only finitely many $n \geq 1$, contradicting Proposition 5.2.3. Thus $\mu(A) = 0$, and

$$\sum_{n=1}^{\infty} \omega_n(x) = \infty, \quad \mu\, a.e.$$

□

5.3 Hurewicz Ergodic Theorem

The following theorem by Hurewicz is a generalization of Birkhoff's Ergodic Theorem to our setting; see also Hurewicz' original paper [H]. We give a new prove, similar to the proof for Birkhoff's Theorem; see Section 2.1 and [KK].

Theorem 5.3.1 *Let $(X, \mathcal{B}, \mu)$ be a probability space, and $T : X \to X$ an invertible, non-singular and conservative transformation. If $f \in L^1(\mu)$, then*

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} f(T^i x)\omega_i(x)}{\sum_{i=0}^{n-1} \omega_i(x)} = f_*(x)$$

exists μ a.e. Furthermore, f_* is T-invariant and

$$\int_X f(x)\,\mathrm{d}\mu(x) = \int_X f_*(x)\,\mathrm{d}\mu(x).$$

Proof Assume with no loss of generality that $f \geq 0$ (otherwise we write $f = f^+ - f^-$, and we consider each part separately). Let

$$f_n(x) = f(x) + f(Tx)\omega_1(x) + \cdots + f(T^{n-1}x)\omega_{n-1}(x),$$

$$g_n(x) = \omega_0(x) + \omega_1(x) + \cdots + \omega_{n-1}(x), \quad \omega_0(x) = g_0(x) = 1,$$

$$\overline{f}(x) = \limsup_{n\to\infty} \frac{f_n(x)}{\sum_{i=0}^{n-1}\omega_i(x)} = \limsup_{n\to\infty} \frac{f_n(x)}{g_n(x)},$$

and

$$\underline{f}(x) = \liminf_{n\to\infty} \frac{f_n(x)}{\sum_{i=0}^{n-1}\omega_i(x)} = \liminf_{n\to\infty} \frac{f_n(x)}{g_n(x)}.$$

By Proposition (5.2.2), one has $g_{n+m}(x) = g_n(x) + g_m(T^n x)$. Using Exercise (5.2.1) and Proposition (5.2.4), we will show that $\overline{f}$ and $\underline{f}$ are T-invariant. To this end,

$$\begin{aligned}
\overline{f}(Tx) &= \limsup_{n\to\infty} \frac{f_n(Tx)}{g_n(T^n x)} \\
&= \limsup_{n\to\infty} \frac{\dfrac{f_{n+1}(x) - f(x)}{\omega_1(x)}}{\dfrac{g_{n+1}(x) - g(x)}{\omega_1(x)}} \\
&= \limsup_{n\to\infty} \frac{f_{n+1}(x) - f(x)}{g_{n+1}(x) - g(x)} \\
&= \limsup_{n\to\infty} \left[\frac{f_{n+1}(x)}{g_{n+1}(x)} \cdot \frac{g_{n+1}(x)}{g_{n+1}(x) - g(x)} - \frac{f(x)}{g_{n+1}(x) - g(x)}\right] \\
&= \limsup_{n\to\infty} \frac{f_{n+1}(x)}{g_{n+1}(x)} \\
&= \overline{f}(x).
\end{aligned}$$

(Similarly $\underline{f}$ is T-invariant).

Now, to prove that f_* exists, is integrable and T-invariant, it is enough to show that

$$\int_X \underline{f}\,\mathrm{d}\mu \geq \int_X f\,\mathrm{d}\mu \geq \int_X \overline{f}\,\mathrm{d}\mu.$$

For since $\overline{f} - \underline{f} \geq 0$, this would imply that $\overline{f} = \underline{f} = f_*$. a.e.

We first prove that $\int_X \overline{f} d\mu \leq \int_X f\,\mathrm{d}\mu$. Fix any $0 < \epsilon < 1$, and let $L > 0$ be any real number. By definition of $\overline{f}$, for any $x \in X$, there exists an integer $m > 0$ such that

$$\frac{f_m(x)}{g_m(x)} \geq \min(\overline{f}(x), L)(1-\epsilon).$$

Now, for any $\delta > 0$ there exists an integer $M > 0$ such that the set

$$X_0 = \{x \in X : \exists\ 1 \leq m \leq M \text{ with } f_m(x) \geq g_m(x)\ \min(\overline{f}(x), L)(1-\epsilon)\}$$

has measure at least $1 - \delta$. Define F on X by

$$F(x) = \begin{cases} f(x) & x \in X_0 \\ L & x \notin X_0. \end{cases}$$

Notice that $f \leq F$ (why?). For any $x \in X$, let $a_n = a_n(x) = F(T^n x)\omega_n(x)$, and $b_n = b_n(x) = \min(\overline{f}(x), L)(1-\epsilon)\omega_n(x)$. We now show that $\{a_n\}$ and $\{b_n\}$ satisfy the hypothesis of Lemma 2.1.1 with $M > 0$ as above. For any $n = 0, 1, 2, \ldots$

–if $T^n x \in X_0$, then there exists $1 \leq m \leq M$ such that

$$f_m(T^n x) \geq \min(\overline{f}(x), L)(1-\epsilon) g_m(T^n x).$$

Hence,

$$\omega_n(x) f_m(T^n x) \geq \min(\overline{f}(x), L)(1-\epsilon) g_m(T^n x)\omega_n(x).$$

Now,

$$\begin{aligned}
b_n + \ldots + b_{n+m-1} &= \min(\overline{f}(x), L)(1-\epsilon) g_m(T^n x)\omega_n(x) \\
&\leq \omega_n(x) f_m(T^n x) \\
&= f(T^n x)\omega_n(x) + f(T^{n+1}x)\omega_{n+1}(x) + \cdots + f(T^{n+m-1}x)\omega_{n+m-1}(x) \\
&\leq F(T^n x)\omega_n(x) + F(T^{n+1}x)\omega_{n+1}(x) + \cdots + F(T^{n+m-1}x)\omega_{n+m-1}(x) \\
&= a_n + a_{n+1} + \cdots + a_{n+m-1}.
\end{aligned}$$

–If $T^n x \notin X_0$, then take $m = 1$ since

$$a_n = F(T^n x)\omega_n(x) = L\omega_n(x) \geq \min(\overline{f}(x), L)(1-\epsilon)\omega_n(x) = b_n.$$

Hence by T-invariance of $\overline{f}$, and Lemma 2.1.1 for all integers $N > M$ one has

$$F(x)+F(Tx)+\omega_1(x)+\cdots+\omega_{N-1}(x)F(T^{N-1}x) \geq \min(\overline{f}(x), L)(1-\epsilon)g_{N-M}(x).$$

Integrating both sides, and using Proposition (5.2.1) together with the T-invariance of $\overline{f}$ one gets

$$\begin{aligned} N\int_X F(x)\,\mathrm{d}\mu(x) &\geq \int_X \min(\overline{f}(x), L)(1-\epsilon)g_{N-M}(x)\,\mathrm{d}\mu(x) \\ &= (N-M)\int_X \min(\overline{f}(x), L)(1-\epsilon)\,\mathrm{d}\mu(x). \end{aligned}$$

Since

$$\int_X F(x)\,\mathrm{d}\mu(x) = \int_{X_0} f(x)\,\mathrm{d}\mu(x) + L\mu(X\setminus X_0),$$

one has

$$\begin{aligned} \int_X f(x)\,\mathrm{d}\mu(x) &\geq \int_{X_0} f(x)\,\mathrm{d}\mu(x) \\ &= \int_X F(x)\,\mathrm{d}\mu(x) - L\mu(X\setminus X_0) \\ &\geq \frac{(N-M)}{N}\int_X \min(\overline{f}(x), L)(1-\epsilon)\,\mathrm{d}\mu(x) - L\delta. \end{aligned}$$

Now letting first $N \to \infty$, then $\delta \to 0$, then $\epsilon \to 0$, and lastly $L \to \infty$ one gets together with the monotone convergence theorem that $\overline{f}$ is integrable, and

$$\int_X f(x)\,\mathrm{d}\mu(x) \geq \int_X \overline{f}(x)\,\mathrm{d}\mu(x).$$

We now prove that

$$\int_X f(x)\,\mathrm{d}\mu(x) \leq \int_X \underline{f}(x)\,\mathrm{d}\mu(x).$$

Fix $\epsilon > 0$, for any $x \in X$ there exists an integer m such that

$$\frac{f_m(x)}{g_m(x)} \leq (\underline{f}(x) + \epsilon).$$

For any $\delta > 0$ there exists an integer $M > 0$ such that the set

$$Y_0 = \{x \in X : \exists\ 1 \leq m \leq M \text{ with } f_m(x) \leq\ (\underline{f}(x) + \epsilon) g_m(x)\}$$

has measure at least $1 - \delta$. Define G on X by

$$G(x) = \begin{cases} f(x) & x \in Y_0 \\ 0 & x \notin Y_0. \end{cases}$$

Notice that $G \leq f$. Let $b_n = G(T^n x)\omega_n(x)$, and $a_n = (\underline{f}(x) + \epsilon)\omega_n(x)$. We now check that the sequences $\{a_n\}$ and $\{b_n\}$ satisfy the hypothesis of Lemma 2.1.1 with $M > 0$ as above.

–if $T^n x \in Y_0$, then there exists $1 \leq m \leq M$ such that

$$f_m(T^n x) \leq\ (\underline{f}(x) + \epsilon) g_m(T^n x).$$

Hence,

$$\omega_n(x) f_m(T^n x) \leq\ (\underline{f}(x)+\epsilon) g_m(T^n x)\omega_n(x) = (\underline{f}(x)+\epsilon)(\omega_n(x)+\cdots+\omega_{n+m-1}(x).$$

By Proposition (5.2.2), and the fact that $f \geq G$, one gets

$$\begin{aligned} b_n + \ldots + b_{n+m-1} &= G(T^n x)\omega_n(x) + \cdots + G(T^{n+m-1}x)\omega_{n+m-1}(x) \\ &\leq f(T^n x)\omega_n(x) + \cdots + f(T^{n+m-1}x)\omega_{n+m-1}(x) \\ &= \omega_n(x) f_m(T^n x) \\ &\leq (\underline{f}(x) + \epsilon)(\omega_n(x) + \cdots + \omega_{n+m+1}(x)) \\ &= a_n + \cdots a_{n+m-1}. \end{aligned}$$

–If $T^n x \notin Y_0$, then take $m = 1$ since

$$b_n = G(T^n x)\omega_n(x) = 0 \leq (\underline{f}(x) + \epsilon)(\omega_n(x)) = a_n.$$

Hence by Lemma 2.1.1 one has for all integers $N > M$

$$G(x) + G(Tx)\omega_1(x) + \ldots + G(T^{N-M-1}x)\omega_{N-M-1}(x) \leq (\underline{f}(x) + \epsilon) g_N(x).$$

Integrating both sides yields

$$(N-M)\int_X G(x)d\mu(x) \le N(\int_X \underline{f}(x)d\mu(x) + \epsilon).$$

Since $f \ge 0$, the measure ν defined by $\nu(A) = \int_A f(x)\,\mathrm{d}\mu(x)$ is absolutely continuous with respect to the measure μ. Hence, there exists $\delta_0 > 0$ such that if $\mu(A) < \delta$, then $\nu(A) < \delta_0$. Since $\mu(X \setminus Y_0) < \delta$, then $\nu(X \setminus Y_0) = \int_{X\setminus Y_0} f(x)d\mu(x) < \delta_0$. Hence,

$$\begin{aligned}\int_X f(x)\,\mathrm{d}\mu(x) &= \int_X G(x)\,\mathrm{d}\mu(x) + \int_{X\setminus Y_0} f(x)\,\mathrm{d}\mu(x) \\ &\le \frac{N}{N-M}\int_X (\underline{f}(x)+\epsilon)\,\mathrm{d}\mu(x) + \delta_0.\end{aligned}$$

Now, let first $N \to \infty$, then $\delta \to 0$ (and hence $\delta_0 \to 0$), and finally $\epsilon \to 0$, one gets

$$\int_X f(x)\,\mathrm{d}\mu(x) \le \int_X \underline{f}(x)\,\mathrm{d}\mu(x).$$

This shows that

$$\int_X \underline{f}\,\mathrm{d}\mu \ge \int_X f\,\mathrm{d}\mu \ge \int_X \overline{f}\,\mathrm{d}\mu,$$

hence, $\overline{f} = \underline{f} = f_*$ a.e., and f_* is T-invariant. □

Remark We can extend the notion of ergodicity to our setting. If T is non-singular and conservative, we say that T is ergodic if for any measurable set A satisfying $\mu(A\Delta T^{-1}A) = 0$, one has $\mu(A) = 0$ or 1. It is easy to check that the proof of Proposition (1.7.1) holds in this case, so that T ergodic implies that each T-invariant function is a constant μ a.e. Hence, if T is invertible, non-singular, conservative and ergodic, then by Hurewicz Ergodic Theorem one has for any $f \in L^1(\mu)$,

$$\lim_{n\to\infty} \frac{\displaystyle\sum_{i=0}^{n-1} f(T^i x)\omega_i(x)}{\displaystyle\sum_{i=0}^{n-1} \omega_i(x)} = \int_X f d\mu \quad \mu \ a.e.$$

Chapter 6

Invariant Measures for Continuous Transformations

6.1 Existence

Suppose X is a compact metric space, and let $\mathcal{B}$ be the Borel σ-algebra i.e., the σ-algebra generated by the open sets. Let $M(X)$ be the collection of all Borel probability measures on X. There is natural embedding of the space X in $M(X)$ via the map $x \to \delta_x$, where δ_X is the Dirac measure concentrated at x ($\delta_x(A) = 1$ if $x \in A$, and is zero otherwise). Furthermore, $M(X)$ is a convex set, i.e., $p\mu + (1-p)\nu \in M(X)$ whenever $\mu, \nu \in M(X)$ and $0 \leq p \leq 1$. Theorem 6.1.2 below shows that a member of $M(X)$ is determined by how it integrates continuous functions. We denote by $C(X)$ the Banach space of all complex valued continuous functions on X under the supremum norm.

Theorem 6.1.1 *Every member of $M(X)$ is regular, i.e., for all $B \in \mathcal{B}$ and every $\epsilon > 0$ there exist an open set U_ϵ and a closed sed C_ϵ such that $C_\epsilon \subseteq B \subseteq U_\epsilon$ such that $\mu(U_\epsilon \setminus C_\epsilon) < \epsilon$.*

Idea of proof Call a set $B \in \mathcal{B}$ with the above property *a regular set*. Let $\mathcal{R} = \{B \in \mathcal{B} : B \text{ is regular }\}$. Show that $\mathcal{R}$ is a σ-algebra containing all the closed sets. $\square$

Corollary 6.1.1 *For any $B \in \mathcal{B}$, and any $\mu \in M(X)$,*

$$\mu(B) = \sup_{C \subseteq B : C \text{ closed}} \mu(C) = \inf_{B \subseteq U : U \text{ open}} \mu(U).$$

Theorem 6.1.2 *Let $\mu, m \in M(X)$. If*

$$\int_X f \, \mathrm{d}\mu = \int_X f \, \mathrm{d}m$$

for all $f \in C(X)$, then $\mu = m$.

Proof From the above corollary, it suffices to show that $\mu(C) = m(C)$ for all closed subsets C of X. Let $\epsilon > 0$, by regularity of the measure m there exists an open set U_ϵ such that $C \subseteq U_\epsilon$ and $m(U_\epsilon \setminus C) < \epsilon$. Define $f \in C(X)$ as follows

$$f(x) = \begin{cases} 0 & x \notin U_\epsilon \\ \frac{d(x, X \setminus U_\epsilon)}{d(x, X \setminus U_\epsilon) + d(x, C)} & x \in U_\epsilon. \end{cases}$$

Notice that $1_C \leq f \leq 1_{U_\epsilon}$, thus

$$\mu(C) \leq \int_X f \, \mathrm{d}\mu = \int_X f \, \mathrm{d}m \leq m(U_\epsilon) \leq m(C) + \epsilon.$$

Using a similar argument, one can show that $m(C) \leq \mu(C) + \epsilon$. Therefore, $\mu(C) = m(C)$ for all closed sets, and hence for all Borel sets. $\square$

This allows us to define a metric structure on $M(X)$ as follows. A sequence $\{\mu_n\}$ in $M(X)$ converges to $\mu \in M(X)$ if and only if

$$\lim_{n \to \infty} \int_X f(x) \, \mathrm{d}\mu_n(x) = \int_X f(x) \, \mathrm{d}\mu(x)$$

for all $f \in C(X)$. We will show that under this notion of convergence the space $M(X)$ is compact, but first we need The Riesz Representation Representation Theorem.

Theorem 6.1.3 *(The Riesz Representation Theorem) Let X be a compact metric space and $J : C(X) \to \mathbb{C}$ a continuous linear map such that J is a positive operator and $J(1) = 1$. Then there exists a $\mu \in M(X)$ such that $J(f) = \int_X f(x) \, \mathrm{d}\mu(x)$.*

Theorem 6.1.4 *The space $M(X)$ is compact.*

Idea of proof Let $\{\mu_n\}$ be a sequence in $M(X)$. Choose a countable dense subset of $\{f_n\}$ of $C(X)$. The sequence $\{\int_X f_1 \, \mathrm{d}\mu_n\}$ is a bounded sequence of complex numbers, hence has a convergent subsequence $\{\int_X f_1 \, \mathrm{d}\mu_n^{(1)}\}$. Now, the sequence $\{\int_X f_2 \, \mathrm{d}\mu_n^{(1)}\}$ is bounded, and hence has a convergent subsequence $\{\int_X f_2 \, \mathrm{d}\mu_n^{(2)}\}$. Notice that $\{\int_X f_1 \, \mathrm{d}\mu_n^{(2)}\}$ is also convergent. We continue in this manner, to get for each i a subsequence $\{\mu_n^{(i)}\}$ of $\{\mu_n\}$ such that for all $j \leq i$, $\{\mu_n^{(i)}\}$ is a subsequence of $\{\mu_n^{(j)}\}$ and $\{\int_X f_j \, \mathrm{d}\mu_n^{(i)}\}$ converges. Consider the diagonal sequence $\{\mu_n^{(n)}\}$, then $\{\int_X f_j \, \mathrm{d}\mu_n^{(n)}\}$ converges for all j, and hence $\{\int_X f \, \mathrm{d}\mu_n^{(n)}\}$ converges for all $f \in C(X)$. Now define $J : C(X) \to \mathbb{C}$ by $J(f) = \lim_{n\to\infty}\{\int_X f \, \mathrm{d}\mu_n^{(n)}\}$. Then, J is linear, continuous ($|J(f)| \leq \sup_{x\in X}|f(x)|$), positive and $J(1) = 1$. Thus, by Riesz Representation Theorem, there exists a $\mu \in M(X)$ such that $J(f) = \lim_{n\to\infty}\{\int_X f \, \mathrm{d}\mu_n^{(n)}\} = \int_X f \, \mathrm{d}\mu$. Therefore, $\lim_{n\to\infty} \mu_n^{(n)} = \mu$, and $M(X)$ is compact. $\square$

Let $T : X \to X$ be a continuous transformation. Since $\mathcal{B}$ is generated by the open sets, then T is measurable with respect to $\mathcal{B}$. Furthermore, T induces in a natural way, an operator $\overline{T} : M(X) \to M(X)$ given by

$$(\overline{T}\mu)(A) = \mu(T^{-1}A)$$

for all $A \in \mathcal{B}$. Then $\overline{T}^i\mu(A) = \mu(T^{-i}A)$. Using a standard argument, one can easily show that

$$\int_X f(x) \, \mathrm{d}(\overline{T}\mu)(x) = \int_X f(Tx) d\mu(x)$$

for all continuous functions f on X. Note that T is measure preserving with respect to $\mu \in M(X)$ if and only if $\overline{T}\mu = \mu$. Equivalently, μ is measure preserving if and only if

$$\int_X f(x) \, \mathrm{d}\mu(x) = \int_X f(Tx) \, \mathrm{d}\mu(x)$$

for all continuous functions f on X. Let

$$M(X,T) = \{\mu \in M(X) : \overline{T}\mu = \mu\}$$

be the collection of all probability measures under which T is measure preserving.

Theorem 6.1.5 *Let $T : X \to X$ be continuous, and $\{\sigma_n\}$ a sequence in $M(X)$. Define a sequence $\{\mu_n\}$ in $M(X)$ by*

$$\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \overline{T}^i \sigma_n.$$

Then, any limit point μ of $\{\mu_n\}$ is a member of $M(X,T)$.

Proof We need to show that for any continuous function f on X, one has $\int_X f(x)\,\mathrm{d}\mu(x) = \int_X f(Tx)\,\mathrm{d}\mu$. Since $M(X)$ is compact there exists a $\mu \in M(X)$ and a subsequence $\{\mu_{n_i}\}$ such that $\mu_{n_i} \to \mu$ in $M(X)$. Now for any f continuous, we have

$$\begin{aligned}
\left|\int_X f(Tx)\,\mathrm{d}\mu(x) - \int_X f(x)\,\mathrm{d}\mu(x)\right| &= \lim_{j\to\infty}\left|\int_X f(Tx)\,\mathrm{d}\mu_{n_j}(x) - \int_X f(x)\,\mathrm{d}\mu_{n_j}(x)\right| \\
&= \lim_{j\to\infty}\left|\frac{1}{n_j}\int_X \sum_{i=0}^{n_j-1}\left(f(T^{i+1}x) - f(T^i x)\right)\mathrm{d}\sigma_{n_j}(x)\right| \\
&= \lim_{j\to\infty}\left|\frac{1}{n_j}\int_X (f(T^{n_j}x) - f(x))\,\mathrm{d}\sigma_{n_j}(x)\right| \\
&\le \lim_{j\to\infty}\frac{2\sup_{x\in X}|f(x)|}{n_j} = 0.
\end{aligned}$$

Therefore $\mu \in M(X,T)$. □

Theorem 6.1.6 *Let T be a continuous transformation on a compact metric space. Then,*

(i) $M(X,T)$ is a compact convex subset of $M(X)$.

(ii) $\mu \in M(X,T)$ is an extreme point (i.e. μ cannot be written in a non-trivial way as a convex combination of elements of $M(X,T)$) if and only if T is ergodic with respect to μ.

Proof (i) Clearly $M(X,T)$ is convex. Now let $\{\mu_n\}$ be a sequence in $M(X,T)$ converging to μ in $M(X)$. We need to show that $\mu \in M(X,T)$.

Since T is continuous, then for any continuous function f on X, the function $f \circ T$ is also continuous. Hence,

$$\begin{aligned}
\int_X f(Tx)\,\mathrm{d}\mu(x) &= \lim_{n\to\infty} \int_X f(Tx)\,\mathrm{d}\mu_n(x) \\
&= \lim_{n\to\infty} \int_X f(x)\,\mathrm{d}\mu_n(x) \\
&= \int_X f(x)\,\mathrm{d}\mu(x).
\end{aligned}$$

Therefore, T is measure preserving with respect to μ, and $\mu \in M(X,T)$.
(ii) Suppose T is ergodic with respect to μ, and assume that

$$\mu = p\mu_1 + (1-p)\mu_2$$

for some $\mu_1, \mu_2 \in M(X,T)$, and some $0 < p \leq 1$. We will show that $\mu = \mu_1$. Notice that the measure μ_1 is absolutely continuous with respect to μ, and T is ergodic with respect to μ, hence by Theorem (2.1.2) we have $\mu_1 = \mu$.

Conversely, (we prove the contrapositive) suppose that T is not ergodic with respect to μ. Then there exists a measurable set E such that $T^{-1}E = E$, and $0 < \mu(E) < 1$. Define measures μ_1, μ_2 on X by

$$\mu_1(B) = \frac{\mu(B \cap E)}{\mu(E)} \text{ and } \mu_1(B) = \frac{\mu\left(B \cap (X \setminus E)\right)}{\mu(X \setminus E)}.$$

Since E and $X \setminus E$ are T-invariant sets, then $\mu_1, \mu_2 \in M(X,T)$, and $\mu_1 \neq \mu_2$ since $\mu_1(E) = 1$ while $\mu_2(E) = 0$. Furthermore, for any measurable set B,

$$\mu(B) = \mu(E)\mu_1(B) + (1 - \mu(E))\mu_2(B),$$

i.e. μ_1 is a non-trivial convex combination of elements of $M(X,T)$. Thus, μ is not an extreme point of $M(X,T)$. $\square$

Since the Banach space $C(X)$ of all continuous functions on X (under the sup norm) is separable i.e. $C(X)$ has a countable dense subset, one gets the following strengthening of the Ergodic Theorem.

Theorem 6.1.7 *If $T : X \to X$ is continuous and $\mu \in M(X,T)$ is ergodic, then there exists a measurable set Y such that $\mu(Y) = 1$, and*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \int_X f(x)\,\mathrm{d}\mu(x)$$

for all $x \in Y$, and $f \in C(X)$.

Proof Choose a countable dense subset $\{f_k\}$ in $C(X)$. By the Ergodic Theorem, for each k there exists a subset X_k with $\mu(X_k) = 1$ and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f_k(T^i x) = \int_X f_k(x)\,\mathrm{d}\mu(x)$$

for all $x \in X_k$. Let $Y = \bigcap_{k=1}^{\infty} X_k$, then $\mu(Y) = 1$, and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f_k(T^i x) = \int_X f_k(x) d\mu(x)$$

for all k and all $x \in Y$. Now, let $f \in C(X)$, then there exists a subsequence $\{f_{k_j}\}$ converging to f in the supremum norm, and hence is uniformly convergent. For any $x \in Y$, using uniform convergence and the dominated convergence theorem, one gets

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) &= \lim_{n\to\infty} \lim_{j\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f_{k_j}(T^i x) \\
&= \lim_{j\to\infty} \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f_{k_j}(T^i x) \\
&= \lim_{j\to\infty} \int_X f_{k_j}\,\mathrm{d}\mu = \int_X f\,\mathrm{d}\mu.
\end{aligned}$$

$\square$

Theorem 6.1.8 *Let $T : X \to X$ be continuous, and $\mu \in M(X, T)$. Then T is ergodic with respect to μ if and only if*

$$\frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i x} \to \mu \ \ a.e.$$

Proof Suppose T is ergodic with respect to μ. Notice that for any $f \in C(X)$,

$$\int_X f(y)\,\mathrm{d}(\delta_{T^i x})(y) = f(T^i x),$$

Hence by theorem 6.1.7, there exists a measurable set Y with $\mu(Y) = 1$ such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_X f(y)\,\mathrm{d}(\delta_{T^i x})(y) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \int_X f(y)\,\mathrm{d}\mu(y)$$

for all $x \in Y$, and $f \in C(X)$. Thus, $\frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i x} \to \mu$ for all $x \in Y$.

Conversely, suppose $\frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i x} \to \mu$ for all $x \in Y$, where $\mu(Y) = 1$. Then for any $f \in C(X)$ and any $g \in L^1(X, \mathcal{B}, \mu)$ one has

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) g(x) = g(x) \int_X f(y)\,\mathrm{d}\mu(y).$$

By the dominated convergence theorem

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_X f(T^i x) g(x)\,\mathrm{d}\mu(x) = \int_X g(x) d\mu(x) \int_X f(y)\,\mathrm{d}\mu(y).$$

Now, let $F, G \in L^2(X, \mathcal{B}, \mu)$. Then, $G \in L^1(X, \mathcal{B}, \mu)$ so that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \int_X f(T^i x) G(x)\,\mathrm{d}\mu(x) = \int_X G(x) d\mu(x) \int_X f(y) d\mu(y)$$

for all $f \in C(X)$. Let $\epsilon > 0$, there exists $f \in C(X)$ such that $||F - f||_2 < \epsilon$ which implies that $|\int F\,\mathrm{d}\mu - \int f d\mu| < \epsilon$. Furthermore, there exists N so that for $n \geq N$ one has

$$\left| \int_X \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) G(x)\,\mathrm{d}\mu(x) - \int_X G d\mu \int_X f\,\mathrm{d}\mu \right| < \epsilon.$$

Thus, for $n \geq N$ one has

$$\begin{aligned}
&\left|\int_X \frac{1}{n}\sum_{i=0}^{n-1} F(T^i x)G(x)\,\mathrm{d}\mu(x) - \int_X G d\mu \int_X F\,\mathrm{d}\mu\right| \\
\leq \; & \int_X \frac{1}{n}\sum_{i=0}^{n-1} \left|F(T^i x) - f(T^i x)\right| |G(x)|\,\mathrm{d}\mu(x) \\
+ \; & \left|\int_X \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x)G(x)d\mu(x) - \int_X G d\mu \int_X f d\mu\right| \\
+ \; & \left|\int_X f d\mu \int_X G d\mu - \int_X F\,\mathrm{d}\mu \int_X G\,\mathrm{d}\mu\right| \\
< \; & \epsilon||G||_2 + \epsilon + \epsilon||G||_2.
\end{aligned}$$

Thus,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \int_X F(T^i x)G(x)\,\mathrm{d}\mu(x) = \int_X G(x)\,\mathrm{d}\mu(x) \int_X F(y)\,\mathrm{d}\mu(y)$$

for all $F, G \in L^2(X, \mathcal{B}, \mu)$ and $x \in Y$. Taking F and G to be indicator functions, one gets that T is ergodic.

□

Exercise 6.1.1 Let X be a compact metric space and $T : X \to X$ be a continuous homeomorphism. Let $x \in X$ be periodic point of T of period n, i.e. $T^n x = x$ and $T^j x \neq x$ for $j < i$. Show that if $\mu \in M(X, T)$ is ergodic and $\mu(\{x\}) > 0$, then $\mu = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i x}$.

6.2 Unique Ergodicity

A continuous transformation $T : X \to X$ on a compact metric space is *uniquely ergodic* if there is only one T-invariant probabilty measure μ on X. In this case, $M(X, T) = \{\mu\}$, and μ is necessarily ergodic, since μ is an extreme point of $M(X, T)$. Recall that if $\nu \in M(X, T)$ is ergodic, then there exists a measurable subset Y such that $\nu(Y) = 1$ and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) = \int_X f(y)\,\mathrm{d}\nu(y)$$

for all $x \in Y$ and all $f \in C(X)$. When T is uniquely ergodic we will see that we have a much stronger result.

Theorem 6.2.1 *Let $T : X \to X$ be a continuous transformation on a compact metric space X. Then the following are equivalent:*

(i) For every $f \in C(X)$, the sequence $\{\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x)\}$ converges uniformly to a constant.

(ii) For every $f \in C(X)$, the sequence $\{\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x)\}$ converges pointwise to a constant.

(iii) There exists a $\mu \in M(X,T)$ such that for every $f \in C(X)$ and all $x \in X$.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) = \int_X f(y)\,\mathrm{d}\mu(y).$$

(iv) T is uniquely ergodic.

Proof (i) $\Rightarrow$ (ii) immediate.
(ii) $\Rightarrow$ (iii) Define $L : C(X) \to \mathbb{C}$ by

$$L(f) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x).$$

By assumption $L(f)$ is independent of x, hence L is well defined. It is easy to see that L is linear, continuous ($|L(f)| \leq \sup_{x\in X} |f(x)|$), positive and $L(1) = 1$. Thus, by Riesz Representation Theorem there exists a probability measure $\mu \in M(X)$ such that

$$L(f) = \int_X f(x)\,\mathrm{d}\mu(x)$$

for all $f \in C(x)$. But

$$L(f \circ T) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^{i+1}x) = L(f).$$

Hence,

$$\int_X f(Tx)\,\mathrm{d}\mu(x) = \int_X f(x)\,\mathrm{d}\mu(x).$$

Thus, $\mu \in M(X,T)$, and for every $f \in C(X)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) = \int_X f(x)\,\mathrm{d}\mu(x)$$

for all $x \in X$.

(iii) $\Rightarrow$ (iv) Suppose $\mu \in M(X,T)$ is such that for every $f \in C(X)$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) = \int_X f(x)\,\mathrm{d}\mu(x)$$

for all $x \in X$. Assume $\nu \in M(X,T)$, we will show that $\mu = \nu$. For any $f \in C(X)$, since the sequence $\{\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x)\}$ converges pointwise to the constant function $\int_X f(x)\,\mathrm{d}\mu(x)$, and since each term of the sequence is bounded in absolute value by the constant $\sup_{x\in X}|f(x)|$, it follows by the Dominated Convergence Theorem that

$$\lim_{n\to\infty} \int_X \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x)\,\mathrm{d}\nu(x) = \int_X\int_X f(x)\,\mathrm{d}\mu(x)d\nu(y) = \int_X f(x)\,\mathrm{d}\mu(x).$$

But for each n,

$$\int_X \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x)\,\mathrm{d}\nu(x) = \int_X f(x)\,\mathrm{d}\nu(x).$$

Thus, $\int_X f(x)\,\mathrm{d}\mu(x) = \int_X f(x)\,\mathrm{d}\nu(x)$, and $\mu = \nu$.

(iv) $\Rightarrow$ (i) The proof is done by contradiction. Assume $M(X,T) = \{\mu\}$ and suppose $g \in C(X)$ is such that the sequence $\{\frac{1}{n}\sum_{j=0}^{n-1} g \circ T^j\}$ does not converge uniformly on X. Then there exists $\epsilon > 0$ such that for each N there exists $n > N$ and there exists $x_n \in X$ such that

$$\Big|\frac{1}{n}\sum_{j=0}^{n-1} g(T^j x_n) - \int_X g\,\mathrm{d}\mu\Big| \geq \epsilon.$$

Let

$$\mu_n = \frac{1}{n}\sum_{j=0}^{n-1} \delta_{T^j x_n} = \frac{1}{n}\sum_{j=0}^{n-1} \overline{T}^j \delta_{x_n}.$$

Then,

$$|\int_X g d\mu_n - \int_X g \,\mathrm{d}\mu| \geq \epsilon.$$

Since $M(X)$ is compact. there exists a subsequence μ_{n_i} converging to $\nu \in M(X)$. Hence,

$$|\int_X g \,\mathrm{d}\nu - \int_X g \,\mathrm{d}\mu| \geq \epsilon.$$

By Theorem (6.1.5), $\nu \in M(X,T)$ and by unique ergodicity $\mu = \nu$, which is a contradiction. □

Example If T_θ is an irrational rotation, then T_θ is uniquely ergodic. This is a consequence of the above theorem and Weyl's Theorem on uniform distribution: for any Riemann integrable function f on $[0,1)$, and any $x \in [0,1)$, one has

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(x + i\theta - \lfloor x + i\theta \rfloor) = \int_X f(y)\,\mathrm{d}y.$$

As an application of this, let us consider the following question. Consider the sequence of *first digits*

$$\{1, 2, 4, 8, 1, 3, 6, 1, \ldots\}$$

obtained by writing the first decimal digit of each term in the sequence

$$\{2^n : n \geq 0\} = \{1, 2, 4, 8, 16, 32, 64, 128, \ldots\}.$$

For each $k = 1, 2, \ldots, 9$, let $p_k(n)$ be the number of times the digit k appears in the first n terms of the *first digit* sequence. The asymptotic relative frequency of the digit k is then $\lim_{n\to\infty} \dfrac{p_k(n)}{n}$. We want to identify this limit for each $k \in \{1, 2, \ldots 9\}$. To do this, let $\theta = \log_{10} 2$, then θ is irrational. For $k = 1, 2, \ldots, 9$, let $J_k = [\log_{10} k, \log_{10}(k+1))$. By unique ergodicity of T_θ, we have for each $k = 1, 2, \ldots, 9$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} 1_{J_k}(T_\theta^j(0)) = \lambda(J_k) = \log_{10}\left(\frac{k+1}{k}\right).$$

Returning to our original problem, notice that the first digit of 2^i is k if and only if

$$k \cdot 10^r \leq 2^i < (k+1) \cdot 10^r$$

for some $r \geq 0$. In this case,

$$r + \log_{10} k \leq i \log_{10} 2 = i\theta < r + \log_{10}(k+1).$$

This shows that $r = \lfloor i\theta \rfloor$, and

$$\log_{10} k \leq i\theta - \lfloor i\theta \rfloor < \log_{10}(k+1).$$

But $T_\theta^i(0) = i\theta - \lfloor i\theta \rfloor$, so that $T_\theta^i(0) \in J_k$. Summarizing, we see that the first digit of 2^i is k if and only if $T_\theta^i(0) \in J_k$. Thus,

$$\lim_{n\to\infty} \frac{p_k(n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{J_k}(T_\theta^i(0)) = \log_{10}\left(\frac{k+1}{k}\right).$$

Chapter 7

Topological Dynamics

In this chapter we will shift our attention from the theory of measure preserving transformations and ergodic theory to the study of topological dynamics. The field of topological dynamics also studies dynamical systems, but in a different setting: where ergodic theory is set in a measure space with a measure preserving transformation, topological dynamics focuses on a topological space with a continuous transformation. The two fields bear great similarity and seem to co-evolve. Whenever a new notion has proven its value in one field, analogies will be sought in the other. The two fields are, however, also complementary. Indeed, we shall encounter a most striking example of a map which exhibits its most interesting topological dynamics in a set of measure zero, right in the blind spot of the measure theoretic eye.
In order to illuminate this interplay between the two fields we will be merely interested in compact metric spaces. The compactness assumption will play the role of the assumption of a finite measure space. The assumption of the existence of a metric is sufficient for applications and will greatly simplify all of our proofs. The chapter is structured as follows. We will first introduce several basic notions from topological dynamics and discuss how they are related to each other and to analogous notions from ergodic theory. We will then devote a separate section to topological entropy. To conclude the chapter, we will discuss some examples and applications.

7.1 Basic Notions

Throughout this chapter X will denote a compact metric space. Unless stated otherwise, we will always denote the metric by d and occasionally write (X, d) to denote a metric space if there is any risk of confusion. We will assume that the reader has some familiarity with basic (analytic) topology. To jog the reader's memory, a brief outline of these basics is included as an appendix. Here we will only summarize the properties of the space X, for future reference.

Theorem 7.1.1 *Let X be a compact metric space, Y a topological space and $f : X \longrightarrow Y$ continuous. Then,*

- *X is a Hausdorff space*
- *Every closed subspace of X is compact*
- *Every compact subspace of X is closed*
- *X has a countable basis for its topology*
- *X is normal, i.e. for any pair of disjoint closed sets A and B of X there are disjoint open sets U, V containing A and B, respectively*
- *If $\mathcal{O}$ is an open cover of X, then there is a $\delta > 0$ such that every subset A of X with $diam(A) < \delta$ is contained in an element of $\mathcal{O}$. We call $\delta > 0$ a Lebesgue number for $\mathcal{O}$.*
- *X is a Baire space*
- *f is uniformly continuous*
- *If Y is an ordered space then f attains a maximum and minimum on X*
- *If A and B are closed sets in X and $[a, b] \subset \mathbb{R}$, then there exists a continuous map $g : X \longrightarrow [a, b]$ such that $g(x) = a$ for all $x \in A$ and $g(x) = b$ for all $x \in b$*

The reader may recognize that this theorem contains some of the most important results in analytic topology, most notably the Lebesgue number lemma, Baire's category theorem, the extreme value theorem and Urysohn's lemma.
Let us now introduce some concepts of topological dynamics: topological transitivity, topological conjugacy, periodicity and expansiveness.

Definition 7.1.1 *Let $T : X \longrightarrow X$ be a continuous transformation. For $x \in X$, the forward orbit of x is the set $FO_T(x) = \{T^n(x)|n \in \mathbb{Z}_{\geq 0}\}$. T is called one-sided topologically transitive if for some $x \in X$, $FO_T(x)$ is dense in X.*
If T is invertible, the orbit of x is defined as $O_T(x) = \{T^n(x)|n \in \mathbb{Z}\}$. If T is a homeomorphism, T is called topologically transitive if for some $x \in X$ $O_T(x)$ is dense in X.

In the literature sometimes a stronger version of topological transitivity is introduced, called minimality.

Definition 7.1.2 *A homeomorphism $T : X \longrightarrow X$ is called minimal if every $x \in X$ has a dense orbit in X.*

Example 7.1.1 Let X be the topological space $X = \{1, 2, \ldots, 1000\}$ with the discrete topology. X is a finite space, hence compact, and the discrete metric

$$d_{disc}(x, y) = \begin{cases} 0 \text{ if } x = y \\ 1 \text{ if } x \neq y \end{cases}$$

induces the discrete topology on X. Thus X is a compact metric space. If we now define $T : X \longrightarrow X$ by the permutation $(12 \cdots 1000)$, i.e. $T(1) = 2$, $T(2) = 3$,. . .,$T(1000) = 1$, then it is easy to see that T is a homeomorphism. Moreover, $O_T(x) = X$ for all $x \in X$, so T is minimal.

Example 7.1.2 Fix $n \in \mathbb{Z}_{>0}$. Let X be the topological space $X = \{0\} \cup \{\frac{1}{n^k}|k \in \mathbb{Z}_{>0}\}$ equipped with the subspace topology inherited from $\mathbb{R}$. Then X is a compact metric space. Indeed, if $\mathcal{O}$ is an open cover of X, then $\mathcal{O}$ contains an open set U which contains 0. U contains all but finitely many points of X, so by adding an open set U_y to U from $\mathcal{O}$ with $y \in U_y$, for each y not in U, we obtain a finite subcover. Define $T : X \longrightarrow X$ by $T(x) = \frac{x}{n}$.

Note that T is continuous, but not surjective as $T^{-1}(\{1\}) = \emptyset$. Observe that $FO_T(1) = X - \{0\}$ and $\overline{X - \{0\}} = X$, so T is one-sided topologically transitive as $x = 1$ has a dense forward orbit.

The following theorem summarizes some equivalent definitions of topological transitivity.

Theorem 7.1.2 *Let $T : X \longrightarrow X$ be a homeomorphism of a compact metric space. Then the following are equivalent:*

1. *T is topologically transitive*
2. *If A is a closed set in X satisfying $TA = A$, then either $A = X$ or A has empty interior (i.e. $X - A$ is dense in X)*
3. *For any pair of non empty open sets U, V in X there exists an $n \in \mathbb{Z}$ such that $T^n U \cap V \neq \emptyset$*

Proof

Recall that $O_T(x)$ is dense in X if and only if for every U open in X, $U \cap O_T(x) \neq \emptyset$.
(1)⇒(2): Suppose that $\overline{O_T(x)} = X$ and let A be as stated. Suppose that A does not have an empty interior, then there exists an open set U such that U is open, non-empty and $U \subset A$. Now, we can find a $p \in \mathbb{Z}$ such that $T^p(x) \in U \subset A$. Since $T^n A = A$ for all $n \in \mathbb{Z}$, $O_T(x) \subset A$ and taking closures we obtain $X \subset A$. Hence, either A has empty interior, or $A = X$.
(2)⇒(3): Suppose U, V are open and non-empty. Then $W = \cup_{n=-\infty}^{\infty} T^n U$ is open, non-empty and invariant under T. Hence, $A = X - W$ is closed, $TA = A$ and $A \neq X$. By (2), A has empty interior and therefore W is dense in X. Thus $W \cap V \neq \emptyset$, which implies (3).
(3)⇒(1): For an arbitrary $y \in X$ we have: $y \in \overline{O_T(x)}$ if and only if every open neighborhood V of y intersects $O_T(x)$, i.e. $T^m(x) \in V$ for some $m \in \mathbb{Z}$. By theorem 7.1.1 there exists a countable basis $\mathcal{U} = \{U_n\}_{n=1}^{\infty}$ for the topology of X. Since every open neighborhood of y contains a basis element U_i such that $y \in U_i$, we see that $\overline{O_T(x)} = X$ if and only if for every $n \in \mathbb{Z}_{>0}$ there is some $m \in \mathbb{Z}$ such that $T^m(x) \in U_n$. Hence,

$$\{x \in X | \overline{O_T(x)} = X\} = \bigcap_{n=1}^{\infty} \bigcup_{m=-\infty}^{\infty} T^m U_n$$

But $\cup_{m=-\infty}^{\infty} T^m U_n$ is T-invariant and therefore intersects every open set in X by (3). Thus, $\cup_{m=-\infty}^{\infty} T^m U_n$ is dense in X for every n. Now, since X is a Baire space (c.f. theorem 7.1.1), we see that $\{x \in X | \overline{O_T(x)} = X\}$ is itself dense in X and thus certainly non-empty (as $X \neq \emptyset$). $\square$

An analogue of this theorem exists for one-sided topological transitivity, see [W].
Note that theorem 7.1.2 clearly resembles theorem (1.6.1) and (2) implies that we can view topological transitivity as an analogue of ergodicity (in some sense). This is also reflected in the following (partial) analogue of theorem (1.7.1).

Theorem 7.1.3 *Let $T : X \longrightarrow X$ be continuous and one-sided topologically transitive or a topologically transitive homeomorphism. Then every continuous T-invariant function is constant.*

Proof

Let $f : X \longrightarrow Y$ be continuous on X and suppose $f \circ T = f$. Then $f \circ T^n = f$, so f is constant on (forward) orbits of points. Let $x_0 \in X$ have a dense (forward) orbit in X and suppose $f(x_0) = c$ for some $c \in Y$. Fix $\varepsilon > 0$ and let $x \in X$ be arbitrary. By continuity of f at x, we can find a $\delta > 0$ such that $d(f(x), f(\tilde{x})) < \varepsilon$ for any $\tilde{x} \in X$ with $d(x, \tilde{x}) < \delta$. But then, since x_0 has a dense (forward) orbit in X, there is some $n \in \mathbb{Z}$ ($n \in \mathbb{Z}_{\geq 0}$) such that $d(x, T^n(x_0)) < \delta$. Hence,

$$d(f(x), c) = d(f(x), f(T^n(x_0))) < \varepsilon$$

Since $\varepsilon > 0$ was arbitrary, our proof is complete. $\square$

Exercise 7.1.1 Let X be a compact metric space with metric d and let $T : X \longrightarrow X$ is a topologically transitive homeomorphism. Show that if T is an isometry (i.e. $d(T(x), T(y)) = d(x, y)$, for all $x, y \in X$) then T is minimal.

Definition 7.1.3 *Let X be a topological space, $T : X \longrightarrow X$ a transformation and $x \in X$. Then x is called a periodic point of T of period n ($n \in \mathbb{Z}_{>0}$) if $T^n(x) = x$ and $T^m(x) \neq x$ for $m < n$. A periodic point of period 1 is called a fixed point.*

Example 7.1.3 The map $T : \mathbb{R} \longrightarrow \mathbb{R}$, defined by $T(x) = -x$ has one fixed point, namely $x = 0$. All other points in the domain are periodic points of period 2.

Definition 7.1.4 *Let X be a compact metric space and $T : X \longrightarrow X$ a homeomorphism. T is said to be expansive if there exist a $\delta > 0$ such that: if $x \neq y$ then there exists an $n \in \mathbb{Z}$ such that $d(T^n(x), T^n(y)) > \delta$. δ is called an expansive constant for T.*

Example 7.1.4 Let X be a finite space with the discrete topology. X is a compact metric space, since the discrete metric d_{disc} induces the topology on X. Now, *any* bijective map $T : X \longrightarrow X$ is an expansive homeomorphism and any $0 < \delta < 1$ is an expansive constant for T.

For a non-trivial example of an expansive homeomorphism, see exercise 7.3.1. Expansiveness is closely related to the concept of a generator.

