COMPUTATIONAL ERGODIC THEORY

Vorlesung im Wintersemester 2008 / 09
von Prof. Dr. Fritz Colonius
Universität Augsburg

Autor: Christoph Kawan

Contents

1 Introduction

Literature: Geon Ho Choe, *Computational Ergodic Theory*, Springer 2005.
Origins: Statistical Mechanics, Boltzmann (1887), Birkhoff's Ergodic Theorem.

Let X be a set and assume that we can associate a probability measure to subsets of X, $\mu(A) \in [0,1]$, $A \subset X$. Let $T : X \circlearrowright$ with

$$\mu(A) = \mu(T^{-1}(A)) \text{ for all } A \subset X.$$

(T preserves μ, μ is invariant under T.) Let $N \in \mathbb{N}$.

$$\frac{1}{N}\#\{n \in \{1,\dots,N\} \mid T^n(x) \in B\} \xrightarrow{N\to\infty} ?$$

Birkhoff's Ergodic Theorem: $\to \mu(B)$, if we cannot decompose X into two subsets with positive probability measure which remain invariant under T (**Ergodic Hypothesis**).

Example: Let $X \subset \mathbb{R}$. One can define probability measures using a density ρ with respect to Lebesgue measure:

$$\mu(A) = \int_A \rho(x)dx$$

if $\rho(x) \geq 0$ and $\int_X \rho(x)dx = 1$. T leaves μ invariant if

$$\int_A \rho(x)dx = \int_{T^{-1}(A)} \rho(x)dx \text{ for all } A \subset X.$$

Let $X = [0,1]$, $T(x) = 4x(1-x)$ (logistic map). Then the invariant density is

$$\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

Thus, for $A \subset [0,1]$

$$\int_A \frac{dx}{\pi\sqrt{x(1-x)}} = \int_{T^{-1}(A)} \frac{dx}{\pi\sqrt{x(1-x)}}.$$

2 Invariant Measures

2.1 σ-Algebras and Probability Measures

In the following, X is a nonvoid set. A *σ-algebra* on X is a family $\mathcal{A}$ of subsets of X with

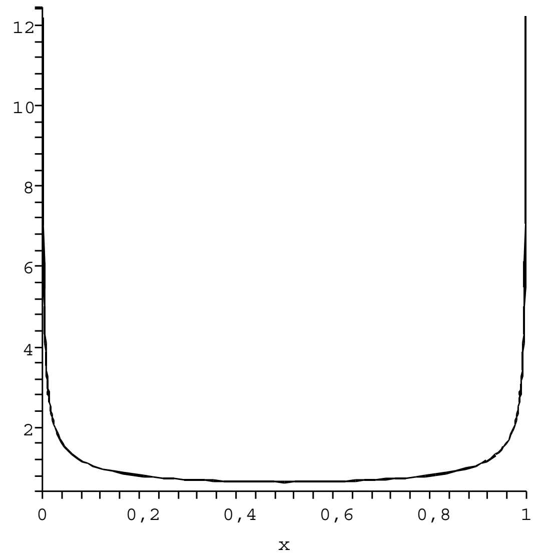

Figure 1: Graph of ρ, $\rho(x) = \left[\pi\sqrt{x(1-x)}\right]^{-1}$.

(i) $\emptyset, X \in \mathcal{A}$,

(ii) $A \in \mathcal{A}$ implies $X \backslash A \in \mathcal{A}$,

(iii) $A_n \in \mathcal{A}$, $n \in \mathbb{N}$, implies $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

A *measure* is a map $\mu : \mathcal{A} \to [0,\infty] = [0,\infty) \cup \{\infty\}$ with

(i) $\mu(\emptyset) = 0$,

(ii) $A_n \in \mathcal{A}$ $(n \in \mathbb{N})$ with $A_n \cap A_m = \emptyset$ for $n \neq m$ implies

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

If $\mu(X) = 1$, then μ is called a *probability measure*.

2.1 Examples:

(i) Counting measure:

$$\mu(A) := \#A \ \text{ (number of elements in } A).$$

(ii) Dirac measure:

$$\delta_x(A) := \begin{cases} 1 & \text{for } x \in A, \\ 0 & \text{for } x \notin A. \end{cases}$$

A pair $(X,\mathcal{A})$ is called a *measurable space*. A triple $(X,\mathcal{A},\mu)$ is called a *measure space*. A measure space is *complete* if

$$A \in \mathcal{A},\ \mu(A) = 0 \text{ and } N \subset A \quad \Rightarrow \quad N \in \mathcal{A} \text{ and } \mu(N) = 0.$$

Fact: Every measure space can be extended to a complete measure space.

An n-dimensional rectangle in $\mathbb{R}^n$ is a set of the form

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

with $a_i < b_i$ for $i = 1, \ldots, n$ (also open and half-open intervals are allowed). Let $\mathcal{R}$ be the set of all rectangles and define

$$\mu : \mathcal{R} \to [0,\infty],\ \ \mu(R) := \prod_{i=1}^{n}(b_i - a_i).$$

We can extend μ to a measure on the smallest σ-algebra containing $\mathcal{R}$. The corresponding complete measure is the familiar Lebesgue measure.

$f : (X,\mathcal{A}) \to \mathbb{R}$ is called *measurable* if $f^{-1}(I) \in \mathcal{A}$ for every open interval $I \subset \mathbb{R}$. A *characteristic function* $s : X \to \mathbb{R}$ is a function defined by

$$s(x) := \begin{cases} 1 & \text{for } x \in E \\ 0 & \text{for } x \notin E \end{cases}$$

for some $E \in \mathcal{A}$. We also write $s = \mathbb{1}_E$. A *simple function* $s : X \to \mathbb{R}$ is a function of the form

$$s(x) = \sum_{i=1}^{n} \alpha_i s_i(x)$$

with $\alpha_i \in \mathbb{R}$, $n \in \mathbb{N}$, and characteristic functions s_i. Every simple function is measurable, as can easily be shown.

Let f be a measurable function with $f(x) \geq 0$ for all $x \in X$. Then there exists an increasing sequence (s_n) of simple functions with $s_n(x) \to f(x)$ and $s_{n+1}(x) \geq s_n(x)$ for all $x \in X$.

Idea of the proof: Let f be a measurable function. Define

$$s_n(x) := \begin{cases} \frac{i-1}{2^n} & \text{if } \frac{i-1}{2^n} \leq f(x) \leq \frac{i}{2^n},\ i = 1, \ldots, n2^n, \\ n & \text{if } f(x) > n. \end{cases}$$

Let X be a metric space. Then the smallest σ-algebra containing all open sets is called the *Borel-σ-algebra*. The corresponding measurable sets are called *Borel measurable* and the measurable functions $f : X \to \mathbb{R}$ are called *Borel measurable functions*. Every continuous function is Borel measurable.

2.2 Example: Let $\mathcal{S} = \{0,1\}$. For every $p \in (0,1)$ define a probability measure on $\mathcal{A} := \mathcal{P}(\mathcal{S})$ by

$$\bar{\mu}_p(\{0\}) := p, \;\; \bar{\mu}_p(\{1\}) := 1 - p.$$

Define

$$X := \prod_1^\infty \mathcal{S} = \mathcal{S}^{\mathbb{N}}$$

with elements $x \in X$, $x = (x_1, x_2, x_2, \ldots)$, where $x_i \in \{0,1\}$. Let

$$[a_1, \ldots, a_n] := \{x \in X \mid x_i = a_i \text{ for } i = 1, \ldots, n\}$$

for $a_i \in \{0,1\}$, $i = 1, \ldots, n$. These sets are called *cylinder sets*. Let $\mathcal{R}$ be the set of all cylinder sets. Define $\mu_p : \mathcal{R} \to [0,1]$ by

$$\mu_p([a_1, \ldots, a_n]) := p^k (1-p)^{n-k},$$

where k is the number of zeros in $(a_1, \ldots, a_n)$. Then μ_p can be extended to a probability measure on the σ-algebra generated by the cylinder sets.

Recall that every element $(b_1, b_2, b_3, \ldots)$ of X represents a real number $x \in [0,1]$ via

$$x = \sum_{i=1}^\infty b_i 2^{-i}.$$

This representation is unique if we exclude tails only consisting of ones. Then μ_p can be considered as a measure on $[0,1]$. For $p = \frac{1}{2}$ this is the Lebesgue measure. In order to show this, note that

$$[a_1, \ldots, a_n] = \left\{ x \in [0,1] \; : \; x = \sum_{i=1}^n a_i b^{-i} + \sum_{i=n+1}^\infty b_i 2^{-i}, \; b_i \in \{0,1\} \right\}.$$

This set has Lebesgue measure

$$\mu([a_1, \ldots, a_n]) = \mu\left(\left\{ 2^{-(n+1)} \sum_{i=0}^\infty b_i 2^{-i} \; : \; b_i \in \{0,1\} \right\} \right) = \mu([0, 2^{-n}]) = 2^{-n}.$$

On the other hand,

$$\mu_{1/2}([a_1, \ldots, a_n]) = \left(\tfrac{1}{2}\right)^k \left(1 - \tfrac{1}{2}\right)^{n-k} = 2^{-n}.$$

μ_p is called the *Bernoulli measure*, see also Halmos [2, Sec. 3.8]. ◇

A measure μ is called *continuous* if $\mu(\{a\}) = 0$ for all $a \in X$.[1]

[1]Sometimes a measure with this property is also called *nonatomic*. An example for a measure which is not continuous is the Dirac measure δ_x.

2.3 Proposition: *The Bernoulli measures μ_p, $p \in (0,1)$, are continuous.*

Proof: For $a = (a_1, a_2, a_3, \ldots) \in X$ we have

$$\mu_p(\{a\}) = \mu_p\left(\bigcap_{n=1}^{\infty} [a_1, \ldots, a_n]\right) \overset{\forall n}{\leq} \mu_p([a_1, \ldots, a_n])$$
$$= p^{k_n}(1-p)^{n-k_n} \leq p^{k_n} \to 0,$$

where k_n is the number of zeros in $(a_1, \ldots, a_n)$. The latter holds, since we have excluded tails consisting of ones. This implies $\mu_p(\{a\}) = 0$. □

Let $p(x)$ be a property whose validity depends on $x \in X$. We say that p holds for *μ-almost all* $x \in X$ if $p(x)$ is true for all $x \in X \setminus N$ where $\mu(N) = 0$.

Integration with respect to a measure μ: Let $E \in \mathcal{A}$. Then $\mathbb{1}_E$ is integrable, if $\mu(E) < \infty$ and we define

$$\int_X \mathbb{1}_E d\mu := \mu(E).$$

Let $s = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{E_i}$ be a simple function with $\mu(E_i) < \infty$ for $i = 1, \ldots, n$. Then we define

$$\int_X s d\mu := \sum_{i=1}^{n} \alpha_i \int_X \mathbb{1}_{E_i} d\mu.$$

We call $f : X \to \mathbb{R}$ with $f(x) \geq 0$ for all $x \in X$ *Lebesgue-integrable* with respect to μ, if there is an increasing sequence of simple functions s_n such that

$$s_n(x) \to f(x) \text{ for } \mu\text{-almost all } x \in X$$

and we define

$$\int_X f d\mu := \lim_{n\to\infty} \int_X s_n dx,$$

provided the limit is finite. For a general $f : X \to \mathbb{R}$ decompose $f = f^+ - f^-$ with

$$f^+(x) := \max\{0, f(x)\}, \quad f^-(x) := \max\{0, -f(x)\}$$

and define

$$\int_X f d\mu := \int_X f^+ d\mu - \int_X f^- d\mu,$$

provided both integrals are finite. Similarly, if $f : X \to \mathbb{C}$, decompose f in real and imaginary parts.

Recall the following properties of the integral:

(i) Monotone convergence:

$$\int_X \lim_{n\to\infty} f_n d\mu = \lim_{n\to\infty} \int_X f_n d\mu,$$

if (f_n) is monotone increasing or decreasing (almost everywhere).

(ii) Fatou's Lemma:

$$\int_X \liminf_{n\to\infty} f_n d\mu \le \liminf_{n\to\infty} \int_X f_n d\mu.$$

(iii) Lebesgue's Theorem on dominated convergence.

The L^p-spaces $L^p(X,\mathbb{R},\mu)$ and $L^p(X,\mathbb{C},\mu)$ for $p \in [1,\infty)$ are the Banach spaces of measurable functions $f : X \to \mathbb{R}$ ($f : X \to \mathbb{C}$) such that the integral over $|f|^p$ exists. The norm is defined by[2]

$$\|f\|_p := \left(\int_X |f|^p d\mu\right)^{1/p}.$$

2.2 Invariant Measures

2.4 Definition: *Let $(X_1,\mathcal{A}_1,\mu_1)$ and $(X_2,\mathcal{A}_2,\mu_2)$ be measure spaces and $T : X_1 \to X_2$ measurable, i.e., $T^{-1}(E) \in \mathcal{A}_1$ for all $E \in \mathcal{A}_2$. The map T is called* **measure preserving** *if*

$$\mu_2(E) = \mu_1(T^{-1}(E)) \ \text{ for all } E \in \mathcal{A}_2.$$

If $X_1 = X_2$, $\mathcal{A}_1 = \mathcal{A}_2$ and $\mu_1 = \mu_2 =: \mu$, then we call T a **transformation***, and μ is called* ***T*-invariant**.

2.5 Theorem: *Let $(X,\mathcal{A},\mu)$ be a measure space and $T : X \circlearrowleft$ a measurable map. Then the following are equivalent:*

(i) *T is a transformation.*

(ii) *For all functions f which are integrable with respect to μ we have*

$$\int_X f d\mu = \int_X f \circ T d\mu.$$

(iii) *Define a linear operator U_T on $L^p(X,\mathbb{C},\mu)$ for $p \in [1,\infty)$ by*

$$U_T f := f \circ T \ \text{ for all } f \in L^p(X,\mathbb{C},\mu).$$

Then U_T is norm-preserving, i.e.,

$$\|f\|_p = \|U_t f\|_p \ \text{ for all } f \in L^p(X,\mathbb{C},\mu).$$

Proof: "(ii) $\Rightarrow$ (i)": Let $f = \mathbb{1}_E$. Then

$$\mu(E) = \int_X \mathbb{1}_E d\mu \overset{\text{(ii)}}{=} \int_X \mathbb{1}_E \circ T d\mu = \int_X \mathbb{1}_{T^{-1}(E)} d\mu = \mu(T^{-1}(E)).$$

[2]Actually the elements of the L^p-spaces are equivalence classes of functions, whereby two functions are considered to be equivalent if they coincide almost everywhere.

“(i) ⇒ (ii)”: We can write $f \in L^1(X, \mathbb{C}, \mu)$ as $f = f_1 - f_2 + i(f_3 - f_4)$ with $f_i \geq 0$. Hence, it suffices to prove (ii) for $f \geq 0$. As above, for $f = \mathbb{1}_E$

$$\begin{aligned} \int_X f d\mu &= \mu(E) \overset{\text{(i)}}{=} \mu(T^{-1}(E)) = \int_X \mathbb{1}_{T^{-1}(E)} d\mu \\ &= \int_X \mathbb{1}_E \circ T d\mu = \int_X f \circ T d\mu. \end{aligned}$$

By linearity of the integral this is also true for simple functions. By our construction every $f \in L^1(X, \mathbb{C}, \mu)$ with $f \geq 0$ can be approximated by an increasing sequence of simple functions s_n: $s_n(x) \to f(x)$ for all $x \in X$. Then also $s_n(T(x)) \to f(T(x))$ for all $x \in X$ and $s_n \circ T$ are also simple functions, and the sequence is monotone increasing. Thus, by monotone convergence

$$\int_X f \circ T d\mu = \lim_{n\to\infty} \int_X s_n \circ T d\mu = \lim_{n\to\infty} \int_X s_n d\mu = \int_X f d\mu.$$

“(i) ⇔ (iii)”: This is proved similarly as the equivalence of (i) and (ii). □

2.3 Examples

2.6 Example: Let $X = [0, 1)$ and $T(x) = x + \theta \pmod 1$, where $\theta \in [0, 1)$. Then the Lebesgue measure is invariant under T. Suppose $\theta = \frac{p}{q} \in \mathbb{Q}$. Then

$$T^q(x) = x + q\Theta \pmod 1 = x + p \pmod 1 = x.$$

Hence, every point is periodic. So the interesting case is when θ is irrational.[3] ◊

2.7 Example: Let $X = [0, 1)$ and $T(x) = 2x \pmod 1$. The preimage of an interval E consists of two intervals, each of them with half the length of E. Again, the Lebesgue measure (i.e., the length of intervals) is an invariant measure. ◊

2.8 Example: Let $X = [0, 1)$ and

$$T(x) = \begin{cases} 2x \pmod 1 & \text{for } 0 \leq x < \frac{1}{2}, \\ 4x \pmod 1 & \text{for } \frac{1}{2} \leq x < 1. \end{cases}$$

Again, the Lebesgue measure is invariant. ◊

[3] Note that for rational θ the Lebesgue measure is not the only invariant measure. Indeed, there is an infinite number of invariant measures.

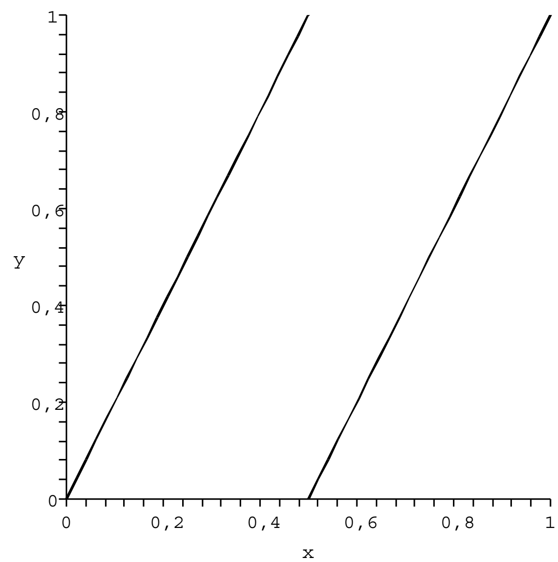

Figure 2: $T(x) = 2x \pmod 1$, see Ex. 2.7.

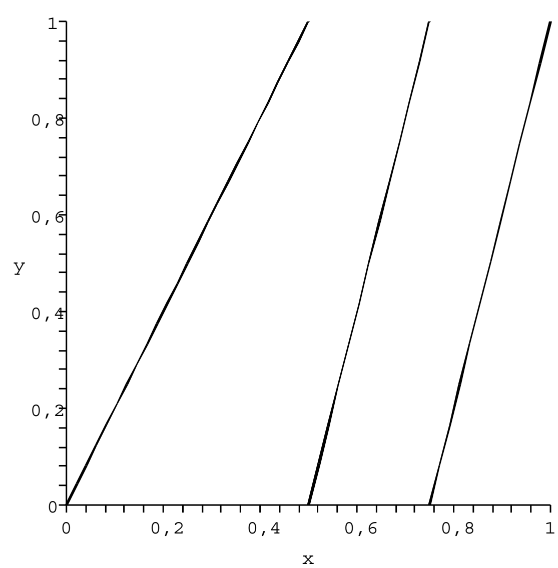

Figure 3: T from Ex. 2.8.

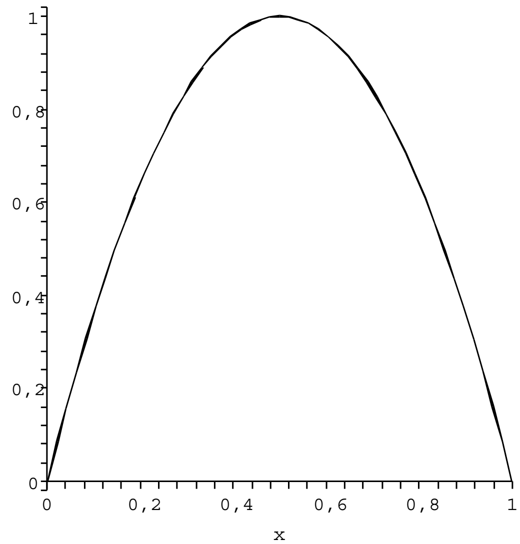

Figure 4: $T(x) = 4x(1-x)$, see Ex. 2.9.

2.9 Example: (*The logistic map with parameter 4*)
Let $X = [0,1)$ and $T(x) = 4x(1-x)$. We claim that there is an invariant measure μ with density

$$\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

with respect to Lebesgue measure, and μ is a probability measure. We have to show that

$$\mu(A) = \int_A \rho(x)dx = \mu(T^{-1}(A)) = \int_{T^{-1}(A)} \rho(x)dx.$$

Proof: Use the `MAPLE` program `logistics1`. ◊

2.10 Example: (β-*transformation*)
Let $\beta := \frac{1}{2}(\sqrt{5}+1)$. This is a solution of

$$0 = \beta^2 - \beta - 1,$$

or equivalently, $\beta - 1 = \frac{1}{\beta}$. Hence, β is the golden section. Define $X = [0,1)$. Then the transformation is given by

$$T(x) = \beta x \text{ (mod 1)}.$$

There is an invariant measure μ with a density with respect to Lebesgue measure, given by

$$\rho(x) = \begin{cases} \frac{\beta^3}{1+\beta^2} & \text{for } 0 \leq x < \frac{1}{\beta}, \\ \frac{\beta^2}{1+\beta^2} & \text{for } \frac{1}{\beta} \leq x. \end{cases}$$

$\Diamond$

2.11 Example: (*The Gauß-Transformation*)
Let $X = [0,1)$ and define

$$T(x) := \begin{cases} \frac{1}{x} \pmod 1 & \text{for } 0 < x < 1, \\ 0 & \text{for } x = 0. \end{cases}$$

[GRAPH]

Gauß (1812, in a letter to Laplace): There is an invariant probability measure with a density with respect to Lebesgue measure, given by

$$\rho(x) = \frac{1}{\ln(2)} \frac{1}{x+1}.$$

For the proof it suffices to show that

(i) $\int_{T^{-1}((0,a))} \frac{dx}{x+1} = \int_{(0,a)} \frac{dx}{x+1}$,

(ii) $\int_0^1 \frac{dx}{x+1} = \ln(2)$.