Definition 7.1.5 *Let X be a compact metric space and $T : X \longrightarrow X$ a homeomorphism. A finite open cover α of X is called a generator for T if the set $\cap_{n=-\infty}^{\infty} T^{-n} A_n$ contains at most one point of X, for any collection $\{A_n\}_{n=-\infty}^{\infty}$, $A_i \in \alpha$.*

Theorem 7.1.4 *Let X be a compact metric space and $T : X \longrightarrow X$ a homeomorphism. Then T is expansive if and only if T has a generator.*

Proof

Suppose that T is expansive and let $\delta > 0$ be an expansive constant for T. Let α be the open cover of X defined by $\alpha = \{B(a, \frac{\delta}{2}) | a \in X\}$. By compactness, there exists a finite subcover β of α. Suppose that $x, y \in \cap_{n=-\infty}^{\infty} T^{-n} B_n$, $B_i \in \beta$ for all i. Then $d(T^m(x), T^m(y)) \leq \delta$, for all $m \in \mathbb{Z}$. But by expansiveness, there exists an $l \in \mathbb{Z}$ such that $d(T^l(x), T^l(y)) > \delta$, a contradiction. Hence, $x = y$. We conclude that β is a generator for T.
Conversely, suppose that α is a generator for T. By theorem 7.1.1, there exists a Lebesgue number $\delta > 0$ for the open cover α. We claim that $\delta/4$ is an expansive constant for T. Indeed, if $x, y \in X$ are such that $d(T^m(x), T^m(y)) \leq \delta/4$, for all $m \in \mathbb{Z}$. Then, for all $m \in \mathbb{Z}$, the closed ball $\overline{B}(T^m(x), \delta/4)$ contains $T^m(y)$ and has diameter $\delta/2 < \delta$. Hence,

$$\overline{B}(T^m(x), \delta/4) \subset A_m,$$

for some $A_m \in \alpha$. It follows that $x, y \in \cap_{m=-\infty}^{\infty} T^{-m} A_m$.But α is a generator, so $\cap_{m=-\infty}^{\infty} T^{-m} A_m$ contains at most one point. We conclude that $x = y$, T is expansive. $\square$

Exercise 7.1.2 In the literature, our definition of a generator is sometimes called a weak generator. A generator is then defined by replacing $\cap_{n=-\infty}^{\infty} T^{-n} A_n$ by $\cap_{n=-\infty}^{\infty} T^{-n} \overline{A}_n$ in definition 7.1.5. Show that in a compact metric space both concepts are in fact equivalent.

Exercise 7.1.3 Prove the following basic properties of an expansive homeomorphism $T : X \longrightarrow X$:

a. T is expansive if and only if T^n is expansive ($n \neq 0$)

b. If A is a closed subset of X and $T(A) = A$, then the restriction of T to A, $T|_A$, is expansive

c. Suppose $S : Y \longrightarrow Y$ is an expansive homeomorphism. Then the product map $T \times S : X \times Y \longrightarrow X \times Y$ is expansive with respect to the metric $d = max\{d_X, d_Y\}$.

Definition 7.1.6 *Let X, Y be compact spaces and let $T : X \longrightarrow X$, $S : Y \longrightarrow Y$ be homeomorphisms. T is said to be topologically conjugate to S if there exists a homeomorphism $\phi : X \longrightarrow Y$ such that $\phi \circ T = S \circ \phi$. ϕ is called a (topological) conjugacy.*

Note that topological conjugacy defines an equivalence relation on the space of all homeomorphisms.
The term 'topological conjugacy' is, in a sense, a misnomer. The following theorem shows that topological conjugacy can be considered as the counterpart of a measure preserving isomorphism.

Theorem 7.1.5 *Let X, Y be compact spaces and let $T : X \longrightarrow X$, $S : Y \longrightarrow Y$ be topologically conjugate homeomorphisms. Then,*

1. *T is topologically transitive if and only if S is topologically transitive*

2. *T is minimal if and only if S is minimal*

3. *T is expansive if and only if S is expansive*

Proof

The proof of the first two statements is trivial. For the third statement we note that

$$\bigcap_{n=-\infty}^{\infty} T^{-n}(\phi^{-1}A_n) = \bigcap_{n=-\infty}^{\infty} \phi^{-1} \circ S^{-n}(A_n) = \phi^{-1}(\bigcap_{n=-\infty}^{\infty} S^{-n}A_n)$$

The lefthandside of the equation contains at most one point if and only if the righthandside does, so a finite open cover α is a generator for S if and only if $\phi^{-1}\alpha = \{\phi^{-1}A | A \in \alpha\}$ is a generator for T. Hence the result follows from theorem 7.1.4. $\square$

7.2 Topological Entropy

Topological entropy was first introduced as an analogue of the succesful concept of measure theoretic entropy. It will turn out to be a conjugacy invariant and is therefore a useful tool for distinguishing between topological dynamical systems. The definition of topological entropy comes in two flavors. The first is in terms of open covers and is very similar to the definition of measure theoretic entropy in terms of partitions. The second (and chronologically later) definition uses (n,ε) separating and spanning sets. Interestingly, the latter definition was a topological dynamical discovery, which was later engineered into a similar definition of measure theoretic entropy.

7.2.1 Two Definitions

We will start with a definition of topological entropy similar to the definition of measure theoretic entropy introduced earlier. To spark the reader's recognition of the similarities between the two, we use analogous notation and terminology. Before kicking off, let us first make a remark on the assumptions about our space X. The first definition presented will only require X to be compact. The second, on the other hand, only requires X to be a metric space. We will obscure from these subtleties and simply stick to the context of a compact metric space, in which the two definitions will turn out to be equivalent. Nevertheless, the reader should be aware that both definitions represent generalizations of topological entropy in different directions

and are therefore interesting in their own right.
Let X be a compact metric space and let α, β be open covers of X. We say that β is a *refinement* of α, and write $\alpha \leq \beta$, if for every $B \in \beta$ there is an $A \in \alpha$ such that $B \subset A$. The *common refinement* of α and β is defined to be $\alpha \vee \beta = \{A \cap B | A \in \alpha, B \in \beta\}$. For a finite collection $\{\alpha_i\}_{i=1}^n$ of open covers we may define $\bigvee_{i=1}^n \alpha_i = \{\cap_{i=1}^n A_{j_i} | A_{j_i} \in \alpha_i\}$. For a continuous transformation $T : X \longrightarrow X$ we define $T^{-1}\alpha = \{T^{-1}A | A \in \alpha\}$. Note that these are all again open covers of X. Finally, we define the diameter of an open cover α as $diam(\alpha) := \sup_{A \in \alpha} diam(A)$, where $diam(A) = \sup_{x,y \in A} d(x,y)$.

Exercise 7.2.1 Show that $T^{-1}(\alpha \vee \beta) = T^{-1}(\alpha) \vee T^{-1}(\beta)$ and show that $\alpha \leq \beta$ implies $T^{-1}\alpha \leq T^{-1}\beta$.

Definition 7.2.1 *Let α be an open cover of X and let $N(\alpha)$ be the number of sets in a finite subcover of α of minimal cardinality. We define the entropy of α to be $H(\alpha) = \log(N(\alpha))$.*

The following proposition summarizes some easy properties of $H(\alpha)$. The proof is left as an exercise.

Proposition 7.2.1 *Let α be an open cover of X and let $H(\alpha)$ be the entropy of α. Then*

1. $H(\alpha) \geq 0$
2. $H(\alpha) = 0$ *if and only if* $N(\alpha) = 1$ *if and only if* $X \in \alpha$
3. *If* $\alpha \leq \beta$, *then* $H(\alpha) \leq H(\beta)$
4. $H(\alpha \vee \beta) \leq H(\alpha) + H(\beta)$
5. *For* $T : X \longrightarrow X$ *continuous we have* $H(T^{-1}\alpha) \leq H(\alpha)$. *If T is surjective then* $H(T^{-1}\alpha) = H(\alpha)$.

Exercise 7.2.2 Prove proposition 7.2.1 (Hint: If T is surjective then $T(T^{-1}A) = A$).

We will now move on to the definition of topological entropy for a continuous transformation with respect to an open cover and subsequently make this definition independent of open covers.

Definition 7.2.2 *Let α be an open cover of X and let $T : X \longrightarrow X$ be continuous. We define the topological entropy of T with respect to α to be:*

$$h(T,\alpha) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha)$$

We must show that $h(T,\alpha)$ is well-defined, i.e. that the right hand side exists.

Theorem 7.2.1 $\lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=1}^{n-1} T^{-i}\alpha)$ *exists.*

Proof

Define

$$a_n = H(\bigvee_{i=0}^{n-1} T^{-i}\alpha)$$

Then, by proposition (4.2.2),, it suffices to show that $\{a_n\}$ is subadditive. Now, by (4) of proposition 7.2.1 and exercise 7.2.1

$$\begin{aligned} a_{n+p} &= H(\bigvee_{i=0}^{n+p-1} T^{-i}\alpha) \\ &\leq H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) + H(T^{-n}\bigvee_{j=0}^{p-1} T^{-j}\alpha) \\ &\leq a_n + a_p \end{aligned}$$

This completes our proof. $\square$

Proposition 7.2.2 $h(T,\alpha)$ *satisfies the following:*

1. $h(T,\alpha) \geq 0$
2. *If* $\alpha \leq \beta$, *then* $h(T,\alpha) \leq h(T,\beta)$
3. $h(T,\alpha) \leq H(\alpha)$

Proof

These are easy consequences of proposition 7.2.1. We will only prove the third statement. We have:

$$
\begin{aligned}
H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) &\leq \sum_{i=0}^{n-1} H(T^{-i}\alpha) \\
&\leq nH(\alpha)
\end{aligned}
$$

Here we have subsequently used (4) and (5) of the proposition. □

Finally, we arrive at our sought definition:

Definition 7.2.3 *(I) Let $T : X \longrightarrow X$ be continuous. We define the topological entropy of T to be:*

$$h_1(T) = \sup_{\alpha} h(T, \alpha)$$

where the supremum is taken over all open covers of X.

We will defer a discussion of the properties of topological entropy till the end of this chapter.

Let us now turn to the second approach to defining topological entropy. This approach was first taken by E.I. Dinaburg.
We shall first define the main ingredients. Let d be the metric on the compact metric space X and for each $n \in \mathbb{Z}_{\geq 0}$ define a new metric by:

$$d_n(x, y) = \max_{0 \leq i \leq n-1} d(T^i(x), T^i(y))$$

Definition 7.2.4 *Let $n \in \mathbb{Z}_{\geq 0}$, $\varepsilon > 0$ and $A \subset X$. A is called (n, ε)-spanning for X if for all $x \in X$ there exists a $y \in A$ such that $d_n(x, y) < \varepsilon$. Define span(n, ε, T) to be the minimum cardinality of an (n, ε)-spanning set. A is called (n, ε)-separated if for any $x, y \in A$ $d_n(x, y) > \varepsilon$. We define sep(n, ε, T) to be the maximum cardinality of an (n, ε)-separated set.*

Figure 7.1: An (n, ε)-separated set (left) and (n, ε)-spanning set (right). The dotted circles represent open balls of d_n-radius $\frac{\varepsilon}{2}$ (left) and of d_n-radius ε (right), respectively.

Note that we can extend this definition by letting $A \subset K$, where K is a compact subset of X. This generalization to the context of non-compact metric spaces was devised by R.E. Bowen. See [W] for an outline of this extended framework.
We can also formulate the above in terms of open balls. If $B(x, r) = \{y \in X | d(x, y) < r\}$ is an open ball in the metric d, then the open ball with centre x and radius r in the metric d_n is given by:

$$B_n(x, r) := \bigcap_{i=0}^{n-1} T^{-i} B(T^i(x), r)$$

Hence, A is (n, ε)-spanning for X if:

$$X = \bigcup_{a \in A} B_n(a, \varepsilon)$$

and A is (n, ε)-separated if:

$$(A - \{a\}) \cap B_n(a, \varepsilon) = \emptyset \text{ for all } a \in A$$

Definition 7.2.5 *For $n \in \mathbb{Z}_{\geq 0}$ and $\varepsilon > 0$ we define $cov(n, \varepsilon, T)$ to be the minimum cardinality of a covering of X by open sets of d_n-diameter less than ε.*

The following theorem shows that the above notions are actually two sides of the same coin.

Theorem 7.2.2 *$cov(n, 3\varepsilon, T) \leq span(n, \varepsilon, T) \leq sep(n, \varepsilon, T) \leq cov(n, \varepsilon, T)$. Furthermore, $sep(n, \varepsilon, T)$, $span(n, \varepsilon, T)$ and $cov(n, \varepsilon, T)$ are finite, for all n, $\varepsilon > 0$ and continuous T.*

Proof

We will only prove the last two inequalities. The first is left as an exercise to the reader. Let A be an (n, ε)-separated set of cardinality $sep(n, \varepsilon, T)$. Suppose A is not (n, ε)-spanning for X. Then there is some $x \in X$ such that $d_n(x, a) \geq \varepsilon$, for all $a \in A$. But then $A \cup \{x\}$ is an (n, ε)-separated set of cardinality larger than $sep(n, \varepsilon, T)$. This contradiction shows that A is an (n, ε)-spanning set for X. The second inequality now follows since the cardinality of A is at least as large as $span(n, \varepsilon, T)$.
To prove the third inequality, let A be an (n, ε)-separated set of cardinality $sep(n, \varepsilon, T)$. Note that if α is an open cover of d_n-diameter less than ε, then no element of α can contain more than 1 element of A. This holds in particular for an open cover of minimal cardinality, so the third inequality is proved.
The final statement follows for $span(n, \varepsilon, T)$ and $cov(n, \varepsilon, T)$ from the compactness of X and subsequently for $sep(n, \varepsilon, T)$ by the last inequality. □

Exercise 7.2.3 Finish the proof of the above theorem by proving the first inequality.

Lemma 7.2.1 *The limit $cov(\varepsilon, T) := \lim_{n\to\infty} \frac{1}{n} \log(cov(n, \varepsilon, T))$ exists and is finite.*

Proof

Fix $\varepsilon > 0$. Let α, β be open covers of X consisting of sets with d_m-diameter and d_n-diameter, smaller than ε and with cardinalities $cov(m, \varepsilon, T)$ and $cov(n, \varepsilon, T)$, respectively. Pick any $A \in \alpha$ and $B \in \beta$. Then, for $x, y \in A \cap T^{-m}B$,

$$\begin{aligned} d_{m+n}(x, y) &= \max_{0\leq i\leq m+n-1} d(T^i(x), T^i(y)) \\ &\leq \max\{\max_{0\leq i\leq m-1} d(T^i(x), T^i(y)), \max_{m\leq j\leq m+n-1} d(T^j(x), T^j(y))\} \\ &< \varepsilon \end{aligned}$$

Thus, $A \cap T^{-m}B$ has d_{m+n}-diameter less than ε and the sets $\{A \cap T^{-m}B | A \in \alpha, B \in \beta\}$ form a cover of X. Hence,

$$cov(m+n, \varepsilon, T) \leq cov(m, \varepsilon, T) \cdot cov(n, \varepsilon, T)$$

Now, since log is an increasing function, the sequence $\{a_n\}_{n=1}^{\infty}$ defined by $a_n = \log cov(n, \varepsilon, T)$ is subadditive, so the result follows from proposition 5. $\square$

Motivated by the above lemma and theorem, we have the following definition

Definition 7.2.6 *(II) The topological entropy of a continuous transformation $T : X \longrightarrow X$ is given by:*

$$\begin{aligned} h_2(T) &= \lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log(cov(n, \varepsilon, T)) \\ &= \lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log(span(n, \varepsilon, T)) \\ &= \lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log(sep(n, \varepsilon, T)) \end{aligned}$$

Note that there seems to be a concealed ambiguity in this definition, since $h_2(T)$ still depends on the chosen metric d. As we will see later on, in compact metric spaces $h_2(T)$ is independent of d, as long as it induces the topology of X. However, there is definitely an issue at this point if we drop our assumption of compactness of the space X.

7.2.2 Equivalence of the two Definitions

We will now show that definitions (I) and (II) of topological entropy coincide on compact metric spaces. This will justify the use of the notation $h(T)$ for both definitions of topological entropy.
In the meantime, we will continue to write $h_1(T)$ for topological entropy in the definition using open covers and $h_2(T)$ for the second definition of entropy.
After all the work done in the last section, our only missing ingredient is the following lemma:

Lemma 7.2.2 *Let X be a compact metric space. Suppose $\{\alpha_n\}_{n=1}^{\infty}$ is a sequence of open covers of X such that $\lim_{n\to\infty} diam(\alpha_n) = 0$, where diam is the diameter under the metric d. Then $\lim_{n\to\infty} h_1(T, \alpha_n) = h_1(T)$.*

Proof

We will only prove the lemma for the case $h_1(T) < \infty$. Let $\varepsilon > 0$ be arbitrary and let β be an open cover of X such that $h_1(T, \beta) > h_1(T) - \varepsilon$. By theorem 7.1.1 there exists a Lebesgue number $\delta > 0$ for β. By assumption, there exists an $N > 0$ such that for $n \geq N$ we have $diam(\alpha_n) < \delta$. So, for such n, we find that for any $A \in \alpha_n$ there exists a $B \in \beta$ such that $A \subset B$, i.e. $\beta \leq \alpha_n$. Hence, by proposition 7.2.2 and the above,

$$h_1(T) - \varepsilon < h_1(T, \beta) \leq h_1(T, \alpha_n) \leq h_1(T)$$

for all $n \geq N$. □

Exercise 7.2.4 Finish the proof of lemma 7.2.2. That is, show that if $h_1(T) = \infty$, then $\lim_{n\to\infty} h_1(T, \alpha_n) = \infty$.

Theorem 7.2.3 *For a continuous transformation $T : X \longrightarrow X$ of a compact metric space X, definitions (I) and (II) of topological entropy coincide.*

Proof

Step 1: $h_2(T) \leq h_1(T)$.
For any n, let $\{\alpha_k\}_{k=1}^{\infty}$ be the sequence of open covers of X, defined by $\alpha_k = \{B_n(x, \frac{1}{3k}) | x \in X\}$. Then, clearly, the d_n-diameter of α_k is smaller than $\frac{1}{k}$. Now, $\bigvee_{i=0}^{n-1} T^{-i}\alpha_k$ is an open cover of X by sets of d_n-diameter smaller than $\frac{1}{k}$, hence $cov(n, \frac{1}{k}, T) \leq N(\bigvee_{i=0}^{n-1} T^{-i}\alpha_k)$. Since $\lim_{k\to\infty} diam(\alpha_k) = 0$, by lemma 7.2.2, we can subsequently take the log, divide by n, take the limit for $n \to \infty$ and the limit for $k \to \infty$ to obtain the desired result.
Step 2: $h_1(T) \leq h_2(T)$.
Let $\{\beta_k\}_{k=1}^{\infty}$ be defined by $\beta_k = \{B(x, \frac{1}{k}) | x \in X\}$. Note that $\delta_k = \frac{2}{k}$ is a Lebesgue number for β_k, for each k. Let A be an $(n, \frac{\delta_k}{2})$-spanning set for X of cardinality $span(n, \frac{\delta_k}{2}, T)$. Then, for each $a \in A$ the ball $B(T^i(a), \frac{\delta_k}{2})$ is contained in a member of β_k (for all $0 \leq i < n$), hence $B_n(T^i(a), \frac{\delta_k}{2})$ is contained in a member of $\bigvee_{i=0}^{n-1} T^{-i}\beta_k$. Thus, $N(\bigvee_{i=0}^{n-1} T^{-i}\beta_k) \leq span(n, \frac{1}{4k}, T)$. Proceeding in the same way as in step 1, we obtain the desired inequality.□

Properties of Entropy

We will now list some elementary properties of topological entropy.

Theorem 7.2.4 *Let X be a compact metric space and let $T : X \longrightarrow X$ be continuous. Then $h(T)$ satisfies the following properties:*

1. *If the metrics d and $\tilde{d}$ both generate the topology of X, then $h(T)$ is the same under both metrics*
2. *$h(T)$ is a conjugacy invariant*
3. *$h(T^n) = n \cdot h(T)$, for all $n \in \mathbb{Z}_{>0}$*
4. *If T is a homeomorphism, then $h(T^{-1}) = h(T)$. In this case, $h(T^n) = |n| \cdot h(T)$, for all $n \in \mathbb{Z}$*

Proof

(1): Let (X,d) and $(X,\tilde{d})$ denote the two metric spaces. Since both induce the same topology, the identity maps $i : (X,d) \longrightarrow (X,\tilde{d})$ and $j : (X,\tilde{d}) \longrightarrow (X,d)$ are continuous and hence by theorem 7.1.1 uniformly continuous. Fix $\varepsilon_1 > 0$. Then, by uniform continuity, we can subsequently find an $\varepsilon_2 > 0$ and $\varepsilon_3 > 0$ such that for all $x, y \in X$:

$$\begin{aligned} d(x,y) < \varepsilon_1 &\;\;\Rightarrow\;\; \tilde{d}(x,y) < \varepsilon_2 \\ \tilde{d}(x,y) < \varepsilon_2 &\;\;\Rightarrow\;\; d(x,y) < \varepsilon_3 \end{aligned}$$

Let A be an (n,ε_1)-spanning set for T under d. Then A is also (n,ε_2)-spanning for T under $\tilde{d}$. Analogously, every (n,ε_2)-spanning set for T under $\tilde{d}$ is (n,ε_3)-spanning for T under d. Therefore,

$$span(n,\varepsilon_3,T,d) \leq span(n,\varepsilon_2,T,\tilde{d}) \leq span(n,\varepsilon_1,T,d)$$

where the fourth argument emphasizes the metric. By subsequently taking the log, dividing by n, taking the limit for $n \to \infty$ and the limit for $\varepsilon_1 \downarrow 0$ (so that $\varepsilon_2, \varepsilon_3 \downarrow 0$) in these inequalities, we obtain the desired result.
(2): Suppose that $S : Y \longrightarrow Y$ is topologically conjugate to T, where Y is a

compact metric space, with conjugacy $\psi : Y \longrightarrow X$. Let d_X be the metric on X. Then the map $\tilde{d_Y}$ defined by $\tilde{d_Y}(x, y) = d_X(\psi(x), \psi(y))$ defines a metric on Y which induces the topology of Y. Indeed, let $B_{d_Y}(y_0, \eta)$ be a basis element of the topology of Y. Then $\psi(B_{d_Y}(y_0, \eta))$ is open in X, as ψ is an open map. Hence, there is an open ball $B_{d_X}(\psi(y_0), \tilde{\eta}) \subset \psi(B_{d_Y}(y_0, \eta))$, since the open balls are a basis for the topology of X. Now, $\psi^{-1}(B_{d_X}(\psi(y_0), \tilde{\eta})) \subset B_{d_Y}(y_0, \eta)$ and

$$y \in \psi^{-1}(B_{d_X}(\psi(y_0), \tilde{\eta})) \Leftrightarrow \tilde{d_Y}(y_0, y) = d_X(\psi(y), \psi(y_0)) < \tilde{\eta}$$

hence, $B_{\tilde{d_Y}}(y_0, \tilde{\eta}) \subset B_{d_Y}(y_0, \eta)$. Analogously, for any $y_0 \in Y$ and $\tilde{\delta} > 0$, there is a $\delta > 0$ such that

$$B_{d_Y}(y_0, \delta) \subset B_{\tilde{d_Y}}(y_0, \tilde{\delta})$$

Thus, $\tilde{d_Y}$ induces the topology of Y.
Now, for $x_1, x_2 \in X$, we have:

$$\begin{aligned} d_X(T(x_1), T(x_2)) &= d_X(T \circ \psi(y_1), T \circ \psi(y_2)) \\ &= d_X(\psi \circ S(y_1), \psi \circ S(y_2)) \\ &= \tilde{d_Y}(S(y_1), S(y_2)) \end{aligned}$$

Where we have used that $x_1 = \psi(y_1)$, $x_2 = \psi(y_2)$ for some $y_1, y_2 \in Y$, since ψ is a bijection. Thus, we see that the $n-$diameters (in the metrics d_X and $\tilde{d_Y}$) remain the same under ψ. Hence, $cov(n, \varepsilon, T, d_X) = cov(n, \varepsilon, S, \tilde{d_Y})$, for all $\varepsilon > 0$ and $n \in \mathbb{Z}_{\geq 0}$. By (1), $h(S)$ is the same under d_Y and $\tilde{d_Y}$, so it now easily follows that $h(S) = h(T)$.
(3): Observe that

$$\max_{1 \leq i \leq m-1} d(T^{ni}(x), T^{ni}(y)) \leq \max_{1 \leq j \leq nm-1} d(T^j(x), T^j(y))$$

therefore, $span(m, \varepsilon, T^n) \leq span(nm, \varepsilon, T)$. This implies $\frac{1}{m} span(m, \varepsilon, T^n) \leq \frac{n}{nm} span(nm, \varepsilon, T)$, thus $h(T^n) \leq nh(T)$.
Fix $\varepsilon > 0$. Then, by uniform continuity of T^i on X (c.f. theorem 7.1.1), we can find a $\delta > 0$ such that $d(T^i(x), T^i(y)) < \varepsilon$ for all $x, y \in X$ satisfying $d(x, y) < \delta$ and $i = 0, \ldots, n-1$. Hence, for $x, y \in X$ satisfying $d(x, y) < \delta$, we find $d_n(x, y) < \varepsilon$. Let A be an (m, δ)-spanning set for T^n. Then, for all $x \in X$ there exists some $a \in A$ such that

$$\max_{0 \leq i \leq m-1} d(T^{ni}(x), T^{ni}(a)) < \delta$$

But then, by the above,

$$\max_{0\leq k\leq n-1}\max_{0\leq i\leq m-1} d(T^{ni+k}(x), T^{ni+k}(a)) = \max_{0\leq j\leq nm-1} d(T^j(x), T^j(a)) < \varepsilon$$

Thus, $span(mn, \varepsilon, T) \leq span(m, \delta, T^n)$ and it follows that $h(T^n) \geq nh(T)$
(4): Let A be an (n, ε)-separated set for T. Then, for any $x, y \in A$ $d_n(x, y) > \varepsilon$. But then,

$$\max_{0\leq i\leq n-1} d(T^{-i}(T^{n-1}(x)), T^{-i}(T^{n-1}(y))) = \max_{0\leq i\leq n-1} d(T^i(x), T^i(y)) > \varepsilon$$

so $T^{n-1}A$ is an (n, ε)-separated set for T^{-1} of the same cardinality. Conversely, every (n, ε)-separated set B for T^{-1} gives the (n, ε)-separated set $T^{-(n-1)}B$ for T. Thus, $sep(n, \varepsilon, T) = sep(n, \varepsilon, T^{-1})$ and the first statement readily follows. The second statement is a consequence of (3). □

Exercise 7.2.5 Show that assertion (4) of theorem 7.2.4, i.e. $h(T^{-1}) = h(T)$, still holds if X is merely assumed to be compact.

Exercise 7.2.6 Let (X, d_X), (Y, d_Y) be compact metric spaces and $T : X \longrightarrow X$, $S : Y \longrightarrow Y$ continuous. Show that $h(T \times S) = h(T) + h(S)$. (Hint: first show that the metric $d = max\{d_X, d_Y\}$ induces the product topology on $X \times Y$).

7.3 Examples

In this section we provide a few detailed examples to get some familiarity with the material discussed in this chapter. We derive some simple properties of the dynamical systems presented below, but leave most of the good stuff for you to discover in the exercises. After all, the proof of the pudding is in the eating!

Example. (The Quadratic Map) Consider the following biological population model. Suppose we are studying a colony of penguins and wish to model the evolution of the population through time. We presume that there is a certain limiting population level, P^*, which cannot be exceeded. Whenever the population level is low, it is supposed to increase, since there is plenty of fish to go around. If, on the other hand, population is near the upper limit level P^*, it is expected to decrease as a result of overcrowding. We arrive at the following simple model for the population at time n, denoted by P_n:

$$P_{n+1} = \lambda P_n(P^* - P_n)$$

Where λ is a 'speed' parameter. If we now set $P^* = 1$, we can think of P_n as the population at time n as a percentage of the limiting population level. Then, writing $x = P_0$, the population dynamics are described by iteration of the following map:

Definition 7.3.1 *The Quadratic Map with parameter λ, $Q_\lambda : \mathbb{R} \longrightarrow \mathbb{R}$, is defined to be*

$$Q_\lambda(x) = \lambda x(1 - x)$$

Of course, in the context of the biological model formulated above, our attention is restricted to the interval $[0, 1]$ since the population percentage cannot lie outside this interval.
The quadratic map is deceivingly simple, as it in fact displays a wide range of interesting dynamics, such as periodicity, topological transitivity, bifurcations and chaos.

Let us take a closer look at the case $\lambda > 4$. For $x \in (-\infty, 0) \cup (1, \infty)$ it can be shown that $Q^n_\lambda(x) \to -\infty$ as $n \to \infty$. The interesting dynamics occur in the interval $[0, 1]$. Figure 7.2 shows a phase portrait for Q_λ on $[0, 1]$, which is nothing more than the graph of Q_λ together with the graph of the line $y = x$. In a phase portrait the trajectory of a point under iteration of the map Q_λ is easily visualized by iteratively projecting from the line $y = x$ to the graph and vice versa. From our phase portrait for Q_λ it is immediately visible that there is a 'leak' in our interval, i.e. the subset $A_1 = \{x \in [0, 1] | Q_\lambda(x) \notin [0, 1]\}$ is non-empty. Define $A_n = \{x \in [0, 1] | Q^i_\lambda(x) \in [0, 1] \text{ for } 0 \leq i \leq n - 1, Q^n_\lambda(x) \notin [0, 1]\}$ and set $C = [0, 1] - \cup_{n=1}^{\infty} A_n$. This procedure is reminiscent of the construction of the 'classical' Cantor set, by removing middle-thirds in subsequent steps. In the exercises below it is shown that C is indeed a Cantor set.

Figure 7.2: Phase portrait of the quadratic map Q_λ for $\lambda > 4$ on $[0, 1]$. The arrows indicate the (partial) trajectory of a point in $[0, 1]$. The dotted lines divide $[0, 1]$ in three parts: I_0, A_1 and I_1 (from left to right).

Example 7.3.1 . (Circle Rotations) Let S^1 denote the unit circle in $\mathbb{C}$. We define the rotation map $R_\theta : S^1 \longrightarrow S^1$ with parameter θ by $R_\theta(z) = e^{i\theta}z = e^{i(\theta+\omega)}$, where $\theta \in [0, 2\pi)$ and $z = e^{i\omega}$ for some $\omega \in [0, 2\pi)$. It is easily seen that R_θ is a homeomorphism for every θ. The rotation map is very similar to the earlier introduced shift map T_θ. This is not surprising, considering the fact that the interval $[0, 1]$ is homeomorphic to S^1 after identification of the endpoints 0 and 1.

Proposition 7.3.1 *Let $\theta \in [0, 2\pi)$ and $R_\theta : S^1 \longrightarrow S^1$ be the rotation map. If θ is rational, say $\theta = \frac{a}{b}$, then every $z \in S^1$ is periodic with period b. R_θ is minimal if and only if θ is irrational.*

Proof

The first statement follows trivially from the fact that $e^{2ni\pi} = 1$ for $n \in \mathbb{Z}$.

Suppose that θ is irrational. Fix $\varepsilon > 0$ and $z \in S^1$, so $z = e^{i\omega}$ for some $\omega \in [0, 2\pi)$. Then

$$\begin{aligned} R_\theta^n(z) = R_\theta^m(z) &\Leftrightarrow e^{i(\omega+n\theta)} = e^{i(\omega+m\theta)} \\ &\Leftrightarrow e^{i(n-m)\theta} = 1 \\ &\Leftrightarrow (n-m)\theta \in \mathbb{Z} \end{aligned}$$

Thus, the points $\{R_\theta^n(z) | n \in \mathbb{Z}\}$ are all distinct and it follows that $\{R_\theta^n(z)\}_{n=1}^\infty$ is an infinite sequence. By compactness, this sequence has a convergent subsequence, which is Cauchy. Therefore, we can find integers $n > m$ such that

$$d(R_\theta^n(z), R_\theta^m(z)) < \varepsilon$$

where d is the arc length distance function. Now, since R_θ is distance preserving with respect to this metric, we can set $l = n - m$ to obtain $d(R_\theta^l(z), z) < \varepsilon$. Also, by continuity of R_θ^l, R_θ^l maps the connected, closed arc from z to $R_\theta^l(z)$ onto the connected closed arc from $R_\theta^l(z)$ to $R_\theta^{2l}(z)$ and this one onto the arc connecting $R_\theta^{2l}(z)$ to $R_\theta^{3l}(z)$, etc. Since these arcs have positive and equal length, they cover S^1. The result now follows, since the arcs have length smaller than ε, and $\varepsilon > 0$ was arbitrary. $\square$

As mentioned in the proof, R_θ is an isometry with respect to the arc length distance function and this metric induces the topology on S^1. Hence, we see that $span(n, \varepsilon, R_\theta) = span(1, \varepsilon, R_\theta)$ for all $n \in \mathbb{Z}_{\geq 0}$ and it follows that $h(R_\theta) = 0$, for any $\theta \in [0, 2\pi)$.

Example 7.3.2 . (Bernoulli Shift Revisited) In this example we will reintroduce the Bernoulli shift in a topological context. Let $X_k = \{0, 1, ., k-1\}^{\mathbb{Z}}$ (or $X_k^+ = \{0, ., k-1\}^{\mathbb{N} \cup \{0\}}$) and recall that the *left shift* on X_k is defined by $L : X_k \longrightarrow X_k$, $L(x) = y$, $y_n = x_{n+1}$. We can define a topology on X_k (or X_k^+) which makes the shift map continuous (in fact, a homeomorphism in the case of X_k) as follows: we equip $\{0, \ldots, k-1\}$ with the discrete topology and subsequently give X_k the corresponding product topology. It is easy to see that the cylinder sets form a basis for this topology. By the Tychonoff theorem (see e.g. [Mu]), which asserts that an arbitrary product of compact spaces is compact in the product topology, our space X_k is compact. In the exercises you are asked to show that the metric

$$d(x,y) = 2^{-l} \text{ if } l = min\{|j| : x_j \neq y_j\}$$

induces the product topology, i.e. the open balls $B(x,r)$ with respect to this metric form a basis for the product topology. Here we set $d(x,y) = 0$ if $x = y$, this corresponds to the case $l = \infty$.
We will end this example by calculating the topological entropy of the full shift L.

Proposition 7.3.2 *The topological entropy of the full shift map is equal to* $h(L) = \log(k)$.

Proof

We will prove the proposition for the left shift on X_k^+, the case for the left shift on X_k is similar. Fix $0 < \varepsilon < 1$ and pick any $x = \{x_i\}_{i=0}^{\infty}$, $y = \{y_i\}_{i=0}^{\infty} \in X_k^+$. Notice that if at least one of the first n symbols in the itineraries of x and y differ, then

$$d_n(x,y) = \max_{0 \leq i \leq n-1} d(T^i(x), T^i(y)) = 1 > \varepsilon$$

where we have used the metric d on X_k^+ defined above. Thus, the set A_n consisting of sequences $a \in X_k^+$ such that $a_i \in \{0,\ldots,k-1\}$ for $0 \leq i \leq n-1$ and $a_j = 0$ for $j \geq n$ is (n,ε)-separated. Explicitly, $a \in A_n$ is of the form

$$a_0 a_1 a_2 \ldots a_{n-1} 000 \ldots$$

Since there are k^n possibilities to choose the first n symbols of an itinerary, we obtain $sep(n,\varepsilon,L) \geq k^n$. Hence,

$$h(T) = \lim_{\varepsilon \downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log(sep(n,\varepsilon,L)) \geq \log(k)$$

To prove the reverse inequality, take $l \in \mathbb{Z}_{>0}$ such that $2^{-l} < \varepsilon$. Then A_{n+l} is an (n,ε,L)-spanning set, since for every $x \in X_k^+$ there is some $a \in A_{n+l}$ for which the first $n+l$ symbols coincide. In other words, $d(x,a) < 2^{-l} < \varepsilon$. Therefore,

$$h(T) = \lim_{\varepsilon \downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log(span(n,\varepsilon,L)) \leq \lim_{\varepsilon \downarrow 0} \lim_{n\to\infty} \frac{n+l}{n} \log(k) = \log(k)$$

and our proof is complete. □

Figure 7.3: Phase portrait of the teepee map T on $[0, 1]$.

Example 7.3.3 . (Symbolic Dynamics) The Bernoulli shift on the bisequence space X_k is a topological dynamical system whose dynamics are quite well understood. The subject of symbolic dynamics is devoted to using this knowledge to unravel the dynamical properties of more difficult systems. The scheme works as follows. Let $T : X \longrightarrow X$ be a topological dynamical system and let $\alpha = \{A_0, \dots, A_{k-1}\}$ be a partition of X such that $A_i \cap A_j = \emptyset$ for $i \neq j$ (Note the difference with the earlier introduced notion of a partition). We assume that T is invertible. Now, for $x \in X$, we let $\phi_i(x)$ be the index of the element of α which contains $T^i(x)$, $i \in \mathbb{Z}$. This defines a map $\phi : X \longrightarrow X_k$, called the *Itinerary map generated by T on α*,

$$\phi(x) = \{\phi_i(x)\}_{i=-\infty}^{\infty}$$

The image of x under ϕ is called the *itinerary of x*. Note that ϕ satisfies $\phi \circ T = L \circ \phi$. Of course, if T is not invertible, we may apply the same procedure by replacing X_k by X_k^+.

We will now apply the above procedure to determine the number of periodic points of the teepee (or tent) map $T : [0, 1] \longrightarrow [0, 1]$, defined by (see figure 7.3)

$$T(x) = \begin{cases} 2x \text{ if } 0 \leq x \leq \frac{1}{2} \\ 2(1-x) \text{ if } \frac{1}{2} \leq x \leq 1 \end{cases}$$

Let $I_0 = [0, \frac{1}{2}]$, $I_1 = [\frac{1}{2}, 1]$. We would like to define an itinerary map

$\psi : [0,1] \longrightarrow X_2^+$ generated on $\{I_0, I_1\}$. Unfortunately, the fact that $I_0 \cap I_1 \neq \emptyset$ causes some complications. Indeed, we have an ambiguous choice of the itinerary for $x = \frac{1}{2}$, since it can be described by either $(01000\ldots)$ or $(11000\ldots)$, or any other point in $[0,1]$ which will end up in $x = \frac{1}{2}$ upon iteration of T. We will therefore identify any pair of sequences of the form $(a_0a_1\ldots a_l01000\ldots)$ and $(a_0a_1\ldots a_l11000\ldots)$ and replace the two sequences by a single equivalence class. The resulting space will be denoted by $\tilde{X}_2^+$ and the resulting itinerary map again by $\psi : [0,1] \longrightarrow \tilde{X}_2^+$.
We will first show that ψ is injective. Suppose that there are $x, y \in [0,1]$ with $\psi(x) = \psi(y)$. Then $T^n(x)$ and $T^n(y)$ lie on the same side of $x = \frac{1}{2}$ for all $n \in \mathbb{Z}_{\geq 0}$. Suppose that $x \neq y$. Then, for all $n \in \mathbb{Z}_{\geq 0}$, $T^n([x,y])$ does not contain the point $x = \frac{1}{2}$. But then no rational numbers of the form $\frac{p}{2^n}$, $p \in \{1, 2, 3, \ldots, 2^n - 1\}$, are in $[x,y]$ for all $n \in \mathbb{Z}_{>0}$. Since these form a dense set in $[0,1]$, we obtain a contradiction. We conclude that $x = y$, ψ is one-to-one.
Now let us show that ψ is also surjective. Let $a = \{a_i\}_{i=0}^{\infty} \in \tilde{X}_2^+$. If a is one of the aforementioned equivalence classes, we use the '0' element to represent a. Now, define

$$\begin{aligned} A_n &= \{x \in [0,1] | x \in I_{a_0}, T(x) \in I_{a_1}, \ldots, T^n(x) \in I_{a_n}\} \\ &= I_{a_0} \cap T^{-1}I_{a_1} \cap \ldots \cap T^{-n}I_{a_n} \end{aligned}$$

Since T is continuous, $T^{-i}I_{a_i}$ is closed for all i and therefore A_n is closed for $n \in \mathbb{Z}_{\geq 0}$. As $[0,1]$ is a compact metric space, it follows from theorem 7.1.1 that A_n is compact. Moreover, $\{A_n\}_{n=0}^{\infty}$ forms a decreasing sequence, in the sense that $A_{n+1} \subset A_n$. Therefore, the intersection $\cap_{n=0}^{\infty} A_n$ is not empty and a is precisely the itinerary of $x \in \cap_{n=0}^{\infty} A_n$. We conclude that ψ is onto.
Now, since $\psi \circ T = L \circ \psi$, with $L : \tilde{X}_2^+ \longrightarrow \tilde{X}_2^+$ the (induced) left shift, we have found a bijection between the periodic points of L on $\tilde{X}_2^+$ and those of T on $[0,1]$. The periodic points for T are quite difficult to find, but the periodic points of L are easily identified. First observe that there are no periodic points in the earlier defined equivalence classes in $\tilde{X}_2^+$ since these points are eventually mapped to 0. Now, any periodic point a of L of period n in X_2^+ must have an itinerary of the form

$$(a_0a_1\ldots a_{n-1}a_0a_1\ldots a_{n-1}a_0a_1\ldots)$$

Hence, T has 2^n periodic points of period n in $[0,1]$.

Exercise 7.3.1 In this exercise you are asked to prove some properties of the shift map. Let $X_k = \{0, 1, ., k-1\}^{\mathbb{Z}}$ and $X_k^+ = \{0, 1, ., k-1\}^{\mathbb{N}\cup\{0\}}$.

a. Show that the map $d : X_k \times X_k \longrightarrow X_k$, defined by

$$d(x, y) = 2^{-l} \text{ if } l = min\{|j| : x_j \neq y_j\}$$

is a metric on X_k that induces the product topology on X_k.

b. Show that the map $d^* : X_k \times X_k \longrightarrow X_k$, defined by

$$d^*(x, y) = \sum_{n=-\infty}^{\infty} \frac{|x_n - y_n|}{2^{|n|}}$$

is a metric on X_k that induces the product topology on X_k.

c. Show that the one-sided shift on X_k^+ is continuous. Show that the two-sided shift on X_k is a homeomorphism.

d. Show that the two-sided shift $L : X_k \longrightarrow X_k$ is expansive.

Exercise 7.3.2 Let Q_λ be the quadratic map and let C, A_1 be as in the first example. Set $[0, 1] - A_1 = I_0 \cup I_1$, where I_0 is the interval left of A_1 and I_1 the interval to the right. We assume that $\lambda > 2+\sqrt{5}$, so that $|Q'_\lambda(x)| > 1$ for all $x \in I_0 \cup I_1$ (check this!). Let $X_2^+ = \{0, 1\}^{\mathbb{N}}$ and let $\phi : C \longrightarrow X_2^+$ be the itinerary map generated by Q_λ on $\{I_0, I_1\}$. This exercise serves as another demonstration of the usefulness of symbolic dynamics.

a. Show that ϕ is a homeomorphism.

b. Show that ϕ is a topological conjugacy between Q_λ and the one-sided shift map on X_2^+.

c. Prove that Q_λ has exactly 2^n periodic points of period n in C. Show that the periodic points of Q_λ are dense in C. Finally, prove that Q_λ is topologically transitive on C.

Exercise 7.3.3 Let Q_λ be the quadratic map and let C be as in the first example. This exercise shows that C is a Cantor set, i.e. a closed, totally disconnected and perfect subset of $[0, 1]$. As in the previous exercise, we will assume that $\lambda > 2 + \sqrt{5}$.

a. Prove that C is closed.

b. Show that C is totally disconnected, i.e. the only connected subspaces of C are one-point sets.

c. Demonstrate that C is perfect, i.e. every point is a limit point of C.

Chapter 8

The Variational Principle

In this chapter we will establish a powerful relationship between measure theoretic and topological entropy, known as the *Variational Principle*. It asserts that for a continuous transformation T of a compact metric space X the topological entropy is given by $h(T) = \sup\{h_\mu(T) | \mu \in M(X,T)\}$. To prove this statement, we will proceed along the shortest and most popular route to victory, provided by [M][1]. The proof is at times quite technical and is therefore divided into several digestable pieces.

8.1 Main Theorem

For the first part of the proof of the Variational Principle we will only use some properties of measure theoretic entropy. All the necessary ingredients are listed in the following lemma:

Lemma 8.1.1 *Let α, β, γ be finite partitions of X and T a measure preserving transformation of the probability space $(X, \mathcal{F}, \mu)$. Then,*

1. *$H_\mu(\alpha) \leq \log(N(\alpha))$, where $N(\alpha)$ is the number of sets in the partition of non-zero measure. Equality holds only if $\mu(A) = \frac{1}{N(\alpha)}$, for all $A \in \alpha$ with $\mu(A) > 0$*

2. *If $\alpha \leq \beta$, then $h_\mu(T, \alpha) \leq h_\mu(T, \beta)$*

[1] Our exposition of Misiurewicz's proof follows [W] to a certain extent

3. $h_\mu(T,\alpha) \le h_\mu(T,\gamma) + H_\mu(\alpha|\gamma)$

4. $h_\mu(T^n) = n \cdot h_\mu(T)$, *for all* $n \in \mathbb{Z}_{>0}$

Proofs

(1): This is a straighforward consequence of the concavity of the function $f : [0,\infty) \longrightarrow \mathbb{R}$, defined by $f(t) = -t\log(t)$ $(t > 0)$, $f(0) = 0$. In this notation, $H_\mu(\alpha) = \sum_{A\in\alpha} f(\mu(A))$.
(2): $\alpha \le \beta$ implies that for every $B \in \beta$, there some $A \in \alpha$ such that $B \subset A$ and hence also $T^{-i}B \subset T^{-i}A$, $i \in \mathbb{Z}_{\ge 0}$. Therefore, for $n \ge 1$

$$\bigvee_{i=0}^{n-1} T^{-i}\alpha \le \bigvee_{i=0}^{n-1} T^{-i}\beta$$

The result now follows from proposition .
(3): By proposition (4.2.1)(e) and (b), respectively,

$$\begin{aligned} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha) &\le H(\bigvee_{i=0}^{n-1} T^{-i}\gamma \vee \bigvee_{i=0}^{n-1} T^{-i}\alpha) \\ &= H(\bigvee_{i=0}^{n-1} T^{-i}\gamma) + H(\bigvee_{i=0}^{n-1} T^{-i}\alpha | \bigvee_{i=0}^{n-1} T^{-i}\gamma) \end{aligned}$$

Now, by successively applying proposition (4.2.1) (d), (g) and finally using the fact that T is measure preserving, we get

$$\begin{aligned} H(\bigvee_{i=0}^{n-1} T^{-i}\alpha | \bigvee_{i=0}^{n-1} T^{-i}\gamma) &\le \sum_{i=0}^{n-1} H(T^{-i}\alpha | \bigvee_{i=0}^{n-1} T^{-i}\gamma) \\ &\le \sum_{i=0}^{n-1} H(T^{-i}\alpha | T^{-i}\gamma) \\ &= nH(\alpha|\gamma) \end{aligned}$$

Combining the inequalities, dividing both sides by n and taking the limit for $n \to \infty$ gives the desired result.

(4): Note that

$$\begin{aligned}
h(T^n, \bigvee_{i=0}^{n-1} T^{-i}\alpha) &= \lim_{k\to\infty} \frac{1}{k} H(\bigvee_{j=0}^{k-1} T^{-nj}(\bigvee_{i=0}^{n-1} T^{-i}\alpha)) \\
&= \lim_{k\to\infty} \frac{n}{kn} H(\bigvee_{i=0}^{kn-1} T^{-i}\alpha) \\
&= \lim_{m\to\infty} \frac{n}{m} H(\bigvee_{i=0}^{m-1} T^{-i}\alpha) \\
&= nh(T,\alpha)
\end{aligned}$$

Notice that $\{\bigvee_{i=0}^{n-1} T^{-i}\alpha \mid \alpha \text{ a partition of } X\} \subset \{\beta \mid \beta \text{ a partition of } X\}$. Hence,

$$\begin{aligned}
nh(T) &= n \sup_{\alpha} h(T,\alpha) \\
&= \sup_{\alpha} h(T^n, \bigvee_{i=0}^{n-1} T^{-i}\alpha) \\
&\le \sup_{\alpha} h(T^n, \alpha) \\
&= h(T^n)
\end{aligned}$$

Finally, since $\alpha \le \bigvee_{i=0}^{n-1} T^{-i}\alpha$, we obtain by (2),

$$h(T^n,\alpha) \le h(T^n, \bigvee_{i=0}^{n-1} T^{-i}\alpha) = nh(T,\alpha)$$

So $h(T^n) \le nh(T)$. This concludes our proof. $\square$

Exercise 8.1.1 Use lemma 8.1.1 and proposition (4.2.1) to show that for an invertible measure preserving transformation T on $(X, \mathcal{F}, \mu)$:

$$h(T^n) = |n| h(T), \text{for all } n \in \mathbb{Z}$$

We are now ready to prove the first part of the theorem.