Statement (ii) is proved by

$$\int_0^1 \frac{dx}{x+1} = [\ln(x+1)]_0^1 = \ln(2) - \underbrace{\ln(1)}_{=0} = \ln(2).$$

For the proof of (i) note that $T^{-1}((0,a))$ is the disjoint union of the intervals $(\frac{1}{n+a}, \frac{1}{n}]$, $n \in \mathbb{N}$, since for $x \in [0,1)$ we have

$$\begin{aligned} \tfrac{1}{x} \pmod 1 \in (0,a) &\Leftrightarrow \exists n \in \mathbb{N} : \tfrac{1}{x} \in (n, n+a) \\ &\Leftrightarrow \exists n \in \mathbb{N} : x \in \left(\tfrac{1}{n+a}, \tfrac{1}{n}\right). \end{aligned}$$

This implies

$$\begin{aligned}
\int_{T^{-1}((0,a))} \frac{dx}{x+1} &= \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\frac{1}{n+a}}^{\frac{1}{n}} \frac{dx}{x+1} = \lim_{N\to\infty} \sum_{n=1}^{N} \left[\ln(x+1)\right]_{\frac{1}{n+a}}^{\frac{1}{n}} \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \left[\ln\left(\frac{n+1}{n}\right) - \ln\left(\frac{n+a+1}{n+a}\right)\right] \\
&= \lim_{N\to\infty} \left[\ln(N+1) - \ln(N+a+1) + \ln(1+a)\right] \\
&= \lim_{N\to\infty} \left[-\ln\left(\frac{N+a+1}{N+1}\right) + \ln(1+a)\right] \\
&= \ln(1+a) = \int_0^a \frac{dx}{x+1}.
\end{aligned}$$

◊

2.12 Example: Let $X = \mathbb{R}$ and $T(x) = x - \frac{1}{x}$. Then T preserves Lebesgue measure, i.e., $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} f(x - \frac{1}{x})dx$ for all $f \in L^1(\mathbb{R}, \mathbb{C}, \mu)$.
In order to show that, let $y \in \mathbb{R}$. Then the preimage of y is given by

$$y = x - \tfrac{1}{x} = \tfrac{x^2-1}{x} \Leftrightarrow x^2 - xy - 1 = 0 \Leftrightarrow x = \tfrac{y}{2} \pm \tfrac{1}{2}\sqrt{y^2+4}.$$

Thus, T^{-1} of an interval (a, b) is the union of the two intervals

$$\left(\tfrac{1}{2}(a - \sqrt{a^2+4}, \tfrac{1}{2}(b - \sqrt{b^2+4})\right), \quad \left(\tfrac{1}{2}(a + \sqrt{a^2+4}, \tfrac{1}{2}(b + \sqrt{b^2+4})\right).$$

The sum of their lengths is $b - a$, which proves the assertion. ◊

2.13 Example: Let $X = \mathbb{R}$, $T(x) = \frac{1}{2}(x - \frac{1}{x})$. Then T has an invariant probability measure with density

$$\rho(x) = \frac{1}{\pi(1+x^2)}$$

with respect to Lebesgue measure. The preimage $T^{-1}((a, b))$ is the union of the two intervals

$$\left(a - \sqrt{a^2+1}, b - \sqrt{b^2+1}\right), \quad \left(a + \sqrt{a^2+1}, b + \sqrt{b^2+1}\right).$$

Now

$$\mu((a,b)) = \frac{1}{\pi} \int_a^b \frac{dx}{1+x^2} = \frac{1}{\pi}\left(\arctan(b) - \arctan(a)\right)$$

and

$$\begin{aligned}
\mu(T^{-1}((a,b))) &= \frac{1}{\pi} \int_{a-\sqrt{a^2+1}}^{b-\sqrt{b^2+1}} \frac{dx}{1+x^2} \\
&= \tfrac{1}{\pi}\big(\arctan(b - \sqrt{b^2+1}) - \arctan(a - \sqrt{a^2+1}) \\
&+ \arctan(b + \sqrt{b^2+1}) - \arctan(a + \sqrt{a^2+1})\big).
\end{aligned}$$

By using the trigonometric identity

$$\arctan\left(u+\sqrt{u^2+1}\right)+\arctan\left(u-\sqrt{u^2+1}\right)\equiv\arctan(u)$$

we obtain

$$\mu(T^{-1}((a,b)))=\tfrac{1}{\pi}\left(\arctan(b)-\arctan(a)\right).$$

Furthermore,

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{dx}{1+x^2}=\frac{1}{\pi}\lim_{x\to\infty}\left(\arctan(x)-\arctan(-x)\right)=1.$$

This transformation T comes from Newton's method applied to $f(x)=1+x^2$:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}=x_n-\frac{1+x_n^2}{2x_n}=\frac{2x_n^2-1-x_n^2}{2x_n}=\tfrac{1}{2}\left(x_n-\tfrac{1}{x_n}\right).$$

$\diamondsuit$

As a motivation for the following example consider again the map $T(x)=x+\theta \pmod 1$ on $X=[0,1)$. The interval $[0,1)$ can be identified with $S^1=\{z\in\mathbb{C}\mid |z|=1\}$ via the map $x\mapsto e^{2\pi ix}$, $[0,1)\to S^1$. Addition modulo one defines a group structure on S^1, where the neutral element is 0 and the inverse of $x\in[0,1)$ is given by $-x+1$, since $x+(-x+1) \pmod 1 = 1 \pmod 1 = 0$. Addition and inversion are continuous on S^1.

2.14 Example: (*Endomorphisms of compact groups)*
A topological group G is a group which also is a topological space such that the group operations are continuous, i.e., the maps

$$(g_1,g_2)\mapsto g_1g_2,\quad G\times G\to G,$$
$$g\mapsto g^{-1},\quad G\to G,$$

are continuous. We also require that G has the Hausdorff property: For $g_1,g_2\in G$ with $g_1\neq g_2$ there are disjoint open sets A_1 and A_2 with $g_1\in A_1$ and $g_2\in A_2$. An endomorphism of a topological group G is a map $\varphi: G\to G$ which is a group homomorphism and continuous. We will use the following

THEOREM: *For a compact topological group G there is a unique measure μ (on the Borel-σ-algebra of G) such that $\mu(G)=1$ and for all open sets $A\subset G$ and all $x\in G$*

$$\mu(A)=\mu(xA) \tag{1}$$

where $xA=\{xa\mid a\in A\}$.

A measure with the property (1) is also called *left invariant*, and the unique measure μ of the theorem is called the *Haar measure* on G. An example for

Haar measure is Lebesgue measure on $[0,1]/\sim\;\cong S^1 \cong \mathbb{R}/\mathbb{Z}$, where $\sim$ is the equivalence relation which identifies 0 and 1 and every other point only with itself. This space is an abelian compact group with the addition modulo one.

CLAIM: A surjective endomorphism Φ of a compact topological group preserves the Haar measure (Reference: Pedersen [3]).

Proof: Let μ be the Haar measure. Define

$$\nu(E) := \mu(\Phi^{-1}(E))$$

for all measurable sets $E \subset G$. Then ν is a probability measure as can easily be verified. We want to show that $\nu = \mu$. Due to the theorem it suffices to show that ν is left invariant: Write an arbitrary element of G as $\Phi(x)$. Then

$$\begin{aligned} y \in \Phi^{-1}(\Phi(x)A) &\Leftrightarrow \Phi(y) \in \Phi(x)A \\ &\Leftrightarrow \Phi(x)^{-1}\Phi(y) \in A \\ &\Leftrightarrow \Phi(x^{-1}y) \in A \\ &\Leftrightarrow x^{-1}y \in \Phi^{-1}(A) \\ &\Leftrightarrow y \in x\Phi^{-1}(A). \end{aligned}$$

Hence, $\Phi^{-1}(\Phi(x)A) = x\Phi^{-1}(A)$. Thus,

$$\nu(\Phi(x)A) \stackrel{\text{def}}{=} \mu(\Phi^{-1}(\Phi(x)A))) = \mu(x\Phi^{-1}(A)) = \mu(\Phi^{-1}(A)) \stackrel{\text{def}}{=} \nu(A).$$

Since $\Phi(x)$ is an arbitrary element of G, ν is left invariant. ◊

2.15 Example: (*The baker's transformation*)
Let $X = [0,1] \times [0,1]$ and

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

This map preserves Lebesgue measure on the unit square. (Note that T is not continuous. The first component can also be written as $2x \pmod 1$.) See also the `MAPLE` program `image_baker`. ◊

2.16 Example: (*Arnold's Cat Map, a toral automorphism*)
Let $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$. Then the matrix $A := \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ defines a transformation from $\mathbb{T}^2$ to $\mathbb{T}^2$ by

$$T(x,y) = (2x + y \pmod 1, x + y \pmod 1).$$

Indeed: A defines a linear map $L_A : \mathbb{R}^2 \circlearrowleft (x \mapsto Ax)$. Since all entries of A are integers, L_A maps $\mathbb{Z}^2$ to $\mathbb{Z}^2$. Since $\det A = 1$, A is invertible and also A^{-1} has only integer entries. Hence, also $L_{A^{-1}} = L_A^{-1}$ maps $\mathbb{Z}^2$ to $\mathbb{Z}^2$. Therefore, A induces a bijective map $T : \mathbb{T}^2 \circlearrowleft$ by

$$\pi \circ L_A = T \circ \pi,$$

where $\pi : \mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ is the natural projection, mapping x to its equivalence class $x + \mathbb{Z}^2$. That is, the following diagram commutes.

$$\begin{array}{ccc} \mathbb{R}^2 & \xrightarrow{L_A} & \mathbb{R}^2 \\ \pi\downarrow & & \downarrow\pi \\ \mathbb{T}^2 & \xrightarrow[T]{} & \mathbb{T}^2 \end{array}$$

Let $x = x' + m$ with $m \in \mathbb{Z}^2$, i.e., $\pi(x) = \pi(x')$. Then $\pi(L_A(x)) = \pi(L_A(x') + L_a(m)) = \pi(L_A(x))$, since $L_a(m) \in \mathbb{Z}^2$. The same is true for T^{-1}. This proves that T is well-defined. The invariant measure is the two-dimensional Lebesgue measure. See the `MAPLE` program `Image_Arnold`. $\Diamond$

2.17 Example: (*The* Λ*-transformation*)
Let $X = [0, 1]$. For $0 < c < 1$ define the Λ-transformation

$$\tau_c(x) := \begin{cases} \frac{1}{c}x & \text{for } 0 \le x \le c, \\ -\frac{1}{1-c}x + \frac{1}{1-c} & \text{for } c < x \le 1. \end{cases}$$

The Lebesgue measure is invariant, since the preimage of an interval E is the union of two intervals with total length equal to the length of E. $\Diamond$

2.18 Example: (*The truncated* Λ*-transformation*)
Let $X = [0, 1]$ and for $\frac{1}{2} < a < 1$ define $b := \frac{2a-1}{a}$. Let

$$T_a(x) := \begin{cases} \frac{1-a}{b}x + a & \text{for } 0 \le x \le b, \\ \frac{a}{1-a}(-x + 1) & \text{for } b < x \le 1. \end{cases}$$

Then $T_a(b) = 1$, $T_a(0) = a$, $T_a(1) = 0$ and $T_a(a) = a$. The latter follows from $b < a$ which – by definition of b – is equivalent to $(a-1)^2 > 0$. We have $T_a^{-1}(\{a\}) = \{0, a\}$. Therefore, δ_a is an invariant measure. There is a more interesting invariant measure μ which has a density with respect to Lebesgue measure. We have

$$\begin{aligned} T_a([0, a]) &= [a, 1] \text{ and } T([a, 1]) = [0, a], \\ T_a^{-1}([0, a]) &= [a, 1] \cup \{0\}, \\ T_a^{-1}([a, 1]) &= [0, a]. \end{aligned}$$

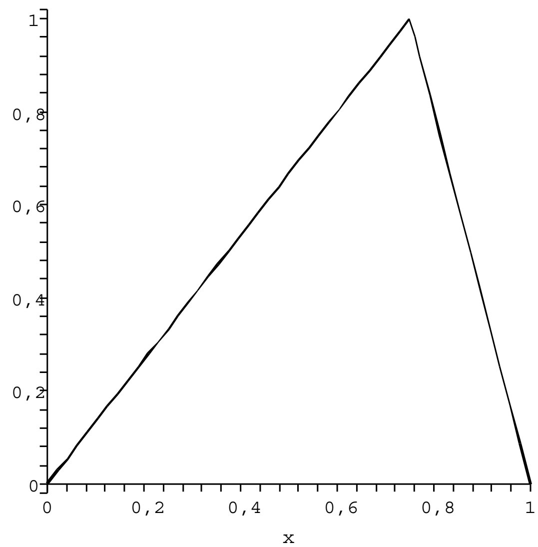

Figure 5: The Lambda Transformation from Ex. 2.17 for $c = \frac{3}{4}$.

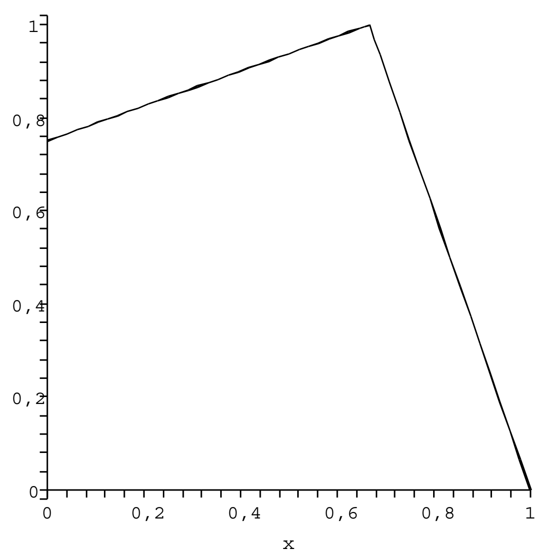

Figure 6: The Truncated Lambda Transformation from Ex. 2.18 for $a = \frac{3}{4}$.

Hence, if μ is an invariant measure with a density with respect to Lebesgue measure, then

$$\mu([0,a]) = \mu([a,1] \cup \{0\}) = \mu([a,1]) + \underbrace{\mu(\{0\})}_{=0} = \mu([a,1]).$$

Since $X = [0,1] = [0,a] \cup [a,1]$, it holds that $\mu([0,a]) = \frac{1}{2} = \mu([a,1])$. If μ is invariant under T_a, then it is also invariant under T_a^2, since[4]

$$\mu(T_a^{-2}(E)) = \mu(T_a^{-1}(E)) = \mu(E).$$

The restrictions of T_a^2 to $[0,a]$ and $[a,1]$ are well-defined with invariant measures $\frac{1}{a}dx$ and $\frac{1}{1-a}dx$, respectively. We obtain an invariant measure for T_a^2 with density

$$\rho(x) = \begin{cases} \frac{1}{2a} & \text{for } 0 \le x \le a, \\ \frac{1}{2(1-a)} & \text{for } a < x \le 1. \end{cases}$$

An easy calculation shows that this is also an invariant measure for T_a. $\Diamond$

2.4 Shift Transformations

Consider the set of symbols $\{1,\dots,k\}$. Define the set

$$X := \prod_1^\infty \{1,\dots,k\}$$

of sequences with entries in $\{1,\dots,k\}$ ($k = 2$: binary sequences). Let $p_1,\dots,p_k \ge 0$ with $\sum_{j=1}^k p_j = 1$. This defines a probability measure on $\{1,\dots,k\}$. For $t \ge 1$ define a *block* or *cylinder set* of length n by

$$[a_1,\dots,a_n]_{t,\dots,t+n-1} := \{(x_1,x_2,\dots) \mid x_t = a_1,\ x_{t+1} = a_2,\dots,x_{t+n-1} = a_n\}.$$

Define μ on cylinder sets by

$$\begin{aligned} \mu\left([a_1,\dots,a_n]_{t,\dots,t+n-1}\right) &:= p_{a_1} p_{a_2} \dots p_{a_n} \in [0,1], \\ \mu(\emptyset) &:= 0. \end{aligned}$$

We have

$$X = \bigcup_{i=1}^k [i]_1 \ \Rightarrow\ \mu(X) = \sum_{i=1}^k p_i = 1.$$

[4]For an arbitrary map $T : X \circlearrowleft$, $n \in \mathbb{N}$, and $A \subset X$ we define $T^{-n}(A) := \{x \in X \mid f^n(x) \in A\}$, i.e., the preimage of A under the n^{th} iterate of T.

Then we can extend μ to the σ-algebra generated by the cylinder sets and get a probability measure. The natural map to consider on the set X is the shift:

$$\theta : X \circlearrowleft, \ \ (x_1, x_2, x_2, \ldots) \mapsto (x_2, x_3, \ldots).$$

θ is called the *Bernoulli shift* and X is called the *Bernoulli shift space*. The measure μ is shift-invariant:

$$\mu\left(\theta^{-1}([a_1, a_2, \ldots, a_n]_{t,\ldots,t+n-1})\right) = \mu\left([a_1, a_2, \ldots, a_n]_{t+1,\ldots,t+n}\right) = p_{a_1} \cdot \ldots \cdot p_{a_n}.$$

Analogous for $X = \prod_{-\infty}^{\infty}\{1, \ldots, k\}$.

A *stochastic $k \times k$-matrix* $P = (p_{ij})$ is a matrix with entries $p_{ij} \geq 0$ and

$$\sum_j p_{ij} = 1 \ \ \forall i \ \ \text{(the row sums are equal to one).}$$

If P is a stochastic matrix, then P^n ($n \in \mathbb{N}$) is a stochastic matrix. In fact,

$$\sum_j (P^2)_{ij} = \sum_j \sum_k p_{ik} p_{kj} = \sum_k \sum_j p_{ik} p_{kj} = \sum_k p_{ik} \underbrace{\sum_j p_{kj}}_{=1} = 1.$$

Analogously this works for general n. Interpretation: p_{ij} is the probability to go from i to j. $(P^n)_{ij}$ is the probability to go from i to j in n steps.

A stochastic matrix P is called *irreducible* if for all i, j there is $m \in \mathbb{N}$ with $(P^m)_{ij} > 0$.

Convention: We write vP for the product of a row vector v and a matrix P.

2.19 Lemma: *Let $P \in \mathbb{R}^{k \times k}$ be irreducible. Then every eigenvector $w \geq 0$ for a positive eigenvalue λ satisfies $w_j > 0$ for all j.*

Proof: Since w is an eigenvector, there is at least one component which is positive, say $w_\mu > 0$. For all j there is $m \in \mathbb{N}$ with $(P^m)_{\mu j} > 0$. Since $wP^m = \lambda^m w$, we have

$$\sum_i w_i (P^m)_{ij} = \lambda^m w_j$$

and

$$0 < \underbrace{w_\mu}_{>0} \underbrace{(P^m)_{\mu j}}_{>0} \leq \sum_i w_i (P^m)_{ij} = \lambda^m w_j \Rightarrow w_j > 0.$$

This proves the assertion. □

2.20 Theorem: *Let $P \in \mathbb{R}^{k \times k}$ be a stochastic matrix. Then the following statements hold true:*

(i) *P has the eigenvalue* 1.

(ii) *There is a vector $v \geq 0$ (i.e., all entries of v are nonnegative) with $v \neq 0$ and $vP = v$.*

(iii) *Let P be irreducible. Then there is a unique vector $\pi = (\pi_1, \ldots, \pi_k)$ such that $\pi P = \pi$, $\sum_{i=1}^k \pi_i = 1$ and $\pi_i \geq 0$.*

Proof:

(i) Take $u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. Then $Pu = u$. So P has the eigenvalue 1.

(ii) Define $f(v) = vP$ for all $v \in \mathbb{R}^k$ and

$$S := \left\{ v \in \mathbb{R}^k \;:\; 1 = \|v\|_1 = \sum_{i=1}^k v_i \text{ and } v_1, \ldots, v_k \geq 0 \right\}.$$

For $v \in S$

$$\|f(v)\|_1 = \|vP\|_1 = \sum_j (vP)_j = \sum_j \sum_i v_i p_{ij} = \sum_i v_i \underbrace{\sum_j p_{ij}}_{=1} = 1.$$

Hence, f defines a map, again denoted by f, which maps S into S. The set S is compact and convex. Since f is continuous, we can apply Brouwer's fixed point theorem and conclude that there is a fixed point of f, i.e., $v = f(v) = vP$, $v \in S$.

(iii) Let π be a fixed point of the map $f : S \circlearrowleft$. Let $v \neq 0$ be an eigenvector for $\lambda = 1$. Consider for every $t \in \mathbb{R}$ the vector $\pi + tv$. This is an eigenvector for $\lambda = 1$. By Lemma 2.19 all entries of π are positive. Choose $t_0 \in \mathbb{R}$ such that all entries are nonnegative, but at least one component is equal to zero. By Lemma 2.19 this implies $\pi + t_0 v = 0$ and hence $v = \frac{1}{t_0}\pi$. This shows uniqueness of π.

□

The eigenvalue $\lambda = 1$ is called the *Frobenius-Perron eigenvalue* and π the *Frobenius-Perron eigenvector*.[5]

Now let $X = \prod_1^\infty \{1, \ldots, k\}$ and fix an irreducible stochastic $k \times k$-matrix P. Consider the Perron-Frobenius eigenvector $\pi = (\pi_i)$ ($\pi > 0$ and $\sum_i \pi_i = 1$, $\pi P = \pi$). Define μ by

$$\mu\left([a_1, \ldots, a_n]_{t,\ldots,t+n-1}\right) = \pi_{a_1} p_{a_1 a_2} p_{a_2 a_3} \cdots p_{a_{n-1} a_n}.$$

This generates a shift-invariant probability measure on X, again denoted by μ. This is called a *Markov measure* and X is called the *Markov shift space*.

[5] It can also be shown that the generalized eigenspace for $\lambda = 1$ is one-dimensional.

Suppose

$$p_{1j} = p_{ij} \text{ for all } i, j.$$

Then we get back a *Bernoulli shift*. ($p_j = p_{1j}, j = 1, \ldots, k$).

2.5 Isomorphic Transformations

Recall from Linear Algebra: Two matrices $A, B \in \mathbb{R}^{n\times n}$ are similar if $A = S^{-1}BS$ for some $S \in \mathrm{Gl}(n, \mathbb{R})$, or equivalently $SA = BS$:

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{A} & \mathbb{R}^n \\ {\scriptstyle S}\downarrow & & \downarrow{\scriptstyle S} \\ \mathbb{R}^n & \xrightarrow[B]{} & \mathbb{R}^n \end{array}$$

What is the appropriate definition of similarity for measure preserving transformations?

2.21 Definition:

(i) *Let (X_1, μ_1) and (X_2, μ_2) be measure spaces. A map $\Phi : X_1 \to X_2$ is said to be* **almost everywhere bijective***, if there are $E_1 \subset X_1$ and $E_2 \subset X_2$ with $\mu_1(E_1) = \mu_2(E_2) = 0$ such that*

$$\Phi|_{X_1\setminus E_1} : X_1\setminus E_1 \to X_2\setminus E_2$$

is bijective.

(ii) *(X_1, μ_1) and (X_2, μ_2) are called* **isomorphic** *with isomorphism Φ, if Φ is an almost everywhere bijective map with Φ, Φ^{-1} measurable and measure preserving.*

2.22 Example: The spaces $\prod_1^\infty \{0, 1\}$ with $(\frac{1}{2}, \frac{1}{2})$-Bernoulli measure and $[0, 1]$ with Lebesgue measure are isomorphic. See Exercise 2 on Sheet 2. $\diamondsuit$

2.23 Definition: *Let T_1 on (X_1, μ_1) and T_2 on (X_2, μ_2) be measure preserving. They are called* **isomorphic** *or* **conjugate***, if there is an isomorphism $\Phi : (X_1, \mu_1) \to (X_2, \mu_2)$ such that the following diagram commutes.*

$$\begin{array}{ccc} (X_1, \mu_1) & \xrightarrow{T_1} & (X_1, \mu_1) \\ {\scriptstyle \Phi}\downarrow & & \downarrow{\scriptstyle \Phi} \\ (X_2, \mu_2) & \xrightarrow[T_2]{} & (X_2, \mu_2) \end{array}$$

Note: Here we assume that $T_1(X_1\setminus E_1) \subset X_1\setminus E_1$ and $T_2(X_2\setminus E_2) \subset X_2\setminus E_2$ for the null sets E_1 and E_2 outside of which Φ is bijective.

2.24 Remarks:

(i) Conjugation is an equivalence relation.

(ii) If Φ is only measure preserving, then Φ is called a **semi-conjugacy**, T_2 is called a **factor** of T_1, and T_1 is called an **extension** of T_2.

(iii) Analogous definitions can be given in a topological setting, where X_1, X_2 are topological spaces and T_1, T_2 are continuous. Then it is required that Φ is a homeomorphism with $\Phi \circ T_1 = T_2 \circ \Phi$ (*topological conjugacy*).

2.25 Example: The measure spaces $X_1 := \prod_1^\infty \{0,1\}$ with $(\frac{1}{2}, \frac{1}{2})$-Bernoulli measure μ_1 and $X_2 := [0,1]$ with Lebesgue measure dx are isomorphic via

$$\Phi : (x_1, x_2, x_2, \ldots) \mapsto \sum_{n=1}^{\infty} x_n 2^{-n}.$$

Let $T_1 : X_1 \circlearrowleft$ be the shift

$$T_1((x_1, x_2, x_3, \ldots)) = (x_2, x_3, \ldots)$$

and $T_2 : X_2 \circlearrowleft$ the map

$$T_2(x) = 2x \ (\mathrm{mod}\ 1).$$

Then $\Phi \circ T_1 = T_2 \circ \Phi$, which is proven by

$$\begin{aligned} T_2\left(\sum_{n=1}^{\infty} x_n 2^{-n}\right) &= x_1 + \underbrace{\sum_{n=2}^{\infty} x_n 2^{-n+1}}_{\leq 1} \ (\mathrm{mod}\ 1) \\ &= \sum_{n=1}^{\infty} x_{n+1} 2^{-n} = \Phi(x_2, x_3, \ldots) = \Phi(T_1(x_1, x_2, \ldots)). \end{aligned}$$

$\Diamond$

2.26 Example: Let $X = [0,1]$ and $T(x) = 4x(1-x)$ (logistic map) and

$$\Lambda(x) = \begin{cases} 2x & \text{for } x \in [0, \frac{1}{2}], \\ 2 - 2x & \text{for } x \in (\frac{1}{2}, 1]. \end{cases} \quad \text{(tent map)}$$

Λ preserves Lebesgue measure and T preserves $d\mu = \rho(x)dx$ with density $\rho(x) = \frac{1}{\pi\sqrt{x(1-x)}}$. Define

$$\Phi(x) := \sin^2\left(\frac{\pi}{2}x\right) = \frac{1}{2}\left(1 - \cos(\pi x)\right).$$

Clearly, Φ is bijective and $\Phi : (X, dx) \to (X, \mu)$ is measure preserving: The

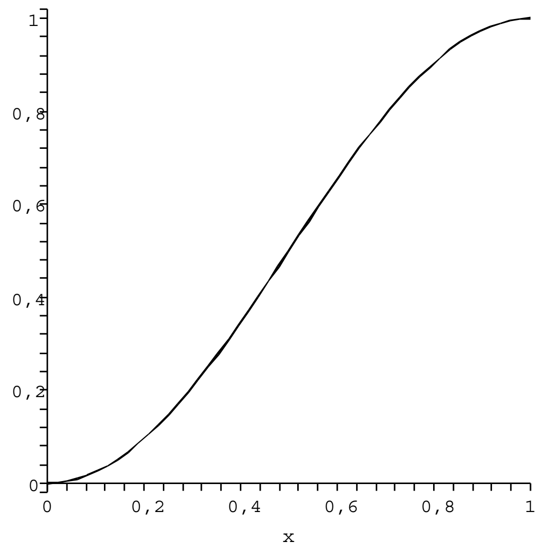

Figure 7: Graph of Φ from Ex. 2.26.

length (= Lebesgue measure) of $\Phi^{-1}([0, \Phi(a)])$ equals the length of $[0, a]$ which equals a and

$$\mu([0, \Phi(a)]) = \int_0^{\sin^2(\frac{\pi}{2}a)} \rho(x)dx \overset{(\star)}{=} a.$$

Both sides of the equality $(\star)$ coincide, since they coincide for $a = 0$ and their derivatives coincide: The derivative of the right-hand side obviously equals 1. For the left-hand side the chain rule gives

$$\begin{aligned} \frac{d}{da}\int_0^{\sin^2(\frac{\pi}{2}a)} \frac{dx}{\pi\sqrt{x(1-x)}} &= \frac{2\sin(\frac{\pi}{2}a)\cos(\frac{\pi}{2}a)\frac{\pi}{2}}{\pi\sqrt{\sin^2(\frac{\pi}{2}a)(1-\sin^2(\frac{\pi}{2}a))}} \\ &= \frac{\sin(\frac{\pi}{2}a)\cos(\frac{\pi}{2}a)}{\sqrt{\sin^2(\frac{\pi}{2}a)\cos^2(\frac{\pi}{2}a)}} = \frac{\sin(\frac{\pi}{2}a)\cos(\frac{\pi}{2}a)}{\sin(\frac{\pi}{2}a)\cos(\frac{\pi}{2}a)} = 1. \end{aligned}$$

It remains to show that Φ conjugates T and Λ. For $T \circ \Phi$ we obtain

$$T(\Phi(x)) = 4\sin^2(\tfrac{\pi}{2}x)(1-\sin^2(\tfrac{\pi}{2}x)) = 4\sin^2(\tfrac{\pi}{2}x)\cos^2(\tfrac{\pi}{2}x) = \sin^2(\pi x).$$

In the last equality we used the trigonometric identity $2\sin(\cdot)\cos(\cdot) = \sin(2\cdot)$. For $\Phi \circ \Lambda$ we get

$$\Phi(\Lambda(x)) = \sin^2(\tfrac{\pi}{2}\Lambda(x)) = \sin^2(\tfrac{\pi}{2}2x) = \sin^2(\pi x)$$

for $x \in [0, \frac{1}{2}]$ and

$$\begin{aligned}\Phi(\Lambda(x)) &= \sin^2(\tfrac{\pi}{2}\Lambda(x)) = \sin^2(\tfrac{\pi}{2}(2-2x)) \\ &= \sin^2(\pi(1-x)) = \sin^2(\pi - \pi x) = \sin^2(\pi x)\end{aligned}$$

for $x \in (\frac{1}{2}, 1]$. $\Diamond$

2.27 Example: Let $X = [0,1]$ and consider $S(x) = 2x$ (mod 1) and $T(x) = 4x(1-x)$. (S preserves Lebesgue measure and T preserves $d\mu = \frac{dx}{\pi\sqrt{x(1-x)}}$.) Define $\Psi : (X, dx) \to (X, d\mu)$ by

$$\Psi(x) = \sin^2(\pi x).$$

For almost all x Ψ is two-to-one, hence not an isomorphism. So it can only be a semi-conjugacy. Ψ is surjective and it is measure-preserving: The length of $\Psi^{-1}([0, \Psi(a)])$ is $2a$. We have to show that it equals

$$\int_0^{\sin^2(\pi a)} \rho(x)dx = \mu([0, \psi(a)]).$$

This is proven with the same arguments as in the preceding example. It is left to show that the conjugation property holds:

$$T(\Psi(x)) = 4\sin^2(\pi x)(1 - \sin^2(\pi x)) = \sin^2(2\pi x) = \Psi(S(x)),$$

since for $x \in [0, \frac{1}{2}]$ we have $\Psi(S(x)) = \sin^2(2\pi x)$ and for $x \in (\frac{1}{2}, 1]$

$$\begin{aligned}\Psi(S(x)) &= \sin^2(\pi(2x-1)) = \sin^2(2\pi x - \pi) \\ &= (-\sin(2\pi x))^2 = \sin^2(2\pi x).\end{aligned}$$

$\Diamond$

2.28 Example: Consider again $S(x) = 2x$ (mod 1) on $[0,1)$, identified with the unit circle, with Lebesgue measure and $T(x) = \frac{1}{2}(x - \frac{1}{x})$ on $\mathbb{R}$ with invariant measure $\frac{dx}{\pi(1+x^2)}$. It can be shown that they are conjugate via

$$\Phi(x) = -\cot(\pi x).$$

$\Diamond$

2.6 Coding Maps

Idea: Use Shift Transformations to describe transformations.

2.29 Definition: *Let T be measure preserving on a probability space (X, μ). A partition $\mathcal{P} = \{E_0, E_1, \dots, E_k\}$ is called* **generating** *if the subsets of the form*

$$E_{i_1} \cap T^{-1}(E_{i_2}) \cap \cdots \cap T^{-(n-1)}(E_{i_n}), \quad i_j \in \{0, 1, \dots, k\},$$

generate the σ-algebra of X.

Example: Consider a partition $\{E_0, E_1\}$ of $[0,1]$ into subintervals and look at

$$E_0, E_1, E_0 \cap T^{-1}(E_0), E_0 \cap T^{-1}(E_1), E_1 \cap T^{-1}(E_0), E_1 \cap T^{-1}(E_1), \dots$$

Then the smallest σ-algebra containing all these sets should be the Borel-σ-algebra. In the following: $\mathcal{P} = \{E_0, E_1\}$.

Let $Y = \prod_1^\infty \{0, 1\}$ and let $\Sigma : Y \circlearrowleft$ be the shift

$$\Sigma : (i_1, i_2, i_3, \dots) \mapsto (i_2, i_3, i_4, \dots).$$

For $x \in X$ define the coding map $\Phi : X \to Y$ by

$$\Phi(x) = (i_1, i_2, i_3, \dots),$$

where $T^{n-1}(x) \in E_{i_n}$ for all $n \geq 1$. Thus, i_n is uniquely determined by x and T. The coding map is almost everywhere injective, if (what we assume)

$$\bigcap_{n=1}^{\infty} T^{-(n-1)}(E_{i_n})$$

contains at most one element x with μ-probability one. Next we define a probability measure on Y: Let $[i_1, \dots, i_n]$ denote a cylinder set in Y, i.e.,

$$[i_1, \dots, i_n] = \{(y_1, y_2, \dots) \in Y \mid y_k = i_k, \ 1 \leq k \leq n\}.$$

Then

$$\Phi\left(E_{i_1} \cap T^{-1}(E_{i_2}) \cap \cdots \cap T^{-(n-1)}(E_{i_n})\right) = [i_1, \dots, i_n].$$

Define ν on Y by

$$\nu([i_1, \dots, i_n]) = \mu\left(E_{i_1} \cap T^{-1}(E_{i_2}) \cap \cdots \cap T^{-(n-1)}(E_{i_n})\right).$$

Then $\Phi : X \to Y$ is measure preserving and

$$\Phi \circ T = \Sigma \circ \Phi.$$

Note that only those cylinder sets get positive measure whose preimage have positive μ-measure. Hence, Φ is an isomorphism. How can one visualize Φ and the measure ν?

$$\begin{array}{ccc} (X,\mu) & \xrightarrow{T} & (X,\mu) \\ \Phi \downarrow & & \downarrow \Phi \\ (Y,\nu) & \xrightarrow[\Sigma]{} & (Y,\nu) \\ \gamma \downarrow & & \downarrow \gamma \\ [0,1] & \xrightarrow[S]{} & [0,1] \end{array}$$

We use that we can get from $Y = \prod_1^\infty \{0,1\}$ to $[0,1]$ via binary expansion: Define $\gamma : Y \to [0,1]$ by $\gamma([i_1, i_2, \ldots]) = \sum_{n=1}^\infty i_n 2^{-n}$ and a measure ν_0 on cylinder sets

$$\nu_0(\gamma([i_1, \ldots, i_n]) := \nu([i_1, \ldots, i_n]).$$

Then, with $S(x) = 2x \pmod 1$ on $[0,1)$, we have

$$\gamma \circ \Sigma = S \circ \gamma.$$

We can visualize ν_0 (and hence μ) using “many points”.

2.30 Example: $X = [0,1]$, $T(x) = 2x \pmod 1$. $\mathcal{P} = \{E_0, E_1\}$ with $E_0 = [0, \frac{1}{2})$, $E_1 = [\frac{1}{2}, 1]$. Then $\Phi(x) = (b_1, b_2, b_3, \ldots)$ for $x = \sum_{n=1}^\infty b_n 2^{-n}$, $b_n \in \{0,1\}$. Then $\gamma \circ \phi = \mathrm{id}$, and ν_0 is Lebesgue measure (i.e., the invariant measure for T). The proof is left as an exercise. $\Diamond$

2.31 Example: Take $X = [0,1] \times [0,1]$ and (Baker)

$$T(x) = \begin{cases} (2x, \frac{1}{2}y) & \text{for } x \in [0, \frac{1}{2}), \\ (2x-1, \frac{1}{2}y + \frac{1}{2}) & \text{for } x \in [\frac{1}{2}, 1]. \end{cases}$$

Take $\mathcal{P} = \{E_0, E_1\}$ with $E_0 = [0, \frac{1}{2}) \times [0,1]$ and $E_1 = [\frac{1}{2}, 1] \times [0,1]$.

FIGURE

Let $Y = \prod_{-\infty}^\infty \{0,1\}$ be the two-sided $(\frac{1}{2}, \frac{1}{2})$-Bernoulli shift space and Σ the shift transformation. Define $\Phi : X \to Y$ by

$$\Phi(x) = (\ldots, i_{-1}, i_0, i_1, i_2, \ldots),$$

where $T^n(x) \in E_{i_n}$. Equivalently, for $(a,b) \in [0,1] \times [0,1]$ with

$$a = \sum_{j=0}^\infty a_j 2^{-j}, \quad b = \sum_{j=0}^\infty b_j 2^{-j},$$

$$\Phi(x) = \Phi((a,b)) = (\ldots, b_{-2}, b_{-1}, a_0, a_1, a_2, \ldots).$$

Thus, Φ is an isomorphism. $\Diamond$

2.32 Example: *(Coding map for the logistic map)*
Let $T(x) = 4x(1-x)$ on $X = [0,1]$. Let $E_0 = [0, \frac{1}{2})$, $E_1 = [\frac{1}{2}, 1]$. Binary sequence (b_n): $T^{n-1}(x) \in E_{b_n}$ ($\Leftrightarrow x \in T^{-(n-1)}(E_{b_n})$). $\Diamond$

3 Birkhoff's Ergodic Theorem

Aim: Let μ be an invariant measure for $T : X \circlearrowleft$ and $f : X \to \mathbb{R}$. We want to compare $\int f d\mu$ and the average value of f along a trajectory $f(T^n(x))$.

Birkhoff (1931): Ergodicity necessary.

3.1 Ergodicity

The following definition is fundamental for the whole theory.

3.1 Definition: *Let $(X, \mathcal{A}, \mu)$ be a probability space and $T : X \circlearrowleft$ a μ-preserving map, i.e., $\mu(T^{-1}(E)) = \mu(E)$ for all $E \in \mathcal{A}$. Then T is called* **ergodic** *if for $E \in \mathcal{A}$ one has*

$$\mu\Big((T^{-1}(E) \setminus E) \cup (E \setminus T^{-1}(E))\Big) = 0 \;\;\Rightarrow\;\; \mu(E) = 0 \text{ or } \mu(E) = 1.$$

For a better understanding of this property it is useful to introduce the following convention: Two sets are said to be equal if they only differ by null sets, formally:

$$A = B \;\text{ if }\; \mu(\underbrace{A \setminus B \cup B \setminus A}_{=:A\Delta B}) = 0.$$

In these terms: *T is ergodic if*

$$T^{-1}(E) = E \;\;\Rightarrow\;\; E = \emptyset \text{ or } E = X.$$

We also call the measure μ ergodic (with respect to T).

3.2 Definition: *If, for a transformation T on $(X, \mathcal{A}, \mu)$, there are disjoint measurable E_j with $\mu(E_j) > 0$,*

$$X = \bigcup_j E_j, \;\; T(E_j) \subset E_j,$$

such that $T|_{E_j} : E_j \to E_j$ is ergodic with respect to the conditional measure μ_{E_j}, then such E_j is called an **ergodic component** *of T.*

The following observation will be useful for a measure μ. Let $E \in \mathcal{A}$ with $\mu(E) > 0$. Then

$$\mu_E(A) := \frac{\mu(A \cap E)}{\mu(E)}, \;\; A \in \mathcal{A},$$

defines a measure, called the **conditional measure**. If μ is a probability measure, then also μ_E is a probability measure. Furthermore, we also may consider μ_E as a measure on E.

3.3 Theorem: *Let T be a transformation on (X,μ). Then the following are equivalent:*

(i) *T is ergodic.*

(ii) *If $\mu(A)>0$, then $\bigcup_{n=1}^{\infty} T^{-n}(A)=X$.*

(iii) *If $\mu(A)>0$ and $\mu(B)>0$, then $\mu(T^{-n}(A)\cap B)>0$ for some $n\geq 1$.*

(iv) *If a measurable function $f:X\to\mathbb{C}$ satisfies $f(T(x))=f(x)$ for almost every $x\in X$, then f is constant almost everywhere.*

Proof: "(i) $\Rightarrow$ (ii)": Put $E:=\bigcup_{n=1}^{\infty} T^{-n}(A)$ for $\mu(A)>0$. Then $T^{-1}(E)=\bigcup_{n=2}^{\infty} T^{-n}(A)\subset E$ and

$$\mu(E\Delta T^{-1}(E))=\mu(E\backslash T^{-1}(E))=\mu(E)-\mu(T^{-1}(E))=0$$

by invariance of μ. Hence, $E=T^{-1}(E)$. Since T is ergodic, $E=\emptyset$ or $E=X$. Since $E\supset T^{-1}(A)$ and $\mu(E)\geq\mu(T^{-1}(A))=\mu(A)>0$, we conclude $\mu(E)=1$, i.e., $E=X$.

"(ii) $\Rightarrow$ (iii)": Let $\mu(A),\mu(B)>0$. Since $\bigcup_{n=1}^{\infty} T^{-n}(A)=X$, $B=\bigcup_{n=1}^{\infty}(T^{-n}(A)\cap B)$, there is $n\geq 1$ with $\mu(T^{-n}(A)\cap B)>0$.

"(iii) $\Rightarrow$ (i)": Suppose $T^{-1}(B)=B$ and $\mu(B)>0$. Let $A:=X\backslash B$. Then $T^{-n}(A)=X\backslash T^{-n}(B)=X\backslash B$. Hence, $\mu(T^{-n}(A)\cap B)=0$ for all $n\geq 1$. Thus, by (iii), $\mu(A)=0$ and hence $\mu(B)=1$.

"(i) $\Rightarrow$ (iv)": Let $f:X\to\mathbb{C}$ be measurable with $f(T(x))=f(x)$ for almost all x. By considering real and imaginary parts separately, we may assume that f is real-valued. Put, for $n\geq 1$, $k\in\mathbb{Z}$,

$$E_{n,k}:=\left\{x\in X\ :\ 2^{-k}k\leq f(x)<2^{-n}(k+1)\right\}.$$

Then $\{E_{n,k}\mid k\in\mathbb{Z}\}$ is a partition of X for every n. Note that

$$\begin{aligned}T^{-1}(E_{n,k}) &= \left\{x\mid 2^{-n}k\leq f(T(x))<2^{-n}(k+1)\right\}\\ &\overset{\text{ass.}}{=} \left\{x\mid 2^{-n}k\leq f(x)<2^{-n}(k+1)\right\}=E_{n,k}.\end{aligned}$$

Hence, by ergodicity, $E_{n,k}$ has measure 0 or 1. More precisely, for each n there is a unique $k_n\in\mathbb{Z}$ such that

$$\mu(E_{n,k_n})=1 \text{ and } \mu(E_{n,k})=0 \text{ for } k\neq k_n.$$

Let $X_0:=\bigcap_{n=1}^{\infty} E_{n,k_n}$. Then $\mu(X_0)=1$ and f is constant on X. (Since all values are contained in an interval of length 2^{-n}, $n\in\mathbb{N}$).

"(iv) $\Rightarrow$ (i)": Suppose $T^{-1}(E) = E$. Then, with $f(x) = \mathbb{1}_E(x)$

$$f(x) = \mathbb{1}_E(x) = \mathbb{1}_E(T(x)) = \begin{cases} 1 & \text{if } T(x) \in E \Leftrightarrow x \in T^{-1}(E) = E, \\ 0 & \text{if } T(x) \notin E \Leftrightarrow x \notin T^{-1}(E) = E. \end{cases}$$

By (iv) $\mathbb{1}_E(x)$ is constant. Hence, either $E = X$ or $E = \emptyset$. $\square$

3.4 Example: Let $X = [0,1]$, $T(x) = x + \theta \pmod 1$, where $\theta \in [0,1]$. The Lebesgue measure is invariant.

Assertion: The Lebesgue measure is ergodic for T iff $\theta \notin \mathbb{Q}$.

Proof: Let $\theta \in \mathbb{Q}$, i.e., $\theta = \frac{p}{q}$, $p, q \in \mathbb{N}$ (w.l.o.g.). Define

$$f(x) := e^{2\pi i q x}, \quad x \in [0,1].$$

f is obviously not constant.

$$f(T(x)) = e^{2\pi i q(x+\theta)} = e^{2\pi i q x} \underbrace{e^{2\pi i p}}_{=1} = e^{2\pi i q x} = f(x).$$

Hence, by Theorem 3.3 (iv) T is not ergodic.

Now let $\theta \in \mathbb{R} \backslash \mathbb{Q}$. We show: For all $f \in L^2(X, \mathbb{C})$ with $f(T(x)) = f(x)$ for all $x \in X$ it follows that f is constant, which implies ergodicity. $L^2(X, \mathbb{C})$ has an inner product, defined by

$$\langle f, g \rangle = \int_0^1 f(x) \overline{g(x)} dx$$

for $f, g \in L^2(X, \mathbb{C})$. Fact: This is a Hilbert space. The norm is $\|f\|_2 = \sqrt{\langle f, f \rangle}$. The following set of elements in $L^2(X, \mathbb{C})$ is orthonormal:

$$f_n(x) := e^{2\pi i n x}, \quad x \in [0,1], \quad n \in \mathbb{Z}.$$

We compute

$$\langle f_n, f_m \rangle = \int_0^1 e^{2\pi i n x} \overline{e^{2\pi i m x}} dx = \int_0^1 e^{2\pi i (n-m) x} dx = \begin{cases} 1 & \text{for } n = m, \\ 0 & \text{for } n \neq m. \end{cases}$$

The latter is true since for $n \neq m$

$$\int_0^1 e^{2\pi i (n-m) x} dx = \frac{1}{2\pi i (n-m)} \left[e^{2\pi i (n-m) x} \right]_0^1 = \frac{1}{2\pi i (n-m)} (1 - 1) = 0.$$

$L^2(X, \mathbb{C})$ is infinite-dimensional, but every element $f \in L^2(X, \mathbb{C})$ can uniquely be written as

$$f = \sum_{n \in \mathbb{Z}} c_n f_n \qquad \text{(Fourier Series)}$$

with $c_n = \langle f, f_n \rangle$. $\{f_n\}_{n\in\mathbb{Z}}$ is a complete ON-system (see also Bachman and Narici [4, pp. 155–157]). Let $f \in L^2(X, \mathbb{C})$. Then

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}, \quad x \in [0,1] = X.$$

We compute

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

Since the c_n are unique, it follows that

$$c_n = c_n e^{2\pi i n\theta} \text{ for all } n \in \mathbb{Z}.$$

If $c_n \neq 0$, then $e^{2\pi i n\theta} = 1$, which implies $n = 0$, since θ is irrational. Hence, $f(x) = c_0$, a constant. $\diamondsuit$

A generalization:

3.5 Theorem: *Let F be a compact Abelian group with Haar measure μ ($\mu(H) = \mu(gH)$ for $H \subset G$, $g \in G$). For each $g \in G$ define*

$$T_g : G \circlearrowleft, \quad T_g(x) = gx, \quad x \in G.$$

Then T_g is ergodic with respect to μ iff

$$\{g^n \mid n \in \mathbb{Z}\}$$

is dense in G.