Theorem 8.1.1 *Let $T : X \longrightarrow X$ be a continuous transformation of a compact metric space. Then $h(T) \geq \sup\{h_\mu(T) | \mu \in M(X,T)\}$.*

Proof

Fix $\mu \in M(X,T)$. Let $\alpha = \{A_1, \ldots, A_n\}$ be a finite partition of X. Pick $\varepsilon > 0$ such that $\varepsilon < \frac{1}{n \log n}$. By theorem 16, μ is regular, so we can find closed sets $B_i \subset A_i$ such that $\mu(A_i - B_i) < \varepsilon$, $i = 1, \ldots, n$. Define $B_0 = X - \cup_{i=1}^n B_i$ and let β be the partition $\beta = \{B_0, \ldots, B_n\}$. Then, writing $f(t) = -t\log(t)$,

$$\begin{aligned} H_\mu(\alpha|\beta) &= \sum_{i=0}^n \sum_{j=1}^n \mu(B_i) f(\frac{\mu(B_i \cap A_j)}{\mu(B_i)}) \\ &= \mu(B_0) \sum_{j=1}^n f(\frac{\mu(B_0 \cap A_j)}{\mu(B_0)}) \end{aligned}$$

in the final step we used the fact that for $i \geq 1$,

$$\begin{aligned} B_i \cap A_i &= B_i \\ B_i \cap A_j &= \emptyset \text{ if } i \neq j \end{aligned}$$

We can define a σ-algebra on B_0 by $\mathcal{F} \cap B_0 = \{F \cap B_0 | F \in \mathcal{F}\}$ and define the conditional measure $\mu(\cdot|B_0) : \mathcal{F} \cap B_0 \longrightarrow [0,1]$ by

$$\mu(A|B_0) = \frac{\mu(A \cap B_0)}{\mu(B_0)}$$

It is easy to check that $(B_0, \mathcal{F} \cap B_0, \mu(\cdot|B_0))$ is a probability space and $\mu(\cdot|B_0) \in M(B_0, T|_{B_0})$. Now, noting that $\alpha_0 := \{A \cap B_0 | A \in \alpha\}$ is a partition of B_0 under $\mu(\cdot|B_0)$, we get

$$H_\mu(\alpha|\beta) = \mu(B_0) \sum_{j=1}^n f(\frac{\mu(B_0 \cap A_j)}{\mu(B_0)}) = \mu(B_0) H_{\mu(\cdot|B_0)}(\alpha_0)$$

Hence, we can apply (1) of lemma 8.1.1 and use the fact that

$$\mu(B_0) = \mu(X - \cup_{i=1}^n B_i) = \mu(\cup_{i=1}^n A_i - \cup_{i=1}^n B_i) = \mu(\cup_{i=1}(A_i - B_i)) < n\varepsilon$$

to obtain

$$H_\mu(\alpha|\beta) \le \mu(B_0)\log(n) < n\varepsilon\log(n) < 1$$

Define for each i, $i = 1, \ldots, n$, the open set C_i by $C_i = B_0 \cup B_i = X - \cup_{j\neq i} B_j$. Then we can define an open cover of X by $\gamma = \{C_1, \ldots, C_n\}$. By (1) of lemma 8.1.1 we have for $m \ge 1$, in the notation of the lemma, $H_\mu(\bigvee_{i=0}^{m-1} T^{-i}\beta) \le \log(N(\bigvee_{i=0}^{m-1} T^{-i}\beta))$. Note that γ is not necessarily a partition, but by the proof of (1) of lemma 8.1.1, we still have $H_\mu(\bigvee_{i=0}^{m-1} T^{-i}\beta) \le \log(2^m N(\bigvee_{i=0}^{m-1} T^{-i}\gamma))$, since β contains (at most) one more set of non-zero measure.
Thus, it follows that

$$h_\mu(T,\beta) \le h(T,\gamma) + \log 2 \le h(T) + \log 2$$

and by (3) of lemma 8.1.1 we get obtain

$$h_\mu(T,\alpha) \le h_\mu(T,\beta) + H_\mu(\alpha|\beta) \le h(T) + \log 2 + 1$$

Note that if $\mu \in M(X,T)$, then also $\mu \in M(X,T^m)$, so the above inequality holds for T^m as well. Applying (4) of lemma 8.1.1 and (3) of theorem 7.2.4 leads us to $mh_\mu(T) \le mh(T) + \log 2 + 1$. By dividing by m, taking the limit for $m \to \infty$ and taking the supremum over all $\mu \in M(X,T)$, we obtain the desired result. □

We will now finish the proof of the Variational Principle by proving the opposite inequality. This is the hard part of the proof, since it does not only require more machinery, but also a clever trick.

Lemma 8.1.2 *Let X be a compact metric space and α a finite partition of X. Then, for any $\mu, \nu \in M(X,T)$ and $p \in [0,1]$ we have $H_{p\mu+(1-p)\nu}(\alpha) \ge pH_\mu(\alpha) + (1-p)H_\nu(\alpha)$*

Proof

Define the function $f : [0,\infty) \longrightarrow \mathbb{R}$ by $f(t) = -t\log(t)$ $(t > 0)$, $f(0) = 0$. Then f is concave and so, for A measurable, we have:

$$0 \le f(p\mu(A) + (1-p)\nu(A)) - pf(\mu(A)) - (1-p)f(\nu(A))$$

The result now follows easily. □

Exercise 8.1.2 Suppose X is a compact metric space and $T : X \longrightarrow X$ a continuous transformation. Use (the proof of) lemma 8.1.2 to show that for any $\mu, \nu \in M(X,T)$ and $p \in [0,1]$ we have $h_{p\mu+(1-p)\nu}(T) \geq ph_\mu(T) + (1-p)h_\nu(T)$.

Exercise 8.1.3 Improve the result in the exercise above by showing that we can replace the inequality by an equality sign, i.e. for any $\mu, \nu \in M(X,T)$ and $p \in [0,1]$ we have $h_{p\mu+(1-p)\nu}(T) = ph_\mu(T) + (1-p)h_\nu(T)$.

Recall that the boundary of a set A is defined by $\partial A = \overline{A} - A$.

Lemma 8.1.3 *Let X be a compact metric space and $\mu \in M(X)$. Then,*

1. *For any $x \in X$ and $\delta > 0$, there is a $0 < \eta < \delta$ such that $\mu(\partial B(x,\eta)) = 0$*
2. *For any $\delta > 0$, there is a finite partition $\alpha = \{A_1, \ldots, A_n\}$ of X such that $diam(A_j) < \delta$ and $\mu(\partial A_j) = 0$, for all j*
3. *If $T : X \longrightarrow X$ is continuous, $\mu \in M(X,T)$ and $A_j \subset X$ measurable such that $\mu(\partial A_j) = 0$, $j = 0, \ldots, n-1$, then $\mu(\partial(\cap_{j=0}^{n-1} T^{-j} A_j)) = 0$*
4. *If $\mu_n \to \mu$ in $M(X)$ and A is a measurable set such that $\mu(\partial A) = 0$, then $\lim_{n\to\infty} \mu_n(A) = \mu(A)$*

Proof

(1): Suppose the statement is false. Then there exists an $x \in X$ and $\delta > 0$ such that for all $0 < \eta < \delta$ $\mu(\partial B(x,\eta)) > 0$. Let $\{\eta_i\}_{i=1}^\infty$ be a sequence of distinct real numbers satisfying $0 < \eta_i < \delta$ and $\eta_i \uparrow \delta$ for $i \to \infty$. Then, $\mu(\cup_{i=1}^\infty \partial B(x,\eta_i)) = \sum_{i=1}^\infty \mu(\partial B(x,\eta_i)) = \infty$, contradicting the fact that μ is a probability measure on X.
(2): Fix $\delta > 0$. By (1), for each $x \in X$, we can find an $0 < \eta_x < \delta/2$ such that $\mu(\partial B(x,\eta_x)) = 0$. The collection $\{B(x,\eta_x) | x \in X\}$ forms an open cover of X, so by compactness there exists a finite subcover which we denote by $\beta = \{B_1, \ldots, B_n\}$. Define α by letting $A_1 = \overline{B_1}$ and for $0 < j \leq n$ let $A_j = \overline{B_j} - (\cup_{k=1}^{j-1} \overline{B_k})$. Then α is a partition of X, $diam(A_j) < diam(B_j) < \delta$ and $\mu(\partial A_j) \leq \mu(\cup_{i=1}^n \partial B_i) = 0$, since $\partial A_j \subset \cup_{i=1}^n \partial B_i$.
(3): Let $x \in \partial(\cap_{j=0}^{n-1} T^{-j} A_j)$. Then $x \in \overline{\cap_{j=0}^{n-1} T^{-j} A_j}$, but $x \notin \cap_{j=0}^{n-1} T^{-j} A_j$. That is, every open neighborhood of x intersects every $T^{-j} A_j$, but $x \notin T^{-k} A_k$

for some $0 \leq k \leq n-1$. Hence, $x \in \overline{T^{-k}A_k} - T^{-k}A_k$ and by continuity of T^k,

$$T^k(x) \in T^k(\overline{T^{-k}A_k}) - A_k \subset \overline{A_k} - A_k = \partial A_k$$

Thus, $\partial(\cap_{j=0}^{n-1} T^{-j}A_j) \subset \cup_{j=0}^{n-1} T^{-j}\partial A_j$ and the statement readily follows.
(4): Recall that $\mu_n \to \mu$ in $M(X)$ for $n \to \infty$ if and only if:

$$\lim_{n\to\infty} \int_X f(x)d\mu_n(x) = \int_X f(x)d\mu(x)$$

for all $f \in C(X)$. Let A be as stated and define $A_k = \{x \in X | d(x, \overline{A}) > \frac{1}{k}\}$, $k \in \mathbb{Z}_{>0}$. Then, since $\overline{A}$ and $X - A_k$ are closed, it follows from theorem 7.1.1 that for every k there exists a function $f_k \in C(X)$ such that $0 \leq f_k \leq 1$, $f_k(x) = 1$ for all $x \in \overline{A}$ and $f_k(x) = 0$ for all $x \in X - A_k$. Now, for each k,

$$\lim_{n\to\infty} \mu_n(\overline{A}) = \lim_{n\to\infty} \int_X 1_{\overline{A}}(x)d\mu_n(x) \leq \lim_{n\to\infty} \int_X f_k(x)d\mu_n(x) = \int_X f_k(x)d\mu(x)$$

Note that $f_k \downarrow 1_{\overline{A}}$ in μ-measure for $k \to \infty$, so by the Dominated Convergence Theorem,

$$\lim_{n\to\infty} \mu_n(\overline{A}) \leq \lim_{k\to\infty} \int_X f_k(x)d\mu(x) = \mu(\overline{A}) = \mu(A)$$

where the final inequality follows from the assumption $\mu(\partial A) = 0$. The proof of the opposite inequality is similar. □

Lemma 8.1.4 *Let q, n be integers such that $1 < q < n$. Define, for $0 \leq j \leq q-1$, $a(j) = \lfloor \frac{n-j}{q} \rfloor$, where $\lfloor\rfloor$ means taking the integer part. Then we have the following*

1. $a(0) \geq a(1) \geq \cdots \geq a(q-1)$
2. *Fix $0 \leq j \leq q-1$. Define*

$$S_j = \{0, 1, \ldots, j-1, j + a(j)q, j + a(j)q + 1, \ldots, n-1\}$$

Then

$$\{0, 1, \ldots, n-1\} = \{j + rq + i | 0 \leq r \leq a(j) - 1, 0 \leq i \leq q-1\} \cup S_j$$

and $card(S_j) \leq 2q$.

3. *For each* $0 \leq j \leq q-1$, $(a(j)-1)q+j \leq \lfloor \frac{n-j}{q} - 1 \rfloor q + j \leq n-q$. *The numbers* $\{j+rq | 0 \leq j \leq q-1, 0 \leq r \leq a(j)-1\}$ *are distinct and no greater than* $n-q$.

The three statements are clear after a little thought.

Example 8.1.1 Let us work out what is going on in lemma 8.1.4 for $n = 10$, $q = 4$. Then $a(0) = \lfloor \frac{10-0}{4} \rfloor = 2$ and similarly, $a(1) = 2$, $a(2) = 2$ and $a(3) = 1$. This gives us $S_0 = \{8,9\}$, $S_1 = \{0,9\}$, $S_2 = \{0,1\}$ and $S_3 = \{0,1,2,7,8,9\}$. For example, S_3 is obtained by taking all nonnegative integers smaller than 3 and adding the numbers $3+a(3)q = 7$, $3+a(3)q+1 = 8$ and $3+a(3)q+2 = 9$. Note that $card(S_j) \leq 8$. One can readily check all properties stated in the lemma, e.g. $(a(3)-1)q+3 \leq \lfloor \frac{10-3}{4} - 1 \rfloor \cdot 4 + 3 = 3 \leq 6 = n-q$.

We shall now finish the proof of our main theorem. The proof is presented in a logical order, which is in this case not quite the same as *thinking* order. Therefore, before plunging into a formal proof, we will briefly discuss the main ideas.
We would like to construct a Borel probability measure μ with measure theoretic entropy $h_\mu(T) \geq sep(\varepsilon, T) := \lim_{n\to\infty} \frac{1}{n} \log(sep(n,\varepsilon,T))$. To do this, we first find a measure σ_n with $H_{\sigma_n}(\bigvee_{i=0}^{n-1} T^{-i}\alpha)$ equal to $\log(sep(n,\varepsilon,T))$, where α is a suitably chosen partition. This is not too difficult. The problem is to find an element of $M(X,T)$ with this property. Theorem (6.1.5) suggest a suitable measure μ which can be obtained from the σ_n. The trick in lemma 8.1.4 plays a crucial role in getting from an estimate for H_{σ_n} to one for H_μ. The idea is to first remove the 'tails' S_j of the partition $\bigvee_{i=0}^{n-1} T^{-i}\alpha$, make a crude estimate for these tails and later add them back on again. Lemma 8.1.3 fills in the remaining technicalities.

Theorem 8.1.2 *Let* $T : X \longrightarrow X$ *be a continuous transformation of a compact metric space. Then* $h(T) \leq \sup\{h_\mu(T) | \mu \in M(X,T)\}$.

Proof

Fix $\varepsilon > 0$. For each n, let E_n be an (n,ε)-separated set of cardinality $sep(n,\varepsilon,T)$. Define $\sigma_n \in M(X)$ by $\sigma_n = (1/sep(n,\varepsilon,T)) \sum_{x\in E_n} \delta_x$,

where δ_x is the Dirac measure concentrated at x. Define $\mu_n \in M(X)$ by $\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \sigma_n \circ T^{-i}$. By theorem 19, $M(X)$ is compact, hence there exists a subsequence $\{n_j\}_{j=1}^{\infty}$ such that $\{\mu_{n_j}\}$ converges in $M(X)$ to some $\mu \in M(X)$. By theorem (6.1.5), $\mu \in M(X,T)$. We will show that $h_\mu(T) \geq sep(\varepsilon, T)$, from which the result clearly follows.
By (2) of lemma 8.1.3, we can find a μ-measurable partition $\alpha = \{A_1, \ldots, A_k\}$ of X such that $diam(A_j) < \varepsilon$ and $\mu(\partial A_j) = 0$, $j = 1, \ldots, k$. We may assume that every $x \in E_n$ is contained in some A_j. Now, if $A \in \bigvee_{i=0}^{n-1} T^{-i}\alpha$, then A cannot contain more than one element of E_n. Hence, $\sigma_n(A) = 0$ or $1/sep(n, \varepsilon, T)$. Since $\cup_{j=1}^{k} A_j$ contains all of E_n, we see that $H_{\sigma_n}(\bigvee_{i=0}^{n-1} T^{-i}\alpha) = \log(sep(n, \varepsilon, T))$.
Fix integers q, n with $1 < q < n$ and define for each $0 \leq j \leq q-1$ $a(j)$ as in lemma 8.1.4. Fix $0 \leq j \leq q-1$. Since

$$\bigvee_{i=0}^{n-1} T^{-i}\alpha = (\bigvee_{r=0}^{a(j)-1} T^{-(rq+j)}(\bigvee_{i=0}^{q-1} T^{-i}\alpha)) \vee (\bigvee_{i\in S_j} T^{-i}\alpha)$$

we find that

$$\begin{aligned}
\log(sep(n,\varepsilon,T)) &= H_{\sigma_n}(\bigvee_{i=0}^{n-1} T^{-i}\alpha) \\
&\leq \sum_{r=0}^{a(j)-1} H_{\sigma_n}(T^{-rq-j}(\bigvee_{i=0}^{q-1} T^{-i}\alpha)) + \sum_{i\in S_j} H_{\sigma_n}(T^{-i}\alpha) \\
&\leq \sum_{r=0}^{a(j)-1} H_{\sigma_n \circ T^{-(rq+j)}}(\bigvee_{i=0}^{q-1} T^{-i}\alpha) + 2q\log k
\end{aligned}$$

Here we used proposition (4.2.1)(d) and lemma 8.1.1. Now if we sum the above inequality over j and divide both sides by n, we obtain by (3) and (1) of lemma 8.1.4

$$\begin{aligned}
\frac{q}{n}\log(sep(n,\varepsilon,T)) &\leq \frac{1}{n}\sum_{l=0}^{n-1} H_{\sigma_n \circ T^{-l}}(\bigvee_{i=0}^{q-1} T^{-i}\alpha) + \frac{2q^2}{n}\log k \\
&\leq H_{\mu_n}(\bigvee_{i=0}^{q-1} T^{-i}\alpha) + \frac{2q^2}{n}\log k
\end{aligned}$$

By (3) of lemma 8.1.3, each atom A of $\bigvee_{i=0}^{q-1} T^{-i}\alpha$ has boudary of μ-measure zero, so by (4) of the same lemma, $\lim_{j\to\infty} \mu_{n_j}(A) = \mu(A)$. Hence, if we replace n by n_j in the above inequality and take the limit for $j \to \infty$ we obtain $q sep(\varepsilon, T) \leq H_\mu \bigvee_{i=0}^{q-1} T^{-i}\alpha$. If we now divide both sides by q and take the limit for $q \to \infty$, we get the desired inequality. □

Corollary 8.1.1 (The Variational Principle) *The topological entropy of a continuous transformation $T : X \longrightarrow X$ of a compact metric space X is given by $h(T) = \sup\{h_\mu(T) | \mu \in M(X,T)\}$*

To get a taste of the power of this statement, let us recast our proof of the invariance of topological entropy under conjugacy ((2) of theorem 7.2.4). Let $\phi : X_1 \to X_2$ denote the conjugacy. We note that $\mu \in M(X_1, T_1)$ if and only if $\mu \circ \phi^{-1} \in M(X_2, T_2)$ and we observe that $h_\mu(T_1) = h_{\mu\circ\phi^{-1}}(T_2)$. The result now follows from the Variational Principle. It is that simple.

8.2 Measures of Maximal Entropy

The Variational Principle suggests an educated way of choosing a Borel probability measure on X, namely one that maximizes the entropy of T.

Definition 8.2.1 *Let X be a compact metric space and $T : X \longrightarrow X$ be continuous. A measure $\mu \in M(X,T)$ is called a measure of maximal entropy if $h_\mu(T) = h(T)$. Let $M_{max}(X,T) = \{\mu \in M(X,T) | h_\mu(T) = h(T)\}$. If $M_{max}(X,T) = \{\mu\}$ then μ is called a unique measure of maximal entropy.*

Example. Recall that the topological entropy of the circle rotation R_θ is given by $h(R_\theta) = 0$. Since $h_\mu(R_\theta) \geq 0$ for all $\mu \in M(X, R_\theta)$, we see that *every* $\mu \in M(X, R_\theta)$ is a measure of maximal entropy, i.e. $M_{max}(X, R_\theta) = M(X, R_\theta)$. More generally, we have $h(T) = 0$ and hence $M_{max}(X,T) = M(X,T)$ for any continuous isometry $T : X \longrightarrow X$.

Measures of maximal entropy are closely connected to (uniquely) ergodic measures, as will become apparent from the following theorem.

Theorem 8.2.1 *Let X be a compact metric space and $T : X \longrightarrow X$ be continuous. Then*

- *$M_{max}(X,T)$ is a convex set.*
- *If $h(T) < \infty$ then the extreme points of $M_{max}(X,T)$ are precisely the ergodic members of $M_{max}(X,T)$.*
- *If $h(T) = \infty$ then $M_{max}(X,T) \neq \emptyset$. If, moreover, T has a unique measure of maximal entropy, then T is uniquely ergodic.*
- *A unique measure of maximal entropy is ergodic. Conversely, if T is uniquely ergodic, then T has a measure of maximal entropy.*

Proofs

(1): Let $p \in [0,1]$ and $\mu, \nu \in M_{max}(X,T)$. Then, by exercise 8.1.3,

$$h_{p\mu+(1-p)\nu}(T) = ph_\mu(T) + (1-p)h_\nu(T) = ph(T) + (1-p)h(T) = h(T)$$

Hence, $p\mu + (1-p)\nu \in M_{max}(X,T)$.
(2): Suppose $\mu \in M_{max}(X,T)$ is ergodic. Then, by theorem 6.1.6, μ cannot be written as a non-trivial convex combination of elements of $M(X,T)$. Since $M_{max}(X,T) \subset M(X,T)$, μ is an extreme point of $M_{max}(X,T)$. Conversely, suppose μ is an extreme point of $M(X,T)$ and suppose there is a $p \in (0,1)$ and $\nu_1, \nu_2 \in M(X,T)$ such that $\mu = p\nu_1 + (1-p)\nu_2$. By exercise 8.1.3, $h(T) = h_\mu(T) = ph_{\nu_1}(T) + (1-p)h_{\nu_2}(T)$. But by the Variational Principle, $h(T) = \sup\{h_\mu(T) | \mu \in M(X,T)\}$, so $h(T) = h_{\nu_1}(T) = h_{\nu_2}(T)$. In other words, $\nu_1, \nu_2 \in M_{max}(X,T)$, thus $\mu = \nu_1 = \nu_2$. Therefore, μ is also an extreme point of $M(X,T)$ and we conclude that μ is ergodic.
For the second part, suppose that $M_{max}(X,T) \neq \emptyset$.
(3): By the Variational Principle, for any $n \in \mathbb{Z}_{>0}$, we can find a $\mu_n \in M(X,T)$ such that $h_{\mu_n}(T) > 2^n$. Define $\mu \in M(X,T)$ by

$$\mu = \sum_{n=1}^{\infty} \frac{\mu_n}{2^n} = \sum_{n=1}^{N} \frac{\mu_n}{2^n} + \sum_{n=N}^{\infty} \frac{\mu_n}{2^n}$$

But then,

$$h_\mu(T) \geq \sum_{n=1}^{N} \frac{\mu_n}{2^n} > N$$

This holds for arbitrary $N \in \mathbb{Z}_{>0}$, so $h_\mu(T) = h(T) = \infty$ and μ is a measure of maximal entropy.

Now suppose that T has a unique measure of maximal entropy. Then, for any $\nu \in M(X,T)$, $h_{\mu/2+\nu/2}(T) = \frac{1}{2}h_\mu(T) + \frac{1}{2}h_\nu(T) = \infty$. Hence, $\mu = \nu$, $M(X,T) = \{\mu\}$.
(4): The first statement follows from (2), for $h(T) < \infty$, and (3), for $h(T) = \infty$. If T is uniquely ergodic, then $M(X,T) = \{\mu\}$ for some μ. By the Variational Principle, $h_\mu(T) = h(T)$. □

Example 8.2.1 For θ irrational R_θ is uniquely ergodic with respect to Lebesgue measure. By the theorem above, it follows that Lebesgue measure is also a unique measure of maximal entropy for R_θ.

Exercise 8.2.1 Let $X = \{0,1,\ldots,k-1\}^{\mathbb{Z}}$ and let $T : X \longrightarrow X$ be the full two-sided shift. Use proposition 7.3.2 to show that the uniform product measure is a unique measure of maximal entropy for T.

We end this section with a generalization of the above exercise.

Proposition 8.2.1 *Every expansive homeomorphism of a compact metric space has a measure of maximal entropy.*

Proof

Let $T : X \longrightarrow X$ be an expansive homeomorphism and let $\delta > 0$ be an expansive constant for T. Fix $0 < \varepsilon < \delta$. Define $\mu \in M(X,T)$ as in the proof of theorem 8.1.2. Then, by the proof, $h_\mu(T) \geq sep(\varepsilon, T)$. We will show that $h(T) = sep(\varepsilon, T)$. It then immediately follows that $\mu \in M_{max}(X,T)$.
Pick any $0 < \eta < \varepsilon$ and let A be an (n,η)-separated set of cardinality $sep(n,\eta,T)$. By expansiveness, we can find for any $x,y \in A$ some $k = k_{x,y} \in \mathbb{Z}$ such that

$$d(T^k(x), T^k(y)) > \varepsilon$$

By theorem 7.2.2, A is finite, so $l := \max\{|k_{x,y}| : x,y \in A\}$ is in $\mathbb{Z}_{>0}$. Now, by our choice of l, for $x,y \in T^{-l}A$ we have

$$\max_{0 \leq i \leq 2l+n-1} d(T^i(x), T^i(y)) > \varepsilon$$

So $T^{-l}A$ is an $(2l+n, \varepsilon, T)$-separated set of cardinality $sep(n, \eta, T)$. Hence,

$$\begin{aligned} sep(\eta, T) &= \lim_{n\to\infty} \frac{1}{n}\log(sep(n,\eta,T)) \\ &\le \lim_{n\to\infty} \frac{1}{n}\log(sep(2l+n,\varepsilon,T)) \\ &\le \lim_{n\to\infty} \frac{1}{2l+n}\log(sep(2l+n,\varepsilon,T)) \\ &= sep(\varepsilon, T) \end{aligned}$$

Conversely, if A is (n, ε)-separated then A is also (n, η)-separated, so

$$sep(\varepsilon, T) \le sep(\eta, T).$$

We conclude that $sep(\varepsilon, T) = sep(\eta, T)$. Since $0 < \eta < \varepsilon$ was arbitrary, this shows that $h(T) = \lim_{\eta\downarrow 0} sep(\eta, T) = sep(\varepsilon, T)$. This completes our proof. $\square$

From the proof we extract the following corollary.

Corollary 8.2.1 *Let $T : X \longrightarrow X$ be an expansive homeomorphism and let δ be an expansive constant for T. Then $h(T) = sep(\varepsilon, T)$, for any $0 < \varepsilon < \delta$.*

At this point we would like to make a remark on the proof of the above proposition. One might be inclined to believe that the measure μ constructed in the proof of theorem 8.1.2 is a measure of maximal entropy for *any* continuous transformation T. The reader should note, though, that μ still depends on ε and it is therefore only because of the above corollary that μ is a measure of maximal entropy for expansive homeomorphisms.

Bibliography

[B] Michael Brin and Garrett Stuck, *Introduction to Dynamical Systems*, Cambridge University Press, 2002.

[D] K. Dajani and C. Kraaikamp, *Ergodic theory of numbers*, Carus Mathematical Monographs, 29. Mathematical Association of America, Washington, DC 2002.

[DF] Dajani, K., Fieldsteel, A. *Equipartition of Interval Partitions and an Application to Number Theory*, Proc. AMS Vol 129, **12**, (2001), 3453 - 3460.

[De] Devaney, R.L., 1989, An Introduction to Chaotic Dynamical Systems, Second Edition, Addison-Wesley Publishing Company, Inc.

[G] Gleick, J., 1998, Chaos: Making a New Science, Vintage

[Ho] Hoppensteadt, F.C., 1993, Analysis and Simulation of Chaotic Systems, Springer-Verlag New York, Inc.

[H] W. Hurewicz, *Ergodic theorem withour invariant measure*. Annals of Math., **45** (1944), 192–206.

[KK] Kamae, Teturo; Keane, Michael, *A simple proof of the ratio ergodic theorem*, Osaka J. Math. 34 (1997), no. 3, 653–657.

[KT] Kingman and Taylor, *Introduction to measure and probability*, Cambridge Press, 1966.

[M] Misiurewicz, M., 1976, A short proof of the variational principle for a $\mathbb{Z}_+^n$ action on a compact space, Asterique, 40, pp. 147-187

[Mu] Munkres, J.R., 2000, Topology, Second Edition, Prentice Hall, Inc.

[Pa] William Parry, *Topics in Ergodic Theory*, Reprint of the 1981 original. Cambridge Tracts in Mathematics, 75. Cambridge University Press, Cambridge, 2004.

[P] Karl Petersen, *Ergodic Theory*, Cambridge Studies in Advanced Mathematics, 2. Cambridge University Press, Cambridge, 1989.

[W] Peter Walters, *An Introduction to Ergodic Theory*, Graduate Texts in Mathematics, 79. Springer-Verlag, New York-Berlin, 1982.

Index

Notes on Ergodic Theory.

H. Bruin

November 5, 2014

Abstract

These are notes in the making for the course VO 250059: Ergodic Theory 1, Spring Semester 2013-2014, University of Vienna

1 Notation

Throughout, (X,d) will be a metric space, possibly compact, and $T : X \to X$ will be a (piecewise) continuous map. The combination (X,T) defines dynamical systems by means of iteration. The **orbit** of a point $x \in X$ is the set

$$\operatorname{orb}(x) = \{x, T(x), T \circ T(x), \dots, \underbrace{T \circ \cdots \circ T}_{n \text{ times}}(x) =: T^n(x), \cdots\} = \{T^n(x) : n \geqslant 0\},$$

and if T is invertible, then $\operatorname{orb}(x) = \{T^n(x) : n \in \mathbb{Z}\}$ where the negative iterates are defined as $T^{-n} = (T^{inv})^n$. In other words, we consider $n \in \mathbb{N}$ (or $n \in \mathbb{Z}$) as discrete time, and $T^n(x)$ is the position the point x takes at time n.

Definition 1. *We call x a* **fixed point** *if $T(x) = x$;* **periodic** *if there is $n \geqslant 1$ such that $T^n(x) = x$;* **recurrent** *if $x \in \overline{\operatorname{orb}(x)}$.*

In general *chaotic* dynamical systems most orbits are more complicated than periodic (or quasi-periodic as the irrational rotation R_α discussed below). The behaviour of such orbits is hard to predict. Ergodic Theory is meant to help in predicting the behaviour of **typical** orbits, where typical means: almost all points x for some (invariant) measure μ.

To define measures properly, we need a σ-algebra $\mathcal{B}$ of “measurable” subsets. σ-algebra means that the collection $\mathcal{B}$ is closed under taking complements, countable unions and countable intersections, and also that $\varnothing, X \in \mathcal{B}$. Then a measure μ is a function

$\mu : \mathcal{B} \to \mathbb{R}^+$ that is countably subadditive: $\mu(\cup_i A_i) \leqslant \sum_i \mu(A)_i$ (with equality if the sets A_i are pairwise disjoint).

Example: For a subset $A \subset X$, define

$$\nu(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_A \circ T^i x,$$

for the **indicator function** 1_A, assuming for the moment that this limit exists. We call this the visit frequency of x to the set A. We can compute

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_A \circ T^i x &= \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=0}^{n-1} 1_A \circ T^{i+1} x + 1_A x - 1_A(T^n x) \right) \\
&= \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=0}^{n-1} 1_{T^{-1}A} \circ T^i x + 1_A x - 1_A(T^n x) \right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{T^{-1}A} \circ T^i x = \nu(T^{-1}(A))
\end{aligned}$$

That is, visit frequency measures, when well-defined, are **invariant** under the map. This allows us to use invariant measure to make statistical predictions of what orbit do "on average".

Let $\mathcal{B}_0$ be the collection of subsets $A \in \mathcal{B}$ such that $\mu(A) = 0$, that is: $\mathcal{B}_0$ are the **null-sets** of μ. We say that an event happens almost surely (a.s.) or μ-almost everywhere (μ-a.e.) if it is true for all $x \in X \setminus A$ for some $A \in \mathcal{B}_0$.

A measure μ on $(X, T, \mathcal{B})$ is called

- non-singular if $A \in \mathcal{B}_0$ implies $T^{-1}(A) \in \mathcal{B}_0$.
- non-atomic if $\mu(\{x\}) = 0$ for every $x \in X$
- T-invariant if $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$.
- finite if $\mu(X) < \infty$. In this case we can always rescale μ so that $\mu(X) = 1$, *i.e.,* μ is a probability measure.
- σ-finite if there is a countable collection X_i such that $X = \cup_i X_i$ and $\mu(X_i) \leqslant 1$ for all i. In principle, finite measures are also σ-finite, but we would like to reserve the term σ-finite only for infinite measures (*i.e.,* $\mu(X) = \infty$).

Example: Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$T \begin{pmatrix} x \\ y \end{pmatrix} = M \begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{for matrix} \qquad M = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

2 What are its invariant measures?

Note that T is a bijection of $\mathbb{R}^2$, with 0 as single fixed point. Therefore the Dirac measure δ_0 is T-invariant. However, also Lebesgue measure m is invariant because (using coordinate transformation $x = T^{-1}(y)$)

$$m(T^{-1}A) = \int_{T^{-1}A} dm(x) = \int_A \det(M^{-1}) dm(y) = \int_A \frac{1}{\det(M)} dm(y) = m(A)$$

because $\det(M) = 1$. This is a general fact: If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a bijection with Jacobian $J = |\det(DT)| = 1$, then Lebesgue measure is preserved. However, Lebesgue measure is not a probability measure (it is σ-finite). In the above case of the integer matrix with determinant 1, T preserves (and is a bijection) on $\mathbb{Z}^2$. Therefore we can factor out over $\mathbb{Z}^2$ and obtain a map on the two-torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$:

$$\begin{aligned} T \; : \; & \mathbb{T}^2 \to \mathbb{T}^2 \\ & \begin{pmatrix} x \\ y \end{pmatrix} \mapsto M \begin{pmatrix} x \\ y \end{pmatrix} \pmod 1 \end{aligned}$$

This map is called Arnol'd's cat-map, and it preserves Lebesgue measure, which on $\mathbb{T}^2$ is a probability measure.

A special case of the above is:

Proposition 1. *If $T : U \subset \mathbb{R}^n \to U$ is an isometry (or piecewise isometric bijection), then T preserves Lebesgue measure.*

Let $\mathcal{M}(X, T)$ denote the set of T-invariant Borel[1] probability measures. In general, there are always invariant measures.

Theorem 1 (Krylov-Bogol'ubov)**.** *If $T : X \to X$ is a continuous map on a nonempty compact metric space X, then $\mathcal{M}(T) \neq \emptyset$.*

Proof. Let ν be any probability measure and define Cesaro means:

$$\nu_n(A) = \frac{1}{n} \sum_{j=0}^{n-1} \nu(T^{-j}A),$$

these are all probability measures. The collection of probability measures on a compact metric space is known to be compact in the weak* topology, *i.e.*, there is limit probability measure μ and a subsequence $(n_i)_{i \in \mathbb{N}}$ such that for every continuous function $\psi : X \to \mathbb{R}$:

$$\int_X \psi \, d\nu_{n_i} \to \int \psi \, d\mu \text{ as } i \to \infty. \tag{1}$$

[1] that is, sets in the σ-algebra of sets generated by the open subsets of X.

On a metric space, we can, for any $\varepsilon > 0$ and closed set A, find a continuous function $\psi_A : X \to [0,1]$ such that $\psi_A(x) = 1$ if $x \in A$ and $\mu(A) \leqslant \int_X \psi_A d\mu \leqslant \mu(A) + \varepsilon$ and similarly $\mu(T^{-1}A) \leqslant \int_X \psi_A \circ T \, d\mu \leqslant \mu(T^{-1}A) + \varepsilon$. Now

$$\begin{aligned}
|\mu(T^{-1}(A)) - \mu(A)| &\leqslant \left| \int \psi_A \circ T \, d\mu - \int \psi_A \, d\mu \right| + 2\varepsilon \\
&= \lim_{i\to\infty} \left| \int \psi_A \circ T \, d\nu_{n_i} - \int \psi_A \, d\nu_{n_i} \right| + 2\varepsilon \\
&= \lim_{i\to\infty} \frac{1}{n_i} \left| \sum_{j=0}^{n_i-1} \left(\int \psi_A \circ T^{-(j+1)} \, d\nu - \int \psi_A \circ T^{-j} \, d\nu \right) \right| + 2\varepsilon \\
&\leqslant \lim_{i\to\infty} \frac{1}{n_i} \left| \int \psi_A \circ T^{-n_i} \, d\nu - \int \psi_A \, d\nu \right| + 2\varepsilon \\
&\leqslant \lim_{i\to\infty} \frac{1}{n_i} 2\|\psi_A\|_\infty + 2\varepsilon = 2\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, and because the closed sets form a generator of the Borel sets, we find that $\mu(T^{-1}(A)) = \mu(A)$ as required. □

3 Ergodicity and unique ergodicity

Definition 2. *A measure is called* **ergodic** *if $T^{-1}(A) = A \pmod{\mu}$ for some $A \in \mathcal{B}$ implies that $\mu(A) = 0$ or $\mu(A^c) = 0$.*

Proposition 2. *The following are equivalent:*

(i) μ is ergodic;

(ii) If $\psi \in L^1(\mu)$ is T-invariant, i.e., *$\psi \circ T = \psi$ μ-a.e., then ψ is constant μ-a.e.*

Proof. (i) $\Rightarrow$ (ii): Let $\psi : X \to \mathbb{R}$ be T-invariant μ-a.e., but not constant. Thus there exists $a \in \mathbb{R}$ such that $A := \psi^{-1}((-\infty, a])$ and $A^c = \psi^{-1}((a, \infty))$ both have positive measure. By T-invariance, $T^{-1}A = A \pmod{\mu}$, and we have a contradiction to ergodicity.
(ii) $\Rightarrow$ (i): Let A be a set of positive measure such that $T^{-1}A = A$. Let $\psi = 1_A$ be its indicator function; it is T-invariant because A is T-invariant. By (ii), ψ is constant μ-a.e., but as $\psi(x) = 0$ for $x \in A^c$, it follows that $\mu(A^c) = 0$. □

The rotation $R_\alpha : \mathbb{S}^1 \to \mathbb{S}^1$ is defined as $R_\alpha(x) = x + \alpha \pmod 1$.

Theorem 2 (Poincaré)**.** *If $\alpha \in \mathbb{Q}$, then every orbit is periodic.*

If $\alpha \notin \mathbb{Q}$, then every orbit is dense in $\mathbb{S}^1$. In fact, for every interval J and every $x \in \mathbb{S}^1$, the visit frequency

$$v(J) := \lim_{n\to\infty} \frac{1}{n}\#\{0 \leqslant i < n : R_\alpha^i(x) \in J\} = |J|.$$

Proof. If $\alpha = \frac{p}{q}$, then clearly

$$R_\alpha^q(x) = x + q\alpha \pmod 1 = x + q\frac{p}{q} \pmod 1 = x + p \pmod 1 = x.$$

Conversely, if $R_\alpha^q(x) = x$, then $x = x + q\alpha \pmod 1$, so $q\alpha = p$ for some integer p, and $\alpha = \frac{p}{q} \in \mathbb{Q}$.

Therefore, if $\alpha \notin \mathbb{Q}$, then x cannot be periodic, so its orbit is infinite. Let $\varepsilon > 0$. Since $\mathbb{S}^1$ is compact, there must be $m < n$ such that $0 < \delta := d(R_\alpha^m(x), R_\alpha^n(x)) < \varepsilon$. Since R_α is an isometry, $|R_\alpha^{k(n-m)}(x) - R_\alpha^{(k+1)(n-m)}(x)| = \delta$ for every $k \in \mathbb{Z}$, and $\{R_\alpha^{k(n-m)}(x) : k \in \mathbb{Z}\}$ is a collection of points such that every two neighbours are exactly δ apart. Since $\varepsilon > \delta$ is arbitrary, this shows that $\mathrm{orb}(x)$ is dense, but we want to prove more.

Let $J_\delta^0 = [R_\alpha^m(x), R_\alpha^n(x))$ and $J_\delta^k = R_\alpha^k(J_\delta)$. Then for $K = \lfloor 1/\delta \rfloor$, $\{J_\delta^k\}_{k=0}^K$ is a cover $\mathbb{S}^1$ of adjacent intervals, each of length δ, and $R_\alpha^{j(n-m)}$ is an isometry from J_δ^i to J_δ^{i+j}. Therefore the visit frequencies

$$\underline{v}_k = \liminf_n \frac{1}{n}\#\{0 \leqslant i < n : R_\alpha^i(x) \in J_\delta^k\}$$

are all the same for $0 \leqslant k \leqslant K$, and together they add up to at most $1 + \frac{1}{K}$. This shows for example that

$$\frac{1}{K+1} \leqslant \underline{v}_k \leqslant \overline{v}_k := \limsup_n \frac{1}{n}\#\{0 \leqslant i < n : R_\alpha^i(x) \in J_\delta^k\} \leqslant \frac{1}{K},$$

and these inequalities are **independent** of the point x. Now an arbitrary interval J can be covered by $\lfloor |J|/\delta \rfloor + 2$ such adjacent J_δ^k, so

$$v(J) \leqslant \left(\frac{|J|}{\delta} + 2\right)\frac{1}{K} \leqslant (|J|(K+1) + 2)\frac{1}{K} \leqslant |J| + \frac{3}{K}.$$

A similar computation gives $v(J) \geqslant |J| - \frac{3}{K}$. Now taking $\varepsilon \to 0$ (hence $\delta \to 0$ and $K \to \infty$), we find that the limit $v(J)$ indeed exists, and is equal to $|J|$. $\square$

Definition 3. *A transformation (X, T) is called* **uniquely ergodic** *if there is exactly one invariant probability measure.*

The above proof shows that Lebesgue measure is the only invariant measure if $\alpha \notin \mathbb{Q}$, so $(\mathbb{S}^1, R_\alpha)$ is uniquely ergodic. However, there is a missing step in the logic, in that we didn't show yet that Lebesgue measure is ergodic. This will be shown in Example 1 and also Theorem 6.

Questions: Does R_α preserve a σ-finite measure? Does R_α preserve a **non-atomic** σ-finite measure?

Lemma 1. *Let X be a compact space. A transformation $(X, \mathcal{B}, \mu, T)$ is uniquely ergodic if and only if, for every continuous function, the Birkhoff averages $\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x)$ converge* **uniformly** *to a constant function.*

Remark 1. *Every continuous map on a compact space has an invariant measure, as we showed in Theorem 1. Theorem 6 later on shows that if there is only one invariant measure, it has to be ergodic as well.*

Proof. If μ and ν were two different ergodic measures, then we can find a continuous function $f : X \to \mathbb{R}$ such that $\int f d\mu \neq \int f d\nu$. Using the Ergodic Theorem for both measures (with their own typical points x and y), we see that

$$\lim_n \frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k(x) = \int f d\mu \neq \int f d\nu = \lim_n \frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k(y),$$

so there is not even convergence to a constant function.

Conversely, we know by the Ergodic Theorem that $\lim_n \frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k(x) = \int f d\mu$ is constant μ-a.e. But if the convergence is not uniform, then there are sequences $(x_i), (y_i) \subset X$ and $(m_i), (n_i) \subset \mathbb{N}$, such that $\lim_i \frac{1}{m_i}\sum_{k=0}^{m_i-1} f \circ T^k(x) := A \neq B =: \lim_i \frac{1}{n_i}\sum_{k=0}^{n_i-1} f \circ T^k(y_i)$. Take functionals $\mu_i(g) = \liminf_i \frac{1}{m_i}\sum_{k=0}^{m_i-1} g \circ T^k(x)$ and $\nu_i(g) = \liminf_i \frac{1}{n_i}\sum_{k=0}^{n_i-1} g \circ T^k(x)$. Both sequences have weak accumulation points μ and ν which are easily shown to be T-invariant measures, see the proof of Theorem 1. But they are not the same because $\mu(f) = A \neq B = \nu(f)$. $\square$

4 The Ergodic Theorem

Theorem 2 is an instance of a very general fact observed in ergodic theory:

Space Average = Time Average (for typical points).

This is expressed in the

Theorem 3 (Birkhoff Ergodic Theorem)**.** *Let μ be a probability measure and $\psi \in L^1(\mu)$. Then the* **ergodic average**

$$\overline{\psi}(x) := \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \psi \circ T^i(x)$$

exists μ-a.e. (everywhere if ψ is continuous), and $\overline{\psi}$ is T-invariant, i.e., $\overline{\psi} \circ T = \overline{\psi}$ *μ-a.e. If in addition μ is ergodic then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \psi \circ T^i(x) = \int_X \psi \, d\mu \qquad \mu\text{-a.e.} \tag{2}$$

Remark 2. *A point $x \in X$ satisfying* (2) *is called* **typical** *for μ. To be precise, the set of μ-typical points also depends on ψ, but for different functions $\psi, \tilde{\psi}$, the (μ, ψ)-typical points and $(\mu, \tilde{\psi})$-typical points differ only on a null-set.*

Corollary 1. *Lebesgue measure is the only R_α-invariant probability measure.*

Proof. Suppose R_α had two invariant measures, μ and ν. Then there must be an interval J such that $\mu(J) \neq \nu(J)$. Let $\psi = 1_J$ be the indicator function; it will belongs to $L^1(\mu)$ and $L^1(\nu)$. Apply Birkhoff's Ergodic Theorem for some μ-typical point x and ν-typical point y. Since their visit frequencies to J are the same, we have

$$\begin{aligned} \mu(J) &= \int \psi \, d\mu = \lim_n \frac{1}{n} \#\{0 \leqslant i < n : R_\alpha(x) \in J\} \\ &= \lim_n \frac{1}{n} \#\{0 \leqslant i < n : R_\alpha(y) \in J\} = \int \psi \, d\nu = \nu(J), \end{aligned}$$

a contradiction to μ and ν being different. □

5 Absolute continuity and invariant densities

Definition 4. *A measure μ is called* **absolutely continuous** *w.r.t. the measure ν (notation: $\mu \ll \nu$ if $\nu(A) = 0$ implies $\mu(A) = 0$. If both $\mu \ll \nu$ and $\nu \ll \mu$, then μ and ν are called equivalent.*

Proposition 3. *If $\mu \ll \nu$ are both T-invariant probability measures, with a common σ-algebra $\mathcal{B}$ of measurable sets. If ν is ergodic, then $\mu = \nu$.*

Proof. First we show that μ is ergodic. Indeed, otherwise there is a T-invariant set A such that $\mu(A) > 0$ and $\mu(A^c) > 0$. By ergodicity of ν at least one of A or A^c must have ν-measure 0, but this would contradict that $\mu \ll \nu$.

Now let $A \in \mathcal{B}$ and let $Y \subset X$ be the set of ν-typical points. Then $\nu(Y^c) = 0$ and hence $\mu(Y^c) = 0$. Applying Birkhoff's Ergodic Theorem to μ and ν separately for $\psi = 1_A$ and some μ-typical $y \in Y$, we get

$$\mu(A) = \lim_n \frac{1}{n} \sum_{i=0}^{n-1} \psi \circ T(y) = \nu(A).$$

But $A \in \mathcal{B}$ was arbitrary, so $\mu = \nu$. □

Theorem 4 (Radon-Nikodym)**.** *If μ is a probability measure and $\mu \ll \nu$ then there is a function $h \in L^1(\nu)$ (called* **Radon-Nikodym derivative** *or* **density***) such that $\mu(A) = \int_A h(x)\, d\nu(x)$ for every measurable set A.*

Sometimes we use the notation: $h = \frac{d\mu}{d\nu}$.

Proposition 4. *Let $T : U \subset \mathbb{R}^n \to U$ be (piecewise) differentiable, and μ is absolutely continuous w.r.t. Lebesgue. Then μ is T-invariant if and only if its density $h = \frac{d\mu}{dx}$ satisfies*

$$h(x) = \sum_{T(y)=x} \frac{h(y)}{|\det DT(y)|} \tag{3}$$

for every x.

Proof. The T-invariance means that $d\mu(x) = d\mu(T^{-1}(x))$, but we need to beware that T^{-1} is multivalued. So it is more careful to split the space U into pieces U_n such that the restrictions $T_n := T|U_n$ are diffeomorphism (onto their images) and write $y_n = T_n^{-1}(x) = T^{-1}(x) \cap U_n$. Then we obtain (using the change of coordinates)

$$\begin{aligned} h(x)\ dx &= d\mu(x) = d\mu(T^{-1}(x)) = \sum_n d\mu \circ T_n^{-1}(x) \\ &= \sum_n h(y_n)|\det(DT_n^{-1})(x)|dy_n = \sum_n \frac{h(y_n)}{\det |DT(y_n)|} dy_n, \end{aligned}$$

and the statement follows.

Conversely, if (3) holds, then the above computation gives $d\mu(x) = d\mu \circ T^{-1}(x)$, which is the required invariance. $\square$

Example: If $T : [0,1] \to [0,1]$ is (countably) piecewise linear, and each branch $T : I_n \to [0,1]$ (on which T is affine) is onto, then T preserves Lebesgue measure. Indeed, the intervals I_n have pairwise disjoint interiors, and their lengths add up to 1. If s_n is the slope of $T : I_n \to [0,1]$, then $s_n = 1/|I_n|$. Therefore $\sum_n \frac{1}{DT(y_n)} = \sum_n 1/s_n = \sum_n |I_n| = 1$.

Example: The map $T : \mathbb{R} \setminus \{0\} \to \mathbb{R},\ T(x) = x - \frac{1}{x}$ is called the **Boole transformation**. It is 2-to-1; the two preimages of $x \in \mathbb{R}$ are $y_\pm = \frac{1}{2}(x \pm \sqrt{x^2+4})$. Clearly $T'(x) = 1 + \frac{1}{x^2}$. A tedious computation shows that

$$\frac{1}{|T'(y_-)|} + \frac{1}{|T'(y_+)|} = 1.$$

Indeed, $|T'(y_\pm)| = 1 + \frac{2}{x^2+2\pm x\sqrt{x^2+4}}$, $1/|T'(y_\pm)| = \frac{x^2+2\pm x\sqrt{x^2+4}}{x^2+4\pm x\sqrt{x^2+4}}$, and

$$\begin{aligned}
\frac{1}{|T'(y_-)|} + \frac{1}{|T'(y_+)|} &= \frac{x^2+2-x\sqrt{x^2+4}}{x^2+4-x\sqrt{x^2+4}} + \frac{x^2+2+x\sqrt{x^2+4}}{x^2+4+x\sqrt{x^2+4}} \\
&= \frac{(x^2+2-x\sqrt{x^2+4})(x^2+4+x\sqrt{x^2+4})}{(x^2+4)^2 - x^2(x^2+4)} \\
&\qquad + \frac{(x^2+2+x\sqrt{x^2+4})(x^2+4-x\sqrt{x^2+4})}{(x^2+4)^2 - x^2(x^2+4)} \\
&= \frac{(x^2+2)^2 - x^2(x^2+4) + 2(x^2+2) - 2x\sqrt{x^2+4}}{4(x^2+4)} + \\
&\qquad \frac{(x^2+2)^2 - x^2(x^2+4) + 2(x^2+2) + 2x\sqrt{x^2+4}}{4(x^2+4)} \\
&= \frac{4(x^2+2)+8}{4(x^2+4)} = 1.
\end{aligned}$$

Therefore $h(x) \equiv 1$ is an invariant density, so Lebesgue measure is preserved.

Example: The **Gauß map** $G : (0,1] \to [0,1)$ is defined as $G(x) = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor$. It has an invariant density $h(x) = \frac{1}{\log 2}\frac{1}{1+x}$. Here $\frac{1}{\log 2}$ is just the normalising factor (so that $\int_0^1 h(x)dx = 1$).

Let $I_n = (\frac{1}{n+1}, \frac{1}{n}]$ for $n = 1,2,3,\dots$ be the domains of the branches of G, and for $x \in (0,1)$, and $y_n := G^{-1}(x) \cap I_n = \frac{1}{x+n}$. Also $G'(y_n) = -\frac{1}{y_n^2}$. Therefore

$$\begin{aligned}
\sum_{n\geqslant 1} \frac{h(y_n)}{|G'(y_n)|} &= \frac{1}{\log 2}\sum_{n\geqslant 1} \frac{y_n^2}{1+y_n} = \frac{1}{\log 2}\sum_{n\geqslant 1} \frac{\frac{1}{(x+n)^2}}{1+\frac{1}{x+n}} \\
&= \frac{1}{\log 2}\sum_{n\geqslant 1} \frac{1}{x+n}\frac{1}{x+n+1} \\
&= \frac{1}{\log 2}\sum_{n\geqslant 1} \frac{1}{x+n} - \frac{1}{x+n+1} \qquad \text{telescoping series} \\
&= \frac{1}{\log 2}\frac{1}{x+1} = h(x).
\end{aligned}$$

Exercise 1. *Show that for each integer $n \geqslant 2$, the interval map given by*

$$T_n(x) = \begin{cases} nx & \text{if } 0 \leqslant x \leqslant \frac{1}{n}, \\ \frac{1}{x} - \lfloor \frac{1}{x} \rfloor & \text{if } \frac{1}{n} < x \leqslant 1, \end{cases}$$

has invariant density $\frac{1}{\log 2}\frac{1}{1+x}$.

Theorem 5. *If $T : \mathbb{S}^1 \to \mathbb{S}^1$ is a C^2 expanding circle map, then it preserves a measure μ equivalent to Lebesgue, and μ is ergodic.*

Expanding here means that there is $\lambda > 1$ such that $|T'(x)| \geqslant \lambda$ for all $x \in \mathbb{S}^1$. The above theorem can be proved in more generality, but in the stated version it conveys the ideas more clearly.

Proof. First some estimates on derivatives. Using the Mean Value Theorem twice, we obtain

$$\begin{aligned}
\log \frac{|T'(x)|}{|T'(y)|} &= \log(1 + \frac{|T'(x)| - |T'(y)|}{|T'(y)|}) \leqslant \frac{|T'(x)| - |T'(y)|}{|T'(y)|} \\
&= \frac{|T''(\xi)| \cdot |x-y|}{|T'(y)|} = \frac{|T''(\xi)|}{|T'(y)|} \frac{|Tx - Ty|}{\lambda} \\
&\leqslant \sup_\zeta \frac{|T'(\zeta)|}{|T'(y)|} \frac{|Tx - Ty|}{\lambda} \leqslant K|Tx - Ty|
\end{aligned}$$

for some constant K. The chain rule then gives:

$$\log \frac{|DT^n(x)|}{|DT^n(y)|} = \sum_{i=0}^{n-1} \log \frac{|T'(T^i x)|}{|T'(T^i y)|} \leqslant K \sum_{i=1}^{n} |T^i(x) - T^i(y)|.$$

Since T is a continuous expanding map of the circle, it wraps the circle d times around itself, and for each n, there are d^n pairwise disjoint intervals Z such that $T^i Z \to \mathbb{S}^1$ is onto, with slope at least λ^i. If we take x, y above in one such Z, then $|x - y| < \lambda^{-n}|T^n(x) - T^n(y)|$ and in fact $|T^i(x) - T^i(y)| < \lambda^{-(n-i)}|T^n(x) - T^n(y)|$. Therefore we obtain

$$\log \frac{|DT^n(x)|}{|DT^n(y)|} = K \sum_{i=1}^{n} \lambda^{-(n-i)} |T^n(x) - T^n(y)| \leqslant \frac{K}{\lambda - 1} |T^n(x) - T^n(y)| \leqslant \log K'$$

for some $K' \in (1, \infty)$. This means that if $A \subset Z$ (so $T^n : A \to T^n(A)$ is a bijection), then

$$\frac{1}{K'} \frac{m(A)}{m(Z)} \leqslant \frac{m(T^n A)}{m(T^n Z)} = \frac{m(T^n A)}{m(\mathbb{S}^1)} \leqslant K' \frac{m(A)}{m(Z)}, \tag{4}$$

where m is Lebesgue measure.

Now we construct the T-invariant measure μ. Take $B \subset \mathcal{B}$ arbitrary, and set $\mu_n(B) = \frac{1}{n} \sum_{i=0}^{n-1} m(T^{-i} B)$. Then by (4),

$$\frac{1}{K'} m(B) \leqslant \mu_n(B) \leqslant K' m(B).$$

We can take a weak* limit of the μ_n's, call it μ, then

$$\frac{1}{K'} m(B) \leqslant \mu(B) \leqslant K' m(B),$$

and therefore μ and m are equivalent. The T-invariance of μ proven in the same way as Theorem 1.

Now for the ergodicity of μ, we need the Lebesgue Density Theorem, which says that if $m(A) > 0$, then for m-a.e. $x \in A$, the limit

$$\lim_{\varepsilon \to 0} \frac{m(A \cap B_\varepsilon(x))}{m(B_\varepsilon(x))} = 1,$$

where $B_\varepsilon(x)$ is the ε-balls around x. Points x with this property are called **(Lebesgue) density points** of A. (In fact, the above also holds, if $B_\varepsilon(x)$ is just a one-sided ε-neighbourhood of x.)

Assume by contradiction that μ is not ergodic. Take $A \in \mathcal{B}$ a T-invariant set such that $\mu(A) > 0$ and $\mu(A^c) > 0$. By equivalence of μ and m, also $\delta := m(A^c) > 0$. Let x be a density point of A, and Z_n be a neighbourhood of x such that $T^n : Z \to \mathbb{S}^1$ is a bijection. As $n \to \infty$, $Z \to \{x\}$, and therefore we can choose n so large (hence Z so small) that

$$\frac{m(A \cap Z)}{m(Z)} > 1 - \delta/K'.$$

Therefore $\frac{m(A^c \cap Z)}{m(Z)} < \delta/K'$, and using (4),

$$\frac{m(T^n(A_c \cap Z))}{m(T^n(Z))} \leqslant K' \frac{m(A^c \cap Z)}{m(Z)} < K'\delta/K' = \delta.$$

Since $T^n : A^c \cap Z \to A^c$ is a bijection, and $m(T^n Z) = m(\mathbb{S}^1) = 1$, we get $\delta = m(A^c) < \delta$, a contraction. Therefore μ is ergodic. □

6 The Choquet Simplex and the Ergodic Decomposition

Throughout this section, let $T : X \to X$ a **continuous** transformation of a compact metric space. Recall that $\mathcal{M}(X)$ is the collection of probability measures defined on X; we saw in (1) that it is compact in the weak* topology. In general, X carries many T-invariant measures. The set $\mathcal{M}(X,T) = \{\mu \in \mathcal{M}(X) : \mu \text{ is } T\text{-invariant}\}$ is called the **Choquet simplex** of T. Let $\mathcal{M}_{erg}(X,T)$ be the subset of $\mathcal{M}(X,T)$ of ergodic T-invariant measures.

Clearly $\mathcal{M}(X,T) = \{\mu\}$ if (X,T) is uniquely ergodic. The name "simplex" just reflects the convexity of $\mathcal{M}(X,T)$: if $\mu_1, \mu_2 \in \mathcal{M}(X,T)$, then also $\alpha\mu_1 + (1-\alpha)\mu_2 \in \mathcal{M}(X,T)$ for every $\alpha \in [0,1]$.

Lemma 2. *The Choquet simplex $\mathcal{M}(X,T)$ is a compact subset of $\mathcal{M}(X)$ w.r.t. weak* topology.*

Proof. Suppose $\{\mu_n\} \subset \mathcal{M}(X,T)$, then by the compactness of $\mathcal{M}(X)$, see (1), there is $\mu \in \mathcal{M}(X)$ and a subsequence $(n_i)_i$ such that for every continuous function $f : X \to \mathbb{R}$ such that $\int f d\mu_{n_i} \to \int f d\mu$ as $i \to \infty$. It remains to show that μ is T-invariant, but this simply follows from continuity of $f \circ T$ and

$$\int f \circ T \, d\mu = \lim_i \int f \circ T \, d\mu_{n_i} = \lim_i \int f \, d\mu_{n_i} = \int f \, d\mu.$$

□

Theorem 6. *The ergodic measures are exactly the extremal points of the Choquet simplex.*

Proof. First assume that μ is not ergodic. Hence there is a T-invariant set A such that $0 < \mu(A) < 1$. Define

$$\mu_1(B) = \frac{\mu(B \cap A)}{\mu(A)} \qquad \text{and} \qquad \mu_2(B) = \frac{\mu(B \setminus A)}{\mu(X \setminus A)}.$$

Then $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ for $\alpha = \mu(A) \in (0,1)$ so μ is not an extremal point.

Suppose now that μ is ergodic but that $\mu = \alpha\mu_1 + (1-\alpha)\mu_2$ for some $\alpha \in (0,1)$. We need to show that $\mu_1 = \mu_2 = \mu$. From the definition, it is clear that $\mu_1 \ll \mu$, so a Radon-Nikodym derivative $\frac{d\mu_1}{d\mu}$ exists in $L^1(\mu)$. Let $A^- = \{x \in X : \frac{d\mu_1}{d\mu} < 1\}$. Then

$$\begin{aligned}\int_{A^- \cap T^{-1}A^-} \frac{d\mu_1}{d\mu} \, d\mu + \int_{A^- \setminus T^{-1}A^-} \frac{d\mu_1}{d\mu} \, d\mu = \mu_1(A^-)& \\ = \mu_1(T^{-1}A^-) = \int_{T^{-1}A^- \cap A^-} \frac{d\mu_1}{d\mu} \, d\mu + \int_{T^{-1}A^- \setminus A^-} \frac{d\mu_1}{d\mu} \, d\mu.&\end{aligned}$$

Canceling the term $\int_{A^- \cap T^{-1}A^-} \frac{d\mu_1}{d\mu} \, d\mu$ gives

$$\int_{A^- \setminus T^{-1}A^-} \frac{d\mu_1}{d\mu} \, d\mu = \int_{T^{-1}A^- \setminus A^-} \frac{d\mu_1}{d\mu} \, d\mu. \tag{5}$$

But also $\mu(T^{-1}A^- \setminus A^-) = \mu(T^{-1}A^-) - \mu(T^{-1}A^- \cap A^-) = \mu(A^- \setminus T^{-1}A^-)$. Therefore, in (5), both integrations are over sets of the same measure, but in the left-hand side, the integrand < 1 while in the right-hand side, the integrand $\geqslant 1$. Therefore $\mu(T^{-1}A^- \setminus A^-) = \mu(A^- \setminus T^{-1}A^-) = 0$, and hence A^- is T-invariant. By assumed ergodicity of μ, $\mu(A^-) = 0$ or 1. In the latter case,

$$1 = \mu_1(X) = \int_X \frac{d\mu_1}{d\mu} \, d\mu = \int_{A^-} \frac{d\mu_1}{d\mu} \, d\mu < \mu(A^-) = 1,$$

a contradiction. Therefore $\mu(A^-) = 0$. But then we can repeat the argument for $A^+ = \{x \in X : \frac{d\mu_1}{d\mu} > 1\}$ and find that $\mu(A^+) = 0$ as well. Therefore $\frac{d\mu_1}{d\mu} = 1$ μ-a.e. and hence $\mu_1 = \mu$. But then also $\mu_2 = \mu$, which finishes the proof. □

The following fundamental theorem implies that for checking the properties of any measure $\mu \in \mathcal{M}(X,T)$, it suffices to verify the properties for ergodic measures:

Theorem 7 (Ergodic Decomposition). *For every $\mu \in \mathcal{M}(X,T)$, there is a measure ν on the spaces of ergodic measures such that $\nu(\mathcal{M}_{erg}(X,T)) = 1$ and*

$$\mu(B) = \int_{\mathcal{M}_{erg}(X,T)} m(B)\ d\nu(m)$$

for all Borel sets B.

7 Poincaré Recurrence

Theorem 8 (Poincaré's Recurrence Theorem). *If (X,T,μ) is a measure preserving system with $\mu(X) = 1$, then for every measurable set $U \subset X$ of positive measure, μ-a.e. $x \in U$ returns to U,* i.e., *there is $n = n(x)$ such that $T^n(x) \in U$.*

Proof of Theorem 8. Let U be an arbitrary measurable set of positive measure. As μ is invariant, $\mu(T^{-i}(U)) = \mu(U) > 0$ for all $i \geqslant 0$. On the other hand, $1 = \mu(X) \geqslant \mu(\cup_i T^{-i}(U))$, so there must be overlap in the backward iterates of U, *i.e.,* there are $0 \leqslant i < j$ such that $\mu(T^{-i}(U) \cap T^{-j}(U)) > 0$. Take the j-th iterate and find $\mu(T^{j-i}(U) \cap U) \geqslant \mu(T^{-i}(U) \cap T^{-j}(U)) > 0$. This means that a positive measure part of the set U returns to itself after $n := j - i$ iterates.

For the part U' of U that didn't return after n steps, assuming U' has positive measure, we repeat the argument. That is, there is n' such that $\mu(T^{n'}(U') \cap U') > 0$ and then also $\mu(T^{n'}(U') \cap U) > 0$.

Repeating this argument, we can exhaust the set U up to a set of measure zero, and this proves the theorem. □

Definition 5. *A system $(X,T,\mathcal{B},\mu)$ is called* conservative *if for every set $A \in \mathcal{B}$ with $\mu(A) > 0$, there is $n \geqslant 1$ such that $\mu(T^n(A) \cap A) > 0$. The system is called* dissipative *otherwise, and it is called* totally dissipative *if $\mu(T^n(A) \cap A) = 0$ for very set $A \in \mathcal{B}$.*

We call the transformation T **recurrent** *w.r.t. μ if $B \setminus \cup_{i \in \mathbb{N}} T^{-i}(B)$ has zero measure for every $B \in \mathcal{B}$. In fact, this is equivalent to μ being conservative.*

The Poincaré Recurrence Theorem thus states that probability measure preserving systems are conservative.

Lemma 3 (Kac Lemma). *Let (X,T) preserve an ergodic measure μ. Take $Y \subset X$ measurable such that $0 < \mu(Y) \leqslant 1$, and let $\tau : Y \to \mathbb{N}$ be the first return time to Y.*

Then

$$\int \tau d\mu = \sum_{k \geqslant 1} k\mu(Y_k) = \begin{cases} 1 & \textit{if } \mu \textit{ is a probability measure ,} \\ \infty & \textit{if } \mu \textit{ is a conservative } \sigma\textit{-finite measure.} \end{cases}$$

for $Y_k := \{y \in Y : \tau(y) = k\}$.

Proof. Build a tower over Y by defining levels $L_0 = Y$, $L_1 = T(Y) \setminus Y$ and recursively $L_{j+1} = T(L_j) \setminus Y$. Then $L_j = \{T^j(y) : y \in Y, T^k(x) \notin Y \text{ for } 0 < k < j\}$. In particular, all the L_j are disjoint and $T(L_j) \subset L_{j+1} \cup Y$.

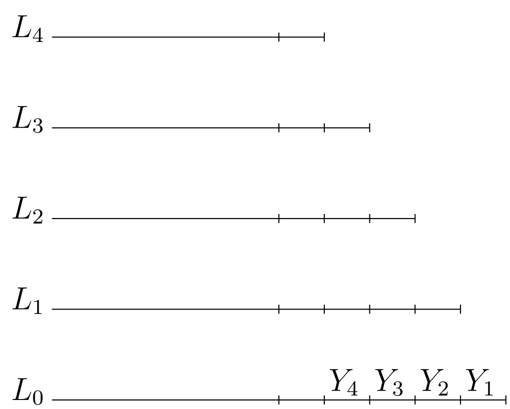

Figure 1: The tower consisting of levels L_j, $j \geqslant 0$.

We claim that $B := \cup_{j \geqslant 0} L_j$ is T-invariant (up to measure zero). Clearly $T^{-1}(L_j) \subset L_{j-1}$ for $j \geqslant 1$. Hence, we only need to show that $T^{-1}(Y) \subset B \pmod{\mu}$. Set $A := T^{-1}(Y) \setminus B$. We consider the two cases:

- $\mu(X) = 1$: if $x \in A$, then $T^{-j}(x) \notin Y$ for all $j \geqslant 0$, because if $j \geqslant 0$ were the minimal value such that $T^j(z) = x$ for some $z \in Y$, then $x \in L_j$.

  The sets $T^{-j}(A)$, $j \geqslant 0$, are in fact pairwise disjoint because if $x \in T^{-j}(A) \cap T^{-k}(A)$ for some minimal $0 \leqslant j < k$, then $T^{j-k}(A) \subset L_{2k-j-1}$, contradicting the previous paragraph.

  But this means that if $\mu(A) > 0$, then not only $\mu(T^{-j}(A)) = \mu(A) > 0$, but by disjointness, $\mu(\cup_j T^{-j}(A) = \sum_j \mu(T^{-j}(A)) = \infty$, contradicting that μ is a probability measure.

  This proves that $\mu(A) = 0$, so $T^{-1}(B) = B \pmod{\mu}$ and by ergodicity, $\mu(B) = 1$.

- $\mu(X) = \infty$ and μ is conservative: Note that $T^k(A) \cap A = \varnothing$ for all $k \geqslant 1$. Therefore, if $\mu(A) > 0$, we have a contradiction to conservativity.

The sets Y_k are clearly pairwise disjoint. Since $\tau(y) < \infty$ for μ-a.e. $y \in Y$, $\cup_k Y_k = Y$ (mod μ). Furthermore, $T^j(Y_k)$ are pairwise disjoint subsets of L_j for $j < k$ and $L_j = \cup_{k>j} T^j(Y_k)$ (mod μ). Finally, $T^{-1}(T^j(Y_k) \cap L_j) = T^{j-1}(Y_k) \cap L_{j-1}$ for $1 \leqslant j < k$. By T-invariance,

$$\mu(T^j(Y_k) \cap L_j) = \mu(T^{j-1}(Y_k) \cap L_{j-1}) = \cdots = \mu(T(Y_k) \cap L_1) = \mu(Y_k)$$

for $0 \leqslant j < k$.

Therefore (swapping the order of summation in the second line)

$$\begin{aligned}
\mu(X) = \mu(B) &= \sum_{j \geqslant 0} \mu(L_j) = \sum_{j \geqslant 0} \sum_{k>j} \mu(T^j(Y_k) \cap L_j) \\
&= \sum_{k \geqslant 1} \sum_{0 \leqslant j < k} \mu(T^j(Y_k) \cap L_j) \\
&= \sum_{k \geqslant 1} \sum_{0 \leqslant j < k} \mu(Y_k) = \sum_{k \geqslant 1} k\mu(Y_k),
\end{aligned}$$

as required. □

8 The Koopman operator

Given a probability measure preserving dynamical system $(X, \mathcal{B}, \mu, T)$, we can take the space of complex-valued square-integrable observables $L^2(\mu)$. This is a Hilbert space, equipped with inner product $\langle f, g \rangle = \int_X f(x) \cdot \overline{g(x)}\, d\mu$.

The Koopman operator $U_T : L^2(\mu) \to L^2(\mu)$ is defined as $U_T f = f \circ T$. By T-invariance of μ, it is a unitary operator. Indeed

$$\langle U_T f, U_T g \rangle = \int_X f \circ T(x) \cdot \overline{g \circ T(x)}\, d\mu = \int_X (f \cdot \overline{g}) \circ T(x)\, d\mu = \int_X f \cdot \overline{g}\, d\mu = \langle f, g \rangle,$$

and therefore $U_T^* U_T = U_T U_T^* = I$. This has several consequences, common to all unitary operators. First of all, the spectrum $\sigma(U_T)$ of U_T is a closed subset of the unit circle.

Secondly, we can give a (continuous) decomposition of U_T in orthogonal projections, called spectral decomposition. For a fixed eigenfunction ψ (with eigenvalue $\lambda \in \mathbb{S}^1$, we let $\Pi_\lambda : L^2(\mu) \to L^2(\mu)$ be the orthogonal projection onto the span of ψ. More generally, if $S \subset \sigma(U_T)$, we define Π_S as the orthogonal projection on the largest closed subspace V such that $U_T|_V$ has spectrum contained in S. As any orthogonal projection, we have the properties:

- $\Pi_S^2 = \Pi_S$ (Π_S is idempotent);

- $\Pi_S^* = \Pi_S$ (Π_S is self-adjoint);
- $\Pi_S \Pi_{S'} = 0$ if $S \cap S' = \varnothing$;
- The kernel $\mathcal{N}(\Pi_S)$ equals the orthogonal complement, $V^\perp$, of V.

Theorem 9 (Spectral Decomposition of Unitary Operators)**.** *There is a measure ν_T on $\mathbb{S}^1$ such that*

$$U_T = \int_{\sigma(U_T)} \lambda\, \Pi_\lambda d\nu_T(\lambda),$$

and $\nu_T(\lambda) \neq 0$ if and only if λ is an eigenvalue of U_T. Using the above properties of orthogonal projections, we also get

$$U_T^n = \int_{\sigma(U_T)} \lambda^n \Pi_\lambda d\nu_T(\lambda).$$

9 Bernoulli shifts

Let (Σ, σ, μ) be a Bernoulli shift, say with alphabet $\mathcal{A} = \{1, 2, \dots, N\}$. Here $\Sigma = \mathcal{A}^{\mathbb{Z}}$ (two-sided) or $\Sigma = \mathcal{A}^{\mathbb{N}\cup\{0\}}$ (one-sided), and μ is a stationary product measure with probability vector $(p_1, \dots, p_N)$. Write

$$Z_{[k+1,k+N]}(a_1 \dots a_N) = \{x \in \Sigma : x_{k+1} \dots x_{k+N} = a_1 \dots a_N\}$$

for the cylinder set of length N. If $C = Z_{[k+1,k+R]}$ and $C' = Z_{[l+1,l+S]}$ are two cylinders fixing coordinates on disjoint integer intervals (*i.e.,* $[k+1, k+R] \cap [l+1, l+S] = \varnothing$), then clearly $\mu(C \cap C') = \mu(C)\mu(C')$. This just reflects the independence of disjoint events in a sequence of Bernoulli trials.

Definition 6. *Two measure preserving dynamical systems $(X, \mathcal{B}, T, \mu)$ and $(Y, \mathcal{C}, S, \nu)$ are called* **isomorphic** *if there are $X' \in \mathcal{B}$, $Y' \in \mathcal{C}$ and $\phi : Y' \to X'$ such that*

- *$\mu(X') = 1$, $\nu(Y') = 1$;*
- *$\phi : Y' \to X'$ is a bi-measurable bijection;*
- *ϕ is measure preserving: $\nu(\phi^{-1}(B)) = \mu(B)$ for all $B \in \mathcal{B}$.*
- *$\phi \circ S = T \circ \phi$.*

Clearly invertible systems cannot be isomorphic to non-invertible systems. But there is a construction to make a non-invertible system invertible, namely by passing to the natural extension.

Definition 7. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving dynamical system. A system $(Y, \mathcal{C}, S, \nu)$ is a* **natural extension** *of $(X, \mathcal{B}, \mu, T)$ if there are $X' \in \mathcal{B}$, $Y' \in \mathcal{C}$ and $\phi : Y' \to X'$ such that*

- *$\mu(X') = 1$, $\nu(Y') = 1$;*
- *$S : Y' \to Y'$ is* **invertible***;*
- *$\phi : Y' \to X'$ is a measurable surjection;*
- *ϕ is measure preserving: $\nu(\phi^{-1}(B)) = \mu(B)$ for all $B \in \mathcal{B}$;*
- *$\phi \circ S = T \circ \phi$.*

Any two natural extensions can be shown to be isomorphic, so it makes sense to speak of **the** natural extension. Sometimes natural extensions have explicit formulas (such as the baker transformation being the natural extension of the angle doubling map). There is also a general construction: Set

$$Y = \{(x_i)_{i \geqslant 0} : T(x_{i+1}) = x_i \in X \text{ for all } i \geqslant 0\}$$

with $S(x_0, x_1, \dots) = T(x_0), x_0, x_1, \dots$. Then S is invertible (with the left shift $\sigma = S^{-1}$) and

$$\nu(A_0, A_1, A_2, \dots) = \inf_i \mu(A_i) \quad \text{for } (A_0, A_1, A_2 \dots) \subset S,$$

is S-invariant. Now defining $\phi(x_0, x_1, x_2, \dots) := x_0$ makes the diagram commute: $T \circ \phi = \phi \circ S$. Also ϕ is measure preserving because, for each $A \in \mathcal{B}$,

$$\phi^{-1}(A) = (A, T^{-1}(A), T^{-2}(A), T^{-3}(A), \dots)$$

and clearly $\nu(A, T^{-1}(A), T^{-2}(A), T^{-3}(A), \dots) = \mu(A)$ because $\mu(T^{-i}(A)) = \mu(A)$ for every i by T-invariance of μ.

Definition 8. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving dynamical system.*

- *If T is invertible, then the system is called* **Bernoulli** *if it is isomorphic to a Bernoulli shift.*
- *If T is non-invertible, then the system is called* **one-sided Bernoulli** *if it is isomorphic to a one-sided Bernoulli shift.*
- *If T is non-invertible, then the system is called* **Bernoulli** *if its natural extension is isomorphic to a one-sided Bernoulli shift.*

The Bernoulli property is quite general, even though the isomorphism ϕ may be very difficult to find explicitly. Expanding circle maps that satisfy the conditions of Theorem 5 are also Bernoulli, *i.e.,* have a Bernoulli natural extension, see [11]. Being **one-sided Bernoulli**, on the other hand quite, is special. If $T : [0,1] \to [0,1]$ has N linear surjective branches I_i, $i = 1, \dots, N$, then Lebesgue measure m is invariant, and $([0,1], \mathcal{B}, m, T)$ is isomorphic to the one-sided Bernoulli system with probability vector $(|I_1|, \dots, |I_N|)$. If T is piecewise C^2 but not piecewise linear, then it has to be C^2-conjugate to a piecewise linear expanding map to be one-sided Bernoulli, see [6].

10 Mixing and weak mixing

Whereas Bernoulli trials are totally independent, mixing refers to an **asymptotic independence**:

Definition 9. *A probability measure preserving dynamical systems* $(X, \mathcal{B}, \mu, T)$ *is* **mixing** *(or* **strong mixing***) if*

$$\mu(T^{-n}(A) \cap B) \to \mu(A)\mu(B) \text{ as } n \to \infty \tag{6}$$

for every $A, B \in \mathcal{B}$*.*

Proposition 5. *A probability preserving dynamical systems* $(X, \mathcal{B}, T, \mu)$ *is mixing if and only if*

$$\int_X f \circ T^n(x) \cdot \overline{g(x)} \, d\mu \to \int_X f(x) \, d\mu \cdot \int_X \overline{g(x)} \, d\mu \text{ as } n \to \infty \tag{7}$$

for all $f, g \in L^2(\mu)$*, or written in the notation of the Koopman operator* $U_T f = f \circ T$ *and inner product* $\langle f, g \rangle = \int_X f(x) \cdot \overline{g(x)} \, d\mu$*:*

$$\langle U_T^n f, g \rangle \to \langle f, 1 \rangle \langle 1, g \rangle \text{ as } n \to \infty. \tag{8}$$

Proof. The "if"-direction follows by taking indicator functions $f = 1_A$ and $g = 1_B$. For the "only if"-direction, general $f, g \in L^2(\mu)$ can be approximated by linear combinations of indicator functions. □

Definition 10. *A probability measure preserving dynamical systems* $(X, \mathcal{B}, \mu, T)$ *is* **weak mixing** *if in average*

$$\frac{1}{n} \sum_{i=0}^{n-1} |\mu(T^{-i}(A) \cap B) - \mu(A)\mu(B)| \to 0 \text{ as } n \to \infty \tag{9}$$

for every $A, B \in \mathcal{B}$*.*

We can express ergodicity in analogy of (6) and (9):

Lemma 4. *A probability preserving dynamical systems $(X, \mathcal{B}, T, \mu)$ is ergodic if and only if*

$$\frac{1}{n}\sum_{i=0}^{n-1} \mu(T^{-i}(A) \cap B) - \mu(A)\mu(B) \to 0 \text{ as } n \to \infty,$$

for all $A, B \in \mathcal{B}$. (Compared to (9), note the absence of absolute value bars.)

Proof. Assume that T is ergodic, so by Birkhoff's Ergodic Theorem $\frac{1}{n}\sum_{i=0}^{n-1} 1_A \circ T^i(x) \to \mu(A)$ for μ-a.e. x. Multiplying by 1_B gives

$$\frac{1}{n}\sum_{i=0}^{n-1} 1_A \circ T^i(x) 1_B(x) \to \mu(A) 1_B(x) \quad \mu\text{-a.e.}$$

Integrating over x (using the Dominated Convergence Theorem to swap limit and integral), gives $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} \int_X 1_A \circ T^i(x) 1_B(x)\, d\mu = \mu(A)\mu(B)$.

Conversely, assume that $A = T^{-1}A$ and take $B = A$. Then we obtain $\mu(A) = \frac{1}{n}\sum_{i=0}^{n-1} \mu(T^{-i}(A)) \to \mu(A)^2$, hence $\mu(A) \in \{0, 1\}$. □

Theorem 10. *We have the implications:*

Bernoulli ⇒ mixing ⇒ weak mixing ⇒ ergodic ⇒ recurrent.

None of the reverse implications holds in generality.

Proof. Bernoulli ⇒ mixing holds for any pair of cylinder sets C, C' because $\mu(\sigma^{-n}(C) \cap C) = \mu(C)\mu(C')$ for n sufficiently large. The property carries over to all measurable sets by the Kolmogorov Extension Theorem.

Mixing ⇒ weak mixing is immediate from the definition.

Weak mixing ⇒ ergodic: Let $A = T^{-1}(A)$ be a measurable T-invariant set. Then by weak mixing $\mu(A) = \frac{1}{n}\sum_{i=0}^{n-1} \mu(T^{-i}(A) \cap A) \to \mu(A)\mu(A) = \mu(A^2)$. This means that $\mu(A) = 0$ or 1.

Ergodic ⇒ recurrent. If $B \in \mathcal{B}$ has positive measure, then $A := \cup_{i \in \mathbb{N}} T^{-i}(B)$ is T-invariant up to a set of measure 0, see the Poincaré Recurrence Theorem. By ergodicity, $\mu(A) = 1$, and this is the definition of recurrence, see Definition 5. □

We say that a subset $E \subset \mathbb{N} \cup \{0\}$ has **density** zero if $\lim_n \frac{1}{n}\#(E \cap \{0, \dots, n-1\}) = 0$.

Lemma 5. *Let $(a_i)_{i \geqslant 0}$ be a bounded non-negative sequence of real numbers. Then $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i = 0$ if and only if there is a sequence E of zero density in $\mathbb{N} \cup \{0\}$ such that $\lim_{E \not\ni n \to \infty} a_n = 0$.*

Proof. $\Leftarrow$: Assume that $\lim_{E\not\ni n\to\infty} a_n = 0$ and for $\varepsilon > 0$, take N such that $a_n < \varepsilon$ for all $E \not\ni n \geqslant N$. Also let $A = \sup a_n$. Then

$$\begin{aligned} 0 &\leqslant \frac{1}{n}\sum_{i=0}^{n-1} a_i = \frac{1}{n}\sum_{E\not\ni i=0}^{n-1} a_i + \frac{1}{n}\sum_{E\ni i=0}^{n-1} a_i \\ &\leqslant \frac{NA + (n-N)\varepsilon}{n} + A\frac{1}{n}\#(E\cap\{0,\dots,n-1\}) \to \varepsilon, \end{aligned}$$

as $n\to\infty$. Since $\varepsilon > 0$ is arbitrary, $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i = 0$.

$\Rightarrow$: Let $E_m = \{n : a_n \geqslant \frac{1}{m}\}$. Then clearly $E_1 \subset E_2 \subset E_3 \subset \dots$ and each E_m has density 0 because

$$0 = m\cdot\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i \geqslant \lim_n \frac{1}{n}\sum_{i=0}^{n-1} 1_{E_m}(i) = \lim_n \frac{1}{n}\#(E_m\cap\{0,\dots n-1\}).$$

Now take $0 = N_0 < N_1 < N_2 < \dots$ such that $\frac{1}{n}\#(E_m\cap\{0,\dots,n-1\}) < \frac{1}{m}$ for every $n \geqslant N_{m-1}$. Let $E = \cup_m (E_m \cap \{N_{m-1},\dots,N_m - 1\})$.

Then, taking $m = m(n)$ maximal such that $N_{m-1} < n$,

$$\begin{aligned} &\frac{1}{n}\,\#\,(E\cap\{0,\dots,n-1\}) \\ &\quad\leqslant \frac{1}{n}\#(E_{m-1}\cap\{0,\dots,N_{m-1}-1\}) + \frac{1}{n}\#(E_m\cap\{N_{m-1},\dots,n-1\}) \\ &\quad\leqslant \frac{1}{N_{m-1}}\#(E_{m-1}\cap\{0,\dots,N_{m-1}-1\}) + \frac{1}{n}\#(E_m\cap\{0,\dots,n-1\}) \\ &\quad\leqslant \frac{1}{m-1} + \frac{1}{m} \to 0 \end{aligned}$$

as $n\to\infty$. $\square$

Corollary 2. *For a non-negative sequence $(a_n)_{n\geqslant 0}$ of real numbers, $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i = 0$ if and only if $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i^2 = 0$.*

Proof. By the previous lemma, $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i = 0$ if and only if $\lim_{E\not\ni n\to\infty} a_n = 0$ for a set E of zero density. But the latter is clearly equivalent to $\lim_{E\not\ni n\to\infty} a_n^2 = 0$ for the same set E. Applying the lemma again, we have $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} a_i^2 = 0$. $\square$

Example 1. *Let $R_\alpha : \mathbb{S}^1 \to \mathbb{S}^1$ be an irrational circle rotation; it preserves Lebesgue measure. We claim that R_α is not mixing or weak mixing, but it is ergodic. To see why R_α is not mixing, take an interval A of length $\frac{1}{4}$. There are infinitely many n such that $R_\alpha^{-n}(A)\cap A = \varnothing$, so $\liminf_n \mu(R^{-n}(A)\cap A) = 0 \neq (\frac{1}{4})^2$. However, R_α has a non-constant eigenfunction $\psi : \mathbb{S}^1 \to \mathbb{C}$ defined as $\psi(x) = e^{2\pi i x}$ because $\psi\circ R_\alpha(x) = e^{2\pi i(x+\alpha)} = e^{2\pi i\alpha}\psi(x)$. Therefore R_α is not weak mixing, see Theorem 11 below. To* *prove ergodicity, we show that every T-invariant function $\psi \in L^2(m)$ must be constant. Indeed, write $\psi(x) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi i n x}$ as a Fourier series. The T-invariance implies that $a_n e^{2\pi i\alpha} = a_n$ for all $n \in \mathbb{Z}$. Since $\alpha \notin \mathbb{Q}$, this means that $a_n = 0$ for all $n \neq 0$, so $\psi(x) \equiv a_0$ is indeed constant.*

Theorem 11. *Let $(X, \mathcal{B}, \mu, T)$ be a probability measure preserving dynamical system. Then the following are equivalent:*

1. *$(X, \mathcal{B}, \mu, T)$ is weak mixing;*
2. *$\lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle f\circ T^i, g\rangle - \langle f, 1\rangle\langle 1, g\rangle| = 0$ for all $L^2(\mu)$ functions f, g;*
3. *$\lim_{E\not\ni n\to\infty} \mu(T^{-n}A \cap B) = \mu(A)\mu(B)$ for all $A, B \in \mathcal{B}$ and a subset E of zero density;*
4. *$T \times T$ is weak mixing;*
5. *$T \times S$ is ergodic on (X, Y) for every ergodic system $(Y, \mathcal{C}, \nu, S)$;*
6. *$T \times T$ is ergodic;*
7. *The Koopman operator U_T has no measurable eigenfunctions other than constants.*

Proof. 2. $\Rightarrow$ 1. Take $f = 1_A$, $g = 1_B$.
1. $\Leftrightarrow$ 3. Use Lemma 5 for $a_i = |\mu(T^{-i}(A) \cap B) - \mu(A)\mu(B)|$.
3. $\Rightarrow$ 4. For every $A, B, C, D \in \mathcal{B}$, there are subsets E_1 and E_2 of $\mathbb{N}$ of zero density such that

$$\lim_{E_1\not\ni n\to\infty} \mu(T^{-n}(A)\cap B) - \mu(A)\mu(B)| = \lim_{E_2\not\ni n\to\infty} \mu(T^{-n}(C)\cap D) - \mu(C)\mu(D)| = 0.$$

The union $E = E_1 \cup E_2$ still has density 0, and

$$\begin{aligned} 0 \leqslant \lim_{E\not\ni n\to\infty} &|\ \mu\times\mu\ ((T\times T)^{-n}(A\times C)\cap(B\times D)) - \mu\times\mu(A\times B)\cdot\mu\times\mu(C\times D)| \\ = &\lim_{E\not\ni n\to\infty} |\mu(T^{-n}(A)\cap B)\cdot\mu(T^{-n}(C)\cap D) - \mu(A)\mu(B)\mu(C)\mu(D)| \\ \leqslant &\lim_{E\not\ni n\to\infty} \mu(T^{-n}(A)\cap B)\cdot|\mu(T^{-n}(C)\cap D) - \mu(C)\mu(D)| \\ &+ \lim_{E\not\ni n\to\infty} \mu(C)\mu(D)\cdot|\mu(T^{-n}(A)\cap B) - \mu(A)\mu(B)| = 0. \end{aligned}$$

4. $\Rightarrow$ 5. If $T\times T$ is weakly mixing, then so is T itself. Suppose $(Y, \mathcal{C}, \nu, S)$ is an ergodic

system, then, for $A, B \in \mathcal{B}$ and $C, D \in \mathcal{C}$ we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} \mu\left(T^{-i}(A)\cap B\right)\nu(S^{-i}(C)\cap D) \\
&\qquad = \frac{1}{n}\sum_{i=0}^{n-1} \mu(A)\mu(B)\nu(S^{-i}(C)\cap D) \\
&\qquad\quad + \frac{1}{n}\sum_{i=0}^{n-1} (\mu(T^{-i}(A)\cap B) - \mu(A)\mu(B))\nu(S^{-i}(C)\cap D).
\end{aligned}$$

By ergodicity of S (see Lemma 4), $\frac{1}{n}\sum_{i=0}^{n-1} \nu(S^{-i}(C)\cap D) \to \mu(C)\mu(D)$, so the first term in the above expression tends to $\mu(A)\mu(B)\mu(C)\mu(D)$. The second term is majorised by $\frac{1}{n}\sum_{i=0}^{n-1} |\mu(T^{-i}(A)\cap B) - \mu(A)\mu(B)|$, which tends to 0 because T is weak mixing.

5. $\Rightarrow$ 6. By assumption $T \times S$ is ergodic for the trivial map $S : \{0\} \to \{0\}$. Therefore T itself is ergodic, and hence $T \times T$ is ergodic.

6. $\Rightarrow$ 7. Suppose f is an eigenfunction with eigenvalue λ. The Koopman operator is an isometry (by T-invariance of the measure), so $\langle f, f\rangle = \langle U_T f, U_T f\rangle = \langle \lambda f, \lambda f\rangle = |\lambda|^2\langle f, f\rangle$, and $|\lambda| = 1$. Write $\psi(x, y) = f(x)\overline{f}(y)$. Then

$$\psi \circ (T \times T)(x, y) = \psi(Tx, Ty) = f(Tx)\overline{f(Ty)} = |\lambda|^2\psi(x, y) = \psi(x, y),$$

so ψ is $T \times T$-invariant. By ergodicity of $T \times T$, ψ must be constant $\mu \times \mu$-a.e. But then also f must be constant μ-a.e.

7. $\Rightarrow$ 2. This is the hardest step; it relies on spectral theory of unitary operators. If ψ is an eigenfunction of U_T, then by assumption, ψ is constant, so the eigenvalue is 1. Let $V = \text{span}\,(\psi)$ and Π_1 is the orthogonal projection onto V; clearly $V^\perp = \{f \in L^2(\mu) : \int f\, d\mu = 0\}$. One can derive that the spectral measure ν_T cannot have any atoms, except possibly at Π_1.

Now take $f \in V^\perp$ and $g \in L^2(\mu)$ arbitrary. Using the Spectral Theorem 9, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i f, g\rangle|^2 &= \frac{1}{n}\sum_{i=0}^{n-1} \left| \int_{\sigma(U_T)} \lambda^i \langle \Pi_\lambda f, g\rangle \; d\nu_T(\lambda) \right|^2 \\
&= \frac{1}{n}\sum_{i=0}^{n-1} \int_{\sigma(U_T)} \lambda^i \langle \Pi_\lambda f, g\rangle \; d\nu_T(\lambda) \overline{\int_{\sigma(U_T)} \kappa^i \langle \Pi_\kappa f, g\rangle \; d\nu_T(\kappa)} \\
&= \frac{1}{n}\sum_{i=0}^{n-1} \int\int_{\sigma(U_T)\times\sigma(U_T)} \lambda^i \overline{\kappa}^i \; \langle \Pi_\lambda f, g\rangle \overline{\langle \Pi_\kappa f, g\rangle} \; d\nu_T(\lambda)\; d\nu_T(\kappa) \\
&= \int\int_{\sigma(U_T)\times\sigma(U_T)} \frac{1}{n}\sum_{i=0}^{n-1} \lambda^i \overline{\kappa}^i \; \langle \Pi_\lambda f, g\rangle \overline{\langle \Pi_\kappa f, g\rangle} \; d\nu_T(\lambda)\; d\nu_T(\kappa) \\
&= \int\int_{\sigma(U_T)\times\sigma(U_T)} \frac{1}{n}\frac{1 - (\lambda\overline{\kappa})^n}{1 - \lambda\overline{\kappa}} \; \langle \Pi_\lambda f, g\rangle \overline{\langle \Pi_\kappa f, g\rangle} \; d\nu_T(\lambda)\; d\nu_T(\kappa),
\end{aligned}$$

where in the final line we used that the diagonal $\{\lambda = \kappa\}$ has $\nu_T \times \nu_T$-measure zero, because ν is non-atomic (except possibly the atom Π_1 at $\lambda = 1$, but then $\Pi_1 f = 0$). Now $\frac{1}{n}\frac{1-(\lambda\overline{\kappa})^n}{1-\lambda\overline{\kappa}}$ is bounded (use l'Hôpital's rule) and tends to 0 for $\lambda \neq \kappa$, so by the Bounded Convergence Theorem, we have

$$\lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i f, g\rangle|^2 = 0.$$

Using Corollary 2, we derive that also $\lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i f, g\rangle| = 0$ (*i.e.*, without the square). Finally, if $f \in L^2(\mu)$ is arbitrary, then $f - \langle f, 1\rangle \in V^\perp$. We find

$$\begin{aligned} 0 &= \lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i (f - \langle f, 1\rangle), g\rangle| \\ &= \lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i f - \langle f, 1\rangle, g\rangle| \\ &= \lim_n \frac{1}{n}\sum_{i=0}^{n-1} |\langle U_T^i f, g\rangle - \langle f, 1\rangle\langle 1, g\rangle| \end{aligned}$$

and so property 2. is verified. □

11 Cutting and Stacking

The purpose of **cutting and stacking** is to create invertible maps of the interval that preserve Lebesgue measure, and have further good properties such as "unique ergodicity", "not weak mixing", or rather the opposite "weak mixing but not strong mixing". Famous examples due to Kakutani and to Chacon achieve this, and we will present them here.

The procedure is as follows:

- Cut the unit interval into several intervals, say $A, B, C, \dots$ (these will become the **stacks**), and a remaining interval S (called the **spacer**).
- Cut each interval into parts (a fix finite number for each stack), and also cut of some intervals from the spacer.
- Pile the parts of the stacks and the cut-off pieces of the spacer on top of the stacks, according to some fixed rule. By choosing the parts in the previous step of the correct size, we can ensure that all intervals in each separate stack have the same size; they can therefore be neatly aligned vertically.

- Map every point on a level of a stack directly to the level above. Then every point has a well-defined image (except for points at the top levels in a stack and points in the remaining spacer), and also a well-defined preimage (except for points at a bottom level in a stack and points in the remaining spacer). Where defined, Lebesgue measure is preserved.

- Repeat the process, now slicing vertically through whole stacks and stacking whole stacks on top of other stacks, possibly putting some intervals of the spacer in between. Wherever the map was defined at a previous step, the definition remains the same.