Observe: $G = \mathbb{R}/\mathbb{Z}$ becomes a topological group under addition modulo one. It is also compact and Abelian. The Haar measure is the Lebesgue measure. A character of G is a homomorphism $\chi : G \to \mathbb{C}\backslash\{0\}$ such that $|\chi(g)| = 1$ for all $g \in G$ ($\chi(g_1 g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$). The characters form a complete ON-system in $L^2(G, \mathbb{C}, \mu)$. In other words: For each $f \in L^2(G, \mathbb{C}, \mu)$ there are unique numbers $\hat{f}(\chi)$, χ a character, such that

$$f(x) = \sum_{\chi} \hat{f}(\chi)\chi(x),$$

where the sum runs over all characters.

3.6 Example: Let $X = [0,1)$ and $T(x) = 2x \pmod 1$. The Lebesgue measure is an invariant ergodic measure.

<u>Proof:</u> Let $f \in L^2(X, \mathbb{C})$ be invariant, i.e.,

$$f(x) = f(T(x)) \text{ for almost all } x \in [0,1).$$

Then $f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}$, $x \in [0,1)$, and the coefficients are unique. Compute

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

Hence, $c_n = 0$, if n is odd. Computation of $f(T^2(x)) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i(4n)x}$ shows that all coefficients c_n with n not a multiple of 4 are equal to zero. Going on this way we find that all c_n are equal to zero except possibly c_0. Hence, $f(x)$ is constant. $\diamondsuit$

3.7 Theorem: *Consider $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$, the n-dimensional torus. Define a multiplication*

$$(y_1,\ldots,y_n)\cdot(z_1,\ldots,z_n) := (y_1+z_1 \text{ (mod 1)}, \ldots, y_n+z_n \text{ (mod 1)}).$$

This makes $\mathbb{T}^n$ into a compact Abelian group. Let $\Phi : \mathbb{T}^n \circlearrowleft$ be a surjective homomorphism given by

$$\Phi(x) = Ax, \quad A \in \mathbb{Z}^{n\times n}.$$

Then Φ is ergodic with respect to Lebesgue measure iff no eigenvalue of A is a root of unity.

3.8 Example: Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. The eigenvalues are given by

$$0 = (2-\lambda)(1-\lambda) - 1 = \lambda^2 - 3\lambda + 1 \Leftrightarrow \lambda_{1/2} = \frac{1}{2}\left(3 \pm \sqrt{5}\right).$$

Hence, the eigenvalues are <u>not</u> roots of unity ($\lambda_1^n, \lambda_2^n \neq 1$ for all $n \in \mathbb{N}$). $\diamondsuit$

3.2 Birkhoff's Ergodic Theorem

3.9 Theorem: *Let (X,μ) be a probability space. If $T : X \circlearrowleft$ preserves the measure μ and $f : X \to \mathbb{R}$ is integrable, then*

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

for all $x \in X$ and for some $f^* \in L^1(X,\mathbb{R},\mu)$ with

$$f^*(T(x)) = f^*(x) \text{ for almost all } x \in X.$$

If T is ergodic, then f^* is constant and

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

for almost all $x \in X$.

Discussion: Let $f = \mathbb{1}_E$, $E \subset X$ measurable. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_E(T^k(x))$$

counts how often $T^k(x)$ visits E in average. If T is ergodic, then this limit equals $\mu(E)$ ("time average = average over space").
Conversely: If E is a measurable invariant set, then $\mathbb{1}_E$ is an invariant function: $\mathbb{1}_E(x) = \mathbb{1}_E(T(x))$ for almost all $x \in X$.

For the proof of Theorem 3.9 we need the following lemma.

3.10 Lemma (Maximal Ergodic Theorem): *Let $T : X \circlearrowleft$ be measure preserving and consider $f : X \to \mathbb{R}$ integrable. Define $f_0 :\equiv 0$, $f_n := f + f \circ T + \cdots + f \circ T^{n-1}$, $n \geq 1$, and $F_N(x) := \max_{0\leq n\leq N} f_n(x)$, $x \in X$. Then*

$$\int_{\{x:\, F_N(x)>0\}} f d\mu > 0 \ \textit{ for all } N \in \mathbb{N}.$$

Proof: Observe that $F_N \in L^1(X,\mu)$ since $f, f\circ T, \ldots$ are integrable. For $0 \leq n \leq N$, $F_N \geq f_n$ and hence,

$$F_N \circ T \geq f_n \circ T$$

Thus,

$$F_N \circ T + f \geq f + f_n \circ T = f_{n+1} \ \text{ for } n = 0, 1, \ldots, N-1.$$

This shows

$$F_N(T(x)) + f(x) \geq \max_{1\leq n\leq N} f_n(x) \ \text{ for all } x \in X.$$

If $F_N(x) > 0$, then the right hand side equals $\max_{0\leq n\leq N} f_n(x) = F_N(x)$. We find

$$f(x) \geq F_N(x) - F_N(T(x)) \ \text{ on } \ \{x \mid F_N(x) > 0\} =: A_N.$$

We compute

$$\int_{A_N} f d\mu \geq \int_{A_N} F_N d\mu - \int_{A_N} F_N \circ T d\mu = 0.$$

The latter is true since T preserves μ. □

3.11 Corollary: *Let $T : X \circlearrowleft$ be measure preserving. If $g : X \to \mathbb{R}$ is integrable and*

$$B_\alpha := \left\{ x \in X \ : \ \sup_{n\geq 1} \frac{1}{n} \sum_{k=0}^{n-1} g(T^k(x)) > \alpha \right\}, \ \ \alpha \in \mathbb{R},$$

then
$$\int_{B_\alpha} g d\mu \geq \alpha\mu(B_\alpha).$$
If $T^{-1}(A) = A$ *for some* $A \subset X$ *(measurable), then*
$$\int_{B_\alpha \cap A} g d\mu \geq \alpha\mu(B_\alpha \cap A).$$

Proof: The second assertion is immediate from the first one, if we apply it to A instead of X. Apply the Maximal Ergodic Theorem to
$$f := g - \alpha.$$
Then
$$B_\alpha := \bigcup_{N=0}^{\infty} \{x \in X \mid F_N(x) > 0\}$$
and
$$\int_{\{x\colon F_n(x)>0\}} f d\mu \geq 0 \text{ for all } N \quad \Rightarrow \quad \int_{B_\alpha} f d\mu \geq 0.$$
Hence, $\int_{B_\alpha} g d\mu - \int_{B_\alpha} \alpha d\mu \geq 0$, i.e.,
$$\int_{B_\alpha} g d\mu \geq \alpha\mu(B_\alpha).$$

$\square$

Proof of Birkhoff's Ergodic Theorem:
Define
$$f^*(x) := \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)),$$
$$f_*(x) := \liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)).$$

This implies
$$\begin{aligned} f^*(T(x)) &= \limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^{k+1}(x)) \\ &= \limsup_{n\to\infty} \frac{n+1}{n+1}\frac{1}{n} \left[f(T(x)) + \cdots + f(T^n(x))\right] \\ &= \limsup_{n\to\infty} \frac{n+1}{n+1}\frac{1}{n} \left[x + f(T(x)) + \cdots + f(T^n(x))\right] - \underbrace{\frac{x}{n}}_{\to 0} \\ &= \limsup_{n\to\infty} \underbrace{\frac{n+1}{n}}_{\to 1} \frac{1}{n+1} \sum_{k=0}^{n} f(T^k(x)) = f^*(x). \end{aligned}$$

We still have to show that $f^*(x) = f_*(x)$ and they are integrable. Put

$$E_{\alpha,\beta} := \{x \in X \mid f_*(x) < \beta \text{ and } \alpha < f^*(x)\}, \quad \alpha, \beta \in \mathbb{Q}.$$

We want to show that

$$\{x \mid f_*(x) < f^*(x)\}$$

has measure 0. Then for $\alpha, \beta \in \mathbb{Q}$

$$\{x \mid f_*(x) < f^*(x)\} = \bigcup_{\beta<\alpha} E_{\alpha,\beta} \text{ (countable union).}$$

We find

$$\begin{aligned} T^{-1}(E_{\alpha,\beta}) &= \{x \mid f_*(T(x)) < \beta \text{ and } \alpha < f^*(T(x))\} \\ &= \{x \mid f_*(x) < \beta \text{ and } \alpha < f^*(x)\} = E_{\alpha,\beta}. \end{aligned}$$

Put

$$B_\alpha := \left\{ x \ : \ \sup_{n\geq 1} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) > \alpha \right\}.$$

Then $E_{\alpha,\beta} \subset B_\alpha$. By Corollary 3.11

$$\int_{E_{\alpha,\beta}} f d\mu = \int_{B_\alpha \cap E_{\alpha,\beta}} f d\mu \geq \alpha\mu(B_\alpha \cap E_{\alpha,\beta}) = \alpha\mu(E_{\alpha,\beta}).$$

Note that $(-f)^* = -f_*$, $(-f)_* = -f^*$ and

$$E_{\alpha,\beta} = \{x \mid (-f)^*(x) > -\beta \text{ and } -\alpha > (-f)_*(x)\}.$$

Replace f, α, β by $-f, -\beta, -\alpha$. Then

$$\int_{E_{\alpha,\beta}} (-f) d\mu \geq -\beta\mu(E_{\alpha,\beta}) \ \Rightarrow \ \int_{E_{\alpha,\beta}} f d\mu \leq \beta\mu(E_{\alpha,\beta}).$$

If $\alpha > 0$, then $\mu(E_{\alpha,\beta}) = 0$.

$$\{x \mid f_*(x) < f^*(x)\} = \bigcup_{\beta<\alpha} E_{\alpha,\beta} \ \Rightarrow \ \mu(\{x \mid f_*(x) < f^*(x)\}) = 0.$$

Hence,

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

Next we show: f^* is integrable: Let

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

Then (since T leaves μ invariant)

$$\int_X g_n d\mu \le \frac{1}{n}\sum_{k=0}^{n-1}\underbrace{\int_X |f(T^k(x))|d\mu}_{=\int_X |f|d\mu} = \int_X |f|d\mu < \infty.$$

Now (by Fatou's Lemma)

$$\int_X |f^*|d\mu = \int_X |f_*|d\mu = \int_X \liminf_{n\to\infty} g_n d\mu \le \liminf_{n\to\infty}\int_X g_n d\mu.$$

Hence, f^* is integrable. It remains to show that

$$\int_X f d\mu = \int_X f^* d\mu. \tag{2}$$

Since μ is ergodic and $f^*(T(x)) = f^*(x)$, by Theorem 3.3 we get that f^* is constant almost everywhere. This implies

$$\int_X f^* d\mu = f^*(x)\underbrace{\mu(X)}_{=1} = \int_X f d\mu.$$

For the proof of (2) define for all $n \ge 1$ and $k \in \mathbb{Z}$ the set

$$D_{n,k} := \left\{ x \in X \mid \tfrac{k}{n} \le f^*(x) < \tfrac{k+1}{n} \right\}.$$

For fixed n, X is the disjoint union of the set $D_{n,k}$, $k \in \mathbb{Z}$. $D_{n,k}$ is invariant, since $f^*(x) = f^*(T(x))$ implies

$$T^{-1}(D_{n,k}) = \{x \in X \mid T(x) \in D_{n,k}\} = D_{n,k}.$$

For $\varepsilon > 0$ small enough

$$D_{n,k} \subset B_{\frac{k}{n}-\varepsilon} = \left\{ x \in X \ : \ \sup_{n\ge 1}\frac{1}{n}\sum_{i=0}^{k-1} f(T^i(x)) > \frac{k}{n} - \varepsilon \right\}.$$

By Corollary 3.11 we obtain

$$\int_{D_{n,k}} f d\mu \ge \left(\tfrac{k}{n} - \varepsilon\right)\mu(D_{n,k}) \ \text{ for all } \varepsilon > 0.$$

Thus,

$$\int_{D_{n,k}} f d\mu \ge \tfrac{k}{n}\mu(D_{n,k}).$$

Together with the definition of $D_{n,k}$ this yields

$$\int_{D_{n,k}} f^* d\mu \le \tfrac{k+1}{n}\mu(D_{n,k}) = \tfrac{1}{n}\mu(D_{n,k}) + \int_{D_{n,k}} f d\mu.$$

Summation over k yields

$$\int_X f^* d\mu \leq \tfrac{1}{n} + \int_X f d\mu \text{ for all } n \in \mathbb{N}.$$

For n tending to ∞ we get the inequality

$$\int_X f^* d\mu \leq \int_X f d\mu.$$

The same procedure for $-f$ gives

$$\int_X (-f)^* d\mu \leq \int_X (-f) d\mu \quad \Rightarrow \quad \int_X f^* d\mu \geq \int_X f d\mu.$$

This finishes the proof. □

3.12 Remarks:

- We have shown:
$$\int_X f d\mu = \int_X f^* d\mu \text{ for all } \mu.$$
Hence, by Lebesgue's Theorem on Dominated Convergence we obtain for f bounded:
$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) - f^*(x) \right\|_1 = \int_X \left| \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) - f^*(x) \right| d\mu \xrightarrow{n\to\infty} 0,$$
i.e., convergence in $L^1(X, \mu)$.
- A stochastic interpretation of Birkhoff's Ergodic Theorem: Let (Ω, P) be a probability space. $A, B \subset \Omega$ are called independent if $P(A \cap B) = P(A)P(B)$. $A_i \subset \Omega$, $i \in \mathbb{N}$, are called independent if for all $1 \leq i_1 < i_2 < \cdots < i_k$
$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$
A sequence of integrable functions $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ is independent, if
$$\left\{ X_i^{-1}(B_i) \right\}_{i=1}^{\infty}, \quad B_i \subset \mathbb{R},$$
are independent for all $\{B_i\}$ in $\mathbb{R}$.
The distribution of an integrable function $X : \Omega \to \mathbb{R}$ is
$$P_X(A) := P(X^{-1}(A)), \quad A \subset \mathbb{R}.$$
The *Strong Law of Large Numbers* says: Let $X_1, X_2, \ldots$ be independent integrable functions from Ω to $\mathbb{R}$ with identical distribution P_X. Let the mean be $\int_{\mathbb{R}} x dP_X(x)$. Then
$$\frac{1}{n} [X_1(\omega) + \cdots + X_n(\omega)] \to \int_{\mathbb{R}} x dP_X(x)$$
for P-almost all $\omega \in \Omega$.

3.13 Theorem: *Let $T : X \circlearrowleft$ be μ-preserving for a probability measure μ. Then μ is ergodic if and only if for all measurable $A, B \subset X$*

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

i.e., convergence in average.

Proof: "$\Rightarrow$": Take $f := \mathbb{1}_A$ in Birkhoff's Theorem. Then

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

Thus,

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

for almost all $x \in X$. By Lebesgue's Theorem on Dominated Convergence (LDC) we get

$$\begin{aligned} \mu(A)\mu(B) &= \int_X \mu(A)\mathbb{1}_B(x)d\mu \\ &\overset{\text{LDC}}{=} \lim_{n\to\infty} \frac{1}{n} \int_X \underbrace{\mathbb{1}_A(T^i(x))\mathbb{1}_B(x)}_{=\mathbb{1}_{T^{-i}(A)\cap B}(x)} d\mu \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(T^{-i} \cap B). \end{aligned}$$

"$\Leftarrow$": Suppose that $T^{-1}(E) = E$. Taking $A = B = E$ we get

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

Hence, $\mu(E) = 0$ or $\mu(E) = 1$. □

The following theorem is stated without proof.

3.14 Theorem (The Mean Ergodic Theorem, von Neumann): *Let (X, μ) be a measure space with $X = \bigcup_{i=1}^{\infty} X_i$ with $\mu(X_i) < \infty$ (σ-finite). Let $T : X \circlearrowleft$ be measure-preserving for $f \in L^2(X, \mu)$. Then there is a T-invariant function $\tilde{f} \in L^2(X, \mu)$ with*

$$\lim_{n\to\infty} \left\| \frac{1}{n} \sum_{i=1}^{n} f \circ T^{i-1} - \tilde{f} \right\|_2 = 0.$$

3.15 Theorem (Borel): *Suppose T is continuous on a compact metric space X, and let $\{T^n\}_{n\in\mathbb{N}}$ be uniformly equicontinuous, i.e.,*

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall n \in \mathbb{N} : \forall x, y \in X : d(x,y) < \delta \Rightarrow d(T^n(x), T^n(y)) < \varepsilon.$$

If μ is ergodic for T and $\mu(U) > 0$ for all nonempty open sets $U \subset X$, then for all continuous $f : X \to \mathbb{R}$ and every $x \in X$

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

Proof: By Birkhoff's Theorem the assertion is true for all x outside of N with $\mu(N) = 0$. Since μ is positive on nonempty open sets, the interior of N is empty, hence the assertion holds on a dense set $X_0 \subset X$. Let $x \in X$ and $\varepsilon > 0$. Since $\{T^n\}_{n\in\mathbb{N}}$ is uniformly equicontinuous by assumption and f is uniformly continuous, since X is compact, there is $\delta > 0$ such that

$$d(x,y) < \delta \quad \Rightarrow \quad \sup_{k\geq 0} |f(T^k(x)) - f(T^k(y))| < \varepsilon.$$

Choose a point $y \in X_0$ such that $d(x,y) < \delta$. Then for all $n \geq 0$

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) - \int_X f d\mu \right| \leq \underbrace{\left| \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) - f(T^k(y)) \right|}_{<\varepsilon} + \left| \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(y)) - \int_X f d\mu \right| \xrightarrow{n\to\infty} 0.$$

The second summand tends to zero, since $y \in X_0$. Since ε is arbitrary, the assertion holds. □

3.16 Theorem (Kronecker-Weyl): *For an irrational number $\theta \in (0,1)$ we have*

$$\lim_{n\to\infty} \frac{1}{n} card\,\{k \in [0,n] \cap \mathbb{Z} \mid \{k\theta\} \in I\} = length\ of\ I,$$

for each interval $I \subset (0,1)$, where $\{\cdot\}$ denotes the fractional part of a real number.

Proof: The map $T(x) = x + \theta \pmod 1$, $T : [0,1) \circlearrowleft$, is ergodic with respect to Lebesgue measure (see Exercise 3.4). We identify $[0,1)$ with the (compact) unit circle. The family $(T^n)_{n\in\mathbb{N}}$ of iterates of T is uniformly equicontinuous, since

$$|T^n(x) - T^n(y)| = |(x + n\theta) - (y + n\theta)| = |x - y|$$

for all $x, y \in [0,1)$ and $n \in \mathbb{N}$. Moreover, the Lebesgue measure has the property that all nonempty open sets have positive measure. Hence, Theorem 3.15 implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_I(T^k(x)) = \lambda(I)$$

for all $x \in [0,1)$ and all intervals $I \subset [0,1)$. For $x = 0$ we have

$$\begin{aligned} \sum_{k=0}^{n-1} \mathbb{1}_I(T^i(0)) &= \sum_{k=0}^{n-1} \mathbb{1}_I(k\theta \pmod 1) \\ &= \operatorname{card}\{k \in \{0,1,\ldots,n-1\} \mid \{k\theta\} \in I\}. \end{aligned}$$

This implies the assertion. □

Question: When are Markov shifts ergodic?

Let P be a $N \times N$-stochastic matrix, i.e., $P = (p_{ij})$, $p_{ij} \geq 0$, $\sum_j p_{ij} = 1$ for all j (the row sums are 1). p_{ij} is interpreted as the probability to go from i to j. Then $X = \prod_{i=1}^{\infty}\{1,\ldots,N\}$ with the shift $\theta : X \circlearrowleft$, $\theta(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots)$. An invariant measure μ is defined by its values on cylinder sets, i.e., by $\mu([a_1, a_2, \ldots, a_n])_{t,t+1,\ldots,t+n-1}$. By Theorem 2.20 (ii) there is a vector $\pi = (\pi_1, \ldots, \pi_N)$ with $\pi \geq 0$, $\sum_{i=1}^N \pi_i = 1$, such that $\pi P = \pi$. π is unique, if P is irreducible, i.e., for all i, j there is $m \in \mathbb{N}$ with

$$(P^m)_{ij} =: p_{ij}^m > 0.$$

The measure μ of cylinder sets is then defined by

$$\mu([a_1, a_2, \ldots, a_N])_{t,t+1,\ldots,t+n-1} = \pi_{a_1} p_{a_1 a_2} p_{a_2 a_3} \cdot \ldots \cdot p_{a_{n-1} a_n}.$$

3.17 Theorem: *A Markov shift is ergodic iff P is irreducible.*

Proof: We only prove the backward direction "⇐": First recall that $p_{ij}^k = (P^k)_{ij}$ is the probability of $\{i_n = j \mid i_0 = i\}$. Define

$$E_i := \{x \in X \mid x_0 = i\}, \quad i = 1, \ldots, N.$$

Birkhoff's Theorem implies that $\frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}_{E_i}(\theta^k(x))$ exists for almost all $x \in X$ and the limit is integrable. Hence, using dominated convergence, there exist

$$\begin{aligned} q_{ij} &:= \frac{1}{\pi_i} \int_X \left[\lim_{n\to\infty} \mathbb{1}_{E_j}(\theta^k(x)) \cdot \mathbb{1}_{E_i}(x)\right] d\mu \\ &= \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu\left(\theta^{-k}(E_j) \cap E_i\right) \\ &= \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \pi_i p_{ij}^k = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^k. \end{aligned}$$

The matrix $Q = (q_{ij})$ is stochastic, i.e., $q_{ij} \geq 0$,

$$\sum_j q_{ij} = \sum_j \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^k = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \underbrace{\sum_j p_{ij}^k}_{=1} = 1,$$

since P is stochastic. Furthermore $QP = PQ = Q$, since $Q = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k$, and $Q^2 = Q$. The latter follows from

$$Q \frac{1}{n} \sum_{k=0}^{n-1} P^k = \frac{1}{n} \sum_{k=0}^{n-1} \underbrace{QP^k}_{=Q} = Q$$

by letting n tend to infinity.

<u>Claim:</u> If P is irreducible, then all entries q_{ij} of Q are positive and all rows of Q are identical and each row of Q equals π.

<u>Proof:</u> $Q = QP$ implies that for fixed i and j $q_{ij} = \sum_k q_{ik} p_{kj}^n \geq q_{ik} p_{kj}^n$ for all k and n. Define

$$F_i := \{ j \mid q_{ij} > 0 \}.$$

Then

$$k \in F_i \text{ and } p_{kj}^n > 0 \quad \Rightarrow \quad j \in F_i \tag{3}$$

and $F_i \neq \emptyset$, since some q_{ik} is positive. By irreducibility there is an n with $p_{kj}^n > 0$. Again, by irreducibility of P, (3) implies that $F_i = \{1, \ldots, N\}$. All rows of Q are identical: If not, there are j_0, k_0 such that $q_{j_0 k_0} < \max_i q_{ik_0} =: q$. Since $Q^2 = Q$ we have for all i:

$$q_{ik_0} = \sum_j q_{ij} \underbrace{q_{jk_0}}_{\leq q} < q \underbrace{\sum_j q_{ij}}_{=1} = q.$$

This is impossible. Next we show that for all i and j $q_{ij} = \pi_j$. Compute

$$(\pi Q)_j = \sum_i \pi_i q_{ij} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \sum_j \pi_i p_{ij}^k = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (\pi P^k)_j = \pi_j.$$

The latter equality holds true since $\pi P = \pi$, $(\pi Q)_j = \sum_{i=1}^N \pi_i q_{ij}$, and q_{ij} is independent of i. Hence,

$$(\pi Q)_j = \underbrace{\left(\sum_{i=1}^N \pi_i \right)}_{=1} q_{ij} = q_{ij},$$

which implies $q_{ij} = \pi_j$ for all i and j.