- Keep repeating. Eventually, the measure of points where the map is not defined tends to zero. In the end, assuming that the spacer will be entirely spent, there will only be one point for each stack without image and one points in each stack without preimage. We can take an arbitrary bijection between them to define the map everywhere.

- The resulting transformation of the interval is invertible and preserves Lebesgue measure. The number of stacks used is called the **rank** of the transformation.

Example 2 (Kakutani)**.** *Take one stack, so start with $A = [0, 1]$. Cut it in half and put the right half on top of the left half. Repeat this procedure. Let us call the result limit map $T : [0, 1] \to [0, 1]$ the Kakutani map. The resulting formula is:*

$$T(x) = \begin{cases} x + \frac{1}{2} & \text{if } x \in [0, \frac{1}{2}); \\ x - \frac{1}{4} & \text{if } x \in [\frac{1}{2}, \frac{3}{4}); \\ x - \frac{1}{2} - \frac{1}{8} & \text{if } x \in [\frac{3}{4}, \frac{7}{8}); \\ \vdots & \vdots \\ x - (1 - \frac{1}{2^n} - \frac{1}{2^{n+1}}) & \text{if } x \in [1 - \frac{1}{2^n}, 1 - \frac{1}{2^{n+1}}), n \geqslant 1, \end{cases}$$

see Figure 2. If $x \in [0, 1)$ is written in base 2, i.e.,

$$x = 0.b_1 b_2 b_3 \ldots \qquad b_i \in \{0, 1\}, \quad x = \sum_i b_i 2^{-i},$$

then T acts as the **adding machine***: add* 0.1 *with carry. That is, if $k = \min\{i \geqslant 1 : b_i = 0\}$, then $T(0.b_1 b_2 b_3 \ldots) = 0.001 b_{k+1} b_{k+2} \ldots$. If $k = \infty$, so $x = 0.111111\ldots$, then $T(x) = 0.0000\ldots$.*

Proposition 6. *The Kakutani map $T : [0, 1] \to [0, 1]$ of cutting and stacking is uniquely ergodic, but not weakly mixing.*

Proof. The map T permutes the dyadic intervals cyclically. For example $T((0, \frac{1}{2})) = (\frac{1}{2}, 1))$ and $T((\frac{1}{2}, 1)) = (0, \frac{1}{2})$. Therefore, $f(x) = 1_{(0,\frac{1}{2})} - 1_{(\frac{1}{2},1)}$ is an eigenfunction for

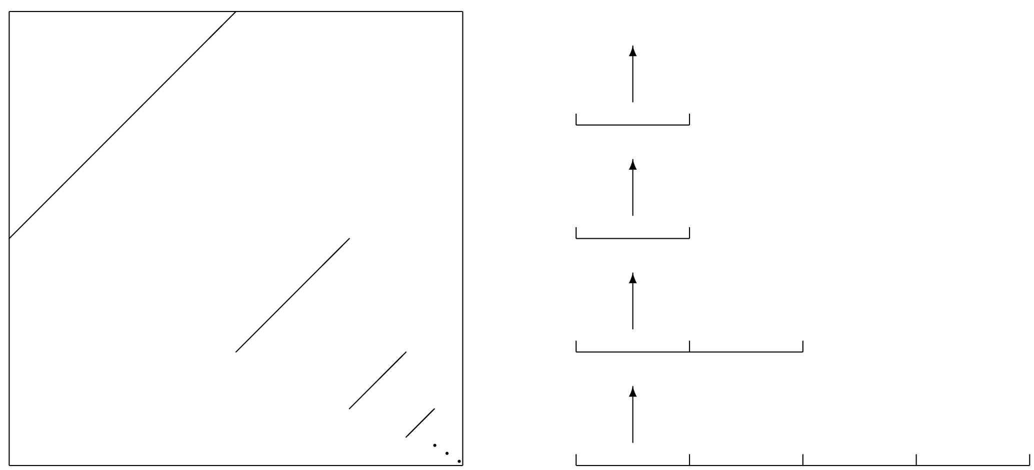

Figure 2: The Kakutani map $T : [0,1] \to [0,1]$ resulting from cutting and stacking.

eigenvalue -1. Using four intervals, we can construct (complex-valued) eigenfunctions for eigenvalues $\pm i$. In generality, all the numbers $e^{2\pi i m/2^n}$, $m, n \in \mathbb{N}$ are eigenvalues, and the corresponding eigenfunctions span $L^2(m)$. This property is called **pure point spectrum**. In any case, T is not weakly mixing.

Now for unique ergodicity, we use the fact again that T permutes the dyadic intervals cyclically. Call these intervals $D_{j,N} = [\frac{j}{2^N}, \frac{j+1}{2^N})$ for $N \in \mathbb{N}$ and $j = \{0, 1, \dots, 2^N - 1\}$, and if $x \in [0,1)$, we indicate the dyadic interval containing it by $D_{j,N}(x)$. Let

$$\begin{cases} \overline{f}_N(x) = \sup_{t \in D_{j,N}(x)} f(t), \\ \underline{f}_N(x) = \inf_{t \in D_{j,N}(x)} f(t), \end{cases}$$

be step-functions that we can use to compute the Riemann integral of f. That is:

$$\int \overline{f}_N(s) ds := \frac{1}{2^N} \sum_{j=0}^{2^N-1} \sup_{t \in D_{j,N}} f(t) \geqslant \int f(s) ds \geqslant \int \underline{f}_N(s) ds := \frac{1}{2^N} \sum_{j=0}^{2^N-1} \inf_{t \in D_{j,N}} f(t).$$

For continuous (or more generally Riemann integrable) functions, $\int \overline{f}_N dx - \int \underline{f}_N dx \to 0$ as $N \to \infty$, and their common limit is called the Riemann integral of f.

According to Lemma 1, we need to show that $\frac{1}{n} \sum_{i=0}^{N-1} f \circ T^i(x)$ converges uniformly to a constant (for each continuous function f) to show that T is uniquely ergodic, *i.e.,* Lebesgue measure is the unique invariant measure.

Let $f : [0,1] \to \mathbb{R}$ be continuous and $\varepsilon > 0$ be given. By uniform continuity, we can find N such that $\max_j(\sup_{t \in D_{j,N}} f(t) - \inf_{t \in D_{j,N}} f(t)) < \varepsilon$. Write $n = m2^N + r$. Any orbit x will visit all intervals $D_{j,N}$ cyclically before returning close to itself, and hence visit each $D_{j,N}$ exactly m times in the first $m2^N$ iterates. Therefore

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n-1} f\circ T^i(x) &\leqslant \frac{1}{m2^N+r}\left(\sum_{j=0}^{2^N-1} m \sup_{t\in D_{j,N}} f(t) + r\|f\|_\infty\right) \\
&\leqslant \frac{1}{2^N}\sum_{j=0}^{2^N-1} \sup_{t\in D_{j,N}} f(t) + \frac{r\|f\|_\infty}{m2^N+r} = \int \overline{f}_N(s)ds + \frac{r\|f\|_\infty}{m2^N+r} \to \int \overline{f}_N(s)ds,
\end{aligned}$$

as $m\to\infty$. A similar computation gives $\frac{1}{n}\sum_{i=0}^{n-1} f\circ T^i(x) \geqslant \int \underline{f}_N(x)dx$. As $\varepsilon\to 0$ (and hence $N\to\infty$), we get convergence to the integral $\int f(s)ds$, independently of the initial point x. $\square$

Example 3 (Chacon)**.** *Take one stack and one stack: $A_0 = [0,\frac{2}{9})$ and $S = [\frac{2}{3},1)$. Cut A_0 is three equal parts and cut $[\frac{2}{3},\frac{8}{9})$ from spacer S. Pile the middle interval $[\frac{2}{9},\frac{4}{9})$ on the left, then the cut-off piece $[\frac{2}{3},\frac{8}{9})$ of the spacer, and then remaining interval $[\frac{4}{9},\frac{2}{3})$. The stack can now be coded upward as $A_1 = A_0A_0SA_0$.*

Repeat this procedure: cut the stack vertically in three stacks (of width $\frac{2}{27}$), cut an interval $[\frac{8}{9},\frac{26}{27})$ from the spacer, and pile them on top of one another: middle stack on left, then the cut-off piece of the spacer, and then the remaining third of the stack. The stack can now be coded upward as $A_2 = A_1A_1SA_1$.

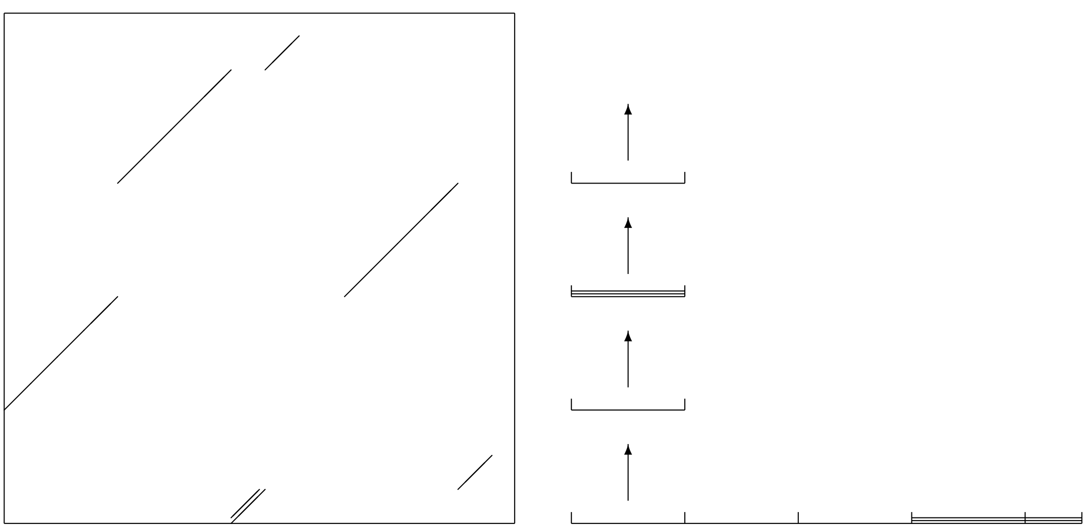

Figure 3: The Chacon map $T:[0,1]\to[0,1]$ resulting from cutting and stacking.

Proposition 7. *The Chacon map $T:[0,1]\to[0,1]$ of cutting and stacking is uniquely ergodic, weakly mixing but not strongly mixing.*

Sketch of Proof. First some observations on the symbolic pattern that emerges of the Chacon cutting and stacking. When stacking intervals, their labels follow the following pattern

$$\underbrace{\underbrace{A_0A_0SA_0}_{A_1}\underbrace{A_0A_0SA_0}_{A_1}\,S\,\underbrace{A_0A_0SA_0}_{A_1}}_{A_2}\underbrace{\underbrace{A_0A_0SA_0}_{A_1}\underbrace{A_0A_0SA_0}_{A_1}\,S\,\underbrace{A_0A_0SA_0}_{A_1}}_{A_2}\,S\,\underbrace{\underbrace{A_0A_0SA_0}_{A_1}\underbrace{A_0A_0SA_0}_{A_1}\,S\,\underbrace{A_0A_0SA_0}_{A_1}}_{A_2}$$

$$\underbrace{A_2A_2SA_2}_{A_3}$$

This pattern is the same at every level; we could have started with A_n, grouped together as $A_{n+1} = A_nA_nSA_n$, etc. At step n in the construction of the tower, the width of the stack is $w_n = \frac{2}{3}(3^{-(n+1)}$ and the length of the the word A_n is $l_n = \frac{1}{2}(3^{n+1} - 1)$.

The frequency of each block $\sigma^k(A_n)$ is almost the same in every block huge block B, regardless where taken in the infinite string. This observation leads to unique ergodicity (similar although a bit more involved as in the case of the Kakutani map), but we will skip the details.

Instead, we focus on the weak mixing. Clearly the word A_n appears in triples, and also as $A_nA_nA_nSA_nA_nA_n$. To explain the idea behind the proof, pretend that an eigenfunction (with eigenvalue $e^{2\pi i\lambda}$) were constant on any set E whose code is A_n (or σ^kA_n for some $0 \leqslant k < l_n$, where σ denotes the left-shift). Such set E are intervals of width w_n. Then

$$f \circ T^{l_n}|_E = e^{2\pi i\lambda l_n} f|_E \text{ and } f \circ T^{2l_n+1}|_E = e^{2\pi i\lambda l_n} f|_E$$

This gives $1 = e^{2\pi i\lambda l_n} = e^{2\pi i\lambda l_n}$, so $\lambda = 0$, and the eigenvalue is 1 after all.

The rigorous argument is as follows. Suppose that $f(x) = e^{2\pi i\vartheta(x)}$ were an eigenfunction for eigenvalue $e^{2\pi i\lambda}$ and a measurable function $\vartheta : \mathbb{S}^1 \to \mathbb{R}$. By Lusin's Theorem, we can find a subset $F \subset \mathbb{S}^1$ of Lebesgue measure $\geqslant 1 - \varepsilon$ such that ϑ is uniformly continuous on F. Choose $\varepsilon > 0$ arbitrary, and take N so large that the variation of ϑ is less that ε on any set of the form $E \cap F$, where points in E have code starting as $\sigma^k(A_N)$, $0 \leqslant k < l_N$. Sets of this type fill a set E^* with mass at least half of the unit interval.

Because of the frequent occurrence of $A_NA_NA_NSA_NA_NA_N$, a definite proportion of E^* is covered by set E with the property that such that $T^{2l_N+1} \cap T^{l_N}E \cap E \neq \varnothing$, because they have codes of length l_N that reappear after both l_N and $2l_N + 1$ shifts. For x in this intersection,

$$\begin{cases} \vartheta \circ T^{2l_N+1}(x) = (l_N + 1)\lambda + \vartheta \circ T^{l_N}(x) \pmod 1 \\ \vartheta \circ T^{l_N}(x) = l_N\lambda + \vartheta(x) \pmod 1 \end{cases}$$

where all three point $x, T^{l_N}(x), T^{2l_N+1}(x)$ belong to the same copy E. Subtracting the two equations gives

$$\lambda \bmod 1 = \vartheta \circ T^{2l_N+1}(x) - \vartheta \circ T^{l_N}(x) + \vartheta(x) - \vartheta \circ T^{l_N}(x) \leqslant 2\varepsilon.$$

But ε is arbitrary, so $\lambda = 0 \mod 1$ and the eigenvalue is 1.

Now for the strong mixing, consider once more the sets $E = E_{k,n}$ of points whose codes starts as the k-th cyclic permutation of A_n for some $0 \leqslant k < l_n$, that is: the first l_n symbols of $\sigma^k(A_nA_n)$. Their measure is $\mu(E) = w_n$, and for different k, they are disjoint. Furthermore, the only l_n-block appearing are cyclic permutations of A_n or cyclic permutations with a spacer S inserted somewhere. At least half of these appearances are of the first type, so $\mu(\cup_{k=0}^{l_n-1} E_{k,n}) \geqslant \frac{1}{2}$ for each n.

The basic idea is that $\mu(E \cap T^{-l_n} E) \geqslant \frac{1}{3}\mu(E)$ because at least a third of the appearances of A_n is followed by another A_n. But $\frac{1}{3}\mu(E) \gg \mu(E)^2$, as one would expect for mixing. Of course, mixing only says that $\lim_l \mu(Y \cap T^{-l}(E)) = \mu(Y)^2$ only for sets Y not depending on l.

However, let $Y_m = [m/8, (m+1)/8] \subset [0,1]$, $m = 0, \dots, 7$ be the eight dyadic intervals of length 1/8. For each n, at least one Y_m is covered for at least half by sets E of the above type, say a set $Z \subset Y_m$ of measure $\mu(Z) \geqslant \frac{1}{2}\mu(Y_m))$ such that $Z \subset \cup_k E_{k,n}$. That means that

$$\mu(Y_m \cap T^{-l_n}(Y_m)) \geqslant \mu(Z \cap T^{-l_n}(Z)) \geqslant \frac{1}{3}\mu(Z) \geqslant \frac{1}{6}\mu(Y_m) > \mu(Y_m)^2.$$

Let Y be one of the Y_m's for which the above holds for infinitely many n. Then $\limsup_n \mu(Y_m \cap T^{-l_n}(Y_m)) > \mu(Y)^2$, contradicting strong mixing. $\square$

12 Toral automorphisms

The best known example of a toral automorphism (that is, an invertible linear map on the torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$) is the Arnol'd cat map. This map $T_C : \mathbb{T}^2 \to \mathbb{T}^2$ is defined as

$$T_C(x,y) = C\begin{pmatrix} x \\ y \end{pmatrix} \pmod 1 \qquad \text{for the matrix } C = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

The name come from the illustration in Arnol'd's book [3] showing how the head of a cat, drawn on a torus, is distorted by the action of the map[2]. Properties of T_C are:

- C preserves the integer lattice, so T_C is well-defined an continuous.
- $\det(C) = 1$, so Lebesgue measure m is preserved (both by C and T_C). Also C and T_C are invertible, and C^{-1} is still an integer matrix.
- The eigenvalues of C are $\lambda_\pm = (3 \pm \sqrt{5})/2$, and the corresponding eigenspaces $E_\pm$ are spanned $(-1, (\sqrt{5}+1)/2)^T$ and $(1, (\sqrt{5}-1)/2)^T$. These are orthogonal

[2]Arnol'd didn't seem to like cats, but see the applet https://www.jasondavies.com/catmap/ how the cat survives

(naturally, since C is symmetric), and have irrational slopes, so they wrap densely in the torus.

- Every rational point in $\mathbb{T}^2$ is periodic under T (as their denominators cannot increase, so T acts here as an invertible map on a finite set). This gives many invaraint measures: the equidistribution on each periodic orbit. Therefore T_C is not uniquely ergodic.

The properties are common to all maps T_A, provided they satisfy the following definition.

Definition 11. *A toral automorphism $T : \mathbb{T}^d \to \mathbb{T}^d$ is an invertible linear map on the (d-dimensional) torus $\mathbb{T}^d$. Each such T is of the form $T_A(x) = Ax \pmod 1$, where the matrix A satisfies:*

- *A is an integer matrix with $\det(A) = \pm 1$;*
- *the eigenvalues of A are not on the unit circle; this property is called* **hyperbolicity**.

Somewhat easier to treat that the cat map is T_A for $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which is an orientation reversing matrix with $A^2 = C$. The map T_A has a **Markov partition**, that is a partition $\{R_i\}_{i=1}^N$ for sets such that

1. The R_i have disjoint interiors and $\cup_i R_i = \mathbb{T}^d$;
2. If $T_A(R_i) \cap R_j \neq \varnothing$, then $T_A(R_i)$ stretches across R_j in the unstable direction (*i.e.,* the direction spanned by the unstable eigenspaces of A).
3. If $T_A^{-1}(R_i) \cap R_j \neq \varnothing$, then $T_A^{-1}(R_i)$ stretches across R_j in the stable direction (*i.e.,* the direction spanned by the stable eigenspaces of A).

In fact, every hyperbolic toral automorphism has a Markov partition, but in general they are fiendishly difficult to find explicitly. In the case of A, a Markov partition of three rectangles R_i for $i = 1, 2, 3$ can be constructed, see Figure 4.

The corresponding transition matrix is

$$B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{where } B_{ij} = \begin{cases} 1 & \text{if } T_A(R_i) \cap R_j \neq \varnothing \\ 0 & \text{if } T_A(R_i) \cap R_j = \varnothing. \end{cases}$$

Note that the characteristic polynomial of B is

$$\det(B - \lambda I) = -\lambda^3 + 2\lambda + 1 = -(\lambda + 1)(\lambda^2 - \lambda - 1) = -(\lambda + 1) \det(A - \lambda I).$$

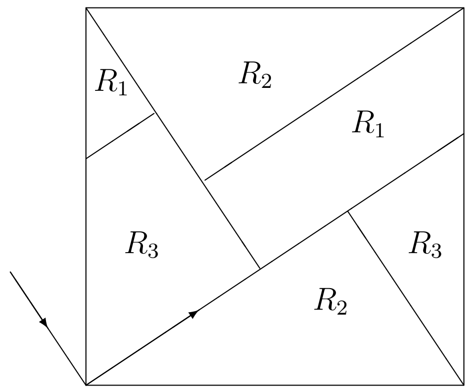

Figure 4: The Markov partition for the toral automorphism T_A. The arrows indicate the stable and unstable directions at $(0,0)$.

so B has the eigenvalues of A (no coincidence!), together with $\lambda = -1$. The transition matrix B generates a **subshift of finite type**:

$$\Sigma_B = \{(x_i)_{i \in \mathbb{Z}} : x_i \in \{1,2,3\}, B_{x_i x_{i+1}} = 1 \ \forall \ i \in \mathbb{Z}\},$$

equipped with the left-shift σ. That is, Σ_B contains only sequences in which each $x_i x_{i+1}$ indicate transitions from Markov partition elements that are allowed by the map T_A.

It can be shown that $(\mathbb{T}^d, \mathcal{B}, T, Leb)$ is isomorphic to the shift space $(\Sigma_B, \mathcal{C}, \sigma, \mu)$ where

$$\mu([x_k x_{k+1} \dots x_n]) = m_{x_k} \Pi_{x_k x_{k+1}} \Pi_{x_{k+1} x_{k+2}} \dots \Pi_{x_{n-1} x_n},$$

for $m_i = Leb(R_i)$, $i = 1, \dots, d\}$, and weighted transition matrix Π where

$$\Pi_{ij} = \frac{Leb(T_A(R_i) \cap R_j)}{Leb(R_i)} \text{ is the relative mass that } T_A \text{ transports from } R_i \text{ to } R_j.$$

Finally $\mathcal{C}$ the σ-algebra of set generated by allowed cylinder sets.

Theorem 12. *For every hyperbolic toral automorphism, Lebesgue measure is ergodic and mixing.*

Proof. We only give the proof for dimension 2. The higher dimensional case goes similarly. Consider the Fourier modes (also called **characters**)

$$\chi_{(m,n)} : \mathbb{T}^2 \to \mathbb{C}, \qquad \chi_{(m,n)}(x,y) = e^{2\pi i(mx+ny)}.$$

These form an orthogonal system (w.r.t. $\langle \varphi, \psi \rangle = \int \varphi \overline{\psi} d\lambda$), spanning $L^2(\lambda)$ for Lebesgue measure λ. We have

$$U_{T_A} \chi_{(m,n)}(x,y) = \chi_{(m,n)} \circ T_A(x,y) = \chi_{m,n}(x,y) = e^{2\pi i((am+cn)x+(bm+dn)y)} = \chi_{A^t(m,n)}(x,y).$$

In other words, U_{T_A} maps the character with index (m,n) to the character with index $A^t(m,n)$, where A^t is the transpose matrix.

For the proof of ergodicity, assume that φ is a T_A-invariant L^2-function. Write it as Fourier series:

$$\varphi(x,y) = \sum_{m,n\in\mathbb{Z}} \varphi_{(m,n)}\chi_{(m,n)}(x,y),$$

where the Fourier coefficients $\varphi_{m,n} \to 0$ as $|m|+|n| \to \infty$ By T_A-invariance, we have

$$\varphi(x,y) = \varphi \circ T_A(x,y) = \sum_{m,n\in\mathbb{Z}} \varphi_{(m,n)}\chi_{A^t(m,n)}(x,y),$$

and hence $\varphi_{(m,n)} = \varphi_{A^t(m,n)}$ for all m,n. For $(m,n)=(0,0)$ this is not a problem, but this only produces constant functions. If $(m,n) \neq (0,0)$, then the A^t-orbit of (m,n), so infinitely many equal Fourier coefficients

$$\varphi_{(m,n)} = \varphi_{A^t(m,n)} = \varphi_{(A^t)^2(m,n)} = \varphi_{(A^t)^3(m,n)} = \varphi_{(A^t)^4(m,n)} \cdots$$

As the Fourier coefficients converge to zero as $|m|+|n| \to \infty$, they all must be equal to zero, and hence φ is a constant function. This proves ergodicity.

For the proof of mixing, we need a lemma, which we give without proof.

Lemma 6. *A transformation (X,T,μ) is mixing if and only if for all φ, ψ in a complete orthogonal system spanning $L^2(\mu)$, we have*

$$\int_X \varphi \circ T^N(x)\overline{\psi(x)}\,d\mu \to \int_X \varphi(x)\,d\mu \cdot \int_X \overline{\psi(x)}\,d\mu$$

as $N \to \infty$.

To use this lemma on $\varphi = \chi_{(m,n)}$ and $\psi = \chi_{(k,l)}$, we compute

$$\int_X \chi_{(m,n)} \circ T^N(x)\overline{\chi_{(k,l)}(x)}\,d\lambda = \int_X \chi_{(A^t)^N(m,n)}\overline{\chi_{(k,l)}(x)}\,d\lambda.$$

If $(m,n)=(0,0)$, then $(A^t)^N(m,n)=(0,0)=(m,n)$ for all N. Hence, the integral is non-zero only if $(k,l)=(0,0)$, but then the integral equals 1, which is the same as $\int_X \chi_{(0,0)}\,d\lambda \int_X \overline{\chi_{(0,0)}(x)}\,d\lambda$. If $(k,l)=(0,0)$, then the integral is zero, but so is $\int_X \chi_{(0,0)}\,d\lambda \int_X \overline{\chi_{(0,0)}(x)}\,d\lambda$.

If $(m,n)\neq(0,0)$, then, regardless what (k,l) is, there is N such that $(A^t)^M(m,n) \neq (k,l)$ for all $M \geqslant N$. Therefore

$$\int_X \chi_{(m,n)} \circ T^M(x)\overline{\chi_{(k,l)}(x)}\,d\lambda = 0 = \int_X \chi_{(m,n)}\,d\lambda \int_X \overline{\chi_{(k,l)}(x)}\,d\lambda.$$

The lemma therefore guarantees mixing. $\square$

13 Topological entropy and topological pressure

Topological entropy was first defined in 1965 by Adler et al. [1], but the form that Bowen [4] and Dinaburg [8] redressed it in is commonly used nowadays.

We will start by start giving the original definition, because the idea of joints of covers easily relates to joints of partitions as used in measure-theoretic entropy. After that, we will give Bowen's approach, since it readily generalises to topological pressure as well.

13.1 The orginal definition

Let (X, d, T) be a continuous map on compact metric space (X, d). We say that $\mathcal{U} = \{U_i\}$ is an *open ε-cover* if all U_i are open sets of diamter $\leqslant \varepsilon$ and $X \subset \bigcup_i U_i$. Naturally, compactness of X guarantees that for every open cover, we can select a finite subcover. Thus, let $\mathcal{N}(\mathcal{U})$ the the minimal possible cardinality of subcovers of $\mathcal{U}$. We say that $\mathcal{U}$ *refines* $\mathcal{V}$ (notation $\mathcal{U} \succeq \mathcal{V}$) if evey $U \in \mathcal{U}$ is contained in a $V \in \mathcal{V}$. If $\mathcal{U} \succeq \mathcal{V}$ then $\mathcal{N}(\mathcal{U}) \geq \mathcal{N}(\mathcal{V})$.

Given two cover $\mathcal{U}$ and $\mathcal{V}$, the joint

$$\mathcal{U} \vee \mathcal{V} := \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\}$$

is an open cover again, and one can verify that $\mathcal{N}(\mathcal{U} \vee \mathcal{V}) \leqslant \mathcal{N}(\mathcal{U})\mathcal{N}(\mathcal{V})$. Since T is continuous, $T^{-1}(\mathcal{U})$ is an open cover as well, although in this case it need not be an ε-cover; However, $\mathcal{U} \vee T^{-1}(\mathcal{U})$ is an ε-cover, and it refines $T^{-1}(\mathcal{U})$.

Define the *topological entropy* as

$$h_{top}(T) = \lim_{\varepsilon \to 0} \sup_{\mathcal{U}} \lim_{n} \frac{1}{n} \log \mathcal{N}(\mathcal{U}^n) \qquad \text{for } \mathcal{U}^n := \bigvee_{i=0}^{n-1} T^{-i}(\mathcal{U}), \tag{10}$$

where the supremum is taken over all open ε-covers $\mathcal{U}$. Because $\mathcal{N}(\mathcal{U} \vee \mathcal{V}) \leqslant \mathcal{N}(\mathcal{U})\mathcal{N}(\mathcal{V})$, the sequence $\log \mathcal{N}(\mathcal{U}^n)$ is subadditive, so the limit $\lim_n \frac{1}{n} \log \mathcal{N}(\mathcal{U}^n)$ exists. We have the following properties:

Lemma 7.

- *$h_{top}(T^k) = k h_{top}(T)$ for $k \geqslant 0$. If T is invertible, then also $h_{top}(T^{-1}) = h_{top}(T)$.*
- *If (Y, S) is semiconjugate to (X, T), then $h_{top}(S) \leqslant h_{top}(T)$. In particular, conjugate systems (on compact spaces!) have the same entropy.*

Proof. □

13.2 Topological entropy of interval maps

If $X = [0,1]$ with the usual Euclidean metric, then there are various shortcuts to compute the entropy of a continuous map $T : [0,1] \to [0,1]$. Let us call any maximal interval on which T is monotone a *lap*; the number of laps is denoted as $\ell(T)$. Also, the *variation* of T is defined as

$$Var(T) = \sup_{0 \leqslant x_0 < \dots\ x_N \leqslant N} \sum_{i=1}^{N} |T(x_i) - T(x_{i-1})|,$$

where the supremum runs over all finite collections of points in $[0,1]$. The following result is due to Misurewicz & Szlenk [12].

Proposition 8. *Let $T : [0,1] \to [0,1]$ have finitely many laps. Then*

$$\begin{aligned} h_{top}(T) &= \lim_{n\to\infty} \frac{1}{n} \log \ell(T^n) \\ &= \limsup_{n\to\infty} \frac{1}{n} \log \#\{\textit{clusters of } n\textit{-periodic points}\} \\ &= \max\{0, \lim_{n\to\infty} \frac{1}{n} \log\, Var(T^n)\}. \end{aligned}$$

where two n-periodic points are in the same cluster if they belong to the same lap of T^n.

Proof. Since the variation of a monotone function is given by $\sup T - \inf T$, and due to the definition of "cluster" of n-periodic points, the inequalities

$$\#\{\text{clusters of } n\text{-periodic points}\}, \mathrm{Var}(T^n) \leqslant \ell(T^n)$$

are immediate. For a lap I of T^n, let $\gamma := |T^n(I)|$ be its *height*. We state without proof (cf. [5, Chapter 9]):

$$\begin{array}{l} \text{For every } \delta > 0, \text{ there is } \gamma > 0 \text{ such that the number of} \\ \text{laps } I \text{ of } T^n \text{ with the property that } I \text{ belongs to a lap of } T^j \\ \text{of height} \geqslant \gamma \text{ for all } 1 \leqslant j \leqslant n \text{ is at least } (1-\delta)^n \ell(T^n). \end{array} \tag{11}$$

This means that $Var(T^n) \geqslant \gamma(1-\delta)^n \ell(T^n)$, and therefore

$$-2\delta + \lim_n \frac{1}{n} \log \ell(T^n) \leqslant \lim_n \frac{1}{n} \log \mathrm{Var}(T^n) \leqslant \lim_n \frac{1}{n} \log \ell(T^n).$$

Since δ is arbitrary, both above quantities are all equal.

Making the further assumption (without proof[3]) that there is $K = K(\gamma)$ such that $\cup_{i=0}^{K} T^i(J) = X$ for every interval of length $|J| \geqslant \gamma$, we also find that

$$\#\{\text{clusters of } n+i\text{-periodic points}, 0 \leqslant i \leqslant K\} \geqslant (1-\delta)^n \ell(T^n).$$

[3] In fact, it is not entirely true if T has an invariant subset attracting an open neighbourhood. But it suffices to restrict T to its *nonwandering set*, that is, the set $\Omega(T) = \{x \in X : x \in \cup_{n\geqslant 1} T^n(U))$ for every neighbourhood $U \ni x\}$, because $h_{top}(T) = h_{top}(T|_{\Omega(T)})$.

This implies that

$$-2\delta + \lim_n \frac{1}{n} \log \ell(T^n) \leqslant \limsup_n \frac{1}{n} \max_{0 \leqslant i \leqslant K} \log \#\{\text{clusters of } n+i\text{-periodic points}\}$$

so also $\lim_n \frac{1}{n} \log \ell(T^n) = \limsup_{n\to\infty} \frac{1}{n} \log \#\{\text{clusters of } n\text{-periodic points}\}$

If $\varepsilon > 0$ is so small that the width of every lap is greater than 2ε, then for every ε-cover $\mathcal{U}$, every subcover of $\mathcal{U}^n$ has at least one element in each lap of T^n. Therefore $\ell(T^n) \leqslant \mathcal{N}(\mathcal{U}^n)$ for every ε-cover, so $\lim_n \frac{1}{n} \log \ell(T^n) \leqslant h_{top}(T)$. □

13.3 Bowen's approach

Let T be map of a compact metric space (X, d). If my eyesight is not so good, I cannot distinguish two points $x, y \in X$ if they are at a distance $d(x, y) < \varepsilon$ from one another. I may still be able to distinguish there orbits, if $d(T^k x, T^k y) > \varepsilon$ for some $k \geqslant 0$. Hence, if I'm willing to wait $n-1$ iterations, I can distinguish x and y if

$$d_n(x, y) := \max\{d(T^k x, T^k y) : 0 \leqslant k < n\} > \varepsilon.$$

If this holds, then x and y are said to be (n, ε)**-separated**. Among all the subsets of X of which all points are mutually (n, ε)-separated, choose one, say $E_n(\varepsilon)$, of maximal cardinality. Then $s_n(\varepsilon) := \#E_n(\varepsilon)$ is the maximal number of n-orbits I can distinguish with ε-poor eyesight.

The **topological entropy** is defined as the limit (as $\varepsilon \to 0$) of the exponential growth-rate of $s_n(\varepsilon)$:

$$h_{top}(T) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon). \tag{12}$$

Note that $s_n(\varepsilon_1) \geqslant s_n(\varepsilon_2)$ if $\varepsilon_1 \leqslant \varepsilon_2$, so $\limsup_n \frac{1}{n} \log s_n(\varepsilon)$ is a decreasing function in ε, and the limit as $\varepsilon \to 0$ indeed exists.

Instead of (n, ε)-separated sets, we can also work with (n, ε)**-spanning** sets, that is, sets that contain, for every $x \in X$, a y such that $d_n(x, y) \leqslant \varepsilon$. Note that, due to its maximality, $E_n(\varepsilon)$ is always (n, ε)-spanning, and no proper subset of $E_n(\varepsilon)$ is (n, ε)-spanning. Each $y \in E_n(\varepsilon)$ must have a point of an $(n, \varepsilon/2)$-spanning set within an $\varepsilon/2$-ball (in d_n-metric) around it, and by the triangle inequality, this $\varepsilon/2$-ball is disjoint from $\varepsilon/2$-ball centred around all other points in $E_n(\varepsilon)$. Therefore, if $r_n(\varepsilon)$ denotes the minimal cardinality among all (n, ε)-spanning sets, then

$$r_n(\varepsilon) \leqslant s_n(\varepsilon) \leqslant r_n(\varepsilon/2). \tag{13}$$

Thus we can equally well define

$$h_{top}(T) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log r_n(\varepsilon). \tag{14}$$

Examples: Consider the β-transformation $T_\beta : [0,) \to [0,1)$, $x \mapsto \beta x \pmod 1$ for some $\beta > 1$. Take $\varepsilon < 1/(2\beta^2)$, and $G_n = \{\frac{k}{\beta^n} \ : \ 0 \leqslant k < \beta^n\}$. Then G_n is (n,ε)-separating, so $s_n(\varepsilon) \geqslant \beta^n$. On the other hand, $G'_n = \{\frac{2k\varepsilon}{\beta^n} \ : \ 0 \leqslant k < \beta^n/(2\varepsilon)\}$ is (n,ε)-spanning, so $r_n(\varepsilon) \leqslant \beta^n/(2\varepsilon)$. Therefore

$$\log \beta = \limsup_n \frac{1}{n} \log \beta^n \leqslant h_{top}(T_\beta) \leqslant \limsup_n \log \beta^n/(2\varepsilon) = \log \beta.$$

Circle rotations, or in general isometries, T have zero topological entropy. Indeed, if $E(\varepsilon)$ is an ε-separated set (or ε-spanning set), it will also be (n,ε)-separated (or (n,ε)-spanning) for every $n \geqslant 1$. Hence $s_n(\varepsilon)$ and $r_n(\varepsilon)$ are bounded in n, and their exponential growth rates are equal to zero.

Finally, let (X,σ) be the full shifts on N symbols. Let $\varepsilon > 0$ be arbitrary, and take m such that $2^{-m} < \varepsilon$. If we select a point from each $n+m$-cylinder, this gives an (n,ε)-spanning set, whereas selecting a point from each n-cylinder gives an (n,ε)-separated set. Therefore

$$\begin{aligned} \log N = \limsup_n \frac{1}{n} \log N^n &\leqslant \limsup_n \frac{1}{n} \log s_n(\varepsilon) \leqslant h_{top}(T_\beta) \\ &\leqslant \limsup_n \frac{1}{n} \log r_n(\varepsilon) \leqslant \limsup_n \log N^{n+m} = \log N. \end{aligned}$$

Proposition 9. *For a continuous map T on a compact metric space (X,d), the three definitions* (10)*,* (12) *and* (14) *give the same outcome.*

Proof. The equality of the limits (12) and (14) follows directly from (13).

If $\mathcal{U}$ is an ε-cover, every $A \in \mathcal{U}^n$ can contain at most one point in an (n,ε)-separated set, so $s(n,\varepsilon) < \mathcal{N}(\mathcal{U}^n)$, whence $\limsup_n \frac{1}{n} \log s(n,\varepsilon) \leqslant \lim_n \frac{1}{n} \log \mathcal{N}(\mathcal{U}^n)$.

Finally, in a compact metric space, every open cover $\mathcal{U}$ has a numbr (called its *Lebesgue number*) such that for every $x \in X$, there is $U \imath \mathcal{U}$ such that $B_\delta(x) \subset U$. Clearly $\delta < \varepsilon$ if $\mathcal{U}$ is an ε-cover. Now if an open ε-cover $\mathcal{U}$ has Lebesgue number δ, and E is an $(n,\delta$-spanning set of cardinality $\#E = r(n,\delta)$, then $X \subset \cup_{x\in E}$ $cap_{i=0}^{n-1} T^{-i}(B_\delta(T^i))$. Since each $B_\delta(T^i(x))$ is contained in some $U \in \mathcal{U}$, we have $\mathcal{N}(\mathcal{U}^n) \leqslant r(n,\delta)$. Since $\delta \to 0$ as $\varepsilon \to 0$, also

$$\lim_{\varepsilon\to 0} \lim_n \frac{1}{n} \log \mathcal{N}(\mathcal{U}^n) \leqslant \lim_{\delta \to 0} \limsup_n \frac{1}{n} \log r(n,\delta).$$

□

13.4 Topological pressure

The topological pressure $P_{top}(T,\psi)$ combines entropy with a potential function $\psi : X \to \mathbb{R}$. By definition, $h_{top}(T) = P_{top}(T,\psi)$ if $\psi(x) \equiv 0$. Denote the n-th ergodic sum of ψ

by

$$S_n\psi(x) = \sum_{k=0}^{n-1} \psi \circ T^k(x).$$

Next set

$$\begin{cases} K_n(T,\psi,\varepsilon) = \sup\{\sum_{x\in E} e^{S_n\psi(x)} \ : \ E \text{ is } (n,\varepsilon)\text{-separated}\}, \\ L_n(T,\psi,\varepsilon) = \inf\{\sum_{x\in E} e^{S_n\psi(x)} \ : \ E \text{ is } (n,\varepsilon)\text{-spanning}\}. \end{cases} \tag{15}$$

For reasonable choices of potentials, the quantities $\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log K_n(T,\psi,\varepsilon)$ and $\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log L_n(T,\psi,\varepsilon)$ are the same, and this quantity is called the **topological pressure**. To give an example of an unreasonable potential, take X_0 be a dense T-invariant subset of X such that $X \setminus X_0$ is also dense. Let

$$\psi(x) = \begin{cases} 100 & \text{if } x \in X_0, \\ 0 & \text{if } x \notin X_0. \end{cases}$$

Then $L_n(T,\psi,\varepsilon) = r_n(\varepsilon)$ whilst $K_n(T,\psi,\varepsilon) = e^{100n}s_n(\varepsilon)$, and their exponential growth rates differ by a factor 100. Hence, some amount of continuity of ψ is necessary to make it work.

Lemma 8. *If $\varepsilon > 0$ is such that $d(x,y) < \varepsilon$ implies that $|\psi(x) - \psi(y)| < \delta/2$, then*

$$e^{-n\delta}K_n(T,\psi,\varepsilon) \leqslant L_n(T,\psi,\varepsilon/2) \leqslant K_n(T,\psi,\varepsilon/2).$$

Exercise 2. *Prove Lemma 8. In fact, the second inequality holds regardless of what ψ is.*

Theorem 13. *If $T : X \to X$ and $\psi : X \to \mathbb{R}$ are continuous on a compact metric space, then the topological pressure is well-defined by*

$$P_{top}(T,\psi) := \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log K_n(T,\psi,\varepsilon) = \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log L_n(T,\psi,\varepsilon).$$

Exercise 3. *Show that $P_{top}(T^R, S_R\psi) = R \cdot P_{top}(T,\psi)$.*

14 Measure-theoretic entropy

Entropy is a measure for the complexity of a dynamical system (X,T). In the previous sections, we related this (or rather topological entropy) to the exponential growth rate of the cardinality of $\mathcal{P}_n = \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ for some partition of the space X. In this section, we look at the measure theoretic entropy $h_\mu(T)$ of an T-invariant measure μ, and this amounts to, instead of just counting $\mathcal{P}_n$, taking a particular weighted sum of the elements $Z_n \in \mathcal{P}_n$. However, if the mass of μ is equally distributed over the all the $Z_n \in$

$\mathcal{P}_n$, then the outcome of this sum is largest; then μ would be the measure of maximal entropy. In "good" systems (X, T) is indeed the supremum over the measure theoretic entropies of all the T-invariant probability measures. This is called the **Variational Principle**:

$$h_{top}(T) = \sup\{h_\mu(T) \ : \ \mu \text{ is } T\text{-invariant probability measure}\}. \tag{16}$$

In this section, we will skip some of the more technical aspect, such as **conditional entropy** (however, see Proposition 10) and σ-algebras (completing a set of partitions), and this means that at some points we cannot give full proofs. Rather than presenting more philosophy what entropy should signify, let us first give the mathematical definition.

Define

$$\varphi : [0,1] \to \mathbb{R} \quad \varphi(x) = -x \log x$$

with $\varphi(0) := \lim_{x \downarrow 0} \varphi(x) = 0$. Clearly $\varphi'(x) = -(1+\log x)$ so $\varphi(x)$ assume its maximum at $1/e$ and $\varphi(1/e) = 1/e$. Also $\varphi''(x) = -1/x < 0$, so that φ is **strictly concave**:

$$\alpha\varphi(x) + \beta\varphi(y) \leqslant \varphi(\alpha x + \beta y) \qquad \text{for all } \alpha + \beta = 1, \alpha, \beta \geqslant 0, \tag{17}$$

with equality if and only if $x = y$.

Theorem 14 (Jensen's Inequality)**.** *For every strictly concave function $\varphi : [0, \infty) \to \mathbb{R}$ we have*

$$\sum_i \alpha_i \varphi(x_i) \leqslant \varphi(\sum_i \alpha_i x_i) \text{ for } \alpha_i > 0, \sum_i \alpha_i = 1 \text{ and } x_i \in [0, \infty), \tag{18}$$

with equality if and only if all the x_i are the same.

Proof. We prove this by induction on n. For $n = 2$ it is simply (17). So assume that (18) holds for some n, and we treat the case $n + 1$. Assume $\alpha_i > 0$ and $\sum_{i=1}^{n+1} \alpha_i = 1$ and write $B = \sum_{i=1}^n \alpha_i$.

$$\begin{aligned}
\varphi(\sum_{i=1}^{n+1} \alpha_i x_i) &= \varphi(B \sum_{i=1}^n \frac{\alpha_i}{B} x_i + \alpha_{n+1} x_{n+1}) \\
&\geqslant B\varphi(\sum_{i=1}^n \frac{\alpha_i}{B} x_i) + \varphi(\alpha_{n+1} x_{n+1}) \qquad \text{by (17)} \\
&\geqslant B \sum_{i=1}^n \frac{\alpha_i}{B} \varphi(x_i) + \varphi(\alpha_{n+1} x_{n+1}) \qquad \text{by (18) for } n \\
&= \sum_{i=1}^{n+1} \alpha_i \varphi(x_i)
\end{aligned}$$

as required. Equality also carries over by induction, because if x_i are all equal for $1 \leqslant i \leqslant n$, (17) only preserves equality if $x_{n+1} = \sum_{i=1}^n \frac{\alpha_i}{B} x_i = x_1$. □

This proof doesn't use the specific form of φ, only its (strict) concavity. Applying it to $\varphi(x) = -x \log x$, we obtain:

Corollary 3. *For $p_1 + \cdots + p_n = 1$, $p_i > 0$, then $\sum_{i=1}^n \varphi(p_i) \leqslant \log n$ with equality if and only if all p_i are equal,* i.e., $p_i \equiv \frac{1}{n}$.

Proof. Take $\alpha_i = \frac{1}{n}$, then by Theorem 14,

$$\frac{1}{n}\sum_{i=1}^n \varphi(p_i) = \sum_{i=1}^n \alpha_i \varphi(p_i) \leqslant \varphi(\sum_{i=1}^n \frac{1}{n} p_i) = \varphi(\frac{1}{n}) = \frac{1}{n}\log n.$$

Now multiply by n. □

Corollary 4. *For real numbers a_i and $p_1 + \cdots + p_n = 1$, $p_i > 0$, $\sum_{i=1}^n p_i(a_i - \log p_i) \leqslant \log \sum_{i=1}^n e^{a_i}$ with equality if and only if $p_i = e^{a_i}/\sum_{i=1}^n e^{a_i}$ for each i.*

Proof. Write $\mathcal{Z} = \sum_{i=1}^n e^{a_i}$. Put $\alpha_i = e^{a_i}/\mathcal{Z}$ and $x_i = p_i\mathcal{Z}/e^{a_i}$ in Theorem 14. Then

$$\begin{aligned}\sum_{i=1}^n p_i(a_i - \log \mathcal{Z} - \log p_i) &= -\sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \left(\frac{p_i\mathcal{Z}}{e^{a_i}} \log \frac{p_i\mathcal{Z}}{e^{a_i}} \right) \\ &\leqslant -\sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \frac{p_i\mathcal{Z}}{e^{a_i}} \log \sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \frac{p_i\mathcal{Z}}{e^{a_i}} = \varphi(1) = 0.\end{aligned}$$

Rearranging gives $\sum_{i=1}^n p_i(a_i - \log p_i) \leqslant \log Z$, with equality only if $x_i = p_i Z/e^{a_i}$ are all the same, *i.e.*, $p_i = e^{a_i}/Z$. □

Exercise 4. *Reprove Corollaries 3 and 4 using Lagrange multipliers.*

Given a finite partition $\mathcal{P}$ of a probability space (X, μ), let

$$H_\mu(\mathcal{P}) = \sum_{P\in\mathcal{P}} \varphi(\mu(P)) = -\sum_{P\in\mathcal{P}} \mu(P)\log(\mu(P)), \tag{19}$$

where we can ignore the partition elements with $\mu(P) = 0$ because $\varphi(0) = 0$. For a T-invariant probability measure μ on $(X, \mathcal{B}, T)$, and a partition $\mathcal{P}$, define the **entropy of μ w.r.t. $\mathcal{P}$** as

$$H_\mu(T, \mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}). \tag{20}$$

Finally, the **measure theoretic entropy** of μ is

$$h_\mu(T) = \sup\{H_\mu(T, \mathcal{P}) \ : \ \mathcal{P} \text{ is a finite partition of } X\}. \tag{21}$$

Naturally, this raises the questions:

Does the limit exist in (20)?
How can one possibly consider **all** partitions of X?

We come to this later; first we want to argue that entropy is a characteristic of a measure preserving system. That is, two measure preserving systems $(X, \mathcal{B}, T, \mu)$ and $(Y, \mathcal{C}, S, \nu)$ that are **isomorphic**, *i.e.,* there are full-measured sets $X_0 \subset X$, $Y_0 \subset Y$ and a bi-measurable invertible measure-preserving map $\pi : X_0 \to Y_0$ (called **isomorphism**) such that the diagram

$$\begin{array}{ccc} (X_0, \mathcal{B}, \mu) & \xrightarrow{T} & (X_0, \mathcal{B}, \mu) \\ \pi \downarrow & & \downarrow \pi \\ (Y_0, \mathcal{C}, \nu) & \xrightarrow{S} & (Y_0, \mathcal{C}, \nu) \end{array}$$

commutes, then $h_\mu(T) = h_\nu(S)$. This holds, because the bi-measurable measure-preserving map π preserves all the quantities involved in (19)-(21), including the class of partitions for both systems.

A major class of systems where this is very important are the Bernoulli shifts. These are the standard probability space to measure a sequence of i.i.d. events each with outcomes in $\{0, \dots, N-1\}$ with probabilities $p_0, \dots, p_{N-1}$ respectively. That is: $X = \{0, \dots, N-1\}^{\mathbb{N}_0}$ or $\{0, \dots, N-1\}^{\mathbb{Z}}$, σ is the left-shift, and μ the Bernoulli measure that assigns to every cylinder set $[x_m \dots x_n]$ the mass

$$\mu([x_m \dots x_n]) = \prod_{k=m}^{n} \rho(x_k) \qquad \text{where } \rho(x_k) = p_i \text{ if } x_k = i.$$

For such a Bernoulli shift, the entropy is

$$h_\mu(\sigma) = -\sum_i p_i \log p_i, \tag{22}$$

so two Bernoulli shifts (X, p, μ_p) and $(X', p', \mu_{p'})$ can only be isomorphic if $-\sum_i p_i \log p_i = -\sum_i p'_i \log(p'_i)$. The famous theorem of Ornstein showed that entropy is a complete invariant for Bernoulli shifts:

Theorem 15 (Ornstein 1974 [14], cf. page 105 of [16])**.** *Two Bernoulli shifts (X, p, μ_p) and $(X', p', \mu_{p'})$ are isomorphic if and only if $-\sum_i p_i \log p_i = -\sum_i p'_i \log p'_i$.*

Exercise 5. *Conclude that the Bernoulli shift $\mu_{(\frac14,\frac14,\frac14,\frac14)}$ is isomorphic to $\mu_{(\frac18,\frac18,\frac18,\frac18,\frac12)}$, but that no Bernoulli measure on four symbols can be isomorhic to $\mu_{(\frac15,\frac15,\frac15,\frac15,\frac15)}$*

Let us go back to the definition of entropy, and try to answer the outstanding questions.

Definition 12. *We call a real sequence $(a_n)_{n\geqslant 1}$* **subadditive** *if*

$$a_{m+n} \leqslant a_m + a_n \quad \textit{for all } m, n \in \mathbb{N}.$$

Theorem 16. *If* $(a_n)_{n\geqslant 1}$ *is subadditive, then* $\lim_n \frac{a_n}{n} = \inf_{r\geqslant 1} \frac{a_r}{r}$.

Proof. Every integer n can be written uniquely as $n = i\cdot r + j$ for $0 \leqslant j < r$. Therefore

$$\limsup_{n\to\infty} \frac{a_n}{n} = \limsup_{i\to\infty} \frac{a_{i\cdot r+j}}{i\cdot r+j} \leqslant \limsup_{i\to\infty} \frac{ia_r + a_j}{i\cdot r+j} = \frac{a_r}{r}.$$

This holds for all $r \in \mathbb{N}$, so we obtain

$$\inf_r \frac{a_r}{r} \leqslant \liminf_n \frac{a_n}{n} \leqslant \limsup_n \frac{a_n}{n} \leqslant \inf_r \frac{a_r}{r},$$

as required. □

Definition 13. *Motivated by the conditional measure* $\mu(P|Q) = \frac{\mu(P\cap Q)}{\mu(Q)}$*, we define* **conditional entropy** *of a measure* μ *as*

$$H_\mu(\mathcal{P}|\mathcal{Q}) = -\sum_j \mu(Q_j) \sum_i \frac{\mu(P_i\cap Q_j)}{\mu(Q_j)} \log \frac{\mu(P_i\cap Q_j)}{\mu(Q_j)}, \tag{23}$$

where i *runs over all elements* $P_i \in \mathcal{P}$ *and* j *runs over all elements* $Q_j \in \mathcal{Q}$.

Avoiding philosophical discussions how to interpret this notion, we just list the main properties that are needed in this course that rely of condition entropy:

Proposition 10. *Given measures* μ, μ_i *and two partitions* $\mathcal{P}$ *and* $\mathcal{Q}$*, we have*

1. $H_\mu(\mathcal{P} \vee \mathcal{Q}) \leqslant H_\mu(\mathcal{P}) + H_\mu(\mathcal{Q})$;
2. $H_\mu(T, \mathcal{P}) \leqslant H_\mu(T, \mathcal{Q}) + H_\mu(\mathcal{P} \mid \mathcal{Q})$.
3. $\sum_{i=1}^n p_i H_{\mu_i}(\mathcal{P}) \leqslant H_{\sum_{i=1}^n p_i\mu_i}(\mathcal{P})$ *whenever* $\sum_{i=1}^n p_1 = 1$, $p_i \geqslant 0$,

Subadditivity is the key to the convergence in (20). Call $a_n = H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$. Then

$$\begin{aligned}
a_{m+n} &= H_\mu(\bigvee_{k=0}^{m+n-1} T^{-k}\mathcal{P}) && \text{use Proposition 10, part 1.}\\
&\leqslant H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=m}^{m+n-1} T^{-k}\mathcal{P}) && \text{use } T\text{-invariance of } \mu\\
&= H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})\\
&= a_m + a_n.
\end{aligned}$$

Therefore $H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ is subadditive, and the existence of the limit of $\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ follows.

Proposition 11. *Entropy has the following properties:*

1. *The identity map has entropy* 0*;*
2. $h_\mu(T^R) = R \cdot h_\mu(T)$ *and for invertible systems* $h_\mu(T^{-R}) = R \cdot h_\mu(T)$*.*

Proof. Statement 1. follows simply because $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P} = \mathcal{P}$ if T is the identity map, so the cardinality of $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ doesn't increase with n.

For statement 2. set $\mathcal{Q} = \bigvee_{j=0}^{R-1} T^{-j}\mathcal{P}$. Then for $k \geqslant 1$,

$$\begin{aligned} R \cdot H_\mu(T, \mathcal{P}) &= \lim_{n\to\infty} R \cdot \frac{1}{nR} H_\mu(\bigvee_{j=0}^{nR-1} T^{-k}\mathcal{P}) \\ &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\bigvee_{j=0}^{n-1} (T^R)^{-j}\mathcal{Q}) \\ &= H_\mu(T^R, \mathcal{Q}). \end{aligned}$$

Taking the supremum over all $\mathcal{P}$ or $\mathcal{Q}$ has the same effect. □

The next theorem is the key to really computing entropy, as it shows that a single well-chosen partition $\mathcal{P}$ suffices to compute the entropy as $h_\mu(T) = H_\mu(T, \mathcal{P})$.

Theorem 17 (Kolmogorov-Sinaĭ)**.** *Let* $(X, \mathcal{B}, T, \mu)$ *be a measure-preserving dynamical system. If partition* $\mathcal{P}$ *is such that*

$$\begin{cases} \bigvee_{j=0}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is non-invertible,} \\ \bigvee_{j=-\infty}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is invertible,} \end{cases}$$

then $h_\mu(T) = H_\mu(T, \mathcal{P})$*.*

We haven't explained properly what "generates $\mathcal{B}$ means, but the idea you should have in mind is that (up to measure 0), every two points in X should be in different elements of $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ (if T is non-invertible), or of $\bigvee_{k=-n}^{n-1} T^{-k}\mathcal{P}$ (if T is invertible) for some sufficiently large n. The partition $\mathcal{B} = \{X\}$ fails miserably here, because $\bigvee_{j=-n}^{n} T^{-k}\mathcal{P} = \mathcal{P}$ for all n and no two points are ever separated in $\mathcal{P}$. A more subtle example can be created for the doubling map $T_2 : \mathbb{S}^1 \to \mathbb{S}^1$, $T_2(x) = 2x \pmod 1$. The partition $\mathcal{P} = \{[0, \frac{1}{2}), [\frac{1}{2}, 1)\}$. is separating every two points, because if $x \neq y$, say $2^{-(n+1)} < |x - y| \leqslant 2^{-n}$, then there is $k \leqslant n$ such that $T_2^k x$ and $T_2^k y$ belong to different partition elements.

On the other hand, $\mathcal{Q} = \{[\frac{1}{4}, \frac{3}{4}), [0, \frac{1}{4}) \cup [\frac{3}{4}, 1)\}$ does **not** separate points. Indeed, if $y = 1 - x$, then $T_2^k(y) = 1 - T_2^k(x)$ for all $k \geqslant 0$, so x and y belong to the same partition element, $T_2^k(y)$ and $T_2^k(x)$ will also belong to the same partition element!

In this case, $\mathcal{P}$ can be used to compute $h_\mu(T)$, while $\mathcal{Q}$ in principle cannot (although here, for all Bernoulli measure $\mu = \mu_{p,1-p}$, we have $h_\mu(T_2) = H_\mu(T, \mathcal{P}) = H_\mu(T, \mathcal{Q})$).

Proof of Theorem 17. Let $\mathcal{A}$ be the generating partition. Then $h_\mu(T, \mathcal{A}) \leqslant h_\mu(T)$ because the RHS is the supremum over all partitions. Let $\varepsilon > 0$ be arbitrary, and let $\mathcal{C}$ be a finite partition, say $\#\mathcal{C} = N$, such that $H_\mu(T, \mathcal{C}) \geqslant h_\mu(T) - \varepsilon$. We have

$$\begin{aligned} h_\mu(T|\bigvee_{i=-k}^{k} T^i\mathcal{A}) &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\bigvee_{j=0}^{n-1} T^{-j}(\bigvee_{i=-k}^{k} T^i\mathcal{A})) = \lim_{n\to\infty} \frac{1}{n} H_\mu(T^k(\bigvee_{j=0}^{n+2k} T^{-j}\mathcal{A})) \\ &= \lim_{n\to\infty} \frac{n+2k+1}{n} \frac{1}{n+2k+1} H_\mu(\bigvee_{j=0}^{n+2k} T^{-j}\mathcal{A}) = h_\mu(T\mathcal{A}). \end{aligned}$$

Using this and Proposition 10, part 2, we compute:

$$h_\mu(T|\mathcal{C}) \leqslant h_\mu(T|\bigvee_{i=-k}^{k} T^i\mathcal{A}) + H_\mu(\mathcal{C}|\bigvee_{i=-k}^{k} T^i\mathcal{A}) = h_\mu(T|\mathcal{A}) + H_\mu(\mathcal{C}|\bigvee_{i=-k}^{k} T^i\mathcal{A}).$$

Since $\mathcal{A}$ is generating, the measure of points in X for which the element $A \in \bigvee_{i=-k}^{k} T^ik\mathcal{A}$ that contains x is itself contained in a single element of $\mathcal{C}$ tends to one as $k \to \infty$. Therefore we can find k so large that if we set

$$\mathcal{A}^* = \{A \in \bigvee_{i=-k}^{k} T^ik\mathcal{A} : A \cap C \neq \varnothing \neq A \cap C' \text{ for some } C \neq C' \in \mathcal{C}\},$$

then $\mu(\bigcup_{A\in\mathcal{A}^*} A) \leqslant \varepsilon/(N \log N)$. This gives

$$\begin{aligned} H_\mu(\mathcal{C}|\bigvee_{i=-k}^{k} T^ik\mathcal{A}) &= \sum_{A\in\bigvee_{i=-k}^{k} T^ik\mathcal{A}} \sum_{C\in\mathcal{C}} \mu(A)\varphi\left(\frac{\mu(A\cap C)}{\mu(A)}\right) \\ &= \sum_{A\in\mathcal{A}^*} \mu(A) \sum_{C\in\mathcal{C}} \varphi\left(\frac{\mu(A\cap C)}{\mu(A)}\right) \leqslant \sum_{A\in\mathcal{A}^*} \mu(A) N \log N < \varepsilon. \end{aligned}$$

Combining all, we get $h_\mu(T, \mathcal{A}) \geqslant h_\mu(T, \mathcal{C}) - \varepsilon \geqslant h_\mu(T) - 2\varepsilon$ completing the proof. $\square$

We finish this section with computing the entropy for a Bernoulli shift on two symbols, *i.e.,* we will prove (22) for two-letter alphabets and any probability $\mu([0]) =: p \in [0, 1]$. The space is thus $X = \{0, 1\}^{\mathbb{N}_0}$ and each $x \in X$ represents an infinite sequence of coin-flips with an unfair coin that gives head probability p (if head has the symbol 0). Recall from probability theory

$$\mathbb{P}(k \text{ heads in } n \text{ flips}) = \binom{n}{k} p^k (1-p)^{n-k},$$

so by full probability:

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = 1.$$

Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are the binomial coefficients, and we can compute

$$\begin{cases} k\binom{n}{k} = \frac{n!}{(k-1)!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1} \\ (n-k)\binom{n}{k} = \frac{n!}{(k)!(n-k-1)!} = n\frac{(n-1)!}{k!(n-k-1)!} = n\binom{n-1}{k} \end{cases} \tag{24}$$

This gives all the ingredients necessary for the computation.

$$\begin{aligned}
H_\mu(\bigvee_{k=0}^{n-1}\sigma^{-k}\mathcal{P}) &= -\sum_{x_0,\dots,x_{n-1}=0}^{1} \mu([x_0,\dots,x_{n-1}])\log\mu([x_0,\dots,x_{n-1}]) \\
&= -\sum_{x_0,\dots,x_{n-1}=0}^{1}\prod_{j=0}^{n-1}\rho(x_j)\log\prod_{j=0}^{n-1}\rho(x_j) \\
&= -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\log\left(p^k(1-p)^{n-k}\right) \\
&= -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}k\log p \\
&\qquad -\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}(n-k)\log(1-p)
\end{aligned}$$

In the first sum, the term $k = 0$ gives zero, as does the term $k = n$ for the second sum. Thus we leave out these terms and rearrange by (24):

$$\begin{aligned}
&= -p\log p\sum_{k=1}^{n}k\binom{n-1}{k}p^{k-1}(1-p)^{n-k} \\
&\qquad -(1-p)\log(1-p)\sum_{k=0}^{n-1}(n-k)\binom{n}{k}p^k(1-p)^{n-k-1} \\
&= -p\log p\sum_{k=1}^{n}n\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} \\
&\qquad -(1-p)\log(1-p)\sum_{k=0}^{n-1}n\binom{n-1}{k}p^k(1-p)^{n-k-1} \\
&= n\left(-p\log p-(1-p)\log(1-p)\right).
\end{aligned}$$

The partition $\mathcal{P} = \{[0],[1]\}$ is generating, so by Theorem 17,

$$h_\mu(\sigma) = H_\mu(\sigma,\mathcal{P}) = \lim_n\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1}\sigma^{-k}\mathcal{P}) = -p\log p-(1-p)\log(1-p)$$

as required.

15 The Variational Principle

The Variational Principle claims that topological entropy (or pressure) is achieved by taking the supremum of the measure-theoretic entropies over all invariant probability measures. But in the course of these notes, topological entropy has seen various definitions. Even $\sup\{h_\mu(T) : \mu \text{ is a } T\text{-invariant probability measure}\}$ is sometimes used as definition of topological entropy. So it is high time to be more definite.

We will do this by immediately passing to topological pressure, which we will base on the definition in terms of (n,δ)-spanning sets and/or (n,ε)-separated sets. Topological entropy then simply emerges as $h_{top}(T) = P_{top}(T, 0)$.

Theorem 18 (The Variational Principle)**.** *Let (X, d) be a compact metric space, $T : X \to X$ a continuous map and $\psi : X \to \mathbb{R}$ as continuous potential. Then*

$$P_{top}(T,\psi) = \sup\{h_\mu(T) + \int_X \psi \, d\mu \ : \ \mu \textit{ is a } T\textit{-invariant probability measure}\}. \quad (25)$$

Remark 3. *By the ergodic decomposition, every T-invariant probability measure can be written as convex combination (sometimes in the form of an integral) of ergodic T-invariant probability measures. Therefore, it suffices to take the supremum over all ergodic T-invariant probability measures.*

Proof. First we show that for every T-invariant probability measure, $h_\mu(T)+\int_X \psi \, d\mu \leqslant P_{top}(T,\psi)$. Let $\mathcal{P} = \{P_0, \dots, P_{N-1}\}$ be an arbitrary partition with $N \geqslant 2$ (if $\mathcal{P} = \{X\}$, then $h_\mu(T,\mathcal{P}) = 0$ and there is not much to prove). Let $\eta > 0$ be arbitrary, and choose $\varepsilon > 0$ so that $\varepsilon N \log N < \eta$.

By "regularity of μ", there are compact sets $Q_i \subset P_i$ such that $\mu(P_i \setminus Q_i) < \varepsilon$ for each $0 \leqslant i < N$. Take $Q_N = X \setminus \cup_{i=0}^{N-1} Q_i$. Then $\mathcal{Q} = \{Q_0, \dots, Q_N\}$ is a new partition of X, with $\mu(Q_N) \leqslant N\varepsilon$. Furthermore

$$\frac{\mu(P_i \cap Q_j)}{\mu(Q_j)} = \begin{cases} 0 & \text{if } i \neq j < N, \\ 1 & \text{if } i = j < N. \end{cases}$$

whereas $\sum_{i=0}^{N-1} \frac{\mu(P_i \cap Q_N)}{\mu(Q_N)} = 1$. Therefore the conditional entropy

$$\begin{aligned} H_\mu(\mathcal{P}|\mathcal{Q}) &= \sum_{j=0}^{N} \sum_{i=0}^{N-1} \mu(Q_j) \underbrace{\varphi\left(\frac{\mu(P_i \cap Q_j)}{\mu(Q_j)}\right)}_{=\ 0 \text{ if } j<N} \\ &= -\mu(Q_N) \sum_{i=0}^{N-1} \frac{\mu(P_i \cap Q_N)}{\mu(Q_N)} \log(\frac{\mu(P_i \cap Q_N)}{\mu(Q_N)}) \\ &\leqslant \mu(Q_N) \log N \qquad \text{by Corollary 3} \\ &\leqslant \varepsilon N \log N < \eta. \end{aligned}$$

Choose $0 < \delta < \frac{1}{2} \min_{0 \leqslant i < j < N} d(Q_i, Q_j)$ so that

$$d(x,y) < \delta \text{ implies } |\psi(x) - \psi(y)| < \varepsilon. \tag{26}$$

Here we use uniform continuity of ψ on the compact space X. Fix n and let $E_n(\delta)$ be an (n,δ)-spanning set. For $Z \in \mathcal{Q}_n := \bigvee_{k=0}^{n-1} T^{-k}\mathcal{Q}$, let $\alpha(Z) = \sup\{S_n\psi(x) \ : \ x \in Z\}$. For each such Z, also choose $x_Z \in \overline{Z}$ such that $S_n\psi(x) = \alpha(Z)$ (again we use continuity of ψ here), and $y_Z \in E_n(\delta)$ such that $d_n(x_Z, y_Z) < \delta$. Hence

$$\alpha(Z) - n\varepsilon \leqslant S_n\psi(y_Z) \leqslant \alpha(Z) + n\varepsilon.$$

This gives

$$H_\mu(\mathcal{Q}_n) + \int_X S_n\psi \ d\mu \leqslant \sum_{Z \in \mathcal{Q}_n} \mu(Z)(\alpha(Z) - \log \mu(Z)) \leqslant \log \sum_{Z \in \mathcal{Q}_n} e^{\alpha(Z)} \tag{27}$$

by Corollary 4.

Each δ-ball intersects the closure of at most two elements of $\mathcal{Q}$. Hence, for each $y \in E_n(\delta)$, the cardinality $\#\{Z \in \mathcal{Q}_n : y_Z = y\} \leqslant 2^n$. Therefore

$$\sum_{Z \in \mathcal{Q}_n} e^{\alpha(Z) - n\varepsilon} \leqslant \sum_{Z \in \mathcal{Q}_n} e^{S_n\psi(y_Z)} \leqslant 2^n \sum_{y \in E_n(\delta)} e^{S_n\psi(y)}.$$

Take the logarithm and rearrange to

$$\log \sum_{Z \in \mathcal{Q}_n} e^{\alpha(Z)} \leqslant n(\varepsilon + \log 2) + \log \sum_{y \in E_n(\delta)} e^{S_n\varphi(y)}.$$

By T-invariance of μ we have $\int S_n\psi \ d\mu = n \int \psi \ d\mu$. Therefore

$$\begin{aligned} \frac{1}{n} H_\mu(\mathcal{Q}_n) + \int_X \psi \ d\mu &\leqslant \frac{1}{n} H_\mu(\mathcal{Q}_n) + \frac{1}{n}\int_X S_n\psi \ d\mu \\ &\leqslant \frac{1}{n} \log \sum_{Z \in \mathcal{Q}_n} e^{\alpha(Z)} \\ &\leqslant \varepsilon + \log 2 + \frac{1}{n} \log \sum_{y \in E_n(\delta)} e^{S_n\varphi(y)}. \end{aligned}$$

Taking the limit $n \to \infty$ and recalling that $E_n(\delta)$ is an arbitrary (n,δ)-spanning set, gives

$$H_\mu(T, \mathcal{Q}) + \int_X \psi \ d\mu \leqslant \varepsilon + \log 2 + P_{top}(T, \psi).$$

By Proposition 10, part 2., and recalling that $\varepsilon < \eta$, we get

$$H_\mu(T, \mathcal{P}) + \int_X \psi \ d\mu = H_\mu(T, \mathcal{Q}) + H_\mu(\mathcal{P}|\mathcal{Q}) + \int_X \psi \ d\mu \leqslant 2\eta + \log 2 + P_{top}(T, \psi).$$

We can apply the same reasoning to T^R and $S_R\psi$ instead of T and ψ. This gives

$$\begin{aligned} R \cdot \left(H_\mu(T, \mathcal{P}) + \int_X \psi \, d\mu \right) &= H_\mu(T^R, \mathcal{P}) + \int_X S_R\psi \, d\mu \\ &\leqslant 2\eta + \log 2 + P_{top}(T^R, S_R\psi) \\ &= 2\eta + \log 2 + R \cdot P_{top}(T, \psi). \end{aligned}$$

Divide by R and take $R \to \infty$ to find $H_\mu(T, \mathcal{P}) + \int_X \psi \, d\mu \leqslant P_{top}(T, \psi)$. Finally take the supremum over all partitions $\mathcal{P}$.

Now the other direction, we will work with (n, ε)-separated sets. After choosing $\varepsilon > 0$ arbitrary, we need to find a T-invariant probability measure μ such that

$$h_\mu(T) + \int_X \psi \, d\mu \geqslant \limsup_{n\to\infty} \frac{1}{n} \log K_n(T, \psi, \varepsilon) := P(T, \psi, \varepsilon).$$

Let $E_n(\varepsilon)$ be an (n, ε)-separated set such that

$$\log \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} \geqslant \log K_n(T, \psi, \varepsilon) - 1. \tag{28}$$

Define Δ_n as weighted sum of Dirac measures:

$$\Delta_n = \frac{1}{\mathcal{Z}} \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} \delta_y,$$

where $\mathcal{Z} = \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)}$ is the normalising constant. Take a new probability measure

$$\mu_n = \frac{1}{n} \sum_{k=0}^{n-1} \Delta_n \circ T^{-k}.$$

Therefore

$$\begin{aligned} \int_X \psi \, d\mu_n &= \frac{1}{n} \sum_{k=0}^{n-1} \int_X \psi \, d(\Delta_n \circ T^{-k}) = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{y \in E_n(\varepsilon)} \psi \circ T^k(y) \frac{1}{\mathcal{Z}} e^{S_n\psi(y)} \\ &= \frac{1}{n} \sum_{y \in E_n(\varepsilon)} S_n\psi(y) \frac{1}{\mathcal{Z}} e^{S_n\psi(y)} = \frac{1}{n} \int_X S_n\psi \, d\Delta_n. \end{aligned} \tag{29}$$

Since the space of probability measures on X is compact in the weak topology, we can find a sequence $(n_j)_{j \geqslant 1}$ such that for every continuous function $f : X \to \mathbb{R}$

$$\int_X f \, d\mu_{n_j} \to \int_X f \, d\mu \quad \text{as } j \to \infty.$$

Choose a partition $\mathcal{P} = \{P_0, \dots, P_{N-1}\}$ with $\mathrm{diam}(P_i) < \varepsilon$ and $\mu(\partial P_i) = 0$ for all $0 \leqslant i < N$. Since $Z \in \mathcal{P}_n := \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ contains at most one element of an (n, ε)-separated set, we have

$$\begin{aligned} H_{\Delta_n}(\mathcal{P}_n) + \int_X S_n\psi \; d\Delta_n &= \sum_{y \in E_n(\varepsilon)} \Delta_n(\{y\}) \left(S_n\psi(y) - \log \Delta_n(\{y\})\right) \\ &= \log \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} = \log \mathcal{Z}. \end{aligned}$$

by Corollary 4

Take $0 < q < n$ arbitrary, and for $0 \leqslant j < q$, let

$$U_j = \{j, j+1, \dots, a_j q + j - 1\} \qquad \text{where } a_j = \lfloor \frac{n-j}{q} \rfloor.$$

Then

$$\{0, 1, \dots, n-1\} = U_j \cup \underbrace{\{0, 1, \dots, j-1\} \cup a_j q + j, a_j q + j + 1, \dots, n-1\}}_{V_j}$$

where V_j has at most $2q$ elements. We split

$$\begin{aligned} \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P} &= \left(\bigvee_{r=0}^{a_j-1} \bigvee_{i=0}^{q-1} T^{-(rq+j+i)}\mathcal{P} \right) \vee \bigvee_{l \in V_j} T^{-l}\mathcal{P} \\ &= \left(\bigvee_{r=0}^{a_j-1} T^{-(rq+j)} \bigvee_{i=0}^{q-1} T^{-i}\mathcal{P} \right) \vee \bigvee_{l \in V_j} T^{-l}\mathcal{P}. \end{aligned}$$

Therefore,

$$\begin{aligned} \log \mathcal{Z} &= H_{\Delta_n}(\mathcal{P}_n) + \int_X S_n\psi \; d\Delta_n \\ &\leqslant \sum_{r=0}^{a_j-1} H_{\Delta_n}(T^{-(rq+j)} \bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + H_{\Delta_n}(\bigvee_{l \in V_j} T^{-l}\mathcal{P}) + \int_X S_n\psi \; d\Delta_n \\ &\leqslant \sum_{r=0}^{a_j-1} H_{\Delta_n \circ T^{-(rq+j)}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q \log N + \int_X S_n\psi \; d\Delta_n, \end{aligned}$$

because $\bigvee_{l \in V_j} T^{-l}\mathcal{P}$ has at most N^{2q} elements and using Corollary 3. Summing the above inequality over $j = 0, \dots, q-1$, gives

$$\begin{aligned} q \log \mathcal{Z} &= \sum_{j=0}^{q-1} \sum_{r=0}^{a_j-1} H_{\Delta_n \circ T^{-rq+j}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n \\ &\leqslant n \sum_{k=0}^{n-1} \frac{1}{n} H_{\Delta_n \circ T^{-k}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n. \end{aligned}$$

Proposition 10, part 3., allows us to swap the weighted average and the operation H:

$$q \log \mathcal{Z} \leqslant nH_{\mu_n}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n.$$

Dividing by n and recalling (28) for the left hand side, and (29) to replace Δ_n by μ_n, we find

$$\frac{q}{n} \log K_n(T,\psi,\varepsilon) - \frac{q}{n} \leqslant H_{\mu_n}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + \frac{2q^2}{n} \log N + q \int_X \psi \; d\mu_n.$$

Because $\mu(\partial P_i) = 0$ for all i, we can replace n by n_j and take the weak limit as $j \to \infty$. This gives

$$qP(T,\psi,\varepsilon) \leqslant H_\mu(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + q \int_X \psi \; d\mu.$$

Finally divide by q and let $q \to \infty$:

$$P(T,\psi,\varepsilon) \leqslant h_\mu(T) + \int_X \psi \; d\mu.$$

This conclues the proof. □

16 Measures of maximal entropy

For the full shift (Ω, σ) with $\Omega = \{0, \dots, N-1\}^{\mathbb{N}_0}$ or $\Omega = \{0, \dots, N-1\}^{\mathbb{Z}}$, we have $h_{top}(\sigma) = \log N$, and the $(\frac{1}{N}, \dots, \frac{1}{N})$-Bernoulli measure μ indeed achieves this maximum: $h_\mu(\sigma) = h_{top}(\sigma)$. Hence μ is a (and in this case unique) **measure of maximal entropy**. The intuition to have here is that for a measure to achieve maximal entropy, it should distribute its mass as evenly over the space as possible. But how does this work for subshifts, where it is not immediately obvious how to distribute mass evenly?

For subshifts of finite type, Parry [15] demonstrated how to construct the measure of maximal entropy, which is now called after him. Let (Σ_A, σ) be a subshift of finite type on alphabet $\{0, \dots, N-1\}$ with transition matrix $A = (a_{i,j})_{i,j=0}^{N-1}$, so $x = (x_n) \in \Sigma_n$ if and only if $a_{x_n, x_{n+1}} = 1$ for all n. Let us assume that A is aperiodic and irreducible. Then by the Perron-Frobenius Theorem for matrices, there is a unique real eigenvalue, of multiplicity one, which is larger in absolute value than every other eigenvalue, and $h_{top}(\sigma) = \log \lambda$. Furthermore, by irreducibility of A, the left and right eigenvectors $u = (u_0, \dots, u_{N-1})$ and $v = (v_0, \dots, v_{N-1})^T$ associated to λ are unique up to a multiplicative factor, and they can be chosen to be strictly positive. We will scale them such that

$$\sum_{i=0}^{N-1} u_i v_i = 1.$$

Now define the **Parry measure** by

$$\begin{aligned} p_i &:= u_i v_i = \mu([i]), \\ p_{i,j} &:= \frac{a_{i,j} v_j}{\lambda v_i} = \mu([ij] \mid [i]), \end{aligned}$$

so $p_{i,j}$ indicates the conditional probability that $x_{n+1} = j$ knowing that $x_n = i$. Therefore $\mu([ij]) = \mu([i])\mu([ij] \mid [i]) = p_i p_{i,j}$. It is stationary (*i.e.*, shift-invariant) but not quite a product measure, but $\mu([i_m \dots i_n]) = p_{i_m} \cdot p_{i_m, i_{m+1}} \cdots p_{i_{n-1}, i_n}$.

Theorem 19. *The Parry measure μ is the unique measure of maximal entropy for a subshift of finite type with irreducible transition matrix.*

Proof. In this proof, we will only show that $h_\mu(\sigma) = h_{top}(\sigma) = \log \lambda$, and skip the (more complicated) uniqueness part.

The definitions of mass of 1-cylinders and 2-cylinders are compatible, because (since v is a right eigenvector)

$$\sum_{j=0}^{N-1} \mu([ij]) = \sum_{j=0}^{N-1} p_i p_{i,j} = p_i \sum_{j=0}^{N-1} \frac{a_{i,j} v_j}{\lambda v_i} = p_i \frac{\lambda v_i}{\lambda v_i} = p_i = \mu([i]).$$

Summing over i, we get $\sum_{i=0}^{N-1} \mu([i]) = \sum_{i=0}^{N-1} u_i v_i = 1$, due to our scaling.

To show that μ is shift-invariant, we take any cylinder set $Z = [i_m \dots i_n]$ and compute

$$\begin{aligned} \mu(\sigma^{-1} Z) &= \sum_{i=0}^{N-1} \mu([i i_m \dots i_n]) = \sum_{i=0}^{N-1} \frac{p_i p_{i,i_m}}{p_{i_m}} \mu([i_m \dots i_n]) \\ &= \mu([i_m \dots i_n]) \sum_{i=0}^{N-1} \frac{u_i v_i \ a_{i,i_m} v_{i_m}}{\lambda v_i \ u_{i_m} v_{i_m}} \\ &= \mu(Z) \sum_{i=0}^{N-1} \frac{u_i a_{i,i_m}}{\lambda u_{i_m}} = \mu(Z) \frac{\lambda u_{i_m}}{\lambda u_{i_m}} = \mu(Z). \end{aligned}$$

This invariance carries over to all sets in the σ-algebra $\mathcal{B}$ generated by the cylinder sets.

Based on the interpretation of conditional probabilities, the identity

$$\sum_{\substack{i_{m+1}, \dots, i_{n-1} = 0 \\ a_{i_k, i_{k+1}} = 1}}^{N-1} p_{i_m} p_{i_m, i_{m+1}} \cdots p_{i_{n-1}, i_n} = p_{i_m} \tag{30}$$

follows because the left hand side indicates the total probability of starting in state i_m and reach some state after $n - m$ steps.

To compute $h_\mu(\sigma)$, we will confine ourselves to the partition $\mathcal{P}$ of 1-cylinder sets; this partition is generating, so this restriction is justified by Theorem 17.

$$\begin{aligned}
H_\mu(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) &= -\sum_{\substack{i_0,\dots,i_{n-1}=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} \mu([i_0\dots i_{n-1}]) \log \mu([i_0\dots i_{n-1}]) \\
&= -\sum_{\substack{i_0,\dots,i_{n-1}=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} p_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n} \left(\log p_{i_0} + \log p_{i_0,i_1} + \cdots + \log p_{i_{n-2},i_{n-1}}\right) \\
&= -\sum_{i_0=0}^{N-1} p_{i_0} \log p_{i_0} - (n-1) \sum_{i,j=0}^{N-1} p_i p_{i,j} \log p_{i,j},
\end{aligned}$$

by (30) used repeatedly. Hence

$$\begin{aligned}
h_\mu(\sigma) &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) \\
&= -\sum_{i,j=0}^{N-1} p_i p_{i,j} \log p_{i,j} \\
&= -\sum_{i,j=0}^{N-1} \frac{u_i a_{i,j} v_j}{\lambda} \left(\log a_{i,j} + \log v_j - \log v_i - \log \lambda\right).
\end{aligned}$$

The first term in the brackets is zero because $a_{i,j} \in \{0,1\}$. The second term (summing first over i) simplifies to

$$-\sum_{j=0}^{N-1} \frac{\lambda u_j v_j}{\lambda} \log v_j = -\sum_{j=0}^{N-1} u_j v_j \log v_j,$$

whereas the third term (summing first over j) simplifies to

$$\sum_{i=0}^{N-1} \frac{u_i \lambda v_i}{\lambda} \log v_i = \sum_{i=0}^{N-1} u_i v_i \log v_i.$$

Hence these two terms cancel each other. The remaining term is

$$\sum_{i,j=0}^{N-1} \frac{u_i a_{i,j} v_j}{\lambda} \log \lambda = \sum_{i=0}^{N-1} \frac{u_i \lambda v_i}{\lambda} \log \lambda = \sum_{i=0}^{N-1} u_i v_i \log \lambda = \log \lambda.$$

$\square$

Remark 4. *There are systems without maximising measure, for example among the "shifts of finite type" on infinite alphabets. To give an example (without proof, see [7]), if* $\mathbb{N}$ *is the alphabet, and the infinite transition matrix* $A = (a_{i,j})_{i,j\in\mathbb{N}}$ *is given by*

$$a_{i,j} = \begin{cases} 1 & \text{if } j \geqslant i-1, \\ 0 & \text{if } j < i-1, \end{cases}$$

then $h_{top}(\sigma) = \log 4$*, but there is no measure of maximal entropy.*

Exercise 6. *Find the maximal measure for the Fibonacci subshift of finite type. What is the limit frequency of the symbol zero in* μ*-typical sequences* x*?*

References

[1] R. L. Adler, A. G. Konheim, M. H. McAndrew, *Topological entropy,* Trans. Amer. Math. Soc. **114** (1965), 309-319.

[2] V. I. Arnol'd, A. Avez, *Ergodic Problems in Classical Mechanics,* New York: Benjamin (1968).

[3] V. I. Arnold, *Mathematical Methods of Classical Mechanics,* Springer-Verlag (1989).

[4] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms,* Lect. Math. **470** Spring 1974 and second revised edition 2008.

[5] K. M. Brucks, H. Bruin, *Topics from one-dimensional dynamics,* London Mathematical Society, Student Texts **62** Cambridge University Press 2004.

[6] H Bruin, J. Hawkins, *Rigidity of smooth one-sided Bernoulli endomorphisms,* New York Journ. Maths. **15** (2009), 451–483.

[7] H. Bruin, M. Todd, *Transience and thermodynamic formalism for infinitely branched interval maps,* J. London Math. Soc. **86** 2012, 171–194

[8] E. I. Dinaburg, *The relation between topological entropy and metric enropy,* Dokl. Akad. Nauk SSSR., **190** (1970), 19–22, Soviet Math. Dok. **11** (1970), 13-16.

[9] P. and T. Ehrenfest, *Begriffliche Grundlagen der statistischen Auffassung in der Mechanik,* Enzy. d. Mathem. Wiss. IV, 2, II (1912), 3-90. English translation: *The conceptual foundations of the statistical approach in mechanics,* Cornell Univ. Press (1959).

[10] G. Keller, *Equilibrium states in Ergodic Theory,* London Math. Society, Student Texts **42** Cambridge Univ. Press, Cambridge 1998.

[11] F. Ledrappier, *Some properties of absolutely continuous invariant measures on an interval,* Erg. Th. and Dyn. Sys. **1** (1981), 77–94.

[12] M. Misiurewicz and W. Szlenk, *Entropy of piecewise monotone mappings,* Studia Math. **67** (1980), no. 1, 45-63.

[13] K. Petersen, *Ergodic theory,* Cambridge University Press, 1983.

[14] D. Ornstein, *Ergodic theory, randomness and dynamical systems,* Yale Math. Monographs, **5** Yale Univ. New Haven, 1974.

[15] W. Parry, *Intrinsic Markov chains*, Trans. AMS. **112** (1965), 55-65.

[16] P. Walters, *An introduction to ergodic theory,* Springer Verlag (1982).

Ergodic Theory
(Lecture Notes)

Joan Andreu Lázaro Camí

Department of Mathematics
Imperial College London
January 2010

Contents

Chapter 1

Introduction

Ergodic theory lies in somewhere among measure theory, analysis, probability, dynamical systems, and differential equations and can be motivated from many different angles. We will choose one specific point of view but there are many others. Let

$$\dot{x} = f(x)$$

be an *ordinary differential equation*. The problem of studying differential equations goes back centuries and, throughout the years, many different approaches and techniques have been developed. The most classical approach is that of finding explicit analytic solutions. This approach can provide a great deal of information but is essentially only applicable to an extremely restricted class of differential equations. From the very beginning of the 20th century, there has been a great development on topological methods to obtain qualitative topological information such as the existence of periodic solutions. Again, this can be a very successful approach in certain situations but there are a lot of equations which have, for example, infinitely many periodic solutions, possibly intertwined in very complicated ways, to which these methods do not really apply. Finally there are numerical methods to approximate solutions. In the last few decades, with the increase of computing power, there has been hope that numerical methods could play an important role. Again, while this is true in some situations, there are also a lot of equations for which the numerical methods have very limited applicability because the approximation errors grow exponentially and quickly become uncontrollable. Moreover, the *sensitive dependence on initial conditions* is now understood to be an intrinsic feature of certain equations than cannot be resolved by increasing the computing power.

Example 1 Lorenz's equations were introduced by the meteorologist E.

Lorenz in 1963 as an extremely simplified model of the Navier-Stokes equations for a fluid flow:

$$\left\{\begin{array}{l} \dot{x}_1 = \sigma\left(x_2 - x_1\right) \\ \dot{x}_2 = x_1\left(\rho - x_3\right) - x_2 \\ \dot{x}_3 = x_1 x_2 - \beta x_3, \end{array}\right. \tag{1.1}$$

where σ is the ***Prandlt number*** and ρ the ***Rayleigh number***. Usually $\sigma = 10$, $\beta = 8/3$, and ρ varies. However, for $\rho = 28$, the systems exhibits a chaotic behaviour. This is a very good example of a relatively simple ODE which is quite intractable from many angles. It does not admit any explicit analytic solutions; the topology is extremely complicated with infinitely many periodic solutions which are knotted in many different ways (there are studies of the structure of the periodic solutions of Lorenz's equations from the point of view of knot theory); on the other hand, numerical integration has very limited use since nearby solutions diverge very quickly.

Using classical methods, one can prove that the solutions of Lorenz's equations, eventually, end up in some bounded region $U \subset \mathbb{R}^3$. This simplifies our approach significantly since it means that it is sufficient to concentrate on the solutions inside U. A combination of results obtained over almost 40 years by several different mathematicians can be formulated in the following theorem which can be thought of essentially as a statement in ergodic theory. We give here a precise but slightly informal statement as some of the terms will be defined more precisely later on these notes.

Theorem 2 *For every ball $B \subset \mathbb{R}^3$, there exists a "probability" $p(B) \in [0,1]$ such that, for "almost every" initial condition $x_0 \in \mathbb{R}^3$, we have*

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{1}_B\left(x_t\right) dt = p\left(B\right), \tag{1.2}$$

where x_t is the solution of (1.1) with initial condition x_0.

First of all, recall that $\mathbf{1}_B$ is the characteristic function of the set B defined by

$$\mathbf{1}_B\left(x\right) = \left\{\begin{array}{ll} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{array}\right.$$

The integral $\int_0^T \mathbf{1}_B\left(x_t\right) dt$ is simply the amount of time that the solution x_t spends inside the ball B between time 0 and time T, and $T^{-1} \int_0^T \mathbf{1}_B\left(x_t\right) dt$ is therefore the *proportion* of time that the solution spends in B from $t = 0$ to T. Theorem 2 makes two highly non trivial assertions:

1. that the proportion $T^{-1}\int_0^T \mathbf{1}_B(x_t)\,dt$ converges as $T \to \infty$;
2. that this limit is independent of the initial condition x_0.

There is no a priori reason why the limit (1.2) should exist. But perhaps the most remarkable fact is that the limit is the same for *almost all* initial conditions (the concept *almost all* will be made precise later). This says that the asymptotic time averages of the solution x_t with initial condition x_0 are actually independent of this initial condition. Therefore, independently of the initial condition, the proportion of time that the systems spends on B is $P(B)$. In other words, there exists a way of *measuring* the balls B such that the measure $P(B)$ gives us information on the amount of time that the system, on average, spends on B. Theorem 2 is just a particular case of the more general Birkoff's Ergodic Theorem which we will state and prove in Chapter 4.

The moral of the story is that even though Lorenz's equations are difficult to describe from an analytic, numerical, or topological point of view, they are very well behaved from a probabilistic point of view. The tools and methods of probability theory are therefore very well suited to study and understand these equations and other similar dynamical systems. This is essentially the point of view on ergodic theory that we will take in these lectures. Since this is an introductory course, we will focus on the simplest examples of dynamical systems for which there is already an extremely rich and interesting theory, which are one-dimensional maps of the interval or the circle. However, the ideas and methods which we will present often apply in much more general situations and usually form the conceptual foundation for analogous results in higher dimensions. Indeed, results about interval maps are applied directly to higher dimensional systems. For example, Lorenz's equations can be studied taking a cross section for the flow and using Poincaré's first return map, which essentially reduces the system to a one dimensional map.

Chapter 2

Measure Theory

In this chapter, we will introduce the minimal requirements of Measure Theory which will be needed later. In particular, we will review one of the pillars of measure theory, namely, the concept of integral with respect to an arbitrary measure. For a more extensive exposition, the reader is encouraged to check, for example, with [2].

2.1 Motivation: Positive measures and Cantor sets

The notion of measure is, in the first instance, a generalization of the standard idea of length, or, in general, volume. Indeed, while we know how to define the length λ of an interval $[a, b]$, namely, $\lambda([a, b]) = b - a$, we do not a priori know how to measure the size of sets which contain no intervals but which, logically, have *positive measure*. For example, let $\{r_i\}_{i=0}^{\infty}$ be a sequence of positive numbers such that $\sum_i r_i < 1$. Define a set $C \subset [0, 1]$ recursively removing open subintervals from $[0, 1]$ in the following way. To start with, we remove an open subinterval I_0 of length r_0 from the interior of $[0, 1]$ so that $[0, 1] \setminus I_0$ has two connected components. Then we remove intervals I_1 and I_2 of lengths r_1 and r_2 respectively from the interior of these components so that $[0, 1] \setminus (I_0 \cup I_1 \cup I_2)$ has 4 connected components. Now remove intervals $I_3, ..., I_6$ from each of the interiors of these components and continue in this way. Let

$$C = [0, 1] \setminus \bigcup_{i=0}^{\infty} I_i \tag{2.1}$$

be a ***Cantor set***. By construction, C does not contain any intervals since every interval is eventually subdivided by the removal of one of the subintervals I_k from its interior. Therefore, it seems that it does not make sense to talk about C as having any *length*. Nevertheless, the total length of the intervals removed is $\sum_{i\geq 0} r_i < 1$ so it would make sense to say that the size of C is $1 - \sum_{i\geq 0} r_i$. Measure theory formalizes this notion in a rigorous way and makes it possible to assign a size to sets such as C.

Remark 3 *If $\sum_{i\geq 0} r_i = 1$ is exactly 1, then C is an example of a non-countable set of zero Lebesgue measure.*

Non-measurable sets

The example above shows that it is desirable to generalise the notion of *length* so that we can apply it to *measure* more complicated subsets which are not intervals. In particular, we would like to say that the Cantor set defined above has positive measure. It turns out that, in general, it is not possible to define a measure in a consistent way on all possible subsets. In 1924 Banach and Tarski showed that it is possible to divide the unit ball in 3-dimensional space into 5 parts and re-assemble these parts to form two unit balls, thus apparently doubling the volume of the original set. This implies that it is impossible to consistently assign a well defined volume to any subset in an additive way. See a very interesting discussion on wikipedia on this point (Banach-Tarski paradox).

Consider the following simpler example. Let $\mathbb{S}^1$ be the unit circle and let $f_\alpha : \mathbb{S}^1 \to \mathbb{S}^1$ be an irrational circle rotation. We will see that, in this case, every orbit is dense in $\mathbb{S}^1$ (Theorem 27) Let $A \subset \mathbb{S}^1$ be a set containing exactly one point from each orbit. Suppose that we have defined a general notion of a measure m on $\mathbb{S}^1$ that generalises the notion of length of an interval so that the measure $m(A)$ has a meaning. In particular, in order to be well-defined, such a measure will be translation invariant in the sense that the measure of a set cannot be changed by simply translating this set. Therefore, since a circle rotation f_α is just a translation, we have $m(f_\alpha^n(A)) = m(A)$ for every $n \in \mathbb{Z}$, where $f_\alpha^n := f_\alpha \circ \overset{n \text{ times}}{\cdots} \circ f_\alpha$. Moreover, since A contains only one single point from each orbit and all points on a given orbit are distinct, we have $f_\alpha^n(A) \cap f_\alpha^m(A) = \emptyset$ if $n \neq m$. Consequently,

$$m\left(f_\alpha^n(A) \cup f_\alpha^m(A)\right) = m(f_\alpha^n(A)) + m(f_\alpha^m(A)).$$

Therefore,

$$1 = m\left(\mathbb{S}^1\right) = m\left(\bigcup_{n=-\infty}^{\infty} f^n(A)\right) = \sum_{n=-\infty}^{\infty} m(f^n(A)) = \sum_{n=-\infty}^{\infty} m(A)$$

which is clearly impossible as the right hand side is zero if $m(A) = 0$ or infinity if $m(A) > 0$. In order to overcome this difficulty, one has to restrict the family of subsets which can be assigned a length consistently. This subsets will be called *measurable sets* and the family a *σ-algebra*.

Remark 4 *The previous counterexample depends on the Axiom of Choice to ensure that the set constructed by choosing a single point from each of an uncountable family of subsets exists.*

2.2 Measures and σ-algebras

Let X be a set and / a collection of (not necessarily disjoint) subsets of X.