By Theorem 3.13 ergodicity of the Markov shift follows if

$$\frac{1}{n}\sum_{k=0}^{n-1}\mu(\theta^{-k}(E)\cap F) \xrightarrow{n\to\infty} \mu(E)\mu(F)$$

for all measurable sets E and F. It suffices to prove this property for cylinder sets E and F: Let

$$\begin{aligned} E &= \{x \mid (x_r,\dots,x_{r+l}) = (i_0,i_1,\dots,i_l)\}, \\ F &= \{x \mid (x_s,\dots,x_{s+m}) = (j_0,j_1,\dots,j_m)\}, \end{aligned}$$

for given symbols $i_0,i_1,\dots,i_r,j_0,j_1,\dots,j_m \in \{1,\dots,N\}$. For k large enough we have

$$(\{r,r+1,\dots,r+l\}+k)\cap\{s,s+1,\dots,s+m\} = \emptyset.$$

Then

$$\mu\left(\theta^{-k}(E)\cap F\right) = \pi_{j_0}p_{j_0j_1}\cdot\ldots\cdot p_{j_{m-1}}p_{j_m}\left(p_{j_m}^{k-m}p_{i_0i_1}\cdot\ldots\cdot p_{i_{l-1}i_l}\right)$$

and

$$\frac{1}{n}\sum_{k=0}^{n-1}\mu\left(\theta^{-k}(E)\cap F\right) = \pi_{j_0}p_{j_0j_1}\cdot\ldots\cdot p_{j_{m-1}j_m}p_{i_0i_1}\cdot\ldots\cdot p_{i_{l-1}i_l}\underbrace{\left(\frac{1}{n}\sum_{k=0}^{n-1}p_{j_m i_0}^{k-m}\right)}_{\to q_{j_m i_0}=\pi_{i_0}}$$

The right hand side is

$$\mu(E)\mu(F) = \pi_{i_0}p_{i_0i_1}\cdot\ldots\cdot p_{i_{l-1}i_l}\pi_{j_0}p_{j_0j_1}\cdot\ldots\cdot p_{j_{m-1}j_m}.$$

This finishes the proof. □

3.3 Absolutely Continuous and Singular Invariant Measures

3.18 Theorem: *Let (X,μ) be a probability space and let $T : X \circlearrowleft$ be a μ-preserving ergodic transformation. Suppose that $\rho \in L^1(X,\mu)$ satisfies $\rho(x) \geq 0$ for μ-almost all $x \in X$ and $\int_X \rho d\mu = 1$. If T is also ergodic with respect to the measure $d\nu = \rho d\mu$, then $\rho(x) = 1$ for almost all $x \in X$.*

Proof: Let $E \subset X$ be measurable and let

$$\begin{aligned} X_1 &= \left\{x\in X \,:\, \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{1}_E(T^k(x)) = \mu(E)\right\}, \\ X_2 &= \left\{x\in X \,:\, \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{1}_E(T^k(x)) = \nu(E)\right\}. \end{aligned}$$

Then by Birkhoff's Ergodic Theorem

$$\mu(X_1) = 1 \text{ and } \nu(X_2) = 1.$$

By definition

$$1 = \nu(X_2) = \int_{X_2} \rho d\mu.$$

Hence, $\mu(X_2) > 0$. Since $\mu(X_1 \cap X_2) = \mu(X_2) > 0$ it follows that $X_1 \cap X_2 \neq \emptyset$. Choose $x \in X_1 \cap X_2$. Then

$$\mu(E) = \lim_{n\to\infty} \sum_{k=0}^{n-1} \mathbb{1}_E(T^k(x)) = \nu(E) = \int_E \rho d\mu.$$

This holds for all E, hence $\rho(x) = 1$ μ-almost everywhere. □

3.19 Example: (*Solenoid*)
Let $S^1 = \{\phi \mid 0 \le \phi < 1\}$ be the unit interval identified with the unit circle, and let

$$D = \{(u, v) \in \mathbb{R}^2 \mid u^2 + v^2 \le 1\}$$

be the unit disk. Observe that S^1 with addition modulo 1 is a group. Consider

$$X := S^1 \times D$$

identified with the solid torus in $\mathbb{R}^3$. For $0 < a < \frac{1}{2}$ define the *solenoid map* $T : X \circlearrowleft$ by

$$T(\phi, u, v) := \left(2\phi, au + \tfrac{1}{2}\cos(2\pi\phi), av + \tfrac{1}{2}\sin(2\pi\phi)\right).$$

The image of T is contained in X, since

$$\begin{aligned}\left(au + \tfrac{1}{2}\cos(2\pi\phi)\right)^2 + \left(av + \tfrac{1}{2}\sin(2\pi\phi)\right)^2 &= a^2(\underbrace{u^2 + v^2}_{\le 1}) \\ &\quad + a(\underbrace{u\cos(2\pi\phi) + v\sin(2\pi\phi)}_{\le 1}) + \tfrac{1}{4} \\ &\le a^2 + a + \frac{1}{4} < 1.\end{aligned}$$

In fact $T(x) \subset \operatorname{int}(X)$. T is injective: Suppose $T(\phi_1, u_1, v_1) = T(\phi_2, u_2, v_2)$. Hence, $2\phi_1 = 2\phi_2$ (mod 1). If $\phi_1 = \phi_2$, then $au_1 = au_2$ and $av_1 = av_2$, hence $u_1 = u_2$ and $v_1 = v_2$. Else $2\phi_1 = 2\phi_2 \pm 1$. Hence, $\phi_1 - \phi_2 = \pm\frac{1}{2}$.

$$\begin{aligned}au_1 + \tfrac{1}{2}\cos(2\pi\phi_1) &= au_2 + \tfrac{1}{2}\cos(2\pi\phi_2) \\ &= au_2 + \tfrac{1}{2}\cos(2\pi(\phi_1 \pm \tfrac{1}{2})) = au_2 - \tfrac{1}{2}\cos(2\pi\phi_1).\end{aligned}$$

Analogously, $av_1 + \frac{1}{2}\sin(2\pi\phi_1) = av_2 - \frac{1}{2}\sin(2\pi\phi_1)$. Thus, $a(u_1 - u_2) = -\cos(2\pi\phi_1)$ and $a(v_1 - v_2) = -\sin(2\pi\phi_1)$. Thus,

$$a^2\left[(u_1 - u_2)^2 + (v_1 - v_2)^2\right] = 1.$$

This is impossible, since $a^2 < \frac{1}{4}$ and $(u_1 - u_2)^2 + (v_1 - v_2)^2 \leq 1$. We have

$$T^{n+1}(X) \subset \operatorname{int} T^n(x) \ \ \forall n \geq 0.$$

The solenoid is $S := \bigcap_{n=0}^{\infty} T^n(X)$. S is nonempty, since it is the intersection of a decreasing sequence of compact sets. Then $T|_S$ is bijective.

FIGURE

Question: Does there exist an invariant measure on S?

This question is answered by the following Theorem. $\Diamond$

3.20 Theorem (Krylov-Bogolyubov): *Let $T : X \circlearrowleft$ be continuous on a compact metric space X. Then there exists a T-invariant probability measure on X (i.e., on the Borel-σ-algebra of X).*

Proof: The proof is based on the following facts from Functional Analysis:

(i) Let $L : C(X, \mathbb{R}) \to \mathbb{R}$ be a continuous linear operator with $Lf \geq 0$ if $f \geq 0$. Then there exists a (unique) finite measure on X such that for every $f \in C(X, \mathbb{R})$

$$Lf = \int_X f d\mu.$$

(*Riesz Representation Theorem*).

(ii) Let $\mathrm{pm}(X)$ be the set of all probability measures on X and (μ_n) a sequence in $\mathrm{pm}(X)$. Then there are $\mu \in \mathrm{pm}(X)$ and a subsequence (μ_{n_k}) such that for every $f \in C(X, \mathbb{R})$

$$\int_X f d\mu_{n_k} \xrightarrow{k\to\infty} \int_X f d\mu.$$

(Weak compactness of $\mathrm{pm}(X)$).

(iii) If $g : X \to \mathbb{C}$ is integrable, then for every $\varepsilon > 0$ there is a set $N \subset X$ with $\mu(N) < \varepsilon$ and a continuous function $f : X \to \mathbb{C}$ such that $g(x) = f(x)$ for all $x \in X \backslash N$. (*Lusin's Theorem*).

Fix $n \in \mathbb{N}$ and $x \in X$ and define

$$L_n : C(X, \mathbb{R}) \to \mathbb{R}, \ \ L_n f := \frac{1}{n}\sum_{k=0}^{n-1} f(T^k(x)), \ \ f \in C(X, \mathbb{R}).$$

L_n is linear and continuous with $L_n f \geq 0$ for $f \geq 0$. By (i) there is a finite measure μ_n with

$$L_n f = \int_X f d\mu_n \text{ for all } f \in C(X, \mathbb{R}).$$

Since

$$\mu(X) = \int_X d\mu = L_n 1 = 1,$$

μ_n is a probability measure. By (ii) there is a probability measure μ and a subsequence (μ_{n_k}) with

$$L_{n_k} f = \int_X f d\mu_{n_k} \overset{k\to\infty}{\longrightarrow} \int_X f d\mu \text{ for all } f \in C(X, \mathbb{R}). \tag{4}$$

It remains to show that μ is T-invariant. Let $f \in C(X, \mathbb{R})$. Then for all $k \in \mathbb{N}$

$$\left| \frac{1}{n_k} \sum_{j=0}^{n_k-1} f(T^j(x)) - \frac{1}{n_k} \sum_{j=0}^{n_k-1} f(T^j(T(x))) \right| \leq \frac{1}{n_k} |f(x) - f(T^{n_k}(x))| \leq \frac{2\|f\|_\infty}{n_k}. \tag{5}$$

This implies that for all $\varepsilon > 0$ and $k \in \mathbb{N}$ large enough

$$\begin{aligned} \left| \int_X f d\mu - \int_X f \circ T d\mu \right| \leq & \left| \int_X f d\mu - \int_X f d\mu_{n_k} \right| \\ & + \left| \int_X f d\mu_{n_k} - \int_X f \circ T d\mu_{n_k} \right| \\ & + \left| \int_X f \circ T d\mu_{n_k} - \int_X f \circ T d\mu \right|. \end{aligned}$$

The first and third summand can be made smaller than $\frac{\varepsilon}{3}$ by (4), and the second summand by (5). This shows that $\int_X f d\mu = \int_X f \circ T d\mu$ for all $f \in C(X, \mathbb{R})$. Let $A \subset X$ be measurable. From (iii) we can conclude that

$$\int_X \mathbb{1}_A d\mu = \int_X \mathbb{1}_{T^{-1}(A)} d\mu.$$

holds, which implies $\int_X f d\mu = \int_X f \circ T d\mu$ for all integrable f. By Theorem 2.5 this proves that μ is T-invariant. □

4 More on Ergodicity

4.1 Mixing

Recall Theorem 3.13: An invariant measure is ergodic for $T : X \circlearrowleft$ iff for all measurable $A, B \subset X$

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

4.1 Definition: *A measure-preserving transformation T on (X,μ) is called* **mixing***, if for all measurable sets $A, B \subset X$*

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

Clearly, mixing transformations are ergodic. Question: Is mixing stronger than ergodicity?

We will show that $T : [0,1) \circlearrowleft$, $x \mapsto x + \theta$ (mod 1), θ irrational, which we know is ergodic, is not mixing.

<u>Recall:</u> μ is T-invariant iff $U_T f = f \circ T$ is norm-preserving on $L^2(X,\mathbb{C},\mu)$, i.e., $\int_X |f|^2 d\mu = \int_X |f \circ T|^2 d\mu$ for all $f \in L^2(X,\mathbb{C},\mu)$. The inner product on $L^2(X,\mathbb{C},\mu)$ is given by

$$(f,g)_{L^2} = \int_X f(x)\overline{g(x)} d\mu(x).$$

4.2 Theorem: *Let (X,μ) be a probability space and $T : X \circlearrowleft$ μ-preserving. Then the following are equivalent:*

(i) *T is mixing.*

(ii) *For $f \in L^2(X,\mathbb{C},\mu)$: $\lim_{n\to\infty}(U_T^n f, f) = (f,1)(1,f)$.*

(iii) *For $f, g \in L^2(X,\mathbb{C},\mu)$: $\lim_{n\to\infty}(U_T^n f, g) = (f,1)(1,g)$.*

Proof: "(i) $\Rightarrow$ (ii)": Write U instead of U_T. First let f be a simple function, i.e., $f = \sum_{i=1}^k c_i \mathbb{1}_{E_i}$, $E_i \subset X$ measurable, $c_i \in \mathbb{C}$. Then

$$Uf = \sum_{i=1}^k c_i \mathbb{1}_{T^{-1}(E_i)}$$

and

$$\begin{aligned} (U^n f, f) &= \int_X U^n f \overline{f} d\mu = \sum_{i,j=1}^k c_i \overline{c_j} \mu(T^{-n}(E_i) \cap E_j) \\ &\overset{n\to\infty}{\longrightarrow} \sum_{i,j=1}^k c_i \overline{c_j} \mu(E_i)\mu(E_j) = \left(\int_X f \cdot 1 d\mu\right)\left(\int_X 1 \cdot \overline{f} d\mu\right) \\ &= (f,1)\cdot(1,f). \end{aligned}$$

In order to prove this for general $f \in L^2$, we need the Cauchy-Schwarz Inequality:

$$|(g,h)| \le \|g\|_2 \|h\|_2 \ \text{ with equality iff } c_1|g(x)|^2 = c_2|h(x)|^2 \ \forall x \in X.$$

Apply Cauchy-Schwarz to $h, 1$. Then

$$|(h,1)| \le \|h\|_2 \underbrace{\|1\|_2}_{=1} = \|h\|_2,$$

since $\mu(X) = 1$. Take $g \in L^2(X, \mathbb{C}, \mu)$. Since the simple functions are dense in $L^2(X, \mathbb{C}, \mu)$, there is a simple function f with $\|g - f\|_e < \varepsilon$, $\varepsilon > 0$ arbitrary. Let $n \in \mathbb{N}$ with

$$|(U^n f, f) - (f, 1)(1, f)| < \varepsilon.$$

We also have

$$\|U^n f - U^n g\|_2 = \|f - g\|_2 < \varepsilon.$$

Cauchy-Schwarz implies:

$$\begin{aligned}
|(U^n f, f) - (U^n g, g)| &= |(U^n f, f) - (U^n f, g) + (U^n f, g) - (U^n g, g)| \\
&\leq |(U^n f, f - g)| + |(U^n (f - g), g)| \\
&\leq \|U^n f\|_2 \|f - g\|_2 + \|U^n (f - g)\|_2 \|g\|_2 \\
&\leq \|f\|_2 \varepsilon + \varepsilon \|g\|_2 = \varepsilon(\|f\|_2 + \|g\|_2).
\end{aligned}$$

Since $\|f\|_2 = \|f - g\|_2 + \|g\|_2$, we have $\|f\|_2 + \|g\|_2 \leq 2\|g\|_2 + \varepsilon$.

$$\begin{aligned}
|(f, 1)(1, f) - (g, 1)(1, g)| &= \left| |(f, 1)|^2 - |(g, 1)|^2 \right| \\
&= |(|f, 1)| + |(g, 1)|)(|(f, 1) - (g, 1)|)| \\
&\leq (\|f\|_2 + \|g\|_2) \underbrace{((f, 1) - (g, 1))}_{=(f-g,1) \leq \|f-g\|_2} \\
&\leq (\|f\|_2 + \|g\|_2) \|f - g\|_2 \leq (2\|g\|_2 + \varepsilon)\varepsilon.
\end{aligned}$$

Hence,

$$\begin{aligned}
|(U^n g, g) - (g, 1)(1, g)| &= |(U^n g, g) - (U^n f, f) + (U^n f, f) \\
&- (f, 1)(1, f) + (f, 1)(1, f) - (g, 1)(1, g)| \\
&\leq (2\|g\|_2 + \varepsilon) + \varepsilon + (2\|g\|_2 + \varepsilon)\varepsilon.
\end{aligned}$$

This implies (ii).

"(ii)$\Rightarrow$ (iii)": We use (ii) for $f + g$:

$$\begin{aligned}
(U^n (f + g), f + g) &\to (f + g, 1)(1, f + g) = (f, 1)(1, f) + (g, 1)(1, g) \\
&+ (f, 1)(1, g) + (g, 1)(1, f).
\end{aligned}$$

Since by (ii) $(U^n f, f) \to (f, 1)(1, f)$ and $(U^n g, g) \to (g, 1)(1, g)$, we have

$$(U^n f, g) + (U^n g, f) \to (f, 1)(1, g) + (g, 1)(1, f). \tag{6}$$

For if instead of f we obtain

$$i(U^n f, g) - i(U^n g, f) \to i(f, 1)(1, g) - i(g, 1)(1, f). \tag{7}$$

Dividing (7) by i and adding it to (6) gives $2(U^n f, g) \to 2(f, 1)(1, g)$.

"(iii)⇒ (i)": Let $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$. Then

$$\begin{aligned}(U_T^n f, g)_{L^2}^2 &= \int_X f(T^n(x))\overline{g(x)}d\mu(x) = \int_X \mathbb{1}_{T^{-n}(A)}(x)\mathbb{1}_B(x)d\mu(x) \\ &= \int_X \mathbb{1}_{T^{-n}(A)\cap B}d\mu = \mu(T^{-n}(A)\cap B)\end{aligned}$$

and similarly

$$(f,1)_{L^2}(1,g)_{L^2} = \mu(A)\mu(B).$$

Hence, (iii) implies

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

which finishes the proof. □

4.3 Corollary: *Let T be a mixing transformation T on a probability space (X,μ). Then U_T has no eigenvalues on the unit circle except for* 1.

Proof: Let $U = U_T$. Suppose that $\lambda \neq 1$, $|\lambda| = 1$, is an eigenvalue of U_T, i.e., there exists a nonconstant $f \in L^2(X,\mathbb{C},\mu)$ with

$$(Uf)(x) = \lambda f(x) \text{ for almost all } x \in X.$$

W.l.o.g. assume that $\|f\|_{L^2} = 1$. We have to show that T is <u>not</u> mixing. For all $n \in \mathbb{N}$ we have

$$|(U^n f, f)_{L^2}| = |(\lambda^n f, f)_{L^2}| = |\lambda|^n\,|(f,f)| = 1.$$

On the other hand $|(1,f)|^2 < 1$, since by Cauchy-Schwarz

$$|(f,1)_{L^2}|^2 < (f,f)_{L^2}(1,1)_{L^2} = \|f\|_2^2\|1\|_2^2 = 1.$$

This is a contradiction to Theorem 4.2 (ii). □

Consider again the irrational translation $T(x) = x+\theta$, θ irrational. Then

$$U_T(\underbrace{e^{2\pi i(\cdot)}}_{=f})(x) = e^{2\pi i(x+\theta)} = e^{2\pi i\theta}\underbrace{e^{2\pi i x}}_{=f(x)}.$$

So $e^{2\pi i\theta} \neq 1$ is an eigenvalue of U_T, which lies on the unit circle. Hence, T is not mixing.

4.4 Definition: *Let T be measure-preserving on a probability space (X,μ). For a real-valued function $f \in L^2(X,\mu)$ we call*

$$r_n(f) := \left|\int_X f(T^n(x))f(x)d\mu(x) - \left(\int_X f(x)d\mu(x)\right)^2\right|$$

the ***n^{th}* correlation coefficient** *of f.*

Clear: For a mixing transformation one has $r_n(f) \to 0$ for $n \to \infty$.

4.5 Example: Consider $T(x) = x + \sqrt{3} - 1 \pmod 1$ on $[0,1)$. The Lebesgue measure is invariant. For $f(x) = x$ we have $\int_X f(x)dx = \frac{1}{2}$, and

$$\int_X f(T^n(x))f(x)dx \approx \frac{1}{s}\sum_{i=0}^{s} f(T^n(x_i))x_i, \quad x_i = \frac{i}{s}.$$

See the `MAPLE` program `Correlation_Irr`. ◊

4.2 Recurrence and First Return Time

4.6 Theorem (Poincaré's Recurrence Theorem): *Let T be measure-preserving on a probability space (X, μ). Consider $E \subset X$ with $\mu(E) > 0$. Then almost all $x \in E$ are recurrent, i.e.,*

$$T^{n_k}(x) \in E \text{ for a sequence } n_k \xrightarrow{k\to\infty} \infty.$$

Special situation: X metric space, $\mu(A) > 0$ for each nonempty open set A. Again, T is μ-preserving. Choose $x \in X$. Poincaré implies: For every $\varepsilon > 0$ $T^{n_k}(B_\varepsilon(x)) \cap B_\varepsilon(x) \neq \emptyset$ for a sequence $n_k \to \infty$.

Proof (of Poincaré's Recurrence Theorem): For every $n \in \mathbb{N}_0$ let

$$E_n := \bigcup_{k=n}^{\infty} T^{-k}(E).$$

Then $\bigcap_{n=0}^{\infty} E_n$ is the set of all points $x \in X$ such that $T^n(x) \in E$ infinitely often. Put $F := E \cap (\bigcap_{n=0}^{\infty} E_n)$. We have to show that $\mu(F) = \mu(E)$. If $x \in F$ there are $0 < n_1 < n_2 < \cdots$ with $T^{n_i}(x) \in E$. Fix n_i. Then for $j > i$

$$T^{n_j}(x) = T^{n_j - n_i}(T^{n_i}(x)) \in E.$$

Hence, $T^{n_i}(x) \in F$. So we know that x returns to F infinitely often. Note that $T^{-1}(E_n) = E_{n+1}$. Hence,

$$\mu(E_n) = \mu(T^{-1}(E_n)) = \mu(E_{n+1}).$$

Furthermore, $E_0 \supset E_1 \supset E_2 \supset \cdots$. Thus,

$$\mu\left(\bigcap_{n=0}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n) = \mu(E_0).$$

Similarly,

$$E_0 \cap E \supset E_1 \cap E \supset E_2 \cap E \supset \cdots,$$

which implies

$$\mu(F) = \mu\left(E \cap \bigcap_{n=0}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n \cap E) = \mu(\underbrace{E_0 \cap E}_{=E}) = \mu(E),$$

since $E \subset E_0$. □

4.7 Definition: *Let $T : X \circlearrowleft$ be measure-preserving on a probability space (X, μ). Suppose $\mu(E) > 0$, fix $x \in E$ and define the* **first return time** *in E by*

$$R_E(x) := \min\{n \in \mathbb{N} \mid T^n(x) \in E\}.$$

Poincaré guarantees that $R_E(x) < \infty$ for almost all $x \in E$. Define the **first return time transformation** *by*

$$T_E(x) := T^{R_E(x)}(x), \quad T_E : E \to E.$$

Question: Are the maps $x \mapsto R_E(x)$ and $x \mapsto T_E(x)$ measurable and can we describe their properties and relate them to properties of T?