Definition 5 *We say that / is an* ***algebra*** *(of subsets of X) if*

1. *- # / and X # /,*
2. *A # / implies A^c # /,*
3. *for any finite collection $A_1, ..., A_n$ of subsets in / we have that $\dot{}_{i=1}^{n} A_i$ # /.*

*We say that / is a σ-**algebra** if, additionally,*

3'. *for any countable collection &A_i'$_{i\%\mathbb{N}}$ of subsets in /, we have*

$$(_{i\%\mathbb{N}} A_i \# / .$$

The family of all subsets of a set X is obviously a σ-algebra. Given 0 a family of subsets of X we define the ***σ-algebra*** $\sigma(0)$ ***generated by*** 0 as the smallest σ-algebra containing 0. That is, as the intersection of all the σ-algebras containing 0. This is always well defined and is in general smaller than the σ-algebra of all subsets of X.

Exercise 6 Prove that the intersection of all the σ-algebras containing 0 is indeed a σ-algebra.

If X is a topological space, the σ-algebra generated by open sets is called the ***Borel σ-algebra*** and is denoted by $1(X)$. Observe that a Cantor set C introduced in (2.1) is the complement of a countable union of open intervals and, therefore, belongs to Borel σ-algebra $1([0,1])$.

Definition 7 *A real-valued set function μ on a class of sets $\mathcal{C}$ is called*

1. ***additive*** *if*
$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$$
for any finite sequence $\{A_1, ..., A_n\} \subset \mathcal{C}$ of pairwise disjoint sets such that $\bigcup_{i=1}^{n} A_i \in \mathcal{C}$.

2. ***countably additive*** *(or σ-additive) if*
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$
for any countably collection $\{A_i\}_{i\geq 1} \subset \mathcal{A}$ of pairwise disjoint sets such that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Definition 8 *Let $\mathcal{A}$ be a σ-algebra of X. A **measure** μ is a function*
$$\mu : \mathcal{A} \longrightarrow [0, \infty]$$
which is countably additive.

This definition shows that the σ-algebra is as intrinsic to the definition of a measure as the space itself. In general, we refer to a ***measure space*** as a triple $(X, \mathcal{A}, \mu)$. The elements in the σ-algebra $\mathcal{A}$ are called ***measurable sets***. We say that μ is ***finite*** if $\mu(X) < \infty$ and that μ is a ***probability measure*** if $\mu(X) = 1$. A measure μ is called ***σ-finite*** if $X = \bigcup_{i=1}^{\infty} A_i$ such that $A_i \in \mathcal{A}$ and $\mu(A_i) < \infty$ for any i. For example, the Lebesgue measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is σ-finite because
$$\mathbb{R} = \bigcup_{n\in\mathbb{Z}} [n, n+1]$$
because the Lebesgue measure of an interval is its length. On the other hand, observe that if $\hat{\mu}$ is a finite measure we can easily define a probability measure μ by
$$\mu(A) = \frac{\hat{\mu}(A)}{\hat{\mu}(X)}, \quad A \in \mathcal{A}.$$

Exercise 9 Let $\mathcal{A}$ be a σ-algebra and let $\mu : \mathcal{A} \to [0, \infty]$ be an additive positive function, $\mu \neq \infty$. Then,

1. μ is σ-additive $\iff$ For any increasing sequence $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ (i.e., $A_n \subset A_{n+1}$) we have
$$\lim_{n\to\infty} \mu(A_n) = \mu(A) \quad \text{where } A := \bigcup_{n\in\mathbb{N}} A_n.$$
2.
   (a) μ is σ-additive $\Longrightarrow$ For any decreasing sequence $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ (i.e., $A_{n+1} \subset A_n$) such that $\mu(A_1) < \infty$ we have
   $$\lim_{n\to\infty} \mu(A_n) = \mu(A) \quad \text{where } A := \bigcap_{n\in\mathbb{N}} A_n.$$
   (b) If for any decreasing sequence $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{F}$ such that $A_n \downarrow \emptyset$, we have $\lim_{n\to\infty} \mu(A_n) = 0$, then μ is σ-additive.

Defining a countably additive function on σ-algebras is non-trivial. It is usually easier to define countably additive functions on algebras because the class of sequences $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$ such that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ is smaller than in σ-algebras. Observe that, unlike what happens in σ-algebras, $\bigcup_{n\in\mathbb{N}} A_n$ needs not belong to $\mathcal{A}$ if $\mathcal{A}$ is only an algebra. For example, the standard length is a countably additive function on the algebra generated by finite unions of intervals. The fact that this extends to a countably additive function on the corresponding σ-algebra (and therefore, that we can measure Cantor sets) is guaranteed by the following fundamental result.

Theorem 10 (Carathéodory's Theorem, [3, Theorem 1.5.6]) *Let μ_0 be a countably additive function defined on an algebra $\mathcal{A}$ of subsets. Then μ_0 can be extended in a unique way to a countably additive function μ on the σ-algebra $\mathcal{F} = \sigma(\mathcal{A})$.*

Remark 11 The σ-additivity of μ_0 cannot be removed as the following counter-example shows. Let $\mathcal{A}$ be the algebra of sets $A \subset \mathbb{N}$ such that either A or $\mathbb{N}\setminus A$ is finite. For finite A, let $\mu(A) = 0$, and for A with a finite complement let $\mu(A) = 1$. Then μ is an additive, but not countably additive set function.

Proof. It is clear that A is indeed an algebra. $\mu(A \cup B) = \mu(A) + \mu(B)$ is obvious for disjoint sets A and B if A is finite. Finally, A and B in $\mathcal{A}$ cannot be infinite simultaneously being disjoint. If μ was countably additive, we would have
$$\mu(\mathbb{N}) = \sum_{n=1}^{\infty} \mu(\{n\}) = 0,$$
which is clearly a contraction. ■

Nevertheless, defining measures on $\mathbb{R}$ is easier, as the next subsection summarises.

2.2.1 Measures on $\mathbb{R}$

In this subsection, we are going to gather some definitions and results that, roughly speaking, state that a measure on $\mathbb{R}$ is completely determined by the value of that measure on intervals of the form $(a,b]$, $a \leqslant b$.

Definition 12 *A* ***distribution function*** $F : \mathbb{R} \to \mathbb{R}$ *is a right-continuous increasing function. That is,*

$$x \leqslant y \Rightarrow F(x) \leqslant F(y) \quad \text{and} \quad \lim_{x \to a^+} F(x) = F(a).$$

Let now $\mathfrak{J} := \{(a,b] : a \leqslant b \in \mathbb{R}\}$ and let $F : \mathbb{R} \to \mathbb{R}$ be a distribution function. Define

$$\begin{aligned} \mu : \mathfrak{J} &\longrightarrow [0,\infty] \\ (a,b] &\longmapsto F(b) - F(a). \end{aligned} \tag{2.2}$$

Then, one can prove that μ thus defined is a σ-additive and σ-finite function on $\mathfrak{J}$. Moreover, there exists a unique σ-additive extension of μ onto $\mathcal{B}(\mathbb{R})$. That is,

Theorem 13 *A distribution function* $F : \mathbb{R} \to \mathbb{R}$ *determines a measure* μ *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *by means of the formula*

$$\mu((a,b]) = F(b) - F(a), \quad (a,b] \in \mathfrak{J}.$$

A natural question now arises. Can we obtain any measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ from a distribution function? The answer is no, but *almost* any of them. Observe that the measure defined through (2.2) is finite on any bounded interval. These are precisely the measures we can generate by means of distribution functions. They are called Lebesgue-Stieljes measures.

Definition 14 *A* ***Lebesgue-Stieljes measure*** *is a measure* μ *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *such that, for any bounded* $A \in \mathcal{B}(\mathbb{R})$, $\mu(A) < \infty$.

Proposition 15 *Let* μ *be a Lebesgue-Stieljes measure. Then, there exists a distribution function* F *such that*

$$\forall a < b \in \mathbb{R}, \quad F(b) - F(a) = \mu((a,b]). \tag{2.3}$$

Proof. Define $F(0) = c \in \mathbb{R}$ any arbitrary value and

$$F(x) := \begin{cases} F(0) + \mu((0,x]) & \text{if } x > 0 \\ F(0) - \mu((x,0]) & \text{if } x < 0. \end{cases}$$

F is a distribution function. It is clearly increasing and, by definition, satisfies (2.3). In order to check the right-continuity, let $b \geq 0$. Then

$$\lim_{x \to b^+} F(x) = F(0) + \lim_{x \to b^+} \mu((0,x]) = F(0) + \mu((0,b]) = F(b),$$

where in the second equality we have used that

$$\lim_{x \to b} \mu(A_x) = \mu(A_b) \quad \text{where } A_x = (0,x],\ A_x \uparrow A_b \text{ as } x \downarrow b$$

(see Exercise 9.2(a). The case $b < 0$ is analogous. ■

2.2.2 Examples

1. **Dirac delta measures**. Dirac measures δ_a, $a \in \mathbb{R}$, are defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as

$$\delta_a(A) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{if } a \notin A, \end{cases} \qquad A \in \mathcal{B}(\mathbb{R}).$$

In this case, we say that the *entire mass* is concentrated at the single point a. The distribution function of δ_a is

$$F(x) = \begin{cases} 1 & \text{if } x \geq a \\ 0 & \text{if } x < a. \end{cases}$$

An immediate generalization is the case of a measure concentrated on finite set of points $\{a_1, ..., a_n\}$ each of which carries some proportion $\rho_1, ..., \rho_n$ of the total mass, i.e., $\mu := \sum_{i=1}^n \rho_i \delta_{a_i}$ with $\rho_1 + ... + \rho_n = 1$. Then, given $A \in \mathcal{B}(\mathbb{R})$,

$$\mu(A) = \sum_{\{i : a_i \in A\}} \rho_i$$

is the sum of the weights carried by those points contained in A.

2. **Lebesgue measure**. Lebesgue measure is defined on $\mathcal{B}(\mathbb{R})$ and assigns to any subinterval $I \subset \mathbb{R}$ its length. Lebesgue measure λ is characterised by the distribution function $F(x) = x$. That is,

$$\lambda((a,b]) = b - a.$$

3. **Absolutely continuous measures**. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. For any subinterval $I \subset \mathbb{R}$ define
$$\mu(I) := \int_I f(y)\, dy.$$
Then μ defines a σ-finitely additive function on the algebra of finite unions of subintervals of $\mathbb{R}$ and thus extends uniquely to a measure on $\mathfrak{B}(\mathbb{R})$. Indeed, a possible distribution function F associated to μ is
$$F(x) = \int_0^x f(y)\, dy.$$

4. **Normal law**. The probability measure given by the distribution function
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2}\, dy$$
is called the ***standard normal law*** and is denoted by $N(0,1)$.

5. **Measures on spaces of sequences**. Let Σ_k^+ denote the set of infinite sequences of k symbols. That is, and element $a \in \Sigma_k^+$ is a sequence $a = (a_0, a_1, ...)$ with $a_i \in \{0, 1, ..., k-1\}$. For any given *finite block* $(x_0, ..., x_{n-1})$ of length n with $x_i \in \{0, 1, ..., k-1\}$, let
$$I_{x_0 \dots x_{n-1}} := \left\{ a \in \Sigma_k^+ : a_i = x_i,\ i = 0, ..., n-1 \right\}$$
denote the set of all infinite sequences which start precisely with the prescribed finite block $(x_0, ..., x_{n-1})$. We call this a ***cylinder set*** of order n. Let
$$\mathcal{A} = \{\text{finite unions of cylinder sets}\}.$$

Exercise 16 Show that $\mathcal{A}$ is an algebra of subsets of Σ_k^+.

Fix now k numbers $\{p_0, ..., p_{k-1}\} \subset [0,1]$ such that $p_0 + ... + p_{k-1} = 1$ and define a function $\mu : \mathcal{A} \to \mathbb{R}_+$ on the algebra of cylinder sets by
$$\mu\left(I_{x_0 \dots x_{n-1}}\right) := \prod_{i=0}^{n-1} p_{x_i}.$$

Exercise 17 Prove that μ is σ-additive.

Therefore, the function μ extends uniquely to a measure on the σ-algebra $\mathfrak{B} = \sigma(\mathcal{A})$.

2.3 Integration

Integration with respect to a measure can be regarded as a powerful generalization of the standard Riemann integral. In this section, we are going to review the basics of integration with respect to an arbitrary measure. Before, we need to introduce some definitions.

Let $(X, \mathcal{A}, \mu)$ be a measure space and let $A \in \mathcal{A}$. We define the ***characteristic function*** $\mathbf{1}_A$ as

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

On the other hand, a ***simple*** or ***elementary function*** $\zeta : X \to \mathbb{R}$ is a function that can be written in the form

$$\zeta = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$$

for some constants $c_i \in \mathbb{R}$ and some disjoint measurable sets $A_i \in \mathcal{A}$, $i = 1, ..., n$. The integral of a simple function ζ with respect to the measure μ is defined in a straightforward manner as

$$\int \zeta d\mu := \sum_{i=1}^{n} c_i \mu(A_i).$$

The idea is to extend this integral to more general functions. More concretely, we can define the integral of a measurable function. Recall that a function $f : X \to \mathbb{R}$ is ***measurable*** if $f^{-1}(I) \in \mathcal{A}$ for any $I \in \mathcal{B}(\mathbb{R})$. If $(X, \mathcal{A}, \mu)$ is a probability space (i.e., μ is a probability), measurable functions are usually called ***random variables***.

Exercise 18 *Let $f : X \to \mathbb{R}_+$ be a measurable function. Show that f is the (pointwise) limit of an increasing sequence of elementary functions.* ***Hint:*** *define, for any $n \in \mathbb{N}$,*

$$\zeta_n := \sum_{k=1}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \leq f < \frac{k+1}{2^n}\right\}} + n\mathbf{1}_{\{n \leq f\}}.$$

The integral of a general, measurable, *non-negative* function $f : \mathbb{R} \to \mathbb{R}_+$ can be defined in two equivalent ways. On the one hand,

$$\int f d\mu := \sup\left\{\int \zeta d\mu : \zeta \text{ is simple, } \zeta \leq f\right\}. \tag{2.4}$$

On the other hand, one can prove that

$$\int f d\mu = \lim_{n \to \infty} \int \zeta_n d\mu$$

where $\{\zeta_n\}_{n \in \mathbb{N}}$ is an increasing sequence of elementary functions converging to f. The integral is usually called the ***Lebesgue integral*** of the function f with respect to the measure μ (even if μ is not Lebesgue measure). Observe that $\int f d\mu$ may be ∞.

Remark 19 *Note that, unlike the Riemann integral which is defined by a limiting process that may or may not converge, the supremum in (2.4) is always well defined, though it needs not be finite.*

In general, let $f : X \to \mathbb{R}$ be a measurable function and write

$$f = f^+ - f^-$$

where $f^+ = \max\{f, 0\}$ and $f^- = \min\{-f, 0\}$. It is not difficult to prove that both f^+ and f^- are measurable, non-negative functions.

Definition 20 *Let $f : X \to \mathbb{R}$ be a measurable function. If*

$$\int f^+ d\mu < \infty \quad \text{and} \quad \int f^- d\mu < \infty$$

*we say that f is **μ-integrable** and we define*

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu.$$

The set of all μ-integrable functions is denoted by $L^1(X, \mu)$.

2.3.1 Properties of the Lebesgue integral

Let $f, g : X \to \mathbb{R}$ be two arbitrary measurable functions. The Lebesgue integral has the following properties (that, in general, are not difficult to prove):

1. $$\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$$

   for any $a, b \in \mathbb{R}$.

2. If $A \in \Sigma$ is such that $\mu(A) = 0$, then

$$\int_A f d\mu := \int \mathbf{1}_A f d\mu = 0.$$

That is, the integral of f over a set of measure 0 is 0. This is true even if f takes the values $\pm\infty$ on A. That is, if A contains singularities of f. Recall that we say that a point $x = a$ is a ***singularity*** if $f(a) = \pm\infty$. For example, if a non-negative function $f \geq 0$ is integrable, $\int f d\mu < \infty$, then we can say that $\mu(\{x : f(x) = \infty\}) = 0$.

3. If $f \geq 0$ and $\int f d\mu = 0$, then $\mu(\{x : f(x) > 0\}) = 0$.

4. If $f \leq g$, then
$$\int f d\mu \leq \int g d\mu.$$

5.
$$\left|\int f d\mu\right| \leq \int |f| d\mu. \tag{2.5}$$

Indeed, $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$ and

$$\begin{aligned}\left|\int f d\mu\right| &= \left|\int f^+ d\mu - \int f^- d\mu\right| \leq \left|\int f^+ d\mu\right| + \left|\int f^- d\mu\right| \\ &= \int f^+ d\mu + \int f^- d\mu = \int |f| d\mu.\end{aligned}$$

By (2.5), we can characterise $L^1(X, \mu)$ as the space of measurable functions $f : X \to \mathbb{R}$ such that

$$\int |f| d\mu < \infty.$$

In general, for $p \geq 1$, we introduce the spaces $L^p(X, \mu)$ as the space of measurable functions such that $\int_X |f|^p d\mu < \infty$, where two functions are identified if they differ, at most, on a set of zero measure. $L^p(X, \mu)$ is a *Banach space* with the norm $\|f\|_p := (\int_X |f|^p d\mu)^{1/p}$.

Example 21 Let $f : [0, 1] \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{Q} \\ 1 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

It is well known that this function is not Riemann integrable because the limit of the upper and lower Riemann sums do not coincide. However, as a function measurable with respect to the Lebesgue measure λ, f is simple: it values 0 on the measurable set $\mathbb{Q}$ and values 1 on the measurable set $[0,1]\setminus\mathbb{Q}$. The set of rational numbers $\mathbb{Q}$ has zero Lebesgue measure because $\mathbb{Q}$ is countable. Therefore $\lambda([0,1]\setminus\mathbb{Q}) = 1$ and

$$\int_{[0,1]} f d\lambda = \lambda\left([0,1]\setminus\mathbb{Q}\right) = 1.$$

Chapter 3

Invariant measures

3.1 Invariant measures: definitions and examples

Let $(X, 2\,, \mu)$ and $(Y, >, \nu)$ be two measure spaces. A map $T : X\ \$\ Y$ is called ***measurable*** if the preimage $T^{\#1}(A)$ of any measurable set $A\ \#\ >$ is measurable, i.e., $T^{\#1}(A)\ \#\ 2\,$. A measurable map T is ***non-singular*** if the preimage of every set of measure 0 has measure 0. The map $T : X\ \$\ Y$ is ***measure-preserving*** if $\mu\ T^{\#1}(A)' = \nu(A)$ for any $A\ \#\ >$. A non-singular map from a measure space $(X, 2\,, \mu)$ into itself is called a ***non-singular transformation***, or simply a ***transformation***. If a transformation $T : X\ \$\ X$ preserves a measure μ, then μ is called T***-invariant***. Usually, we will deal with measurable maps between topological spaces. In that case, the σ-algebras involved will be always the corresponding Borel σ-algebras.

A set has ***full measure*** if its complement has measure 0. We say that a property holds for μ***-almost every*** x (μ-a.e.) or μ***-almost surely*** (μ-a.s.) if it holds on a subset of full μ-measure.

Let λ be the Lebesgue measure on $\mathbb{R}$. A ***flow*** $\&T^t\,'_{t\%J}$ on a measurable space $(X, 2\,, \mu)$ is a family of measurable maps $T^t : X\ \$\ X$ where, usually, J equals $\mathbb{R}$ (time-continuous flows) or $\mathbb{N}$ (discrete flows). If $J = \mathbb{N}$, we say that the flow $\&T^n\,'_{n\%\mathbb{N}}$ is ***measurable*** if T^n is measurable for any n. When $J = \mathbb{R}$, $\&T^t\,'_{t\%\mathbb{R}}$ is ***measurable*** if the product map $T : X\ ?\ \mathbb{R}\ \$\ X$ given by $T(x, t) = T^t(x)$ is measurable with respect to the product σ-algebra $\sigma(2\ ?\ 1(\mathbb{R}))$ on $X\,?\,\mathbb{R}$, and $T^t : X\ \$\ X$ is a non-singular measurable transformation for any $t\ \#\ \mathbb{R}$. A measurable flow T^t is a ***measure-preserving flow*** if each T^t is a measure-preserving transformation. Discrete flows are usually built from measurable maps $T : X\ \$\ X$ as follows: for any $n\ \#\ \mathbb{N}$, we define

$$T^n(x) = T + \overset{n)}{\ldots} + T(x)$$

and $T^0 = \mathrm{Id}$, the identity on X.

3.1.1 Examples

1. **Dirac measures on fixed points**. If $T : X \ \$ \ X$ is a measurable map and p a fixed point of T, $T(p) = p$, then the Dirac measure δ_p is invariant. Indeed, let $A \# 2$ be a arbitrary measurable set. We have to prove that

   $$\delta_p \overset{+}{\ } T^{\#1}(A)' = \delta_p(A). \tag{3.1}$$

   We consider two cases. First of all, suppose $p \# A$ so that $\delta_p(A) = 1$. In this case $p \# T^{\#1}(A)$ clearly so $\delta_p \overset{+}{\ } T^{\#1}(A)' = 1$ and (3.1) holds. Secondly, suppose that $p \not\# A$. Then $\delta_p(A) = 0$ and we also have $p \not\# T^{\#1}(A)$ because if $p \# T^{\#1}(A)$ then $p = T(p) \# T \overset{+}{\ } T^{\#1}(A)' + (A$, which would be a contradiction. Therefore $p \not\# T^{\#1}(A)$, $\delta_p \overset{+}{\ } T^{\#1}(A)' = 0$, and (3.1) holds again.

2. **Dirac measures on periodic orbits**. Let $T : X \ \$ \ X$ be a measurable map and let $P = \&a_1, ..., a_n'$ be a periodic orbit with minimal period n. That is, $T(a_i) = a_{i+1}$ for $i = 1, ..., n \, ! \, 1$, and $T(a_n) = a_1$. Let $\rho_1, ..., \rho_n$ be constants such that $\rho_i \# (0, 1)$ and $\sum_{i=1}^n \rho_i = 1$. Consider the measure

   $$\delta_P(A) = \sum_{\&i:a_i\%A'} \rho_i.$$

   Exercise 22 Show that δ_P is invariant if and only if $\rho_i = 1/n$ for every $i = 1, ..., n$.

3. **Circle rotations**.

   Proposition 23 *Let $T : \mathbb{S}^1 \ \$ \ \mathbb{S}^1$ be a circle rotation, $T(x) = x + \alpha$ for some $\alpha \# \mathbb{R}$. The Lebesgue measure is invariant.*

   Proof. T is just a translation and Lebesgue measure is invariant under translations. ■

   However, depending on the value of α, there may be other invariant measures. If $2\pi/\alpha$ is rational then all points $x \# \mathbb{S}^1$ are periodic of the same period and, therefore, T admits also infinitely many distinct

Dirac measures on the periodic orbits (see Example 26). If $2\pi/\alpha$ is irrational, then all orbits are dense in $\mathbb{S}^1$ (Example 26) and the Lebesgue measure is the unique invariant measure of T.

4. **Measure-preserving flows in** $\mathbb{R}^n$. Let $U \subset \mathbb{R}^n$ be an open set and $v : U \to \mathbb{R}^n$ a C^r vector field, $r \geq 1$. Consider the differential equation

$$\dot{x} = v(x). \tag{3.2}$$

Suppose that, for every $p \in U$, there exists a (unique) solution $x : \mathbb{R} \to U$ of (3.2) with initial condition p, which means that, $\dot{x}_t = v(x_t)$ and $x_{t=0} = p$. For any $t \in \mathbb{R}$, we define the map $\varphi_t : U \to U$ by $\varphi_t(p) = x_t$ where $x : \mathbb{R} \to U$ is the solution of (3.2) with initial condition $x_0 = p$. Basic results of ordinary differential equations show that, for every t, the map φ_t is a C^r diffeomorphism and the family of maps $\varphi_t : U \to U$ defines a one-parameter group, i.e., $\varphi_{t=0} = \mathrm{Id}$ (the identity) and $\varphi_{t+s} = \varphi_t \circ \varphi_s$ for any $t, s \in \mathbb{R}$. Moreover, by Liouville's formula,

$$\det \left| \frac{\partial \varphi_t}{\partial x^i}(p) \right| = \exp \left(\int_0^t \operatorname{div} v\,(\varphi_s(p))\, ds \right)$$

for any $p \in U$ and t. Hence, if we assume $\operatorname{div} v = 0$, we have $\det \left| \frac{\partial \varphi_t}{\partial x^i}(p) \right| = 1$ and φ_t preserves the *n-dimensional volume* (or Lebesgue measure). Hamiltonian vector fields are examples of vector fields that satisfy $\operatorname{div} v = 0$. Recall that a vector field is called ***Hamiltonian*** if $n = 2m$ is an even number and there exists a function $H : U \to \mathbb{R}$ such that, denoting the points in $\mathbb{R}^n$ as $(q_1, ..., q_m, p_1, ..., p_m)$,

$$v = \left(\frac{\partial H}{\partial p_1}, ..., \frac{\partial H}{\partial p_m}, -\frac{\partial H}{\partial q_1}, ..., -\frac{\partial H}{\partial q_m} \right).$$

Exercise 24 Complete the proof and show that $\operatorname{div} v = 0$ implies that the flow φ_t associated to v preserves the Lebesgue measure.

3.2 Poincaré's recurrence Theorem

Invariant measures play a fundamental role in dynamics. As a first example, we state and prove the following famous result by Poincaré which implies that recurrence is a generic property of orbits of measure-preserving dynamical systems.

Theorem 25 (Poincaré's Recurrence Theorem) *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T : X \to X$ be a measure-preserving map. Let $A \in \mathcal{B}$ such that $\mu(A) > 0$. Then, μ-almost every point $x \in A$, there exists some $n \in \mathbb{N}$ such that $T^n(x) \in A$. Consequently, there are infinitely many $k \in \mathbb{N}$ for which $T^k(x) \in A$; a point $x \in A$ returns to A infinitely often.*

Proof. Let

$$B := \{x \in A : T^k(x) \notin A \text{ for all } k \in \mathbb{N}\} = A \setminus \bigcup_{k \in \mathbb{N}} T^{-k}(A).$$

Then $B \in \mathcal{B}$ and all the preimages $T^{-k}(B)$ are measurable, have the same measure as B, and *disjoint*. Indeed, suppose that

$$T^{-n}(B) \bigcap T^{-m}(B) \neq \emptyset, \quad n \neq m, \; n > m.$$

That is, there exists some $x \in T^{-n}(B) \bigcap T^{-m}(B)$ such that

$$\begin{aligned} T^m(x) \in T^m\left(T^{-n}(B) \bigcap T^{-m}(B)\right) \\ = T^m\left(T^{-n}(B)\right) \bigcap T^m\left(T^{-m}(B)\right) = T^{n-m}(B) \bigcap B. \end{aligned}$$

But this implies that $T^{n-m}(B) \bigcap B \neq \emptyset$ which contradicts the definition of B.

Now, since X has finite total measure, it follows that B has measure 0. Actually,

$$\infty > \mu(X) \geq \mu\left(\bigcup_{k \in \mathbb{N}} T^{-k}(B)\right) = \sum_{k \in \mathbb{N}} \mu\left(T^{-k}(B)\right) = \sum_{k \in \mathbb{N}} \mu(B),$$

which implies $\mu(B) = 0$. In other words, $\mu(A) = \mu(A \setminus B)$ and every point in $A \setminus B$ returns to A, which proves the first assertion.

To show that almost every point of A returns to A infinitely often let

$$\tilde{B}_n := \{x \in A : T^n(x) \in A \text{ and } T^k(x) \notin A, \; k > n\}, \quad n \geq 1,$$

denote the set of points which return to A for the last time after exactly n iterations. We will show that $\mu(\tilde{B}_n) = 0$ for any $n \geq 1$ so that the set

$$\tilde{B} := \bigcup_{n \geq 1} \tilde{B}_n \subset A$$

of the points with return to A only finitely many times has measure 0 as well. Indeed, consider the set $T^n(\tilde{B}_n) \subset B$, which is by definition contained in A

and consists of points that never return to A. Therefore, $\mu(T^n(\tilde{B}_n)) = 0$. But

$$\tilde{B}_n \subset T^{-n}\left(T^n(\tilde{B}_n)\right).$$

Consequently, using that μ is T-invariant we have

$$\mu(\tilde{B}_n) \leq \mu\left(T^{-n}\left(T^n(\tilde{B}_n)\right)\right) = \mu\left(T^n(\tilde{B}_n)\right) = 0.$$

■

The conclusion of Poincaré's Recurrence Theorem may be useless if the preserved measure μ has no *physical* meaning. For example, if $p \in X$ is a fixed point, i.e., $T(p) = (p)$, then the Dirac measure δ_p is invariant. However, with respect to this measure, any set that does not contain p has measure zero so we cannot state anything about the recurrence properties of the systems (except for p, which is a fixed point).

On the other hand, Poincaré's Recurrence Theorem leads us to some paradoxical conclusions. For example, particle dynamics are ruled by Hamiltonian vector fields, which preserve the Lebesgue volume of the phase space. If we open a partition separating a chamber containing gas and a chamber with a vacuum, then Poincaré's Theorem implies that, after a while, the gas molecules will again collect in the first chamber. This is because there exists a set of strictly positive Lebesgue measure in $\mathbb{R}^{3N} \times \mathbb{R}^{3N}$ whose points correspond to the positions and velocities of N particles in the first chamber. The resolution of this paradox lies in the fact that *a while* may be longer than the duration of the solar's system existence. And, of course, that particle dynamics are described at a microscopic level by quantum mechanics, whose effects cannot be taken into account deterministically.

Example 26 Let $T : \mathbb{S}^1 \to \mathbb{S}^1$ be a circle rotation of angle $\alpha \in \mathbb{R}$. If $\alpha = 2\pi\frac{m}{n}$, $m, n \in \mathbb{N}$, then T^n is the identity and Theorem 25 is obvious. If α is not commensurable with 2π, then Poincaré's Recurrence Theorem gives

$$\forall\delta > 0,\ \exists n \in \mathbb{N} \text{ such that } \|T^n(x) - x\| < \delta.$$

It easily follows that

Theorem 27 *If $\alpha \neq 2\pi\frac{m}{n}$, then the orbit $\{T^k(x),\ k = 1, 2, \ldots\}$ is dense on the circle $\mathbb{S}^1$.*

Exercise 28 Prove Theorem 27.

3.3 Invariant measures for continuous maps

In this section, we show that a continuous map $T : X \to X$ of a compact metric space X into itself has at least one invariant Borel probability measure.

Theorem 29 (Krylov-Bogolubov) *Let X be a compact metric space and $T : X \to X$ a continuous map. Then there exists a T-invariant Borel probability μ on X.*

Exercise 30 The compactness condition is essential here. Consider the open interval $I = (0,1)$ and the map $T : (0,1) \to (0,1)$ given by $T(x) = x/2$. Show that T admits no invariant probabilities.

We will prove Krylov-Bogolubov's Theorem in several steps. Before proceeding, we need to introduce some definitions.

Let $\mathcal{M}$ denote the set of all Borel probability measures on a topological space $(X, \mathcal{B})$, $\mathcal{B} = \sigma(X)$. A sequence of measures $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}$ converges in the ***weak* topology*** to a measure $\mu \in \mathcal{M}$ if $\int_X f d\mu_n \to \int_X f d\mu$ for any continuous function $f \in C(X)$ as $n \to \infty$. A measurable map $T : X \to X$ induces a map $T_* : \mathcal{M} \to \mathcal{M}$ defined by $(T_*\mu)(A) := \mu\left(T^{-1}(A)\right)$, $A \in \mathcal{B}$. We call $T_*\mu$ the ***push-forward*** of μ. Similarly, we can define $(T_*^n\mu)(A) = \mu(T^{-n}(A))$. Obviously, μ is T-invariant if and only if $T_*\mu = \mu$.

This notion of convergence is called *weak-star convergence* because the space of finite Borel measures can be canonically identified with the space of linear functionals on the space of continuous functions, i.e. with the dual space of continuous functions. Actually, every finite Borel measure μ on X defines a bounded linear functional $L_\mu(f) = \int_X f d\mu$ on the space $C_c(X)$ of continuous functions on X with compact support. Furthermore, L_μ is positive in the sense that $L_\mu(f) \geq 0$ if $f \geq 0$. The Riesz Representation Theorem (see [9, Theorem 2.14]) states that the converse is also true: for every positive bounded linear functional L on $C_c(X)$, there is a finite Borel measure μ on X such that $L(f) = \int_X f d\mu$, $f \in C_c(X)$. If X is compact, then trivially $C(X) = C_c(X)$.

For an arbitrary measure $\mu_0 \in \mathcal{M}$, we define the sequence of measures

$$\mu_n := \frac{1}{n}\sum_{i=0}^{n-1} T_*^i \mu_0. \tag{3.3}$$

Our aim is to show that the sequence μ_n converges in the weak* topology to a probability measure μ and that this measure is invariant. In order to do that, we will need a couple of Lemmas.

Lemma 31 $\mathcal{M}$ *is compact in the weak* topology.*

Proof (sketch). $C(X)$ is a Banach space with the norm of the supremum

$$\|f\|_\infty = \sup_{x\in X} f(x), \quad f \in C(X).$$

Its dual space, $C(X)^*$ is a normed space with the norm

$$\|L\| = \sup\{|L(f)| : \|f\|_\infty = 1, \ f \in C(X)\}$$

such that the set of probabilities $\mathcal{M} \subset C(X)^*$ is contained in the unit ball

$$\{L \in C(X)^* : \|L\| = 1\}.$$

Indeed, if $\|f\|_\infty = 1$ then $f \leq 1$ a.s. so that, for any measure $\mu \in C(X)^*$,

$$L_\mu(f) = \int_X f d\mu \leq \int_X 1 d\mu, \quad \|f\|_\infty = 1,$$

by monotonicity of the Lebesgue integral. Therefore, since the constant function 1 has norm $\|1\|_\infty = 1$ one,

$$\|\mu\| = \sup\left\{\int_X f d\mu : \|f\|_\infty = 1, \ f \in C(X)\right\} = \int_X 1 d\mu = \mu(X),$$

which implies that the set of probability measures $\mathcal{M}$ is contained in the unit ball of $C(X)^*$.

Moreover, $\mathcal{M}$ is closed. To prove this statement, let $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{M}$ be a sequence of probabilities that, regarded as a sequence of linear functionals $\left(L_{\mu_n}\right)_{n\in\mathbb{N}}$, converge weakly to a linear map $L \in C^*(X)$. We need to show that $L = L_\mu$ for some $\mu \in \mathcal{M}$. Since L_{μ_n} are positive,

$$L_{\mu_n}(f) \geq 0 \quad \forall n \quad \text{if} \quad f \geq 0,$$

then taking the limit $n \to \infty$ we conclude $L(f) \geq 0$ if $f \geq 0$, so L is positive. By the Riesz Representation Theorem, there exists a finite Borel measure μ such that $L = L_\mu$. But $\{\mu_n\}_{n\in\mathbb{N}}$ are probabilities, so

$$1 = L_{\mu_n}(1) = \mu_n(X) \quad \forall n.$$

Letting $n \to \infty$, we also conclude that $1 = L_\mu(1)$, which implies that μ is a probability as required.

Finally, by the Banach-Alaoglu Theorem, we have that the unit ball of $C^*(M)$ is compact. Since $\mathcal{M}$ is a closed set of a compact one, $\mathcal{M}$ is compact too:

Theorem 32 (Banach-Alaoglu, [6, Chapter V, §4]) *Let E be a normed space. Then, the closed balls in E^* are compact in the weak* topology.*

■

Lemma 33 *For all integrable functions $f : X \to \mathbb{R}$ we have*

$$\int_X f\,d\,(T_*\mu) = \int_X (f\circ T)d\mu. \tag{3.4}$$

Proof. We first prove that the statement holds for characteristic functions. If $f = \mathbf{1}_A$, $A \in \mathcal{B}$, then

$$\int_X \mathbf{1}_A d\,(T_*\mu) = (T_*\mu)\,(A) = \mu\left(T^{-1}\,(A)\right) = \int_X \mathbf{1}_{T^{-1}(A)}d\mu = \int_X \mathbf{1}_A\circ T d\mu.$$

Obviously, (3.4) also holds if f is a simple function, that is, a linear combination of characteristic functions. Now suppose that f is a non-negative integrable function. By Exercise 18, there exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions converging to f. Moreover, the sequence $(f_n\circ T)_{n\in\mathbb{N}}$ is clearly an increasing sequence of simple functions converging to $f\circ T$. By the definition of Lebesgue integral, we have

$$\int_X f_n\circ T d\mu \xrightarrow[n\to\infty]{} \int_X f\circ T d\mu \quad \text{and} \quad \int_X f_n d\,(T_*\mu) \xrightarrow[n\to\infty]{} \int_X f d\,(T_*\mu)\,.$$

Since we already proved that $\int_X f_n\circ T d\mu = \int_X f_n d\,(T_*\mu)$ for simple functions and the limit of a sequence is unique, we conclude that

$$\int_X f\circ T d\mu = \int_X f d\,(T_*\mu)\,.$$

For a general measurable f, we repeat the same argument for the positive f^+ and negative f^- parts of f. ■

Lemma 34 $T_* : \mathcal{M} \to \mathcal{M}$ *is continuous.*

Proof. Suppose $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{M}$ is a sequence converging to μ, i.e., $\mu_n \to \mu$ as $n \to \infty$. Then, by the definition of convergence in the weak* topology, for any continuous function $f : X \to \mathbb{R}$ we have

$$\int_X f d\,(T_*\mu_n) = \int_X f\circ T d\mu_n \xrightarrow[n\to\infty]{} \int_X f\circ T d\mu = \int_X f d\,(T_*\mu)$$

where we have used Lemma 33. In other words, $T_*\mu_n \to T_*\mu$ as $n \to \infty$, which, in turn, implies that T is continuous. ■

Proof of Theorem 29. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be the sequence of measure defined in (3.3). The compactness of A implies, in particular, that the sequence $\{\mu_n\}_{n\in\mathbb{N}}$ has a converging subsequence $\{\mu_{n_j}\}_{j\in\mathbb{N}}$. Define

$$\mu := \lim_{j\to\infty} \mu_{n_j}.$$

We will show that μ is invariant.

One the one hand, using the linearity of T_*, we have

$$\begin{aligned}
T_*\mu_{n_j} &= T_*\left(\frac{1}{n_j}\sum_{i=0}^{n_j-1} T_*^i\mu_0\right) = \frac{1}{n_j}\sum_{i=0}^{n_j-1} T_*^{i+1}\mu_0 \\
&= \frac{1}{n_j}\sum_{i=0}^{n_j-1} T_*^i\mu_0 - \frac{1}{n_j}\mu_0 + \frac{1}{n_j}T_*^{n_j}\mu_0 \\
&= \mu_{n_j} - \frac{1}{n_j}\mu_0 + \frac{1}{n_j}T_*^{n_j}\mu_0.
\end{aligned}$$

Since the last two terms of this expression tend to 0 as $j \to \infty$, we conclude that

$$T_*\mu_{n_j} \xrightarrow[n\to\infty]{} \mu.$$

On the other hand, from the continuity of T (Lemma 34), we have $T_*\mu_{n_j} \to T_*\mu$. That is, $T_*\mu = \mu$ and μ is an invariant measure. ■

Chapter 4

Birkhoff's Ergodic Theorem

In this chapter, we will state and prove Birkhoff's Ergodic Theorem which, in the case of ergodic measures, gives a much more sophisticated statement about recurrence than Poincaré's Recurrence Theorem. We will review the examples of the previous chapter and discuss the ergodicity of their measures. Finally, we will conclude with a section on the existence of ergodic measures in general.

4.1 Ergodic transformations.

Throughout this section, $(X, 2, \mu)$ will denote a measure space and $T : X \; \$ \; X$ a measure-preserving transformation.

Definition 35 *We say that T or μ is* ***ergodic*** *if, for any $A \# 2$ such that $T^{\#1}(A) = A$, then A has either measure 0 or full measure.*

Exercise 36 Show that $T^{\#1}(A) = A$ implies $T(A) = A$ but that the converse is not true in general.

A set $A \# 2$ such that $T^{\#1}(A) = A$ is called ***T-invariant***. A measurable function $f : X \; \$ \; \mathbb{R}$ is called ***T-invariant*** if $f + T = f$ almost everywhere.

Proposition 37 *A measurable transformation is ergodic if and only if every invariant measurable function is constant a.e..*

Proof. Exercise. (**Hint:** see the proof of Proposition 38). ■

Therefore, the ergodicity a transformation $T : X \; \$ \; X$ can be characterised saying that T-invariant measurable functions are constant a.e..

However, when $(X, 2, \mu)$ is a finite measure space, we can use this characterisation in a smaller set of functions.

Proposition 38 *Suppose that* $\mu(X) < \%$. *The following properties are equivalent:*

1. *T is ergodic.*

2. *If f # $L^p(X, \mu)$ is T-invariant, p 9 1, then f is constant almost everywhere.*

Proof. 2 =4 1. If A # 2 is T-invariant, the characteristic function $\mathbf{1}_A$ is T-invariant and belongs to $L^p(X, \mu)$. Therefore, $\mathbf{1}_A$ is constant a.e.. That is, $\mu(A) = 0$ or 1.

1 =4 2. If f # $L^p(X, \mu)$ is T-invariant, the set $A_c :=$ &$x : f(x)$ 6 c' is invariant for each c # $\mathbb{R}$. Since T is ergodic, this means that $\mu(A_c)$ is either 0 or 1.

Exercise 39 *Show that this implies that f is constant almost everywhere.*

■

4.2 Conditional Expectation

Let $(X, 2, \mu)$ be a probability space. That is, $\mu(X) = 1$. A measurable function $f : X$ \$ $\mathbb{R}$ between the measurable spaces $(X, 2)$ and $(\mathbb{R}, 1(\mathbb{R}))$ is called a ***random variable***. If f is a random variable, $\mathrm{E}[f]$ will denote & $_X f d\mu$ and we will call this integral the ***expectation*** or ***mean value*** of f.

Definition 40 *Let $f : X$ \$ $\mathbb{R}$ be a real-valued random variable such that* $\mathrm{E}[$<f<$] < \%$ *and let* > *be a sub σ-algebra of* 2, > (2. *The* ***conditional expectation*** *of f with respect to* > *is a* >*-measurable random variable f^* such that*

$$\$_A f d\mu = \$_A f^* d\mu \quad \text{for any } A \ \# > . \tag{4.1}$$

We denote f^* by $\mathrm{E}[f$<>$]$.

The existence of $\mathrm{E}[f$<>$]$ is not a trivial issue and it is based on the Radon-Nikodym's Theorem, one of the most important theorems in measure theory. The explicit computation of $\mathrm{E}[f$<>$]$ can be carried out in some particular situations.

Example 41 Let $\mathcal{G}$ be the σ-algebra generated by a finite partition $A_1, ..., A_n$ of X and suppose that $\mu(A_i) > 0$ for any $i = 1, ..., n$. Then

$$\mathrm{E}[f|\mathcal{G}] = \sum_{i=1}^{n} \frac{\mathrm{E}[f\mathbf{1}_{A_i}]}{\mu(A_i)} \mathbf{1}_{A_i}.$$

In particular if $\mathcal{G} = \{\emptyset, X, A, A^c\}$ is the σ-algebra generated by $A \in \mathcal{F}$, then

$$\mathrm{E}[f|\mathcal{G}] = \frac{\mathrm{E}[f\mathbf{1}_A]}{\mu(A)} \mathbf{1}_A + \frac{\mathrm{E}[f\mathbf{1}_{A^c}]}{\mu(A^c)} \mathbf{1}_{A^c}.$$

Example 42 If $\mathcal{G}$ is the trivial σ-algebra, $\mathcal{G} = \{\emptyset, X\}$, then $\mathrm{E}[f|\mathcal{G}] = \mathrm{E}[f]$ for any random variable f. Indeed, on the one hand, only constants are measurable with respect to the trivial σ-algebra; on the other hand, (4.1) implies the conditional expectation to be equal to the mean value of f.

We review Radon-Nikodym's Theorem for the benefit of a clearer exposition. Recall that given two finite measures μ and ν on a measurable space $(X, \mathcal{F})$, we say that ν is ***absolutely continuous*** with respect to μ if $\mu(A) = 0$ implies $\nu(A) = 0$, where $A \in \mathcal{F}$. We will write $\nu \ll \mu$.

Theorem 43 (Radon-Nikodym's Theorem) *Let μ and ν be two finite measures on $(X, \mathcal{F})$ such that ν is absolutely continuous with respect to μ. Then, there exists an essentially unique measurable function $g : X \to \mathbb{R}$ such that*

$$\nu(A) = \int_A g d\mu. \tag{4.2}$$

*The density g is denoted by $\frac{d\nu}{d\mu}$ and is usually called the **Radon-Nikodym derivative**.*

Essentially unique in Radon-Nykodym's Theorem means that any two functions satisfying (4.2) may only differ on a set of μ-measure 0.

Proposition 44 *With the same notation as in Definition 40, the conditional expectation $\mathrm{E}[f|\mathcal{G}]$ exists and is essentially unique.*

Proof. We continue denoting by μ the restriction of μ to $\mathcal{G}$ and define the measure ν on $\mathcal{G}$ by

$$\nu(A) = \int_A f d\mu, \quad A \in \mathcal{G}.$$

It is clear that ν is absolutely continuous with respect to μ. Its Radon-Nikodym derivative is then the required conditional expectation. The uniqueness follows from the uniqueness statement in the Radon-Nikodym's Theorem. ■

4.2.1 Properties of the conditional expectation

1. **Linearity:** for any two random variables $f, g : X \to \mathbb{R}$ and two real numbers $a, b \in \mathbb{R}$,
$$\mathrm{E}\left[af + bg|\mathcal{G}\right] = a\,\mathrm{E}\left[f|\mathcal{G}\right] + b\,\mathrm{E}\left[g|\mathcal{G}\right].$$
This property follows from the definition of the conditional expectation and that of the integral.

2. **monotonicity:** If $g \leq f$ then $\mathrm{E}\left[g|\mathcal{G}\right] \leq \mathrm{E}\left[f|\mathcal{G}\right]$.

   Exercise 45 Show the monotonicity property using that, if two $\mathcal{G}$-measurable random variables $f_1, f_2 : X \to \mathbb{R}$ satisfy $\mathrm{E}\left[f_1\mathbf{1}_A\right] \leq \mathrm{E}\left[f_2\mathbf{1}_A\right]$ for any $A \in \mathcal{G}$, then $f_1 \leq f_2$.

3. The mean value of a random variable is the same as that of its conditional expectation:
$$\mathrm{E}\left[\mathrm{E}\left[f|\mathcal{G}\right]\right] = \mathrm{E}\left[f\right].$$
This is a consequence of (4.1) with $A = X \in \mathcal{G}$.

4. If $f : X \to \mathbb{R}$ is a $\mathcal{G}$-measurable random variable, then $\mathrm{E}\left[f|\mathcal{G}\right] = f$. Indeed, f is already $\mathcal{G}$-measurable and satisfies (4.1).

5. Two elements $A, B \in \mathcal{F}$ are ***independent*** if $\mu(A \cap B) = \mu(A)\,\mu(B)$. Two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are independent if, for any $A \in \mathcal{F}$ and any $B \in \mathcal{G}$, A and B are independent. We say that a random variable $f : X \to \mathbb{R}$ is independent of a σ-algebra $\mathcal{G}$ if the σ-algebra generated by f, $\mathcal{F} = \sigma\left(\left\{f^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\right\}\right)$, is independent of $\mathcal{G}$. Finally, we say that two random variables are independent it the σ-algebras they generate are independent.

   If $f : X \to \mathbb{R}$ is independent of $\mathcal{G}$, then $\mathrm{E}\left[f|\mathcal{G}\right] = \mathrm{E}\left[f\right]$.

   Exercise 46 Prove this statement using that, if $g, f : X \to \mathbb{R}$ are two independent random variables, $\mathrm{E}\left[fg\right] = \mathrm{E}\left[f\right]\mathrm{E}\left[g\right]$.