4.8 Remarks:

- The map $R_E : E \to \mathbb{R}$ is measurable: Consider the set

  $$R_E^{-1}((-\infty, \alpha]) = \{x \in E \mid R_E(x) \le \alpha\}, \quad \alpha \in \mathbb{R}.$$

  For $\alpha < 1$ this is the empty set. For $\alpha \ge 1$ let $k = \lceil \alpha \rceil$ (the smallest integer greater or equal than α). Then

  $$\{x \in E \mid R_E(x) \le \alpha\} = E \cap \left(T^{-1}(E) \cup \cdots \cup T^{-k}(E)\right).$$

  Since T is measurable, the sets $T^{-k}(E)$ are measurable, and hence also $R_E^{-1}((-\infty, \alpha])$ is measurable. This proves the assertion.

- The map $T_E : E \to E$, $x \mapsto T^{R_E(x)}(x)$, is also measurable: Let

  $$E_k := \{x \in E \mid R_E(x) = k\}, \quad k \in \mathbb{N}.$$

  Let $C \subset E$ be measurable. Then

  $$T_E^{-1}(C) = \bigcup_{k=1}^{\infty} \left(E_k \cap T^{-k}(C)\right).$$

  Hence, $T^{-1}(C)$ is measurable, since $T^{-1}(C)$ is measurable and E_k is measurable, since R_E is measurable.

Recall that for E with $\mu(E) > 0$ the conditional measure on E is $\mu_E(C) = \frac{\mu(C)}{\mu(E)}$, $C \subset E$.

4.9 Theorem:

(i) *If T is measure-preserving on a probability space (X, μ) and if $\mu(E) > 0$, then T_E preserves μ_E.*

(ii) *If T is ergodic, then also T_E is ergodic.*

Proof:

(i) Only for invertible T with T^{-1} measurable (for general T the proof is more technical):

$$\mu(T^{-1}(A)) = \mu(A) \text{ for all } A \quad \Leftrightarrow \quad \mu(A) = \mu(T(A)) \text{ for all } A,$$

since $T^{-1}(T(A)) = T(T^{-1}(A)) = A$ for all $A \subset X$. Define for every $n \in \mathbb{N}$

$$A_n := \{x \in A \mid R_E(x) = n\}.$$

Then A_n is measurable and

$$A = \bigcup_{n=1}^{\infty} A_n \text{ (disjoint union).}$$

Note that $T_E(A_n) = T^n(A_n)$. We find

$$\begin{aligned} \mu_E(T_E(A)) &= \mu_E\left(T_E\left(\bigcup_{n=1}^{\infty} A_n\right)\right) = \sum_{n=1}^{\infty} \mu_E(T^n(A_n)) \\ &= \sum_{n=1}^{\infty} \frac{\mu(T^n(A_n))}{\mu(E)} = \sum_{n=1}^{\infty} \frac{\mu(A_n)}{\mu(E)} = \frac{\mu(A)}{\mu(E)} = \mu_E(A). \end{aligned}$$

(ii) Ergodicity: Let $B \subset E$ be invariant for T_E and suppose $\mu_E(B) > 0$. By invariance

$$B = T_E^{-1}(B) = T_E^{-2}(B) = \dots$$

Hence,

$$B = \left(\bigcup_{n=0}^{\infty} T^{-n}(B)\right) \cap E.$$

Since μ is ergodic, we have $\mu(\bigcup_{n=0}^{\infty} T^{-n}(B)) = 1$. Hence,

$$\bigcup_{n=0}^{\infty} T^{-n}(B) = X.$$

It follows that $B = E$, so $\mu_E(B) = \frac{\mu(B)}{\mu(E)} = 1$.

□

4.10 Theorem (Kać Lemma): *If T is an ergodic measure-preserving map on a probability space (X,μ), and if $\mu(E)>0$, then*

$$\int_E R_E d\mu = 1,$$

i.e., $\int_E R_E d\mu_E = \frac{1}{\mu(E)}$.

Proof: We give two proofs:

(i) For invertible T: For $n \geq 1$ let

$$E_n := \{x \in E \mid R_E(x) = n\}.$$

Then $E_n \cap E_m = \emptyset$ if $n \neq m$, and

$$E = \bigcup_{n=1}^{\infty} E_n$$

by Poincaré. Since T is ergodic, for all $A, B \subset X$ with $\mu(A), \mu(B) > 0$ there is $k \in \mathbb{N}$ with $\mu(T^{-n}(A) \cap B) > 0$. Hence, there exists no set $A \subset X$ of positive measure with $\mu(T^{-k}(A) \cap E) = 0$ for all $k \in \mathbb{N}$, which implies that for almost all $x \in X$ we find $k \in \mathbb{N}$ with $T^{-k}(x) \in E$. This implies

$$X = \bigcup_{n=1}^{\infty} \bigcup_{k=0}^{n-1} T^k(E_n). \tag{8}$$

Observe that the sets $E_n, T(E_n), T^2(E_n), \ldots, T^{n-1}(E_n)$ are disjoint. By ergodicity and injectivity of T

$$\mu(T^k(E_n)) = \mu(E_n) \text{ for all } k.$$

Hence, (8) is a disjoint union. We compute

$$\int_E R_E d\mu = \sum_{n=1}^{\infty} \int_{E_n} R_E d\mu = \sum_{n=1}^{\infty} n\mu(E_n) \overset{(8)}{=} \mu(X) = 1.$$

(ii) Proof for not necessarily invertible T: Take $x \in E$ and consider

$$x, T_E(x), \ldots, T_E^l(x), \ldots, T_E^L(x), \quad L \in \mathbb{N}.$$

Let

$$N = \sum_{l=0}^{L-1} R_E(T_E^l(x)).$$

Then N is the time duration for the iterates $T^n(x)$, $n = 1, \ldots, N$, to come back to E exactly L times, i.e.,

$$\sum_{n=1}^{N} \mathbb{1}_E(T^n(x)) = L.$$

Now apply Birkhoff's Ergodic Theorem to the map T_E and $f = R_E$. Then

$$\int_E R_E d\mu_E = \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} R_E(T_E^l(x)) = \lim_{N\to\infty} \frac{N}{\sum_{n=1}^N \mathbb{1}_E(T^n(x))}$$
$$= \left(\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_E(T^n(x)) \right)^{-1}.$$

Hence, applying Birkhoff's Ergodic Theorem for T and $f = \mathbb{1}_E$ gives

$$\int_E R_E d\mu_E = \frac{1}{\mu(E)}.$$

□

4.3 Mixing Markov Shift Transformations

Let $\mathcal{A} = \{1, \ldots, k\}$ (symbols, alphabet) and $X = \prod_1^\infty \mathcal{A}$. The shift $T = \theta : X \circlearrowleft$ is defined by $(x_1, x_2, x_3, \ldots) \mapsto (x_2, x_3, \ldots)$. The $(\frac{1}{2}, \frac{1}{2})$-Bernoulli Shift on $\{0, 1\}$:

$$\mu\left([a_1, \ldots, a_n]_{t,\ldots,t+k-1}\right) = \left(\tfrac{1}{2}\right)^n.$$

Markov measures: Are given by a stochastic $k \times k$-matrix $P = (p_{ij})$ ($\sum_j p_{ij} = 1$, $p_{ij} \geq 0$). There exists an eigenvector $\pi P = \pi$, $\pi \geq 0$, $\sum_i \pi_i = 1$. All π_i are positive if P is irreducible, i.e., for all i and j there exists $m \in \mathbb{N}$ with $(P^m)_{ij} > 0$. Markov measure:

$$\mu([a_1, \ldots, a_n]_{t,\ldots,t+n-1}) = \pi_{a_1} p_{a_1 a_2} \cdots p_{a_{n-1} a_n}.$$

This is shift-invariant. μ is ergodic iff P is irreducible (then π is unique) (Theorem 3.17).

Question: Can we characterize the mixing property of μ via the matrix P? (i.e., when is $\mu(\theta^{-n}(A) \cap B) \to \mu(A)\mu(B)$ for $n \to \infty$ satisfied?)

If P is irreducible, then (cp. proof of Theorem 3.17)

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

exists and each row of Q equals π, and $\pi_j > 0$ for all j.

4.11 Definition: *A stochastic matrix A is called* **eventually positive** *if for all n large enough*

$$(A^n)_{ij} > 0 \quad \textit{for all } i, j = 1, \ldots, k.$$

4.12 Proposition: *If A is an eventually positive stochastic matrix, then the eigenvalue $\lambda = 1$ is simple (the algebraic multiplicity equals 1) and all eigenvalues $\mu \neq 1$ satisfy $|\mu| < 1$.*

Proof: See Robinson [5] or Gantmacher [6]. □

4.13 Theorem: *Let P be a stochastic $k \times k$-matrix with eigenvector $\pi = (\pi_i)$ satisfying $\pi P = \pi$, $\pi \geq 0$, $\sum \pi_i = 1$. For $\mathcal{A} = \{1, \dots, k\}$ let T be the associated Markov shift transformation on $X = \prod_1^\infty \mathcal{A}$ with shift invariant Markov measure μ. Suppose P is irreducible. Then the following are equivalent:*

(i) *T is mixing.*

(ii) *$(P^n)_{ij}$ converges to π_j for $n \to \infty$ for all $i, j = 1, \dots, k$.*

(iii) *P is eventually positive.*

Proof: Put $Q := \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n$. Since P is irreducible, μ is ergodic, and Q exists, each row of Q equals π, $\pi_j > 0$ for all j.
"(i) ⇒ (ii)": Suppose T is mixing, i.e.,

$$\mu(T^{-n}(A) \cap B) \to \mu(A)\mu(B) \text{ for all } A, B.$$

Let $A := [j]_1$ and $B := [i]_1$. Claim:

$$\mu(T^{-n}(A) \cap B) = \pi_j (P^n)_{ji} \to \mu(A)\mu(B) = \pi_j \pi_i.$$

This shows that $(P^n)_{ji} \to \pi_i$ for every i and j, i.e., $P^n \to Q$ for $n \to \infty$.

"(ii) ⇒ (iii)": By (ii) $P^n \to Q$ and all entries of Q are positive. Hence, for n large enough, $(P^n)_{ij} > 0$ for all i and j. Thus, P is eventually positive.

"(iii) ⇒ (i)": It suffices to show

$$\mu(T^n(A) \cap B) \to \mu(A)\mu(B)$$

for cylinder sets A, B. Let $A = [i_1, \dots, i_r]_a^{a+r-1}$ and $B = [j_1, \dots, j_s]_b^{b+s-1}$. Let J be the Jordan canonical form of P. Since $\lambda = 1$ is simple and all eigenvalues $\mu \neq 1$ satisfy $|\mu| < 1$, we get

$$J = \begin{pmatrix} 1 & & & \\ & M_{\mu_1} & & \\ & & \ddots & \\ & & & M_{\mu_l} \end{pmatrix}, \quad M_{\mu_i} = \begin{pmatrix} \mu_i & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & \mu_i \end{pmatrix}$$

with $M_{\mu_i}^n \to 0$ for $n \to \infty$. This implies $J^n \to \operatorname{diag}(1, 0, \dots, 0)$. Hence, P^n converges for $n \to \infty$. Since $\frac{1}{n} \sum_{k=0}^{n-1} P^k$ converges to Q, also P^n must converge to Q.

$$\mu(T^{-n}(A)\cap B) = \underbrace{\pi_{j_1}P_{j_1j_2}\cdots P_{j_{s-1}j_s}}_{=\mu(B)} \underbrace{(P^n)_{j_si_1}}_{\to Q_{j_si_1}=\pi_{i_1}} \underbrace{P_{i_1i_2}\cdots P_{i_{r-1}i_r}}_{=\frac{1}{\pi_{i_1}}\mu(A)} \stackrel{n\to\infty}{\longrightarrow} \mu(B)\mu(A).$$

□

4.14 Theorem: *In Theorem 4.13 the speed of convergence of P^n to Q is exponential, i.e.,*

$$\|P^n - Q\| \le \alpha\beta^n$$

with constants $\alpha > 0$, $\beta \in (0,1)$, for some (and then for all) norms in $\mathbb{R}^{n\times n}$.

Proof: All norms on the vector space $\mathbb{C}^{n\times n}$ are equivalent: For any two norms $\|\cdot\|$ and $\|\cdot\|'$ there are constants $c_1, c_2 > 0$ with

$$c_1\|A\| \le \|A\|' \le c_2\|A\|.$$

Let S be invertible with $S^{-1}PS = J$, the Jordan canonical form. Then

$$\|A\|' := \|S^{-1}AS\|, \quad A \in \mathbb{C}^{n\times n},$$

defines a norm, and

$$c_1\|A\| \le \|A\|' \le c_2\|A\|$$

for constants $c_1, c_2 > 0$. Hence,

$$c_1\|P^n - Q\| \le \underbrace{\|S^{-1}(P^n - Q)S\|}_{=\|P^n-Q\|'} \le c_2\|P^n - Q\|.$$

Recall: Since P is eventually positive, 1 is an algebraically simple eigenvalue and all other eigenvalues satisfy $|\mu| < 1$. Thus,

$$J = \begin{pmatrix} 1 & & & \\ & M_{\mu_1} & & \\ & & \ddots & \\ & & & M_{\mu_k} \end{pmatrix}$$

Claim:

$$S^{-1}(P^n - Q)S = S^{-1}P^nS - S^{-1}QS = J^n - \operatorname{diag}(1,0,\ldots,0).$$

Observe that, if all Jordan blocks are one-dimensional, exponential convergence of J^n to $S^{-1}QS$ is clear. We only have to deal with the problem that the Jordan blocks may be higher dimensional. Now use the norm

$$\|A\|_\infty := \max_{i,j}|A_{ij}|.$$

For a Jordan block $M_\mu = \mu I + N$ with $|\mu| < 1$,

$$N = \begin{pmatrix} 0 & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & 0 \end{pmatrix}$$

we have $N^k = 0$. Hence,

$$M_\mu^n = \sum_{j=0}^{k} \binom{k}{j} \mu^{n-j} N^j, \quad n \geq k.$$

Let $\eta := \max\{\|I\|_\infty, \|N\|_\infty, \ldots, \|N^{k-1}\|_\infty\}$. Then for $n \geq k$

$$\|M_\mu^n\|_\infty \leq \eta \sum_{j=0}^{k-1} \binom{n}{j} |\mu|^{n-j} \leq \eta k n^{k-1} |\mu|^{n-k+1}.$$

Observe that η and k are fixed and

$$n^{k-1}|\mu|^n = e^{(k-1)\ln(n)} e^{n \ln|\mu|} = e^{n[(k-1)\frac{\ln(n)}{n} + \ln|\mu|]}.$$

Since $\frac{\ln(n)}{n} \to 0$ for $n \to \infty$ and $\ln|\mu| < 0$, this is bounded above for n large enough by

$$e^{n\ln(\beta)} = \beta^n \ \text{ for } \ \beta \text{ with } |\mu| < \beta < 1.$$

Together

$$\|M_\mu^n\|_\infty \leq \alpha\beta^n \ \text{ for a constant } \alpha > 0 \text{ and } \beta \in (0, 1).$$

This shows that

$$\|J^n - \operatorname{diag}(1, \ldots, 0)\|_\infty \leq \alpha\beta^n,$$

hence the same for $\|P^n - Q\|$ holds. □

Note the difference between irreducibility and eventual positivity:

$$P := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ \to \ P^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Hence, P is irreducible, but it is not eventually positive, since

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ \text{ for all } n \in \mathbb{N}.$$

5 Entropy

5.1 Definition and Elementary Properties

Simple situation (without map): Consider an experiment with uncertainty described by a set $\mathcal{A} = \{a_1, \ldots, a_k\}$. Let p_i be the probability of the outcome a_i. Then

$$p_1 + \cdots + p_k = 1.$$

If p_1 is close to 1, we would mostly obtain a_1. If we measure the surprise or information that we get from some outcome, it would be close to 0, if the outcome is a_1 with probability close to 1. How to measure the information? For the outcome a_i the magnitude of information is $\frac{1}{p_i}$. Instead we take $\log \frac{1}{p_i}$ ($p_1 \approx 1 \to \log \frac{1}{p_1} \approx 0$). The expected information from an experiment is

$$\sum_{i=1}^{k} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{k} p_i \log p_i.$$

This is called the *entropy of* $\mathcal{A}$ (the expected information from an experiment). Usually log is taken as logarithm with base 2.

In general: An experiment corresponds to a measurable partition

$$\mathcal{P} = \{E_1, \ldots, E_n\}$$

of a probability space $(X, \mathcal{A}, \mu)$. The entropy of this partition is defined as

$$H(\mathcal{P}) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log p_i,$$

with $p_i = \mu(E_i)$, where $p_i \log p_i := 0$ if $p_i = 0$. Consider a map $T : X \circlearrowleft$ which is μ-preserving. Idea: How much information do we gain by applying T?

First a simple estimate for the entropy of a partition:

5.1 Lemma: *If a partition $\mathcal{P}$ consists of k subsets, then $H(\mathcal{P}) \leq \log k$.*

Proof: Recall for $a > 0$

$$\frac{\ln a}{\ln 2} = \log_2 a.$$

We have to show

$$-\sum_{i=1}^{k} p_i \ln p_i \leq \ln k$$

for all $p_1, \ldots, p_k \in (0, 1)$, $\sum_{i=1}^{k} p_i = 1$. We show that

$$\max \left(-\sum_{i=1}^{k} p_i \ln p_i \right)$$

over $p_1, \dots, p_k > 0$ with $\sum_i p_i = 1$ equals $\ln k$. A necessary condition for a maximum is that there is $\lambda \in \mathbb{R}$ such that

$$f(p_1, \dots, p_k) = -\sum_{i=1}^{k} p_i \ln p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

has Jacobian equal to zero, and $\sum_{i=1}^{k} p_i = 1$.

$$\frac{\partial f}{\partial p_j} = -\ln p_j - p_j \frac{1}{p_j} + \lambda = \lambda - 1 - \ln p_j = 0$$

for $j = 1, \dots, k$. Together with $p_1 + \cdots + p_k = 1$ this shows that the maximum can only be attained if $p_1 = \cdots = p_k = \frac{1}{k}$. Then

$$-\sum_{i=1}^{k} p_i \ln p_i = -\sum_{i=1}^{k} \frac{1}{k} \ln \frac{1}{k} = -\ln \frac{1}{k} = \ln k,$$

as claimed. (Since there is a maximum and it cannot be attained on the boundary, the necessary condition is also sufficient.) □

Given two partitions $\mathcal{P}$ and $\mathcal{Q}$, the *join* of $\mathcal{P}$ and $\mathcal{Q}$ is the partition $\mathcal{P} \vee \mathcal{Q}$ consisting of all sets of the form $B \cap C$ with $B \in \mathcal{P}$ and $C \in \mathcal{Q}$. Analogously, the join $\bigvee_{i=1}^{n} \mathcal{P}_i$ of finitely many measurable partitions $\mathcal{P}_1, \dots, \mathcal{P}_n$ is defined. Fix a partition $\mathcal{P}$ and consider $T : X \circlearrowleft$. Let

$$T^{-j}\mathcal{P} = \{T^{-j}E_1, \dots, T^{-j}E_k\}, \ \ \mathcal{P} = \{E_1, \dots, E_k\}.$$

This again is a partition. Let

$$\mathcal{P}_n := \mathcal{P} \vee T^{-1}\mathcal{P} \vee \ldots \vee T^{-(n-1)}\mathcal{P}.$$

Define the entropy of T with respect to $\mathcal{P}$ as

$$h(T, \mathcal{P}) := \lim_{n \to \infty} \frac{1}{n} H(\mathcal{P}_n). \tag{9}$$

Finally, the entropy of T is

$$h(T) := \sup_{\mathcal{P}} h(T, \mathcal{P}),$$

where the supremum is taken over all finite measurable partitions of X. We have to show that the limit in (9) exists. To this end, we use the following two lemmas.

5.2 Lemma: *Let $\mathcal{P} = \{C_1, \dots, C_r\}$ and $\mathcal{Q} = \{D_1, \dots, D_s\}$ be measurable partitions of X. Then*

$$H(\mathcal{P} \vee \mathcal{Q}) \le H(\mathcal{P}) + H(\mathcal{Q}).$$

Proof: We have

$$
\begin{aligned}
H(\mathcal{P} \vee \mathcal{Q}) &= -\sum_{i,j} \mu(C_i \cap D_j) \log \mu(C_i \cap D_j) \\
&= -\sum_{i,j} \mu(C_i \cap D_j) \log \left[\mu(C_i) \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \right] \\
&= -\sum_{i,j} \mu(C_i \cap D_j) \log \mu(C_i) \\
&- \sum_{i,j} \mu(C_i) \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \\
&= \underbrace{-\sum_{i} \mu(C_i) \log \mu(C_i)}_{=H(\mathcal{P})} \\
&- \sum_{j} \left[\sum_{i} \mu(C_i) \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \right].
\end{aligned}
$$

Consider the map $\varphi(x) = x \log x$. φ is convex. *Jensen's Inequality* (Elstrodt [7]): Let $f : X \to \mathbb{R}$ be integrable on a probability space (X, μ) and $\varphi : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\varphi\left(\int_X f d\mu\right) \leq \int_X \varphi \circ f d\mu.$$

<u>Claim:</u>

$$-\sum_{i} \mu(C_i) \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \leq -\mu(D_j) \log \mu(D_j). \tag{10}$$

This implies $H(\mathcal{P} \vee \mathcal{Q}) \leq H(\mathcal{P}) + H(\mathcal{Q})$. Define

$$f(x) := \sum_{i} \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \mathbb{1}_{C_i}(x), \quad f : X \to \mathbb{R}.$$

Then $\int_X f d\mu = \mu(D_j)$ and therefore

$$\varphi\left(\int_X f d\mu\right) = \int_X f d\mu \log \int_X f d\mu = \mu(D_j) \log \mu(D_j).$$

On the other hand:

$$
\begin{aligned}
\int_X \varphi \circ f d\mu &= \int_X f(x) \log f(x) d\mu(x) \\
&= \int_X \sum_{i} \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \mathbb{1}_{C_i}(x) \log \left[\sum_{k} \frac{\mu(C_k \cap D_j)}{\mu(C_k)} \mathbb{1}_{C_k}(x) \right]
\end{aligned}
$$

$$= \sum_i \int_{C_i} \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \left[\sum_k \frac{\mu(C_k \cap D_j)}{\mu(C_k)} \mathbb{1}_{C_k}(x) \right] d\mu(x)$$
$$= \sum_i \int_{C_i} \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \left[\frac{\mu(C_i \cap D_j)}{\mu(C_i)} \right] d\mu$$
$$= \sum_i \mu(C_i \cap D_j) \log \frac{\mu(C_i \cap D_j)}{\mu(C_i)}.$$

This proves (10). □

5.3 Remark: The equality $H(\mathcal{P} \vee \mathcal{Q}) = H(\mathcal{P}) + H(\mathcal{Q})$ holds if for all i and j we have $\mu(C_i \cap D_j) = \mu(C_i)\mu(D_j)$. Then it follows

$$-\sum_i \mu(C_i) \frac{\mu(C_i \cap D_j)}{\mu(C_i)} \log \frac{\mu(C_i \cap D_j)}{\mu(C_i)} = -\underbrace{\sum_i \mu(C_i)}_{=1} \mu(D_j) \log \mu(D_j).$$

Two partitions with this property are called *independent*. Actually $H(\mathcal{P} \vee \mathcal{Q}) = H(\mathcal{P}) + H(\mathcal{Q})$ holds if and only if $\mathcal{P}$ and $\mathcal{Q}$ are independent.