6. **Factorization:** If g is a bounded, $\mathcal{G}$-measurable random variable,
$$\mathrm{E}\left[gf|\mathcal{G}\right] = g\,\mathrm{E}\left[f|\mathcal{G}\right].$$

7. If $\mathcal{I}_i$, $i = 1, 2$, are σ-algebras such that with $\mathcal{I}_1 \subset \mathcal{I}_2$,
$$\mathrm{E}\left[\mathrm{E}\left[f|\mathcal{I}_1\right]|\mathcal{I}_2\right] = \mathrm{E}\left[\mathrm{E}\left[f|\mathcal{I}_2\right]|\mathcal{I}_1\right] = \mathrm{E}\left[f|\mathcal{I}_1\right].$$

8. Let f be a random variable independent of $\mathcal{I}$ and let g be a $\mathcal{I}$-measurable random variable. Then, for any measurable function $h : \mathbb{R}^2 \to \mathbb{R}$ such that the random variable
$$\begin{aligned} h(f,g) : X &\longrightarrow \mathbb{R} \\ x &\longmapsto h(f(x), g(x)) \end{aligned}$$
is in $L^1(X, \mu)$, we have
$$\mathrm{E}\left[h(f,g)|\mathcal{I}\right] = \mathrm{E}\left[h(f,x)\right]|_{x=g}.$$
In this expression, $\mathrm{E}[h(f,x)]$ denotes the random variable that, to any fixed $x \in X$, associates the expectation $\mathrm{E}[h(f,x)]$. $\mathrm{E}[h(f,x)]|_{x=g}$ denotes the composition of this random variable with g.

4.3 Birkhoff's Ergodic Theorem

Let $(X, \mathcal{F}, \mu)$ be a probability space and $T : X \to X$ a probability preserving map. The results of this sections are also true if we replace the probability μ with a finite measure, $\mu(X) < \infty$. We define the σ-algebra of T-invariant sets $\mathcal{I} = \sigma\left\{A \in \mathcal{F} : T^{-1}(A) = A\right\}$.

Exercise 47 Show that any $\mathcal{I}$-measurable random variable $f : X \to \mathbb{R}$ is T-invariant.

Theorem 48 (Birkhoff's Ergodic Theorem) *Let $f : X \to \mathbb{R}$ be an integrable random variable (i.e., $\mathrm{E}[|f|] < \infty$). With the notation introduced so far,*
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) = \mathrm{E}[f|\mathcal{I}](x) \quad a.s.. \tag{4.3}$$

Remark 49 *Birkhoff's Ergodic Theorem implies that the limit*
$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x))$$
exists a.s. and, moreover, defines a T-invariant integrable function function because $\mathrm{E}[f|\mathcal{I}]$ is $\mathcal{I}$-measurable (Exercise 47). In the literature, these important consequences of Birkhoff's Ergodic Theorem are sometimes explicitly stated.

Birkhoff's Ergodic Theorem has an important corollary when T is ergodic. Observe that, if T is ergodic, then any subset in the σ-algebra $\mathcal{I}$ of invariant sets has probability either 0 or 1. Roughly speaking, one might think of $\mathcal{I}$ as the trivial σ-algebra so, by Example 42, $\mathrm{E}[f|\mathcal{I}] = \mathrm{E}[f]$. We give a rigorous proof of this fact in the next corollary:

Corollary 50 *If T is ergodic with respect to μ,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) = \mathrm{E}[f] \quad a.s..$$

Proof. Let $f^* = \mathrm{E}[f|\mathcal{I}]$ and define the sets $A_+ := \{x \in X : f^*(x) > \mathrm{E}[f]\}$, $A_0 := \{x \in X : f^*(x) = \mathrm{E}[f]\}$, and $A_- := \{x \in X : f^*(x) < \mathrm{E}[f]\}$. These three sets are T-invariant (f^* is T-invariant by Exercise 47) and, therefore, belong to $\mathcal{I}$. They form a partition of X and, consequently, exactly one of them must have measure 1 and the other two probability 0. If $\mu(A_+) = 1$, then $\mathrm{E}[f^*] = \int_{A_+} f^* d\mu$ and, using the monotonicity of the integral,

$$\mathrm{E}\left[f^*\right] = \int_{A_+} f^* d\mu > \int_{A_+} \mathrm{E}[f]\, d\mu = \mathrm{E}[f]\, \mu(A_+) = \mathrm{E}[f],$$

which is clearly a contradiction because, by definition, $\mathrm{E}[f^*] = \mathrm{E}[f]$. Similarly, $\mu(A_-)$ must also be 0. Consequently, $\mu(A_0) = 1$ and $f^* = \mathrm{E}[f]$ a.s..

■

Corollary 50 is often referred to as Birkhoff's Theorem in the literature. Its physical interpretation is the following. An integrable function $f : X \to \mathbb{R}$ is sometimes called an *observable* since it can be thought of as the result of a *measurement* which depends on the point x of the phase space X at which f is evaluated. The integral $\int_X f d\mu$ is sometimes called the ***space average*** of f (with respect to the measure μ) whereas, for a given point $x \in X$, the averages $\frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x))$ are often referred to as the ***time averages*** of f along the orbit of x. Corollary 50 claims that, when μ is ergodic, *time averages converge to space averages.*

In order to prove Theorem 48, we need an auxiliary result.

Lemma 51 (Maximal Ergodic Theorem) *Define*

$$S_N(x) = \sum_{n=0}^{N-1} f(T^n(x)) \quad and\ M_N(x) := \max\{S_0(x), ..., S_N(x)\}$$

with the convention $S_0 = 0$. Then $\int_{\{M_N>0\}} f d\mu \geq 0$.

Proof. For every $0 \leq k \leq N$ and every $x \in X$, by definition, one has $M_N(T(x)) \geq S_k(T(x))$ and $f(x) + M_N(T(x)) \geq f(x) + S_k(T(x)) = S_{k+1}(x)$. Therefore,

$$f(x) \geq \max\{S_1(x), ..., S_N(x)\} - M_N(T(x)).$$

Furthermore, $\max\{S_1(x), ..., S_N(x)\} = M_N(x)$ on the set $\{M_N > 0\}$, so that

$$\begin{aligned}\int_{\{M_N>0\}} f d\mu &\geq \int_{\{M_N>0\}} (M_N - M_N \circ T)\, d\mu \\ &\geq \mathbb{E}[M_N] - \int_{\{M_N>0\}} (M_N \circ T) d\mu \end{aligned} \tag{4.4}$$

where $\int_{\{M_N>0\}} M_N d\mu \geq \mathbb{E}[M_N]$ because $M_N \geq 0$. Now,

$$\begin{aligned}\int_{\{M_N>0\}} (M_N \circ T) d\mu &= \int_X \mathbf{1}_{\{M_N>0\}} (M_N \circ T) d\mu = \int_X \mathbf{1}_{\{T(x) \circ M_N(x)>0\}} M_N d(T_* \mu) \\ &= \int_{\{T(x) \circ M_N(x)>0\}} M_N d\mu\end{aligned}$$

because T is measure-preserving. Since $M_N \geq 0$, $\int_B M_N d\mu \leq \mathbb{E}[M_N]$ for any $B \in \mathcal{B}$, so that (4.4) implies

$$\int_{\{M_N>0\}} f d\mu \geq \mathbb{E}[M_N] - \int_{\{T(x) \circ M_N(x)>0\}} (M_N \circ T) d\mu \geq 0,$$

which is the required result. ■

Proof of Theorem 48. First of all, observe that proving (4.3) is equivalent to proving that

$$0 = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) - \mathbb{E}[f|\mathcal{I}] = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} (f - \mathbb{E}[f|\mathcal{I}])(T^n(x))$$

where we have used that $\mathbb{E}[f|\mathcal{I}]$ is T-invariant. Therefore, replacing f by $f - \mathbb{E}[f|\mathcal{I}]$ in the statement of Birkhoff's Ergodic Theorem, we can assume without loss of generality that $\mathbb{E}[f|\mathcal{I}] = 0$. Define $\overline{S} = \limsup_{n \to \infty} S_n/n$ and $\underline{S} = \liminf_{n \to \infty} S_n/n$. We want to show that $\overline{S} = \underline{S} = 0$. It is enough to show that $\overline{S} \leq 0$ a.s. since this implies (by considering $-f$ instead of f) that $\underline{S} \geq 0$. Therefore $0 \leq \underline{S} \leq \overline{S} \leq 0$, which means $\overline{S} = \underline{S} = 0$ a.s..

It is clear that $\overline{S}(T(x)) = \overline{S}(x)$ for every $x \in X$, so that, if $\varepsilon > 0$, one has $A^\varepsilon := \left\{x \in X : \overline{S}(x) > \varepsilon\right\} \in \mathcal{I}$. That is, A^ε belongs to the σ-algebra $\mathcal{I}$ of T-invariant sets. We want to show that $\mu(A^\varepsilon) = 0$. Define

$$f^\varepsilon := (f - \varepsilon)\,\mathbf{1}_{A^\varepsilon},$$

and S_N^ε and M_N^ε according to Lemma 51. With these definitions, we have

$$\frac{S_N^\varepsilon}{N} = \begin{cases} 0 & \text{if } \overline{S}(x) \leq \varepsilon \\ \frac{S_N}{N} - \varepsilon & \text{otherwise.} \end{cases} \tag{4.5}$$

The sequence of sets $\{M_N^\varepsilon > 0\}$ increases to the set $B^\varepsilon := \{\sup_N S_N^\varepsilon > 0\} = \{\sup_N \frac{S_N^\varepsilon}{N} > 0\}$. From (4.5),

$$\sup_N \frac{S_N^\varepsilon(x)}{N} > 0 \iff \exists N \in \mathbb{N} : \frac{S_N}{N} - \varepsilon > 0 \iff \overline{S}(x) > \varepsilon.$$

Therefore

$$B^\varepsilon = \left\{\sup_N \frac{S_N}{N} > \varepsilon\right\} = \left\{\overline{S} > \varepsilon\right\} = A^\varepsilon.$$

Now, on the one hand, $\int_{\{M_N^\varepsilon > 0\}} f^\varepsilon d\mu \geq 0$ for any $N \geq 1$ from Lemma 51; on the other hand, $\mathrm{E}[|f^\varepsilon|] \leq \mathrm{E}[|f|] + \varepsilon < \infty$. In this situation, the Dominated Convergence Theorem implies that

$$0 \leq \lim_{N \to \infty} \int_{\{M_N^\varepsilon > 0\}} f^\varepsilon d\mu = \int_{A^\varepsilon} f^\varepsilon d\mu$$

and, therefore,

$$\begin{aligned} 0 \leq \int_{A^\varepsilon} f^\varepsilon d\mu &= \int_{A^\varepsilon} (f - \varepsilon)\, d\mu = \int_{A^\varepsilon} f d\mu - \varepsilon\mu(A^\varepsilon) \\ &= \int_{A^\varepsilon} \mathrm{E}[f|\mathcal{I}]\, d\mu - \varepsilon\mu(A^\varepsilon) = -\varepsilon\mu(A^\varepsilon) \end{aligned}$$

because $A^\varepsilon \in \mathcal{I}$ and we assumed that $\mathrm{E}[f|\mathcal{I}] = 0$. In conclusion, one must have $\mu(A^\varepsilon) = 0$ for any $\varepsilon > 0$, which implies that $\overline{S} \leq 0$ almost surely. ■

Corollary 52 *Let $T : X \to X$ be a measurable transformation and μ a T-invariant ergodic probability. Then, for any $A \in \mathcal{B}$*

$$\frac{\#\left\{1 \leq j \leq N : T^j(x) \in A\right\}}{N} \xrightarrow[N \to \infty]{} \mu(A) \quad a.s..$$

Proof. It is a straightforward consequence of Birkhoff's Ergodic Theorem applied to the characteristic function $f = \mathbf{1}_A$. ■

Example 53 Dirac measures on fixed points and periodic orbits. Let $T : X \ \$ \ X$ be a measurable transformation and let $P = \&a_1, ..., a_n$' be a periodic orbit. Let $\delta_P = \frac{1}{n} \ {}^{n}_{i=1} \delta_{a_i}(A)$ be the Dirac measure uniformly distributed on P (see Subsection 3.1.1). We already know that δ_P is T-invariant.

Proposition 54 *δ_P is ergodic.*

Proof. If $P = \&a_1$' is a fixed point, the statement is trivial because every measurable set $A \# 2$ has measure 0 or 1 with respect to δ_{a_1}. In particular, this is true for any backward invariant set. If P is a periodic orbit with $n \ 9 \ 2$ points, then every measurable set A such that $T^{\#1}(A) = A$ must contain either all points of P or none of them. Therefore A has measure either 0 or 1. ■

Now, let $p \neq q$ be two fixed points for T and define the measure

$$\mu = \frac{1}{2} (\delta_p + \delta_q) .$$

Proposition 55 *μ is not ergodic.*

Proof. Consider the set $A = {}^{\cdot}{}_{n\%\mathbb{N}} T^{\#1}(p)$ of all the preimages of the point p. Then clearly $T^{\#1}(A) = A$. Moreover, $q \not\# A$ since q is a fixed point and therefore cannot be sent to p under a forward iteration. Therefore $\mu(A) = 1/2$ and μ is not ergodic. ■

4.4 Structure of the set of invariant measures

Let X be a topological space and $T : X \ \$ \ X$ a measurable transformation. Recall that A denotes the space of all Borel probabilities on $(X, 1(X))$. The larger space of finite Borel measures is a vector space since for any two measures μ_1 and μ_2 and any two scalars $a, b \# \mathbb{R}$ we have that $a\mu_1 + b\mu_2$ also defines a finite measure. Let A_T " A be the subset of all T-invariant Borel probability measures on X. A subset of a linear space is ***convex*** if, for any $t \# [0, 1]$ and every $\mu_1, \mu_2 \#$ A we have $t\mu_1 + (1 \ ! \ t)\mu_2 \#$ A.

Exercise 56 A and A_T are convex.

Moreover, $\mathcal{A}_T \subset \mathcal{A}$ is closed. If, additionally, $\mathcal{A}$ is compact (for instance if X is compact, see Lemma 31) then $\mathcal{A}_T$ is also compact in the weak* topology. We say that $\mu \in \mathcal{A}_T$ is an ***extremal point*** of $\mathcal{A}_T$ if it cannot be written as a linear combination of any two other points of $\mathcal{A}_T$, i.e., if $\mu = t\mu_1 + (1 - t)\,\mu_2$ for some $\mu_1, \mu_2 \in \mathcal{A}_T$, then necessarily $t = 0$ or 1. The Krein-Milman Theorem claims that a convex set is the convex hull of its extremal points. In particular, a convex set has always extremal points. As we will see in Proposition 59, ergodic probabilities correspond to extremal points and, therefore, the existence of ergodic measures is always guaranteed provided that $\mathcal{A}_T \neq \emptyset$. For example, if X is compact and $T : X \to X$ is continuous, then $\mathcal{A}_T \neq \emptyset$ by Krylov-Bogolubov's Theorem (Theorem 29). Consequently,

Proposition 57 *If X is compact and $T : X \to X$ is continuous, then there exists one ergodic probability at least.*

We say that two measures μ and ν are ***equivalent***, and we will write $\mu = \nu$, if $\mu(A) = \nu(A)$ for any $A \in \mathcal{B}$. In particular, this means that $\int_X f d\mu = \int_X f d\nu$ for any bounded measurable function $f : X \to \mathbb{R}$. We say that two measures μ and ν are ***mutually singular*** if there exists a measurable set $A \in \mathcal{B}$ such that $\mu(A) = 1$ and $\nu(A) = 0$.

Lemma 58 *If $\mu_1, \mu_2 \in \mathcal{A}_T$ are ergodic measures such that $\mu_1 \ll \mu_2$ then $\mu_1 = \mu_2$.*

Proof. Let $f : X \to \mathbb{R}$ be an arbitrary bounded measurable function (and thus in particular integrable with respect to any invariant probability measure). Since μ_2 is ergodic, by Birkhoff's Ergodic Theorem,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f\left(T^n(x)\right) = \int_X f d\mu_2$$

for any $x \in \Omega$ on a measurable set of full μ_2-measure, i.e., $\mu_2(\Omega) = 1$. Since $\mu_1 \ll \mu_2$ and $\mu_2(\Omega^c) = 0$, we have $\mu_1(\Omega^c) = 0$ and, consequently, $\mu_1(\Omega) = 1$ as well. Therefore,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f\left(T^n(x)\right) = \int_X f d\mu_2 \quad \mu_1\text{-a.s..}$$

However, applying Birkhoff's Ergodic Theorem to μ_1, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f\left(T^n(x)\right) = \int_X f d\mu_1 \quad \mu_1\text{-a.s..}$$

In other words,

$$\int_X f d\mu_2 = \int_X f d\mu_1 \tag{4.6}$$

for any measurable bounded function f. Writing (4.6) for a characteristic function $\mathbf{1}_A$, we obtain $\mu_1(A) = \mu_2(A)$ for any measurable set $A \in \mathsf{2}$ and $\mu_1 = \mu_2$. ■

Proposition 59 *$\mu \in \mathsf{A}_T$ is ergodic if and only if it is an extremal point of A_T.*

Proof. $\mu \in \mathsf{A}_T$ extremal point $\Longrightarrow$ $\mu \in \mathsf{A}_T$ ergodic. Suppose that μ is not ergodic. Then there exists a T-invariant measurable subset $A \subset X$ with $0 < \mu(A) < 1$. Define the measures $\mu_A(B) = \mu(B \cap A)/\mu(A)$ and $\mu_{X \setminus A}(B) = \mu(B \cap (X \setminus A))/\mu(X \setminus A)$ where $B \in \mathsf{2}$. Then μ_A and $\mu_{X \setminus A}$ are T-invariant and

$$\mu = \mu(A)\,\mu_A + \mu(X \setminus A)\,\mu_{X \setminus A},$$

so μ is not an extreme point.

$\mu \in \mathsf{A}_T$ ergodic $\Longrightarrow$ $\mu \in \mathsf{A}_T$ extremal point. Suppose by contradiction that μ is not extremal so that $\mu = t\mu_1 + (1 - t)\mu_2$ for two invariant different probability measures $\mu_1, \mu_2 \in \mathsf{A}_T$ and some $t \in (0,1)$. Since $\mu(A) = 0$ always implies $\mu_1(A) = 0$ and $\mu_2(A) = 0$ both μ_1 and μ_2 are absolutely continuous with respect to μ and, moreover, they are ergodic. Indeed, if $\mu(A) = 1$, then necessarily $\mu_1(A) = \mu_2(A) = 1$. Therefore, by Lemma 58, we have $\mu_1 = \mu = \mu_2$ contradicting thus our assumption. ■

Ergodic measures are not only extremal points but also mutually singular each other.

Proposition 60 *Let μ_1 and μ_2 be distinct ergodic invariant measures. Then μ_1 and μ_2 are mutually singular.*

Proof. By Lemma 58, μ_1 and μ_2 cannot be absolutely continuous. Therefore, there exists a measurable set E such that $\mu_1(E) > 0$ and $\mu_2(E) = 0$. Define

$$A = \bigcap_{m=0}^{\infty} \bigcup_{j=m}^{\infty} T^{-j}(E).$$

We will show that $\mu_1(A) = 1$ and $\mu_2(A) = 0$ which will imply that μ_1 and μ_2 are mutually singular.

First, we claim that $T^{-1}(A) = A$. Indeed, if $x \in A$ then $x \in \bigcup_{j=m}^{\infty} T^{-j}(E)$ for any $m \in \mathbb{N}$. That is, $T^j(x) \in E$ for infinitely many values of $j \in \mathbb{N}$.

If x satisfies this property, then so do $T^{-1}(x)$ and $T(x)$, which implies $T^{-1}(A) = A$. Therefore, it is sufficient to show that $\mu_1(A) > 0$ to imply $\mu_1(A) = 0$ by the ergodicity of μ_1.

By the invariance of both μ_1 and μ_2 we have

$$\mu_1\left(\bigcap_{j=0}^{\infty} T^{-j}(E)\right) \geq \mu_1(E) > 0 \tag{4.7}$$

and

$$\mu_2\left(\bigcap_{j=0}^{\infty} T^{-j}(E)\right) = 0. \tag{4.8}$$

Observe that (4.8) implies that $\mu_2(A) = 0$. On the other hand,

$$\bigcap_{j=m}^{\infty} T^{-j}(E) = T^{-m}\left(\bigcap_{j=0}^{\infty} T^{-j}(E)\right)$$

and, consequently,

$$\mu_i\left(\bigcap_{j=m}^{\infty} T^{-j}(E)\right) = \mu_i\left(T^{-m}\left(\bigcap_{j=0}^{\infty} T^{-j}(E)\right)\right) = \mu_i\left(\bigcap_{j=0}^{\infty} T^{-j}(E)\right)$$

for $i = 1, 2$. In particular, the measure of each $\bigcap_{j=m}^{\infty} T^{-j}(E)$ is constant. Since the sets $\bigcap_{j=m}^{\infty} T^{-j}(E)$ are nested, i.e.,

$$\bigcap_{j=m+1}^{\infty} T^{-j}(E) \subseteq \bigcap_{j=m}^{\infty} T^{-j}(E),$$

A is the countable intersection of a nested sequence of sets all of them with the same measure (strictly positive by (4.7)). It follows that $\mu_1(A) > 0$ and $\mu_2(A) = 0$ as required.

Exercise 61 Prove this last sentence.

■

Chapter 5

Circle rotations

In this chapter, we are going to deal with a very important example, that of circle rotations

$$\begin{aligned} T : \mathbb{S}^1 &\longrightarrow \mathbb{S}^1 \\ x &\longmapsto x + \alpha, \quad \alpha \in \mathbb{R}. \end{aligned} \tag{5.1}$$

More concretely, we are going to prove that the *Lebesgue measure* λ on $\left(\mathbb{S}^1, \mathcal{B}(\mathbb{S}^1)\right)$ is ergodic if and only if α is a irrational multiple of 2π, i.e., $\alpha \neq 2\pi\frac{m}{n}$, $m, n \in \mathbb{Z}$. Indeed if $\alpha = 2\pi\frac{m}{n}$ with $m, n \in \mathbb{Z}$ such that m and n have no common factors, then $T^n = \mathrm{Id}$ is the identity and, for any $x \in \mathbb{S}^1$, $O_x = \left\{x, T(x), ..., T^{n-1}(x)\right\}$ is a periodic orbit of period n. Then any set built as a family of arcs $\left\{B_\varepsilon\left(T^i(x)\right)\right\}_{i=0...n-1}$ of length $\varepsilon > 0$ centered at the points of an orbit O_x is invariant and of strictly positive Lebesgue measure. Therefore, the Lebesgue measure is not ergodic. Furthermore, according to Example 53, the Dirac measure supported on O_x

$$\delta_{O_x} = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i(x)}$$

is ergodic. Therefore, a rational rotation admits infinitely many ergodic measures.

5.1 Irrational case

We will prove that the Lebesgue measure is ergodic when α in (5.1) is an irrational multiple of 2π in two different ways. The first proof is rather easy and uses Fourier analysis. The second one is longer and requires some non-trivial results such as the Lebesgue Density Theorem. Nevertheless, this theorem is quite important in order to prove the ergodicity of some concrete maps, hence the reason why we choose to prove the ergodicity of an irrational rotation by this slightly more sophisticated way. To start with, we will give the easier proof.

Proposition 62 *The Lebesgue measure is ergodic with respect to the rotation (5.1) if and only if α is an irrational multiple of 2π.*

Proof. By Proposition 38, it is enough to prove that any T-invariant $f \in L^2\left(\mathbb{S}^1, \lambda\right)$ is constant a.e.. Identify $\mathbb{S}^1$ with the unit interval $[0,1]/_{0\sim 1}$ and think of α as an irrational number between 0 and 1. The Fourier series $\sum_{n=-\infty}^{n=\infty} a_n \mathrm{e}^{2n\pi ix}$ of f converges to f in the L^2 norm. The series $\sum_{n=-\infty}^{n=\infty} a_n \mathrm{e}^{2n\pi i(x+\alpha)}$ converges to $f \circ T$. Since $f = f \circ T$ a.e., uniqueness of Fourier coefficients implies that $a_n = a_n \mathrm{e}^{2n\pi i\alpha}$ for all $n \in \mathbb{Z}$. Since $\mathrm{e}^{2n\pi i\alpha} \neq 1$ for $n \neq 0$, we conclude that $a_n = 0$ for $n \neq 0$, so f is constant a.e..

The converse is immediate because if $\alpha = \frac{n}{m}$ is rational, $n, m \in \mathbb{Z}$, then we already showed that the Lebesgue measure is not ergodic. ■

For the alternative proof, we need first to define new concepts and give additional results.

Lemma 63 ([9, Lemma 7.3]) *Let $W \subset \mathbb{R}^n$ be a measurable set that is contained in a finite union of balls $B_{r_i}(x_i)$ where $x_i \in \mathbb{R}^n$ and $r_i > 0$, $i = 1, ..., N$. Then there is a set $S \subset \{1, ..., N\}$ so that*

(a) the balls $B_{r_i}(x_i)$ with $i \in S$ are disjoint,

(b) $W \subset \bigcup_{i\in S} B_{3r_i}(x_i)$, and

(c) $\lambda(W) \leq 3^n \sum_{i\in S} \lambda(B_{r_i}(x_i))$.

Proof. Order the balls $B_i = B_{r_i}(x_i)$ so that $r_1 \geq r_2 \geq ... \geq r_N$. Put $i_1 = 1$. Discard all B_j that intersect B_{i_i}. Let B_{i_2} be the first of the remaining B_j if there are any. Discard all B_j with $j > i_2$ that intersect B_{i_2}, let B_{i_3} be the first of the remaining ones, and so on as long as possible. This process stops after a finite number of steps and gives $S = \{i_1, i_2, ...\}$. It is clear that (a) holds. Every discarded B_j is a subset of $B_{3r_i}(x_i)$ for some $i \in S$, for if $r' < r$ and $B_{r'}(x')$ intersects $B_r(x)$, then $B_{r'}(x') \subset B_r(x)$. This proves (b). (c) follows from (b) because, in $\mathbb{R}^n$,

$$\lambda\left(B_{3r}(x)\right) = 3^n \lambda\left(B_r(x)\right).$$

■

Corollary 64 *Let $T:\mathbb{S}^1\to\mathbb{S}^1$ be an irrational circle rotation. Then, for every $x\in\mathbb{S}^1$, there exists a sequence of arc neighbourhoods J_n of x with $\lambda(J_n)\to 0$ as $n\to\infty$ and a sequence of finite sets $S_n\subset\mathbb{N}$ such that*

1. $\mathbb{S}^1\subseteq\bigcup_{i\in S_n}T^i(J_n)$;
2. $\sum_{i\in S_n}\lambda\left(T^i(J_n)\right)\leq 3\left(1+\frac{2}{n}\right)$.

Observe that while the first statement is relatively intuitive, the second is highly non-trivial beforehand. Nevertheless, this is a consequence of Lemma 63. The number three is not as important as the fact that there exists a bound on how much the intervals $T^i(J_n)$ can overlap, so that we can give a bound uniform in n. This will be crucially used at the end of the proof of Proposition 68.

Proof of Corollary 64. Identify $\mathbb{S}^1$ with $[0,1]/0\sim 1$ and define the projection

$$\begin{aligned}\pi:\mathbb{R}&\longrightarrow[0,1]/0\sim 1\\ z&\longmapsto[z]\end{aligned}$$

that send any real number to its equivalent class in $[0,1]/0\sim 1$. Let $z\in[0,1]$ and $n\in\mathbb{N}$. Take $J_n:=\pi\left(B_{3/n}(z)\right)$ as the image by π of the open ball of radius $\frac{3}{n}$ centered at z. $\{J_n\}_{n\in\mathbb{N}}$ defines a sequence of arc neighbourhoods of $x=[z]$ such that $\lambda(J_n)=\frac{6}{n}\to 0$ as $n\to 0$.

Exercise 65 *Using that the orbit $\{T^i(x)\}_{i\in\mathbb{N}}$ of x is dense in $\mathbb{S}^1$ for an irrational rotation (Theorem 27), prove that there exists a finite subset $I=\{i_1,...,i_N\}\subset\mathbb{N}$ such that $[0,1]\subseteq\bigcup_{i\in I}B_{1/n}(T^i(x))$.*

Now, we have

$$[0,1]\subseteq\bigcup_{i\in I}B_{1/n}\left(T^i(x)\right).$$

By Corollary 64, there exists a finite set $S\subset I$ such that

$$[0,1]\subseteq\bigcup_{i\in S}B_{3/n}\left(T^i(x)\right)$$

where the balls $B_{1/n}\left(T^i(x)\right)$ are disjoint. Moreover, $\pi\left(B_{3/n}\left(T^i(x)\right)\right)=T^i(J_n)$ and

$$\mathbb{S}^1\subseteq\bigcup_{i\in S}\pi\left(B_{3/n}\left(T^i(x)\right)\right)=\bigcup_{i\in S}T^i(J_n).$$

On the other hand,

$$\sum_{i\in S}\lambda\left(T^i(J_n)\right) = \sum_{i\in S}\lambda\left(\pi\left(B_{3/n}\left(T^i(x)\right)\right)\right)$$
$$\leq \sum_{i\in S}\lambda\left(B_{3/n}\left(T^i(x)\right)\right) \leq 3\sum_{i\in S}\lambda\left(B_{1/n}\left(T^i(x)\right)\right). \tag{5.2}$$

Observe now that the union $\bigcup_{i\in S} B_{1/n}\left(T^i(x)\right)$ is contained in the open interval $(-\frac{1}{n}, 1+\frac{1}{n})$. Since the balls $B_{1/n}\left(T^i(x)\right)$, $i\in S$, are disjoint,

$$\sum_{i\in S}\lambda\left(B_{1/n}\left(T^i(x)\right)\right) = \lambda\left(\bigcup_{i\in S} B_{1/n}\left(T^i(x)\right)\right) \leq 1+\frac{2}{n}.$$

Therefore, (5.2) implies

$$\sum_{i\in S}\lambda\left(T^i(J_n)\right) \leq 3\left(1+\frac{2}{n}\right).$$

■

If $f\in L^1(\mathbb{R}^n,\lambda)$, $n\in\mathbb{N}$, we say that $x\in\mathbb{R}^n$ is a ***Lebesgue point*** of f if

$$f(x) = \lim_{r\to 0}\frac{1}{B_r(x)}\int_{B_r(x)} f\,d\lambda.$$

It is probably far from obvious that every $f\in L^1(\mathbb{R}^n,\lambda)$ has Lebesgue points. But the following remarkable theorem, which we are not going to prove, shows that they always exist. The reader is encouraged to check with [9].

Theorem 66 ([9, Theorem 7.7]) *If $f\in L^1(\mathbb{R}^n,\lambda)$, then almost every $x\in\mathbb{R}^n$ is a Lebesgue point of f.*

One of the most important corollaries of Theorem 66 is what in the literature is sometimes referred to as *Lebesgue's Density Theorem*. This result gives us information about the *density of (Lebesgue) measure* on almost every point of a measurable set.

Corollary 67 (Lebesgue's Density Theorem) *Let $A\in\mathcal{L}(\mathbb{R}^n)$ be a Lebesgue measurable subset of $\mathbb{R}^n$ with positive measure, $\lambda(A)>0$. Then for λ-almost every point $x\in A$,*

$$\lim_{r\to 0}\frac{\lambda(A\cap B_r(x))}{\lambda(B_r(x))} = 1. \tag{5.3}$$

Proof. This result is a consequence of Theorem 66 applied to the characteristic function $f = \mathbf{1}_A$. ■

This result says that in some very subtle way. A priori, one may expect that if $\lambda(A) = 1/2$, then for any subinterval J the ratio between $A \cap J$ and J might be $1/2$, i.e., that the ratio between the measure of the whole interval and the measure of the set A is constant at every scale. This theorem shows that this is not the case. Points $x \in A$ for which (5.3) holds are called ***Lebesgue's density points***.

We are now ready to tackle the prove of ergodicity of Lebesgue measure for irrational circle rotations. Similar arguments of those used in the proof of Proposition 68 will be used later.

Proposition 68 *If $\alpha/2\pi$ is irrational then Lebesgue measure is ergodic.*

Proof. Let $A \in \mathcal{B}\left(\mathbb{S}^1\right)$ satisfy $T^{-1}(A) = A$ and $\lambda(A) > 0$. We want to show that $\lambda(A) = 1$. By Lebesgue's Density Theorem, λ-almost every point of A is a Lebesgue density point of A. Let $x \in A$ be one of such points and fix an arbitrary $\varepsilon > 0$. Choose $n_\varepsilon \in \mathbb{N}$ large enough so that

$$\lambda(A \cap J_{n_\varepsilon}) \geq (1 - \varepsilon)\,\lambda(J_{n_\varepsilon}) \tag{5.4}$$

where J_{n_ε} is a sufficiently small arc neighbourhood of x as in Corollary 64. We shall make three simple statements which combined will give us the desired result. First of all, observe that (5.4) is equivalent to

$$\frac{\lambda(J_{n_\varepsilon} \setminus A)}{\lambda(J_{n_\varepsilon})} \leq \varepsilon. \tag{5.5}$$

Secondly, since T is just a translation and Lebesgue measure is invariant by translations, we have $\lambda\left(T^i(J_{n_\varepsilon})\right) = \lambda(J_{n_\varepsilon})$ and $\lambda\left(T^i(J_{n_\varepsilon} \setminus A)\right) = \lambda(J_{n_\varepsilon} \setminus A)$ for any $i \in \mathbb{N}$ (these equalities stem from the fact that λ is invariant by T^{-1}, which is again a rotation of angle $-\alpha$). In particular,

$$\frac{\lambda\left(T^i(J_{n_\varepsilon} \setminus A)\right)}{\lambda(T^i(J_{n_\varepsilon}))} = \frac{\lambda(J_{n_\varepsilon} \setminus A)}{\lambda(J_{n_\varepsilon})}. \tag{5.6}$$

In third place, using the invariance of A ($T^{-1}(A) = A$ which, in turn, implies $T(A) = A$) and the fact that $\mathbb{S}^1 \subseteq \bigcup_{i \in S_{n_\varepsilon}} T^i(J_{n_\varepsilon})$ we have

$$\mathbb{S}^1 \setminus A \subseteq \left(\bigcup_{i \in S_{n_\varepsilon}} T^i(J_{n_\varepsilon})\right) \setminus A = \bigcup_{i \in S_{n_\varepsilon}} \left(T^i(J_{n_\varepsilon}) \setminus A\right) = \bigcup_{i \in S_{n_\varepsilon}} T^i(J_{n_\varepsilon} \setminus A)$$

so

$$\lambda\left(\mathbb{S}^1\setminus A\right) \leq \sum_{i\in S_{n_\varepsilon}} \lambda\left(T^i\left(J_{n_\varepsilon}\setminus A\right)\right), \tag{5.7}$$

Now, from (5.5), (5.6), and (5.7)

$$\begin{aligned}\lambda\left(\mathbb{S}^1\setminus A\right) &\leq \sum_{i\in S_{n_\varepsilon}} \lambda\left(T^i\left(J_{n_\varepsilon}\setminus A\right)\right) \leq \sum_{i\in S_{n_\varepsilon}} \frac{\lambda\left(J_{n_\varepsilon}\setminus A\right)}{\lambda\left(J_{n_\varepsilon}\right)}\lambda\left(T^i\left(J_{n_\varepsilon}\right)\right) \\ &= \frac{\lambda\left(J_{n_\varepsilon}\setminus A\right)}{\lambda\left(J_{n_\varepsilon}\right)} \sum_{i\in S_{n_\varepsilon}} \lambda\left(T^i\left(J_{n_\varepsilon}\right)\right) \leq 3\varepsilon\left[1+\frac{2}{n_\varepsilon}\right].\end{aligned}$$

by Corollary 64. Since ε is arbitrary, this means that $\lambda\left(\mathbb{S}^1\setminus A\right)=0$ so $\lambda(A)=1$. ■

Chapter 6

Central Limit Theorem

In this chapter we will state a Central Limit Theorem for the random variables $f \circ T^n$ built from an observable $f \in L^1(X, \mu)$ and an ergodic map $T : X \to X$. This Central Limit Theorem is a firs step to give confidence intervals for an estimation of $\mathrm{E}[f]$ by means of Birkhoff's Ergodic Theorem,

$$\frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n(x).$$

The Central Limit Theorem will only hold for mixing maps.

6.1 Mixing maps

Definition 69 *Let $T : X \to X$ be a measurable transformation on a measure space $(X, \mathcal{B}, \mu)$ that preserves μ. We say that T is mixing if, for any two sets $A, B \in \mathcal{B}$ such that $\mu(A) > 0$ and $\mu(B) > 0$, we have*

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

There are two natural interpretations of mixing, one geometrical and one probabilistic. From a geometrical point of view (recall that $\mu\left(T^{-1}(B)\right) = \mu(B)$) one can think of $T^{-n}(B)$ as a *redistribution of mass*. The mixing condition then says that for large n the proportion of $T^{-n}(B)$ which intersects A is just proportional to the measure of A. In other words $T^{-n}(B)$ is spreading itself uniformly with respect to the measure. A more probabilistic point of view is to think of $\mu(T^{-n}(B) \cap A)/\mu(B)$ as the conditional probability of having $x \in A$ given that $T^n(x) \in B$, i.e. the probability that the occurrence of the event B today is a consequence of the occurrence of the event A n steps in the past. The mixing condition then says that this probability converges to the probability of A, i.e., asymptotically, there is no causal relation between the two events.

A classical example by Arnold and Avez ([1]) explains what a mixing map does. Suppose a cocktail shaker X, $\mu(X) = 1$ is filled by 85% of lemon juice and 15% of vodka. Let A be the part of the cocktail shaker originally occupied by the vodka and B any part of the shaker. Let $T^{-1} : X \to X$ be the transformation of the content of the shaker made during one move by the waiter (who is shaking the cocktail repeatedly and redistributes the volume of the two liquids). Then after n moves the fraction of juice in the part B is $\mu(T^{-n}(A) \cap B)/\mu(B)$. As the waiter keeps shaking the cocktail ($n \to \infty$), the fraction of vodka in any part B approaches $\mu(A) = 0.15$, i.e. the vodka spreads uniformly in the mixture.

Proposition 70 *Let $(X, \mathcal{A}, \mu)$ be a probability space. Any mixing map $T : X \to X$ is ergodic.*

Proof. Let $A \in \mathcal{A}$ be any T-invariant measurable set. Then $T^{-n}(A) = A$ and

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

In particular, for $A = B$ we have $\mu(A) = \mu(A)^2$. This means $\mu(A) = 0$ or $\mu(A) = 1$, hence T is ergodic. ■

Proposition 71 *Suppose that $T : X \to X$ is mixing. Then, for any $f, g \in L^2(X, \mu)$,*

$$\lim_{n \to \infty} \int g\,(f \circ T^n)\,d\mu = \int g\,d\mu \int f\,d\mu. \tag{6.1}$$

Proof. Equation (6.1) trivially holds for characteristic functions. Indeed, if $g = \mathbf{1}_A$ and $f = \mathbf{1}_B$ for some sets $A, B \in \mathcal{A}$, then

$$\begin{aligned}\int g\,(f \circ T^n)\,d\mu &= \int \mathbf{1}_A\,(\mathbf{1}_B \circ T^n)\,d\mu = \int \mathbf{1}_A \mathbf{1}_{T^{-n}(B)}\,d\mu = \int \mathbf{1}_{A \cap T^{-n}(B)}\,d\mu \\ &= \mu\left(A \cap T^{-n}(B)\right) \xrightarrow[n \to \infty]{} \mu(A)\,\mu(B) = \int \mathbf{1}_A\,d\mu \int \mathbf{1}_B\,d\mu.\end{aligned}$$

(6.1) is obviously true for elementary functions as well. The general result follows approximating two arbitrary functions f and g by sequences of elementary functions. ■

6.2 Central Limit Theorem

In probability, the ***Strong Law of Large Numbers*** claims that if $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of independent random variables defined on a probability space $(\Omega, \mathcal{B}, P)$ that are identically distributed, then

$$\frac{X_1 + ... + X_n}{n} \xrightarrow[n\to\infty]{} m \ \text{ a.s.,}$$

where $m = \mathrm{E}[X_1]$ is the common expectation. We usually say that $\{X_n\}_{n\in\mathbb{N}}$ is and ***i.i.d. sequence***. Observe that, whenever we have an ergodic map $T : X \to X$ on a probability space $(X, \mathcal{B}, \mu)$, for any $f \in L^1(X, \mu)$, the sequence of random variables $X_n := f \circ T^n$ satisfies the Strong Law of Large Numbers by Birkhoff's Ergodic Theorem,

$$\frac{1}{n}\sum_{i=0}^{n-1} f\left(T^i(x)\right) \xrightarrow[n\to\infty]{} \mathrm{E}[f] \ \text{ a.s..} \tag{6.2}$$

Unlike the standard result in probability, now $X_n = f \circ T^n$ are not independent in general (although they might be in some particular situations).

Let $S_n := \sum_{i=0}^{n-1} f\left(T^i(x)\right)$. Even if we know that $T : X \to X$ is an ergodic map, we would like to know how fast the convergence of the averages S_n/n is to the expected value $\mathrm{E}[f]$. Unfortunately, this convergence is in general very slow except for some concrete functions and for dynamical systems that, in a broad sense, exhibit some *strong mixing properties* (in addition to being ergodic). Since we do not know how many iterates are required in 6.2 to obtain a good approximation of $\mathrm{E}[f]$ and since the rate of convergence to that value may vary from point to point, having a *Central Limit Theorem* for the sequence $X_n = f \circ T^n$ is crucial to estimate confidence intervals for the expectation $\mathrm{E}[f]$.

Definition 72 *Given $f \in L^1(X, \mu)$, we say that the random variables $X_n = f \circ T^n$ satisfy the Central Limit Theorem if*

$$\lim_{n\to\infty} \mu\left(\left\{\frac{S_n - n\,\mathrm{E}[f]}{\sqrt{n}} \leq z\right\}\right) = \frac{1}{\sqrt{2\pi}\sigma_f}\int_{-\infty}^{z} \mathrm{e}^{-\frac{s^2}{2\sigma_f^2}}\, ds$$

for some finite $\sigma_f^2 > 0$. That is, $(S_n - n\,\mathrm{E}[f])/\sqrt{n}$ converges in law to $\mathcal{N}(0, \sigma_f^2)$.

For example, if $f \in L^1(X, \mu)$ is such that $f \circ T^n$ satisfy the Central Limit Theorem, that is,

$$\frac{S_n - n\mathrm{E}[f]}{\sigma_f \sqrt{n}} = \frac{S_n/n - \mathrm{E}[f]}{\sigma_f/\sqrt{n}} \longrightarrow \mathcal{N}(0,1) \text{ as } n \to \infty \tag{6.3}$$

and we suppose we know σ_f then, for n *large enough*,

$$\mathrm{E}[f] \in \left[\frac{S_n}{n} - 1.96\frac{\sigma_f}{\sqrt{n}}, \frac{S_n}{n} + 1.96\frac{\sigma_f}{\sqrt{n}}\right]$$

with an *approximately* 95% confidence level. In general σ_f has to be estimated as well, which means that, strictly speaking, the Gaussian distribution in (6.3) must be replaced with a different law in order to obtain confidence intervals.

In probability, the Central Limit Theorem is proved for i.i.d sequences of square integrable random variables. Observe that, again, we are in completely different context because the random variables $X_n = f \circ T^n$ need not be independent (actually they will not be in general).

From the definition of S_n, it can be argued that, provided that such a σ_f^2 exists, it must be

$$\sigma_f^2 = C_f(0) + 2\sum_{n=1}^{\infty} C_f(n), \tag{6.4}$$

where

$$C_f(n) := \mathrm{Cov}(f, f \circ T^n) = \mathrm{E}[f(f \circ T^n)] - \mathrm{E}[f]^2, \; n \in \mathbb{N},$$

is the ***autocorrelation function***. More generally, given $f, g \in L^2(X, \mu)$, we introduce the ***correlation function***

$$C_{g,f}(n) := \mathrm{Cov}(g, f \circ T^n) = \mathrm{E}[g(f \circ T^n)] - \mathrm{E}[g]\mathrm{E}[f].$$

Exercise 73 Without loss of generality, we can assume that $\mathrm{E}[f] = 0$. Verify the following formula

$$\mathrm{Var}[S_n] = \mathrm{E}\left(S_n^2\right) = nC_f(0) + 2\sum_{i=1}^{n-1}(n-i)\,C_f(n).$$

Observe that in order that σ_f^2 in (6.4) be finite, we must have $C_f(n) \to 0$ as $n \to \infty$. By Proposition 71, this is guaranteed if T is mixing. However, to prove the Central Limit Theorem we need a fast convergence to 0. One can prove that

$$\lim_{n\to\infty} \left(\mathrm{Var}\,[S_n] - n\sigma_f^2\right) = -2\sum_{n=1}^{\infty} nC_f(n)$$

which implies that

$$\lim_{n\to\infty} \mathrm{Var}\left[\frac{S_n}{\sqrt{n}}\right] = \sigma_f^2$$

provided that

$$\sum_{n=1}^{\infty} n\left|C_f(n)\right| < \infty. \tag{6.5}$$

For example, (6.5) holds if there exist constants $K \geq 0$ and $\alpha > 2$ such that $|C_f(n)| \leq Kn^{-\alpha}$ (polynomial decay of correlations) or some constant $\beta > 0$ such $|C_f(n)| \leq \mathrm{e}^{-\beta n}$ (exponential decay). One particular class of functions exhibiting exponential decay are *Hölder continuous functions*:

Definition 74 *A function $f : X \to \mathbb{R}$ defined on a metric space X is called* ***Hölder continuous*** *if there exist constants $\alpha_f \in (0,1]$ and $K_f > 0$ such that*

$$|f(x) - f(y)| \leq K_f \,\mathrm{dist}\,(x,y)^{\alpha_f} \quad \forall x, y \in X.$$

Theorem 75 (exponential decay of correlations) *Let $T : X \to X$ be a mixing map on a metric space. For every pair of Hölder continuous functions f and g, there exist constants $B_{f,g} > 0$ and $\theta_{f,g} < 1$ such that*

$$|C_{f,g}(n)| \leq B_{f,g}\theta_{f,g}^n, \quad n \geq 1.$$

Therefore,

Theorem 76 *Let $f : X \to \mathbb{R}$ be a Hölder continuous functions. Then $X_n = f \circ T^n$ satisfy the Central Limit Theorem with σ_f^2 as in (6.4).*

Bibliography

[1] V. I. Arnold and A. Avez. *Problèmes ergodiques de la Mécanique Classique.* Gauthier-Villars, Paris (1967).

[2] R. G. Bartle. *The Elements of Integration and Lebesgue Measure*. Wiley Classics Library. John Wiley and Sons, Inc. 1995.

[3] V. I. Bogachev. *Measure Theory* (Volume I). Springer, 2007.

[4] M. Brin and G. Stuck. *Introduction to Dynamical Systems.* Cambridge university press, 2002.

[5] N. Chernov and R. Markarian. *Introduction to the Ergodic Theory of Chaotic Billiards.* Monografías del Instituto de Matemática y Ciencias Affines, **19**. Instituto de Matemática y Ciencias Affies; Pontificia Universidad Católica del Perú, Lima, 2001. Available at `premat.fing.edu.uy/papers/2003/75.ps` (Secon Edition) and `http://www.math.uab.edu/chernov/papers/rbook.pdf`.

[6] N. Dunford and J.T. Schwartz. *Linear operators, I. General Theory.* Interscience, New York, 1958.

[7] M. Hairer. *Ergodic Properties of Markov Processes*. Lecture notes, 2006. Available at `www.hairer.org/notes/Markov.pdf`.

[8] S. Luzzatto. Itroduction to Ergodic Theory. Lecture notes, 2009. Available at `http://web.mac.com/stefanoluzzatto/Teaching/M4A36.html`.

[9] W. Rudin. *Real and Complex Analysis*. Third edition. McGraw-Hill Book Co., 1987.

[10] I. G. Sinai. *Introduction to Ergodic Theory*. Princeton University Press, 1976.