5.4 Lemma: *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real numbers with $a_n \geq 0$ and $a_{n+m} \leq a_n + a_m$ for all $n, m \in \mathbb{N}$. Then*

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\in\mathbb{N}} \frac{a_n}{n} =: a.$$

Proof: Fix $\varepsilon > 0$. There is $N \in \mathbb{N}$ with $\frac{a_N}{N} < a + \varepsilon$. For every $n \in \mathbb{N}$ we can write $n = kN + r$ with $k, r \in \mathbb{N}_0$ and $0 \leq r < N$. Then

$$\frac{a_n}{n} \leq \frac{1}{n} [k a_N + a_r] \leq \frac{k}{kN} a_N + \frac{a_r}{n} = \frac{a_N}{N} + \frac{a_r}{n}.$$

Since $\frac{a_r}{n} \to 0$ for $n \to \infty$, we find $n_0 \in \mathbb{N}$ such that $\frac{a_n}{n} < a + 2\varepsilon$ for all $n \geq n_0$. This implies the assertion. □

Now we can conclude that the limit in (9) exists: Consider the sequence $(H(\mathcal{P}_n))$. Note that, since T is μ-preserving, for all $j \in \mathbb{N}$ and every partition $\mathcal{P} = \{E_1, \ldots, E_n\}$ we have

$$\begin{aligned} H(T^{-j}\mathcal{P}_n) &= -\sum_i \mu(T^{-j}(E_i)) \log \mu(T^{-j}(E_i)) \\ &= -\sum_i \mu(E_i) \log \mu(E_i) = H(\mathcal{P}). \end{aligned} \tag{11}$$

Hence, for all $n, m \in \mathbb{N}$ we obtain

$$\begin{aligned} H(\mathcal{P}_{n+m}) &= H(\underbrace{\mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-(n-1)}\mathcal{P}}_{=\mathcal{P}_n} \vee \underbrace{T^{-n}\mathcal{P} \vee \cdots \vee T^{-(n+m-1)}\mathcal{P}}_{=T^{-n}\mathcal{P}_m}) \\ &\overset{\text{Lem. } 5.2}{\leq} H(\mathcal{P}_n) + H(T^{-n}\mathcal{P}_m) \overset{(11)}{=} H(\mathcal{P}_n) + H(\mathcal{P}_m). \end{aligned}$$

By Lemma 5.4 it follows that the limit (9) exists.

5.5 Lemma: *Let $\mathcal{P}$ be a refinement of the partition $\mathcal{Q}$, i.e., the elements of $\mathcal{Q}$ are unions of elements from $\mathcal{P}$. Then*

$$H(\mathcal{P}) \geq H(\mathcal{Q}).$$

Proof: Let $\mathcal{Q} = \{D_1, \ldots, D_r\}$. We have $H(\mathcal{Q}) = -\sum_j \mu(D_j) \log \mu(D_j)$ and D_j is the disjoint union of sets $C_{j_i} \in \mathcal{P}$. Hence, $\mu(D_j) = \sum_i \mu(C_{j_i})$, which implies

$$\begin{aligned} H(\mathcal{Q}) &= -\sum_j \sum_i \mu(C_{j_i}) \log \underbrace{\sum_i \mu(C_{j_i})}_{\geq \mu(C_{j_k}) \text{ for all } k} \\ &\leq -\sum_j \sum_i \mu(C_{j_i}) \log \mu(C_{j_i}) = H(\mathcal{P}). \end{aligned}$$

□

5.2 Conditional Entropy

If we want to compute entropies, we will have to discuss several questions:

- It seems extremely difficult to compute the supremum $h(T, \mathcal{P})$ over all partitions.
  Question: When does there exist a partition $\mathcal{P}$ with $h(T, \mathcal{P}) = h(T)$?
- For doing computation it will be essential to understand precisely what happens when we refine a partition. Up to now, we only know

$$H(\mathcal{P}) \geq H(\mathcal{Q}) \text{ if } \mathcal{P} \text{ refines } \mathcal{Q}.$$

Let us start with the following observation on measurable partitions of a measure space $(X, \mathcal{A}, \mu)$:

$$X = D_1 \dot{\cup} \ldots \dot{\cup} D_k.$$

It does not matter, if we change D_i only in a set of μ-measure 0, hence we consider measurable partitions mod 0: This means: $\mathcal{P}$ and $\mathcal{P}'$ are identified if there is a set A with $\mu(A) = 0$, such that the restrictions of $\mathcal{P}$ and $\mathcal{P}'$ to $X \setminus A$ coincide. Furthermore, partition sets of measure 0 do not play a role, so usually, we will assume that all partition sets have positive measure. Finally, sometimes we will also allow countable partitions instead of finite ones.

Let $\mathcal{P} = \{C_i \mid i \in I\}$ be a measurable partition and recall

$$H_\mu(\mathcal{P}) = -\sum_i \mu(C_i) \log \mu(C_i).$$

For $x \in X$ let $C_{\mathcal{P}}(x)$ be the unique element of $\mathcal{P}$ containing x. The function

$$I_{\mathcal{P}}(x) = -\log \mu(C_{\mathcal{P}}(x))$$

is called the *information function* of $\mathcal{P}$ (defined outside of the set of measure 0 with $\mu(C_{\mathcal{P}}(x)) = 0$). Then

$$H_\mu(\mathcal{P}) = \int_X I_{\mathcal{P}} d\mu = \sum_i \mu(C_i)(-\log \mu(C_i)),$$

since on every element of $\mathcal{P}$ $I_{\mathcal{P}}(x)$ is constant.

Next write for the conditional probability

$$\mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)}.$$

Interpretation: This is the probability of A provided B occurs. A and B are called independent, if $\mu(A \cap B) = \mu(A)\mu(B)$, hence, in this case, $\mu(A|B) = \mu(A)$ (occurrence of B does influence occurrence of A).

Next we introduce the conditional entropy.

5.6 Definition: *Let $\mathcal{P} = \{C_\alpha \mid \alpha \in I\}$ and $\mathcal{Q} = \{D_\beta \mid \beta \in J\}$ be two measurable partitions of (X, μ). The* **conditional entropy** *of $\mathcal{P}$ with respect to $\mathcal{Q}$ is*

$$H(\mathcal{P}|\mathcal{Q}) := -\sum_{\beta \in J} \mu(D_\beta) \sum_{\alpha \in I} \mu(C_\alpha|D_\beta) \log \mu(C_\alpha|D_\beta).$$

The intuitive meaning of the conditional entropy $H(\mathcal{P}|\mathcal{Q})$ is that it is the expected amount of information gained by the experiment $\mathcal{P}$ given the results of the experiment $\mathcal{Q}$.

5.7 Remark: If $\mathcal{Q} = \{X\}$ is the trivial partition, then $H(\mathcal{P}|\mathcal{Q}) = H(\mathcal{P})$. Using an information function, one can write the conditional entropy as

$$H(\mathcal{P}|\mathcal{Q}) = \int_X I_{\mathcal{P},\mathcal{Q}} d\mu,$$

where $I_{\mathcal{P},\mathcal{Q}}$ is the conditional information function

$$I_{\mathcal{P},\mathcal{Q}}(x) = -\log \mu(C_{\mathcal{P}}(x)|D_{\mathcal{Q}}(x)).$$

5.8 Remark: Denote by $\mathcal{P}_{D_\beta}$ the partition of D_β into the sets $D_\beta \cap C_\alpha$, $\alpha \in I$, such that $\mu(D_\beta \cap C_\alpha) > 0$. Then

$$\begin{aligned} H(\mathcal{P}|\mathcal{Q}) &= \sum_{\beta \in J} \mu(D_\beta) H_{\mu_{D_\beta}}(\mathcal{P}_{D_\beta}) \\ &= \sum_{\beta \in J} \mu(D_\beta) \left(-\sum_\alpha \mu_\beta(C_\alpha) \log \mu_\beta(C_\alpha) \right) \\ &= -\sum_{\beta \in J} \mu(D_\beta) \sum_\alpha \frac{\mu(D_\beta \cap C_\alpha)}{\mu(D_\beta)} \log \frac{\mu(D_\beta \cap C_\alpha)}{\mu(D_\beta)}. \end{aligned}$$

Next we collect a number of basic properties.

5.9 Proposition: *Let $(X, \mathcal{A}, \mu)$ be a probability space and let $\mathcal{P} = \{C_\alpha \mid \alpha \in I\}$, $\mathcal{Q} = \{D_\beta \mid \beta \in J\}$ and $\mathcal{R} = \{E_\gamma \mid \gamma \in K\}$ be finite or countable measurable partitions of X. Then the following statements hold:*

(i) $0 \leq H(\mathcal{P}|\mathcal{Q}) \leq H(\mathcal{P})$.

(ii) $H(\mathcal{P}|\mathcal{Q}) = H(\mathcal{P})$ *iff* $\mathcal{P}$ *and* $\mathcal{Q}$ *are independent.*

(iii) $H(\mathcal{P}|\mathcal{Q}) = 0$ *iff* $\mathcal{Q}$ *is finer than* $\mathcal{P}$.

(iv) *If* $\mathcal{R} \geq \mathcal{Q}$, *then* $H(\mathcal{P}|\mathcal{R}) \leq H(\mathcal{P}|\mathcal{Q})$.

Proof:

(i) $\varphi(x) = x \log x$ is a convex function. Hence,

$$\begin{aligned} 0 \quad &\leq \quad H(\mathcal{P}|\mathcal{Q}) = -\sum_{\beta \in J} \mu(D_\beta) \sum_{\alpha \in I} \varphi(\mu(C_\alpha|D_\beta)) \\ &= \quad -\sum_{\alpha \in I} \sum_{\beta \in J} \mu(D_\beta)\varphi(\mu(C_\alpha|D_\beta)) \\ &\overset{\varphi \text{ convex}}{\leq} \quad -\sum_{\alpha \in I} \varphi\left(\sum_{\beta \in J} \mu(D_\beta)\frac{\mu(C_\alpha \cap D_\beta)}{\mu(D_\beta)}\right) \\ &= \quad -\sum_{\alpha \in I} \varphi(\mu(C_\alpha)) = H(\mathcal{P}). \end{aligned}$$

(ii) Recall $\varphi(x) < 0$ iff $x \in (0,1)$. $H(\mathcal{P}|\mathcal{Q}) = 0$ implies for every β ($\mu(D_\beta) > 0$): $\varphi(\mu(C_\alpha|D_\beta)) = 0$, and consequently $\mu(C_\alpha|D_\beta) \in \{0,1\}$. Hence,

$$\mu(C_\alpha \cap D_\beta) = 0 \;\text{ or }\; \mu(C_\alpha \cap D_\beta) = \mu(D_\beta),$$

i.e., $C_\alpha \cap D_\beta = \emptyset (\bmod 0)$ or $D_\beta \subset C_\alpha (\bmod 0)$. Thus, $\mathcal{Q} \geq \mathcal{P}$ (mod 0). The converse is obvious.

(iii) If $H(\mathcal{P}|\mathcal{Q}) = H(\mathcal{P})$, then equality must hold in the inequality used for (i), then equality must hold for every summand, i.e.,

$$\begin{aligned} \varphi(\mu(C_\alpha)) &= \varphi\left(\sum_{\beta \in J,\; \mu(D_\beta) > 0} \mu(D_\beta)\mu(C_\alpha|D_\beta)\right) \\ &= \sum_{\beta \in J,\; \mu(D_\beta) > 0} \mu(D_\beta)\varphi(\mu(C_\alpha|D_\beta)). \end{aligned}$$

By strict convexity of φ this implies that $\mu(C_\alpha|D_\beta)$ must be independent of β and hence $\mu(C_\alpha|D_\beta) = \mu(C_\alpha)$. Hence, $\mathcal{P}$ and $\mathcal{Q}$ are independent. The converse is obvious.

(iv) Suppose $\mathcal{R}$ is a refinement of $\mathcal{Q}$, $\mathcal{R} \geq \mathcal{Q}$ (mod 0). Consider, for $D \in \mathcal{Q}$, the conditional measure

$$\mu_D(\cdot) = \mu(\cdot|D).$$

Then

$$H_{\mu_D}(\mathcal{P}|\mathcal{R}) \leq H_{\mu_D}(\mathcal{P}) \text{ by (i).}$$

Now

$$\begin{aligned} H(\mathcal{P}|\mathcal{R}) &= H(\mathcal{P}|\mathcal{Q} \vee \mathcal{R}) = -\sum_\gamma \mu(E_\gamma) \sum_\alpha \mu(C_\alpha|E_\gamma) \log \mu(C_\alpha|E_\gamma) \\ &= -\sum_\beta \underbrace{\sum_{\gamma:\ E_\gamma \subset D_\beta} \mu(E_\gamma)}_{=\mu(D_\beta)} \underbrace{\sum_\alpha \frac{\mu(C_\alpha \cap D_\beta \cap E_\gamma)}{\mu(D_\beta \cap E_\gamma)} \log \frac{\mu(C_\alpha \cap D_\beta \cap E_\gamma)}{\mu(D_\beta \cap E_\gamma)}}_{=-H_{\mu_{D_\beta}}(\mathcal{P}|\mathcal{R}) \geq -H_{\mu_{D_\beta}}(\mathcal{P})} \\ &\leq \sum_\beta \mu(D_\beta) H_{\mu_{D_\beta}}(\mathcal{P}) = H(\mathcal{P}|\mathcal{Q}). \end{aligned}$$

□

5.10 Proposition: *Under the assumptions of Proposition 5.9 the following statements hold:*

(i) $H(\mathcal{P} \vee \mathcal{Q}|\mathcal{R}) = H(\mathcal{P}|\mathcal{R}) + H(\mathcal{Q}|\mathcal{P} \vee \mathcal{R})$. *In particular,* $H(\mathcal{P} \vee \mathcal{Q}) = H(\mathcal{P}) + H(\mathcal{Q}|\mathcal{P})$.

(ii) $H(\mathcal{P} \vee \mathcal{Q}|\mathcal{R}) \leq H(\mathcal{P}|\mathcal{R}) + H(\mathcal{Q}|\mathcal{R})$. *In particular,* $H(\mathcal{P} \vee \mathcal{Q}) \leq H(\mathcal{P}) + H(\mathcal{Q})$.

(iii) $H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{R}) \geq H(\mathcal{P}|\mathcal{R})$.

(iv) *If λ is another probability measure on X, then for every measurable partition $\mathcal{P}$ and for every $p \in [0,1]$*

$$pH_\mu(\mathcal{P}) + (1-p)H_\lambda(\mathcal{P}) \leq H_{p\mu+(1-p)\lambda}(\mathcal{P}).$$

Proof:

(i) We have

$$\begin{aligned}
H(\mathcal{P}\vee\mathcal{Q}|\mathcal{R}) &= -\sum_{\gamma}\mu(E_\gamma)\sum_{\alpha,\beta}\mu(C_\alpha\cap D_\beta|E_\gamma)\log\mu(C_\alpha\cap D_\beta|E_\gamma)\\
&= -\sum_{\alpha,\beta,\gamma}\mu(C_\alpha\cap D_\beta\cap E_\gamma)\log\frac{\mu(C_\alpha\cap D_\beta\cap E_\gamma)}{\mu(E_\gamma)}\\
&= -\sum_{\alpha,\beta,\gamma}\mu(C_\alpha\cap D_\beta\cap E_\gamma)\log\frac{\mu(C_\alpha\cap E_\gamma)}{\mu(E_\gamma)}\\
&\quad -\sum_{\alpha,\beta,\gamma}\mu(C_\alpha\cap D_\beta\cap E_\gamma)\log\frac{\mu(C_\alpha\cap D_\beta\cap E_\gamma)}{\mu(C_\alpha\cap E_\gamma)}\\
&= -\sum_{\alpha,\gamma}\mu(C_\alpha\cap E_\gamma)\log\frac{\mu(C_\alpha\cap E_\gamma)}{\mu(E_\gamma)}+H(\mathcal{Q}|\mathcal{P}\vee\mathcal{R})
\end{aligned}$$

and

$$\begin{aligned}
H(\mathcal{Q}|\mathcal{P}\vee\mathcal{R}) &= -\sum_{\alpha,\gamma}\mu(C_\alpha\cap E_\gamma)\sum_{\beta}\frac{\mu(C_\alpha\cap E_\gamma\cap D_\beta)}{\mu(C_\alpha\cap E_\gamma)}\log\frac{\mu(C_\alpha\cap E_\gamma\cap D_\beta)}{\mu(C_\alpha\cap E_\gamma)},\\
H(\mathcal{P}|\mathcal{R}) &= -\sum_{\gamma}\mu(E_\gamma)\sum_{\alpha}\frac{\mu(C_\alpha\cap E_\gamma)}{\mu(E_\gamma)}\log\frac{\mu(C_\alpha\cap E_\gamma)}{\mu(E_\gamma)}.
\end{aligned}$$

Hence,

$$H(\mathcal{P}\vee\mathcal{Q}|\mathcal{R}) = H(\mathcal{P}|\mathcal{R})+H(\mathcal{Q}|\mathcal{P}\vee\mathcal{R}).$$

(ii) This follows from (i):

$$H(\mathcal{P}\vee\mathcal{Q}|\mathcal{R}) = H(\mathcal{P}|\mathcal{R})+H(\mathcal{Q}|\mathcal{P}\vee\mathcal{R})\leq H(\mathcal{P}|\mathcal{R})+H(\mathcal{Q}|\mathcal{R}),$$

since $\mathcal{P}\vee\mathcal{R}\geq\mathcal{R}$.

(iii) Note that by (i) and (ii)

$$H(\mathcal{R}|\mathcal{P}\vee\mathcal{Q}) \overset{\text{(i)}}{=} H(\mathcal{P}\vee\mathcal{R}|\mathcal{Q})-H(\mathcal{P}|\mathcal{Q}) \overset{\text{(ii)}}{\leq} H(\mathcal{R}|\mathcal{Q}).$$

Using (i) several times we find

$$\begin{aligned}
H(\mathcal{P}|\mathcal{Q})+H(\mathcal{Q}|\mathcal{R}) &= H(\mathcal{P}\vee\mathcal{Q})-H(\mathcal{Q})+H(\mathcal{R}\vee\mathcal{Q})-H(\mathcal{R})\\
&\overset{\text{(i)}}{=} H(\mathcal{P}\vee\mathcal{Q})+H(\mathcal{R}|\mathcal{Q})-H(\mathcal{R})\\
&\overset{\text{(i)}}{=} H(\mathcal{P}\vee\mathcal{Q}\vee\mathcal{R})-H(\mathcal{R}|\mathcal{P}\vee\mathcal{Q})+H(\mathcal{R}|\mathcal{Q})-H(\mathcal{R})\\
&\geq H(\mathcal{P}\vee\mathcal{Q}\vee\mathcal{R})-H(\mathcal{R})\\
&\geq H(\mathcal{P}\vee\mathcal{R})-H(\mathcal{R}) \overset{\text{(i)}}{=} H(\mathcal{P}|\mathcal{R}).
\end{aligned}$$

(iv) This follows from convexity of φ.

□

5.11 Corollary: *For two finite measurable partitions $\mathcal{P}$ and $\mathcal{Q}$ let*

$$d_R(\mathcal{P},\mathcal{Q}) := H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{P}).$$

Then d_R is a metric on the set of all equivalence classes (mod 0) of finite measurable partitions of X. It is called the **Rokhlin metric**.

Proof: $d_R(\mathcal{P},\mathcal{Q}) \geq 0$ is clear. If $d_R(\mathcal{P},\mathcal{Q}) = 0$, then $H(\mathcal{P}|\mathcal{Q}) = 0$ and $H(\mathcal{Q}|\mathcal{P}) = 0$. Hence, $\mathcal{Q} \leq \mathcal{P}$ and $\mathcal{P} \leq \mathcal{Q}$, which implies $\mathcal{P} = \mathcal{Q}$ (mod 0). Symmetry is clear by definition. Finally, the triangle inequality follows from Proposition 5.10 (iii):

$$\begin{aligned} d_R(\mathcal{P},\mathcal{R}) &= H(\mathcal{P}|\mathcal{R}) + H(\mathcal{R}|\mathcal{P}) \\ &\leq H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{R}) + H(\mathcal{R}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{P}) \\ &= d_R(\mathcal{P},\mathcal{Q}) + d_R(\mathcal{Q},\mathcal{R}). \end{aligned}$$

□

5.3 Properties of Entropy

We analyze properties of the entropy $h(T,\mathcal{P})$ as a function of the partition $\mathcal{P}$.

5.12 Proposition: *Let $T : (X,\mu) \circlearrowleft$ be a measure-preserving map on a probability space and let $\mathcal{P} = \{C_\alpha \mid \alpha \in I\}$ and $\mathcal{Q}$ be finite measurable partitions of X. Then the following statements hold:*

(i) $0 \leq \limsup_{n\to\infty} \left(-\frac{1}{n} \log \sup_{C\in\mathcal{P}_n} \mu(C)\right) \leq h(T,\mathcal{P}) \leq H(\mathcal{P})$.

(ii) $h(T,\mathcal{P}\vee\mathcal{Q}) \leq h(T,\mathcal{P}) + h(T,\mathcal{Q})$.

(iii) $h(T,\mathcal{P}) \leq h(T,\mathcal{Q}) + H(\mathcal{P}|\mathcal{Q})$. *In particular, if $\mathcal{Q}$ is a refinement of $\mathcal{P}$ ($\mathcal{P} \leq \mathcal{Q}$), then $h(T,\mathcal{P}) \leq h(T,\mathcal{Q})$.*

(iv) $|h(T,\mathcal{P}) - h(T,\mathcal{Q})| \leq H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{P}) = d_R(\mathcal{P},\mathcal{Q})$ *(the* Rokhlin Inequality*).*[6]

Proof:

(i) The first inequality is obvious, the last follows from Ex. 1 on Sheet 9:

$$h(T,\mathcal{P}) = \lim_{n\to\infty} H(\mathcal{P}|T^{-1}\mathcal{P}_n) \overset{\forall n}{\leq} H(\mathcal{P}|T^{-1}\mathcal{P}_n) \overset{\text{Prop. 5.9 (i)}}{\leq} H(\mathcal{P}).$$

[6]This shows that $h(T,\cdot)$ is a Lipschitz continuous function with Lipschitz constant 1 on the space of finite measurable partitions with the Rokhlin metric.

The middle inequality follows, since for every partition $\mathcal{R} = \{E_\gamma \mid \gamma \in K\}$

$$-\log \sup_\gamma \mu(E_\gamma) = \inf_{x \in X} \underbrace{I_{\mathcal{R}}(x)}_{=-\log \mu(E_{\mathcal{P}}(x))} .$$

Then

$$H(\mathcal{R}) = \int_X I_{\mathcal{R}} d\mu \geq -\log \sup_\gamma \mu(E_\gamma).$$

This shows that for every $n \geq 1$

$$-\log \sup_{C \in \mathcal{P}_n} \mu(C) \leq H(\mathcal{P}_n).$$

Hence,

$$h(T, \mathcal{P}) = \lim_{n \to \infty} \frac{1}{n} H(\mathcal{P}_n) \geq \limsup_{n \to \infty} \left(-\frac{1}{n} \log \sup_{C \in \mathcal{P}_n} \mu(C) \right).$$

(ii) We have

$$(\mathcal{P} \vee \mathcal{Q})_n = (\mathcal{P} \vee \mathcal{Q}) \vee T^{-1}(\mathcal{P} \vee \mathcal{Q}) \vee \cdots \vee T^{-(n-1)}(\mathcal{P} \vee \mathcal{Q}) = \mathcal{P}_n \vee \mathcal{Q}_n.$$

Hence, by Proposition 5.10 (i)

$$H((\mathcal{P} \vee \mathcal{Q})_n) = H(\mathcal{P}_n \vee \mathcal{Q}_n) = H(\mathcal{P}_n) + H(\mathcal{Q}_n | \mathcal{P}_n)$$

and by Proposition 5.9 (i)

$$\begin{aligned} h(T, \mathcal{P} \vee \mathcal{Q}) &= \lim_{n \to \infty} \frac{1}{n} H((\mathcal{P} \vee \mathcal{Q})_n) \\ &= \lim_{n \to \infty} \frac{1}{n} \left[H(\mathcal{P}_n) + H(\mathcal{Q}_n | \mathcal{P}_n) \right] \\ &\leq \lim_{n \to \infty} \frac{1}{n} H(\mathcal{P}_n) + \lim_{n \to \infty} \frac{1}{n} H(\mathcal{Q}_n) \\ &= h(T, \mathcal{P}) + h(T, \mathcal{Q}). \end{aligned}$$

(iii) The particular case is clear, since $\mathcal{P} \leq \mathcal{Q}$ implies $H(\mathcal{P}|\mathcal{Q}) = 0$ by Proposition 5.9 (iii). Further we obtain

$$H(\mathcal{P}_n) \leq H(\mathcal{P}_n \vee \mathcal{Q}_n) = H(\mathcal{Q}_n) + H(\mathcal{P}_n | \mathcal{Q}_n).$$

Note that

$$\mathcal{P}_n = \mathcal{P} \vee T^{-1}\mathcal{P} \vee \ldots \vee T^{-(n-1)}\mathcal{P} = \mathcal{P} \vee T^{-1}\mathcal{P}_{n-1}.$$

Hence,

$$\begin{aligned} H(\mathcal{P}_n|\mathcal{Q}_n) &= H(\mathcal{P} \vee T^{-1}\mathcal{P}_{n-1}|\mathcal{Q}_n) \\ &= H(\mathcal{P}|\mathcal{Q}_n) + H(T^{-1}\mathcal{P}_{n-1}|\mathcal{P} \vee \mathcal{Q}_n) \\ &\le H(\mathcal{P}|\mathcal{Q}) + H(T^{-1}\mathcal{P}_{n-1}|\mathcal{Q}_n) \\ &\le H(\mathcal{P}|\mathcal{Q}) + \underbrace{H(T^{-1}\mathcal{P}|T^{-1}\mathcal{Q})}_{=H(\mathcal{P}|\mathcal{Q}) \text{ by invariance}} + H(T^{-2}\mathcal{P}_{n-2}|\mathcal{Q}_n) \\ &\le nH(\mathcal{P}|\mathcal{Q}). \end{aligned}$$

The last inequality follows inductively. Thus,

$$\begin{aligned} h(T,\mathcal{P}) &= \lim_{n\to\infty} \frac{1}{n} H(\mathcal{P}_n) \le \lim_{n\to\infty} \frac{1}{n} \left[H(\mathcal{Q}_n) + nH(\mathcal{P}|\mathcal{Q}) \right] \\ &= h(T,\mathcal{Q}) + H(\mathcal{P}|\mathcal{Q}). \end{aligned}$$

(iv) This follows immediately from (iii).

□

5.13 Proposition: *Under the assumptions of Proposition 5.12 the following statements hold:*

(i) $h(T, T^{-1}\mathcal{P}) = h(T,\mathcal{P})$ *and if* T *is invertible,* $h(T,\mathcal{P}) = h(T, T\mathcal{P})$.

(ii) $h(T,\mathcal{P}) = h(T, \bigvee_{i=0}^{k} T^{-i}\mathcal{P})$ *for all* $k \in \mathbb{N}$*, and if* T *is invertible,* $h(T,\mathcal{P}) = h(T, \bigvee_{i=-k}^{k} T^{i}\mathcal{P})$ *for all* $k \in \mathbb{N}$.

Proof:

(i) This follows from the invariance property, since

$$\begin{aligned} H((T^{-1}\mathcal{P})_n) &= H(T^{-1}\mathcal{P} \vee T^{-2}\mathcal{P} \vee \cdots \vee T^{-n}\mathcal{P}) \\ &= H(\mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-(n-1)}\mathcal{P}) = H(\mathcal{P}_n) \end{aligned}$$

and

$$h(T,\mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{P}_n) = \lim_{n\to\infty} H((T^{-1}\mathcal{P})_n) = h(T, T^{-}\mathcal{P}).$$

For invertible T the proof works analogously.

(ii) Observe that

$$\begin{aligned} \left(\bigvee_{i=0}^{k} T^{-i}\mathcal{P} \right)_n &= \left(\mathcal{P} \vee T^{-1}\mathcal{P} \vee \cdots \vee T^{-k}\mathcal{P} \right)_n \\ &= \mathcal{P} \vee \cdots \vee T^{-(n+k-1)}\mathcal{P} = \mathcal{P}_{n+k}, \end{aligned}$$

and hence,

$$h\left(T, \bigvee_{i=0}^{k} T^{-i}\mathcal{P}\right) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{P}_{n+k}) = \lim_{n\to\infty} \underbrace{\frac{n+k}{n}}_{\to 1} \frac{1}{n+k} H(\mathcal{P}_{n+k}) = \lim_{n\to\infty} \frac{1}{n+k} H(\mathcal{P}_{n+k}) = h(T, \mathcal{P}).$$

Again, for invertible T the argument is completely analogous.

□

Recall: The entropy of T is $h(T) = \sup_{\mathcal{P}} h(T, \mathcal{P})$. We want: not all finite measurable partitions, but a subfamily.

5.14 Definition: *A family $\hat{\mathcal{P}}$ of finite measurable partitions is called* **sufficient***, if*

(i) *for noninvertible T the partitions $\mathcal{Q}$ with*

$$\mathcal{Q} \le \bigvee_{i=0}^{k} T^{-i}\mathcal{P} \ \text{ for some } k \in \mathbb{N} \text{ and } \mathcal{P} \in \hat{\mathcal{P}}$$

form a dense subset of the set of all finite measurable partitions with respect to the Rokhlin metric.

(ii) *for invertible T the same holds for the partitions $\mathcal{Q}$ with*

$$\mathcal{Q} \le \bigvee_{i=-k}^{k} T^{-i}\mathcal{P} \ \text{ for some } k \in \mathbb{N} \text{ and } \mathcal{P} \in \hat{\mathcal{P}}.$$

5.15 Theorem: *For every sufficient family $\hat{\mathcal{P}}$ it holds that*

$$h_\mu(T) = \sup_{\mathcal{P}\in\hat{\mathcal{P}}} h(T, \mathcal{P}).$$

Proof: Let T be noninvertible. Let $\mathcal{R}$ be an arbitrary finite measurable partition. Fix $\varepsilon > 0$ and find $\mathcal{P} \in \hat{\mathcal{P}}$ and $k \in \mathbb{N}$ such that for some partition $\mathcal{Q}$ with

$$\mathcal{Q} \le \bigvee_{i=0}^{k} T^{-i}\mathcal{P}$$

one has

$$d_R(\mathcal{R}, \mathcal{Q}) = H(\mathcal{R}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{R}) < \varepsilon.$$

Then, using the Rokhlin Inequality,

$$\begin{aligned} h(T,\mathcal{R}) &\le h(T,\mathcal{Q}) + d_R(\mathcal{R},\mathcal{Q}) \le h(T,\mathcal{Q}) + \varepsilon \\ &\le h\left(T, \bigvee_{i=0}^{k} T^{-i}\mathcal{P}\right) + \varepsilon = h(T,\mathcal{P}) + \varepsilon. \end{aligned}$$

The last equality follows from Proposition 5.13 (ii). The proof for invertible T works analogously. □

5.16 Proposition: *Assume that μ is a non-atomic Borel measure on a compact metric space X, i.e., it is defined on the σ-algebra generated by the open sets and*

$$\mu(\{x\}) = 0 \quad \textit{for all } x \in X.$$

Then every family $(\mathcal{P}^k)_{k\in\mathbb{N}}$ of finite measurable partitions with

$$\max_{C\in\mathcal{P}^k} \operatorname{diam} C \xrightarrow{k\to\infty} 0$$

is a sufficient family.

Proof: Let $\mathcal{R} = \{E_\gamma \mid \gamma \in K\}$, $\mu(E_\gamma) > 0$, be a finite measurable partition of X. Let $\varepsilon > 0$. We show that there is $k \in \mathbb{N}$ such that for a finite measurable partition $\mathcal{Q} \le \mathcal{P}^k$

$$d_R(\mathcal{R},\mathcal{Q}) = H(\mathcal{R}|\mathcal{Q}) + H(\mathcal{Q}|\mathcal{R}) < \varepsilon.$$

Such a partition $\mathcal{Q}$ consists of (finite) unions of elements of $\mathcal{P}^k$. Let $E_\gamma \in \mathcal{R}$. Choosing k large enough one can let

$$\mu(E_\gamma \cap D_\beta)$$

be arbitrarily close to $\mu(D_\beta)$ for some $\mathcal{Q} \le \mathcal{P}^k$, $\mathcal{Q} = \{D_\beta \mid \beta \in J\}$ (Here regularity of Borel measures is used!) Thus,

$$\frac{\mu(E_\gamma \cap D_\beta)}{\mu(D_\beta)}$$

can be made arbitrarily close to 1. Since $\phi(x) = x \log x$ is continuous with $\phi(1) = 0$, one can make

$$\phi\left(\frac{\mu(E_\gamma \cap D_\beta)}{\mu(D_\beta)}\right)$$

be arbitrarily close to 0, for each of the finitely many E_γ. Thus, choosing k large enough,

$$\begin{aligned} H(\mathcal{R}|\mathcal{Q}) &= -\sum_\beta \mu(D_\beta) \sum_\gamma \frac{\mu(E_\gamma \cap D_\beta)}{\mu(D_\beta)} \log \frac{\mu(E_\gamma \cap D_\beta)}{\mu(D_\beta)} \\ &= -\sum_\beta \mu(D_\beta) \sum_\gamma \phi\left(\frac{\mu(E_\gamma \cap D_\beta)}{\mu(D_\beta)}\right) < \frac{\varepsilon}{2}. \end{aligned}$$

Similarly, one can get $H(\mathcal{Q}|\mathcal{R}) < \frac{\varepsilon}{2}$. $\square$

5.17 Definition: *A partition $\mathcal{P}$ is called a* **generator** *if $\hat{\mathcal{P}} = \{\mathcal{P}\}$ is a sufficient family.*

Thus, a generator $\mathcal{P}$ has the property that the partitions $\mathcal{Q}$ with

$$\mathcal{Q} \leq \bigvee_{i=0}^{k} T^{-i}\mathcal{P} = \mathcal{P}_k$$

are dense in the set of all finite partitions.

5.18 Corollary: *If $\mathcal{P}$ is a generator for T, then $h_\mu(T) = h_\mu(T, \mathcal{P})$.*

5.19 Proposition:

(i) *Let $S : (Y, \nu) \circlearrowleft$ be a factor of $T : (X, \mu) \circlearrowleft$ (i.e., S and T are measure-preserving and there is a measure-preserving $\phi : X \to Y$ with $\Phi \circ T = S \circ \Phi$.) Then $h_\nu(S) \leq h_\mu(T)$.*

(ii) *If A is invariant for T with $\mu(A) > 0$, then*

$$h_\mu(T) = \mu(A)h_{\mu_A}(T) + \mu(X \backslash A)h_{\mu_{X\backslash A}}(T),$$

where μ_A and $\mu_{X\backslash A}$ are the conditional measures on A and $X\backslash A$, respectively.

Proof:

(i) For any measurable partition $\mathcal{Q}$ of Y

$$\Phi^{-1}\mathcal{Q} = \left\{\Phi^{-1}(D) \mid D \in \mathcal{Q}\right\}$$

is a measurable partition of X and, since Φ is measure-preserving,

$$H_\mu(\mathcal{R}^{-1}\mathcal{Q}) = H_\nu(\mathcal{Q}), \;\; h_\mu(T, \mathcal{R}^{-1}\mathcal{Q}) = h_\nu(S, \mathcal{Q}).$$

Thus,

$$h_\mu(T) = \sup_{\mathcal{P}} h_\mu(T, \mathcal{P}) \geq \sup_{\mathcal{Q}} h_\mu(T, \Phi^{-1}\mathcal{Q}) = \sup_{\mathcal{Q}} h_\nu(S, \mathcal{Q}) = h_\nu(S).$$

(ii) Let $\mathcal{P}$ be a measurable partition of X and define the partition $\mathcal{Q}$ by $\mathcal{Q} := \{A, X\backslash A\}$. We may replace $\mathcal{P}$ by $\mathcal{P} \vee \mathcal{Q}$ (since we are interested in $\sup_{\mathcal{P}} h(T, \mathcal{P})$). Hence, $\mathcal{P} \geq \mathcal{Q}$ and $T^{-j}(A) \subset A$. Thus,

$$H(\mathcal{P}_n) \;=\; -\sum_{D \in \mathcal{P}_n} \mu(D) \log \mu(D)$$

$$
\begin{aligned}
&= -\sum_{\substack{D\in\mathcal{P}_n\\ D\subset A}} \mu(D)\log\mu(D) + \sum_{\substack{D\in\mathcal{P}_n\\ D\subset X\setminus A}} \mu(D)\log\mu(D)\\
&= -\mu(A)\sum_{\substack{D\in\mathcal{P}_n\\ D\subset A}} \mu_A(D)\log\mu_A(D)\\
&- \mu(X\setminus A)\sum_{\substack{D\in\mathcal{P}_n\\ D\subset X\setminus A}} \mu_{X\setminus A}(D)\log\mu_{X\setminus A}(D)\\
&- \left[\mu(A)\log(A) + \mu(X\setminus A)\log\mu(X\setminus A)\right]\\
&= \mu(A)H_{\mu_A}(\mathcal{P}_n) + \mu(X\setminus A)H_{\mu_{X\setminus A}}(\mathcal{P}_n)\\
&- \left[\mu(A)\log(A) + \mu(X\setminus A)\log\mu(X\setminus A)\right].
\end{aligned}
$$

Multiplying both sides by $\frac{1}{n}$ and letting n go to infinity, yields the assertion.

□

5.4 Examples of Calculation of Entropy

One will expect that the rotation $R : [0,1) \circlearrowleft$, $x \mapsto x+\alpha \pmod 1$, has entropy zero. The easiest way to see this, is to take the family

$$\hat{\mathcal{P}} = \left\{\mathcal{P}^{(N)} \;:\; N\in\mathbb{N}\right\}$$

of partitions into N equal intervals. This family is sufficient by Proposition 5.16. The joint partition

$$\left(\mathcal{P}^{(N)}\right)_n = \bigvee_{i=0}^{n-1} R^{-i}\mathcal{P}^{(N)}$$

has not more than Nn elements (exactly that many for α irrational.) Hence, by Lemma 5.1

$$H\left(\bigvee_{i=0}^{n-1} R^{-i}\mathcal{P}_n\right) \le \log Nn = \log N + \log n.$$

Thus,

$$h\left(R,\mathcal{P}^{(n)}\right) \le \lim_{n\to\infty}\frac{1}{n}\left(\log N + \log n\right) = 0.$$

Analogously one shows that the entropy for the rotation on the 2-torus is zero.

Now: Entropy of Shift Transformations: $\mathcal{A} = \{1,2,\ldots,k\}$, $X = \prod_1^\infty \mathcal{A}$ (sequences of symbols), $T : X \circlearrowleft$, $(x_1,x_2,\ldots) \mapsto (x_2,x_3,\ldots)$. The standard partition of X is $\mathcal{P}_0$, given by

$$[s_i] = \{(x_1,x_2,\ldots) \mid x_1 = s_i\}\,,\;\; s_i = i,\;\; i = 1,\ldots,k.$$

Then $T^{-1}[s_i] = \{(x_1, x_2, \dots) \mid x_2 = s_i\}$. Hence,

$$\mathcal{P}_n = \mathcal{P}_0 \vee T^{-1}\mathcal{P}_0 \vee \cdots \vee T^{-(n-1)}\mathcal{P}_0$$

consists of all cylinder sets of the form

$$[s_1, \dots, s_n] = \{x \in X \mid x_1 = s_1,\ x_2 = s_2,\ \dots,\ x_n = s_n\}$$

for $s_j \in \mathcal{A}$. Suppose the symbol s_i has probability $p_i > 0$, $\sum_{i=1}^k p_i = 1$. Define

$$\mu\left([s_1, \dots, s_n]_t^{t+n-1}\right) = p_1 \cdots p_n.$$

This Bernoulli measure is shift-invariant. What is the entropy of this measure?

$$h_\mu(T) = \sup_{\mathcal{P}} H(T, \mathcal{P}) = \sup_{\mathcal{P}} \lim_{n\to\infty} H(\mathcal{P}_n).$$

Claim: The standard partition is generating.

Proof: See Exercise 1 on sheet 10. □

5.20 Theorem: *The entropy of the* $(p_1, \dots, p_k)$*-Bernoulli Shift is given by*

$$h(T) = -\sum_{i=1}^k p_i \log p_i.$$

Proof: We have to compute the entropy

$$H(\mathcal{P}_{0,n}) = H\left(\mathcal{P}_0 \vee T^{-1}\mathcal{P}_0 \vee \cdots \vee T^{-(n-1)}\mathcal{P}_0\right).$$

The partitions $T^{-i}\mathcal{P}_0$ and $T^{-j}\mathcal{P}_0$ with $i \neq j$ are independent. Hence, Proposition 5.9 (ii) and 5.10 (i) imply

$$H(\mathcal{P}_{0,n}) = H(\mathcal{P}_0) + \underbrace{H(T^{-1}\mathcal{P}_0)}_{=H(\mathcal{P}_0)} + \cdots + \underbrace{H(T^{-(n-1)}\mathcal{P}_0)}_{=H(\mathcal{P}_0)} = nH(\mathcal{P}_0),$$

since $H(T^{-1}\mathcal{P}_0) = H(\mathcal{P}_0)$ by invariance of μ. Now

$$H(\mathcal{P}_0) = -\log_{i=1}^k \mu([s_i]) \log \mu([s_i]) = -\log_{i=1}^k p_i \log p_i$$

and hence,

$$h_\mu(T) = \lim_{n\to\infty} \frac{1}{n} \underbrace{H(\mathcal{P}_{0,n})}_{nH(\mathcal{P}_0)} = H(\mathcal{P}_0) = -\sum_{i=1}^k p_i \log p_i.$$

□

Next we compute the entropy of Markov shifts. Let P be a stochastic matrix ($p_{ij} \geq 0$, $\sum_j p_{ij} = 1$). Assume that P is irreducible, i.e., for all (i,j) there is $n \in \mathbb{N}$ such that $(P^m)_{ij} > 0$. Then Perron-Frobenius implies the existence of a left eigenvector π for the eigenvalue 1 with $\pi_i > 0$ for $i = 1, \ldots, k$. Define a Markov measure μ on the cylinder sets by

$$\mu\left([s_1, \ldots, s_n]_t^{t+n-1}\right) = \pi_{s_1} p_{s_1 s_2} \cdots p_{s_{n-1} s_n}.$$

5.21 Theorem: *The entropy of the Markov shift, given by the matrix P, is*

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

Proof: The standard partition $\mathcal{P}_0$ again is a generator. An element of $\mathcal{P}_{0,n+1}$ is given by $[s_0, \ldots, s_n]$. It has measure

$$\mu\left([s_0, \ldots, s_n]\right) = \pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-1} s_n}.$$

In the following we use the abbreviation $\phi(x) = x \log x$. We obtain

$$\begin{aligned} H(\mathcal{P}_{n+1}) &= -\sum_{s_0,\ldots,s_n=1}^{k} \mu\left([s_0, \ldots, s_n]\right) \log \mu\left([s_0, \ldots, s_n]\right) \\ &= -\sum_{s_0,\ldots,s_n} \pi_{s_0} p_{s_0 s_1} \cdots \left(p_{s_{n-1} s_n} \log\left(\pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-2} s_{n-1}}\right)\right. \\ &+ \left. p_{s_{n-1} s_n} \log p_{s_{n-1} s_n}\right) \\ &= -\sum_{s_0,\ldots,s_{n-1}} \underbrace{\left(\sum_{s_n} p_{s_{n-1} s_n}\right)}_{=1} \phi\left(\pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-2} s_{n-1}}\right) \\ &- \sum_{s_{n-1}, s_n} \left(\sum_{s_0,\ldots,s_{n-2}} \pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-2} s_{n-1}}\right) \phi\left(p_{s_{n-1} s_n}\right). \end{aligned}$$

We have

$$\sum_{s_0,\ldots,s_{n-2}} \pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-2} s_{n-1}} = \pi_{s_{n-1}},$$

since this is the probability to go from some symbol s_0 over some sequence of symbols in $(n-1)$ steps to the symbol s_{n-1}. Hence,

$$H(\mathcal{P}_{n+1}) = \underbrace{-\sum_{s_0,\ldots,s_{n-1}} \phi\left(\pi_{s_0} p_{s_0 s_1} \cdots p_{s_{n-2} s_{n-1}}\right)}_{=H(\mathcal{P}_n)} - \sum_{s_{n-1}, s_n} \pi_{s_{n-1}} \phi\left(p_{s_{n-1} s_n}\right).$$

By induction we obtain

$$H(\mathcal{P}_{n+1}) = H(\mathcal{P}_0) - n \sum_{i,j} \pi_i \phi(p_{ij}) = -\sum_i \pi_i \log \pi_i - n \sum_{i,j} \pi_i p_{ij} \log p_{ij}.$$

This implies the assertion. □

5.22 Remark: The $(\frac{1}{2}, \frac{1}{2})$-Bernoulli Shift (shift on 2 symbols) with $p_0 = p_1 = \frac{1}{2}$ is isomorphic to the doubling map $Tx = 2x \pmod 1$ on $[0, 1)$. Entropy is invariant under isomorphisms. Hence, the entropy of the doubling map is

$$h(T) = -\sum_{i=1}^{2} p_i \log p_i = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = \log 2.$$

5.23 Remark: The main result, which started interest in entropy is due to Ornstein (1970). He could show: If two Bernoulli Shifts have the same entropy, then they are isomorphic.

References

[1] G. H. Choe, *Computational Ergodic Theory*, Springer (2005).

[2] P. Halmos, *Measure Theory*, Springer (1974).

[3] G. K. Pedersen, *Analysis Now*, Springer (1989).

[4] G. Bachman, L. Narici, *Functional Analysis*, DoverPublications.com (2000).

[5] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press (1999).

[6] F. R. Gantmacher, *Matrizentheorie*, Springer (1986).

[7] J. Elstrodt, *Maß- und Integrationstheorie*, Springer (2004).

[8] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge Univ. Press (1995).